Studies in Comput

Weisi Lin
Dacheng Tao
Janusz Kacprzyk
Zhu Li
Ebroul Izquierdo
Haohong Wang (

Multimed

Weisi Lin, Dacheng Tao, Janusz Kacprzyk, Zhu Li, Ebroul Izquierdo, and Haohong Wang (Eds.)

Multimedia Analysis, Processing and Communications

Studies in Computational Intelligence, Volume 346

Editor-in-Chief
Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Vol. 322. Bruno Baruque and Emilio Corchado (Eds.)
Fusion Methods for Unsupervised Learning Ensembles, 2010
ISBN 978-3-642-16204-6

Vol. 324. Alessandro Soro, Vargiu Eloisa, Giuliano Armano, and Gavino Paddeu (Eds.)
Information Retrieval and Mining in Distributed Environments, 2010
ISBN 978-3-642-16088-2

Vol. 325. Quan Bai and Naoki Fukuta (Eds.)
Advances in Practical Multi-Agent Systems, 2010
ISBN 978-3-642-16097-4

Vol. 326. Sheryl Brahnam and Lakhmi C. Jain (Eds.)
Advanced Computational Intelligence Paradigms in Healthcare 5, 2010
ISBN 978-3-642-16094-3

Vol. 327. Slawomir Wiak and Ewa Napieralska-Juszczak (Eds.)
Computational Methods for the Innovative Design of Electrical Devices, 2010
ISBN 978-3-642-16224-4

Vol. 328. Raoul Huys and Viktor K. Jirsa (Eds.)
Nonlinear Dynamics in Human Behavior, 2010
ISBN 978-3-642-16261-9

Vol. 329. Santi Caballé, Fatos Xhafa, and Ajith Abraham (Eds.)
Intelligent Networking, Collaborative Systems and Applications, 2010
ISBN 978-3-642-16792-8

Vol. 330. Steffen Rendle
Context-Aware Ranking with Factorization Models, 2010
ISBN 978-3-642-16897-0

Vol. 331. Athena Vakali and Lakhmi C. Jain (Eds.)
New Directions in Web Data Management 1, 2011
ISBN 978-3-642-17550-3

Vol. 332. Jianguo Zhang, Ling Shao, Lei Zhang, and Graeme A. Jones (Eds.)
Intelligent Video Event Analysis and Understanding, 2011
ISBN 978-3-642-17553-4

Vol. 333. Fedja Hadzic, Henry Tan, and Tharam S. Dillon
Mining of Data with Complex Structures, 2011
ISBN 978-3-642-17556-5

Vol. 334. Álvaro Herrero and Emilio Corchado (Eds.)
Mobile Hybrid Intrusion Detection, 2011
ISBN 978-3-642-18298-3

Vol. 335. Radomir S. Stankovic and Radomir S. Stankovic
From Boolean Logic to Switching Circuits and Automata, 2011
ISBN 978-3-642-11681-0

Vol. 336. Paolo Remagnino, Dorothy N. Monekosso, and Lakhmi C. Jain (Eds.)
Innovations in Defence Support Systems – 3, 2011
ISBN 978-3-642-18277-8

Vol. 337. Sheryl Brahnam and Lakhmi C. Jain (Eds.)
Advanced Computational Intelligence Paradigms in Healthcare 6, 2011
ISBN 978-3-642-17823-8

Vol. 338. Lakhmi C. Jain, Eugene V. Aidman, and Canicious Abeynayake (Eds.)
Innovations in Defence Support Systems – 2, 2011
ISBN 978-3-642-17763-7

Vol. 339. Halina Kwasnicka, Lakhmi C. Jain (Eds.)
Innovations in Intelligent Image Analysis, 2010
ISBN 978-3-642-17933-4

Vol. 340. Heinrich Hussmann, Gerrit Meixner, and Detlef Zuehlke (Eds.)
Model-Driven Development of Advanced User Interfaces, 2011
ISBN 978-3-642-14561-2

Vol. 341. Stéphane Doncieux, Nicolas Bredeche, and Jean-Baptiste Mouret(Eds.)
New Horizons in Evolutionary Robotics, 2011
ISBN 978-3-642-18271-6

Vol. 342. Federico Montesino Pouzols, Diego R. Lopez, and Angel Barriga Barros
Mining and Control of Network Traffic by Computational Intelligence, 2011
ISBN 978-3-642-18083-5

Vol. 343. Kurosh Madani, António Dourado Correia, Agostinho Rosa, and Joaquim Filipe (Eds.)
Computational Intelligence, 2011
ISBN 978-3-642-20205-6

Vol. 344. Atilla Elçi, Mamadou Tadiou Koné, and Mehmet A. Orgun (Eds.)
Semantic Agent Systems, 2011
ISBN 978-3-642-18307-2

Vol. 345. Shi Yu, Léon-Charles Tranchevent, Bart De Moor, and Yves Moreau
Kernel-based Data Fusion for Machine Learning, 2011
ISBN 978-3-642-19405-4

Vol. 346. Weisi Lin, Dacheng Tao, Janusz Kacprzyk, Zhu Li, Ebroul Izquierdo, and Haohong Wang (Eds.)
Multimedia Analysis, Processing and Communications, 2011
ISBN 978-3-642-19550-1

Weisi Lin, Dacheng Tao, Janusz Kacprzyk, Zhu Li,
Ebroul Izquierdo, and Haohong Wang (Eds.)

Multimedia Analysis, Processing and Communications

Springer

Dr. Weisi Lin
School of Computer Engineering
Nanyang Technological University,
Singapore 639798
E-mail: wslin@ntu.edu.sg
http://www3.ntu.edu.sg/home/wslin/

Dr. Dacheng Tao
School of Computer Engineering
Nanyang Technological University,
Singapore 639798
E-mail: dctao@ntu.edu.sg
http://www3.ntu.edu.sg/home/dctao/

Dr. Janusz Kacprzyk
Intelligent Systems Laboratory
Systems Research Institute
Polish Academy of Sciences
E-mail: Janusz.Kacprzyk@ibspan.waw.pl
http://www.ibspan.waw.pl/ kacprzyk/

Dr. Zhu Li
Department of Computing
Hong Kong Polytechnic University,
Hung Hom, Hong Kong
E-mail: zhu.li@ieee.org
http://users.eecs.northwestern.edu/ zli/

Dr. Ebroul Izquierdo
School of Electronic Engineering and
Computer Science, Queen Mary,
University of London, London, U.K.
E-mail: ebroul.izquierdo@elec.qmul.ac.uk
http://www.elec.qmul.ac.uk/mmv/people/ebroul/

Dr. Haohong Wang
TCL-Thomson Electronics
Santa Clara, California
E-mail: haohong@ieee.org
http://users.eecs.northwestern.edu/ haohong/

ISBN 978-3-642-19550-1 e-ISBN 978-3-642-19551-8

DOI 10.1007/978-3-642-19551-8

Studies in Computational Intelligence ISSN 1860-949X

Library of Congress Control Number: 2011923797

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The rapid advances in computing, communication and storage technologies have heralded a new age of explosive growth in multimedia applications, such as online image and video repository, mobile TV and IPTV, video on demand, interactive multimedia game, video blogging, and multimedia based social interaction. These applications open up new opportunities and present new challenges to the technologies in the area of multimedia computing architectures, audio/visual information processing, multimedia analysis and understanding, multimedia retrieval and mining, multimedia coding, communication and networking. During the recent years, considerable amounts of research activities in both industry and academia have been devoted to these topics and a key piece of puzzle is to develop novel and effective approaches in modeling and analyzing, representing and understanding, and encoding and distributing multimedia content, all of which will be the focus of this book.

This edited book provides an excellent forum for experts around the world to present their newest research results, exchange latest experiences and insights, and explore future directions in this important and rapidly evolving field. It aims at increasing the synergy between academic and industry professionals working in the field. It focuses on the state-of-the-art research in various essential areas related to emerging technologies, standards and applications on analysis, processing, computing, and communication of multimedia information.

The target audience of this book will be mainly researchers and engineers as well as graduate students working in various disciplines linked to multimedia analysis, processing and communications, e.g., computer vision, pattern recognition, information technology, image processing, and artificial intelligence. The book is also meant to a broader audience including practicing professionals working in image/video applications such as image processing, video surveillance, multimedia indexing and retrieval, and so on.

Since this book comprises different algorithmic advances and applications, it has been organized into three parts, as outlined and introduced as follows.

Part I: Image Processing and Analysis

The issues related to image processing and analysis are to be discussed in the first eight chapters. In particular, image processing is usually referred to as the mathematical operations on images, generally with digital computers, in order to make modifications, extract certain information or perform understanding and retrieval. The earliest techniques, such as medical imaging, character recognition and image enhancement, can be traced back to 1960s. Throughout the years, with the proliferation of digital cameras and the advances of computing hardware, more and more image processing techniques are developed, and they are attracting the interests of multiple research communities. Nowadays, people are living with the surroundings of images and their processing. After capturing a photo, you can perform many meaningful modifications like segmentation, fusion and matting. When photos are uploaded to Facebook, face detection and recognition technology can help to organize these photos. One can also easily finds images or photos he or she wants via content-based or text-based retrieval techniques. In this part, we cover image database visualization and browsing, human computer interactions in image retrieval, image watermarking, and sketch based face recognition, as well as some low level image processing techniques, e.g., image segmentation and deblur. In each chapter, appropriate evaluation has been included for the introduced techniques and their applications in multimedia services.

Part II: Video Processing and Analysis

We then turn to discussion about video processing and analysis in Chapters 9 to 19. Today's fast developments in digital media processing capabilities and network speeds have made the dissemination of multimedia data more rapid and reliable, and attracted significant research attentions to action recognition, event detection, and video tracking and surveillance. The big improvements in digital multimedia processing are beneficial to the fast and reliable production, processing and dissemination of large amounts of digital data. However, this can easily become a time consuming and cumbersome problem. Therefore, the automated extraction of high level information (such as when and where activities occur, or who and what is in a video) using low-level image and video features (e.g. color, texture, shape, motion) is critical for the future development of multimedia research. Some specific video processing and analysis techniques like video frames localization, video shots detection and segmentation, and activity localization have attracted significant attention. In this part, we focus on the research works on object detection/tracking in surveillance videos, human action recognition based on different types of features, 2D and 3D pose recovery, domain driven video analysis, knowledge extraction with scalable ontological networks for video retrieval, visual quality assessment, and video recommendation. Examples in real world applications and computer simulation results are also presented to give convincing illustrations and help the readers to achieve a deeper insight in the related topics.

Part III: Communications Related Processing

In present age of information technology, multimedia data are ubiquitous in our daily life and work. Thus, multimedia research has become one of the central issues in the relevant research and development. The rapid growth of computer network and communication technology has pushed forward greatly the overall advance of multimedia technology. The study on multimedia data compression is not only important for theoretical studies but also urgently needed in practice. Along with the development of digital video equipment, people take photos and videos with better quality. Moreover, people's continuously pursuing of better visual effect and experience promotes the emergence of high definition (HD) TV, HD video, etc. How to store and transfer these huge data becomes a critical issue for multimedia research. For example, when people need to share photos on the internet, they need a proper compression technique to maintain data quality. Therefore, communication related issues are important for multimedia research. In this part (Chapters 20 to 26), we introduce techniques related to multimedia signal coding, evaluation and transmission.

This book project has brought 26 groups of active researchers together in the areas which we really believe in and with the technology that is expected to have great impact in our work and life. The preparation of this book has been a long, arduous and difficult task. We would like to thank all the authors for their great effort and dedication in preparing their quality contributions. Also, all the reviewers of this Springer book deserve our utmost gratitude. We have enjoyed the whole process, and hope that the researchers, engineers, students and other professionals who read this book would find it informative, useful and inspirational toward their own work in one way or another.

November 2010

Weisi Lin
Dacheng Tao
Janusz Kacprzyk
Zhu Li
Ebroul Izquierdo
Haohong Wang

Contents

Part I: Image Processing and Analysis

Part II: Video Processing and Analysis

Part III: Communications Related Processing

Part I
Image Processing and Analysis

Visualisation and Browsing of Image Databases

William Plant[1] and Gerald Schaefer[2]

[1] School of Engineering and Applied Science
Aston University
Birmingham, U.K.
plantwr1@aston.ac.uk

[2] Department of Computer Science
Loughborough University
Loughborough, U.K.
gerald.schaefer@ieee.org

In this chapter we provide a comprehensive overview of the emerging field of visualising and browsing image databases. We start with a brief introduction to content-based image retrieval and the traditional query-by-example search paradigm that many retrieval systems employ. We specify the problems associated with this type of interface, such as users not being able to formulate a query due to not having a target image or concept in mind. The idea of browsing systems is then introduced as a means to combat these issues, harnessing the cognitive power of the human mind in order to speed up image retrieval. We detail common methods in which the often high-dimensional feature data extracted from images can be used to visualise image databases in an intuitive way. Systems using dimensionality reduction techniques, such as multi-dimensional scaling, are reviewed along with those that cluster images using either divisive or agglomerative techniques as well as graph-based visualisations. While visualisation of an image collection is useful for providing an overview of the contained images, it forms only part of an image database navigation system. We therefore also present various methods provided by these systems to allow for interactive browsing of these datasets. A further area we explore are user studies of systems and visualisations where we look at the different evaluations undertaken in order to test usability and compare systems, and highlight the key findings from these studies. We conclude the chapter with several recommendations for future work in this area.

1 Introduction

Nowadays, the majority of people possess some form of digital camera to use in their everyday lives. Devices range from relatively low quality web cameras, to medium range cameras integrated into mobile devices, to higher quality cameras aimed at the average user, on to high-end cameras used

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 3–57.
springerlink.com

by professional photographers. Affordability of devices and storage media coupled with increased capabilities and the 'to hand' availability of camera equipment has led to a dramatic increase in the number of digital images the average end user creates and stores.

With the reduction in digital photography costs, a shift in the attitude towards photo taking can be observed. Users tend to take more images now than before, particularly of the same objects or scene (e.g. from different perspectives) [25]. This is certainly a change from the past, where one would generally be concerned about the number of exposures left on the current film roll or the cost of developing photographs, whereas a digital camera user not happy with a photo can simply delete it from the camera's memory and images can be printed on home printers.

Personal image collections nowadays are typically in the range of hundreds to thousands of images. The rapid increase in the number of digital images taken by individuals has also caused an exponential growth in the number of images available online. Social networking sites allow users to instantly share images with friends, family or a wider community of users that also have the ability to comment and even 'tag' who or what may be in an image.

Commercially, professional photography companies may store millions of digital images in their databases [50]. These are generally manually annotated image collections used by journalists from a variety of publications to search for images suited to their particular needs. As one can imagine, the search for any particular image in collections of either personal or commercial magnitude can be tiresome and exhaustive. Generally, images are arranged in a one-dimensional linear arrangement, whereby an image has no correlation to any of its neighbours. Images are usually grouped together in a manually named folder or on the basis that they were uploaded to the computer at the same time.

This organisation of images is not ideal for a variety of reasons. Firstly, the cost of storage media has dramatically decreased whilst storage capacity has increased. Therefore an average end user may take many photos of many different events (such as birthdays, holidays etc.) on a camera before uploading them to their computer. If not sorted manually, multiple events may get grouped together, potentially making it difficult for the user to locate specific images in the future.

This leads to a second issue of manually annotating folders. If images of multiple events are stored in the same folder, it is difficult to describe the ambiguity of the content contained within it using just a folder name. Typically the date of the camera upload will be chosen, but this could become rather meaningless after a long period of time. Rodden and Wood [62] demonstrated in their analysis of digital photograph management that users are generally unwilling to annotate their images. Another issue is that words chosen to annotate an image can be highly subjective, with appropriate keywords changing between different users which in turn can render keyword-based search unintuitve and difficult to operate [39].

1.1 Content-Based Image Retrieval

Since textual annotations are not available for most images, searching for particular pictures becomes an inherently difficult task. Luckily a lot of research has been conducted over the last two decades leading to many interesting methods for content-based image retrieval [75, 11]. Content-based image retrieval (CBIR) does not rely on textual attributes but allows search based on features that are directly extracted from the images [75]. This however is, not surprisingly, rather challenging and often relies on the notion of 'visual similarity' between images or parts thereof. While humans are capable of effortlessly matching similar images or objects, machine vision research still has a long way to go before it will reach a similar performance for computers.

Smeulders *et al.* [75] define three primary applications of CBIR systems. A *target search* is undertaken when the user has an absolute target in mind, perhaps of an exact image or images of a specific object or scene. A *category search* is undertaken when a user requires an image that best represents some class of images. Finally, in *search by association*, users have no initial aim other than to search for images of interest. This usually leads to an iterative procedure whereby the search may be focussed on an image which the user finds interesting.

1.2 Query-By-Example

In the early days of CBIR, the general method used by systems such QBIC [15], Virage [20], PhotoBook [52] or NeTra [43], to query an image database was through a query-by-example (QBE) approach. QBE allows a user to specify a query image to the system in order to retrieve images from the database that are deemed similar to that query. Each image is characterised by a feature vector (e.g. the bins of a colour histogram as originally proposed in [76], or a combination of colour, texture and shape features as in [15] - see [75] for a detailed review on image features). An equivalent feature vector is extracted from the query image and compared to all database vectors to arrive at similarity or dissimilarity scores between query and database images (using metrics such as L_1 [76] and L_2 [15] norms or the earth mover's distance [64]).

Upon comparing the database images to the query, the system will present the top N similar images according to their distance from the query image. The presentation of results is typically a one-dimensional linear arrangement, in order of increasing distance (i.e. decreasing similarity) starting from the top left hand corner of a grid.

There are two main drawbacks of QBE-based CBIR. The first one is that users may not deem the images presented by the system as actually being similar to the query. For example, a user may supply the system with a red flower. The system will return all images with a large red content, and the texture and shape similar to a flower. However the user may be searching either for red flowers, or a particular species of flower that happens to be red in their particular picture. This high-level interpretation of an image by the user cannot be satisfied by the low-level feature calculations performed by the computer. This problem is of course not specific to QBE-based retrieval but is common to all similarity-based CBIR approaches and is known as the 'semantic gap' [75].

The second shortcoming of QBE is that a user may not actually have an image to give to the system, thus rendering QBE effectively useless. While potential solutions such as sketch-by-example [28, 38] have been proposed in order to overcome this issue, these have limitations of their own and are hence rarely explored.

1.3 Relevance Feedback

A commonly explored approach to improve the retrieval performance of CBIR systems, and a partial solution to the first issue presented above, is relevance feedback (RF) [86]. This mechanism modifies the underlying parameters of the algorithms of a system in an attempt to learn what a user is searching for. Upon presentation of the initially retrieved images, the user can specify whether they deem a retrieved image useful or not. Multiple images can be selected as either positive or negative examples and these are then used in order to weight the different features according to the user's preference, and update the search results which should now contain more relevant images. This process can be repeated to further improve the retrieved results. In the aforementioned example of the red flower, if the user were to select multiple images of red flowers as positive examples the system is then likely to return more red flowers, weighing the colour feature more highly than shape or texture. On the other hand, if the user selects images of the same species of flower but with varying coloured petals, the system will emphasise shape and texture more than colour. A variety of RF mechanisms exist [86], the most common being a relevant or non-relevant selection (as e.g. used in [15]) or slide mechanisms allowing the user to specify a continuous score of relevance (as employed e.g. in [65]).

The user will generally only select a small amount of positive and negative examples. Therefore, small sample learning methods are required. The most successful of these methods include discriminant analysis and support vector machines (SVMs) [77]. In the work of Tao *et al.* [78], the authors state that SVM based RF has shown promising results in previous studies due to good generalisation abilities, but show that incorporating asymmetric bagging and a random subspace into a SVM, can lead to improved results, while reducing computational complexity. The authors of [77] experiment with variations of discriminant analysis for RF, namely LDA (Fisher linear discriminant analysis) and BDA (biased discriminant analysis) and develop an improved method named directed kernel BDA (DKBDA). The reader is directed to works such as [86, 77, 78] for further information on these and other RF algorithms.

Another variation of RF allows the user to manually drag the system results closer or further away from the query image based on preference [23]. Other browsing-based RF mechanisms are described in Section 3.5.

1.4 Image Browsing Systems

Image browsing systems attempt to provide the user with a more intuitive interface, displaying more images at once in order to harness the cognitive power of the human mind in order to recognise and comprehend an image in seconds. Interaction with a traditional QBE system can often lead to a confusing and frustrating user experience. Formulating queries from images can prove difficult for the user, and the 'black-box' state of such approaches means that users typically cannot derive how the system is retrieving these results, and are thus unable to modify the query in order to improve the results returned by the system.

This is confirmed in a user study presented by Rodden and Wood [62] where the authors provided users with an image retrieval system that offered a variety of querying facilities, including speech recognition and the traditional QBE approach. The authors found (by examining usage logs) that most users did not use the QBE function as the system did not meet their unrealistic expectations of the current state of CBIR. For example, a user explained how he had attempted to find all the images of a new blue car by using a query image, but the images provided were irrelevant. As he had no idea how the system was providing these results, he could not improve the query and thus abandoned the search.

Browsing systems give a useful alternative to QBE. Providing an overview of the database to the user allows for intuitive navigation throughout the system. This is particularly the case when images are arranged according to mutual similarity as has been shown in [59], where a random arrangement of images was compared with a visualisation which positioned images according to their visual similarities, i.e. where images that are visually similar to each other are located close to each other in the visualisation space. It was discovered that during a target search (i.e. looking for a particular image), similarity-based visualisation reduced image retrieval time.

QBE systems cannot be used when the user does not have a specific image in mind, as no query image can be provided. Image browsing systems overcome this problem by showing an overview of the image database. An overview of the collection will give the user a good indication whether or not an image or image class they have in mind might actually be present in the database. In some cases, the entire database will be displayed to the user on a single display. The user can then focus on regions of the visualisations that they are attracted to or believe will harbor a particular concept they have in mind. Browsing such visualisations when arranged according to image similarity, as shown in [59], can increase the rate of retrieval. These visualisations are usually achieved through dimensionality reduction, whereby the relationships between images in a high-dimensional feature space are maintained as best possible in a reduced 2D (or 3D) space which is more comprehensible to the user.

In case image collections are too large to fit to a single display, images can be grouped according to similarity through the application of a clustering procedure. The user is then able to navigate through these clustered groups of images in order to browse the collection. An overview of the database is provided by initially presenting the user with a representative image for each cluster. Clustering can also be performed in a hierarchical manner which in turn allows for visualisation of very large datasets.

Another way in which image databases can be displayed is through graph-based visualisations. In these approaches, links are formed between images that are deemed similar or that share a common concept, while the images themselves form the nodes of the graph. The whole connected graph, or part thereof, is then displayed to the user for visualisation and navigation.

Similarity-based visualisation is not the only useful form of arranging image databases. In particular for personal collections, grouping according to the time images were created has shown to be useful. This approach can be adopted to automatically cluster event images. In cases where time information is not always available or not necessarily reliable, this approach can be combined with similarity-based systems.

The fundamental issue with the development of a browsing system is how to present the user with the images in a database. With image collections ranging in the size of millions, any browsing system needs to utilise the limited screen space provided by a typical computer monitor in a manner which is intuitive and easily navigable by the common user. Immersive environments and virtual reality allow for a completely new way of visualising information with a unique user experience. It is only natural that this approach has also been adopted for visualising image databases. The user is immersed into the actual database, while the addition of a third dimension coupled with the larger visualisation space can lead to a more effective approach of navigation.

While a visualisation of an image collection is useful for providing an overview of the contained images, it provides only part of a useable image database navigation system. Once a collection is visualised, users should have the ability to interact with it in order to arrive at the image(s) they are looking for. Typical operations here include panning and zooming which allow the user to focus on images in a different part of the visualisation space, respectively on images which were previously hidden.

With regards to the three primary CBIR applications of [75], browsing interfaces clearly allow for better *search by association* (searching with no specified target) than QBE approaches. As for *target search* (looking for a particular image), QBE interfaces may provide quicker retrieval times compared to a browsing interface, but of course need a query image to start with. For *category search*, arranging images by similarity creates intuitive groupings of images relating to the same category. On the other hand, formulating a single query image of a category for QBE could prove difficult. For example, suppose a user wanted to enrich a travel article of Australia with a handful of pictures. It is not clear which images in the database would be best suited, without seeing all the Australia related pictures in the database. QBE could only work in this instance if the user knows exactly which aspect of Australia they require (e.g. an image of a kangaroo or the Sydney Opera House). In contrast, allowing the user to browse the database can help cross the 'semantic gap' by allowing the user's cognitive system to play a more active role during image selection.

Browsing systems can provide users with a much less constrained, continuous interface in order to explore an image database. In this chapter, we review a variety of methods used by different researchers in order to arrange and visualise image databases to support intuitive image database navigation. The rest of the chapter is organised as follows: Section 2 focusses on how these databases can be visualised, explaining approaches based on dimensionality reduction, clustering, and graph-based visualisations. Section 3 describes different tools implemented by researchers in order to enable users to browse these visualisations. Section 4 highlights user studies undertaken in the field, how they are performed and what discoveries such studies have found. In each section we provide a critical discussion of the various approaches proposed in the literature. Our observations are summarised and future directions identified in Section 5.

2 Visualisation of Image Databases

In order to browse an image database, the users need to be presented with thumbnails of the images so that they may intuitively navigate the database. The primary issue associated with visualisation is how to best display the images within the limited space of (typically) a 2D screen. A variety of methods have been devised in order to visualise images, whether it be the entire database or a subset of images. In this section we look at the different techniques used in order to visualise image databases for browsing.

2.1 Mapping-Based Visualisation

CBIR systems typically employ high-dimensional features to represent images. Clearly, it is impossible for the human mind to perceive a feature space of this magnitude, and based on the raw data, we are therefore unable to recognise potential relationships within the dataset. In order to visualise this high-dimensional data, various techniques exist which describe the feature space layout within a low-dimensional model which the human mind can more readily understand. For image database browsing, this mapping is typically down to just two dimensions, namely the x and y co-ordinates of a 2D computer display. The main problem is obviously how to perform this mapping so that the relationships of the original data are maintained. In the following, we discuss various approaches that have been employed to this effect.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is the simplest dimensionality reduction approach, working in a linear manner. The starting point for PCA is the symmetric covariance matrix of the feature data from which the eigenvectors and their respective eigenvalues are calculated and ranked in descending order of eigenvalues. The principal components are selected from the top eigenvectors according to the number of dimensions required (i.e. for 2D the top two eigenvectors are selected). These eigenvectors are then used to plot the original data where image thumbnails are plotted at the co-ordinates derived through projection of the orginal feature data into the low-dimensional space. PCA has the advantage that it is relatively simple. However, since it maximises the variance of the captured data it does not necessarily best preserve the mutual relations between the individual data items (this is only the case if the underlying metric in the original feature space is the L_2 norm).

The Personal Digital Historian (PDH) project developed by Mitsubishi Electronics Research Lab (MERL) [45] uses PCA splats in order to visualise images. PDH attempts to bring photo sharing to a round table top, with the system being projected down from above. The authors use colour, texture, and shape features which are then projected, using PCA, to a 2D format whereby similar images appear close together. Keller *et al.* also use a PCA visualisation to present images in a virtual 3D interface based on texture features [31].

Multi-Dimensional Scaling (MDS)

In contrast to PCA, multi-dimensional scaling (MDS) [36] attempts to preserve the original relationships (i.e. distances) of the high dimensional space, as best possible in the low-dimensional projection. MDS starts with a similarity matrix which describes all pair-wise distances between objects in the original, high-dimensional space. The goal is then to best maintain these distances which in turn can be formulated as minimizing a ‘stress’ measure, often defined as [36]

$$\text{STRESS} = \frac{\sum_{i,j} (\hat{\delta}_{ij} - \delta_{ij})^2}{\sum_{i,j} \delta_{ij}^2} \tag{1}$$

where δ_{ij} is the original distance between objects i and j, and $\hat{\delta}_{ij}$ is the distance in the low-dimensional space. Starting from either a random initial configuration, or from the co-ordinates after applying PCA, the algorithm continues to reposition the images in order to reduce the overall stress, until a termination condition has been reached (for example a maximum number of iterations or threshold stress value).

MDS was employed by Rubner *et al.* [64] who suggested using it for browsing image collections. Based on colour signatures of images and the earth mover's distance (EMD) [64], the authors were able to create a representation of the high-dimensional feature space using MDS, placing image thumbnails at the co-ordinates derived by the algorithm. Figure 1 shows an example of a MDS visualisation of an image database.

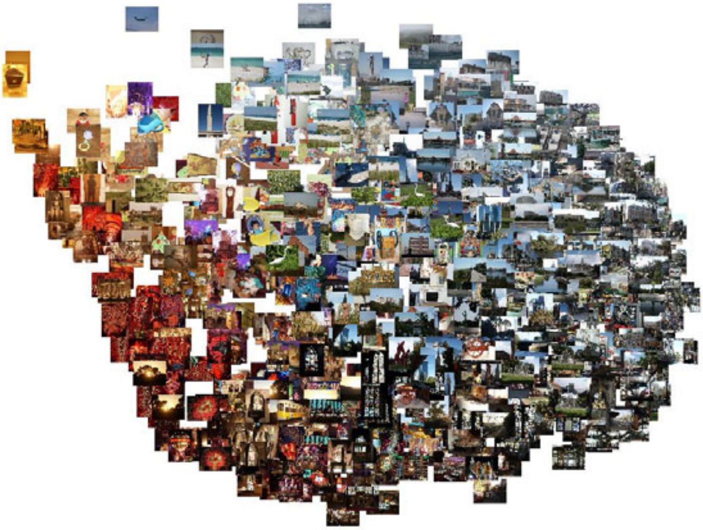


Fig. 1. An MDS visualisation of the UCID image database [74]

MDS provides a more accurate representation of the relationships between images in the feature space compared to PCA. The work of [64] also suggested that MDS can be used for both image query results (local MDS) and to give an overview of a collection of images, providing the user with a general scope of images contained within the database (global MDS). However, MDS comes at the cost of more expensive computation compared with PCA, working in quadratic time. This suggests that image co-ordinates cannot be calculated interactively, and thus that MDS is not well suited to present query results. For global MDS, though image co-ordinates may be calculated off-line in order to browse the data set interactively. Additional difficulties arise when adding images to a collection visualised through MDS, as this typically requires recalculation of the whole dataset and the relocation of image thumbnails in the visualisation.

Rodden *et al.* have investigated the use of MDS for image database visualisation based on the evalution of several user studies [59, 60, 61, 62].

In [59], they compare two approaches, one based on random assortment of images, and one using a similarity-based MDS interface, and conclude that the MDS-based system is faster for locating specific images.

MDS has also been used to measure the effectiveness of particular feature vectors for conveying similarity within a CBIR system. In [40], MDS is employed to manually inspect the similarity derived by using the MPEG-7 texture retrieval descriptor (TRD) and texture browsing descriptor (TBD). They conclude that using the TRD with either the L_1 norm or EMD distances provides more suitable MDS layouts. Besides visual inspection, they also used spatial precision and recall in order to arrive at quantitative conclusions.

These accuracy measures, which are adaptations of the classical precision and recall measures used in information retrieval, were first proposed in [58], where a quantitative comparison between different distance measures is undertaken to examine which provides the best MDS visualisation according to similarity perceived by humans. In order to calculate the average spatial precision and recall, each image in the database is treated as a query image. In Figure 2 the dashed circles represent the increasing levels of recall from the query image (coloured dark gray). The levels of recall are set based on the next closest relevant image (coloured light gray) to the query. The number of relevant images within a circle is divided by the total number of images in that recall circle to calculate spatial precision. This is then averaged for all the recall circles, giving an average level of precision.

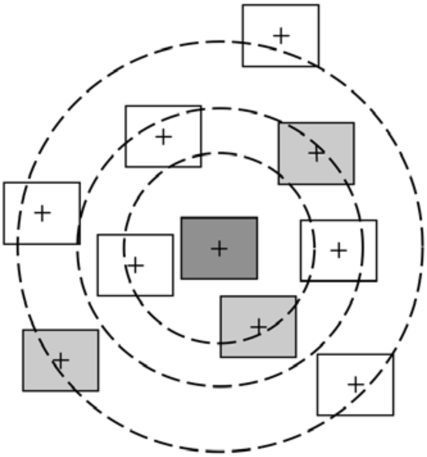


Fig. 2. Illustration of the spatial precision and recall measures used in [58]

Using these measures, [58] examines the quality of visualisations when using different indexing methods and distance measures. They evaluated feature vectors consisting of averages of hue, saturation and value, localised average hue, saturation and value features (where the image is partitioned into 9 regular grid cells), and colour signatures as used in [64], HSV histograms using the χ^2 (chi-squared) and Jeffery Divergence measures, and finally a scheme named IRIS, which is a fairly complex index introduced in the paper. The authors first compared the indexing techniques using a standard QBE system, where results show that the more complex IRIS indexing method achieves the best precision and recall. They then explored how these indexing techniques compare in terms of average spatial recall and precision for an MDS visualisation. Interestingly, they found that here the simplest measure, namely average HSV values, is able to retain roughly 85% of the accuracy, whilst IRIS achieves only around 52%. The authors conclude that, in a reduced dimensionality space, an average HSV MDS visualisation is comparable with a more complex indexing technique, such as IRIS, yet much more simple to compute. They furthermore investigate the computational complexity in more detail and report that the most time consuming indexing technique is the colour signature/EMD method of [64] which takes about 230 times longer to compute a full similarity matrix compared to the average HSV computation.

FastMap

FastMap is an alternative dimensionality reduction technique devised by Faloutsos and Lin [16]. FastMap is able to reduce high-dimensional spaces down to a linear 2D or 3D space. The algorithm selects two pivot objects, an arbitrary image and its furthest possible neighbour. All points are mapped to the line that connects the two pivot points using a hyper-plane located perpendicular to the line that connects the two pivots. The co-ordinates where images appear on the hyper-plane can be used to display the images on the screen, maintaining the relationships which occur in the high-dimensional space. As with MDS, a distance matrix is required as input for the algorithm.

The advantage of FastMap is that it requires less computation compared to MDS, having a linear $O(kn)$ complexity, where n is the number of images and k is the number of dimensions to reduce the data to. In their experiments, the authors tested FastMap against MDS, showing more than comparable results in much shorter times. This suggests that FastMap could potentially be used for computing visualisations ‘on-the-fly’, for example to visualise results of QBE searches. The resultant visualisation however, is not always as accurate as those created by MDS.

FastMap is also employed in the virtual reality system 3D MARS [47] to map images to a 3-dimensional space in which users can virtually navigate themselves around the image database through query selection (see Section 2.4 for more details on virtual reality visualisation systems).

Self-Organising Maps

A self-organising map (SOM) [33] is a neural network which is trained to perform feature extraction and visualisation from the input of raw data.

Using an input layer of neurons, the feature vector of a sample is computed and assigned to a best matching unit (BMU) on a 2D map. Each unit has an associated weight vector, with the same dimensionality of the feature vectors computed from each of the samples in the dataset. A learning rule, typically defined as

$$w_i(t+1) = w_i(t) + \gamma(t)h_{b,i}(t)[x(t) - w_i(t)] \tag{2}$$

where $w_i(t)$ is the weight vector of node i, $\gamma(t)$ is the learning rate and $h_{b,i}$ is a function modifying the weights around the BMU, is then applied to update the weight vectors.

Applied to image databases, employing SOMs leads to similar images being located closer together on the resulting 2D map than less similar images [12]. To avoid a time consuming linear search of what could be an extremely large map (according to the size of the database), hierarchical self-organising maps (HSOMs) can be constructed where only root BMUs need to be compared to the input vector during mapping [12].

An earlier use of SOMs for image database visualisation is the PicSOM system [37]. PicSOM uses layers of parallel SOMs to form a hierarchy, in particular a tree-structured self-organising map (TS-SOM) [34]. Here also, a linear search of all units in the map for the BMU of a given feature vector (constructed in PicSOM using MPEG-7 descriptors for colour, texure and shape) is avoided, by restricting the search for a BMU to a 10×10 unit search below the BMU of the previous level. This reduces the overall BMU search complexity from $O(n)$ to $O(\log(n))$. After training has been implemented on each of the TS-SOM levels (using each image in the test set 100 times), each node is assigned the image most similar from the database. This results in similar images being mapped closer together than dissimilar images on the 2D map. These representative images may then be browsed by the user in a hierarchical manner (see Section 3.2).

While the work of Zhang and Zhong [85] focusses on the development of a content-based HSOM as an indexing structure, Deng *et al.* [12] and Eidenberger [14] implement visualisations that facilitate the browsing of the images in the database. Deng *et al.* [12] train a HSOM using Sammon mapping [67], an MDS variant. The low-level features extracted from the images were regional CIEL*u*v* colour averages, an edge density histogram and texture features extracted through Gabor filters. In their experiments, the system was used to visualise a collection of 3,000 images.

Eidenberger [14] describes a system where HSOMs of video stills are created based on a variety of MPEG-7 descriptors. Each input vector is compared with a BMU representing a cluster of images. When these clusters of images are visualised by the HSOM, a representative image (closest to the BMU weight vector) is displayed. Furthermore, an HSOM is employed for time-based visualisation. Here, each node is required to be visualised by exactly one image, rather than a cluster as with similarity-based visualisation. This is achieved by using the weight vector of each node in the map and assigning it with the closest image feature vector available in the database. The output maps were computed on a hexagonal layout, and images cropped to hexagons.

Other Mapping-based Techniques

A range of more recent techniques for visualising high-dimensional data are investigated by Nguyen and Worring [49]. The three non-linear embedding algorithms employed are ISOMAP (isometric mapping), SNE (stochastic neighbour embedding) and LLE (local linear embedding). In ISOMAP [79], nearest neighbour graphs are formed within the data, and the shortest path between every pair of points is calculated, with the length of the path being used in a distance matrix for MDS. SNE [26] calculates the probability that any two points take each other as nearest neighbours in both the high- and reduced-dimensional space, and attempts to match the two probability distributions. LLE [63] can be seen as an approximation of SNE. The authors further propose to merge ISOMAP with SNE and LLE to form two new techniques, ISOSNE and ISOLLE. In ISOSNE, the distances found through ISOMAP are used to form the probabilities used by SNE, rather than using MDS, and ISOLLE is derived in an analogous way.

In their evaluation they found that both ISOSNE and ISOLLE perform better than MDS. Although ISOSNE performed best, the computation time was reported at being around 10 times that of ISOLLE. The authors therefore concluded that if off-line calculations can be performed, ISOSNE can be used, while for faster visualisations, ISOLLE should be the method of choice.

Milanese *et al.* [44] describe the use of correspondence analysis [30] as a dimensionality reducing mapping technique. Using a data table, a mathematical function is applied in order to create an observation matrix, which can be be used with the eigenvectors of a covariance matrix in order to project the data table into the 2D space. This formulation allows both images and features to be projected onto a common space, and to distinguish which features are closer to a particular cluster of images.

Handling Overlap in Visualisations

Rodden *et al.* [59] observed that the vast majority of users do not like the overlapping and occlusion effects occurring in MDS displays due to images being located too close to each other (see also Figure 1). This issue with partial or even total occlusion is of course not exclusive to MDS, but also occurs in other visualisations such as PCA splats. Co-ordinates that are close together in the feature space will inevitably become even closer in a 2D representation generated through mapping. When image thumbnails are overlaid at these co-ordinates, parts of or indeed entire images are hidden from the user.

In order to combat this, various systems invoke some mechanism which adapts the layout in order to reduce the amount of overlap occurring between images. Much work here has focussed on mapping the visualisation to a regular grid structure. Gomi *et al.* [18] used MDS "as a template" in order to locate images within rectangular regions representing a cluster. Rodden *et al.* [59] developed a method for spreading the images around a grid. First the co-ordinates are used to locate the ideal grid cell for an image. Should this cell be already occupied, a spiral search emanating from the selected cell is performed in order to locate the closest free cell (see Figure 3 on the left). In addition to this basic strategy, where the image is simply mapped to the next closest free cell, a further *swap* strategy was also proposed. Here, an image is moved to the next closest cell, and the new image is placed in the optimal cell. Finally, in a *bump* strategy, the images in the line of cells between the optimum cell and the next closest cell are all moved outwards (from the optimum centre cell) by one cell, with the new image being placed at the centre optimum cell. From experiments it was found that the *bump* strategy produces the lowest average error (i.e. lowest average distance an image is from its optimal cell). The complexity of the algorithm is $O(m^2) + O(n^2)$ where m is the size of the grid and n is the number of images to be located. The three strategies are presented visually in Figure 3 on the right. This technique was also adopted by Schaefer and Ruszala in [73] and [71] to spread out images on an MDS plot and a spherical visualisation space respectively.

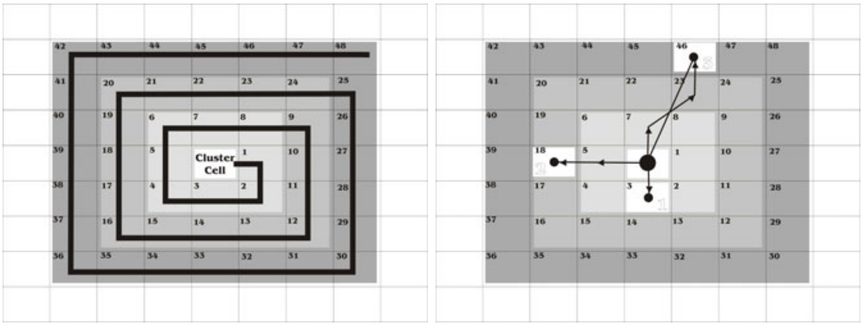


Fig. 3. The spreading strategies proposed in [59]

Liu *et al.* [42] developed two different approaches for overlap reduction in order to present web search engine results. Their first technique also fitted the visualisation to a grid structure, but they comment that the *bump* strategy of [59] works in quadratic time, and is thus not suitable for real-time use. Their method creates an ordered data set, optimised in one dimension while sub-optimising the other and has a complexity of $O(2nlog(n) - nlog(m))$ where m is the number of columns or rows and n is the number of images. Their second technique allows the user to dynamically set the amount of overlap through use of a slider bar. Image co-ordinates are established by

$$P^i_{new} = \gamma P^i_{Sim} + (1 - \gamma) P^i_{Grid} \tag{3}$$

where P^i_{Sim} and P^i_{Grid} represent the locations of the image in the similarity-based and grid-based visualisations respectively, and γ is the overlap ratio controlled through the slider bar.

Nguyen and Worring [49] specify two requirements with regards to dimensionality reduced visualisations, a *structure preservation requirement* and an *image visibility requirement*. The first requirement states that the structure of the relationships between images in the feature space should be retained, while the second demands that images should be visible enough so that the content of the image is distinguishable. It is clear that these two are intrinsically linked. Moving an image in order to make it more visible will detract from the original structure, while maintaining the structure could cause a loss of visibility in certain images.

As a solution to this, Nguyen and Worrring define a cost function which considers both image overlap and structure preservation. In order to detect overlap, a circle is placed about the centre of the image, as it is assumed that an object of focus will be about the centre of an image. If the circles of two images overlap, the position of the images will be modified according to values derived from the cost function. A similar cost function is also used to modify the PCA visualisations in the PDH system of Moghaddam *et al.* [45].

Discussion

From the various works that have employed mapping-based techniques, it is clearly difficult to formulate a direct comparison of which is best. Each individual approach uses a different image database and different underlying features and distance measures to quantify the similarity between images. Ruszala and Schaefer [66] attempt to compare PCA, MDS and FastMap by considering the complexity of the algorithms required. They conclude that if accuracy is of importance then MDS should be used, otherwise FastMap should be implemented when faster visualisation generations are required. However this study does not include the more recent use of local linear embedding algorithms detailed in [49], shown to be faster and as accurate as MDS. Future work could aim at comparing a variety of dimensionality reduction visualisations using the spatial precision and recall measure defined in [60, 58]. The use of approximation algorithms such as FastMap and LLE, operating at lower complexity than more accurate algorithms such as MDS, offers the possibility of visualising dynamically produced data sets such as query results.

From works such as [59, 42, 49] it is clear, that image overlap is an undeniable problem for users who prefer to see images in their entirety. Much research has been undertaken into how is best to resolve this problem. Moving images too far from their mapped location can cause the relationships in the full-dimensional feature space to be distorted, and hence there is a trade-off between image clarity and maintaining the overall structure of the relationships [49]. The placement of images within a grid structure is a visualisation which the general user is familiar with. Hence, arranging images within a grid according to their mutual similarities will typically enhance the general user's browsing experience.

2.2 Clustering-Based Visualisation

Dimensionality reduction techniques applied to image database visualisation are fundamentally limited by the number of pixels displayed on a computer monitor, as this will directly determine the number of images that can be displayed on the screen. Much work has been undertaken in order to reduce the number of images to be displayed to the user at any one time. This is usually achieved by clustering groups of similar images together, so that only a single image for each group is displayed to the user, hence freeing up visualisation space. In this section we describe the principle methods in which images can be grouped automatically for the purpose of image database visualisation, and how each group can be portrayed by representative images.

Content-based Clustering

Content-based clustering uses extracted feature vectors in order to group perceptually similar images together. The advantage of this approach is that no metadata or prior annotation is required in order to arrange images in this manner, although image features or similarity measures which do not model human perception well, can create groupings that may potentially make it difficult for a user to intuitively browse an image database.

Krischnamachari and Abdel-Mottaleb [35] were among the first to propose clustering images by image content. Local colour histograms (extracted from image sub-regions) were used to cluster similar images and each cluster was visualised using a representative image. Schaefer and Ruszala [72] also cluster images based on colour descriptors (the average hue and value in HSV colour space).

Hilliges *et al.* [25] use a combination of colour, texture and roughness features. These are extracted based on a YUV colour histogram, some Haralick texture features and the first four roughness moments of the image. The resulting clustering is utilised in conjunction with an image quality classification technique. The work by Borth *et al.* [5] represents another example of content-based clustering using colour and texture features.

K-means clustering is one of the most commonly used clustering techniques which iteratively approximates cluster centres. Image database navigation approaches that employ k-means include the works by Abdel-Mottaleb *et al.* [1] and Pecenovic *et al.* [51]. Hilliges *et al.* [25] use a variant of k-means named X-means. In their approach, images are first clustered using colour histograms comprised of the u* and v* values of the images in CIEL*u*v* colour space. This way, the system is able to detect series of multiple similar images, which are then classified based on image quality in order for users to only keep their best photographs.

Metadata-based Clustering

Despite the difficulties of manually annotating images, work has been undertaken to visualise images according to this associated metadata. The introduction of systems such as ALIPR (Automatic Linguistic Indexing of Pictures - Real Time) [39] demonstrates that images can be automatically annotated. However, this assignment of high-level semantic meaning by machines is still in its infancy and often not very reliable. Systems such as ImageGrouper [48] and EGO (Effective Group Organisation) [81] allow the user to manually arrange images into clusters and perform the bulk annotation of the contained images.

CAT (Clustered Album Thumbnail) by Gomi *et al.* [18] uses a combination of keyword and content-based clustering. At the top level of the clustered hierarchy, images are clustered by keywords. The user is presented with a list of keywords, of which they can select one or more. Upon keyword selection, all images in the database associated with the chosen keyword(s) are clustered by localised average colour content (average CIEL*u*v* values from grid cells placed over the image) and image texture (calculated through a Daubechies 4 wavelet transform). Each cluster takes a representative image, which in higher levels is sized dependent on the proportion of images from the database that are located in that particular cluster. At lower levels of the structure, images are arranged more uniformly in a grid-like structure using MDS and PCA templates.

The rectangular boxing of clusters employed is similar to that used in PhotoMesa. PhotoMesa [4] has the ability to arrange images in quantum treemaps or bubblemaps. Quantum treemaps are designed to display images of indivisible size regularly, whereas bubblemaps fill the space with indivisable items but generate irregular shaped groups. Images with a shared metadata attribute (e.g. directory, time taken, or keyword) are grouped together. When an image is first loaded into the database, multiple sized thumbnails of the same image are stored in a filesystem and dynamically loaded based on the size of the rectangular sections.

For the quantum treemap algorithm, the input is a list of numbers specifying the size of the rectangles, and the display space. The output is the layout of the rectangles. The algorithm generates rectangles with integer multiples of a given element size, where all the grids of elements align perfectly. When images are assigned to their groups, an evening algorithm is run to re-arrange the images in the boxes. The authors note that a relatively large amount of wasted space may occur on the screen, particulary when the number of images in a group is small. PhotoMesa has three different grid arranging mechanisms in order to irradicate irregular layouts. The size of the rectangle is dependent on the proportion of images from the database that cluster contains. Figure 4 shows an example of a regular layout of images in PhotoMesa.

In an attempt to remove unused space, [4] also introduces the idea of using bubblemaps in order to visualise the database. In this approach, images are

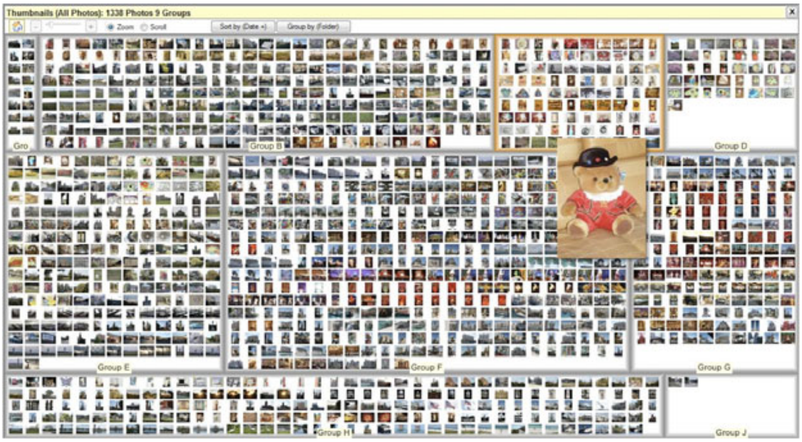


Fig. 4. Clustered images, taken from the UCID dataset [74], visualised as a quantum treemap in PhotoMesa [4]

still displayed on a regular grid, but the surrounding area can be arbitrary in shape.

Time-based and Combined Time/Content-based Clustering

Time-based clustering uses time stamp information associated with an image in order to group images within a collection. This time data may have been provided either by the digital camera when the photograph was taken, or by an operating system when the image was euploaded from the camera or downloaded from the internet, or set manually by the user. The possible ambiguity of when a time stamp may have been attached to an image can indeed be the downfall of this particular method of grouping. Furthermore, some images may contain no time stamp information at all [55].

It has been demonstrated by Rodden and Wood [62] that users find browsing through time-ordered images more intuitive than content-based browsing (see also Section 3.4 for more discussion on this). Graham *et al.* [19] justify their approach of grouping and visualising images according to time with the observation that *"people tend to take personal photographs in bursts"*. Based on this premise, images are clustered according to the time difference between time stamps with images first being clustered by year, then month, day then hour. The authors give an analogy of a birthday party in order to explain sub-clusters in their approach. The event itself will take up an entire day, but different parts of the day may contain different bursts of images, for example blowing out the candles on the cake may have several images attributed to it.

The time-based clustering algorithm employed in Calendar Browser [19] is based on Platt *et al.*'s PhotoTOC system. PhotoTOC (Photo Table of Contents) [55] visualises images in two panes: overview and detail. In the overview pane, a grid of representative images is presented to the user, arranged by month and year, where each image corresponds to a particular time cluster. Images are arranged in a sequential list and new events can be detected by

$$log(g_N) \geq K + \frac{1}{2d+1} \sum_{i=-d}^{d} log(g_{N+1}) \tag{4}$$

Assuming g_i is the time gap between image i and image $i+1$, g_N is considered a gap if it is much longer than the average gap. K is an empirically selected threshold value and d is the window size.

The main difference between the approach by Graham *et al.* and PhotoTOC is how identified events are sub-clustered. In [19], medium sized clusters are first created by a pre-defined time gap. These new clusters are then sub-clustered by the rate at which images are taken for that cluster. This rate can then be matched with the other intra-cluster rates to split and merge clusters. Parent clusters are developed through fixed measurements of time, i.e. events that occurred in the same day, week, month and year. In contrast, PhotoTOC sub-cluster events based on colour content, therefore not completely relying on time stamp information. This approach is hence also applicable to image sets which are only partially time-stamped.

Another approach using a combination of time- and content-based clustering is the PhotoSim system [8]. PhotoSim uses k-means to cluster images already clustered via time, enabling the system to derive clusters that model human perception. For this, they utilise colour histograms based on the U and V components of the YUV colour space. In the example shown in Figure 5, the images in the cluster have been separated into portraits and pictures of buildings taken at either night or day.

Hierarchical Clustering

Hierarchical clustering can be seen as analogous to file structures found in common operating systems with clusters of images corresponding to folders and individual images being mapped to files. Indeed, this is often how users organise their personal collections. The majority of systems that cluster images, arrange clusters in a hierarchical manner. Examples of this can be found in [55, 8, 18, 5].

Hierarchical clustering algorithms are typically divided into agglomerative and divisive methods [29]. Agglomerative, or bottom-up clustering, begins with treating each individual sample as an individual cluster. Using a form of similarity, clusters are merged with their most similar neighbours and this process is repeated until a pre-defined number of clusters remain. These clusters then form the top layer of the generated tree. In contrast, divisive, or top-down, clustering begins with all samples starting as a single large cluster which is then iteratively split into smaller clusters until a termination criterion is met (such as all clusters corresponding to individual samples). In

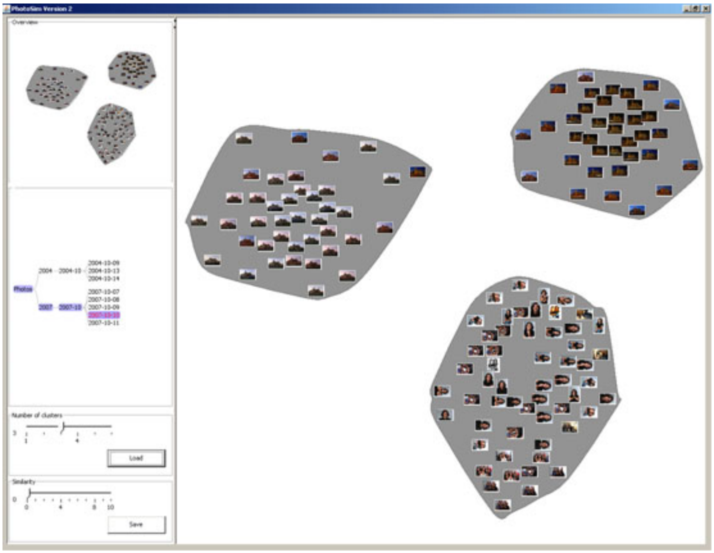


Fig. 5. Images from a cluster in PhotoSim [8] further clustered into portraits, buildings at night and buildings in the day

terms of image database visualisation, the leaf nodes of the tree correspond to the individual images, while the nodes at different levels of the tree form the various image clusters.

Despite being computationally more expensive than partional methods such as k-means clustering, it has been shown that approaches based on agglomerative clustering afford better retrieval accuracy [1]. Krischnamachari and Abdel-Mottaleb [35] use local colour histograms to form a hierarchical structure of images. First, each image is treated as its own cluster, and represents a leaf node of the tree. From all the clusters, the two with the most similar average colour histograms are merged together to form a parent cluster. Consequently each parent node has exactly two child nodes, forming a binary tree.

The CAT system in [18] first uses agglomerative clustering to group images initially by the keywords associated with them, and then creates internal clusters based on colour and texture features, again through the application of agglomerative clustering. Borth *et al.* [5] also use agglomerative clustering for their Navidgator system, which allows browsing through a dataset of video stills.

Pecenovic *et al.* [51] employ a hierarchical form of k-means clustering, where nodes are successively split as proposed in the LBG algorithm [41] to form a tree structure that can be visualised and browsed.

A hierarchical structure can also be derived without the application of an actual clustering algorithm. This is demonstrated by Schaefer and Ruszala in [73] and [72] who perform a uniform quantisation type clustering based on the definition of a grid structure for visualisation. Once the grid is defined, each image in the dataset will fall into one of the grid cells. Each grid cell hence corresponds to an image cluster. A spreading algorithm as in [59] is applied to reduce the number of unused cells and the number of images assigned to one partiular cell. When multiple images are mapped to a particular cell, a tree structure is formed by subdividing each cell into further uniform partitions with the spreading algorithm being applied to the root grid and all child grids in order to prevent the addition of unnecessary levels in the hierarchy. Based on this structure, using a grid of 24 × 30 cells and an assumption that 40% of the cells are assigned images, the system could visualise $((24 \times 30) \times 0.4)^3$, i.e. over 23 million images.

Selection of Representative Images

For visualisation purposes, each clustered group of images needs to be represented either by a single image or perhaps a small group of images. The manner in which these representative images are selected can vary between systems. In many approaches (such as the one in [72]), the centroid image of the cluster is selected. Formally, this is the image with the minimal cumulative distance from all other images in the database. Alternatively, other systems such as CAT [18] select the image closest the the centroid of the cluster in the feature space. A similar approach is adopted in PhotoTOC [55]. Here, to derive the most representative image of a cluster, the Kullback-Leibler divergence between every image histogram in the cluster and the average histogram for all images in the cluster is measured. The image with the colour histogram closest to the average histogram of the cluster is selected to be the representative image.

A cluster may also be visualised using more than one representative image. For example, the clustered visualisation of web search engine results generated by Liu *et al.* [42] displays a cluster preview of 4 images. Another content-based representative image selection scheme with the ability to display several representative images is that of Krischnamachari and Abdel-Mottaleb [35]. Based on a user-defined number of representative images R, a set of representative images R_n is formed. If $R = 1$, then the representative image is the leaf node of the sub-tree with the feature vector closest to the average feature of all images in a conjoined set R_0. When $R > 1$, image selection is taken from several subsets of images. Referring to Figure 6, if the user requires $R = 2$ representative images for cluster 14, the subsets will be $R_0 = 1, 2, 3, 4$ and $R_1 = 5$. The image most similar to the average of R_0 will be selected, together with the sole image from R_1.

While the works of [18, 55, 72] use content-based analysis in order to select a representative image for the cluster, the Calendar Browser in [19] chooses representative image(s) based on time. The system displays a summary of 25 images at any granularity (i.e. year, month or day). This 25 image summary

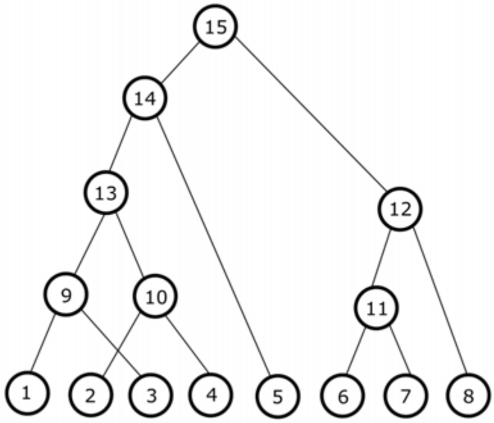


Fig. 6. Example of the hierarchically clustered image database arranged as a binary tree in [35] (© 2009 IEEE)

is created using a two step process. The first step is screen space assignment, where one space is assigned to each cluster. If there are too many clusters, priority is given to large clusters. Any remaining spaces after the allocation of a single space to each cluster are divided amongst the clusters according to their size. This creates a target number of photos for the second step, which performs the actual selection of representative images. The first criteria for selection is based on consecutive images with the smallest time difference, since it is likely that images taken close together describe the same event. One of these two consecutive images is then selected as the representative. If from the first step more than one representative images are required, the largest time difference between images is used, which will typically signify a new event and the second image in this pair will be selected.

Discussion

Clustering-based visualisations have the advantage that a user is given an overview of all images contained within the database at the top level of the hierarchy without displaying each individual image. This gives a good summary of the database. In addition, clustering can be performed in a hierarchical way leading typically to a tree structure representation of the database. As the user traverses this tree, the images become more similar to each other, and hopefully also more suited to the type of image the user is browsing for [5]. One downside of this approach is that if an image is erroneously clustered by the system (i.e. is assigned to a cluster of images that are not very similar to it), it will make that particular image very difficult for the user to locate, effectively making it lost. An example of this would be searching for an image of children playing football in a park, in a database that clusters based on colour similarity. In such a system the image in question might be clustered together with images of plants due to the green colour of the grass the children are playing on. If the chosen representative image is then also one of a plant, intuitively the user may not think to navigate into that cluster.

2.3 Graph-Based Visualisation

Graph-based visualisations utilise links between images to construct a graph where the nodes of the graph are the images, and the edges form the links between similar images. Links can be established through a variety of means including visual similarity between images, or shared keyword annotations. Once a graph has been constructed, it needs to be presented to the user in a visualisation that allows for intuitive browsing of the database.

Mass Spring Visualisation

Dontcheva *et al.* [13] use a mass spring model to generate the visualisation. A spring is formed between two images if they share an associated keyword. The length of the spring is assigned based on the number of images sharing the same keyword and a constant used to control the density of the layout. To generate the layout, the visualisation is evolved using the Runge-Kutta algorithm [3]. The authors conclude that this technique is only suitable for relatively small databases of a few hundred images due to the time required to stablise the arrangement. Worring *et al.* [83] also created a mass spring visualisation [6] based on keyword similarity (the number of keywords a given pair of images have in common). A k-nearest neighbour network is then formed based on this similarity measure. In order to visualise this high-dimensional structure in 2D, connected images are placed closer together while unconnected images are moved further apart. This is achieved by applying attractive or repulsive forces respectively between the images. The authors claim that this visualisation technique aids particularly when implementing a category search (i.e. searching for an image of a particular class), due to the fact that an image selected by the user will have nearest neighbours most relevant based on keyword. For example, selecting a picture of a cat with an associated keyword "pet" could present the user with images of dogs, cats or any other domesticated animal. A set of user interactions are available, designed to reduce the amount of effort required to form a subset of purely relevant images (see Section 3.1). User simulated tests were performed using only relatively small visualisations of up to 300 images. The main difference between the visualisations of [13] and [83] is that in the system of Worring *et al.*, the links between images are explicitly displayed whilst Dontcheva *et al.* do not display such links.

Pathfinder Networks

The use of Pathfinder networks [6] for image browsing was introduced in [7]. The interface, fronting an image database named InfoViz, was used in conjunction with QBIC [15], allowing the user to query and browse the database. Pathfinder networks were originally used to analyse proximity data in psychology, although many other types of high-dimensional data can also be represented using this technique [6]. The underlying theory behind Pathfinder networks is that a link between two items exists if it is the shortest possible link. The Pathfinder algorithm removes all but the shortest links by testing for triangle inequality. In the case that this does not hold, the path is considered redundant and is removed from the network.

For the layout of the network, images with many links between them are considered similar and therefore placed closer together, while images with fewer links are generally located further away. Chen *et al.* inspect the visualisations produced using colour, texture and layout features from the images and state that colour (through use of a colour histogram) provides the best visualisation, achieved using a spring-embedder node placement model. Figure 7 shows an image database visualised using colour histograms in such a Pathfinder network, where images with similar colour histograms form clusters. The experiments in [7] were implemented on a database containing 279 images. With such a small image collection, it is difficult to predict how well Pathfinder network visualisations may scale to larger image database sizes.

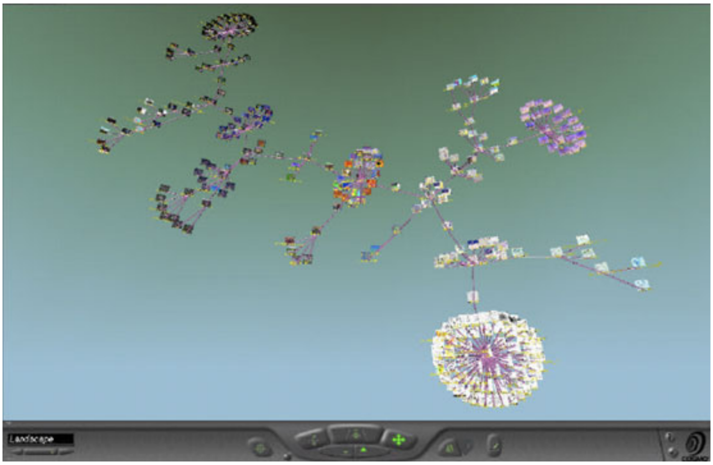


Fig. 7. An image database visualised using a Pathfinder network based on colour histograms [7]

NN^k Networks

NN^k networks, where NN stands for nearest neighbour and k describes a set of different features, were proposed by Heesch and Rüeger [22] to browse through an image database. The basic principle is that a directed graph is formed between every image and its nearest neighbours if there exists at least one possible combination of features for which the image is the top ranked of the other. Seven different features were extracted, including an HSV colour histogram, a colour structure descriptor (for detailing spatial formation of colour), a thumbnail feature (where the image is scaled down and gray values calculated), three texture features and a 'bag of words' (stemmed words taken from text attributted to images) feature.

A weight space is used which is a pre-defined set of weights for each of the features. The number of weight sets for which an image is top ranked, forms the similarity measure between images. For example, assuming that three weight sets are defined together with a query image Q, then if image A is ranked top in the first image set, but image B top in the second and third weight sets, image B will take a higher proportion of the weight space and therefore is deemed more similar to Q than image A.

Each image in the network stores its nearest neighbours, along with the proportion of the weight space for which the image is ranked top. Given a query image, a user-defined number of nearest neighbours will be displayed to the user, as well as links between the neighbours. Images with a higher similarity (i.e. a higher proportion of weight space) are displayed closer to the query which is centralised on the display. The initial display to the user is an overview of the database, generated by clustering the images and displaying the most representative thumbnail from that cluster (as described in Section 2.2). Figure 8 shows an example of how the network is visualised after an image has been selected as a query. In their experiments, the authors used a database containing 34,000 video stills.

Heesch and Rüeger [23] also describe their system's ability to query a database through keywords. In the example of Figure 9, searching the database with the query "airplane taking off" returns a variety of results. The top matching image is placed at the centre of the interface, with the nearest neighbors placed along an Archimedean spiral according to the proportion of the weight space they possess in terms of the query image. Images closer to the centre of the image are larger in size than those on the periphery of the spiral. The user can drag these smaller images closer to the centre where they are dynamically resized and can be inspected more closely by the user. They may then select multiple images to further the query.

Discussion

The use of graph-based visualisations appears to be less common than either mapping-based or clustering-based visualisations. Graph-based visualisations

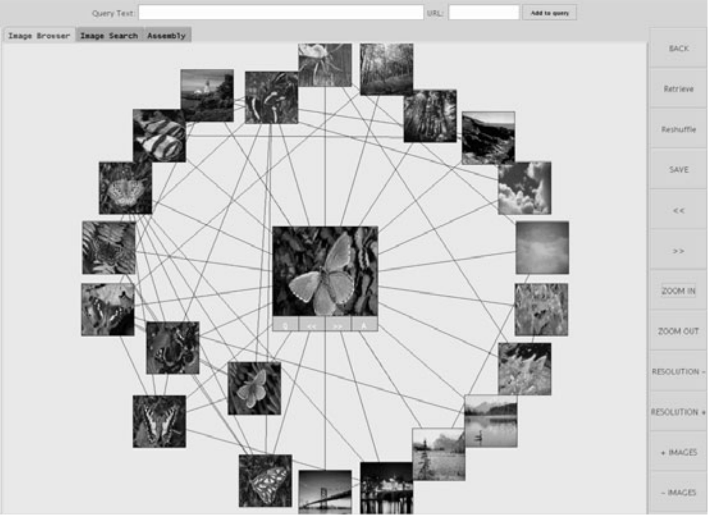
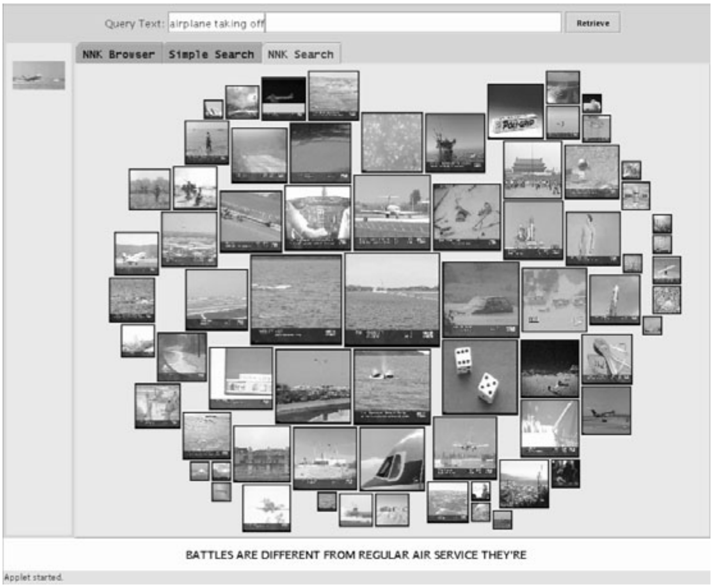


Fig. 8. An example of a an NN^k query selection taken from [22]


Query Text: airplane taking off

BATTLES ARE DIFFERENT FROM REGULAR AIR SERVICE THEY'RE

Applet started.

Fig. 9. An example of a query for "airplane taking off" in the interface devised in [23]

are typically quadratic in complexity, and therefore can only be computed offline in order to allow for real-time browsing. Generating query results ‘on the fly’ is not particularly suited to this style of visualisation. As with dimensionality reduced or clustered visualisations, the introduction of additional images in the database often requires re-calculation of the entire structure.

The major approaches in graph-based visualisations use contrasting visualisation methods. While mass-spring models and the Pathfinder network present a global visualisation similar in form to that of mapping-based techniques, NN^k visualisations present images one by one, allowing users to make an interactive choice on the next image to pursue. This is closer to traditional QBE methods, although the implementation of similarity by proximity should aid the user more than a linear format. Whilst NN^k networks can have a vast number of links (dependent on the size of k), Pathfinder networks attempt to minimise the number of links between images. So far, no study has been undertaken to explore which graph would allow for faster retrieval through browsing. It would also be of interest to see how well Pathfinder and Mass Spring networks are able to visualise larger databases, such as that used for testing the NN^k network.

2.4 Virtual Reality-Based Visualisation

The development of image browsing interfaces has also produced some interesting approaches based on the use of virtual reality (VR) equipment and software. Rather than limiting the user to traditional input hardware such as mouse and keyboard, work has been conducted using more interactive devices such as head tracking equipment [82] and the use of input wands [82, 47]. In general we can divide VR-based image visualisation techniques into two classes: immersive and non-immersive visualisations.

The 3D Mars system [47] visualises an image database in 3 dimensions. Images are projected onto four walls (left, right, front and floor) of a CAVE environment [10] around a user wearing shutter glasses. The interaction with the system begins with a random assortment of images from the database. As the user moves between the walls, the images rotate to face the user in order to prevent images being hidden. A virtual compass is provided on the floor allowing the user to ‘fly’ through the 3D space. The user can select a starting query image from the random assortment via use of an interactive wand. Each of the dimensions in the display represents either colour, texture or structure features. The query image selected by the user is placed at the origin of the axis, and all similar images are visualised in the space dependent on their distance from the query image. This visualisation is generated through the application of FastMap, as described in Section 2.1. The novelty of the system is that it makes interaction much more interesting for the user. However, unfortunately the system does not have the functionality to visualise an overview of the entire database.

In [82] the authors present their system StripBrowser. Images are arranged upon filmstrips and can be ordered using colour content along a rainbow scale or from light to dark. A user can navigate along each filmstrip using a head-tracking device (see also Section 3.1). An issue with projecting all images along one dimension is that the user must browse each image sequentially. This will require more user interactions compared with various approaches that use a 2D or 3D visualisation space.

Non-immersive VR image browsing systems create a virtual environment for users to navigate around in order to view the images in the database. Tian and Taylor [80] use MDS to plot 80 coloured texture images in a 3D space. The images are wrapped to spheres and plotted at the locations derived by MDS based on features vectors comprising a PCA projection of a colour and texture histogram. The user can navigate through the 3D space using a control panel located at the bottom of the screen. An issue not tackled by this system though is the potential overlap of spheres, presumably not occurring within the small database used for testing.

A different non-immersive VR image browsing system is presented by Assfalg *et al.* in [2]. Here, a graphical environment allows the user to move around a virtual world taking photographs of scenes in order to query the database. Upon loading the environment, pre-defined shapes are randomly placed within the scene. The user can 'walk' through the environment using navigation icons located on a panel on screen. The user may then edit the objects in the environment, having the ability to add new pre-defined shapes to the current scenery, and to texture and colour the shapes as desired. Shapes in the prototype system presented include a variety of tree like structures, statues and buildings. The user can select a rectangle over the current view in order to take a photograph, something the authors state as an intuitive metaphor for the user. The selected portion of the scene within the rectangle is used as a starting point for an adjoined QBE system, which retrieves all similar images from the database. Textures can be taken from results retrieved through QBE and applied to the environment in order to achieve modified results. Figure 10 shows a set of possible interactions a user may have with the browser.

3 Browsing Image Databases

In the previous section of this chapter we looked at how the often large image databases can be visualised and presented to the user. Although usually closely related, browsing the database is not the same as visualising it; Webster's dictionary defines browsing as *"to look over casually"* and visualisation as *"putting into visible form"*. A variety of tools have been developed which aid the user in order to interactively browse the images in a database. In this section we review common tools included within image database navigation systems to aid in the task of ultimately arriving at images of interest in an effective and efficient manner. We divide browsing methods into horizontal

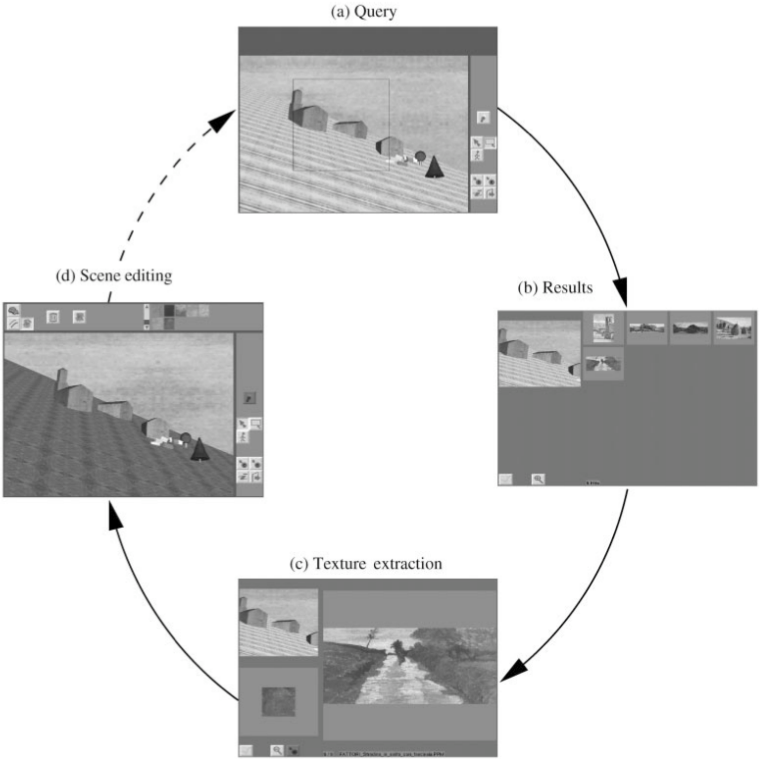


Fig. 10. Interactions available in the VR browsing environment presented of [2]

browsing, which presents images on the same level of a visualisation to the user, and vertical browsing, which can be used to navigate to a different level of the collection. Graph-based visualisations are typically browsed by following the links between images. For systems that organise images based on time stamps, browsing methods should also take this information into account. Finally, browsing can also be usefully employed in relevance feedback mechanisms.

3.1 Horizontal Browsing

We can define horizontal browsing as the navigation within a single plane of visualised images. This type of browsing is often useful when an image database has been visualised either through a mapping scheme (as described in Section 2.1), a single cluster of images (Section 2.2) or through a graph-based visualisation (Section 2.3). Several tools have been developed in order to support this browsing experience.

Panning

If the entire visualised image collection cannot be displayed simultaneously on screen, a panning function is required in order to move around the visualisation. There are a variety of different ways in which panning can be implemented within browsing interfaces. The simplest manner in which a user can pan through an image collection is through the use of traditional scroll bars. This is particularly the case when images are arranged in a regular grid format, such as in the QBIC interface [15]. If possible, scrolling should be limited to one direction only in order to reduce the number of actions required by the user to browse the entire collection [54]. An alternative to scroll bars is the use of a control panel for panning (and zooming) as implemented in various approaches. The systems described in [7, 72, 80] all provide such a navigational toolbar enabling the user to browse through the visualisation space.

The hue sphere system by Schaefer and Ruszala [72] allows for intuitive panning by the user, as it uses the metaphor of a globe in order for users to browse the images in the collection. Images are plotted along the latitude of the globe according to the average hue of the image, whilst the average value is used to plot the image upon the longitude of the globe. The user is able to spin the globe about either horizontal or vertical axes, in order to bring images into view. This is illustrated in Figure 11, showing an image collection after various rotation/panning operations by the user.

StripBrowser [82] also allows for intuitive panning using a head tracking device. As the user looks to the right side, images are being scrolled to the left, while looking to the left causes the scrolling of images to the right. The greater the angle at which a user moves, the faster the scroll motion will be. Scrolling only occurs when the angle reaches a threshold value, otherwise the strip remains stationary. The 3D Mars system [47] allows users to pan the generated visualisation by walking around the 3D space projected on the four CAVE walls.

Zooming

When presenting many images on a single 2D plane, the thumbnail representations of images often have to be reduced to small rectangles which are difficult to distinguish on the screen. This can be seen in the MDS plot in Figure 1. There, although it is possible to see that the images vary in colour, it is not possible to depict the content of each individual image. It would therefore be useful to have a facility to zoom into an area of interest.

For dimensionality reduced visualisations, Rubner *et al.* proposed zoom operations on a global MDS visualisation, using of a joystick in order to *"get closer to the area of interest"* [64]. Another example of a dimensionality reduced visualisation with a browsing interface facilitating zooming is the CIRCUS system presented in [51]. CIRCUS uses multi-dimensional scaling, in particular Sammon mapping [67], in order to present representative images

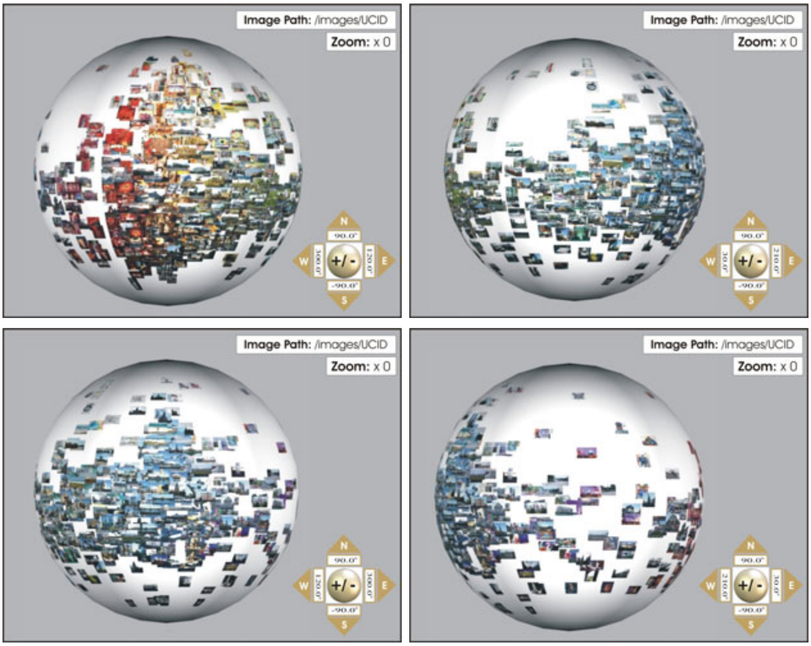
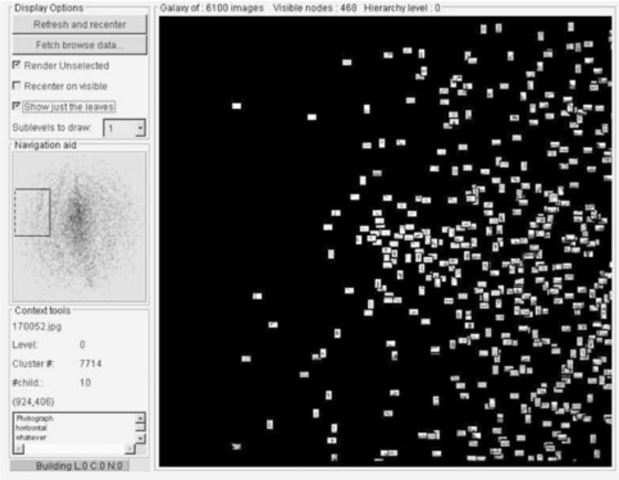


Fig. 11. Browsing through UCID images [74] of different hues in the system proposed in [72]

at each level of the hierarchically clustered database (where clustering is based on content). By selecting a 'Browse Collection' tab, users are presented with a browsing window which, at the minimum zoom factor (i.e. zoomed out as far as possible), shows the Sammon mapping layout of all images at that level of the database.

To reduce the amount of computation required, and to allow the user to browse the database interactively, CIRCUS displays images at the minimum zoom factor simply as dots. This maintains the user's understanding of the relationships between the representative images at this level of the database, while reducing the amount of processing time required by the system. As the user zooms into an area of interest, thumbnails are rendered. When the user focusses on a single image, metadata associated with that image is also presented. Further zooming on a particular image causes CIRCUS to present the next level of images in the hierarchy and hence implements vertical browsing (see Section 3.2).

CIRCUS also implements what is described as a *"fixed small overview"*, preventing the user from becoming lost in a 2D space larger than that of the display area. A detail view is provided, where the images are shown together with an overview displaying a map of the overall visualisation (with



Galaxy of : 6100 images Visible nodes : 468 Hierarchy level : 0

☑ Render Unselected
☐ Recenter on visible
☑ Show just the leaves
Sublevels to draw: 1

170052.jpg
Level: 0
Cluster #: 7714
#child.: 10
(924,406)

Photograph
horizontal
whatever

Building L:0 C:0 N:0

Fig. 12. A screenshot of the CIRCUS browsing interface presented in [51]

the current area of focus highlighted). This is presented to the user in the left hand pane of the CIRCUS browser, shown in Figure 12.

Another system using a similar overview approach is the Photosim system presented in [8], allowing users to view and modify multiple clusters (described in more detail in Section 3.5). A zooming tool is also implemented in the hue sphere system devised by Schaefer and Ruszala [72].

While most zooming interfaces require the use of a computer mouse, a more novel approach to zooming is adopted in the StripBrowser system [82]. Zooming in and out is achieved by moving closer and further away respectively from the screen. The authors note that this is an ideal metaphor for users, as generally to inspect an item in the real world a person will move closer to it. Another implementation of this metaphor is provided in the fully immersive 3D Mars system [47], where the user can zoom in on an image by physically moving closer to the wall on which it is projected.

Hilliges *et al.* [25] provide a clustered visualisation with a zoomable interface. The user may zoom into particular clusters to examine the images within them more closely. In the graph-based system by Chen *et al.* [7] a control panel located at the bottom of browser window has two control buttons allowing the user to zoom the current view in the detail pane in or out. A very similar interface is also implemented by Tian and Taylor [80], allowing the user to zoom in and out of a 3D MDS visualisation of textured images.

Magnification

Although similar to zooming, magnification usually occurs when a cursor is placed over an image. This maintains the overall structure of the visualisation by rendering only small thumbnails for each image in the database at first, while higher resolution images are loaded only when required. An example of a system using mouse over magnification in order to dynamically display a higher resolution image is PhotoMesa [4], illustrated in Figure 4.

The hexagonal browsing system by Eidenberger [14] also provides a high resolution preview image when the mouse cursor is moved over any of the images in the visualisation. For the system created for the user studies conducted by Rodden *et al.* [61], a 3x magnification of an image occurs when the cursor is placed over an area of the MDS visualisation.

Another form of image magnification that can be used for examining image database visualisations is the application of a fisheye lens [69]. Using this magnification mechanism, images located at the centre of the lens are magnified whilst those immediately around the focused image(s) are distorted [54]. Figure 13 shows an example of how a fisheye lens could effect a collection of images fitted to a regular grid.

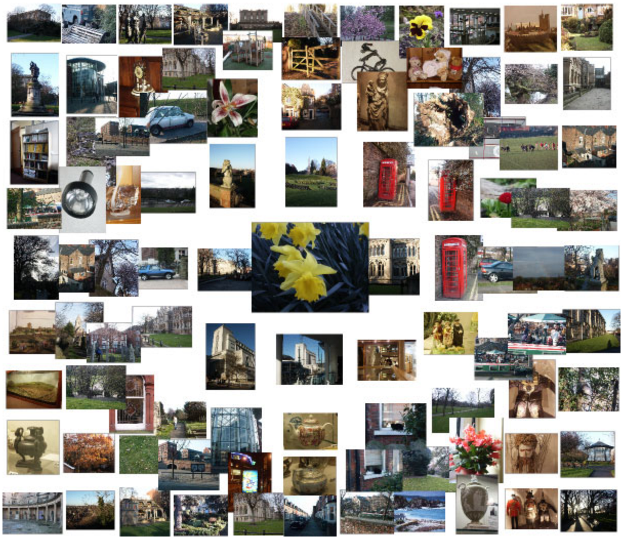


Fig. 13. Example of a fisheye lens browsing over images from the UCID dataset [74]

Scaling

Some browsing systems use scaling, rather than zooming, allowing users to view a particular image in more detail. In the EIB (Elastic Image Browser) system [57], the user may use two slider tools in order to dynamically resize images both horizontally or vertically. This enables the user to display more images in the browser, at the expense of image clarity. Images can also be portrayed as lines, with the colours in the lines becoming the only distinguishing feature between images. The author claims that this could potentially speed up browsing. However, in the EIB visualisation, images are not arranged by mutual similarities; rather they are placed randomly within the grid visualisation leading to a negative effect in terms of the user's browsing experience.

The PDH system [45] also includes a slider tool so that the user may dynamically resize images, whilst maintaining the 2D spatial relationships between images achieved through PCA.

3.2 Vertical Browsing

In visualisation approaches that are based on a hierarchical structure, the contained images can also be navigated using vertical browsing methods. As discussed in Section 2.2, clusters of images are typically visualised through the use of representative images. These images are crucial for vertical browsing as they are typically the reason for which a vertical browsing step into the next level of the hierarchy is initiated by the user.

In the quantum treemap visualisation of image clusters provided in PhotoMesa [4] (shown in Figure 4), the user may click on a highlighted image in order to invoke a smooth zoom into that group of images. The box around the selected cluster remains highlighted to prevent user confusion. Zooming may continue until a single image is displayed at full resolution. The CAT system [18] provides a functionality similar to this.

In systems using a regular grid structure at different levels, such as the hierarchical hue sphere by Schaefer and Ruszala [72], or the hexagonal browsing system by Eidenberger [14], selecting a representative image at a given layer of the hierarchy will present the user with the subsequent layer of the subtree, for which the selected image acts as the root. The user may traverse all layers of the tree. The hue sphere system also displays a visual history of the followed browsing path, while in the hexagonal browsing system, the user is presented with a view of both the previous layer and a preview of the layer described by the currently selected cell. A similar combination of history and preview is included in the PhotoMesa system [4].

A navigational history is also provided in the Navidgator video browsing system [5] shown in Figure 14. Here, the user's most recent image selections are displayed in the top right hand corner of the interface, and may be revisited by selecting one of the thumbnails. The representative images at each level are displayed in the lower portion of the screen. Selecting an image creates a larger preview just above the layer viewing portion of the interface, and also adds a thumbnail of the image to the history. The user may then *zoom* in or out of the levels in the database using arrow buttons. Single arrows move the user up or down a single layer of the database (up to the previous layer, or down to the first layer of the sub-tree for which the currently selected representative image acts as the root) while double arrows enable the user to perform a *multi-level zoom*, whereby every third layer of the tree is displayed. A *max zoom* function is also included which allows the user to navigate directly to the bottom or top layers of the tree.

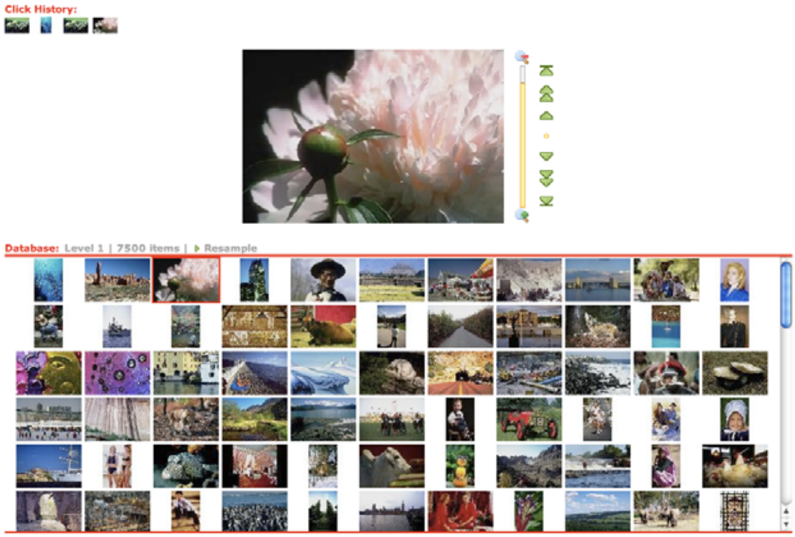


Fig. 14. A screen shot of the Navidgator system detailed in [5]

A different vertical browsing function is implemented in the CIRCUS system [51]. The authors introduce a semantic zoom facility which allows users to zoom into areas of interest. As the user zooms into a representative image beyond a certain zoom factor, the sub-clusters associated with the cluster of interest are automatically displayed.

3.3 Graph-Based Browsing

Operations such as panning and zooming can also be applied to graph-based visualisations. For example, in the Pathfinder network approach by Chen *et al.* [7], a global view of the network is presented (as shown in Figure 7). Displaying this global representation of the structure bears some similarity with some of the mapping-based visualisations of image databases from Section 2.1. In the Pathfinder system, a toolbar is displayed at the bottom of the browsing window, which the user can use to zoom into areas of interest or to pan around the collection. Images found through browsing may then be selected in the interface to be used as a query for the QBIC system [15].

The structure of the graph itself however also allows for different methods to browse from image to image. This is realised in the NN^k network approach by Heesch and Rüeger [22] which exploits the links between images in the graph. First, the user is presented with an overview of the database through representative images of clusters formed through a Markov chain clustering. The user can then select one of these images as a query image. As shown in Figure 8, the selected image is placed at the centre of the screen and a user defined number of nearest neighbours are placed around it based on their similarity to the query image. Furthermore, links between these neighbours are also displayed. Selecting a neighbour will put it as the query image in the centre, with its nearest neighbours then presented in a similar fashion. This process can then be repeated until a required image has been found. The user also has the ability to zoom in or out of the visualisation. As with a typical web browsing interface, a 'Back' button is provided so that the user may return to the previous query, should the newly selected image provide no images of interest.

Browsing the graph-based visualisation developed by Worring *et al.* [83] lies somewhere between the two browsing methods of the NN^k and Pathfinder visualisation styles. While the overview of the database is presented, the user may invoke one of five different actions in order to select a subset of images that are deemed relevant. A user may select a single image, as well as selecting the single image and all of the linked neighbours in the network in a single action. The opposite two actions are also available, whereby the user can deselect a single image (and disable it from future automatic selections) or an image and the associated neighbours. Another action available to the user is the ability to expand the current selection of images by automatically selecting all connected neighbours. A simulated user test showed that the provided interactions can reduce the amount of effort required to select all possible relevant images (compared with selecting images one-by-one). Worring *et al.* conclude that using the functionality of selecting an image and all the nearest neighbours, followed by deselecting all images deemed irrelevant, a higher recall and precision measure can be achieved whilst maintaining the same interaction effort required for a one-by-one selection technique.

3.4 Time-Based Browsing

As described in Section 2.2, time stamp information attached to images can be used to cluster and visualise image collections. Clearly, if a collection is visualised based on temporal concepts, browsing should also be possible in a time-based manner. One of the earliest time-based image browsing systems is the AutoAlbum system introduced in [56] further developed into the PhotoTOC system [55]. Here, a two-level hierarchy based on time is utilised. As can be seen in Figure 15, dates, in monthly intervals, are shown in the overview pane on the left hand side of the interface, with the representative images of the clusters falling into that date also being displayed. Selecting a representative image displays the contents of that cluster in the detail pane, located to the right of the interface.

Whilst AutoAlbum and PhotoTOC restrict the user to monthly intervals, the Calendar Browser in [19] allows the user to 'zoom in' to other time intervals by selecting one of the representative images. At the year level, two controls located at the top of the interface provide a summary of the previous year and next year respectively. This approach is also adapted for the case

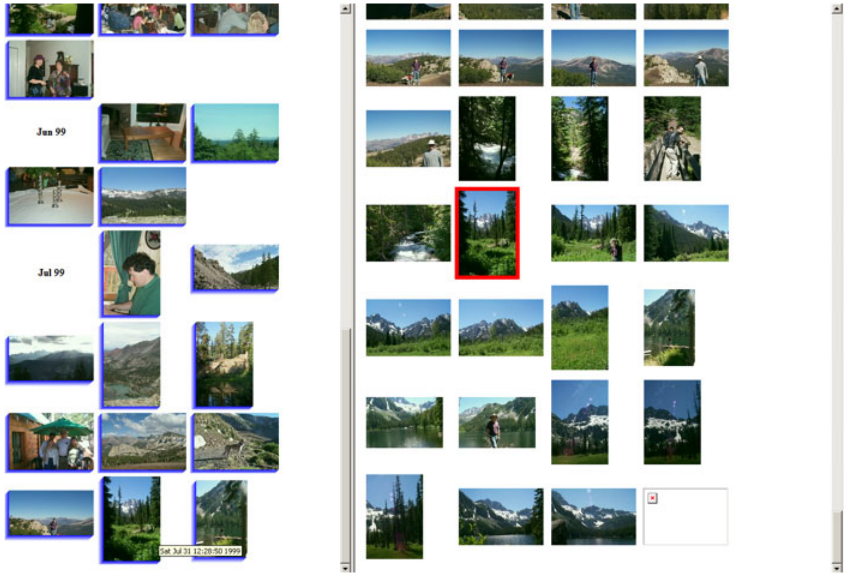


Fig. 15. The PhotoTOC [55] interface. The user has selected the image bottom and centre of the overview pane (left). This image has been highlighted in the detail pane(right), and is amongst visually similar images.

when viewing images at a monthly granularity (i.e. summaries of the previous and next month are displayed). When viewing images with a time stamp attributed to a particular day, images maybe browsed 25 at a time. Selecting an image at this level places it in the centre of the interface, with the images taken immediately before or after displayed around the selected image.

In [19], a modification of the Calendar Browser is also implemented and tested. In the modified interface Hierarchical Browser, a pane located on the left hand side displays a hierarchy of dates. Starting at root nodes representing years, these can be expanded to display monthly nodes, followed by dates and time intervals. Selecting a node from this pane displays the representative images in the detail pane located on the right side of the interface. This is similar to the approach of PhotoTOC [55]. User testing suggested that users could use the Calendar Browser more quickly, but the number of task failures occurring was lower in the Hierarchical Browser.

A different approach to time browsing is presented in the PhotoHelix system [24]. An interactive touch screen table top is used with an interactive pen and a specially developed piece of hardware created using the workings of an optical mouse and an egg timer. By placing the hardware on the interactive screen, a virtual helix is created at the location of the hardware. Images are arranged on the helix according to time, with newer images being located closer to the outside of the spiral. Grouped images, known as piles, are magnified when the spiral is rotated under a fixed lens. The magnified group can then be manipulated by the user through use of an interactive pen. New groups can be formed or individual images scaled, rotated and moved freely around the interactive screen.

The table top PDH system [45] also allows users to browse images according to time. By selecting a 'Calendar' button, images are sorted along a linear timeline.

The hexagonal browsing system of Eidenberger [14] allows users to swap between a content-based and a time-based tree structure. As the time-indexed tree has all the key frames of the collection visualised (as described in Section 2.1), any cell selected in the content-based tree will have a correseponding cell in the time-based tree. However in the content-based tree, images may only occur as leaf nodes if they have not been selected as representative images for clusters. Therefore, when switching between an image in the time-based tree to the content-based tree, the leaf node of the corresponding cell is selected and a message is displayed to the user in order to minimise confusion.

3.5 Browsing-Based Relevance Feedback

As described in Section 1.3, many CBIR systems use some form of relevance feedback (RF) in order to tailor the search towards the current user's needs. The most common mechanisms are the standard relevant/non-relevant classifier, used in QBIC [15], and a slider tool whereby images can be given a continuous score of relevance by the user, as demonstrated by the MARS system [65]. However, the introduction of novel image database visualisations have also led to the development of new RF mechanisms.

In the PDH [45] and the El Niño [68] systems, the intrinsic weightings of feature vectors are modified by allowing the user to manually specify where images should reside in the visualisation. PDH provides the user with a small subset of images to be placed as they wish on a *"user guided display"*. Based on the user layout, PDH uses the location of images in order to estimate feature weights for colour, texture and structure. Using these weightings, a larger image collection is then presented based upon their provided layout. Figure 16 shows a user guided layout on the left, and an automatic layout of a larger set on the right.

The El Niño system [68] allows users to manipulate the entire visualisation, rather than just a subset as in PDH. Images presented to the user may be moved to modify the internal weightings of the system. Each image manually relocated by the user is considered an anchor. The distances between anchors are then used to modify the colour, texture and shape feature weights. The visualisation is then updated based on the new similarity measure. As only a subset of images is shown to the user (typically 100-300), the updated visualisation may lose images which were not selected as anchors by the user. A possible issue with these systems is it may not be clear to the user

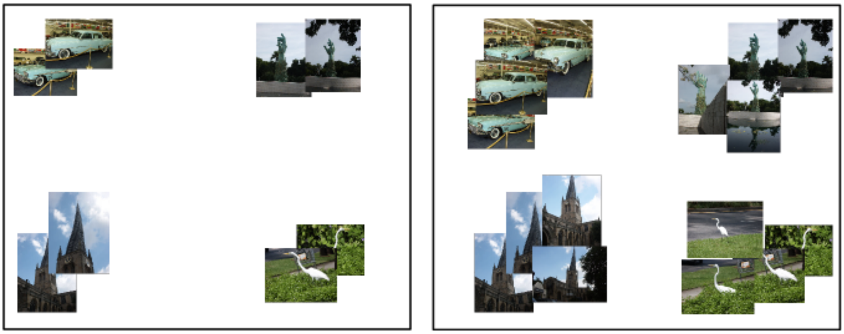


Fig. 16. A user guided layout of UCID images [74] in PDH (Personal Digital Historian) [45]

how far relocate images in order to modify the system to meet their search requirements [48].

Another category of mapping-based visualisations with an example of an RF implementation is the self-organising map based system PicSOM [37] (described in Section 2.1). Here, the user selects images as either relevant or irrelevant. The images, and their user determined relevance, are projected to SOM surfaces in order to find regions of relevant or irrelevant images. A low-pass filtering system is used to expand the regions of relevance on the map. A qualification value is assigned to each image based upon the relevance of the image and surrounding images. Each SOM is searched for the top 100 images with the highest qualification value. The top 20 images from the combined set are returned to the user, from which the process can be repeated if necessary.

While the above systems use RF within a dimensionality reduced visualisation, there have also been clustered visualisations with integrated RF mechanisms. An example of this is Photosim [8], shown in Figure 5. Whilst higher level clusters are created based on time, images within the same time period are clustered on content. Photosim allows users to transfer images between clusters manually, if they are not satisfied with the automatically formed groupings. Furthermore, the user also has the ability to create entirely new clusters. Using a slider tool, the user can alter the degree of similarity in which images are automatically added to the new cluster. Setting the slider to zero creates a cluster with just a single image, dragged from an existing cluster. The higher the threshold value, the degree of similarity required in order to add new images to the cluster is lowered.

Similar approaches with solely manual clustering occurs in the EGO [81] and ImageGrouper [48] systems. In EGO (Effective Group Organisation), a manually created grouping of images retrieved through some search (such as QBE, or keyword-based search) can be defined. EGO then recommends other images in the system by treating each image in the group as a positive training example, and modifying the feature weights to find similar images in the database. ImageGrouper adopts a different approach in that manually created groups can be selected as positive, negative, or neutral examples. The system will then return images based on the positive examples given by the user. Sub-groups can be made within groups in order to narrow the search. For example within a group of car images, the user may narrow the search by selecting only the red cars in the group as positive. The manual groupings in these two systems allow for *bulk annotation*. Instead of labelling each image individually, the user may simply annotate the entire group with keywords, in order to facilitate future keyword searches.

The fully immersive 3D Mars system [47] also has a RF mechanism incorporated, allowing the user to choose positive or negative examples using an interactive wand. The system then modifies the weightings of the features used to query the remainder of images in the database.

3.6 Discussion

The browsing tools described in this section aim to aid the user during the navigation of an image database. While horizontal browsing can be applied to all visualisations where either a selection or all images in the database are displayed to the user on a single plane, vertical browsing is limited to hierarchically organised visualisations. The user is able to select a representative image to view a collection of images similar to their selected image. In this way, the user is presented with a subset of more similar images relating to their intended target. However, unlike horizontal browsing, once the user traverses down a particular path of representative images they can lose the overview of the database. Therefore, to reduce the user's cognitive load and to minimise confusion, systems will often give the user some indication of their current position within the database. An example of this is implemented in the Navidgator system [5], whereby a textual description includes the current level of the database being displayed and the total number of images in the current layer. A potential improvement to this would be a visual map, displaying the user's current location in the database.

The two contrasting styles of graph-based visualisations are providing an overview of the collection (as in the Pathfinder network of Chen *et al.* [7] or the approach by Worring *et al.* [83]) or presenting singular images in the database and their linked neighbours (as implemented in the NN^k network of Heesch and Rüeger [22]). Whilst such an overview may be explored in a similar manner to mapping-based visualisations, the NN^k implementation presents a selected image as a query centralised on the display, with its nearest neighbours displayed around it at distances based on similarity. Unfortuntately, this technique suffers from a drawback similar to that of vertical browsing. Once users enter the database (from selection of an initial query image from an overview formed of representative images of clusters from the database), they may become lost in the network. The only option available to the user is to use a 'Back' button to return to previous query selections or to the initial overview. Presenting the user with the entire network visualisation prevents this problem. The user interactions presented by Worring *et al.* [83] show that such browsing techniques can reduce the amount of user effort required to create a subset of solely relevant images. However, such visualisations have only been tested with up to 300 images and it is not clear how well they would scale for larger datasets.

Time-based browsing as implemented in a variety of systems, such as Calendar Browser [19] or PhotoTOC [55], are typically aimed at personal users, as they can recall the event at which an image was taken in relation to other events in the collection [62]. Browsing such systems assumes that the images in the database are correctly time stamped, which may not be the case for all image collections.

The development of browsing systems has also resulted in some interesting relevance feedback mechanisms in which the user can dynamically update the intrinsic similarity measure by moving the position of the image directly in the visualisation space. However, it is not clear to the user how much effect a particular movement may have upon the system [48]. The EGO [81] and ImageGrouper [48] systems require the user to manually form groups before suggesting to the user possible matches. The approach undertaken in Photosim [8] automatically creates initial clusters before allowing the user to modify them, an approach that typically requires less user effort.

4 User Evaluation of Image Database Navigation Approaches

As is apparent from the previous sections, a lot of research has been conducted aimed at providing intuitive navigation interfaces for users of image collections. Unfortunately, the systems most widely used in practise do not offer any of these approaches. Most users rely solely on a graphical interface of file structure browsers included in common operating systems, whilst others use commercial software such as Apple's iPhoto [27] or Google's Picasa [53] in order to display their personal photos. Professional photography agencies employ staff whose sole responsibility is to manually annotate images with keywords or free text, yet they also do not employ any of the techniques reviewed in this chapter. The low uptake of browsing systems is further hindered by the fact that traditionally only few examples of image management software invoke some use of CBIR techniques, and that CBIR itself still has major challenges to overcome. In this section we explore in more detail the various tasks for which image database navigation systems are particularly useful, reducing the time required to perform them, and review various user evaluation studies that support this argumentation.

4.1 User Tasks

Image browsing systems can be used for a variety of tasks and tests which we will highlight in the following. Typically, different approaches are more suited to particular tasks. For each task, specific data can be extracted in order to measure the performance of a developed system. In addition, the subjective opinion of users can also be measured.

Target Search

Target search [75] is the most commonly employed and tested task in browsing systems, and used in works such as [9, 59, 55, 19, 18]. It is also often used as a method of testing traditional CBIR systems [46]. In target search, the user is shown an image and asked to browse the system in order to locate this target image. When the user has found the image, they perform some test termination action (e.g. click on the target image in the system). The time taken for the user to locate the image can then be recorded for further analysis. A timeout is also often implemented (i.e. when a user is unable to find the image within a specified time limit). Clearly, the gathered timing information can be used to a compare different systems, or to compare some system against a traditional search through a linear list of images.

A variation of this task was used in [48] where users were shown a target image as before, but rather than locating that particular image, were asked to select 10 semantically relevant images from the collection. Apart from the timing information, another measure that can be derived in this test is the error rate, which counts the number of images incorrectly selected by users.

The advantage of a target search task is that it is relatively easy to conduct, and enables a quantitative comparative analysis of two or more systems. It is also more likely to model the more general use of a system, e.g. browsing personal photographs.

Journalistic Task

As outlined in [75], a common use of image retrieval systems for journalistic purposes are *"searches to illustrate a document"*. To replicate this within a user study, participants can be given a short piece of text in which they are instructed to find a set of images from the database which best represent the topic of the text. In the experiments conducted by Rodden *et al.* [61], ten graphic design students were asked to compare an interface with images arranged according to their visual similarity through MDS assigned to a grid structure, and an interface which grouped images according to keywords, namely the geographical location of where the image was taken. Users were issued a travel article based upon some tourist destination (such as New York) and were instructed to browse 100 images from that location and select three that they deemed the most appropriate to accompany the article.

Another study by Rodden *et al.* [61] compared a similarity-based approach with a randomly arranged image set, employing a test population of average computer users. The creators of EGO [81] conducted a similar study with general users.

Graham *et al.* [19] employed a modified version of the journalistic task, providing users with a textual description and 3 minutes to locate as many images in the database relevant to the description. While comparing presentation of image from web search engines in [42], the authors asked the users in the study to find those images out of the top 200 image results that best represent the query terms.

The journalistic task, models a true requirement of a retrieval system. An issue with this test however is, that is difficult to recruit users that would actually employ such a system in the real world (e.g. journalists) and evaluation is hence often performed upon general users who may undertake different search patterns to browse through image collections.

Annotation Task

Rodden and Wood [62] observed that users rarely manually annotate each individual image in a collection (or indeed do not annotate any images at all). One obvious reason for this is the amount of effort required for annotation. Systems such as ImageGrouper [48] and EGO [81] have hence been devised in order to simplify and speed up the annotation of images in a database. Nguyen and Worring [49] run a simulated user study, measuring the number of total interactions required by a user in order to annotate the entire database. They used this method to evaluate a mapping-based visualisation. The baseline number of annotations used is that of a standard linear visualisation which equals the number of images in the database (i.e. one interaction per image). It was shown that the mapping-based visualisation can reduce the number of interactions needed for annotation by up to 94% (dependent on the categories of images in the database and the features used to define similarity). Such a test would obviously be simple to implement in practice, asking users to annotate each image in the database with a keyword from a preset list while measuring the number of interactions or the time spent on the task. In addition, if a ground truth of correct annotations is available, the error rate can also be measured.

Clustering Study

A novel way of measuring the quality of image clustering is presented by Platt [56]. Two users each used their own personal collections (one of which had corrupt time stamps), and were asked to manually cluster the images into albums which acted as the ground truth for database. Each of the personal collections were then automatically clustered by either time, content, or a combination of the two (as well as a control of equally sized clusters). Each image in the database was used as a query, and based on that the automatic clusterings were compared with the ground-truth (i.e. user-based clusterings), using the number of true positives, false positives and false negatives. These were averaged for all the images in the database to generate a percentage known as the $F1$ metric. It was found that a combination of time and content-based clustering achieved the highest $F1$ score for both the corrupt and non-corrupt image collections.

This approach provides an interesting measurement of the quality of automatic clustering algorithms since it directly compares the results of automatic techniques with those derived manually by a user. The drawback is of course, the time involved to generate the ground truth clustering. In [56] the image collections consisted of 294 and 405 images respectively, whereas for collections of 1,000s of images the task will become not only infeasible, but also prone to human error.

User Opinion

After a user study has been conducted, researchers will generally issue the users with a questionnaire in order to gauge their opinions on the different aspects of the system and its user interface. The results of these questionnaires can then be used to modify the system as was done by Rodden *et al.* where the general dislike of users towards image overlapping in MDS visualisations caused the authors to consider fitting the images to a more regular grid structure. When this type of user questioning is included with some other task such as those listed above, it allows to gain an impression on how such a system could be applied in the real world. However, if used without a test such as those listed above, the lack of quantitative statistical data prevents drawing full conclusions upon the true quality of the approach.

4.2 Key Findings from User Studies

User evaluations attempt to prove that the system proposed by the authors improves upon methodologies currently used in the field. Sometimes these studies provide interesting insights into how general users gauge these novel browsing systems and additional functions. Perhaps the most significant user studies have been conducted by Rodden *et al.* In [59], a user study was conducted using target search on a randomly assorted grid of images and an MDS visualisation based on image similarity. The authors were able to show:

- Image retrieval is faster when images are arranged by their mutual similarity.
- Users prefer visualisations that do not overlap.
- More distinct images (i.e. images that are on average less similar to all other images in the database) are easier to find.
- Images located closer to the centre of the screen are retrieved faster than those located closer to the edge.

In a later work of Rodden *et al.* [61], an MDS visualisation fitted to a grid is compared first to a system organising images in groups through keywords, and then with a grid arrangement whereby the images are randomly sorted. The users were asked to perform a journalistic task. The key findings from this work were:

- Users prefer the MDS grid visualisation to one arranged through keywords.
- Users are slower at selecting preferred images within the MDS grid visualisation than the randomly assorted grid.

The authors were surprised that users took longer to select images for a travel article using the MDS grid visualisation rather than the randomly assorted grid of images. As a possible reason, they argue that when images are arranged randomly, images appear to be more distinct as it is unlikely that it will be similar to all of their neighbours. However, the authors state that judging from post-test questionnaires, it appeared that users were generally more satisfied with their image selections when using the MDS based interface. This may be because they have selected an image they were looking for in particular, rather than settling for a related image found quickly using the random arrangement.

Rodden and Wood [62] also explored how users manage their digital photographs. Subjects were supplied with a digital camera and a system called Shoebox, an image browsing system arranging images in folders according to the time they were created. Shoebox also has the added functionality of a QBE search facility and a voice annotation system. Findings from this work include:

- The general user has unrealistic expectations of a QBE system, and can find it difficult to improve a query.
- Users are fairly reluctant to manually annotate images, even when provided with a voice annotated system. Only a small percentage of users in the study changed the title of any image in the system.
- Sorting images according to the time at which they were taken allows users to browse the collection by recalling which particular event the image required is from.
- Displaying many image thumbnails at a time decreases the time required for image retrieval.

The work of Rodden *et al.* has looked at arranging images by similarity as well as time; one of the fundamental conclusions from [62] is that displaying as many images as possible to the user improves retrieval time. However an issue with displaying too many image thumbnails is that the user needs to be able to comprehend what is actually depicted within the image.

A zoom facility as described in Section 3.1, allows thumbnails to be displayed within an overview at a fairly low resolution and zoomed into at the user's discretion. A study by Combs and Bederson [9] investigates solely the effect zooming has on improving the user's browsing experience. A Zoomable Image Browser (ZIB) is compared with a traditional image browsing system, whereby image folders can be selected from the left hand pane, and the contents of the folder are displayed in the right hand pane. Enlargement of images can only be performed by opening a new window. The ZIB system provides a keyword search in a top pane, while the results of the search are displayed in a pane located below. The user has the facility to zoom into the search query results. Despite showing that a target search test was faster using ZIB, this was shown not to be statistically significant. The authors of the study also comment that of the 30 users tested, only 50% actually invoked the zoom facility. Combs and Bederson suggest that the number of query results shown at any time was not enough to warrant a zoom facility, as images were displayed at a resolution distinguishable without the zoom requirement. They conclude that a study into the maximum number of images that can be displayed without zoom should be investigated in future work.

Interesting user studies have also been conducted by the developers of the RF browsing systems EGO [81] and ImageGrouper [48]. Both systems are relatively similar, allowing a user to first query the database, then group the images presented in the results in order to modify the feature weights of the internal similarity measure. A target search was performed, asking the user to select ten semantically relevant images to the target. EGO and ImageGrouper were compared with a slider based and a selection based RF system. The key finding from these studies was:

- Image retrieval took longer using the grouping systems rather than the simple relevant or non-relevant selection system.

The authors attributed this to the fact that the drag-and-drop interfaces require more user actions than a simple selection interface.

In clustering-based visualisations, image groups can be displayed to the user in the form of representative images (as discussed in Section 2.2). In [18], the subjective opinion of users was measured for varying forms of the CAT interface. The authors found that:

- Users preferred the CAT interface when representative images were used rather than when representative images were not included.

Unfortunately the evaluation was performed only with 10 users, making it difficult to conclusively state that representative images indeed do improve a user's browsing experience. Future work could use quantitative tests in order to provide a better insight into the effectiveness of representative images as a tool for browsing.

While the CAT interface is an example of a system invoking hierarchical clustering, user studies conducted by Rodden *et al.* were designed to test the effectiveness of dimensionality reduced visualisations. Relatively little work has been conducted into testing which of these different approaches might perform better in terms of image retrieval. Liu *et al.* [42] however do offer some comparison between a clustered and dimensionality reduced approach. The authors were interested to discover how the results from a web image search engine query can be visualised in order to improve the user's browsing experience. Nine users participated using the standard ranked list interface provided by Google image search, an MDS visualisation that could be manually adjusted to fit images to a grid (as described in Section 2.1) and a clustering-based approach which creates five groups of images based on content. Clusters are represented in a left hand pane through a set of the four most representative images in that group. Should the user select the cluster preview thumbnail, all the images from that cluster are displayed in the right hand pane of the browser. 17 queries were performed in which the top 200 images were displayed on each of the interfaces. Users were instructed to browse the results in order to find the images they deemed most relevant to the query terms. Liu *et al.* found that:

- Both the MDS and the clustering-based visualisations clearly outperform the standard ranked list results.
- Although search times were similar between the MDS and cluster visualisations, users clearly preferred the layout of MDS plots characterising it as *"more intuitive and interesting, also convenient for comparing similar images"*.

4.3 Discussion

Evaluation of image database navigation systems, more often than not, tends to compare the newly proposed browsing system with a more traditional approach. Various studies have confirmed that image databases visualised, as described earlier in this chapter, do indeed allow for faster retrieval than traditional linear approaches [19, 42, 59].

What the majority of user studies have not been able to show is how their browsing system can perform against other browsing systems discussed in this chapter. For example, little work exists in comparing a hierarchically clustered visualisation of a database against the same database visualised using MDS. While the study of Liu *et al.* [42] does offer such a comparison based on a target search scenario, the results were too close to conclude which of the two paradigms offers a more efficient way of searching.

Analysis of questionnaires returned by users after testing allows the collection of personal preferences. However, unlike e.g. the time required to perform a test which can be statistically analysed, user opinion is highly subjective and is often linked to the background and environment of the test population. In addition, the majority of user studies conducted are based on a relatively small number of participants, often no more than ten. Clearly, drawing statistically relevant conclusions from such a small sample size is difficult, both for quantitive measures such as search time and for subjective opinions collected through questionnaires (including those where subjects are asked to assign scores on an ordinal scale).

Future work should focus on developing a standardised benchmark that could be used within the browsing community in order to fully gauge the quality of a newly developed system. This benchmark could comprise both specific search tasks (such as target search) and annotation tasks (similar to the one adopted in [49]) which can be applied to any image database navigation system. Such a benchmark would of course require an underlying dataset with a ground truth (e.g. manual annotations) which in itself is not straightforward to obtain due to the work involved and other factors such as copyright issues.

5 Conclusions

In this chapter we have investigated, in detail, the current state-of-the-art of image retrieval systems that allow a user to visually navigate through an image collection. We first looked at similarity-based methods providing an intuitive visualisation of an image collection and identified three main approaches. Mapping-based visualisations maintain the relationships between images in the database in the high-dimensional feature space. Projection into the (typically 2-dimensional) visualisation space is achieved through application of dimensionality reduction techniques such as PCA or MDS. This type of visualisation has also been adopted in systems that employ virtual reality concepts to provide a more immersive browsing experience. However, the costs associated with the necessary equipment will prevent wide-spread adoption of this approach. Clustering-based methods employ, as the name implies, a clustering algorithm to organise images in a collection. Clustering the images into smaller groups of similar images allows the user to browse down a hierarchy, whereby the further down the tree they delve, the more similar images become. Graph-based visualisations express relationships between images (such as visual similarity or common keyword annotation) as links of a graph structure that is visualised to the user. Image collections can also be displayed based on time stamp information which can prove useful to identify distinct events and display relevant pictures.

Mapping-based visualisations aim to maintain the relationships between images occurring in the high-dimensional feature space, and display them usually within the 2D constraints of a computer display or a 3D virtual environment. It has been shown in [59] that arranging images according to visual similarity can reduce the time required for image retrieval. These visualisations harness the power of the human cognitive system, passing a vast quantity of data processing subconsciously to the user's mind. However, one of the drawbacks of this type of visualisation is that the limited space often causes images to overlap or to occlude each other. Ways to address this issue and reduce overlap include the fitting of images to a regular grid structure or slight adjustments of the visual arrangement in order to preserve the structure. Another problem with mapping-based visualisations is that they are computationally expensive to generate, and are hence rarely suitable for computing 'on-the-fly' visualisations of a large number of query results. Furthermore, the addition of images to the collection typically requires these visualisations to be recalculated in their entirety.

Clustering-based visualisations have the advantage that by dividing the database into smaller entities, only a small subset of images needs to be visualised. This ultimately leads to less processing for both the system and the user. The system needs only to load a section of the database when the user has accessed a particular cluster of images, rather than loading all images as is the case with global mapping-based visualisations. The cognitive load on the user is also reduced as the number of distinct images to be inspected is much lower. However, a disadvantage of clustered visualisations is that the user can become 'trapped' in a subset of the database. This can occur when representative images used at higher levels of the tree either do not represent the images in that subtree well enough, or are not distinct enough from other representative images at the same level. Both scenarios can lead to the user traversing nodes of the structure in vain, leading to excessive time required for retrieval and added frustration to the user. It should be noted that this is not so much a flaw of the visualisation itself but is rather caused by the underlying similarity measure employed, the best of which are still incapable of modeling human perception appropriately.

Indeed, this problem applies to all forms of visualisations including graph-based approaches. If the features extracted and similarity measures do not model human perception well, the links formed between images may impede rather than support the browsing experience. Creating links to multiple images based on a variety of features, as implemented in NN^k networks [22], allows the user to browse through images in the database based upon a particular feature such as colour or texture. However, the user may need to adjust the number of neighbours displayed as an excessive number of links between images will make the visualisation more complex and less intuitive.

The variety of browsing tools available to the user are usually common to all visualisations. Being able to zoom into areas of interest can be applied to any visualisation, although vertical zooming is available only in hierarchically organised visualisations. These structures often come with some overview of the underlying tree and the user's current location within the system in order to help with the navigation task. This kind of overview can also be applied to mapping-based and graph-based visualisations, particularly when the user has increased the zoom factor so that not all of the visualisation is visible within the screen area. In this case a panning function is required to allow the user to navigate the structure without having to zoom out again. Manual scaling of images and magnified previews of images can further enhance the user's browsing experience.

Relatively little work has been performed into investigating which visualisation paradigm may be the most useful, although a variety of user studies have shown that organising image databases in the ways presented in this chapter can reduce the retrieval time when compared with traditional approaches. What is currently missing is a standard benchmark for assessing the effectiveness and efficiency of browsing systems. Such a benchmark would be based on a standardised image set (or several collections of different magnitudes in order to be able to judge scalability) together with a ground truth and a number of pre-defined tasks. One such task which seems particularly interesting is the annotation task defined in [49], a 'real world' task which can be quantitatively measured in order to compare systems. A large, copyright-free image database however is still an issue, although systems such as those presented in [13, 39] use the online image resource Flickr [17] to obtain images. Defining a ground truth is an even bigger challenge (as can e.g. be seen immediately by inspecting the annotations that are given on Flickr).

In addition to advancements in the evaluation of image database navigation systems, further research and new browsing paradigms are likely to be required to harness the true potential of image browsing. One of the coming challenges for browsing systems is the decreasing screen resolution and reduced processing available, that have come as a consequence of mobile computing. More and more people use their mobile phones to explore the internet, and require access to the millions of images available online. Nowadays, mobile phones also act as a primary source of image capture for many. Photographs are often uploaded to the web to either share on social networking sites, or uploaded to a 'cloud' (or server) whereby the user can access their images from any device. Works such as [21, 70, 84] have looked at developing traditional QBE CBIR systems for mobile devices, whilst [32] does briefly look at browsing on a mobile device. With the increasing graphical and processing ability of handheld devices, coupled with the increasing number of images stored locally and online, browsing large image databases in the palm of the users hand will almost certainly be a future requirement.

References

1. Abdel-Mottaleb, M., Krischnamachari, S., Mankovich, N.J.: Performance Evaluation of Clustering Algorithms for Scalable Image Retrieval. In: IEEE Computer Society Workshop on Empirical Evaluation of Computer Vision Algorithms (1998)
2. Assfalg, J., Del-Bimbo, A., Pala, P.: Virtual Reality for Image Retrieval. Journal of Visual Languages and Computing 11(2), 105–124 (2000)
3. Atkinson, K.E.: An Introduction to Numerical Analysis. John Wiley and Sons, Chichester (1989)
4. Bederson, B.: Quantum Treemaps and Bubblemaps for a Zoomable Image Browser. In: ACM Symposium on User Interface Software and Technology, pp. 71–80 (2001)
5. Borth, D., Schulze, C., Ulges, A., Breuel, T.: Navidgator - Similarity Based Browsing for Image and Video Databases. In: German Conference on Advances in Artificial Intelligence, pp. 22–29 (2008)
6. Chen, C.: Information Visualization. Springer, Heidelberg (2004)

7. Chen, C., Gagaudakis, G., Rosin, P.: Similarity-Based Image Browsing. In: International Conference on Intelligent Information Processing, pp. 206–213 (2000)
8. Chen, Y., Butz, A.: Photosim: Tightly Integrating Image Analysis into a Photo Browsing UI. In: International Symposium on Smart Graphics (2008)
9. Combs, T., Bederson, B.: Does Zooming Improve Image Browsing? In: ACM Conference on Digital Libraries, pp. 130–137 (1999)
10. Cruz-Neira, C., Sandin, D., DeFanti, T.: Surround-screen Projection-based Virtual Reality: The Design and Implementation of the CAVE. In: 20th Annual Conference on Computer Graphics and Interactive Techniques, pp. 135–142 (1993)
11. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys 40(2), 1–60 (2008)
12. Deng, D., Zhang, J., Purvis, M.: Visualisation and Comparison of Image Collections based on Self-organised Maps. In: Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation, pp. 97–102 (2004)
13. Dontcheva, M., Agrawala, M., Cohen, M.: Metadata Visualization for Image Browsing. In: ACM Symposium on User Interface Software and Technology (2005)
14. Eidenberger, H.: A Video Browsing Application Based on Visual MPEG-7 Descriptors and Self-organising Maps. International Journal of Fuzzy Systems 6(3) (2004)
15. Faloutsos, C., Equitz, W., Flickner, M., Niblack, W., Petkovic, D., Barber, R.: Efficient and Effective Querying by Image Content. Journal of Intelligent Information Systems 3, 231–262 (1994)
16. Faloutsos, C., Lin, K.: FastMap: A Fast Algorithm for Indexing, Datamining and Visualization of Traditional and Multimedia Datasets. In: ACM SIGMOD International Conference on Management of Data, pp. 163–174 (1995)
17. Flickr (2009), http://www.flickr.com/
18. Gomi, A., Miyazaki, R., Itoh, T., Li, J.: CAT: A Hierarchical Image Browser Using a Rectangle Packing Technique. In: International Conference on Information Visualization, pp. 82–87 (2008)
19. Graham, A., Garcia-Molina, H., Paepcke, A., Winograd, T.: Time as Essence for Photo Browsing through Personal Digital Libraries. In: ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 326–335 (2002)
20. Gupta, A., Jain, R.: Visual Information Retrieval. Communications of the ACM 40(5), 70–79 (1997)
21. Hare, J.S., Lewis, P.H.: Content-based Image Retrieval Using a Mobile Device as a Novel Interface. In: SPIE Storage and Retrieval Methods and Applications for Multimedia, pp. 64–75 (2005)
22. Heesch, D., Rüger, S.M.: NN^k Networks for Content-Based Image Retrieval. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 253–266. Springer, Heidelberg (2004)
23. Heesch, D., Rüger, S.: Three Interfaces for Content-Based Access to Image Collections. In: International Conference on Image and Video Retrieval, pp. 491–499 (2004)
24. Hilliges, O., Baur, D., Butz, A.: Photohelix: Browsing, Sorting and Sharing Digital Photo Collections. In: IEEE Tabletop Workshop on Horizontal Interactive Human-Computer Systems, pp. 87–94 (2007)

25. Hilliges, O., Kunath, P., Pryakhin, A., Butz, A., Kriegel, H.P.: Browsing and Sorting Digital Pictures using Automatic Image Classification and Quality Analysis. In: International Conference on Human-Computer Interaction, pp. 882–891 (2007)
26. Hinton, G., Roweis, S.: Stochastic Neighbor Embedding. In: Advances in Neural Information Processing Systems, vol. 15, pp. 833–840 (2002)
27. Apple iPhoto (2009), http://www.apple.com/ilife/iphoto/
28. Jacobs, C.E., Finkelstein, A., Salesin, D.H.: Fast Multiresolution Image Querying. In: Conference on Computer Graphics and Interactive Techniques, pp. 277–286 (1995)
29. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1988)
30. Jambu, M.: Exploratory and Multivariate Data Analysis. Academic Press, London (1991)
31. Keller, I., Meiers, T., Ellerbrock, T., Sikora, T.: Image Browsing with PCA-Assisted User-Interaction. In: IEEE Workshop on Content-Based Access of Image and Video Libraries, pp. 102–108 (2001)
32. Khella, A., Bederson, B.: Pocket PhotoMesa: A Zooming Image Browser for PDA's. In: International Conference on Mobile and Ubiquitous Multimedia, pp. 19–24 (2004)
33. Kohonen, T.: Self-organizing Maps. Springer, Heidelberg (1997)
34. Koikkalainen, P., Oja, E.: Self-organizing Hierarchical Feature Maps. In: International Joint Conference on Neural Networks, vol. 2, pp. 279–285 (1990)
35. Krischnamachari, S., Abdel-Mottaleb, M.: Image Browsing using Hierarchical Clustering. In: IEEE Symposium Computers and Communications, pp. 301–307 (1999)
36. Kruskal, J.B., Wish, M.: Multidimensional Scaling. Sage, Thousand Oaks (1978)
37. Laaksonen, J., Koskela, M., Oja, E.: PicSOM – Self-organizing Image Retrieval with MPEG-7 Content Descriptors. IEEE Transactions on Neural Networks: Special Issue on Multimedia Processing 13(4), 841–853 (2002)
38. Lew, M.S., Sebe, N.: Visual Websearching Using Iconic Queries. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 788–789 (2000)
39. Li, J., Wang, J.Z.: Real-Time Computerized Annotation of Pictures. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(6), 985–1002 (2008)
40. Lim, S., Chen, L., Lu, G., Smith, R.: Browsing Texture Image Databases. In: International Conference on Multimedia Modelling, pp. 328–333 (2005)
41. Linde, Y., Buzo, A., Gray, R.: An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications 28, 84–94 (1980)
42. Liu, H., Xie, X., Tang, X., Li, Z.W., Ma, W.Y.: Effective Browsing of Web Image Search Results. In: ACM International Workshop on Multimedia Information Retrieval, pp. 84–90 (2004)
43. Ma, W.Y., Manjunath, B.S.: NeTra: A Toolbox for Navigating Large Image Databases. Multimedia Systems 7(3), 184–198 (1999)
44. Milanese, R., Squire, D., Pun, T.: Correspondence Analysis and Hierarchical Indexing for Content-Based Image Retrieval. In: IEEE International Conference on Image Processing, pp. 859–862 (1996)
45. Moghaddam, B., Tian, Q., Lesh, N., Shen, C., Huang, T.: Visualization and User-Modeling for Browsing Personal Photo Libraries. International Journal of Computer Vision 56(1/2), 109–130 (2004)

46. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals. Pattern Recognition Letters 22(5), 593–601 (2001)
47. Nakazato, M., Huang, T.: 3D MARS: Immersive Virtual Reality for Content-Based Image Retrieval. In: IEEE International Conference on Multimedia and Expo., pp. 44–47 (2001)
48. Nakazato, M., Manola, L., Huang, T.: ImageGrouper: A Group-Oriented User Interface for Content-Based Image Retrieval and Digital Image Arrangement. Journal of Visual Language and Computing 14(4), 363–386 (2003)
49. Nguyen, G.P., Worring, M.: Interactive Access to Large Image Collections using Similarity Based Visualization. Journal of Visual Languages and Computing 19, 203–224 (2008)
50. Osman, T., Thakker, D., Schaefer, G., Lakin, P.: An Integrative Semantic Framework for Image Annotation and Retrieval. In: IEEE/WIC/ACM International Conference on Web Intelligence, pp. 366–373 (2007)
51. Pecenovic, Z., Do, M.N., Vetterli, M., Pu, P.: Integrated Browsing and Searching of Large Image Collections. In: Laurini, R. (ed.) VISUAL 2000. LNCS, vol. 1929, pp. 279–289. Springer, Heidelberg (2000)
52. Pentland, A., Picard, W.R., Sclaroff, S.: Photobook: Content-Based Manipulation of Image Databases. International Journal of Computer Vision 18(3), 233–254 (1996)
53. Google Picasa (2009), http://picasa.google.com/
54. Plaisant, C., Carr, D., Shneiderman, B.: Image Browsers: Taxonomy, Guidelines, and Informal Specifications. IEEE Software 12, 21–32 (1995)
55. Platt, J., Czerwinski, M., Field, B.: PhotoTOC: Automatic Clustering for Browsing Personal Photographs. Technical report, Microsoft Research (2002)
56. Platt, J.C.: AutoAlbum: Clustering Digital Photographs using Probalistic Model Merging. In: IEEE Workshop on Content-Based Access of Image and Video Libraries, pp. 96–100 (2000)
57. Porta, M.: New Visualization Modes for Effective Image Presentation. International Journal of Image and Graphics 9(1), 27–49 (2009)
58. Rodden, K.: Evaluating Similarity-Based Visualisations as Interfaces for Image Browsing. PhD thesis, University of Cambridge Computer Laboratory (2001)
59. Rodden, K., Basalaj, W., Sinclair, D., Wood, K.: Evaluating a Visualisation of Image Similarity as a Tool for Image Browsing. In: IEEE Symposium on Information Visualisation, pp. 36–43 (1999)
60. Rodden, K., Basalaj, W., Sinclair, D., Wood, K.: A Comparison of Measures for Visualising Image Similarity. In: The Challenge of Image Retrieval (2000)
61. Rodden, K., Basalaj, W., Sinclair, D., Wood, K.: Does Organisation by Similarity Assist Image Browsing? In: SIGCHI Conference on Human Factors in Computing Systems, pp. 190–197 (2001)
62. Rodden, K., Wood, K.: How Do People Manage Their Digital Photographs? In: SIGCHI Conference on Human Factors in Computing Systems, pp. 409–416 (2003)
63. Roweis, S., Saul, L.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290(5500), 2323–2326 (2000)
64. Rubner, Y., Guibas, L.J., Tomasi, C.: The Earth Movers Distance, Multidimensional Scaling, and Color-based Image Retrieval. In: APRA Image Understanding Workshop, pp. 661–668 (1997)

65. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, M.: Relevance Feedback: A Power Tool for Interactive Content-based Image Retrieval. IEEE Transaction on Circuits and Systems for Video Technology 8(5), 644–655 (1998)
66. Ruszala, S., Schaefer, G.: Visualisation Models for Image Databases: A Comparison of Six Approaches. In: Irish Machine Vision and Image Processing Conference, pp. 186–191 (2004)
67. Sammon, J.W.: A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers 18(5), 401–409 (1969)
68. Santini, S., Jain, R.: Integrated Browsing and Querying for Image Databases. IEEE Multimedia 7, 26–39 (2000)
69. Sarkar, M., Brown, M.: Graphical Fisheye Views. Communications of the ACM 37(12), 73–83 (1994)
70. Sarvas, R., Herrarte, E., Wilhelm, A., Davis, M.: Metadata Creation System for Mobile Images. In: International Conference on Mobile Systems, Applications, and Services, pp. 36–48 (2004)
71. Schaefer, G., Ruszala, S.: Image database navigation: A globe-al approach. In: Bebis, G., Boyle, R., Koracin, D., Parvin, B. (eds.) ISVC 2005. LNCS, vol. 3804, pp. 279–286. Springer, Heidelberg (2005)
72. Schaefer, G., Ruszala, S.: Image Database Navigation on a Hierarchical Hue Sphere. In: International Symposium on Visual Computing, pp. 814–823 (2006)
73. Schaefer, G., Ruszala, S.: Image Database Navigation on a Hierarchical MDS Grid. In: 28th Pattern Recognition Symposium, pp. 304–313 (2006)
74. Schaefer, G., Stich, M.: UCID – An Uncompressed Colour Image Database. In: Storage and Retrieval Methods and Applications for Multimedia, pp. 472–480 (2004)
75. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
76. Swain, M., Ballard, D.: Color Indexing. International Journal of Computer Vision 7(1), 11–32 (1991)
77. Tao, D., Tang, X., Li, X., Rui, Y.: Direct Kernal Biased Discriminant Analysis: A New Content-Based Image Retrieval Relevance Feedback Algorithm. IEEE Transactions on Multimedia 8(4), 716–727 (2006)
78. Tao, D., Tang, X., Li, X., Wu, X.: Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(7), 1088–1099 (2006)
79. Tenenbaum, J., Silva, V., Langford, J.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290(5500), 2319–2322 (2000)
80. Tian, G.Y., Taylor, D.: Colour Image Retrieval Using Virtual Reality. In: IEEE International Conference on Information Visualization, pp. 221–225 (2000)
81. Urban, J., Jose, J.M.: EGO: A Personalized Multimedia Management and Retrieval Tool. International Journal of Intelligent Systems 21(7), 725–745 (2006)
82. van Liere, R., de Leeuw, W.: Exploration of Large Image Collections Using Virtual Reality Devices. In: Workshop on New Paradigms in Information Visualization and Manipulation, held in conjunction with the 8th ACM International Conference on Information and Knowledge Management, pp. 83–86 (1999)
83. Worring, M., de Rooij, O., van Rijn, T.: Browsing Visual Collections Using Graphs. In: International Workshop on Multimedia Information Retrieval, pp. 307–312 (2007)

84. Yeh, T., Tollmar, K., Darrell, T.: Searching the Web with Mobile Images for Location Recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 76–81 (2004)
85. Zhang, H., Zhong, D.: A Scheme for Visual Feature Based Image Indexing. In: SPIE/IS&T Conference on Storage and Retrieval for Image and Video Databases, pp. 36–46 (1995)
86. Zhou, X., Huang, T.: Relevance Feedback in Image Retrieval: A Comprehensive Review. Multimedia Systems 8(6), 536–544 (2003)

Relevance Feedback Models for Content-Based Image Retrieval

Peter Auer and Alex Po Leung

Department Mathematik und Informationstechnologie,
Montanuniversität at Leoben
Franz-Josef-Straße 18, 8700-Leoben, Austria
{auer,alex.leung}@unileoben.ac.at

Abstract. We investigate models for content-based image retrieval with relevance feedback, in particular focusing on the exploration-exploitation dilemma. We propose quantitative models for the user behavior and investigate implications of these models. Three search algorithms for efficient searches based on the user models are proposed and evaluated. In the first model a user queries a database for the most (or a sufficiently) relevant image. The user gives feedback to the system by selecting the most relevant image from a number of images presented by the system. In the second model we consider a filtering task where relevant images should be extracted from a database and presented to the user. The feedback of the user is a binary classification of each presented image as relevant or irrelevant. While these models are related, they differ significantly in the kind of feedback provided by the user. This requires very different mechanisms to trade off exploration (finding out what the user wants) and exploitation (serving images which the system believes relevant for the user).

1 Introduction

In this section we introduce the exploration-exploitation dilemma in the context of content-based image retrieval by giving two examples of exploration-exploitation dilemmas a search engine might face.

Assume that a user is looking for an image of a tiger, and the first images presented to the user are of a dog, a car, and a tree. The user might select the dog as most relevant to her query. From this feedback the search engine might conclude that the user is searching for a specific dog, and continues by presenting images of dogs. Thus narrowing the search space too much in response to the user's feedback, might hinder an efficient search. But another user — giving the same feedback — might indeed be looking for a poodle, such that narrowing the search space is appropriate and efficient.

Another example is a user interested in dogs and hunting. Given images of a dog, a car, and a tree, he might classify only the dog as relevant. If the search engine continues to present images of dogs, images of hunting will rarely be presented. Again, the search space is narrowed too much. But also in this case the user might indeed

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 59–79.
springerlink.com

be interested only in dogs, and exploring other topics will results in a significant number of irrelevant images presented to the user.

These examples show that a search engine needs to trade off between selecting images which are close to the images a user has selected so far, and selecting images which reveal more about the implicit query of the user.

In Section 2 we review some prior work in content-based image retrieval with relevance feedback. Our first model, the comparative feedback model, is presented in Section 3, where we propose also some algorithms for this model and present experimental results. Our second model, the binary feedback model, is considered in Section 4 and some previous results are reviewed.

2 Relation to Previous Work

Content-based image retrieval with relevance feedback can be divided into two subproblems: 1.) how we can conduct a specific search to find a suitable image in as few iterations as possible, and 2.) how we can learn a good similarity measure among images based on long-term user feedback from a large number of user search sessions or user labels from datasets.

In previous work [29, 26, 10, 3], active learning has been used to select images around the decision boundary for user feedback, for speeding up the search process and to boost the amount of information which can be obtained from user feedback. However, images around the decision boundary are usually difficult to label. A user might find it hard to label images in between two categories. Such difficulties and noise from user feedback is not explicitly modeled or taken into account in most previous work.

While active learning tries to boost the amount of information which can be obtained from user feedback — mostly by asking the user about examples which are hard to distinguish — this approach ignores that (a) the user typically is not interested in borderline cases, and (b) the user himself might find it difficult to distinguish between difficult examples, such that the user feedback might be quite noisy. These issues and the noise from user feedback has not been explicitly modeled or taken into account in most previous work. In contrast, we explicitly model the noisy user feedback and select images for presentation to the user, such that — after obtaining the user feedback — the algorithm can efficiently search for suitable images by eliminating images not matching the user's query.

To solve the second of the two sub-problems, i.e. how we can learn a good similarity measure among images, it is necessary to find a reasonable similarity measure among the images. In this paper, we do not address this problem. But, we note that recently user labels are easily obtainable because of the technological advances of the Internet. Large amounts of data for high-level features can be found from databases with user labels, often called image tags, such as Flickr, Facebook and Pbase. The popularity of these databases enhances the accuracies of image search engines. For example, the Yahoo image search engine is using tags from images on Flickr. Thus we will consider a combination low-level visual features and high-level features obtained from user labels, and we assume that a reasonably good similarity measure among images can be defined using this features. In our experiments we will use a similarity measure based on the 2-norm. A combination of keywords and visual features has also be used in [12] and [30].

Traditionally, content-based image retrieval with user feedback is considered a learning problem using data from user feedback and, with visual features most previous work assumes that no label describing images in datasets is available, [26, 4, 24, 23]. Metric functions measuring similarity based on low-level visual features are obtained by discriminative methods. Long-term learning is used with training datasets from the feedback of different users [11, 9, 16, 14, 19, 18, 28, 22]. However, because of different perceptions about the same object, different users may give different kinds of feedback for the same query target. Short-term learning using feedback from a single user in a single search session can be used to deal with the different perceptions of objects. Weighting the importance of different low-level features is often used for short-term learning (e.g. PicSOM [15]).

The use of user feedback as training data has played an important role in most recent work [27, 25, 5, 17, 7]. Feedback is used as positive or negative labels for training. But as the user chooses the most relevant images in any iteration, such an image may be chosen even if the image is rather dissimilar to any suitable image. Furthermore, images predicted to be positive examples by discriminative methods are traditionally selected for presentation in each round. Thus, mistakes of the discriminative method might hinder progress in the search significantly — by ignoring part of the search space with images which are incorrectly predicted as negative.

3 Comparative Feedback

In this section we consider a model in which the search engine supports the user in finding an image which matches her query sufficiently well. In each iteration of the search, the search engine presents a set of images to the user and the user selects the most relevant image from this set. We assume a given database $\mathscr{D}$ of images x, and in each iteration a fixed number k of images is presented to the user. The formal search protocol is as follows:

- For each iteration $i = 1, 2, \ldots$ of the search:
  - The search engine calculates a set of images $x_{i,1}, \ldots, x_{i,k} \in \mathscr{D}$ and presents the images to the user.
  - If one of the presented images matches the user's query sufficiently well, then the user selects this image and the search terminates.
  - Otherwise the user chooses one of the image x_i^* as most relevant, according to a distribution $D\left\{x_i^* = x_{i,j} | x_{i,1}, \ldots, x_{i,k}; t\right\}$ where t denotes the ideal target image for the user's query.

The crucial element of this model is the distribution D assumed for the user's feedback, and how it can be used for an effective search algorithm.

3.1 A Possible User Model

Instead of expecting that the user deterministically chooses the presented image which is most relevant to her query, we assume that this choice is a random process where more relevant images are just more likely to be chosen. This models some sources of noise in the user's choice, in particular it might be difficult for the user to distinguish between images of similar relevance. We assume a similarity measure $S(x_1,x_2)$ between images x_1,x_2, which also measures the relevance of an image x compared to an ideal target image t by $S(x,t)$. Formally, let $0 \leq \alpha \leq 1$ be the uniform noise in the user's choices and we assume that the probability of choosing image $x_{i,j}$ is given by

$$D\left\{x_i^* = x_{i,j}|x_{i,1},\ldots,x_{i,k};t\right\} = (1-\alpha)\frac{S(x_{i,j},t)}{\sum_{j=1}^{k} S(x_{i,j},t)} + \frac{\alpha}{k}.$$

Assuming a distance function $d(\cdot,\cdot)$ on the images, two possible choices for the similarity measure $S(\cdot,\cdot)$ are

$$S(x,t) = \exp\{-ad(x,t)\} \tag{1}$$

and

$$S(x,t) = d(x,t)^{-a} \tag{2}$$

with a parameter $a > 0$. We note that these two similarity measures predict the user behavior in a subtly but significantly different way: Considering only the case $k = 2$, we find for the polynomial similarity measure (2) that

$$D\{x_i^* = x_{i,1}|x_{i,1},x_{i,2};t\} = D\left\{x_i^* = x_{i',1}|x_{i',1},x_{i',2};t\right\}$$

if

$$\frac{d(x_{i,1},t)}{d(x_{i,2},t)} = \frac{d(x_{i',1},t)}{d(x_{i',2},t)}.$$

In contrast, for the exponential similarity measure (1) we find

$$D\{x_i^* = x_{i,1}|x_{i,1},x_{i,2};t\} = D\left\{x_i^* = x_{i',1}|x_{i',1},x_{i',2};t\right\}$$

only if

$$d(x_{i,1},t) - d(x_{i,2},t) = d(x_{i',1},t) - d(x_{i',2},t).$$

Thus for the polynomial similarity measure the user's response depends on the relative size of the distances to the ideal target image, while for the exponential similarity measure it depends on the absolute difference of the distances. As a consequence, the accuracy of the user's response will remain high for the polynomial similarity measure even when all presented images are close to the ideal target image, while the accuracy will significantly deteriorate for the exponential similarity measure. At the current stage it is not clear which model of user behavior is more adequate.

In all the following we use the squared Euclidean norm $d(x,t) = ||x-t||^2$ as distance measure between image x and the ideal target image t.

3.2 Algorithms

In this model the goal of a search algorithm is to present a sufficiently relevant image in as few as possible search iterations. Such a search algorithm will need to continue exploring, since the images which are chosen by the user as most relevant among the presented images, might still be rather irrelevant to the user's query. If the user chooses an image of a dog in the first iteration, the algorithm should not present only images of dogs in the following iterations. Such a greedy exploitation approach where only images close to the already chosen images are presented to the user, is likely to lead to search failures (as the user might be looking for another kind of animals instead). Presenting the images closest to the already chosen images also limits the amount of information obtained from feedback because the presented images are largely similar. Thus, some exploration strategy has to be adopted.

In the following we describe three exploration strategies which serve as early attempts to solve the search problem and which are evaluated in some initial experiments. All three algorithms maintain weights $w(x)$ on the images x in the database $\mathscr{D}$ and calculate images to be presented to the user according to these weights. The first algorithm selects images at random according to their weights. This algorithm is used in the PicSOM system [15]. The second algorithm performs weighted clustering of the images in the database and selects the cluster centers for presentation to the user. The third algorithm is motivated by noise robust binary search algorithms [13, 21]. Our approximate binary search algorithm presents to the user images which divide the search space into two parts of equal weight such that either response of the user will lead to discounting half of the weights.

3.2.1 Weighting Images

All three algorithms described in this section maintain the weights on the images in the database in the same way. Let $w_i(x)$, $x \in \mathscr{D}$, be the weights of the images which are used to calculate the images $x_{i,1}, \ldots, x_{i,k}$ presented to the user in the i-th iteration of the search. Assuming no a priori information about the relevance of images, the weights are initialized as $w_1(x) = 1$ for all $x \in \mathscr{D}$. If the user model were known, e.g. (1) or (2) with known parameter a, then in the i-th iteration the weights $w_i(x)$ could represent the a posteriori likelihood of the images according to their relevance. But in this initial report we do not want to rely too much on a specific user model. Instead, the only information we take from the user feedback is that some images are more likely to be relevant than others, without quantifying how much more likely that would be. This leads to the following weighting scheme which demotes all apparently less relevant images by a constant discount factor $0 \leq \beta < 1$: Let $x_i^* \in \{x_{i,1}, \ldots, x_{i,k}\}$ be the image chosen by the user as most relevant. If the search has not terminated, then all images $x_{i,1}, \ldots, x_{i,k}$ are not sufficiently relevant and thus their weights are set to 0. All images $x \in \mathscr{D}$ which are closer to some $x_{i,j}$ than to x_i^* are demoted by the discount factor β. Formally, we use the following update of the weights:

- Initialize $w_1(x) = 1$ for all $x \in \mathscr{D}$.
- For each iteration $i = 1, 2, \ldots$ of the search:
  - For all $x \in \mathscr{D}$ set

$$w_{i+1}(x) = \begin{cases} w_i(x) & \text{if } d(x_i^*, x) = \min_j d(x_{i,j}, x) \\ \beta \cdot w_i(x) & \text{otherwise} . \end{cases}$$

  - Set $w_{i+1}(x_{i,j}) = 0$ for all $j = 1, \ldots, k$.

Exponential discounting has been proven to be very useful in various learning scenario. An algorithm which uses the very same discounting scheme as above is the weighted majority algorithm [20]. This is an algorithm for online prediction where in each iteration a binary prediction is to be made by the algorithm. After making its prediction the algorithm receives as feedback whether the prediction was correct or not. The weighted majority algorithm relies on a set of hypotheses H where all hypotheses $h \in H$ make binary predictions which are combined into the algorithm's prediction. For this combination the algorithm maintains weights on the hypotheses which are discounted if the prediction of a hypothesis is incorrect. The assumption is that at least one of the hypotheses gives good predictions. In the search scenario with relevance feedback the possible target images can be seen as the set of hypotheses, and the user feedback can be used to discount images which are likely to be less relevant.

More directly related to the proposed weighting scheme are noisy binary search algorithms [13, 21]. Such binary search algorithms tolerate a certain amount of incorrect information about the target value given to them during the search. In Section 3.2.4 we propose such an approximate binary search algorithm for the search with relevance feedback.

3.2.2 The Random Sampling Algorithm

The random sampling algorithm is the simplest algorithm of the algorithms we describe in this section for calculating the sets of images presented to the user in each search iteration. This algorithm randomly selects (without repetition) images from the dataset according to their weights. The rational of this approach — besides it simplicity and efficiency — is that images with higher weights, which are more likely to be relevant, are included in the set presented to the user with a larger probability. Further, this random selection will spread the selected images well across the database, such that a suitable amount of exploration takes place. This algorithm is implemented in the current version of the PicSOM system [15].

3.2.3 The Weighted Clustering Algorithm

The intuition for this algorithm is that various parts of the search space should be represented by the images presented to the user. To calculate k such representatives we use k-means weighted clustering, where the weight of each image gives its influence on the cluster center: the objective is to find cluster centers $y_1, \ldots, y_k \in \mathscr{D}$ which minimize

$$\sum_{x \in \mathscr{D}} w(x) \min_{1 \leq j \leq k} ||x - y_j||^2.$$

The cluster centers calculated by the clustering algorithm are presented to the user.

3.2.4 The Approximate Binary Search Algorithm

In this section we present a search algorithm which is based on robust binary search [13, 21]. For an easy simulation of the binary search we describe our algorithm only for the case of $k = 2$ images presented to the user in each search iteration. The main idea of the algorithm is to present images $x_{i,1}$ and $x_{i,2}$ to the user such that the sum of weights of the images closer to $x_{i,1}$ is about equal to the sum of weights of the images closer to $x_{i,2}$. Thus, whether $x_{i,1}$ or $x_{i,2}$ is more relevant, half of the weights will be discounted in response to the user's feedback.

An important difference between binary search and search with relevance feedback is that in search with relevance feedback the noise of the user feedback depends on the images presented to the user: even if the pairs $x_{i,1}, x_{i,2}$ and $x'_{i,1}, x'_{i,2}$ give the same partition of the search space, the noise of the user feedback might be quite different, depending on the distance of the presented images (and also depending on the target image). To illustrate this, we consider a 1-dimensional search problem for a target $t \in [-1, +1]$ with either the pair of examples (x_1, x_2), $x_1 = -1/2$, $x_2 = +1/2$, presented to the user, or $x'_1 = -1/4$, $x'_2 = +1/4$, presented to the user. Both pairs split the search space at 0, but Figure 1 shows that the noise in the user model behaves quite differently for the two pairs: for a target distant from 0, the pair $(-1/2, +1/2)$ delivers reliable feedback, but for a target close to 0, the pair

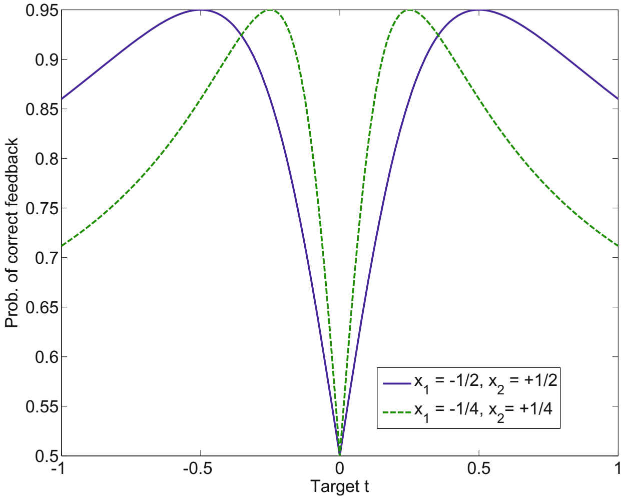


Fig. 1. Probability of correct feedback for different pairs of examples (x_1, x_2) presented to the user, depending on the target t. The feedback is correct if the example closer to the target is chosen by the user.

$(-1/4,+1/4)$ is more reliable[1]. Thus it is important to not only calculate an appropriate partition of the search space but also to present images — inducing this partition — which result in relatively low noise in the user feedback.

Since we are using the squared Euclidean norm $||x-t||^2$ to measure the distance between images x and t, the partition of the search space induced by presented images $x_{i,1}$ and $x_{i,2}$ is given by a hyperplane. For efficient computation we use the following heuristic to find a hyperplane which partitions the search space into parts of approximately equal weight: for a random hyperplane through the centroid of all weighted images, two images are calculated for this hyperplane with about distance $\sigma_i \Delta$ from the hyperplane. Here σ_i^2 is the average weighted distance of all images to the centroid (where distance is measured by the squared Euclidean norm), and Δ is a parameter of the algorithm which we call the gap parameter. Essentially the gap parameter measures the closeness of the presented images and thus influences the amount of noise in the user feedback.

3.3 Experiments

The three search algorithms for content-based image retrieval with user feedback are evaluated with Monte Carlo simulations where randomness is introduced by the user model, the algorithms, and the data themselves, in particular through the randomly selected ideal target of a search. In all experiments the ideal target was an image from the database, and the search terminated as soon as the target image was selected for presentation to the user. In a more realistic scenario it can be expected that searches will terminate earlier since the user will be satisfied with a sufficiently similar image.

In all experiments we use the exponential user model (1). We investigate the influence of the relevant parameters on the number of search iterations. These parameters are the uniform noise rate α, the parameter a of the user model, the discount factor β of the weighting scheme, and for the approximate binary search algorithm also the gap parameter Δ. To reduce statistical fluctuations, each reported plot is averaged over ten repeated experiments with the same set of parameters.

In the first set of experiments we use synthesized data for which the distribution is known such that the experiments are easier to analyze. Very surprisingly, we find in these experiments that the simple random sampling algorithm performs best for a wide range of reasonable parameter settings. We discuss this result in Section 3.4 below. Before we compare the three algorithms in Figures 9–11, we investigate the behavior of the algorithms separately in Figures 2–8.

In a second set of experiments we have simulated actual searches on the VOC2007 dataset [8], and we report qualitative results.

3.3.1 Experiments on Synthetic Data

For this data an image is generated as a 23-dimensional vector with each element uniformly distributed between 0 and 1. The synthetic database contains 10,000 such

[1] Here we used model (2) with the absolute distance and parameters $a = 2$ and $\alpha = 0.1$.

images. The dimensionality and number of data were chosen to match the VOC2007 dataset [8] which contains about 10,000 images from 23 categories. Using the categories as high level features gives image descriptions of the same dimension.

For an easier analysis we set $k = 2$ in these experiments, such that only 2 images are presented to the user in each search iteration. The number of search iterations is expected to be significantly reduced for larger k. All reported results are averaged over 10 searches for randomly selected target images from the dataset.

We first investigate the influence of the user model parameter a and the algorithms' parameters on the number of search iterations. For this, we keep the uniform noise at $\alpha = 0.1$ and vary the user model parameter a and the discount factor β. For the approximate binary search algorithm we report also results for fixed $\beta = 0.5$ and varying a and varying gap parameter Δ.

For the user model parameter a we consider the range $2 \leq a \leq 16$ and $0.1 \leq \alpha \leq 0.3$. This gives an overall noise rate of about 5% to 16% in early search iterations and 17% to 45% close to the end of the search.

Figures 2 and 3 show the performance of the random sampling algorithm and the weighted clustering algorithm for varying a and β. Figure 4 shows the performance of the approximate binary search algorithm for fixed gap parameter Δ and varying a and β, Figure 5 shows the performance for fixed β and varying a and Δ.

In Figures 6, 7, and 8 we investigate the influence of the discount factor β for various uniform noise rates α and fixed user model parameter $a = 8$. For the approximate binary search algorithm we set the gap parameter $\Delta = 2$, which is a reasonable choice given Figure 5.

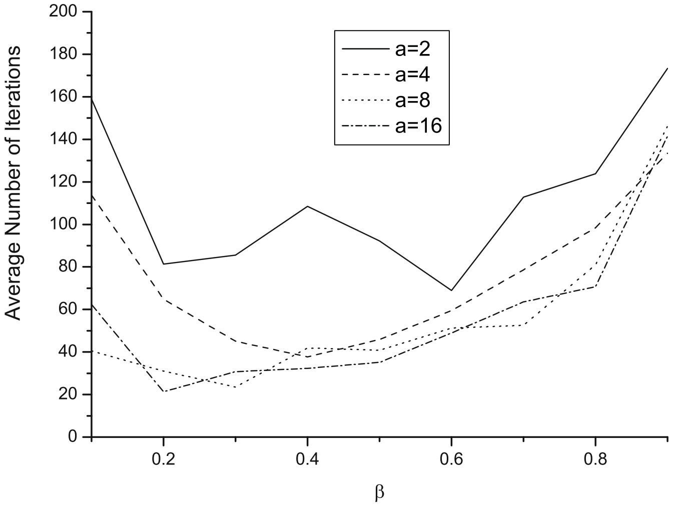


Fig. 2. Average number of search iterations on synthetic data for the **random sampling** algorithm with $\alpha = 0.1$ and varying a and β

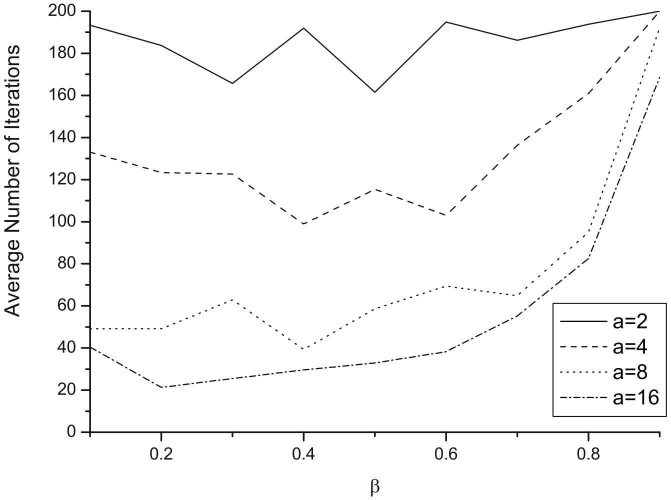


Fig. 3. Average number of search iterations on synthetic data for the **weighted clustering** algorithm with $\alpha = 0.1$ and varying a and β

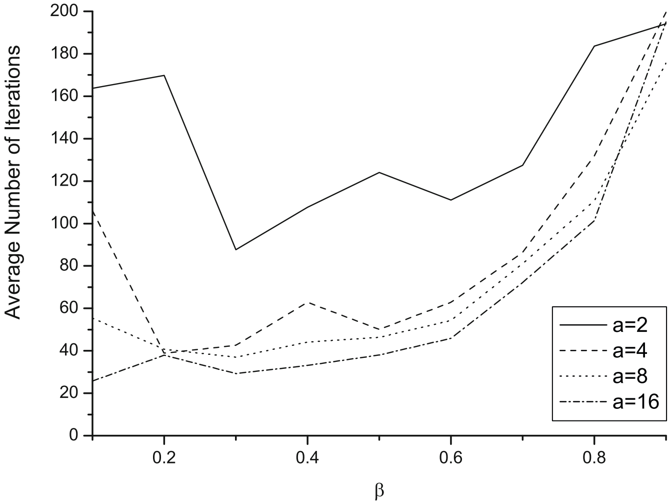


Fig. 4. Average number of search iterations on synthetic data for the approximate **binary search** algorithm with $\alpha = 0.1$, $\Delta = 2$, and varying a and β

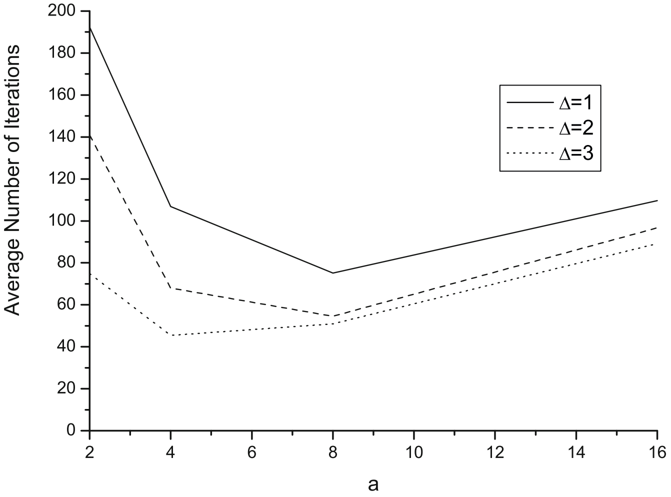


Fig. 5. Average number of search iterations on synthetic data for the approximate **binary search** algorithm with $\alpha = 0.1$, $\beta = 0.5$, and varying a and Δ

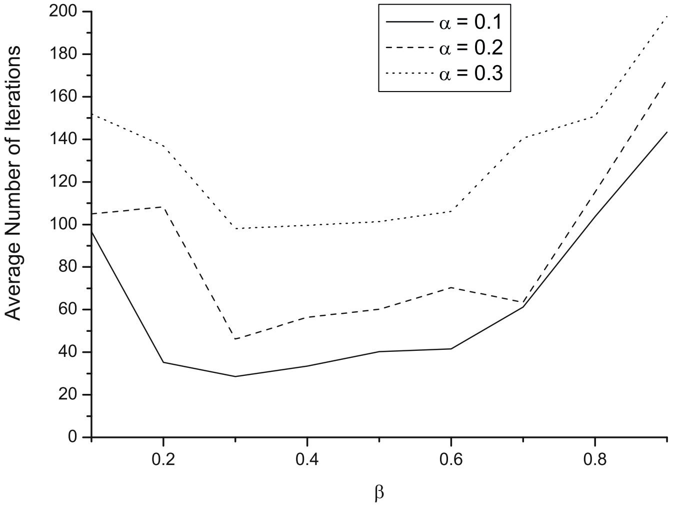


Fig. 6. Average number of search iterations on synthetic data for the **random sampling** algorithm with $a = 8$ and varying α and β

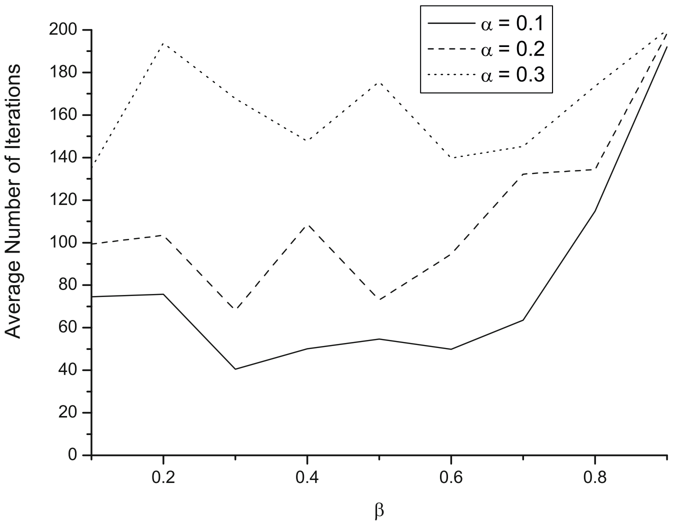


Fig. 7. Average number of search iterations on synthetic data for the **weighted clustering** algorithm with $a = 8$ and varying α and β

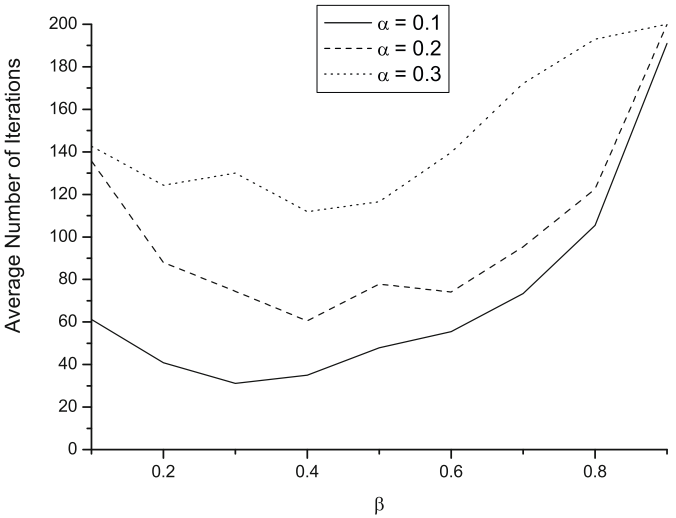


Fig. 8. Average number of search iterations on synthetic data for the approximate **binary search** algorithm with $a = 8$, $\Delta = 2$, and varying α and β

We find that with increasing uniform noise and increasing noise from the user model (i.e. decreasing user model parameter a) the number of search iterations increases as expected. More interestingly, we find that the performance of the algorithms is relatively insensitive in respect to the choice of the discount factor β. For a reasonable range around $\beta = 0.5$ the number of iterations is quite stable. Nevertheless, the number of search iterations can be reduced by an optimal choice of the discount factor. Finally, it seems that a large gap parameter Δ for the approximate binary search algorithm seems advantageous, see also the discussion in Section 3.4.

Finally, we compare the three algorithms for some parameter settings. In Figure 9 we vary the user model parameter a and fix the uniform noise rate $\alpha = 0.1$ and the discount rate $\beta = 0.5$, in Figure 10 we vary the uniform noise rate α and fix $a = 8$ and $\beta = 0.5$, and in Figure 11 we vary the discount factor β and fix $\alpha = 0.1$ and $a = 8$. For the approximate binary search algorithm we set the gap parameter $\Delta = 2$. We find that the simple random sampling algorithm performs best for a wide range of reasonable parameter settings. We discuss this result in Section 3.4 below.

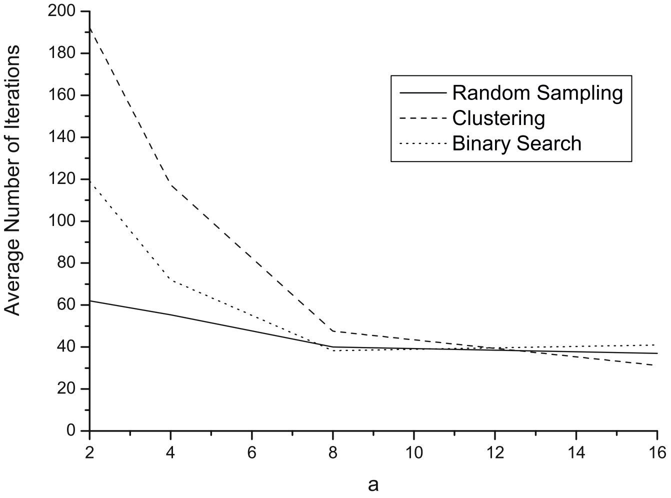


Fig. 9. Average number of search iterations on synthetic data for the three algorithms with $\alpha = 0.1$, $\beta = 0.5$, $\Delta = 2$ and varying a

3.3.2 Results on the VOC2007 Dataset

For the experiments on realistic data we use the VOC2007 dataset with 23 categories and 9963 images. This dataset has been built for the PASCAL Visual Object Classes Challenge 2007 [8]. The goal of the challenge was to recognize objects from several classes in realistic scenes. The 23 object (sub-)classes are Person (person, foot, hand, head), Animal (bird, cat, cow, dog, horse, sheep), Vehicle (aeroplane, bicycle, boat, bus, car, motorbike, train), and Indoor (bottle, chair, dining table, potted plant, sofa, tv/monitor). Each of the 9963 images in the dataset is annotated by a bounding box and class label for each object from the 23 classes which is present in the image. Multiple objects from multiple classes may be present in an image.

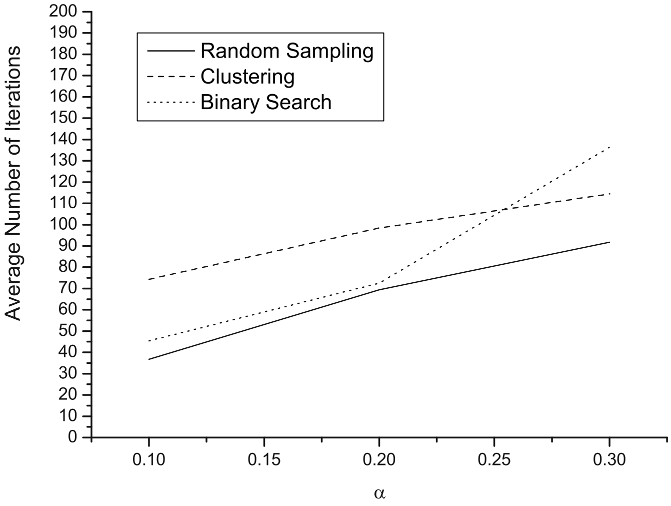


Fig. 10. Average number of search iterations on synthetic data for the three algorithms with $a = 8$, $\beta = 0.5$, $\Delta = 2$, and varying uniform noise rate α

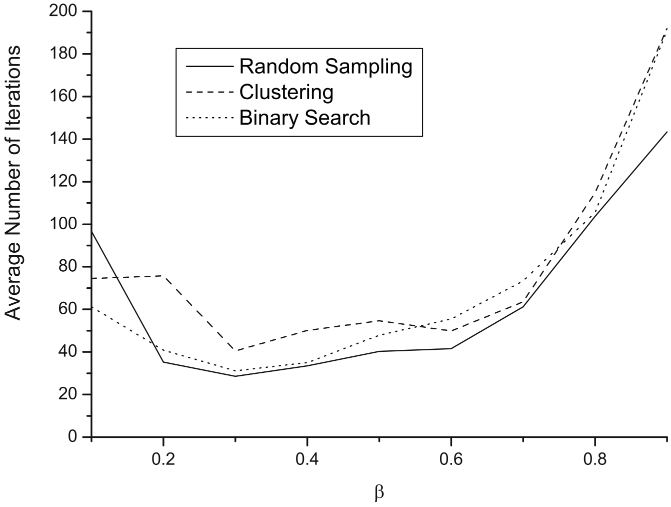


Fig. 11. Average number of search iterations on synthetic data for the three algorithms with $\alpha = 0.1$, $a = 8$, $\Delta = 2$, and varying discount factor β

In our experiments we use 23 high-level features, one feature for each object class, to describe the images. For an image the feature value for an object class is the size (as calculated from the bounding box) of the largest object from this class in the image. If no object from the class is present in the image, then the feature value is 0.

We first replicate an experiment of the previous section: We use the weighted clustering algorithm to search for a target image in the dataset. The results (Figure 12) are quite comparable with the experiments on the synthetic data (Figure 7). Since the results in Figure 12 are averages of only 3 random searches, the fluctuation of the results for the VOC2007 dataset is higher.

In the last set of experiments we perform two realistic searches on the VOC2007 dataset, with a human selecting the most relevant image in each search iteration. In each search iteration 20 images are presented to the user, which are calculated by the weighted clustering algorithm. In addition to the high-level features described above we use also the low-level visual features (color, texture, and edge orientations) calculated by the PiCSOM system [15]. This results in a 761-dimensional feature vector with 23 high-level features and 738 low level features.

The first search was for a car on grass. Figures 13 and 14 show the images presented in the first and second iteration of the search and the images chosen by the user as most relevant in these iterations. Figure 15 shows images chosen by the user as most relevant in subsequent iterations. The second search was for a motorbike on grass, and the images chosen by the user as most relevant are shown in Figure 16. For both searches a good image was found within 10 search iterations.

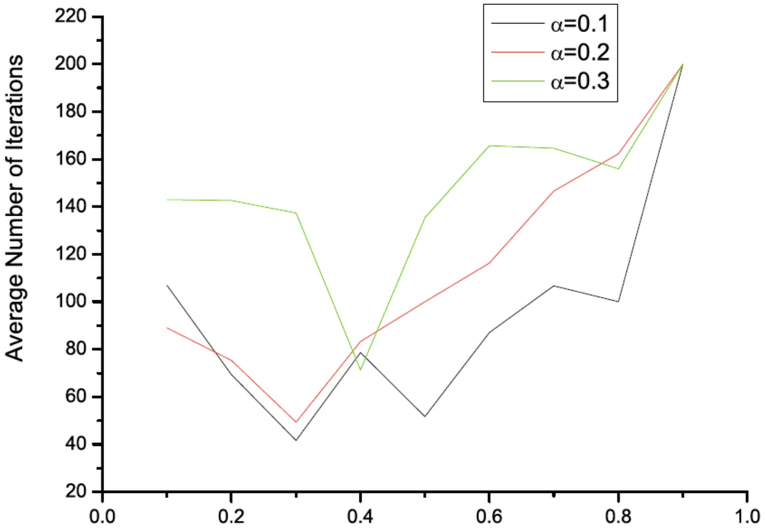


Fig. 12. Average number of search iterations on the VOC2007 dataset for the **weighted clustering** algorithm with varying α and β

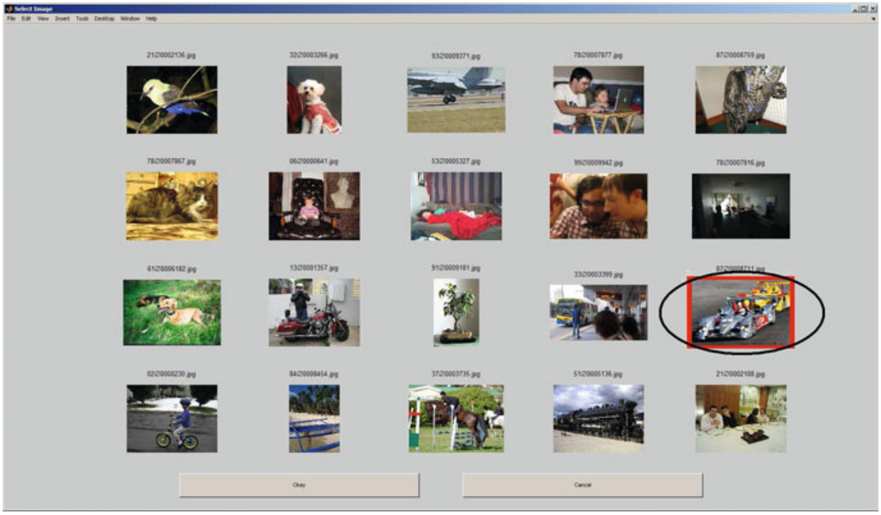


Fig. 13. Search for a car on grass in the VOC2007 dataset by a real user: Iteration 1

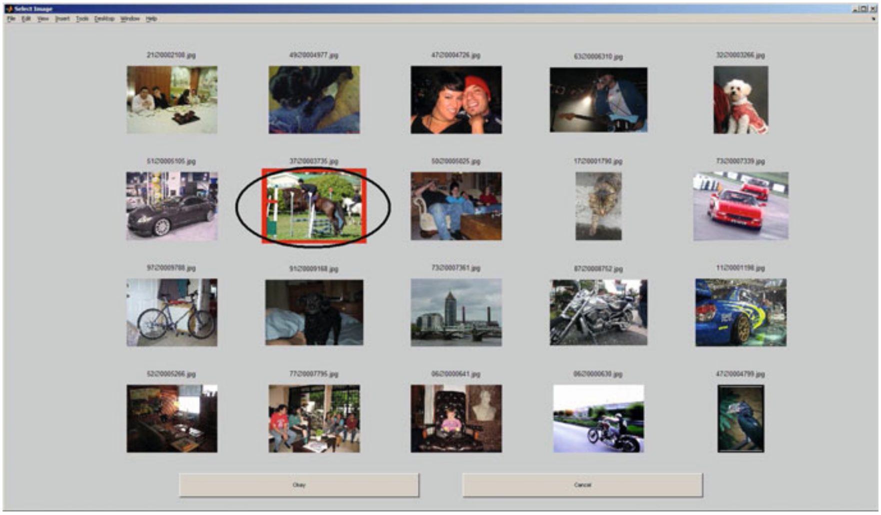


Fig. 14. Search for a car on grass in the VOC2007 dataset by a real user: Iteration 2

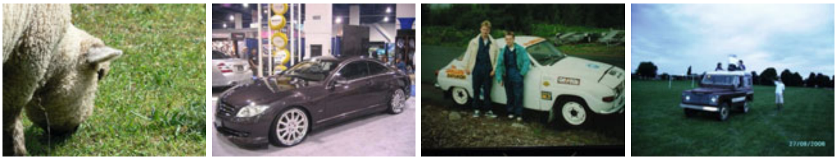


Fig. 15. Search for a car on grass in the VOC2007 dataset by a real user: Most relevant images in iterations 3, 4, 5, and 8

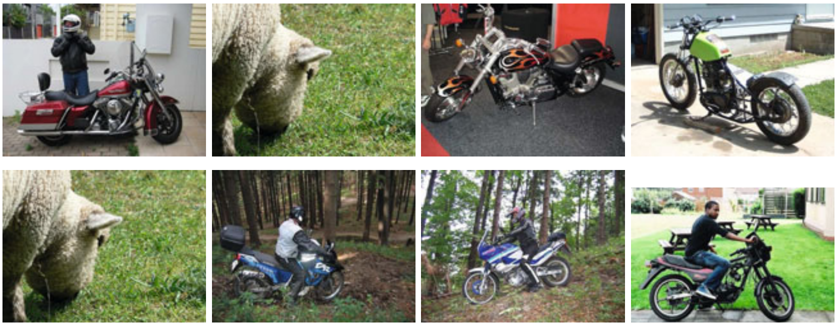


Fig. 16. Search for a motorbike on grass in the VOC2007 dataset by a real user: Most relevant images in iterations 1, 2, 3, 4, 5, 6, 9, and 10

3.4 Discussion

The surprising result of our preliminary experiments is that the simple random sampling algorithm performs significantly better than the algorithms specifically designed for the search task. We are currently investigating possible reasons for this result and we offer a first hypothesis about this below.

As far as the approximate binary search algorithm is concerned, it seems best to present images to the user which are as far as possible from the separating hyperplane, cf. Figure 5. This is plausible given the exponential user model (1) which predicts high noise if the presented images are close. To some extend this observation might explain also the rather poor behavior of the weighted clustering algorithm: the clustering algorithm selects the centroids for presentation to the user while extreme points at the (opposite) boundaries of the clusters might give better performance. By the construction of our 23-dimensional synthetic data, the squared length of most of the random feature vectors is close to the average squared length $\frac{23}{3}$. Thus most of the points are rather extreme points and the sampling algorithm is quite likely to choose such points.

The experiments on the synthetic data show that even with only two images presented to the user a relatively fast search is possible. Presenting 10–20 images should reduce the number of search iterations considerably. This will be verified in further experiments. Initial experiments on the realistic VOC2007 dataset with high-level features already confirm that around 10 search iterations are sufficient for finding a suitable image. Naturally, the search performance depends on appropriate features, and this relation needs also to be investigated further.

4 Binary Feedback

In our second feedback model we are considering a filtering task, where relevant images shall be presented to the user. The user gives a binary classification to each presented image as either relevant or irrelevant. The goal of the search engine in this model is to present mostly relevant images to the user, and only a small number of irrelevant images.

We distinguish two scenarios for this binary feedback model. In the first scenario, in each iteration a set of k images becomes available and the search engine has to decide, which single image out of the k available images should be presented to the user. In the second scenario, the search engine needs to select relevant images $x \in \mathscr{D}$ from a database $\mathscr{D}$. We will argue that the difference between these scenarios is rather minor.

We assume that an image x is represented by a normalized vector of non-negative features, $x \in \mathbf{R}_+^d$, $||x|| = 1$. Furthermore, we assume that the probability of an image x being relevant is given by the inner product $x \cdot t$ with an ideal target image $t \in \mathbf{R}_+^d$, $||t|| = 1$. By using appropriate features — possibly given implicitly by a kernel function — these are reasonable assumptions.

4.1 Selecting a Relevant Image from k Images

The formal search protocol considered in this section is the following:

- The user has an ideal target image t in mind.
- In each iteration $i = 1, 2, \ldots$:
  - There are k images $x_{i,1}, \ldots, x_{i,k}$ given to the search engine.
  - The search engine selects an image $x_i^* \in \{x_{i,1}, \ldots, x_{i,k}\}$ and presents it to the user.
  - The user's feedback is $y_i = +1$ with probability $x_i^* \cdot t$ (the image is relevant to the user), or $y_i = 0$ otherwise.

The goal of the search engine is to maximize the number of relevant images, $\sum_i y_i$. The exploitation-exploration trade-off in this model is more pronounced than in the model discussed in Section 3: Based on the presented images and the received user feedback in previous iterations $< i$, the search engine can calculate an estimate $\hat{t}_i$ for the unknown ideal target image. From a new set of images $x_{i,1}, \ldots, x_{i,k}$, the search engine might select the image which maximizes the estimated probability $x_{i,j} \cdot \hat{t}_i$ of being relevant. But since the estimate $\hat{t}_i$ might be inaccurate, this exploitative choice might be suboptimal. Thus, alternatively, the search engine might exploratively select an image for which the user feedback improves the accuracy of the estimate $\hat{t}_i$ the most.

This model has been analyzed by Auer in [1, Section 4], and an appropriate algorithm based on upper confidence bounds has been proposed. This algorithm implicitly trades off exploration and exploitation. It calculates upper confidence bounds $\hat{p}_{i,j}$ on the probability of image $x_{i,j}$ being relevant, and selects the image with the largest upper confidence bound. Hence, an image is selected (a) if its true probability of being relevant is indeed large, or (b) if the estimates for this probability are rather unreliable and the resulting confidence interval is large. Case (a) gives an exploitative choice, while case (b) improves the estimates of the probabilities and thus is explorative. In [1] it is shown that the proposed algorithm performs almost as well as if the ideal target image t would have been known in advance: in the n iterations the number of presented relevant images is only $O\left(\sqrt{dn\log(kn)}\right)$ less than if t were known in advance.

4.2 Selecting a Relevant Image from a Database

Here we assume a given image database $\mathscr{D}$. The formal search protocol considered in this section is the following:

- The user has an ideal target image t in mind.
- In each iteration $i = 1, 2, \ldots$:
  - The search engine selects an image $x_i^* \in \mathscr{D}$ and presents it to the user.
  - The user's feedback is $y_i = +1$ with probability $x_i^* \cdot t$ (the image is relevant to the user), or $y_i = 0$ otherwise.

Again the goal of the search engine is to maximize the number of relevant images, $\sum_i y_i$. We argue that the algorithm of [1] from the previous section can be adapted to work also for the protocol with a given database. The obvious reduction is to set $k = |\mathscr{D}|$ and give all images from the database to the algorithm. This poses some computational problems and an efficient implementation is needed, but the search performance will degrade at most logarithmically with the size of the database. A rigorous analysis of a variant of this approach has recently be given in in [6].

4.3 Discussion

In this section we have presented a theoretical approach to the filtering problem with binary feedback. The next step will be an empirical evaluation of this approach on realistic data. Since the performance of approaches like the algorithm in [1] depends rather strongly on the number of features, such approaches are indeed much more suitable for filtering a large set of data than for individual search queries considered in Section 3. For individual search queries the amount of information gained by binary feedback seems to be too small for finding good images in few iterations.

5 Conclusion

Two models for the user behavior of content-based image retrieval with relevance feedback are proposed in the this work and the implications of these models are studied. The models can be applied not only to CBIR but also to other information retrieval tasks in general. They require very different mechanisms to trade off exploration and exploitation. Our experimental results show that the performances of our proposed weighted clustering, random sampling, approximate binary search algorithms for the models are promising.

Acknowledgments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 216529.

References

1. Auer, P.: Using Confidence Bounds for Exploitation-Exploration Trade-offs. Journal of Machine Learning Research 3, 397–422 (2002)
2. Leung, A.P., Auer, P.: An Efficient Search Algorithm for Content-Based Image Retrieval with User Feedback. In: 1st Int. Workshop on Video Mining (VM 2008) in association with IEEE International Conference on Data Mining, ICDM 2008 (2008)
3. Chang, E., Tong, S., Goh, K., Chang, C.: Support Vector Machine Concept-Dependent Active Learning for Image Retrieval. IEEE Transactions on Multimedia (2005)
4. Chen, Y., Zhou, X.S., Huang, T.S.: One-class SVM for learning in image retrieval. In: Proc. ICIP (1), pp. 34–37 (2001)
5. Crucianu, M., Ferecatu, M., Boujemaa, N.: Relevance feedback for image retrieval: a short survey. State of the Art in Audiovisual Content-Based Retrieval. Information Universal Access and Interaction, Including Datamodels and Languages, report of the DELOS2 European Network of Excellence, FP6, 20 (2004)
6. Dani, V., Hayes, T.P., Kakade, S.M.: Stochastic Linear Optimization under Bandit Feedback. In: Proc. 21st Ann. Conf. on Learning Theory, pp. 355–366 (2008)
7. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. (2008)
8. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 Results (2007), http://www.pascal-network.org/challenges/VOC/voc2007/workshop
9. Fournier, J., Cord, M.: Long-term similarity learning in content-based image retrieval. In: Proc. ICIP (1), pp. 441–444 (2002)
10. Gosselin, P.-H., Cord, M., Philipp-Foliguet, S.: Active learning methods for Interactive Image Retrieval. IEEE Transactions on Image Processing (2008)
11. He, X., King, O., Ma, W., Li, M., Zhang, H.: Learning a semantic space from user's relevance feedback for image retrieval. IEEE Trans. Circuits Syst. Video Techn., 39–48 (2003)
12. Jing, F., Li, M., Zhang, H., Zhang, B.: A unified framework for image retrieval using keyword and visual features. IEEE Transactions on Image Processing, 979–989 (2005)
13. Karp, R.M., Kleinberg, R.: Noisy binary search and its applications. In: SODA 2007: Proc. 18th Symp. on Discrete Algorithms, pp. 881–890 (2007)
14. Koskela, M., Laaksonen, J.: Using Long-Term Learning to Improve Efficiency of Content-Based Image Retrieval. In: Proc. PRIS, pp. 72–79 (2003)
15. Koskela, M., Laaksonen, J., Oja, E.: Inter-Query Relevance Learning in PicSOM for Content-Based Image Retrieval. In: Kaynak, O., Alpaydın, E., Oja, E., Xu, L. (eds.) ICANN 2003 and ICONIP 2003. LNCS, vol. 2714, Springer, Heidelberg (2003)

16. Tao, D., Tang, X., Li, X., Wu, X.: Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(7), 1088–1099 (2006)
17. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. TOMCCAP, 1–19 (2006)
18. Linenthal, J., Qi, X.: An Effective Noise-Resilient Long-Term Semantic Learning Approach to Content-Based Image Retrieval. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Las Vegas, Nevada, USA, March 30-April 4 (2008)
19. Tao, D., Li, X., Maybank, S.J.: Negative Samples Analysis in Relevance Feedback. IEEE Trans. Knowl. Data Eng. 19(4), 568–580 (2007)
20. Littlestone, N., Warmuth, M.K.: The weighted majority algorithm. Information and Computation, 212–216 (1994)
21. Pelc, A.: Searching games with errors–fifty years of coping with liars. Theoretical Computer Science, 71-109 (2002)
22. Tao, D., Tang, X.: Nonparametric Discriminant Analysis in Relevance Feedback for Content-based Image Retrieval. In: IEEE International Conference on Pattern Recognition (ICPR), pp. 1013–1016 (2004)
23. Rocchio, J.: Relevance Feedback in Information Retrieval. In: Salton: The SMART Retrieval System: Experiments in Automatic Document Processing, ch. 14, pp. 313–323. Prentice-Hall, Englewood Cliffs (1971)
24. Rui, Y., Huang, T.S.: Optimizing Learning in Image Retrieval. In: Proc. CVPR, pp. 1236–1236 (2000)
25. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell., 1349–1380 (2000)
26. Tong, S., Chang, E.Y.: Support vector machine active learning for image retrieval. In: Proc. ACM Multimedia, pp. 107–118 (2001)
27. Veltkamp, R.C., Tanase, M.: Content-based Image Retrieval Systems: a Survey. State-of-the-Art in Content-Based Image and Video Retrieval, 97–124 (1999)
28. Wacht, M., Shan, J., Qi, X.: A Short-Term and Long-Term Learning Approach for Content-Based Image Retrieval. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), Toulouse, France, May 14-19, pp. 389–392 (2006)
29. Zhang, C., Chen, T.: An active learning framework for content-based information retrieval. IEEE Transactions on Multimedia, 260–268 (2002)
30. Zhou, X.S., Huang, T.S.: Unifying Keywords and Visual Contents in Image Retrieval. In: IEEE MultiMedia, pp. 23–33 (2002)

Biological Inspired Methods for Media Classification and Retrieval

Krishna Chandramouli, Tomas Piatrik, and Ebroul Izquierdo

School of Electronic Engineering and Computer Science
Queen Mary, University of London

Summary. Automatic image clustering and classification is a critical and vibrant research topic in the computer vision community over the last couple of decades. However, the performance of the automatic image clustering and classification tools have been hindered by the commonly referred problem of "Semantic Gap", which is defined as the gap between low-level features that can be extracted from the media and the high-level semantic concepts humans are able to perceive from media content. Addressing this problem, recent developments in biologically inspired techniques for media retrieval is presented in this chapter.

The problem of Image clustering and classification has been the subject of active research across the world during the last decade. This is mainly due to the exponential growth of digital pictures and the need for fully automatic annotation and retrieval systems is ever increasing. The goal of image clustering is to group images such that the intra cluster similarity is increased while the inter cluster similarity is decreased. Thus, the aim is to generate classes providing a concise summarization and visualization of the image content. Clustering is the first step for image classification and subsequent labeling of semantic concepts. The optimization of the classes generated is currently studied in three main research avenues [Fog94] : genetic algorithms (GA), evolution strategies and evolutionary programming. GA stresses on chromosomal operators, while evolution strategies emphasize behavioral changes at the level of the individual. On the other hand evolutionary programming stresses behavioral change at the level of the species for natural evolution. However, the optimization solutions generated by classical evolutionary computation algorithms are far-away from the optimal solutions expected. Therefore, research in imitating human cognition or more precisely biological organisms have been increasingly studied for optimizing image clustering and classification problem.

Recent developments in applied and heuristic optimisation have been strongly influenced and inspired by natural and biological systems. Biologically inspired optimisation techniques are partially based on observations of the sociobiologist E.O.Wilson. In particular to his statement [Wil75]:

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 81–109.
springerlink.com

"In theory at least, individual members of the school can profit from discoveries and previous experience of all other members of the school during the search for food. The advantage can become decisive, outweighing the disadvantages of competition for food, whenever, the resource is unpredictably distributed in patches."

Some of the algorithms that are inspired based on such observations having ties to artificial life: $A-life$ are Ant Colony Optimisation (ACO) introduced by Dorigo et al. in [DG96], Particle Swarm Optimisation (PSO) introduced by Kennedy and Eberhart in 1995 [EK95] and Artificial Immune system based optimisation introduced by Dasgupta in [Das99]. The rest of the chapter is organized as follows. In Section 1, the study of Ant Colony Optimization for image clustering is presented followed by the study of Particle Swarm Optimization for image classification in Section 2. Before concluding the chapter in Section 3, a brief discussion on application of clustering and classification algorithms for media retrieval is presented.

1 Ant Colony Optimisation

Some recent studies have pointed out that, the self-organisation of neurons into brain-like structures, and the self-organisation of ants into a swarm are similar in many respects. Ants present a very good natural metaphor for evolutionary computation. With their small size and small number of neurons, they are not capable of dealing with complex tasks individually. The ant colony, on the other hand, can be seen as an "intelligent entity" for its great level of self-organisation and the complexity of the tasks it performs. Their colony system inspired many researchers in the field of Computer Science to develop new solutions for optimisation and artificial intelligence problems.

The ACO *metaheuristic* is a particular metaheuristic inspired by the behaviour of real ants [DG97]. A key feature of ACO is derived form the ability of real ant colonies to find the shortest or optimal paths between food sources and their nest.

1.1 Behaviour of Real Ants

A main means of communication between ants is the use of chemical agents and receptors. The most important of such chemical agents is the pheromone. Pheromones are molecules secreted by glands on the ant's body. Once deposited on the ground, they evaporate at a known rate. Like neurons, ants use pheromone to communicate. The release of a molecule of pheromone by a single ant influences the behaviour of the other ants in the colony.

When one ant traces a pheromone trail to a food source, that trail will be used by other ants reinforcing the pheromone trail each time. Such autocatalytic process will continue until a trail from the ant colony to the food source is established. (see Fig. 1).

Fig. 1. Ants moving in the pheromone trails

In laboratories, several studies have explored how pheromones are used by ants. Deneubourg et al. [DAGP90] used a double bridge connecting a nest of ants and a food source to study pheromone trail laying and following behaviour in controlled experimental conditions. They ran a number of experiments in which they varied the ratio between the length of the two branches of the bridge (see Fig. 2). In this experiment, at the start the ants were left free to move between the nest and the food source and the percentage of ants that chose one or the other of the two branches was observed over time. The outcome was that, although in the initial phase random oscillations could occur, in most experiments all the ants ended up using the shorter branch. In fact, ants do not pursue creation of a trail with the shorter distance from nest to food source. Their goal is rather to bring food to the nest. However, the pheromone trails they create are highly optimised. This collective trail-laying and trail-following behaviour is the inspiring metaphor for ACO.

1.2 Ant System Algorithm

The Ant System approach (AS) was the first attempt to use the natural metaphor of ants to solve a hard combinatorial problem as the traveling salesman problem. The importance of the original AS [DC99] resides mainly in being the prototype of a number of ant algorithms which collectively implement the ACO paradigm. An ant is a simple computational agent, which iteratively constructs a solution for the instance to solve. Partial problem solutions are seen as states. At the core of the AS algorithm lies a loop, where at each iteration t, each ant moves from a state i to another one j, corresponding to the more complete partial solution. For ant k, the probability $p^k_{i,j}(t)$ of moving from state i to state j depends on the combination of two values:

- the attractiveness $\eta_{i,j}$ of the move, as computed by some heuristic indicating the desirability of the move;
- the pheromone level $\tau_{i,j}$ indicating how proficient it has been in the past to make that particular move;

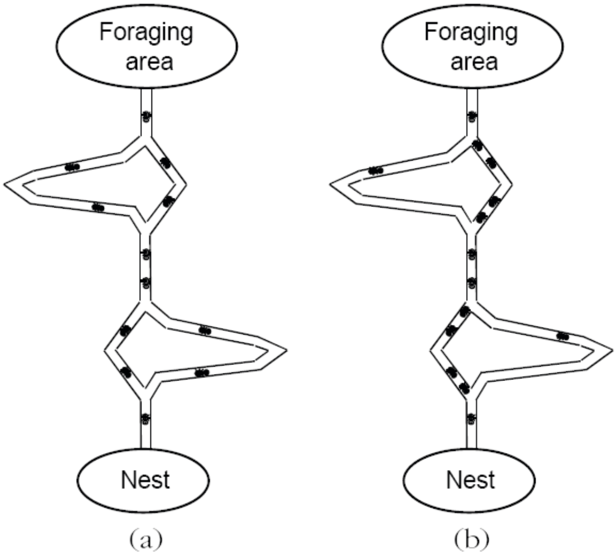


Fig. 2. Double bridge experiment. (a) Ants start exploring the double bridge. (b) Eventually most of the ants choose the shortest path.

The probability of moving from state i to state j is given by following formula, where α and β are heuristicaly estimated parameters ($0<\alpha, \beta<1$):

$$p_{i,j}^{k}(t) = \frac{(\tau_{i,j}(t))^{\alpha}.(\eta_{i,j})^{\beta}}{\sum_{k \in allowed}(\tau_{i,k})^{\alpha}.(\eta_{i,k})^{\beta}} \tag{1}$$

This formula means that the probability of moving ant k depends on both the amount of pheromone τ on that edge and the distance η from i to j. The parameters α and β control the relevance of pheromone and distance in the probabilistic decision. Note that not all edges leaving from i to the next neighbour are allowed.

After each iteration of the algorithm, i.e., when all ants have completed a solution, the pheromone level $\tau_{i,j}(t)$ in the trails is updated by the formula:

$$\tau_{i,j}(t) = \rho \cdot \tau_{i,j}(t-1) + \Delta\tau_{i,j}, \tag{2}$$

where $0<\rho<1$, is another parameter called evaporation coefficient and $\tau_{i,j}$ $(t-\ 1)$ is the previous pheromone concentration. $\Delta\tau_{i,j}$ represents the sum of the pheromone contributions of all ants that used move (i,j) to construct their solution:

$$\Delta\tau_{i,j} = \sum_{k=1}^{m} \Delta\tau_{i,j}^{k}. \tag{3}$$

In (3), m is the number of ants and $\Delta\tau_{i,j}^{k}$ is the amount of pheromone laid on edge (i, j) by ant k. $\Delta\tau_{i,j}^{k}$ can be computed as

$$\Delta\tau_{(i,j)}^{k} = \begin{cases} \frac{Q}{L_k} & , if\ ant\ k\ uses\ edge(i,j) \\ 0 & , otherwise \end{cases}, \tag{4}$$

where Q is a parameter that specifies the amount of pheromones ant k has to distribute through its trail, and L_k is the tour length of ant k. The AS algorithm simply iterates a main loop where m ants construct in parallel their solutions, thereafter updating the pheromone levels according to the quality of their solutions. The original AS algorithm has been further improved with additional strategies. This leads to several other techniques including Ant-Q [DG96], Ant Colony System (ACS) [DC99] and MAX-MIN Ant System [SH96]. Each one of these techniques has been adapted and used in specific domains as economics, data mining and networking.

1.3 Clustering of Images Using Ant Colony Optimisation

Clustering of images according to meaningful classes requires analysis of low-level image attributes including particular combinations of colour, texture or shape descriptors. Image clustering algorithms typically consider several features, or dimensions, of the data in an attempt to learn as much as possible about the object similarities. However, a critical problem is that different low-level image descriptors and similarity measures are not designed to be combined naturally and straightforwardly in a meaningful way. Moreover, particular features are often irrelevant for clustering of specific image classes. Thus, they may lead to counterproductive effects negatively affecting the clustering results. For that reason, it is important to learn associations between complex combinations of low-level features and semantic concepts by conveniently weighting the discriminative power of each low-level feature descriptor.

ACO and its pheromone-driven learning mechanism is used to optimise the performance of feature selection in a clustering process. The proposed algorithm overcomes the limitation originated in the assumption that all the clusters in a dataset can be estimated using the same set of features and by assigning weights to features according to the local correlations of the data along each dimension.

1.4 Data Model

For the sake of completeness, some general terms and notations that will be used throughout the paper are defined next. Let $\mathbf{F}$ be the vector containing m visual low-level features:

$$\mathbf{F} = (F_1,\ F_2,\ ...,\ F_m), \tag{5}$$

where the l^{th} component F_l, $1 \leq l \leq m$, is a column vector belonging to feature space $\mathfrak{F}_\mathfrak{l}$. Therefore, each feature set $\mathbf{F}$ lies on the $m-$fold cartesian product $\mathfrak{F} = \mathfrak{F}_1 \times \mathfrak{F}_2 \times \mathfrak{F}_3 ... \times \mathfrak{F}_\mathfrak{m}$. Let's further assume that each feature space $\mathfrak{F}_\mathfrak{l}$, $1 \leq l \leq m$, is endowed with a similarity function D_l. In ideal case D_l represents a metric. However, in many cases the similarity function D_l is not a metric in the mathematical sense. The problem at hand is how to define a suitable similarity function for $\mathfrak{F}_\mathfrak{l}$. The solution to this problem is not straightforward since the feature spaces $\{\mathfrak{F}_\mathfrak{l}\}_{l=1}^m$ can possess different dimensions and topologies. Due to the complex natures of the visual descriptors, they usually possess non-linear behaviours and their direct combination may easily become meaningless. To harmonise the various natures of visual descriptors representing in the same semantic concepts, a simple solution to the problem can be obtained by linear combination of multiple visual feature spaces.

That is, given m valid similarity distances $\{D_l\}_{l=1}^m$ between the corresponding m component feature vectors $\mathbf{F}$ and $\widetilde{\mathbf{F}}$, we define a weighted similarity measure between $\mathbf{F}$ and $\widetilde{\mathbf{F}}$ as

$$D^\alpha(\mathbf{F}, \widetilde{\mathbf{F}}) = \sum_{l=1}^{m} \alpha_l D_l(F_l, \widetilde{F_l}), \tag{6}$$

where the feature weights $\{\alpha\}_{l=1}^m$ are non-negative and sum to 1. We refer to the vector of weights $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)$ as a feature weighting. Observe that obviously $(\mathfrak{F}, D^\alpha)$ represents a metric space, if D_l is also metric for all l.

1.5 Clustering in High-Dimensional Spaces

Suppose that we are given set of n images $\{x_i\}_{i=1}^n$ represented by feature vectors $\{\mathbf{F}_i\}_{i=1}^n$ and we are interested in partitioning them into k disjoint clusters $\{\pi_u\}_{u=1}^k$. Given a partitioning $\{\pi_u\}_{u=1}^k$, for each partition π_u , we write the corresponding generalised centroid as

$$\mathbf{c}_u = (c_{(u,1)}, c_{(u,2)}, ..., c_{(u,m)}), \tag{7}$$

where, $1 \leq l \leq m$, and $c_{(u,l)} \in \mathfrak{F}_\mathfrak{l}$. As an empirical average, the generalised centroid may be thought of as being the closest in metric D^α to all images in the cluster π_u.

Subspace clustering is defined as feature selection procedure that assigns (local) weights to features according to the local correlations of data along each dimension. Motivated by (6), we measure the distortion of each individual cluster π_u, $1 \leq u \leq k$, as

$$\sum_{x \in \pi_u} D^{\alpha_u}(\mathbf{F}, \mathbf{c}_u). \tag{8}$$

The quality of the entire partitioning $\{\pi_u\}_{u=1}^k$ is defined as the combined distortion of all k clusters:

$$\sum_{u=1}^{k} \sum_{x \in \pi_u} D^{\alpha_u}(\mathbf{F}, \mathbf{c}_u). \tag{9}$$

We would like to find k disjoint clusters $\pi_1^*, \pi_2^*, ..., \pi_k^*$ such that the following is minimised:

$$\{\pi_u\}_{u=1}^k = \arg \min_{\{\pi_u\}_{u=1}^k} (\sum_{u=1}^{k} \sum_{x \in \pi_u} D^{\alpha_u}(\mathbf{F}, \mathbf{c}_u), \tag{10}$$

where $\alpha_u = (\alpha_{(u,1)}, \alpha_{(u,2)}, ..., \alpha_{(u,m)})$ is a local feature weighting for all clusters $\{\pi_u\}_{u=1}^k$.

It's important to note that different feature weightings α_u lead to the different similarities D^{α_u}, hence, the minimisation problem (10) is known to be NP-hard. Now we can turn to the crucial question of how to select the "best" feature weighting. To solve this hard combinatorial problem, a novel approach to image clustering based on ACO meta-heuristic is introduced.

1.6 Subspace Clustering Using Ants

In our proposal, The ACO model plays its part in assigning both images and feature weights to a cluster and each ant is giving its own clustering solution [PI09]. The proposed algorithm is outlined next:

Step 1: Initialisation

The whole process starts by choosing the number of clusters k and the number of ants S . Each ant A , $1 \leq A \leq S$, initialises a random centroid c_u and sets the feature weights equally to $1/m$ for each centroid c_u . The pheromone level $\tau_{(i,u)}$ for each ant A is set to 1. To ensure the minimum comparability, it is required that the similarity distances of all images in all considered features spaces are normalised to the same range using conventional Min-Max Normalisation.

Step 2: Clustering

In this step, each ant assigns each image x_i, represented by feature vectors $\{\mathbf{F}_i\}_{i=1}^n$, to the cluster π_u, $1 \leq u \leq k$, with the probability $P_{(i,u)}$ obtained from:

$$P_{(i,u)} = \frac{\tau_{(i,u)}\eta_{(i,u)}}{\sum_{u=1}^{K} \tau_{(i,u)}\eta_{(i,u)}}. \tag{11}$$

In (11), $\eta_{(i,u)}$ is obtained from the following formula:

$$\eta_{(i,u)} = \frac{Q}{D^{\alpha_u}(\mathbf{F}_i, \mathbf{c}_u)}. \tag{12}$$

As the pheromone level $\tau_{(i,u)}$ is initially set to 1, it does not have any effect on the probability at the beginning. The constant Q is heuristicaly estimated to balance the value of η and τ.

Step 3: Computation of weights

For each centroid $\mathbf{c}_u$, and for each feature F_l, new feature weights are computed as follows:

$$\alpha_{(u,l)} = \frac{e^{-R\cdot\overline{D_l}(\mathbf{F},\mathbf{c}_u)}}{\sqrt{\sum_{s=1}^{m} e^{-2\cdot R\cdot\overline{D_s}(\mathbf{F},\mathbf{c}_u)}}}, \tag{13}$$

where $\overline{D_l}(\mathbf{F},\mathbf{c}_u)$ represents the average distance from the centroid $\mathbf{c}_u$ to all images assigned to the cluster π_u along dimension l. That is:

$$\overline{D_l}(\mathbf{F},\mathbf{c}_u) = \frac{1}{|\pi_u|}\sum_{x_i\in\pi_u} D_l(\mathbf{F}_i,\mathbf{c}_u), \tag{14}$$

where $|\pi_u|$ is the cardinality of set π_u and D_l is similarity measure in corresponding feature space $\mathfrak{F}_l$. We empirically determine the value of R in our experiments with synthetic data.

In (13), we use exponential function for feature weighting, in order to make the weights more sensitive to changes in $\overline{D_l}(\mathbf{F},\mathbf{c}_u)$. Even in the first iteration, each ant sets different feature weights due to the random initialisation of the centroid. To facilitate the interpretation of weight values, we require that $\sum_l \alpha_{(u,l)} = 1 \; \forall u$, by properly adjusting the normalisation factor of the weighting scheme.

Step 4: Computation of Centroid

For each image x_i , $1 \leq i \leq n$, new generalised centroids are computed according to clustering of images [Mac67]. Each ant repeats steps 2, 3 and 4 until the optimisation problem (10) is solved.

Step 5: Pheromone update

A widely adopted definition of optimal clustering is a partitioning that the intra cluster similarity is minimised while the inter cluster similarity is maximised [DB79]. In addition, subspace clustering must limit the scope of the criterion function so as to consider different subspaces for each different cluster. Following above definition, we define the average within-cluster distortion and the average between-cluster distortion, respectively, as

$$\mathit{\Gamma}^A(\alpha_u)=\sum_{l=1}^{m}\alpha_{(u,l)}\cdot\overline{D_l}(\mathbf{F},\mathbf{c}_u), \qquad \mathit{\Lambda}^A(\alpha_u)=\sum_{l=1}^{m}\sum_{i=1}^{n}D_l(F_{(i,l)},c_{(u,l)})-\mathit{\Gamma}^A(\alpha_u). \tag{15}$$

After all ants have done their clustering, the assigned pheromone to each solution is incremented. In order to find optimal feature weightings, the pheromone value is updated according to the quality of the solution. For updating the pheromone to each clustering the following formula is used:

$$\tau_{(i,u)}(t) = \rho \cdot \tau_{(i,u)}(t-1) + \sum_{A=1}^{S} \Delta\tau^{A}_{(i,u)}(t), \tag{16}$$

where ρ is the pheromone trail evaporation coefficient $0 \le \rho \le 1$ which causes vanishing of the pheromone over the iterations. $\tau_{(i,u)}(t-1)$ represents the pheromone value from previous iteration. $\Delta\tau^{A}_{(i,u)}(t)$ is a new amount of pheromone calculated from all S ants that assign image x_i to cluster π_u. This approach of marking solutions by pheromone levels is carried out according to

$$\Delta\tau^{A}_{(i,i)}(t) = \begin{cases} \frac{\Lambda^{A}(\alpha_u)}{n.\Gamma^{A}(\alpha_u)} & , if\ x_i\ belongs\ to\ cluster\ \pi_u \\ 0 & , otherwise \end{cases} \tag{17}$$

Intuitively, we would like to minimise $\Gamma^{A}(\alpha_u)$ and to maximise $\Lambda^{A}(\alpha_u)$, that is we like coherent clusters that are well-separated from each other. In other words, more successful ant will put higher amount of pheromone and influence probability of clustering particular image by other ants. After each solution is marked by the pheromone, each ant will start clustering process with new probability of assigning images to clusters. Whole process stops when all ants choose the same clustering solution.

It is important to note, that the proposed algorithm doesn't assign images to clusters based on simply similarity distances between centroids and images. Indeed, the pheromone value carrying the criterion information from the rest of the ants is another important factor. This means that even if image is closest to the centroid of a cluster, it might be assigned to different cluster according to pheromone feedback from the other ants. This change will affect new setting of feature weights for particular ant; hence, it will enable to explore a new solution. In other words, the number of ants S does not implies that the algorithm operates with only S possible solutions.

1.7 Experimental Evaluation

In this section, the proposed technique is comprehensively evaluated using real-world image datasets. First dataset was obtained from The Corel Image database and includes 600 images divided to 6 categories, each consist of 100 images. Second dataset was obtained from the Caltech Image dataset and consist of 3550 images divided to 40 semantic categories. Third dataset consist of 500 images taken from Flickr which are segmented into regions and manually annotated. In order to investigate the clustering performance of the developed method under varying problem complexity, the supported semantic concepts were divided into two subsets containing 5 and 10 concepts. Representative samples of images for each dataset are depicted in Fig. 3.

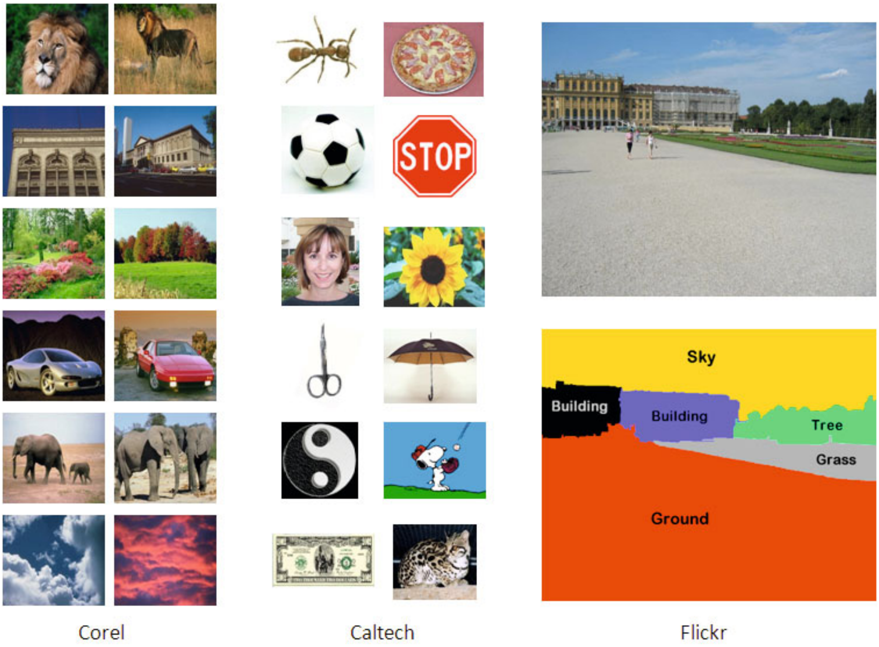
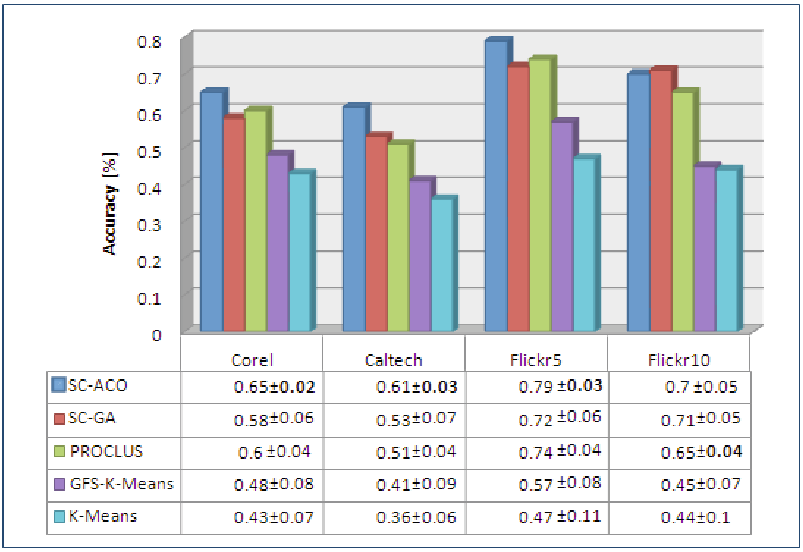


Fig. 3. Several representative images from each database

On each dataset, we compare our subspace clustering approach based on ACO (denoted by SC-ACO) with subspace clustering optimised by GA (denoted by SC-GA), and PROCLUS, and K-Means with global feature selection (denoted by GFS-K-Means), and K-Means with feature weights set equally to $1/m$ for each feature. PROCLUS algorithm is well-known subspace clustering method based on K-Medoid and local feature weighting. Genetic algorithm was integrated with the K-Means for optimising feature weighting.

In general, it is difficult to evaluate the performance of clustering algorithms on high dimensional data. Since there is no universal definition of clustering, there is no universal measure with which to compare clustering results. Cluster quality measures are just heuristics and do not guarantee meaningful clusters, but the clusters found should represent some meaningful pattern in the data in the context of the particular domain, often requiring the analysis of a human expert. Therefore, a gradual approach for the evaluation was taken. First, the proposed algorithm was tested on a small synthetic dataset of images with obvious similarity with respect to given feature but large differences with respect to others. Then we applied our algorithm to the content based image clustering. For evaluation of clustering results, we assume that we are given pre-classified data and benchmark the clustering performance of various approaches against the given ground truth using the clustering accuracy rate. To meaningfully define accuracy rate, we converted the clustering into classification using the following simple rule: identify each cluster with the class that has the largest overlap with the cluster, and assign every image in that cluster to the found class. Since we have the true cluster labels, we can compute clustering accuracy as the number of images correctly classified divided by the total number of images. All tested algorithms were run 10 times and average accuracy rate was computed in order to show the stability of methods. In all our experiments, the number of clusters k was fixed to the number of "true" classes. The number of ants was empirically set to 1000 for experiments on synthetic data and 10000 for experiments with real data. Parameters Q, ρ and R were empirically set in the experiments with synthetic data to be, $Q = 100$, $\rho = 0.2$ and $R = 10$. For visual representation of images, following low-level features (descriptors) were used: Colour Layout (CLD), Colour Structure (CSD), Dominant Colour (DCD), Edge Histogram (EHD) and Grey Level Co-occurrence Matrix (GLC). Observe that the first four are MPEG-7 descriptors [CSP01] while GLC is texture measurement well established from [TJ88].



	Corel	Caltech	Flickr5	Flickr10
🟦 SC-ACO	0.65±0.02	0.61±0.03	0.79 ±0.03	0.7 ±0.05
🟥 SC-GA	0.58±0.06	0.53±0.07	0.72 ±0.06	0.71±0.05
🟩 PROCLUS	0.6 ±0.04	0.51±0.04	0.74 ±0.04	0.65±0.04
🟪 GFS-K-Means	0.48±0.08	0.41±0.09	0.57 ±0.08	0.45±0.07
🟦 K-Means	0.43±0.07	0.36±0.06	0.47 ±0.11	0.44±0.1

Fig. 4. Average accuracy rates for clustering of images/regions by different clustering methods

The clustering performance of all tested algorithms on different image datasets is depicted in Fig. 4. All tested methods depend on initialisation of centroids/medoids, which causes unstable clustering. From experimental results of image/region clustering can be seen that our proposed algorithm performs overall the best and that ACO makes the clustering algorithm more stable. Our experiment results show the need for local feature selection. Furthermore, the proposed algorithm can also provide us better understanding of the underlying process that generates the data.

2 Particle Swarm Optimisation

The initial study on simulating social behaviour of bird flocks and fish schooling were conducted by Reynolds in [Rey87] and Heppner and Grenander in [HG90]. Reynolds was intrigued by the aesthetics of bird flocking choreography, and Heppner was interested in discovering the underlying rules that enabled a large number of birds to flock synchronously, often changing direction. In PSO, the birds in a flock are symbolically represented as particles. These particles are considered flying through a problem space searching for the optimal solution. The location of the particles in a multi-dimensional environment represents the solution to the problem.

The study of Artificial Intelligence (AI) by definition is "the design of intelligent agents" [PMG98], where an intelligent agent is a system that perceives its environment and takes actions which maximise its chances of success [RN03]. The early AI researchers had made an important assumption, so fundamental that it was never stated explicitly nor consciously acknowledged. The researchers assumed that cognition is something inside an individual's head. An AI program was modelled on the vision of a single disconnected person, processing information inside his/her brain. However, humans as species tend to socialise. Thus, in real social interaction, information is exchanged, but also something more important: individuals exchange rules, tips and beliefs about how to process information. Therefore, a social interaction impacts the process of thinking in individuals. Particle Swarm Optimisation has the following two assertions at its heart as discussed in [KE01].

- *Mind is Social.* The notion of cognitivistic perspective of mind as an internal, private thing or process and argue that both function and phenomenon derive from the interactions of individuals in a social world is rejected. Though it is mainstream social science, the statement needs to be made explicit in this age where the cognitivistic view dominates popular as well as scientific thought.
  - *Human intelligence results from social interaction.* Evaluating, comparing and imitating one another, learning from experience and emulating the successful behaviours of others, people are able to adapt to complex environments through the discovery of relatively optimal patterns of attitudes, beliefs and behaviours. Humans predilection for a certain kind of social interaction has resulted in the development of the inherent intelligence of humans.
  - *Culture and cognition are inseparable consequences of human sociality.* Culture emerges as individuals become more similar through mutual social learning. The sweep of culture moves individuals towards more

adaptive patterns of thought and behaviour. The emergent and immergent phenomenon occur simultaneously and inseparably.

- *Particle swarms are a useful computational intelligence methodology.* There are a number of definitions of "computation intelligence" and "soft computing". Computational intelligence and soft computing both include hybrids of evolutionary computation, fuzzy logic, neural networks and artificial life. Central to the concept of computational intelligence is system adaptation that enables or facilitates intelligent behaviour in complex and changing environments.
  - *Swarm intelligence provides a useful paradigm for implementing adaptive systems.* In this sense, it is an extension of evolutionary computation.
  - *Particle swarm optimisation is an extension of, and potentially important new incarnation of, cellular automata* [1]. The topologically structured systems in which the members' topological positions do not vary. Each cell, or location, performs only very simple calculations.

A very simple socio-cognitive theory underlies the Adaptive Culture Model and particle swarms. The process of cultural adaptation comprises a high-level component, seen in the formation of patterns across individuals and the ability to solve problems and a low-level component, the actual and probably universal behaviours of individuals is summarised in terms of three following principles.

1. Evaluate: The tendency to evaluate stimuli - to rate them as positive or negative, attractive or repulsive - is perhaps the most ubiquitous behavioural characteristic of living organisms. Learning cannot occur unless the organism can evaluate, can distinguish features of the environment that attract and features that repel, can tell good from bad. From this point of view, learning could even be defined as a change that enables the organism to improve the average evaluation of its environment.
2. Compare: In almost every aspect of life humans tend to compare with others, whether in evaluating wealth, humour, intelligence or other aspects of opinion and ability. Individuals in the Adaptive Culture Model and particle swarms also compare themselves with their neighbours on the critical measure and imitate only those neighbours who are superior to themselves.
3. Imitate: Humans imitation comprises taking the perspective of the other person, not only imitating a behaviour but realising its purpose, executing the behaviour when it is appropriate. True imitation is central to human sociality and it is central to the acquisition and maintenance of mental abilities.

[1] Cellular Automata consist of a regular grid of cells, each of which can be in only one of a finite number of possible states. The state of a cell is determined by the previous states of a surrounding neighbourhood of cells and is updated synchronously in discrete time steps. [Ros06]

2.1 Algorithm

Originally, PSO was designed for real - valued problems. PSO is initialised with a population of random solutions and each potential solution is assigned a randomised velocity. The potential solutions, referred as particles are flown through hyperspace [EK95]. Each particle has the memory to remember the coordinates in hyperspace which are associated with the best solution (fitness solution it has achieved so far). The value is called *pbest*, another best value is also tracked, which is called *gbest*, it is the global version of the particle swarm optimiser and keeps track of the overall best value. It is the optimal location obtained thus far by any particle in a population.

The PSO consists of at each time step changing the velocity (accelerating) of each particle toward its *pbest* and *gbest*. Acceleration is weighted by a random term, with separate random numbers being generated for acceleration toward *pbest* and *gbest*. The motion of the particles is governed by the velocity update (18) and position update (19).

$$v_{id}(t+1) = v_{id}(t) + c_1(pbest_i(t) - x_{id}(t)) + c_2(gbest_d(t) - x_{id}(t)) \tag{18}$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \tag{19}$$

- $v_{id}(t)$ - represents the velocity of particle
- $pbest_i(t)$ - represents the personal best solution of particle i
- $gbest_d(t)$ - represents the global best solution for $d-$ dimension
- $x_{id}(t)$ - represents the position of the particle
- c_1, c_2 - constant parameters governing cognitive and social interaction respectively.

The trajectory of each individual in the search space is adjusted by dynamically altering the velocity of each particle, according to particles own flying experience and the flying experience of the other particles in the search space. The first summand of (18) represents the velocity at previous time instant, which provides the necessary momentum for particles to roam across the search space. The second summand in (18) is known as cognitive component and represents the knowledge gained by individual particles. The third summand in (18) represents the social modelling of particles, which represents the collaborative effect of the particles, in finding the global optimal solution. The values of the parameters c_2 and c_1 determine the choice between the social and cognitive behaviour of the particles.

The pseudo code for the algorithm is presented below.

1. Initialise a population array of particles with random positions and velocities on d dimensions, in the problem space.
2. For each particle, evaluate the desired optimisation fitness function in d variables.

3. Compare particles fitness evaluation with particles *pbest*. Then set *pbest* value (to make it) equal to the current value, and the *pbest* location equal to the current location in *d*-dimensional space.
4. Compare fitness evaluation with the populations overall previous best. If current value reduces the error to achieve global minima than previous *gbest*, then update *gbest* value.
5. Update the velocity and position of the particle according to (18) and (19).
6. Loop to 2, until the stoping criterion is satisfied.

After some number of iterations the members of the particle swarm populations are found to have congregated around one or more of the optima. In cases where multiple global optima are discovered by the population, topological neighbours tend to cluster in the same regions of the search space. These clusters extend beyond hard-coded neighbourhoods. When an individual finds a relatively optimal combination of elements, it draws its adjacent neighbours toward itself; if the region is superior, the neighbours evaluation will improve as well and they will attract their neighbours and so on. If another subset of the population is attracted to a different but equally good region of the problem space, then a natural separation of groups is seen to emerge, each with its own pattern of coordinates that may easily be thought of as norms or cultures.

When one solution is better than another, it usually ends up absorbing the lesser pattern, though in some cases mediocre “compromise” individuals on the borders of groups prevent the spreading of better solutions through the population. The polarisation of these artificial populations into separate cultures appears very similar to the convergence of human subpopulations on diverse norms of attitude, behaviour and cognition. Interaction results in conformity or convergence on patterns that are similar for proximal individuals and may be different between groups.

2.2 Variants

One of the disadvantages of the PSO is the optimal selection between the two different models. The particles motion can be influenced in one of two ways. The first is called the social behaviour, in which the particle gets attracted to the group centre, i.e. following the group either updating/foregoing the personal best solution. The second behaviour is the cognitive behaviour. In this model, the particle follows the knowledge of self experience without using the knowledge gained as a group. If the social model is followed, then it is likely that, the group could converge at the local minima of the fitness function, without properly exploring the entire problem space. On the other hand, if cognitive model is chosen, the convergence at the sub-optimal value will take a long time and even, may not converge. Since, the particle follows its own path thereby ignoring the knowledge gained as group. To overcome these problems, an optimal choice of both models need to be considered.

Other complex techniques, includes the introduction of inertia weight in the velocity update is given in (20).

$$v_{id}(t) = \omega v_{id}(t-1) + c_1(pbest_i(t-1) - x_{id}(t)) + c_2(gbest_d(t-1) - x_{id}(t)) \quad (20)$$

The time-varying inertia weight (PSO-TVIW) ω [ES01] is introduced to balance between the particles' global and local search abilities. A large inertia weight facilitates global search, while a small inertia weight facilitates local search. It was observed that the optimal solution can be improved by varying the value of ω from 0.9 at the beginning of the search to 0.4 at the end of the search for most problems. However, considering the dynamic nature of real world applications, they have proposed a random inertia weight for tracking dynamic systems, as in (21).

$$\omega = 0.5 + \frac{rand(.)}{2} \quad (21)$$

Where, $rand(.)$ is a uniformly distributed random number within the range [0, 1]. Therefore, the mean value of the inertia weight is 0.75. The disadvantage of PSO-TVIW method is its inability to fine tune the optimum solution, mainly due to the lack of diversity at the end of the search [Ang98]. In [RHW04], a novel parameter automation strategy for the particle swarm algorithm was proposed. Initially, to efficiently control the local search and convergence to the global optimum solution, time varying acceleration coefficients (TVAC) were introduced in addition to the time-varying inertia weight factor. The objective was to enhance the global search in the early part of the optimisation and to encourage the particles to converge toward the global optima at the end of the search. In this new development, authors reduce the cognitive component and increase the social component, by changing the acceleration coefficients c_1 and c_2 with time. With a large cognitive component and small social component at the beginning of the problem space search, particles are exposed to the global search of the problem space. On the other hand, a small cognitive component and a large social component allows the particles to converge to the global optimal in the latter part of the optimisation. The proposed modification is mathematically expressed as in (22) and (23).

$$c_1 = (c_{1f} - c_{1i}\frac{iter}{MAXITE}) + c_1 i \quad (22)$$

$$c_2 = (c_{2f} - c_{2i}\frac{iter}{MAXITE}) + c_2 i \quad (23)$$

Where c_{1i}, c_{1f}, c_{2i} and c_{2f} are constants, $iter$ is the current iteration number and $MAXITE$ is the maximum number of allowable iterations. In PSO, lack of diversity of the population, particularly during the latter stages of the optimisation, was understood as the dominant factor for the convergence of particles to local optimum solutions prematurely. Recently, several attempts on improving the diversity of the population have been reported in the literature, considering the behaviour of the particles in the swarm during the search [LK02] [XZY02]. Furthermore, possible use of the concept "mutation" in PSO as a performance enhancing strategy has also been investigated [HI03]. In evolutionary programming, a mutation function is defined to control the search toward the global optimum solution. However, different forms of mutation functions are used in evolutionary programming and the severity of mutation is decided on the basis of the functional change imposed on the parents. On the other hand, in genetic algorithms, the search toward the global optimum solution is mostly guided by the crossover operation. In PSO, the search toward the global optimum solution is guided by the two stochastic acceleration factors. Therefore, Angeline et al. [Ang98] related these two acceleration factors to the mutation function in evolutionary programming, whereas Shi and Eberhart [YR98] related these two factors to the cross over operation in genetic algorithms.

To control the phenomenon of particles getting caught in the local minima, the authors of [RHW04], enhance the global search via the introduction of a mutation operator, which is conceptually equivalent to the mutation in genetic algorithms. In this new strategy, when the global optimum solution is not improving with the increasing number of generations, a particle is selected randomly and then a random perturbation is added to a randomly selected modulus of the velocity vector of that particle by a predefined probability (mutation probability). However, the mutation step size is set proportionally to the maximum allowable velocity.

In [HJPRY$^+$04], a new hybrid evolutionary-based method combining the particle swarm algorithm and the chaotic search is proposed for optimising. To achieve high performance in optimising, the chaotic search mechanism is embedded in the standard particle swarm algorithm adaptively to avoid the stagnancy of population and increase the speed of convergence. To make the particles escape from stagnancy, the inactive particle should be replaced with freshly created particles adaptively. Chaos is a common phenomenon existing in the non-linear system, which is characterised as ergodicity, randomicity and regularity. The main idea of chaotic search is as follows: chaos queues are generated by iteration of a certain equation, here an equation called Logistic is employed to obtain chaos queues, which are taken into optimisation by carrier wave. It means that chaotic dynamic is amplified into a range where optimisation values are initialised. With the chaotic iteration, the algorithm will find the optimal area effectively. However, chaotic search will sometimes lose its superiority when the search space expands so widely that chaos queues can not reach the optimal area for a short time. So, chaotic search should be restricted into a small time. So, chaotic search should be restricted into a small range in order to obtain high performance in local search.

Liu and Abraham [LAZ07] introduced Turbulence in the particle swarm optimisation (TPSO). The proposed algorithm uses a minimum velocity threshold to control the velocity of particles. TPSO mechanism is similar to a turbulent pump, which supplied some power to the swarm system to explore new search space. The minimum velocity threshold of the particles is tuned adaptively by using a fuzzy logic controller, which is further called as fuzzy adaptive TPSO (FATPSO). The authors discuss the one of the main reason for premature convergence of PSO is due to the stagnation of the particles exploration of a new search space. If a particle's velocity decreases to a threshold a new velocity is assigned to the particles using (24) and (25).

$$v_{ji}(t+1) = \omega \check{v} + c_1 r_1 (pbest_i(t) - x_{id}(t) + c_2 r_2 (gbest_d(t) - x_{id}(t))) \tag{24}$$

$$\check{v} = \begin{cases} v_{ij}, & \text{if } |v_{ij}| \geq v_c; \\ u(-1,1) v_{max}/\rho, & if |v_{ij}| < v_c \end{cases} \tag{25}$$

Where $u(-1,1)$ is the random number, uniformly distributed with the interval [-1,1], and ρ is the scaling factor to control the domain of the particles oscillation according to v_{max} . v_c is the minimum velocity, threshold, a tuneable threshold parameter to limit the minimum of the particles velocity. The performance of the algorithm is directly correlated to two parameter values, v_c and ρ. A large v_c shortens the oscillation period, and it provides a great probability for the particles to leap over local minima using the same number of iterations. But a large v_c compels particles in the quick flying state, which leads them not to search the solution and forcing them not to refine the search.

2.3 Image Classification Using Self Organising Maps

The network architectures and signal processes used to model nervous systems can be categorised as Feedforward, Feedback and competitive. Feedforward networks [RHW86], transform sets of input signals into sets of output signals. The desired input-output transformation is usually determined by external, supervised adjustment of the system parameters. In feedback networks [Hop82], the input information defines the initial activity state of the feedback system, and after state transitions the asymptotic final state is identified as the outcome of the computation. In competitive learning networks, neighbouring cells in a neural network compete in their activities by means of mutual lateral interactions and develop adaptively into specific detectors of different signal patterns.

The basic idea underlying "competitive learning" is briefly presented here: Assume a sequence of statistical samples of a vectorial observable $x = s(t)$ where t is the time coordinate and a set of variable reference vectors $m_i(t) : m_i, \ i = 1, 2, ..., k$. Assume that the $m_i(0)$ have been initialised in some proper way such as random initialisation. If $x(t)$ can be simultaneously compared with each $m_i(t)$ at each successive instant of time, taken here to be integer $t = 1, 2, 3...$, then the best matching $m_i(t)$ is to be updated to match even more closely the current $x(t)$. If the comparison is based on some distance measure $d(x, m_i)$ altering m_i must be such that if $i = c$ is the index of the best-matching reference vector, then $d(x, m_c)$ is decreased, and all the other reference vectors m_i with $i \neq c$ are left intact. In this way, the different reference vectors tend to become specifically "tuned" to different domains of the input variable x.

In competitive neural networks, active neurons reinforce their neighbourhood within certain regions, while suppressing the activities of other neurons [XI05]. This is called on-center/off-surround competition. The objective of SOM is to represent high-dimensional input patterns with prototype vectors that can be visualised in a usually two-dimensional lattice structure [Koh90], [Koh97]. Each unit in the lattice is called a neuron, and adjacent neurons are connected to each other which gives a clear topology of how the network fits itself to the input space. Input patterns are fully connected to all neurons via adaptable weights, and during the training process, neighbouring input patterns are projected into the lattice, corresponding to the adjacent neurons. SOM enjoys the merit of input space density approximation and independence of the order to input patterns.

In the basic SOM training algorithm the prototype vector are trained with (26).

$$m_n(t+1) = m_n(t) + g_{cn}(t)[x - m_n(t)] \tag{26}$$

Where m is the weight of the neurons in the SOM network, $g_{cn}(t)$ is the neighbourhood function that is defined as in (27),

$$g_{cn}(t) = \alpha(t) exp(\frac{||r_c - r_i||^2}{2\alpha^2(t)}) \tag{27}$$

Where, $\alpha(t)$ is the monotonically decreasing learning rate and r represents the position of the corresponding neuron. To further improve the performance of SOM classifier, the weight of the neurons m_d in is optimised with PSO.

Rectangular Self Organising Maps

The initialising topology of the SOM network mesh is a rectangular array of neurons with dimension size equal to the feature vector size. The network topology is shown in Fig. 5. Each neuron represents an image with dimension equal to feature vector. The weights of the neurons are updated by a PSO algorithm. The PSO algorithm updates the weights of the neuron in all the dimensions of the feature vector. The termination of the training step could be achieved in one of two ways. The first approach is to set an iteration limit and once the iteration limit is reached the training process is terminated. The second approach is to iterate until the dissimilarity measure or distance

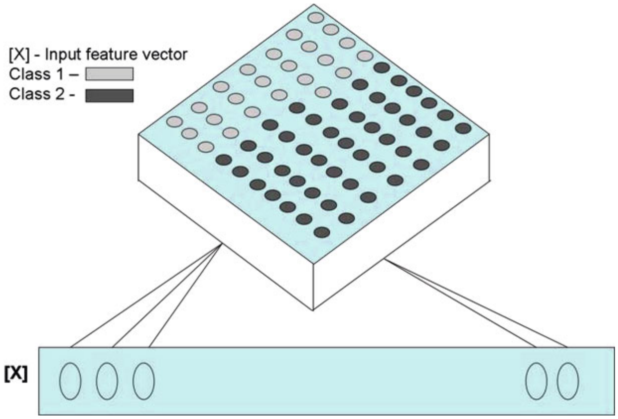


Fig. 5. Rectangular Mesh Self Organising Map

between the input feature vector and particles global best (*gbest*) is minimised below a threshold. The classification step computes the winner node for the corresponding feature vector, which represents the image belonging to specific class. The algorithm for updating the weights of the SOM using PSO [CI06b] is listed below.

1. The rectangular topology of the SOM is initialised with feature vectors $m_i(0)$, $i = 0, 1, 2..., K$randomly, where K is the length of the feature vector.
2. Input feature vector x is presented to the network; choose the winning node J that is closest to x as (28).

$$J = arg_j min||x - m_j|| \tag{28}$$

3. Initialise a population array of particles representing random solutions in d-dimension, of the problem space.
4. For each particle, evaluate the $L1$ norm for x in d dimensions.
5. Compare particle fitness evaluation with particles personal best, *pbest*. Then set *pbest* value equal to the current value, and the *pbest* location equal to the current location in d-dimensional space.
6. Compare fitness evaluation with the populations overall previous best. If current value reduces the error to achieve global minima than previous *gbest*, then update *gbest* value.
7. Update the velocity of each particle as in (29).

$$v_{id}(t+1) = v_{id}(t) + c_1(pbest_i(t) - x_{id}(t)) + c_2(gbest_d(t) - x_{id}(t)) \tag{29}$$

8. Update the position of each particle as in (30).

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \tag{30}$$

9. Loop to Step 2, until the distance between the $m_i(0)$ and x are greater than a threshold value e_{th} .
10. Repeat the Steps 2 and 3 until all the input patterns are exhausted in training.

In steps 2 to 9, the weights of the neuron in SOM are trained to represent the input feature vector. The degree of closeness in pattern matching is determined by the value of e_{th}.

The training set includes both positive and negative samples from the dataset and forms a subset of the testing database. The training models generated online within the network are used to discriminate the image classes. In ideal sense, the network acts as a binary classifier in which one class of images need to be discriminated from the others.

Chaotic Particle Swarm Optimisation

A number of studies have been carried out by various researchers in order to determine the criteria for the choice between social and cognitive models 2. If the social model is followed then it is likely that the group could converge at the local minima of the fitness function, without properly exploring the complete problem space. On the other hand choosing cognitive model has its own disadvantages. If the cognitive model is followed, the convergence at the suboptimal value will be time consuming and may or may not even converge to the optimal solution, because of the fact that each particle follows its own path rejecting the knowledge gained by the group. To overcome this problem, an optimal choice of both models needs to be considered. In combination, the particles personal experience "Personal best (*pbest*)" and its neighbours experience, "Global best (*gbest*)" influence the movement of each particle through a problem space.

In this section, the elementary principle of Chaos is introduced to model the behaviour of particle motion. In our earlier work, the notion of *Chaotic* - PSO was introduced [CI06a] for image classification. The theoretical discussion on Chaotic-PSO with an introduction of wind speed and wind direction to the standard PSO model, in order to model the biological atmosphere for position update of the particles, thus including the dynamics of nature in modelling *Chaos*-PSO. The update of the wind speed is given by the following (31).

$$v_w(t+1) = v_w(t) + v_{op} * rand() + v_{su} * rand() \tag{31}$$

Where v_w is the wind velocity, v_{op} is the opposing direction equal to -1 and v_{su} is the supporting direction equal to 1. The random values are generated to the 6^{th} decimal position. The wind speed has one of two effects. The motion of the particles can be opposed or supported by the wind velocity. The opposing effect slows down the particle in reaching the group global best solution, where the supporting effect increases the particle velocity in reaching in global best solution. Since, each and every particle is separately updated by the wind equation. This is supported by the fact that particles are spatial separated from each other, experiencing different dynamics of atmosphere. When the value of opposing and supporting direction wind velocity equals each other, a static atmosphere is modelled.

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t) + v_w(t) \tag{32}$$

The position update equation for CPSO is given is (32). The introduction of the parameter v_w introduces the elementary theory of chaos in particle swarm optimisation. The initial values of wind speed along the direction plays an important role in determining the final convergence of the particles to the optimal solution. Also, this parameter ensures the optimal searching of the solution space.

Experimental Evaluation

In this section, four different algorithms are evaluated on D700 dataset presented in [DI07]. The details of the algorithms evaluated are listed below:

- SOM - Standard Self Organising Map algorithm, in which the network neurons are trained using the neighbourhood function.
- SOM+GA - Self Organising Maps algorithm, in which the feature vector of the winner node is optimised using Genetic Algorithm.
- SOM+PSO - Self Organising Maps algorithm, in which the feature vector of the winner node is optimised using Standard-Particle Swarm Optimisation technique.
- SOM+Ch-PSO - Self Organising Maps algorithm, in which the feature vector of the winner node is optimised using Chaotic-Particle Swarm Optimisation technique.

The objective of this evaluation is to study the performance of SOM network with and with out the additional optimisation implemented. To this effect, three optimisation techniques are considered, namely Genetic Algorithm, Particle Swarm Optimisation and Chaotic-Particle Swarm Optimisation. The feature set used in the evaluation is MPEG - 7 visual descriptor namely, Colour Layout Descriptor. The feature vector dimension is of size “12”. The visual classifier evaluation is carried out for “7” concepts namely: Building, Car, Cloud, Elephant, Grass, Lion and Tiger. The visual classifier evaluation results is presented in Fig. 6. The evaluation of the framework is presented in terms of “Accuracy”, which is defined as the percentage of the image regions that were assigned to the correct semantic concept.

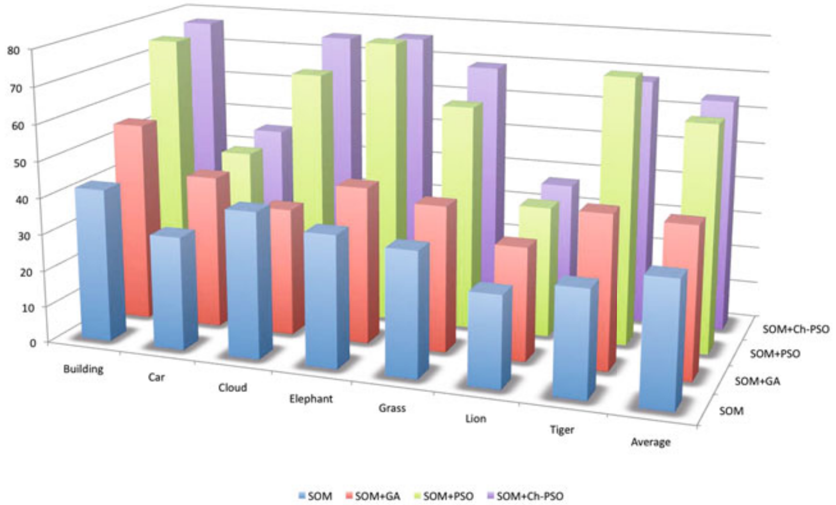


Fig. 6. Classifier Comparison

From the results presented, general observation suggests the application of additional optimisation technique leads to improved performance of the Self Organising Maps network.

Specific observation on the effects of Genetic algorithm and PSO leads to the conclusion that, PSO optimised SOM network provides improved results compared to GA optimised SOM networks. Also, for concepts Building, Car, Cloud, Grass and Lion Chaotic-Particle Swarm Optimisation provides better results compared to standard-PSO technique. However for concepts, Elephant and Tiger, standard-PSO performs better than Chaotic-PSO technique.

The average accuracy for each individual technique SOM, SOM+GA, SOM+s-PSO, SOM+Ch-PSO is 33.9671%, 41.1428%, 62.2% and 64% respectively. From this results, it is evident that, the s-PSO and Ch-PSO optimisation techniques improves the performance of the SOM network under similar test conditions compared to other techniques.

The improved performance of PSO could be attributed to advantages of PSO optimisation technique compared to the genetic algorithm: (i) PSO does not suffer from some of GA's difficulties in interacting with the group members and rather detracts from progress towards the solution; (ii) a particle swarm system has memory, where as the genetic algorithm does not have. Change in genetic populations results in destruction of previous knowledge of the problem, except when elitism is employed, in which case usually one or a small number of individuals retain their identities. In PSO, individuals who fly past optima are tugged to return towards them; and also the knowledge of good solutions are retained by all particles.

2.4 Media Retrieval

Content - based image retrieval (CBIR) exploits visual content descriptions to index and search images from large scale image databases. It has been an active and fast advancing research field over the last decade. CBIR uses visual information extracted from a media such as colour, shape and texture to represent and index the database. In typical CBIR systems, the visual contents of the images in the database are extracted and described by multi-dimensional feature vectors. To retrieve media, users provide the retrieval system with query samples. The system then translates these query(ies) into its internal representation of feature vectors. The similarities/distances between the feature vectors are then calculated and the database is accordingly ranked with the aid of an indexing scheme. The indexing scheme provides an efficient way to search in the image database. Recent retrieval systems have incorporated users' relevance feedback to modify the retrieval process in order to generate perceptually and semantically more meaningful retrieval results.

Unlike textual information, which is human defined and precise in meaning, a picture has a hidden component of creative reasoning of the human brain. This provides the content an overall shape and meaning far beyond capabilities of any language-based representation. Early approaches for image retrieval were based on keywords and manually annotated images inspired by information retrieval [vR79] in text documents. Though manual annotations were developed to preserve knowledge they are burdensome and dependent on subjective interpretations of the professional annotator, thereby resulting in low performance of CBIR system. However, incorporating users judgment on similarity of some media items during a relevance feedback session is a consequence of the knowledge the user has build up through his/her life. Therefore a level of this semantic information is transferred onto the similarity model in order to capture human notion of semantic similarity. Several researchers have worked on building relational base of concepts and content through the use of iterative relevance feedback is presented in [ZLZ01], [MLC98]. The objective of the system is to build a semantic network on top of the keyword association, leading to enhanced deduction and utilisation of semantic content.

In designing a CBIR system, the first and the most important assumption is that discrimination between relevant and non-relevant items is possible with the available features. Without this condition satisfied relevance feedback is futile. There can be established a relatively straightforward transformation between the topology of the feature space and the semantic characteristics of the items the user wants to retrieve. There are relevant items in the archive and they are a small part of the entire available collection. If such items form the majority in the collection the performance of the retrieval process might become limited and sometimes inadequate feedback information is feedback by predominantly labelling positive items and less often negative items.

To simulate human visual perception, multiple low-level features are extracted from image content needs to be considered. The aim is to obtain information from different low-level visual cues at various levels of complexity and to jointly exploit that information to obtain higher levels of conceptual abstraction. Low-level descriptors are very useful to search for patterns of interest and similarities in image database. The proposed system, as shown in Fig. 7, consists of two main subsystems. The first subsystem runs offline and embraces two processing steps. The aim of this step is to extract the different low-level features from the image dataset. The extracted features are stored in the metadata repository. The metadata repository is then further indexed based on the image id's. The second subsystem involves online interaction with the user and comprises a number of processing steps. The second subsystem consists of two online search modules namely "visual search" and "RF system" which are discussed in detail in the following subsections. The reminder of this section will discuss the workflow of the framework.

The interaction is initialised by randomly presenting the user with equal distribution of the database. The user marks only the relevant images from the presented results. The first user interaction inputs are presented to the "visual search module". The visual search module implicitly generates a model for irrelevant model and performs the retrieval. The objective of this step is to infer and predict the user preferences. From the set of results presented from first iteration, the user selects both relevant and irrelevant images and the input is presented to "RF System". The aim of this step is to enhance the inference of the user preferences in order to improve the image retrieval. The user is then iteratively interacts with the system until the user has retrieved all relevant documents or satisfied with the retrieved results. The proposed RF system has been presented in Fig. 7.

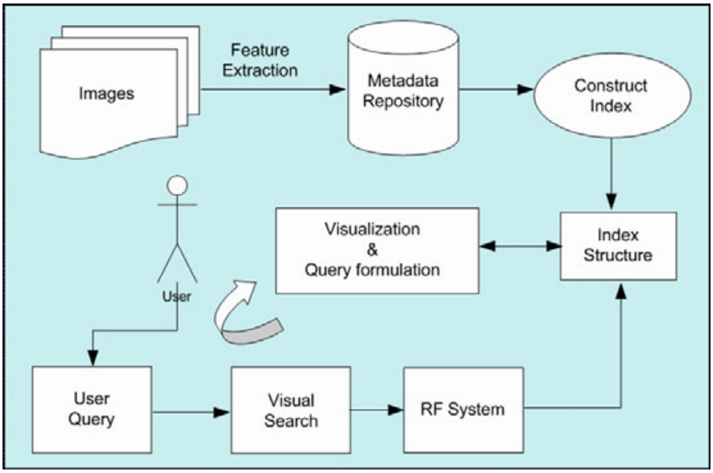


Fig. 7. Proposed RF Framework

2.5 Experimental Results

Feature Set. The MPEG - 7 visual descriptors namely Colour Layout Descriptor (CLD) [MSS03], [MOVY01] and Edge Histogram Descriptor (EHD) are extracted for images in the following datasets. The CLD extracts colour histograms over 8 X 8 image layout. Its similarity measure is a weighted metric with nonlinearly quantised DCT coefficients. The EHD builds on histograms of edges in different directions and scales. Detected edges in a number of directions are used as localised input edge histogram of 80 bins. Its distance is a sum of distances over the original features, as well as global and semi-global histogram values generated by various grouping of local image parts.

PSO Implementation. The PSO model implemented is a combination of cognitive and social behaviour. The structure of the PSO is "fully connected" in which a change in a particle affects the velocity and position of other particles in the group as opposed to partial connectivity, where a change in a particle affects the limited number of neighbourhood in the group. Each dimension of the feature set is optimised with 50 particles. The size of the SOM network is pre-fixed with the maximum number of training samples to be used in the network. The stopping criteria threshold is set to 1.0. The value of the threshold indicated the closeness in solving the optimisation problem.

Corel Dataset. The database used in the experiments is generated from Corel dataset consisting of seven concepts. In [DI07], an overview of the selected dataset is presented. The dataset includes concepts building, cloud, car, elephant, grass, lion and tiger, with the following number of ground truth

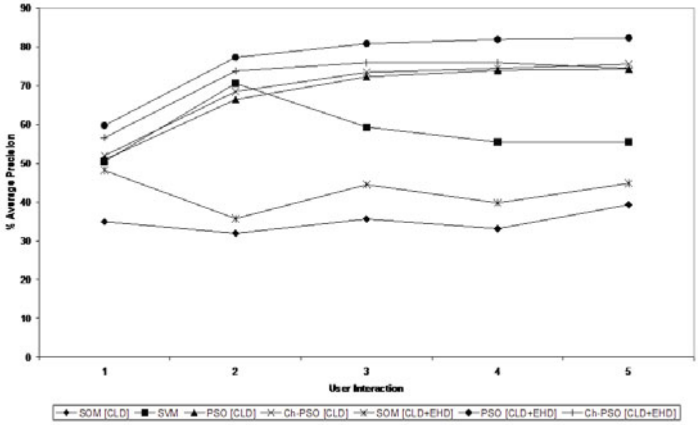


Fig. 8. Average Results

images per concept: 141, 264, 100, 100, 279, 100 and 100 respectively. The Corel database has been specifically modelled for seven concepts and though it is a smaller size it consists of natural images with a variety of background elements with overlapping concepts which make the dataset complex. In Fig. 8, the average precision of the retrieval system has been presented.

3 Conclusion and Future Work

The recent revolutionary development of multimedia processing techniques, combined with the rapid increase in computational capability and decrease in storage and transmission costs, has led to a proliferation of digital multimedia content. One of the most crucial aspects of current interactive multimedia systems is the functionality of indexing and retrieval of the visual information in response to a query. However, the semantic gap between human perception of multimedia content and the automatic interpretation derived from machine remains a formidable challenge. As discussed, the evaluation of the proposed image clustering and classification has shown significant performance improvement over the application of conventional evolutionary computation techniques.

References

[Ang98] Angeline, P.J.: Evolutionary optimization versus particle swarm optimization: Philosophy and the performance difference. In: Porto, V.W., Waagen, D. (eds.) EP 1998. LNCS, vol. 1447, pp. 600–610. Springer, Heidelberg (1998)

[CI06a] Chandramouli, K., Izquierdo, E.: Image classification using chaotic particle swarm optimization. In: IEEE International Conference on Image Processing, Atlanta, USA, pp. 3001–3004 (October 2006)

[CI06b] Chandramouli, K., Izquierdo, E.: Image classification using self organising feature maps and particle swarm optimisation. In: Proc. 7th Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2006), pp. 313–316 (2006)

[CSP01] Chang, S.F., Sikora, T., Purl, A.: Overview of the mpeg-7 standard. IEEE Transactions on Circuits and Systems for Video Technology 11(6), 688–695 (2001)

[DAGP90] Deneubourg, J.L., Aron, S., Goss, S., Pasteels, J.-M.: The self-organizing exploratory pattern of the argentine ant. Journal of Insect Behavior 3, 159–168 (1990)

[Das99] Dasgupta, D.: Artificial Immune Systems and their Applications. Springer, Heidelberg (1999)

[DB79] Davies, D., Bouldin, D.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 224–227 (1979)

[DC99] Dorigo, M., Di Caro, G.: Ant algorithms for discrete optimization. Technical report, Universite Libre de Bruxelles (1999)

[DG96] Dorigo, M., Gambardella, L.: A study of some properties of ant-q. In: Ebeling, W., Rechenberg, I., Voigt, H.-M., Schwefel, H.-P. (eds.) PPSN 1996. LNCS, vol. 1141, pp. 656–665. Springer, Heidelberg (1996)

[DG97] Dorigo, M., Gambardella, L.M.: Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computing 1, 53–66 (1997)

[DI07] Djordjevic, D., Izquierdo, E.: An object- and user- driven system for semantic-based image annotation and retrieval. IEEE Trans. on Circuits and Systems for Video Technology 17(3), 313–323 (2007)

[EK95] Eberhart, R., Kennedy, J.: A new optimizer using particle swarm theory. In: Proc. of the Sixth International Symposium on Micro Machine and Human Science, pp. 39–43 (October 1995)

[ES01] Eberhart, R., Shi, Y.: Particle swarm optimization: Developments, application and resouces. In: Proceedings of the 2001 Congress, vol. 1, pp. 81–86 (2001)

[Fog94] Fogel, L.J.: Evolutionary Programming in perspective. In: Computation Intelligence: Imitating Life, pp. 135–146. IEEE Press, Los Alamitos (1994)

[HG90] Heppner, F., Grenander, U.: A Stochastic nonlinear model for coordinated bird flocks. In: Krasner, S. (ed.) The Ubiquity of Chaos. AAAS Publications, Washington (1990)

[HI03] Higashi, N., Iba, H.: Particle swarm optimization with guasssian mutation. In: Proc. Of the IEEE Swarm Intelligence Symposium, pp. 72–79 (2003)

[HJPRY$^+$04] Hong-Ji, M., Peng, Z., Rong-Yang, W., Jing, X., Zhi, X.: A hybrid particle swarm algorithm with embedded chaotic search. In: IEEE Conference on Cybernatics and Intelligent Systems, vol. 1, pp. 367–371 (2004)

[Hop82] Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. 79, 2554–2558 (1982)

[KE01] Kennedy, J., Eberhart, R.C.: Swarm Intelligence. Morgan Kaufmann, San Francisco (2001)

[Koh90] Kohonen, T.: The self organizing map. Proceedings of IEEE 78(4), 1464–1480 (1990)

[Koh97] Kohonen, T.: Self-Organizing Maps, 2nd edn. Springer, Berlin (1997)

[LAZ07] Liu, H., Abraham, A., Zhang, W.: A fuzzy adaptive turbulent particle swarm optimisation. International Journal of Innovative Computing and Applications 1(1), 39–47 (2007)

[LK02] Lovbjerg, M., Krink, T.: Extending particle swarm optimizers with self organized critically. In: Proc. IEEE Int. Congr. Evolutionary Computation, vol. 2, pp. 1570–1593 (May 2002)

[Mac67] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: LeCam, L.M., Neyman, J. (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability, Berkley, CA, pp. 281–297. University of California Press, Berkeley (1967)

[MLC98] Sclaroff, S., La Cascia, M., Sethi, S.: Combining textual and visual cues for contnet based image retrieval on the world wide web. In: IEEE Workshop on Content based Access of Image and Video Libraries, pp. 24–28 (1998)

[MOVY01] Manjunath, B.S., Ohm, J.-R., Vinod, V.V., Yamada, A.: Color and texture descriptors. IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG - 7 11(6), 703–715 (2001)

[MSS03] Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG - 7, Multimedia Content Description Interface. Wiley, New York (2003)

[PI09] Piatrik, T., Izquierdo, E.: Subspace clustering of images using ant colony optimisation. In: Proceedings of 16th International Conference on Image Processing, ICIP (2009)

[PMG98] Poole, D., Mackworth, A., Goebel, R.: Computational Intelligence: A Logical Approach. Oxford University Press, Oxford (1998)

[Rey87] Reynolds, C.W.: Flocks, herds and schools: a distributed behavioural model. In: Computer Graphics, pp. 25–34 (1987)

[RHW86] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, pp. 318–362 (1986)

[RHW04] Ratnaweera, A., Halgamuge, S.K., Watson, H.C.: Self-organizing hierarchical particle swarm optimizer with time varying acceleration coefficients. IEEE Trans. on Evolutionary Computation 8(3), 240–255 (2004)

[RN03] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd edn. Prentice-Hall, Englewood Cliffs (2003)

[Ros06] Rosin, P.L.: Training cellular automata for image processing. IEEE Trans. on Image Processing 15(7), 2076–2087 (2006)

[SH96] Stutzle, T., Hoos, H.: Improving the ant-system: A detailed report on the max-min ant system. AIDA 66, FG Intellektik (August 1996)

[TJ88] Tuceryan, M., Jain, A.K.: Texture Analysis. The Handbook of Pattern Recognition and Computer Visions, 2nd edn. World Scientific Publishing Co., Singapore (1988)

[vR79] van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworths (1979)

[Wil75] Wilson, E.O.: Sociobiology: The new synthesis. Belknap Press, Cambridge (1975)

[XI05] Xu, R., Wunch II., D.: Survey of clustering algorithms. IEEE Trans. Neural Network 6(3), 645–678 (2005)

[XZY02] Xie, X.F., Zhang, W.J., Yang, Z.L.: A dissipative particle swarm optimization. In: Proc. IEEE Congr Evolutionary Computation, vol. 2, pp. 1456–1461 (May 2002)

[YR98] Shi, Y., Eberhart, R.: Computation between genetic algorithms and particle swarm optimization. In: Porto, V.W., Waagen, D. (eds.) EP 1998. LNCS, vol. 1447, pp. 611–616. Springer, Heidelberg (1998)

[ZLZ01] Zhang, L., Lin, F., Zhang, B.: Support vector machine learning for image retrieval. In: Proceedings of International Conference on Image Processing, vol. 2 (2001)

Robust Image Watermarking Based on Feature Regions

Cheng Deng[1], Xinbo Gao[1], Xuelong Li[2], and Dacheng Tao[3]

[1] School of Electronic Engineering, Xidian University,
Xi'an 710071, P.R. China
chdeng@mail.xidian.edu.cn, xbgao@mail.xidian.edu.cn

[2] State Key Laboratory of Transient Optics and Photonics,
Xi'an Institute of Optics and Precision Mechanics,
Chinese Academy of Sciences, Xi'an 710119, P.R. China
xuelong_li@opt.ac.cn

[3] School of Computer Engineering, Nanyang Technological University,
50 Nanyang Avenue, Blk N4, 639798, Singapore
dacheng.tao@gmail.com

Abstract. In image watermarking, binding the watermark synchronization with the local features has been widely used to provide robustness against geometric distortions as well as common image processing operations. However, in the existing schemes, the problems with random bending attack, nonisotropic scaling, general affine transformation, and combined attacks still remain difficult. In this chapter, we present and discuss the framework of the extraction and selection of the scale-space feature points. We then propose two robust image watermarking algorithms through synchronizing watermarking with the invariant local feature regions centered at feature points. The first algorithm conducts watermark embedding and detection in the *affine covariant regions* (ACRs). The second algorithm is combining the *local circular regions* (LCRs) with Tchebichef moments, and *local Tchebichef moments* (LTMs) are used to embed and detect watermark. These proposed algorithms are evaluated theoretically and experimentally, and are compared with two representative schemes. Experiments are carried out on a set of standard test images, and the preliminary results demonstrate that the developed algorithms improve the performance over these two representative image watermarking schemes in terms of robustness. Towards the overall robustness against geometric distortions and common image processing operations, the LTMs-based method has an advantage over the ACRs-based method.

1 Introduction

With the rapid developments of information technologies, digital media can be accessed, distributed, and copied in many convenient ways, but this also leads to the problem of illegal redistribution and manipulations. Therefore, there is an increasing concern over copyright protection of digital media. Traditionally, cryptography can be used to protect the copyright of digital contents. But once the content is decrypted, there is no way to control its subsequent uses. As a popular and powerful

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 111–137.
springerlink.com

technique, digital watermarking has been widely studied to manage the intellectual property of digital contents. It inserts visible or invisible copyright information, termed watermark, into the content. Ownership of the contents can be verified by retrieving the inserted information. In this chapter, we focus on the digital watermarking technologies to protect the digital content copyright.

For a desired image watermarking system, the watermark is supposed to be robust against various attacks. Usually, these attacks can be classified into two categories: common image processing operations and geometric distortions. Geometric distortions are considered as one of the most difficult attacks to resist. They induce synchronization errors between the extracted watermark and the original watermark, and therefore disable the detector even though the watermark still exists in the watermarked image. Nowadays, a few specialized watermarking approaches have presented to address the geometric distortions. These approaches can be roughly divided into four categories.

The first category is to perform an exhaustive random search for the watermark over the space containing the set of acceptable attack parameters. One concern in the exhaustive search is computational cost. The larger the search space, the more accurate the synchronizer outcome, but it also requires more computation to perform it. Another is the false-positive probability as it increases with the size of the search space [1]-[2].

The second category includes those which embed watermark in the geometrically invariant domain. In [3]-[4], the watermark was embedded in the magnitude part of Fourier-Mellin transform. A rotation, scaling and translation invariant domain is obtained by the Fourier transform after *log-polar mapping*(LPM). However, those watermarking algorithms are known as the implementation difficulties due to the use of LPM. In [5], phase correlation was used to resynchronize the watermarked image and avoid the *inverse log-polar mapping* (ILPM). Watermarking techniques involving invariant domains suffer from implementation issue and are usually vulnerable to cropping.

The watermarking techniques belonging to the third category use a template or insert a periodic watermark pattern for the purpose of resynchronization. In [6] and [7], templates were embedded in the *discrete Fourier transform*(DFT) domain to generate the shape of local peaks. The local peaks are searched in the detection process to identify transformations undergone by the image. The performance of the template-based methods depends on the dimensionality of the attack parameter space. For some complicated attacks, these template-based methods will be incapable of estimating the attack parameters. In [8], a self-reference watermark generated as a special structural pattern was embedded in the spatial domain. In [9], Delannay and Macq designed 2-D cyclic patterns to achieve resynchronization. Dugelay *et al.* added predefined additional information in the useful message bits at the embedding step. During the extraction step, these bits are then used as anchor points to estimate and compensate for global/local geometric distortions [10]. The major drawback of the periodical pattern-based techniques is relatively vulnerable to watermark estimation attack, such as collusion attack.

The fourth category includes those methods which exploit features invariant to geometric distortions. These invariant features may be the whole image, some invariant regions or invariant feature points [11]. By binding the watermark synchronization with the image features, watermark detection can be done without synchronization error. This class of watermark synchronization techniques is also called second generation watermarking [12]. In [13]-[16], image moments were employed to embed watermark. In spite of the robustness against global affine transformations, the moment-based approaches are highly vulnerable to cropping. In [17], a region-based watermarking scheme was proposed by segmenting an image into a set of regions. Two largest regions approximated by ellipsoids are chosen as the embedding area for the watermark. The problem of this method is that the image segmentation depends on the image contents so that image distortions seriously affect the segmentation results [18]. In [19], the Harris detector was used to extract feature points to decompose an image into a set of meshes by using Delaunay tessellation. Both the watermark embedding and detection are conducted in the normalized meshes. In [20], Tang and Hang adopted Mexican-Hat wavelet filtering to extract feature points, and image normalization was then applied to these non-overlapped disks centered at the extracted feature points. In [21],the scale-space theory was applied for feature point extraction. For a chosen feature point, a specific geometric shape is formed and used for embedding watermark.

After surveying the existing watermarking methods that provide a certain degree of robustness against geometric distortions, we have observed that the feature-point-based watermarking methods exhibit more promising than others in terms of robustness. This is mainly because that 1) the extracted feature points are proven to be robust against many common image processing operations and geometric distortions; 2) the watermarks are embedded in a number of local regions centered at these feature points, which increase the ability of the watermark to resist cropping. However, there are some main drawbacks indwelled in current feature-point-based schemes. First, the feature point extraction techniques adopted by the current feature-point-based approaches, such as Harris detector [19] and Mexican-Hat wavelet filtering [20], are sensitive to image modification. Secondly, the fixed value is used to determine the size of local regions so that the watermarking scheme is vulnerable to the scaling of an image [22]. Thirdly, it still remains difficult for the current feature-point-based watermarking schemes to resist *random bending attack*(RBA), nonisotropic scaling, general affine transformation, and combined attacks. These practical problems existed in the current feature-point-based watermarking methods restrict the robustness against geometric distortions and common image processing operations. To this end, two image watermarking based on local feature regions are developed and compared in this chapter.

In the method based on *affine covariant region* (ACRs), the Harris-Affine detector is adopted to extract ACRs. The feature selection criterion based on the graph theoretical clustering algorithm is then employed to select a set of nonoverlapped ACRs for watermark embedding. In order to achieve affine invariance, each region is locally normalized by transforming an ellipse into a circle and rotated to align with its dominant gradient orientation. In watermark embedding, circular watermark pattern is embedded in the normalized patch. For the purpose of imperceptibility after watermarking, an image-dependent visual model is utilized to adjust the embedding strength.

In the method based on *local Tchebichef moments* (LTMs),feature points are extracted by the Harris-Laplace detector and then selected by the proposed feature selection criterion. For each chosen feature point, we construct a *local circular region* (LCR) that is invariant to geometric distortions. The *Tchebichef moments* (TMs) are then employed to describe the global characteristics of the local invariant region. Obviously, the extracted LTMs are independent to the slight changes of pixels in this region. Here we select TMs rather than others because TMs have not only insensitivity to noise but also better feature representation capability and reconstruction accuracy. The magnitudes of LTMs are modified through quantization index modulation, which can achieve the blind detection and improve the detection accuracy. Extensive experimental results show that the proposed schemes are resilient to various geometric distortions as well as common image processing operations and outperform the existing representative works.

The reminder of this paper is organized as follows: In Section 2, some preliminaries are given, including feature points extraction and feature points selection; in Section 3, two image watermarking algorithms based on feature regions are described in detail, respectively. Experimental results and detailed analysis are given in Section 4, and conclusion is drawn finally.

2 Preliminaries

In the framework of the proposed watermarking methods, feature points extraction and feature points selection will play important roles in achieving the desired goal, and they will be discussed in this section.

2.1 Feature Point Extraction

2.1.1 Harris-Laplace Detector

Harris-Laplace detector [23] is the improvement of Harris detector. To obtain the invariance to scale changes, this detector first calculates a set of images represented at different resolution levels for reliable Harris detector. It then selects feature points with an automatic scale selection procedure. For Harris-Laplace detector, the scale-normalized second moment matrix is defined as

$$\mu(\mathbf{x},\sigma_I,\sigma_D)=\sigma_D^2 g(\sigma_I)*\begin{bmatrix} L_x^2(\mathbf{x},\sigma_D) & L_xL_y(\mathbf{x},\sigma_D) \\ L_xL_y(\mathbf{x},\sigma_D) & L_y^2(\mathbf{x},\sigma_D)\end{bmatrix}, \tag{1}$$

where σ_I is the integration scale, σ_D is the differentiation scale, L_s is the derivative computed in the s direction, and $*$ denotes the convolution operator over $\mathbf{x}\in\Re^2$. The scale-space representation is a set of images represented at different levels of resolutions. Given the σ_D, the uniform Gaussian scale-space representation L is

$$L(\mathbf{x}, \sigma_D) = g(\mathbf{x}, \sigma_D) * f(\mathbf{x}), \tag{2}$$

where $g(\mathbf{x}, \sigma_D)$ is the uniform Gaussian kernel with standard deviation σ_D and mean zero, f is an image, and $*$ is convolution operator.

Given σ_I and σ_D, the strength measure of this detector can be computed as

$$m(\mathbf{x}, \sigma_I, \sigma_D) = \det(\mu(\mathbf{x}, \sigma_I, \sigma_D)) - k \cdot \mathrm{trace}^2(\mu(\mathbf{x}, \sigma_I, \sigma_D)), \tag{3}$$

where k is an predefined constant. At each level of the scale space, the local maxima of the strength measure are regarded as the feature points.

The idea of automatic scale selection is to select the characteristic scale of a local structure, for which a given function attains an extremum over scales. Normalized *Laplacian-of-Gaussian*(LoG) operator is used for finding the characteristic scale. The LoG is defined as

$$|LoG(\mathbf{x}, \sigma_I)| = \sigma_I^2 \left| L_{xx}(\mathbf{x}, \sigma_I) + L_{yy}(\mathbf{x}, \sigma_I) \right|, \tag{4}$$

where L_{xx} and L_{yy} are second partial derivatives with respect to x and y, respectively.

For each candidate point, an iterative algorithm is applied for detecting the location and the scale of feature points. The extrema over scales of the LoG are used to select the scale of feature points. Given an initial point $\mathbf{x}$ with scale σ_I , the iteration steps are:

Step 1: Find the local extremum over scale of the LoG for the point $\mathbf{x}_k$, otherwise, reject the point. The investigated range of scales is limited to $\sigma_I^{(k+1)} = t\sigma_I^{(k)}$ with $t \in [0.7, \cdots, 1.4]$.

Step 2: Detect the spatial location $\mathbf{x}_{k+1}$ of a maximum of the strength measure nearest to $\mathbf{x}_k$ for selected $\sigma_I^{(k+1)}$.

Step 3: Go to *Step* 1 if $\sigma_I^{(k+1)} \neq \sigma_I^{(k)}$ or $\mathbf{x}_{k+1} \neq \mathbf{x}_k$.

Since these feature points can only provide position information, the neighborhood of the points are required for watermark embedding and detection. For each feature point, a LCR can be constructed. The radius of the LCR is

$$R = \tau \cdot [\sigma] \tag{5}$$

where $[\cdot]$ is rounding operation, σ is the characteristic scale, and τ is a positive integer, which is used to adjust the size of a LCR. Because the radius of each LCR is in direct proportion to its corresponding characteristic scale, the LCR can be covariant to the image content changes, such as scaling.

2.1.2 Harris-Affine Detector

Lindeberg and Garding [24] extended the notion of the scale space to the affine Gaussian scale space. Then, the rotationally symmetric Gaussian window function in Eq. (2) is substituted by an affine Gaussian scale-space elliptical window function:

$$g(\mathbf{x},\Sigma)=\frac{1}{2\pi\sqrt{\det\Sigma}}\exp\left(\frac{-\mathbf{x}^T\Sigma^{-1}\mathbf{x}}{2}\right), \tag{6}$$

where Σ is a symmetric positive-definite 2×2 matrix. Then the affine Gaussian scale-space representation is defined by

$$L(\mathbf{x},\Sigma)=g(\mathbf{x},\Sigma)*f(\mathbf{x}). \tag{7}$$

In the affine Gaussian scale space, the second moment matrix μ at a given point $\mathbf{x}$ is described by

$$\mu(\mathbf{x},\Sigma_s,\Sigma_t)=g(\mathbf{x},\Sigma_s)*\left[(\nabla L)(\mathbf{x},\Sigma_t)(\nabla L)(\mathbf{x},\Sigma_t)^T\right], \tag{8}$$

where Σ_s and Σ_t are respectively covariance matrices corresponding to the integration scale and local scale.

The Harris-Affine detector is based on the affine normalization around multi-scale Harris points. After a set of feature points are detected by Harris-Laplace detector, an iterative procedure is utilized to estimate elliptical ACRs around the feature points, wherein these regions are relative invariant to arbitrary affine transformations [25]. Consider an image $I(\mathbf{x})$ and its corresponding linearly transformed image $\widetilde{I}=I(\mathbf{Bx})$, the affine scale-space second moment matrices μ and $\widetilde{\mu}$ of I and $\widetilde{I}$ respectively satisfy

$$\mu(\mathbf{x},\Sigma_s,\Sigma_t)=\mathbf{B}^T\widetilde{\mu}(\mathbf{Bx},\mathbf{B}\Sigma_s\mathbf{B}^T,\mathbf{B}\Sigma_t\mathbf{B}^T)\mathbf{B}. \tag{9}$$

It can be proven that if the following property holds for the second moment matrix $\mathbf{M}_L$ at a fixed point $\mathbf{x}_L$ of the image $I(\mathbf{x})$, i.e.,

$$\mu_L(\mathbf{x}_L,\Sigma_{s,L},\Sigma_{t,L})=\mathbf{M}_L,\ \Sigma_{s,L}=s\cdot\mathbf{M}_L^{-1},\ \Sigma_{t,L}=t\cdot\mathbf{M}_L^{-1}, \tag{10}$$

then the second moment matrix at the point $\mathbf{x}_R(\mathbf{x}_R=\mathbf{B}\mathbf{x}_L)$ of the image $\widetilde{I}(\mathbf{x})$ satisfies the following property

$$\widetilde{\mu}_R(\mathbf{x}_R,\Sigma_{s,R},\Sigma_{t,R})=\mathbf{M}_R,\ \Sigma_{s,R}=s\cdot\mathbf{M}_R^{-1},\ \Sigma_{t,R}=t\cdot\mathbf{M}_R^{-1}, \tag{11}$$

where $\mathbf{M}^{-1}$ is the inverse of second moment matrix $\mathbf{M}$ and $s,t\in\Re^+$.

Baumberg [26] employed the square root of the second moment matrix to transform an image to a normalized one. The transformed images of $I(\mathbf{x})$ and $\widetilde{I}(\mathbf{x})$ are respectively defined as

$$I'(\mathbf{M}_L^{-1/2}\mathbf{x})=I(\mathbf{x}),\ \widetilde{I}'(\mathbf{M}_R^{-1/2}\mathbf{x})=\widetilde{I}(\mathbf{x}), \tag{12}$$

where $\mathbf{M}^{-1/2}$ is the square root matrix of $\mathbf{M}$. By using the transformation property, the normalized image can be shown as

$$\mu_L'(\mathbf{M}_L^{-1/2}\mathbf{x}_L,s\mathbf{I},t\mathbf{I})=\mu_R'(\mathbf{M}_R^{-1/2}\mathbf{x}_R,s\mathbf{I},t\mathbf{I})=\mathbf{I}, \tag{13}$$

where $\mathbf{I}$ is the 2×2 identity matrix. Thus, the affine transformation can be expressed as

$$\mathbf{B} = \mathbf{M}_R^{-1/2} \mathbf{R} \mathbf{M}_L^{-1/2}, \tag{14}$$

where $\mathbf{R}$ is a rotational matrix.

If the neighborhood of points $\mathbf{x}_L$ and $\mathbf{x}_R$ are normalized by transformations $\mathbf{x}'_L = \mathbf{M}^{1/2}\mathbf{x}_L$ and $\mathbf{x}'_R = \mathbf{M}^{1/2}\mathbf{x}_R$, respectively, the normalized images are related according to $\mathbf{x}_L = \mathbf{R}\mathbf{x}'_R$, which is simply a rotational transformation.

We further simplify the Harris-Affine detector with four steps: (1) detect initial feature points by using multi-scale Harris detector; (2) select the characteristic scale for each of the located feature points; (3) determine the shape of a point by the eigenvalues and eigenvectors of the second moment matrix; and (4) normalize ACRs according to $\mathbf{x}' = \mathbf{M}^{1/2}\mathbf{x}$.

As mentioned above, the elliptical shape of the ACRs are determined by the second moment matrix $\mathbf{M}$,

$$(\mathbf{x} - \mathbf{x}_0)^T \mathbf{M} (\mathbf{x} - \mathbf{x}_0) \leq 1, \tag{15}$$

where $\mathbf{x}_0$ is the center of the elliptical region, namely, the location of the feature point, $\mathbf{M}$ is the second moment matrix. For an affine transformation, scaling is different in each direction. Nonisotropic scaling has an influence on the spatial location, the scale and the shape of a local structure. Therefore, elliptical ACRs are more adaptive to an anisotropic local structure.

2.2 Feature Point Selection

Scale-space feature detectors, such as Harris-Laplace detector and Harris-Affine detector, was originally developed for matching and recognition. It extracts many feature points that densely cover the whole image. Hence, the local regions centered at feature points are overlapped with each other seriously. These local regions are not directly applicable to watermarking. Therefore, we present a selection criterion based on the graph theoretical clustering algorithm for scale-space feature points. This selection criterion adjusts the number, distribution, and scale of the features and removes those features that are vulnerable to watermark attacks. The framework of this selection criterion is illustrated as Fig. 1.

The scale of feature points derived from scale-space feature detectors is related to the scaling factor of the Gaussian kernel in scale space. Usually, features with small scales have a low repeatability, while features with large scales also have a low probability of being redetected. Moreover, using large-scale features means that local regions centered at these feature points will seriously overlap with each other, which will severely degrade the performance of the watermarked image. As a consequence, we only select features whose scale is in the middle-scale band $(a \leq s \leq b)$. Here, parameters a and b can be adjusted according to the different detectors.

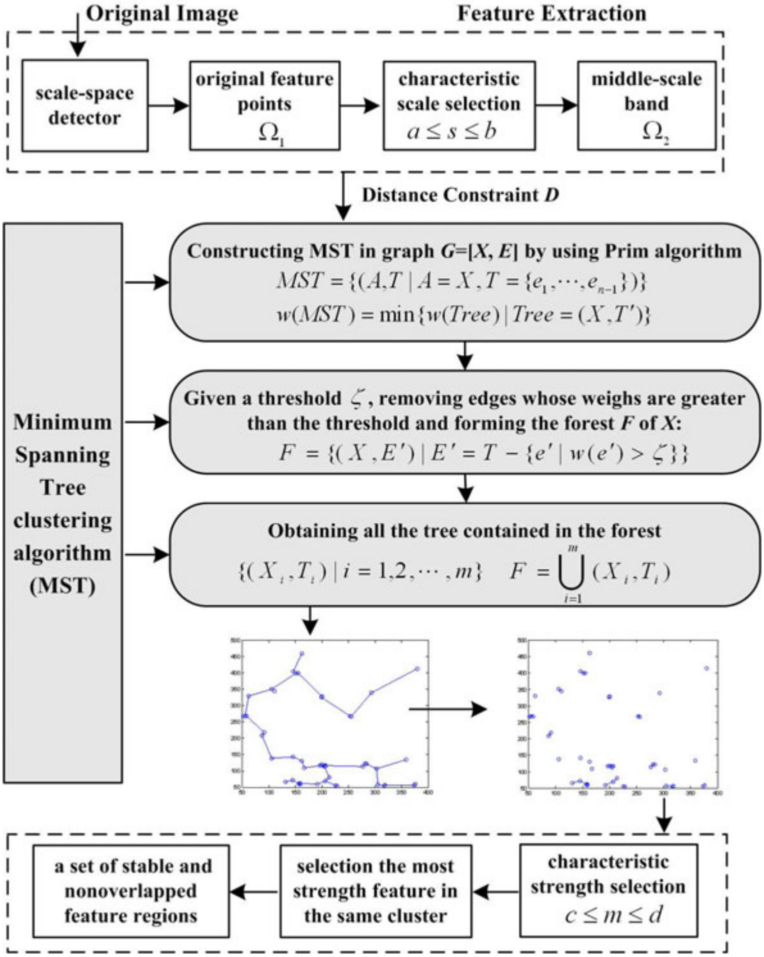

Original Image
Feature Extraction
scale-space detector
original feature points Ω_1
characteristic scale selection $a \le s \le b$
middle-scale band Ω_2
Distance Constraint D
Minimum Spanning Tree clustering algorithm (MST)
Constructing MST in graph G=[X, E] by using Prim algorithm
$MST = \{(A,T \mid A = X, T = \{e_1,\cdots,e_{n-1}\})\}$
$w(MST) = \min\{w(Tree) \mid Tree = (X,T')\}$
Given a threshold ζ, removing edges whose weighs are greater than the threshold and forming the forest F of X:
$F = \{(X,E') \mid E' = T - \{e' \mid w(e') > \zeta\}\}$
Obtaining all the tree contained in the forest
$\{(X_i,T_i) \mid i = 1,2,\cdots,m\}$ $F = \bigcup_{i=1}^{m}(X_i,T_i)$
a set of stable and nonoverlapped feature regions
selection the most strength feature in the same cluster
characteristic strength selection $c \le m \le d$

Fig. 1. Framework of the feature selection criterion based on the graph theoretical clustering algorithm

The distribution of features is also related to the performance of the watermarking system. That is to say, the distance between adjacent features must be determined carefully. If the distance between two features is small, their overlap will be large. On the contrary, if the distance between them is large, the number of local regions will be insufficient. Therefore, distance constraint is adopted to modulate the distribution of the features. We utilize *minimum spanning tree* (MST) clustering algorithm [27] to group these features according to the distance constraint D . In other words, features whose adjacent distance is less than D will be assigned into a cluster. With regard to the same cluster, features whose strength is the largest are used to form local regions. In our experiments, distant constraint D can be adaptively adjusted according to the different detector. At the same time, distant constraint D can be treated as a secret parameter to enhance the security of the watermarking system. That means the receiver will not be able to generate the same local regions if he/she does not know this parameter.

After the above steps, non-overlapped local regions are selected appropriately for watermarking. Fig. 2 and Fig. 3 show the procedures in selecting the ACRs and the LCRs for *Baboon*, *Lena*, and *Plane*, respectively. The original feature points extracted by scale-space feature detectors are illustrated in Fig. 2(a) and Fig. 3(a). While keeping the selected feature points in middle-scale band, as shown in Fig. 2(b) and Fig. 3(b), the final feature points are well chosen after MST-based clustering algorithm. The local regions appropriate for watermarking are shown in Fig. 2(c) and Fig. 3(c).

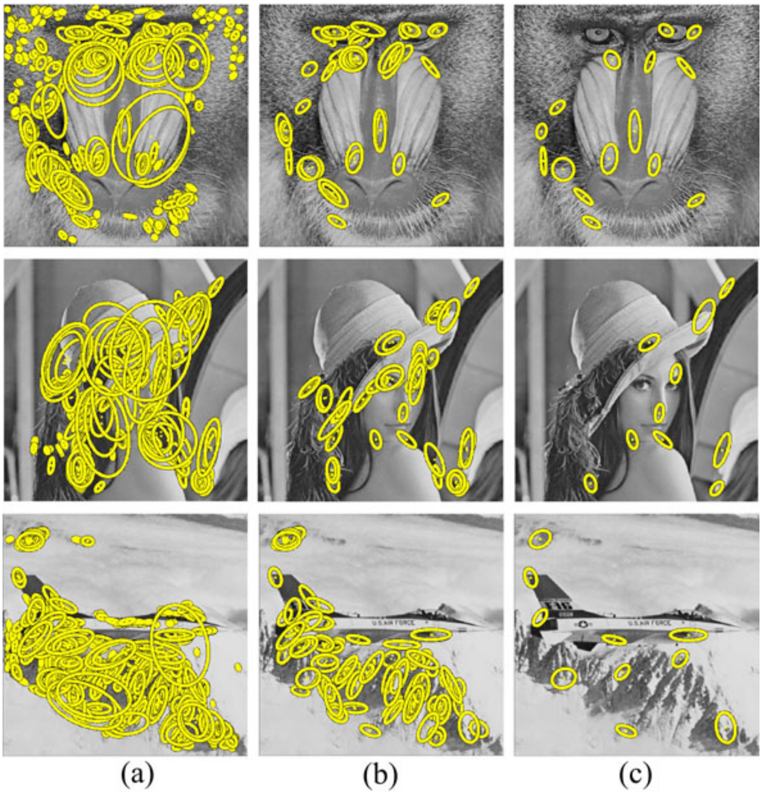


Fig. 2. ACRs selection for *Baboon*, *Lena*, and *Plane*, respectively: (a) original scale-space feature regions, (b) feature regions in middle-scale band, and (c) final selected feature regions for watermarking

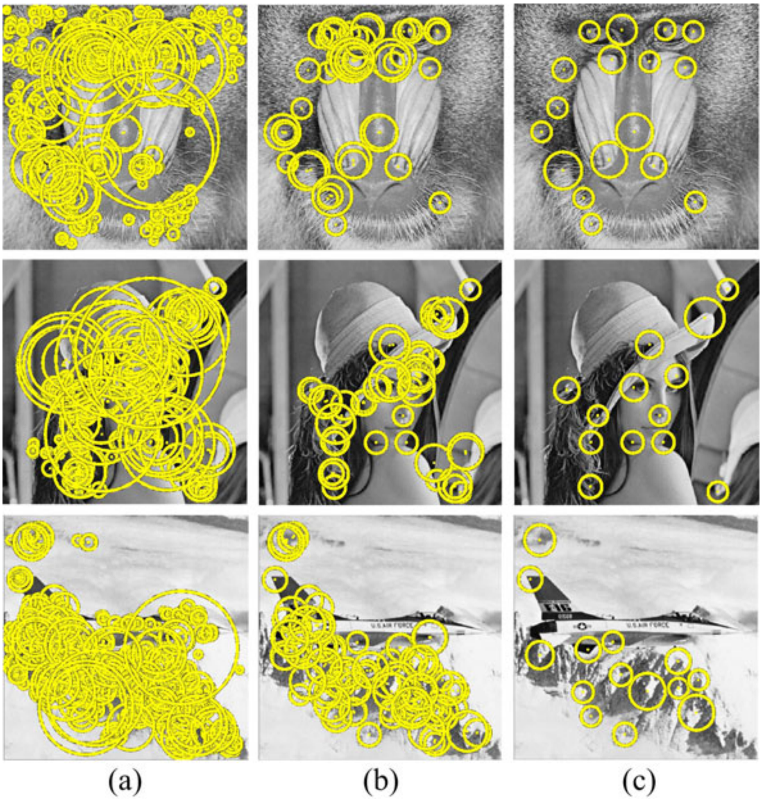


Fig. 3. LCRs selection for *Baboon*, *Lena*, and *Plane*, respectively: (a) original scale-space feature regions, (b) feature regions in middle-scale band, and (c) final selected feature regions for watermarking

3 Robust Image Watermarking Based on Local Features

3.1 Geometrically Invariant Image Watermarking Based on ACRs

Seo and Yoo [21] present a geometrically invariant image watermarking based on ACRs that provides a certain degree of robustness. However, as described previously, there are some main problems in [21]. First, the feature selection process in [21] does not completely solve the issue of the overlapping between the feature regions. In fact embedding watermark into a number of non-overlapped feature regions is the most important pre-requisite for designing a robust image watermarking. In view of the multi-characteristic of scale-space feature point, feature selection process combined with clustering algorithm should be a promising solution to this problem. Secondly, even though the feature regions in [21] have been called invariant regions, in principle they should be termed covariant regions since they deform covariantly with the transformation. The confusion probably arises from the fact that, even though the regions themselves are covariant, the normalized image pattern they covered is typically invariant [28]. So, an ACR directly used as the embedding unit inherently restricts the obtainable robustness against geometric distortions. As to that, combining ACR extraction with geometric invariant region construction will evidently improve the robustness of the watermarking system. Thirdly, 1-D watermark is affinely transformed into an elliptical pattern according to the shape of the elliptical region, which will dramatically reduce the resulting watermark energy in detection end and cause the watermark detection failure.

Take the above problems into consideration, we propose a new image watermarking scheme by incorporating the advantages of the Harris-Affine detector, the image normalization and the orientation alignment seamlessly. The Harris-Affine detector is adopted to extract ACRs. The graph theoretical clustering algorithm is then employed to select a set of separated ACRs for watermark embedding. In order to achieve affine invariance, each region is locally normalized by transforming an ellipse into a circle and rotated to align with its dominant gradient orientation. In watermark embedding, circular watermark pattern is embedded in the normalized patch. For the purpose of imperceptibility after watermarking, an image-dependent visual model is utilized to adjust the embedding strength [29].

3.1.1 Watermark Embedding

For watermark embedding, we first select a set of suitable ACRs according to the procedure mentioned in the Section 2, wherein the selected regions are possible for watermark embedding. Secondly, these regions are transformed into circular patches by normalization and dominant orientation alignment. Finally, the circular watermark pattern is embedded repeatedly in all normalized patches. This process is visualized by Fig. 4, in detail,

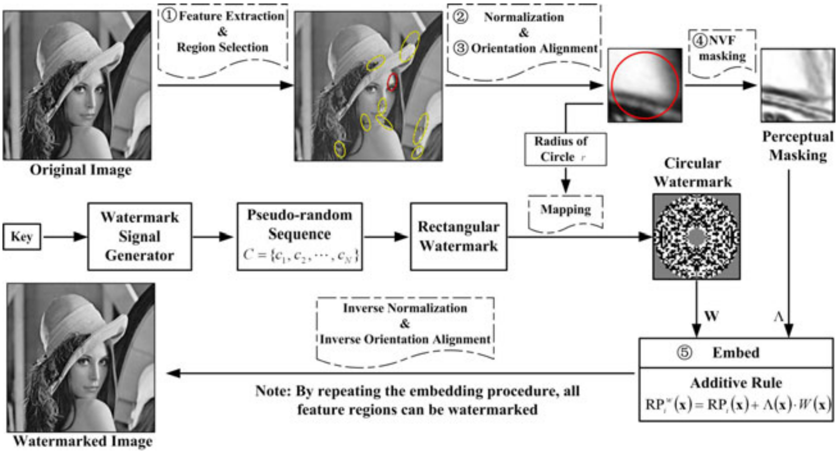

①Feature Extraction & Region Selection
Original Image
② Normalization & ③ Orientation Alignment
④NVF masking
Perceptual Masking
Radius of Circle r
Circular Watermark
Mapping
Key
Watermark Signal Generator
Pseudo-random Sequence $C = \{c_1, c_2, \cdots, c_N\}$
Rectangular Watermark
W
Λ
Inverse Normalization & Inverse Orientation Alignment
⑤ Embed
Additive Rule
$RP_i^w(x) = RP_i(x) + \Lambda(x) \cdot W(x)$
Note: By repeating the embedding procedure, all feature regions can be watermarked
Watermarked Image

Fig. 4. The watermark embedding procedure of the ACRs-based scheme

Step 1: The Harris-Affine detector is used to extract ACRs according to the procedure mentioned in the Section 2. In order to embed watermarks, a set of stable and non-overlapped ACRs are selected according to the scale range, repeatability, and distribution.

Step 2: Each selected ACR P_i is normalized to a circular patch NP_i with a fixed radius r . When all ACRs are normalized, corresponding regions are differed only by a simple rotation.

The size of normalized patch needs to be properly considered. It has been proven that if a large patch is warped into a small patch, which means that the warping process is a multiple-to-one pixel mapping, then one pixel in NP_i represents several pixels in P_i. Under this circumstances, a small normalized patch is beneficial for achieving robustness. In our study, the size of normalized patch is empirically found to be 39×39 $(r = 19)$ for achieving a tradeoff between imperceptibility and robustness.

Step 3: The watermark sequence $C = \{c_1, c_2, \cdots, c_N\}$ is generated by a secret key and then mapped into circular watermark pattern W with the radius of r.

Step 4: For each normalized patch, we can calculate its dominant gradient orientation and align the patch according to this orientation by rotating. Thus, the rotation invariance of the patch can be obtained. Let RP_i denote the rotated circular patches.

In [30], a window centered at these feature points is defined. The gradients of all pixels in a window are calculated by using the first order derivative. Then the histogram of gradients is computed and the peak of the histogram is assigned as the dominant orientation of the feature point. This orientation is usually robust against noise, small local distortions and some displacement of the feature point position.

The gradient of the pixel (x_0, y_0) in the image I is computed as follows

$$\nabla I(x_0, y_0) = [(\partial I/\partial x),\ (\partial I/\partial y)]|_{(x_0, y_0)}. \tag{16}$$

The magnitude of this gradient is $\sqrt{(\partial I/\partial x)^2 + (\partial I/\partial y)^2}$ and its orientation is given by $\tan^{-1}[(\partial I/\partial y)/(\partial I/\partial x)]$.

Step 5: To make the embedded watermark imperceptible, we adopt the following image-dependent visual model [31]

$$\Lambda = (1 - \mathrm{NVF}) \cdot \alpha + \mathrm{NVF} \cdot \beta, \tag{17}$$

where α and β are the watermark strength. For most real-world images, β is set to 3, and α can be adjusted to keep the *peak signal-to-noise ratio* (PSNR) higher than a certain value (in our case, α is set to 15). The *noise visibility function*(NVF) is calculated as follows:

$$\mathrm{NVF}(i, j) = \frac{1}{1 + \theta \cdot \sigma_x^2(i, j)},\ \theta = \frac{D}{\sigma_{x\ max}^2}, \tag{18}$$

where $\sigma_x^2(i, j)$ is the local variance of the neighboring pixels, $\sigma_{x\ max}^2$ is the maximum local variance of the image, and $D \in [50, 100]$ is an experimental constant.

Step 6: Thereafter, the watermarked circular patch is obtained by the additive rule, i.e.,

$$RP_i^w(\mathbf{x}) = RP_i(\mathbf{x}) + \Lambda(\mathbf{x}) \cdot W(\mathbf{x}). \tag{19}$$

Once the watermarked normalized patch RP_i^w is obtained, the inverse orientation alignment and the inverse normalization are used to yield a watermarked ACR. Although *direct inverse normalization* is intuitive, blocking effects caused by the one-to-multiple pixel mapping may degrade the imperceptibility [32]. To deal with this problem, the difference between RP_i and RP_i^w, which is caused by watermarking in the normalized domain, is inversely aligned and then inversely normalized to yield the difference P_i^{diff} in the spatial domain. Hence, the watermarked ACR in the spatial domain can be described as

$$p_i^w = p_i + p_i^{diff}. \tag{20}$$

Finally, by integrating all watermarked ACRs, the watermarked image can be obtained.

3.1.2 Watermark Detection

The watermark detection stage uses the same feature selection process as it in the watermark embedding procedure. Even though the watermarked image undergoes specific affine transformations, the ACRs with the locally largest repeatability can be conserved. The procedure for watermark detection is described in detail as below.

Step 1-3: These three steps are the same as *Steps* 1-3 in the watermark embedding procedure.

Step 4: Because the Wiener filter can separate the image components from the watermark components, we use it to estimate embedded watermark patterns. The estimated watermark is then converted into a sequence $V = \{v_1, v_2, \cdots, v_N\}$.The extracted watermark sequence is then compared with the original embedded watermark to decide whether a watermark exists in an ACR.

3.1.3 False-Positive Probability Analysis

For each ACR, the matching bits between C and T is calculated. If the matching bits is greater than predefined threshold T, it is said that a watermark is existed in a ACR. To determine the threshold, we consider the false-positive probability and the false-negative probability. Usually, it is difficult to analyze the false-negative probability because a wide variety of distortions exist in the procedure of watermark embedding and detection. Hence, it is usual to select the threshold T based on a fixed false-positive probability.

For an unwatermarked image, the extracted bits are treated as independent random variables with probability 0.5. According to Bernoulli trials, the false-positive probability of an ACR is

$$P_{fp} = \sum_{i=T}^{N} (0.5)^N \left(\frac{N!}{i!(N-i)!} \right), \tag{21}$$

where T is the predefined threshold, i is the number of the matching bits, and N is the length of the watermark sequence. Furthermore, an image is claimed to be watermarked if at least m regions are detected. Therefore, the false-positive probability of one image is

$$P_{fp-image} = \sum_{j=m}^{M} \binom{M}{j} (P_{fp})^j (1-P_{fp})^{M-j}, \tag{22}$$

where M is the total number of ACRs in an image.

We can plot $P_{fp-image}$ against various T values, as shown in Fig. 5 using Eq. (22) when the above parameters are set to $T = 160$, $M = 20$, $N = 256$, the false-positive probability of an image for $m = 1,2,3$ are 4.0×10^{-4}, 9.3×10^{-8}, and 1.2×10^{-11}, respectively.

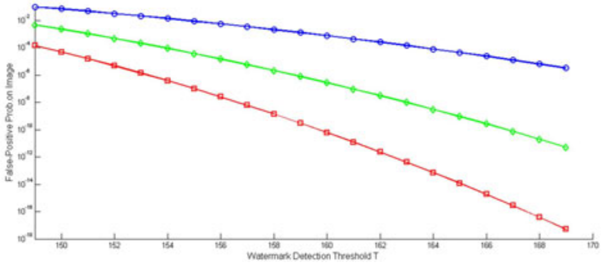


Fig. 5. False-positive probability P_{fp} versus watermark detection threshold T

3.2 Robust Image Watermarking Based on LTMs

The ACRs-based image watermarking scheme has improved the performance in term of robustness. However, it still has some problems: first, normalization and direct inverse normalization of feature regions will lead to blocking effects and degrade the imperceptibility of watermarked image; secondly, transforming watermark into a specific shape (e.g., circular or ellipse) will reduce the resulting watermark energy in detection end; thirdly, if the watermark is directly embedded in the spatial domain, the shift problem may cause the watermark extraction failure. These practical problems existed in the ACRs-based method restrict the robustness against geometric distortions as well as common image processing operations. To this end, we develop an image watermarking approach that has greater robustness against geometric distortions and common image processing operations simultaneously by incorporating the advantages of the moment-based method and the feature-point-based method [33].

In the proposed watermarking scheme, the Harris-Laplace detector is used to extract feature points. For each chosen feature point, a LCR is constructed that is invariant to rotation and scaling. The TMs are then employed to describe the global characteristics of the local invariant regions. Obviously, the extracted LTMs are independent to the slight change of pixels in this region. Here we select TMs rather than ZMs/PZMs based on the fact that TMs have not only insensitivity to noise but also better feature representation capability and reconstruction accuracy. The magnitudes of LTMs are modified through quantization index modulation, which can achieve the blind detection and improve the detection accuracy.

3.2.1 Tchebichef Polynomials and Tchebichef Moments

Unlike Zernike moments(ZMs)/Pseudo Zernike moments(PZMs) polynomials only defined inside the unit circle, discrete Tchebichef polynomials are directly defined in the image coordinate space and preserve the property of orthogonality in a moment set. TMs are hence more suitable for square digital image. The accuracy of image reconstruction with TMs is distinctly better than with ZMs/PZMs [34].

The discrete Tchebichef polynomials[35] are defined as

$$t_n(x) = n! \sum_{k=0}^{n} (-1)^{n-k} \binom{N-1-k}{n-k} \binom{n+k}{n} \binom{x}{k}. \tag{23}$$

and satisfies the following orthogonal condition

$$\sum_{x=0}^{N-1} t_p(x) t_q(x) = \rho(n,N)\delta_{pq} \quad 0 \le p,q \le N-1, \tag{24}$$

where $\rho(n,N) = (2n)! \binom{N+n}{2n+1}$, $n = 0,1,\ldots,N-1$.

For a digital image $f(x,y)$ with size $N \times N$, the $(p+q)$th order scaled Tchebichef moments are given by

$$T_{pq} = \frac{1}{\widetilde{\rho}(p,N)\widetilde{\rho}(q,N)} \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} \widetilde{t}_p(x)\widetilde{t}_q(y) f(x,y) \quad p,q = 0,1,2,\ldots,N-1, \tag{25}$$

where the scaled Tchebichef polynomials $\widetilde{t}_n(x) = t_n(x)/\beta(n,N)$ and $\widetilde{\rho}(n,N) = \rho(n,N)/\beta(n,N)^2$. Here, $\beta(n,N)$ is a suitable constant which is independent of x.

The inverse moment transformation can be defined as

$$f(x,y) = \sum_{p=0}^{N-1}\sum_{q=0}^{N-1} T_{pq}\widetilde{t}_p(x)\widetilde{t}_q(y) \quad x,y = 0,1,\ldots,N-1. \tag{26}$$

However, the TMs tend to exhibit numerical instabilities and propagation of numerical errors with the moment order increasing. Then, the quality of image reconstruction will be affected severely. To solve this problem, the constant $\beta(n,N)$ should be modified as

$$\beta(n,N) = \sqrt{\frac{N(N^2-1)(N^2-2^2)\cdots(N^2-n^2)}{2n+1}}. \tag{27}$$

In this case, $\widetilde{\rho}(n,N) = 1.0$ [36].

Fig. 6 shows the relationship between the maximum moment orders of TMs and reconstruction errors. We use the following measure to evaluate the reconstruction error

$$\varepsilon = \sqrt{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left\{f(i,j) - \widetilde{f}(i,j)\right\}^2}, \tag{28}$$

where $f(x,y)$ is an input gray-level image and $\widetilde{f}(x,y)$ is the reconstructed version.

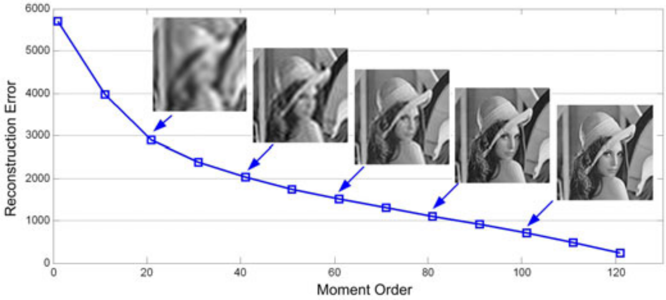


Fig. 6. Reconstruction error with respect to the maximum moment order for *Lena* image (128×128)

3.2.2 Watermark Embedding

The process of watermark embedding is shown in Fig. 7, and the detailed algorithm is given as follows.

Step 1: A set of feature points are extracted by the Harris-Laplace detector according to the proposed selection criterion, as described in Section 2.

Step 2: For each LCR, we need to calculate the gradient direction of all pixels within it by using first order derivative. Then the histogram of the gradient is computed and the peak of the histogram is assigned as the dominant orientation of the feature point. To achieve the rotation invariance, the dominant orientation of the LCR is aligned by rotating.

Ideally, the watermarks are only embedded in the LCRs, as illustrated in Fig. 8(a). But in implementation, as shown in Fig. 8(b) and (c), we actually first obtain original square patches, and the ideal LCRs can then be generated after padding the original square patches with zeros. Considering the fact that discrete TMs polynomials define their moments directly on image coordinate space and the regions for watermarking must preserve the covariant properties, the inscribed square patches of the LCRs, as shown in Fig. 8(d), are extracted as the final embedding regions. Fig. 8 illustrates the formation of the embedding regions.

Step 3: To further enhance the robustness to common image processing operations, such as additive noise and JPEG compression, LTMs are calculated in each inscribed square patch.

Step 4: The set of LTMs whose order are less than or equal to ω_{max} is denoted as $\{\Omega = T_{pq},\ p+q < \omega_{max},\ p,q \neq 0\}$. Watermarks should be embedded in Ω to

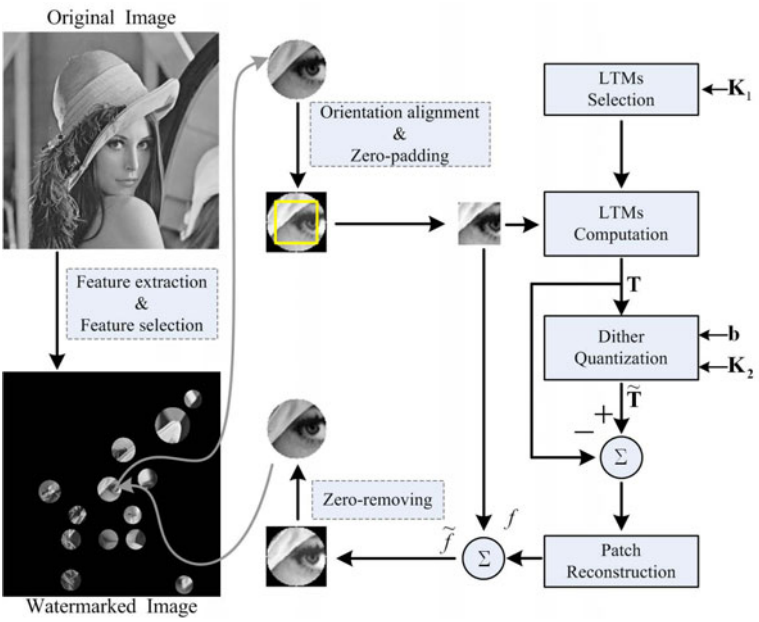


Fig. 7. The watermark embedding procedure of the LTMs-based scheme

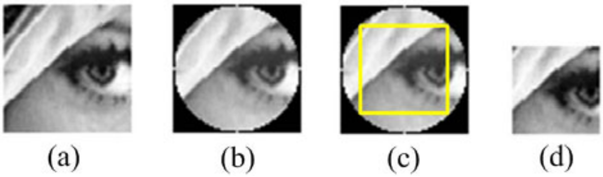


Fig. 8. The formation of the embedding regions: (a) the ideal embedding region, (b) the original square patch, (c) the square patch after orientation alignment and zero-padding, and (d) the inscribed patch for watermarking

balance invisibility and robustness. Dither modulation [37] is adopted for the modification of LTMs to conduct watermark embedding.

Let watermark sequence be $\mathbf{b} = b_1, \ldots, b_L$, and $b_i \in \{0, 1\}$. We first use a secret key $\mathbf{K}_1$ to randomly select L LTMs from Ω to form a TM vector $T = (T_{p_1 q_1}, \ldots, T_{p_L q_L})$. To embed a watermark bit b_i, the magnitude of $T_{p_i q_i}$ is quantized, producing a new vector $\widetilde{T} = (\widetilde{T}_{p_1 q_1}, \ldots, \widetilde{T}_{p_L q_L})$ [38], whose magnitudes satisfy

$$|\widetilde{T}_{p_i q_i}| = \left[\frac{|T_{p_i q_i}| - d_i(b_i)}{\Delta}\right]\Delta + d_i(b_i), \quad i = 1, \ldots L, \tag{29}$$

where $[\cdot]$ is rounding operation, Δ is quantization step and $d_i(\cdot)$ is the dither function for the ith quantizer satisfying $d_i(1) = \Delta/2 + d_i(0)$. The dither vector $(d_1(0), \dots, d_L(0))$, which follows uniform distribution over $[0, \Delta]$, is generated by secret key $\mathbf{K}_2$. Thus, the modified LTMs can be expressed as

$$\widetilde{T}_{p_i q_i} = \frac{|\widetilde{T}_{p_i q_i}|}{|T_{p_i q_i}|} T_{p_i q_i}, \quad i = 1, \dots, L. \tag{30}$$

Step 5: For each watermarked inscribed square patch, it is composed of two parts. One is the reconstructed patch by the LTMs not selected, which is

$$f_{rest}(x,y) = f(x,y) - f_{\mathbf{T}}(x,y), \tag{31}$$

where $f(x,y)$ is the original patch and $f_{\mathbf{T}}(x,y)$ is the reconstructed patch by the selected LTMs before they are changed.

The other is the patch $f_{\widetilde{\mathbf{T}}}(x,y)$, which is reconstructed by those modified LTMs. Consequently, we can obtain a watermarked inscribed square patch by combining these two parts

$$\widetilde{f}(x,y) = f_{rest}(x,y) + f_{\widetilde{\mathbf{T}}}(x,y). \tag{32}$$

After all of the watermarked inscribed square patches replace the original ones, the watermarked image can be obtained.

3.2.3 Watermark Detection

The procedure for watermark detection is illustrated in Fig. 9. In detail,

Step 1-3: These first three steps are identical to *Steps* 1-3 in the watermark embedding procedure.

Step 4: With the same secret key $\mathbf{K}_1$, L relevant LTMs can be chosen, which is denoted by $T' = (T'_{p_1 q_1}, \dots, T'_{p_L q_L})$.

First, with the same key $\mathbf{K}_2$, the same two dither vector $(d_1(0), \dots, d_L(0))$ and $(d_1(1), \dots, d_L(1))$ are regenerated. As the Eq. (29), the magnitude of each $T'_{p_i q_i}$ is then quantized with the above two corresponding dithers, respectively,

$$|T'_{p_i q_i}|_j = \left[\frac{|T'_{p_i q_i}| - d_i(j)}{\Delta} \right] \Delta + d_i(j), \tag{33}$$

where $i = 1, \dots, L$, $j \in 0, 1$ and $[\cdot]$ is the rounding operation.

Finally, by comparing the distances between $|T'_{p_i q_i}|$ with its two quantized versions, we can estimate the watermark bit embedded in $|T_{p_i q_i}|$

$$b'_i = \arg \min_{j \in \{0,1\}} \left(|T'_{p_i q_i}|_j - |T'_{p_i q_i}| \right)^2. \tag{34}$$

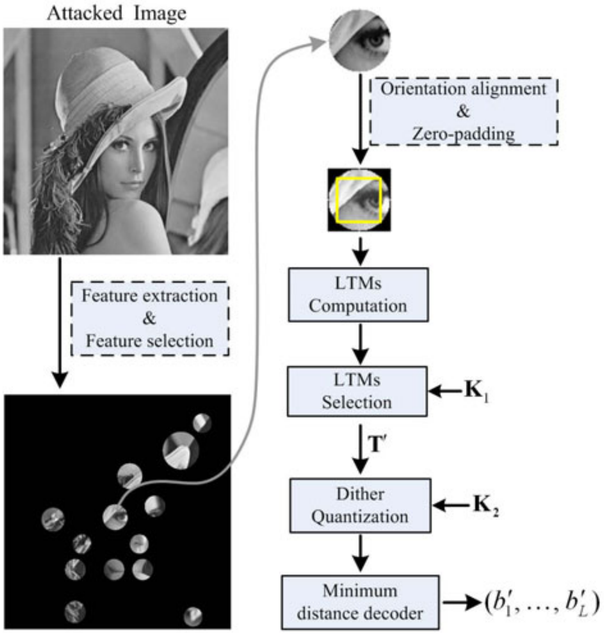


Fig. 9. The watermark detection procedure of the LCRs-based scheme

3.2.4 False-Positive Probability Analysis

For an un-watermarked image, the extraction of watermark bit, analogy with coin-tossing, can be regarded as Bernoulli trials. The extracted bits can be treated as stochastic variable with probability $p = 0.5$. A LCR is claimed to be watermarked if the number of the matching bits is larger than a threshold. Thus, the false-positive error probability of the LCR can then be described as

$$P_{fp} = \sum_{r=T}^{L} (0.5)^L \left(\frac{L!}{r!(L-r)!} \right), \tag{35}$$

where T is a predefined threshold, r is the number of the matching bits and L is the length of watermark bits.

When L is large, the probability density for a binary stochastic variable is known to approximate a Gaussian probability variable with an average $m = Lp$ and a variance $\hat{\sigma}^2 = Lp(1-p)$ [39]. Therefore, Eq. (35) can be rewritten as

$$P_{fp} = \begin{cases} 1 - \frac{1}{2\pi} \frac{e^{-T'^2/2}}{(a-1)T' + a\sqrt{T'^2 + b}} & (T < pL) \\ 0.5 & (T = pL) \\ \frac{1}{2\pi} \frac{e^{-T'^2/2}}{(1-a)T' + a\sqrt{T'^2 + b}} & (T > pL) \end{cases}, \tag{36}$$

where $T' = (T-m)/\hat{\sigma}$, $a = 1/\pi$, and $b = 2\pi$.

For instance, when $L = 224$ and $T = 140$ (140 bits out of 224 match), $P_{fp} = 9.13e-5$, based on Eq. (36). Therefore, when we assume watermark is embedded in a disk, the probability that this assumption is wrong is $9.13e-5$.

4 Experimental Results and Analysis

In this section, the two proposed algorithms are evaluated in two aspects: imperceptibility and robustness. We collect a set of standard test images including images *Baboon*, *Lena*, *Peppers* and *Plane* to build a dataset. The watermark generated by secret key is respectively 256-bit ($N = 256$) and 224-bit ($L = 224$) in the ACRs-based scheme and the LTMs-based scheme.

4.1 Imperceptibility Test

In the ACRs-based scheme, the original images and the watermarked images are shown in Fig. 10(a) and Fig. 10(b). The difference between the original images and the watermarked versions are magnified by a factor 100 and shown in Fig. 10(c). Based on Eq. (17), it is clear that for a region with high local variance (NVF $\rightarrow 0$), a strong watermark is embedded, while for a homogenous region (NVF $\rightarrow 1$), the watermark strength is almost zero. Through NVF, it can modulate the watermark strength adaptively based on local image characteristics and in the meanwhile make the embedded watermark hardly perceptible. In addition, the overall PSNR values of all images in our testing set between the original and watermarking versions are greater than 40 dB.

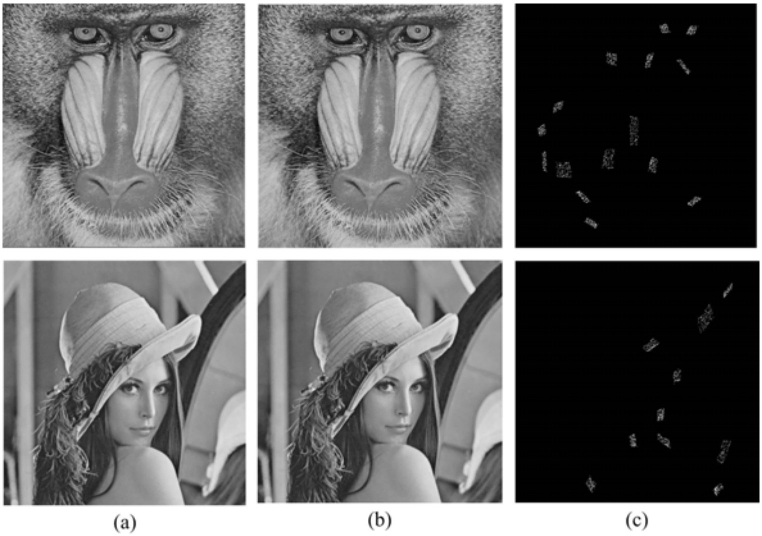


Fig. 10. (a) Original image, (b) watermarked image, and (c) difference image for *Baboon* and *Lena*

In the LTMs-based scheme, the PSNR value between the original image and the watermarked version depends on the following two main factors. On the one hand, given a fixed watermark length, the quantization step Δ in the dither modulation has an impact on the PSNR. A larger value of Δ will increase the watermark strength, but decrease the PSNR, and vice versa. On the other hand, given a fixed Δ or watermark strength, the more watermark bits are embedded, the lower PSNR value is, and vice versa. The relationship between the average PSNR and these two factors are shown in Fig. 11.

In our experiments, we set the quantization step $\Delta = 18$ and the watermark length $L = 224$. Thus, the overall resulting PSNR is greater than 50dB. Fig. 12 shows the original images, the watermarked versions and the residuals between the original and watermarked versions after magnified by a factor 100. As shown in Fig. 12(a) and (b), it is clear that the embedded watermarks are perceptually invisible.

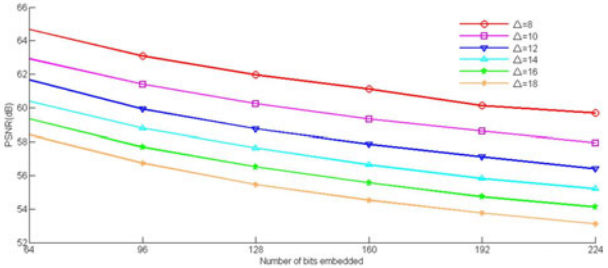


Fig. 11. The relationship between the average PSNR, quantization step Δ and the number of the embedded watermark bits L

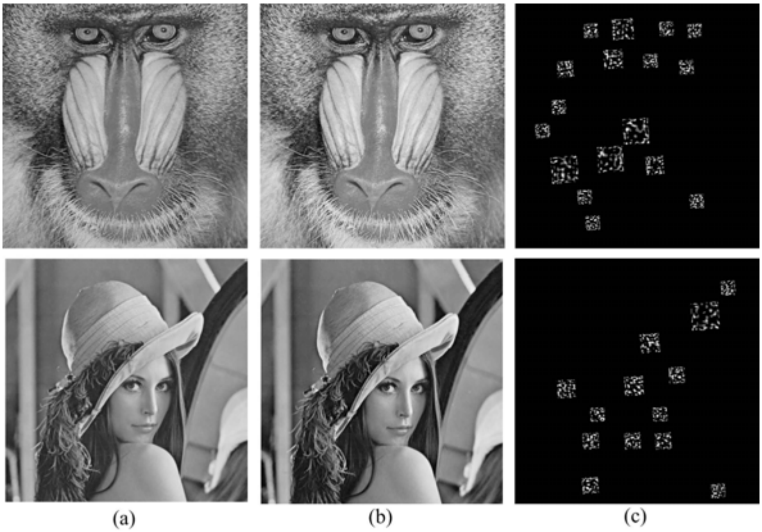


Fig. 12. (a) Original image, (b) watermarked image, and (c) difference image for *Baboon* and *Lena*

4.2 Robustness Test

Apart from the imperceptibility test, two experiments are conducted to measure (1) the performance of the watermark synchronization based on the feature point detectors; (2) the robustness of the proposed watermarking schemes.

4.2.1 Performance of the Feature Detector

In our scheme, only a set of feature points chosen by our feature selection criterion are used to construct local feature regions. The stability of these feature points is important for robustness. To measure the stability, we first extract feature points from the original images and the attacked images, and then compute the repeatability ratio between the number of point-to-point correspondences and the number of points detected in original images.

By utilizing Stirmark 4.0 [40], we apply various attacks to four benchmark images *Baboon*, *Boat*, *Lena*, and *Peppers*. These attacks include median filtering (3×3), Gaussian filtering (3×3), JPEG compression (QF 40 and 50), Cropping (5% and 10% off), Rotation (5° and 10°), Scaling ($\times0.9$, $\times1.1$), and Rotation+Cropping (1° and 5°).

Fig. 13 and Fig. 14 illustrate the results for the ACRs-based scheme and the LTMs-based scheme, respectively. *Corresponding ratio* refers to the ratio of the number of corresponding feature points between the original images and the attacked versions to the number of the feature points extracted from the original images. As shown in Fig. 13(a) and (b), the corresponding ratio is more than 62% on average for common image processing operations, and nearly 50% feature points can be redetected in geometric distortions. As may be seen from

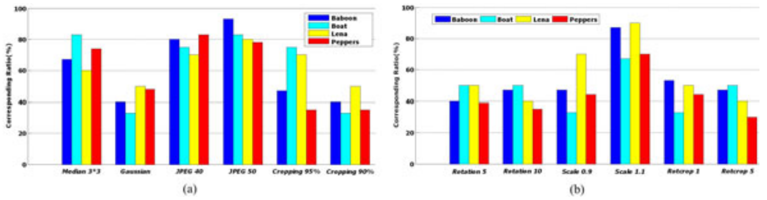


Fig. 13. *Corresponding ratio* in the ACRs-based method: (a) common image processing operations and (b) geometric distortions for *Baboon*, *Boat*, *Lena* and *Peppers*

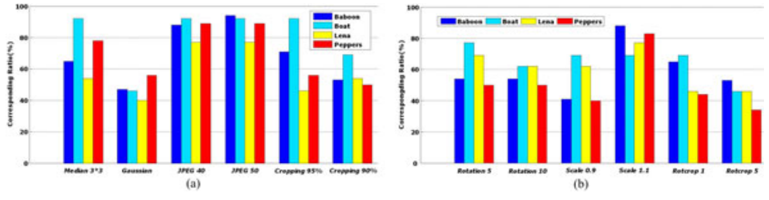


Fig. 14. *Corresponding ratio* in the LTMs-based method: (a) common image processing operations and (b) geometric distortions for *Baboon*, *Boat*, *Lena* and *Peppers*

Fig. 14(a) and (b), the corresponding ratio is nearly 70% on average for common image processing operations, and 60% feature points can be redetected under geometric distortions. Therefore, these selected feature points are stable and useful for robust watermarking against common image processing operations and geometric distortions.

4.2.2 Robustness of the Watermarking Schemes

We use Stirmark 4.0 to evaluate the robustness of the proposed schemes. Table 1 and Table 2 present the detection results of the ACRs-based watermarking scheme in comparison with two representative schemes [20] and [21] under common image processing operations and geometric distortions respectively. Table 3 and Table 4 is the detection results of the LTMs-based watermarking scheme in comparison with the two representative schemes [20] and [21] under the same attacks. The values in main table unites indicate the ratio of the number of regions where watermarks are successfully detected from attacked images to the number of original watermarked regions.

For most of attacks, the proposed watermarking schemes can detect the embedded watermark from a considerable number of feature regions and the ownership can be proven with high confidence. The scale-space feature point detector can provide a set of distinctive and localized feature regions which are covariant to significant affine transformation. Moreover, local feature regions are constructed, which are geometrically invariant. Consequently, the proposed watermarking schemes perform well in common image processing operations, geometric distortions, and even the combined complex attacks.

As shown in Tables 1, 2, 3, and 4, the newly proposed schemes outperform the conventional schemes [20] and [21]. Among these two existing schemes, scheme [21] shows the worst performance. According to the detection results of [20], this scheme can withstand many common image processing operations, e.g., Gaussian

Table 1. Experimental results of the ACRs-based scheme under common image processing operations

Attack Type	*Baboon*			*Lena*			*Peppers*		
	ACRs-based	Ref. [20]	Ref. [21]	ACRs-based	Ref. [20]	Ref. [21]	ACRs-based	Ref. [20]	Ref. [21]
Median filter(3×3)	6/15	2/11	0/13	6/10	1/8	2/11	17/23	1/4	2/13
Gaussian filtering(3×3)	4/15	7/11	1/13	5/10	3/8	1/11	9/23	1/4	2/13
Additive uniform noise(s=0.20)	9/15	5/11	0/13	6/10	1/8	0/11	12/23	1/4	0/13
JPEG 70	10/15	8/11	2/13	4/10	5/8	3/11	14/23	3/4	2/13
JPEG 50	8/15	6/11	0/13	6/10	4/8	1/11	14/23	2/4	1/13
JPEG 30	8/15	4/11	0/13	5/10	2/8	1/11	10/23	0/4	1/13
Median filter(3×3)+JPEG 90	5/15	1/11	0/13	3/10	1/8	2/11	16/23	1/4	2/13
Gaussian filtering(3×3)+JPEG 90	4/15	7/11	1/13	4/10	3/8	1/11	5/23	1/4	2/13

Table 2. Experimental results of the ACRs-based scheme under geometric distortions

Attack Type	*Baboon*			*Lena*			*Peppers*		
	ACRs-based	Ref. [20]	Ref. [21]	ACRs-based	Ref. [20]	Ref. [21]	ACRs-based	Ref. [20]	Ref. [21]
Cropping (5%off)	7/15	2/11	3/13	7/10	2/8	5/11	8/23	2/4	4/13
Cropping (10%off)	6/15	1/11	1/13	5/10	2/8	4/11	6/23	2/4	1/13
Scaling (90%)	5/15	1/11	1/13	5/10	1/8	3/11	8/23	0/4	2/13
Scaling (150%)	7/15	0/11	3/13	6/10	0/8	3/11	13/23	0/4	4/13
Aspect ratio change(0.7, 0.9)	2/15	0/11	0/13	4/10	0/8	1/11	6/23	0/4	0/13
Shearing(1%)	6/15	4/11	0/13	4/10	2/8	2/11	7/23	1/4	0/13
Removed 5 row & 17 column	6/15	2/11	2/13	5/10	1/8	3/11	11/23	0/4	0/13
Rotation 5°	5/15	2/11	0/13	4/10	2/8	3/11	9/23	0/4	2/13
Rotation 30°	4/15	0/11	0/13	4/10	0/8	0/11	7/23	0/4	0/13
Random bend	7/15	3/11	0/13	5/10	2/8	1/11	8/23	0/4	0/13
Cropping 5%+JPEG70	6/15	2/11	1/13	6/10	1/8	3/11	7/23	2/4	2/13
Rotation 5°+Cropping+JPEG70	5/15	0/11	0/13	5/10	0/8	2/11	7/23	0/4	0/13
Line Transform (1.007)	4/15	1/11	0/13	6/10	2/8	1/11	7/23	1/4	0/13

Table 3. Experimental results of the LTMs-based scheme under common image processing operations

Attack Type	*Baboon*			*Lena*			*Peppers*		
	LTMs-based	Ref. [20]	Ref. [21]	LTMs-based	Ref. [20]	Ref. [21]	LTMs-based	Ref. [20]	Ref. [21]
Median filter(3×3)	11/17	2/11	0/13	7/13	1/8	2/11	14/18	1/4	2/13
Gaussian filtering(3×3)	8/17	7/11	1/13	5/13	3/8	1/11	9/18	1/4	2/13
Additive uniform noise(s=0.20)	12/17	5/11	0/13	9/13	1/8	0/11	11/18	1/4	0/13
JPEG 70	15/17	8/11	2/13	9/13	5/8	3/11	16/18	3/4	2/13
JPEG 50	13/17	6/11	0/13	8/13	4/8	1/11	16/18	2/4	1/13
JPEG 30	9/17	4/11	0/13	8/13	2/8	1/11	12/18	0/4	1/13
Median filter(3×3)+JPEG 90	11/17	1/11	0/13	7/13	1/8	2/11	14/18	1/4	2/13
Gaussian filtering(3×3)+JPEG 90	8/17	7/11	1/13	5/13	3/8	1/11	9/18	1/4	2/13

filtering, JPEG compression, and noise contaminating, but fails to rotation, scaling, and particularly nonisotropic scaling. In term of the overall watermark detection rate under geometric distortions and common image processing operations, the LTMs-based algorithm is obviously better than the ACRs-based algorithm.

Table 4. Experimental results of the LTMs-based scheme under geometric distortions

Attack Type	*Baboon*			*Lena*			*Peppers*		
	LTMs-based	Ref. [20]	Ref. [21]	LTMs-based	Ref. [20]	Ref. [21]	LTMs-based	Ref. [20]	Ref. [21]
Cropping (5%off)	10/17	2/11	3/13	6/13	2/8	5/11	9/18	2/4	5/13
Cropping (10%off)	9/17	1/11	1/13	6/13	2/8	4/11	7/18	2/4	4/13
Scaling (90%)	7/17	1/11	1/13	8/13	1/8	3/11	7/18	0/4	2/13
Scaling (150%)	9/17	0/11	3/13	6/13	0/8	3/11	9/18	0/4	4/13
Aspect ratio change(0.7,0.9)	3/17	0/11	0/13	5/13	0/8	1/11	7/18	0/4	0/13
Shearing(1%)	8/17	4/11	0/13	7/13	2/8	2/11	8/18	1/4	0/13
Removed 5 row & 17 column	6/17	2/11	2/13	7/13	1/8	3/11	7/18	0/4	0/13
Rotation 5°	5/17	2/11	0/13	9/13	2/8	3/11	6/18	0/4	2/13
Rotation 30°	5/17	0/11	0/13	8/13	0/8	0/11	4/18	0/4	0/13
Random bend	6/17	3/11	0/13	5/13	2/8	1/11	8/18	0/4	0/13
Cropping 5%+JPEG70	8/17	2/11	1/13	4/13	1/8	3/11	6/18	2/4	2/13
Rotation 5°+Cropping+JPEG70	4/17	0/11	0/13	4/13	0/8	2/11	6/18	0/4	0/13
Line Transform (1.007)	9/17	1/11	0/13	6/13	2/8	1/11	6/18	1/4	0/13

5 Conclusion

This chapter mainly proposes and studies two robust image watermarking algorithms by synchronizing watermarking with the invariant feature regions of the scale-space representation of an image. In the ACRs-based method, Harris-Affine detector is adopted to extract feature points, and watermark embedding and detection are conducted in the ACRs which is invariant to geometric distortions after image normalization and direction alignment. In the LTMs-based method, Harris-Laplace detector is used to detect feature points and construct LCRs, and the magnitudes of LTMs within LCRs are applied for referencing the watermark. Experimental results show these two proposed schemes perform well under various geometric distortions and common image processing operations and outperformed some representative schemes in terms of robustness. With respect to the two developed algorithms, the LTMs-based method outperforms the ACRs-based method in performance on average.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (60771068, 60702061, 60832005), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) in China and the National Laboratory of Automatic Target Recognition, Shenzhen University, China, the Program for Changjiang Scholars and innovative Research Team in University of China (IRT0645), and 100 Talents Program of The Chinese Academy of Sciences.

References

1. Lichtenauer, J., Setyawan, I., Kalker, T., Lagendijk, R.: Exhaustive geometrical search and false positive watermark detection probability. In: Proc. SPIE-Security and Watermarking of Multimedia Contents V, vol. 5020, pp. 203–214 (2003)
2. Miller, M.L., Bloom, J.A.: Computing the probability of false watermark detection. In: Proc. Third Int. Workshop on Information Hiding, pp. 146–158 (1999)
3. O'Ruanaidh, J.J.K., Pun, T.: Rotation, scale and translation invariant digital image watermarking. Signal Processing 66(3), 303–317 (1998)
4. Lin, C.Y., Wu, M., Bloom, J.A., et al.: Rotation, scale, and translation-resilient public watermarking for images. In: Proc. SPIE-Security and Watermarking of Multimedia Contents II, vol. 3971, pp. 90–98 (2000)
5. Zheng, D., Zhao, J., Saddik, A.: RST-invariant digital image watermarking based on log-polar mapping and phase correlation. IEEE Trans. Circuits Syst. Video Technol. 13(8), 753–765 (2003)
6. Pereira, S., Pun, T.: Robust template matching for affine resistant image watermarks. IEEE Trans. Image Process. 9(6), 1123–1129 (2000)
7. Pereira, S., Pun, T.: An iterative template matching algorithm using the chirp-Z transform for digital image watermarking. Pattern Recognit. 33(99), 173–175 (2000)
8. Kutter, M.: Watermarking resisting to translation, rotation and scaling. In: Proc. SPIE-Multimedia Systems and Applications, vol. 3528, pp. 423–431 (1999)
9. Delannay, D., Macq, B.: Generalizaed 2-D cyclic patterns for secret watermark generation. In: Proc. IEEE Int. Conf. Image Process, pp. 77–79 (2000)
10. Dugelay, J.-L., Roche, S., Rey, C., et al.: Still-image watermarking robust to local geometric distortions. IEEE Trans. Image Process. 15(9), 2831–2842 (2006)
11. Xiang, S., Kim, H.J., Huang, J.: Invariant image watermarking based on statistical features in the low-frequency domain. IEEE Trans. Circuits Syst. Video Technol. 18(6), 777–790 (2008)
12. Kutter, M., Bhattacharjee, S.K., Ebrahimi, T.: Towards second generation watermarking schemes. In: Proc. IEEE Int. Conf. Image Process, pp. 320–323 (1999)
13. Alghoniemy, M., Tewfik, A.: Geometric distortion correction through image normalization. In: Proc. IEEE Int. Conf. Multimedia Expo., vol. 3, pp. 1291–1294 (2000)
14. Kim, H.S., Lee, H.K.: Invariant image watermark using Zernike moments. IEEE Trans. Circuits Syst. Video Technol. 13(8), 766–775 (2003)
15. Alghoniemy, M., Tewfik, A.H.: Geometric invariance in image watermarking. IEEE Trans. Image Process. 13(2), 145–153 (2004)
16. Dong, P., Brankov, J.G., Galatsanos, N.P., et al.: Affine transformation resistant watermarking based on image normalization. IEEE Trans. Image Process. 14(12), 2140–2150 (2005)
17. Nikolaidis, A., Pitas, I.: Region-based image watermarking. IEEE Trans. Image Process. 13(2), 145–153 (2004)
18. Lee, H.-Y., Kim, H., Lee, H.-K.: Robust image watermarking using local invariant features. Optical Engineering 45(3), 1–9 (2006)
19. Bas, P., Chassery, J.-M., Macq, B.: Geometrically invariant watermarking using feature points. IEEE Trans. Image Process. 11(9), 1014–1028 (2002)
20. Tang, C.W., Hang, H.M.: A feature-based robust digital image watermarking scheme. IEEE Trans. Signal Process. 51(4), 950–959 (2003)
21. Seo, J.S., Yoo, C.D.: Image watermarking based on invariant regions of scale-space representation. IEEE Trans. Signal Process. 54(4), 1537–1543 (2006)

22. Wang, X., Wu, J., Niu, P.: A new digital image watermarking algorithm resilient to desynchronization attacks. IEEE Trans. Inf. Forens. Security 2(4), 655–663 (2007)
23. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Proc. IEEE Int. Conf. Computer Vision, vol. 1, pp. 525–531 (2001)
24. Lindeberg, T., Garding, J.: Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. Image Vis. Comput. 15, 415–434 (1997)
25. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 128–142. Springer, Heidelberg (2002)
26. Baumberg, A.: Reliable feature matching across widely separated views. In: Proc. IEEE Int. Conf. Computer Vision Pattern Recognition, vol. 1, pp. 774–781 (2000)
27. Gower, J.C., Ross, G.J.S.: Minimum spanning trees and single linkage cluster analysis. Applied Statistics 18(1), 54–64 (1969)
28. Mikolajczyk, K., Schmid, C.: A comparison of affine region detectors. Int. J. Computer Vision 65(1-2), 43–72 (2005)
29. Cheng, D., Gao, X., Li, X., Tao, D.: Geometrically invariant watermarking using affine covariant regions. In: Proc. IEEE Int. Conf. Image Processing, pp. 413–416 (2008)
30. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Computer Vision 60(2), 91–110 (2004)
31. Voloshynovskiy, S., Herrigel, A., Baumgrtner, N., Pun, T.: A Stochastic Approach to Content Adaptive Digital Image Watermarking. In: Proc. Third Int. Workshop Information Hiding, pp. 212–236 (1999)
32. Lu, C.S., Sun, S.W., Hsu, C.Y., Chang, P.C.: Media hash-dependent image watermarking resilient against both geometric attacks and estimation attacks based on false positive-oriented detection. IEEE Trans. Multimedia 84, 668–685 (2006)
33. Cheng, D., Gao, X., Li, X., Tao, D.: A local Tchebichef moments-based robust image watermarking. Signal Processing 89(8), 1531–1539 (2009)
34. Shu, H.Z., Luo, L.M., Coatrieux, J.L.: Moment-based approaches in image-Part I: Basic feature. IEEE Eng. Med. Biol. Mag. 26(5), 70–75 (2007)
35. Mukundan, R., Ong, S.H., Lee, P.A.: Image analysis by Tchebichef moments. IEEE Trans. Image Process. 10(9), 1357–1364 (2001)
36. Mukundan, R.: Some computational aspects of discrete orthonormal moments. IEEE Trans. Image Process. 13(8), 1055–1059 (2004)
37. Chen, B., Wornall, G.W.: Quantization index modulation methods: a class of provably good methods for digital watermarking and information hiding. IEEE Trans. Inf. Theory 47(4), 1423–1443 (2001)
38. Xin, Y.Q., Liao, S., Pawlak, M.: Circularly orthogonal moments for geometrically robust image watermarking. Pattern Recognition 40(12), 3740–3752 (2007)
39. Choi, Y., Aizawa, K.: Watermark detection based on error probability and its applications to video watermarking. Electron. Commun. Jpn. 87(6), 66–76 (2004)
40. Petitcolas, F.A.P.: Watermarking schemes evaluation. IEEE Signal Process. Mag. 17(5), 58–64 (2004)

Image Authentication Using Active Watermarking and Passive Forensics Techniques

Xi Zhao, Philip Bateman, and Anthony T.S. Ho

University of Surrey, UK
{x.zhao,p.bateman,a.ho}@surrey.ac.uk

1 Introduction

The primary reason for the requirement of authenticating images stems from the increasing amount of doctored images that are presented as accurate representations of real-life events, but are later discovered to be faked. The history of manipulating images reaches back almost as far as photography itself, and with the ease of use and availability of image editing software, it has become ubiquitous in the digital age. Image authentication schemes attempt to restore trust in the image by accurately validating the data, positively or negatively. Especially for law enforcement scenarios, images captured at the scene, such as for crime scene investigation and traffic enforcement, potentially be used as evidence in the court of law. If an image presented in court as evidence from a crime scene is to be effectively used by the jury, the integrity of the information must not be in question. The role of a scene of crime officer (SoCOs) is to capture, as much as possible, the left-over evidence at the crime scene by taking photographs and collecting any exhibits found. After the collection of evidence, there is no other way of examining the crime scene as a whole, apart from analysing the collected exhibits and photographs taken [1]. In order to maintain the integrity of the images, not only it is essential to verify that the photographic evidence remains unchanged and authentic, but any manipulated regions should also be localised to help identify which parts of the image cannot be trusted. With the tremendous growth and usage of digital cameras and video devices, the requirement to verify the digital content is paramount, especially if it is to be used as evidence in court [2]. Therefore, digital watermarking technique can be utilised for image content authentication applications to verify or authenticate the integrity of the digital media content.

Digital watermarking is the process of embedding relevant information (such as a logo, fingerprint and serial number), into a media. This technique can be applied to different media types such as video, audio and image content. An example of digital visible watermark is the translucent logos that are often seen embedded at the corner of videos or images, in an attempt to prevent copyright infringement. However, these visible watermarks can be targeted and removed rather simply by

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 139–183.
springerlink.com

cropping the media, or overwriting the logos. Subsequently, the field of digital watermarking is primarily focused on invisible watermarks, which are imperceptible and operate by tweaking the physical data of the media [3, 4]. There are three different classifications associated with digital watermarking, depending on the applications: robust, fragile and semi-fragile. Robust watermarking is primarily designed to provide copyright protection and proof of ownership for digital images. The most important property of robust watermarking is its ability to tolerate certain signal processing operations that usually occur during the lifetime of a media object, as well as preventing any more deliberate attacks.

Fragile and semi-fragile digital watermarking techniques are often utilised for image content authentication. Fragile watermarking schemes are designed to detect any possible manipulations that affect the watermarked image pixel values [5, 6]. In comparison, while fragile watermarking is aptly named because of its sensitivity to any form of attack, semi-fragile watermarking is more robust against attack, and can be used to verify tampered content within images for both malicious and non-malicious manipulations [7–9]. In addition, semi-fragile schemes make it possible to verify the content of the original image, as well as permitting alterations caused by non-malicious (unintentional) modifications such as system processes. Moreover, semi-fragile watermarking is more focused on detecting intentional attacks than validating the originality of the image [10, 11]. During the image transmission, the mild signal processing errors caused by signal reconstruction and storage, such as transmission noise or JPEG compression, are permissible. However, the image content tampering such as copy and paste attack will be identified as a malicious attack. Additionally, in the literature, a significant amount of research has been focused on the design of semi-fragile algorithms that could tolerate JPEG compression and other common non-malicious manipulations [12–18]. However, watermarked images could be compressed by unknown JPEG compression rates of various quality factors (QFs). As a result, in order to authenticate the images, these algorithms have to set a pre-determined threshold that could allow them to tolerate different QF values when extracting the watermarks. To determine the threshold more accurately, the generalised Benford's law can be utilised to estimate the unknown JPEG compression QF, then appropriate thresholds could be adapted for each test image, before initialising the watermark extraction and authentication process. This law has already been successfully used in image forensics technique for JPEG compression evaluation [19]. This adaptive threshold could help to decrease the false alarm and missed detection rates.

In contrast to authenticate the image using active watermarking technique, the image forensics as passive technique has attracted much attention [20–22]. The significant difference is that image forensics seeks to authenticate images based solely on the image data provided in image statistical analysis, meaning it is a passive approach to the problem. As such, no embedded information is loaded into an image, and so the security risks and robustness issues associated with a payload, are avoided. Therefore, image forensics presents itself as an alternative approach to the active insertion of watermarking data to authenticate images. In this chapter, we will review active watermarking techniques, such as fragile and semi-fragile methods as well as passive image forensics techniques such as camera identification and forgery detection methods for image authentication. Furthermore, we will introduce our three proposed image authentication related methods, which are fragile watermarking scheme in Slant transform (SLT) domain, utilising the generalised Benford's Law as image forensics technique to improve semi-fragile watermarking technique and the use of the statistical process control (SPC) for camera identification in image forensics research.

The chapter is organized as follows:

- In Section 2, several fragile and semi-fragile watermarking schemes will be reviewed. Our proposed SLT semi-fragile watermarking algorithm is then introduced. The watermark embedding, detection and authentication processes are described in detail as well as the proposed experimental results are analysed and evaluated by comparing with two other transform based scheme, which in Discrete Cosine transform (DCT) and Pinned Sine transform (PST) domain.
- Section 3 discusses three typical methods of employing predetermined thresholds in semi-fragile watermarking algorithms and the limitations of using predetermined thresholds were highlighted from the literature. Then we proposed a framework incorporating the generalised Benford's Law that could detect unknown JPEG compression QFs in semi-fragile watermarked images to adjust the appropriate threshold with experimental results.
- Section 4 will review image forensics techniques that focus on two main areas, camera identification and image forgery detection and their applications. Then we propose to utilise SPC methods to analyses images captured from different digital camera devices.
- Section 5 gives the conclusion of this chapter and presents some directions for future work of the research.

2 Fragile and Semi-fragile Watermarking

In this section, both fragile and semi-fragile watermarking algorithms for image authentication are reviewed. A detailed discussion on our proposed semi-fragile watermarking schemes in SLT domain to further explain the concept of semi-fragile watermarking is also presented. The results of miss detection rates and false alarm rates are then compared with two existing transforms based on the DCT and PST transforms.

2.1 Literature Review for Fragile and Semi-fragile Watermarking

Fragile Watermarking

As mentioned in Section 1, fragile watermarking schemes should be able to detect any possible manipulations that affect the watermarked image any pixel values. Therefore, it is possible to exploit the inherent weakness of the LSB schemes, and implement a fragile watermarking scheme in the spatial domain. Fridrich [23] proposed a spatial domain based fragile watermarking scheme that could localise tampered regions of a watermarked image, by adapting Wongs method [24]. The watermark embedding process is shown in Figure 1. The original image is first divided into non-overlapping blocks of equal size 8 by 16. In each block, the seven Most Significant Bits (MSB) of each pixel are extracted, and a cryptographic hash function is applied as illustrated in Figure 2. The logo is also divided into 816 blocks and each block contains information about the original block position, image index, original image dimensions (resolution), camera ID and author ID (PIN). The hashed seven MSBs of each block and its corresponding logo block are subjected to an Exclusive-OR (XOR) operation and then encrypted using a key. Finally, the LSBs of the original image are replaced with the result of the XOR operation and encrypted watermark bits, and the watermarked image is created. In the authentication process, the LSBs of the test image are extracted, and the seven MSBs from each block are hashed as shown in Figure 3. For each block, the LSBs are decrypted with a key, along with its corresponding hashed seven MSBs using the XOR operation. Finally, the authentication process itself is achieved by comparing each block of the image with the corresponding block from the logo. If this set of the block is not the same, the block of the image is flagged as a tampered block.

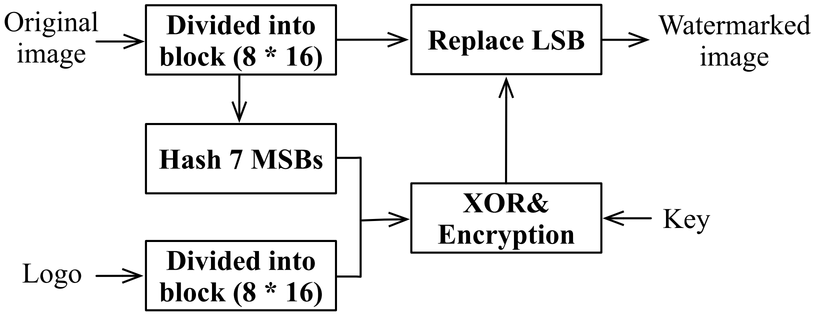


Fig. 1. Fridrich's fragile watermark embedding algorithm

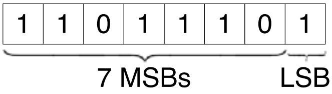


Fig. 2. MSBs and LSB of pixel value 221 in 8 bits binary sequence

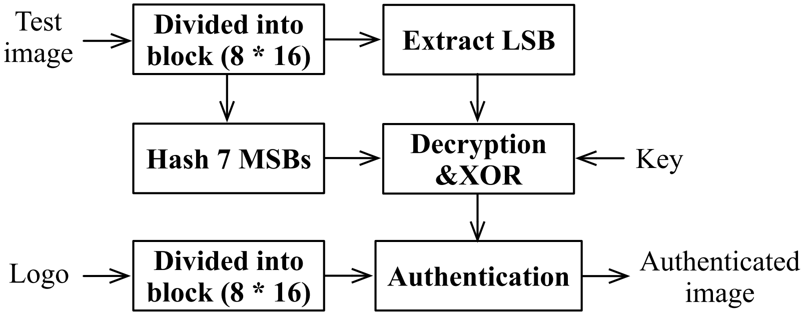


Fig. 3. Fridrich's fragile watermark detection algorithm

Zhang and Wang [25] proposed a statistical scheme of fragile watermarking scheme that embed a folded version of the authentication data derived from five most significant bits (5MSBs) of the original image along with other additional data into the image with acceptable watermarked image quality PSNR as $37.9dB$. Their results showed their algorithm could localized the tampered pixels accurately. Then they further improved their method in [25] that could restore the tampered image content after localized the tampered area without any errors [26]. He *et al.* [27] proposed a conventional self-embedding fragile watermarking scheme based on adjacent-block based statistical detection method (SDM) that could against copy-paste attack and collage attack. Their algorithm could identify the tampered blocks with a probability more than 98% even the tampered area is up to 70% of the host image.

Fragile watermarking scheme can also be applied in transform domain. Li and Shi [5] proposed a fragile watermarking algorithm in Discrete Wavelet Transform (DWT) to achieve the requirements of high security, low distortion, and high accuracy of tamper localization for authenticating JPEG2000 images. Their algorithm could also tolerate vector quantization attack, Holliman-Memon attack, college attack and transplantation attack. Aslantas *et.al* [28] proposed intelligent optimization algorithms (IOA) to improve fragile watermarking schemes in discrete cosine transform (DCT) domain. They used IOA which including four genetic algorithm (GA), clonal selection algorithm (CSA), particle swarm optimization (PSO), and Differential Evolution (De) to correct rounding errors caused by transforming an image from the frequency domain to the spatial domain with the objective of improving DCT-based fragile watermarking. The experimental results showed that the CSA produces better PSNR results whereas DE has lower computational time than other algorithms. Yeh and Lee [29] proposed reversible fragile watermarking by utilizing the pyramidal structure method. They select appropriate embedding areas by analysing the pyramid-structure of the image for embed watermark bits in wavelet domain. The experimental results showed that their scheme could successfully localized even when 50% of the watermarked image is tampered as well as detect counterfeiting attack.

Semi-fragile Watermarking

Many semi-fragile watermarking techniques have been already proposed by researchers. Lin *et al.* [31] proposed embedding algorithm that first applied Discrete Cosine Transform (DCT) to 16 by 16 blocks of the cover image, then embed the watermarks in middle to low frequency (except DC coefficient) of each block. Their scheme could identify the tampered area with 75% accuracy under moderate compression and with near 90% accuracy under light compression. Ho *et al.* [7] proposed a semi-fragile watermarking scheme in Pinned Sine Transform (PST) domain. In their algorithm, the original image is applied by using PST to get the pinned and boundary fields in 8 by 8 blocks. The watermark bits are then inserted into middle to high frequency of each block in the pinned field. The scheme also used a self-restoration method, originally proposed by Fridrich and Goljan [33] to recover the tampered regions. Their scheme could tolerate some common image processing manipulations such as JPEG and wavelet compression, and the detection rate is higher than DCT-based scheme. The algorithm has been further improved by using irregular Sampling instead of the LSB method [15],which aimed to improve the robustness of tampering restoration. Kundur and Hatziankos [18] proposed a DWT based algorithm called *telltale tamper-proofing*, which made it possible to determine tampered regions in multi-resolutions. Unlike other schemes that use DCT, this method does not require a block division process to detect the tampered regions due to the localisation ability of the wavelet transform. The localization ability of the wavelets in both spatial and frequency domains would potentially indicate a good candidate for semi-fragile watermarking.

Maeno *et al.* [34] presented two algorithms that focused on signature generation techniques. The first algorithm used random bias to enhance the block based DCT watermarking scheme proposed by Lin and Chang [12]. The second algorithm used nonuniform quantisation on a non-block based semi-fragile watermarking scheme in the wavelet domain. Their experimental results showed their method was fragile to malicious manipulations, but robust to non-malicious manipulations such as JPEG and JPEG2000 compression. Ding *et al.* [35] also proposed a method by using DWT. In their algorithm, chaos was used to generate a pseudo-random sequence as a watermark, in an effort to improve the overall security. This made an improvement to the more traditional methods of generating a pseudo-random sequence. The sub-bands (HL_2,LH_2,HH_2) were used for embedding the watermark after applying a 2-level wavelet decomposition of the original image. The normalized cross-correlation (NC) was used to evaluate their algorithm by comparing between the original watermark and the extracted watermark after applying JPEG compression and Additive white Gaussian noise (AWGN) manipulations. Ni *et al.* [30] proposed a robust lossless data hiding technique that could be employed into semi-fragile watermarking scheme. The different bit-embedding strategies for groups of pixels with different pixel grayscale value distributions and error correction codes are utilized in their scheme. They analyzed their results into two modules, which are lossless and lossy. If the watermarked image has experienced losslessly compression, the watermark bits can be extracted correctly and the image will be classified as authentic and the original image can be recovered exactly. If this losslessly compressed watermarked image has been further undergone lossy compression, the original image will not be able to be recovered and will be rendered authentic as long as the compression is not so severe that the content has been changed.

2.2 Proposed Slant Transform (SLT) Semi-fragile Watermarking

This section will discuss our proposed method [36] in detail, which consist of the embedding, detection and authentication processes associated with watermarking.

Slant Transform (SLT)

The Slant Transform has been applied to image coding in the past [37] and was recently adopted for robust image watermarking [38]. The SLT can be considered as a fast computational algorithm provides a significant bandwidth reduction and result in a lower mean-square error for moderate size image blocks [37]. In addition, for textured images, the quality of the Slant Transformed images is higher than images coded by using other transforms such as DCT and Hadamard [39]. Moreover, as a similar image processing application to Walsh-Hadamard transform, Slant transform can be identified as a sub-optimum for energy compaction, which is essential for digital watermarking as the robust information hiding can be ensured by capitalizing the spread of middle to higher frequency bands. Furthermore, Slant transform is simpler, faster and especially suitable for highly textured images [38]. Hence, the Slant Transform is proposed for semi-fragile watermarking and authentication of images in this section. The authentication as the method to corroborate the genuineness of an object is mainly focusing on examining whether the image has been tempered or not, the location(s) of tampered region(s) and to what extent it has been changed can also be identified. Furthermore, the SLT can also be used for compressing the original image [39], providing a means to self-recovering the tampered regions by embedding the compressed cover image into the LSBs of the watermarked image [33]. The forward and inverse of SLT [37–39] can be expressed as follows:

$$[\mathbf{V}] = [\mathbf{S}_N][\mathbf{U}][\mathbf{S}_N]^T \qquad [\mathbf{U}] = [\mathbf{S}_N]^T[\mathbf{V}][\mathbf{S}_N] \tag{1}$$

where $[\mathbf{U}]$ represents the original image of size $N \times N$, $[\mathbf{V}]$ represents the transformed components and $[\mathbf{S}_N]$ is the $N \times N$ unitary Slant matrix given by

$$[\mathbf{S}_N] = \frac{1}{\sqrt{2}} \left[\begin{array}{cc:c:cc:c} 1 & 0 & \mathbf{0} & 1 & 0 & \mathbf{0} \\ a_N & b_N & & -a_N & b_N & \\ \hdashline \multicolumn{2}{c:}{\mathbf{0}} & \mathbf{I}_{(N/2)-2} & \multicolumn{2}{c:}{\mathbf{0}} & \mathbf{I}_{(N/2)-2} \\ \hdashline 0 & 1 & \mathbf{0} & 0 & -1 & \mathbf{0} \\ -b_N & a_N & & b_N & a_N & \\ \hdashline \multicolumn{2}{c:}{\mathbf{0}} & \mathbf{I}_{(N/2)-2} & \multicolumn{2}{c:}{\mathbf{0}} & -\mathbf{I}_{(N/2)-2} \end{array} \right] \left[\begin{array}{c:c} \mathbf{S}_{N/2} & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{S}_{N/2} \end{array} \right]$$

where $\mathbf{I}_{(N/2)-2}$ is the identity matrix of dimension $(N/2)-2$ and

$$a_{2N} = \left(\frac{3N^2}{4N^2-1}\right)^{1/2}, \qquad b_{2N} = \left(\frac{N^2-1}{4N^2-1}\right)^{1/2}$$

are constants.

Watermark Embedding

A novel semi-fragile Slant Transform digital watermarking method is adopted based on previous work relating to PST [7] and self-restoration method [33]. The entire embedding process using the Slant Transform is illustrated in Figure 4, which consists of two parts. The first 7 bits of the cover image are extracted and divided into 8×8 blocks, SLT method is then applied to each block. The watermark embedding algorithm is then utilised, which is illustrated in the pseudo-code below. The watermarks for each block are then random generated by input a key as a seed. The obtained watermarks are embedded into the midband of each 8×8 block. After watermark embedding, frequency coefficients of each block of the watermarked image are converted back by using the inverse Slant Transform. Consequently, the first 7 bits of final watermark image is obtained. The SLT watermark embedding algorithm in pseudo-code form is shown as follows:

```
If $w == 1$ And $x \geq \tau$, Then $y = x$, Else $y = \alpha$.
If $w == 0$ And $x < -\tau$, Then $y = x$, Else $y = -\alpha$.
```

where w is the watermark bit, x is the SLT coefficient of the host, y is the modified SLT coefficient, τ is the threshold which controls the perceptual quality of the watermarked image and α is a constant. Similar to part 1, the original image is divided into 8×8 sub-blocks and also undergoes the same Slant Transform; compression for each sub-block is then achieved by discarding the high frequency coefficients. Accordingly, 64 bits information for each block is acquired after compression and then encrypted by utilizing a key as a seed. Obtained blocks are then shuffled, e.g. the value of block 1 moves to block 50, the value of block 35 moves to block 10. Therefore, LSBs of the final watermark image are then gained. Finally, the combination of part1 and part 2 forms the final watermarked image and the key file is generated, which contains information that mentioned previously.

Watermark Detection, Authentication and Restoration

The proposed semi-fragile Slant Transform for image authentication and restoration method is shown in Figure 5. Similar to embedding process, the first 7 bits of the test image are extracted and divided into 8×8 blocks by applying SLT and then apply the detection algorithm to the first 7 bits, which is explained in the paragraph below. Meanwhile, the LSBs are extracted from the test image and only the LSBs of the detected regions are quantized back for recovery by according to authentication result. Consequently, authenticated and recovered images can be output.

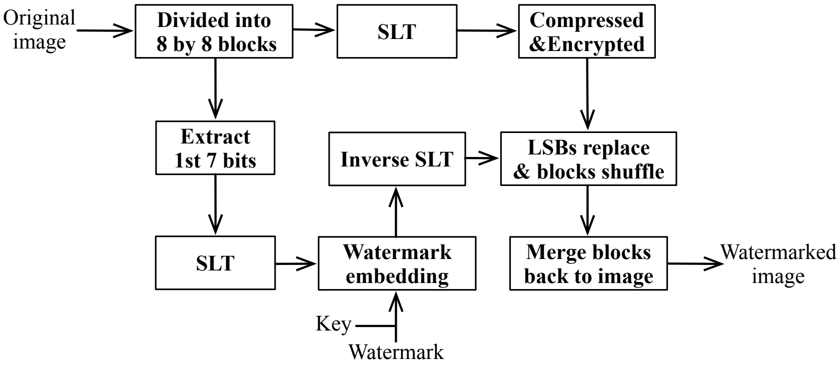


Fig. 4. Our proposed SLT watermark embedding process

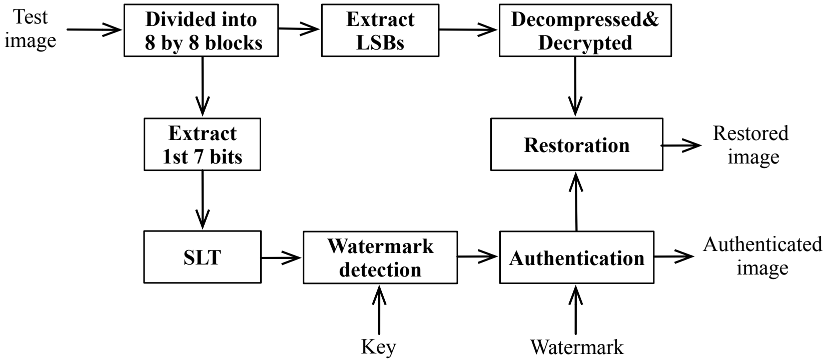


Fig. 5. Our proposed SLT watermark detection, authentication and restoration process

The watermark bits can be detected by extracting the watermarked coefficients y. If y larger than 0, the watermark bit value is 1; if y smaller than 0, the watermark bit value is 0. The retrieved watermark needs to be compared with the watermark that exists in the key file. After the watermark bits from the entire block have been retrieved, the comparison between the watermark bits can be accomplished by using the correlation coefficient ρ, computed as follows:

$$\rho = \frac{\sum\sum\left(\mathbf{w}' - \bar{\mathbf{w}}'\right)\left(\mathbf{w} - \bar{\mathbf{w}}\right)}{\sqrt{\sum\sum\left(\mathbf{w}' - \bar{\mathbf{w}}'\right)^2 \sum\sum\left(\mathbf{w} - \bar{\mathbf{w}}\right)^2}} \tag{2}$$

where $\mathbf{w}$ is the original and $\mathbf{w}'$ is the retrieved watermarks corresponding to the block. For error correction, the correlation coefficient ρ can be compared with a pre-determined threshold value λ. If $\rho < \lambda$, which indicates that the block has been tampered as authentication, and which is followed by restoration of the tampered regions based on the decompression and extraction of the LSBs for the watermarked image.

2.3 Results and Evaluation

A number of experiments have been carried out to evaluate the performance of the proposed SLT watermarking scheme. The proposed watermarking scheme is compared with two other watermarking schemes: the PST-based [7] and the DCT-based [31]. For a fair comparison, the embedding strength of the watermark in each scheme is adjusted such that the peak signal-to-noise ratio (PSNR) of the watermarked images is around 33 dB, which is subjectively considered as acceptable. The performance of the watermarking schemes is measured in terms of the false positive detection rate (P_{FP}), false negative detection rate (P_{FN}) and the average detection rate (P_{avg}), defined as:

$$P_{FP} = \frac{\text{Number of pixels in the untampared region as detected as tampered}}{\text{Total number of pixels in the untampered region}}$$

$$P_{FN} = \frac{\text{Number of pixels in the tampared region as detected as untampered}}{\text{Total number of pixels in the tampered region}}$$

and

$$P_{avg} = \left(1 - \frac{P_{FP} + P_{FN}}{1 + N}\right) \times 100. \tag{3}$$

where N is the number of area(s) have been tampered with. A number of standard test images are used in the experiments and the results for 6 images, each of size 512×512 are reported.

JPEG Compression Attack

Table 1 shows that SLT, DCT and PST are compared by applying JPEG compression attack to 6 different grayscale images (512×512) in order to determine the false positive rate, i.e. over detected rate. As can be seen from the table below, SLT, DCT and PST have similar error detection rates when $QF = 85$. After experiencing 75% JPEG compression attack, the over detection rate of SLT is still considerably low with average rate of 1.2, whereas PST and DCT have the higher average over detection rates of 86 and 31.4, respectively. Although the over detection rates of all three methods have increased when $QF = 65$, SLT still has the lowest increased rate of 30.8 comparing with the average value of over detection rates of PST and DCT, of 92.2 and 88.2 respectively. The reason for the relatively better results using the Slant Transform was that the embedding locations concentrated mainly in the middle frequency band, which is considered to be more robust, whereas DCT and PST mainly concentrated more on high frequencies. Overall, the results indicate that the SLT watermarking method achieves lower errors than PST and DCT based on the JPEG compression attack.

Table 1. Comparative performance of the watermarking schemes against JPEG compression with varying quality factor

Test Image	$\mathbf{QF = 85}$			$\mathbf{QF = 75}$			$\mathbf{QF = 65}$		
	SLT	PST	DCT	SLT	PST	DCT	SLT	PST	DCT
Lena	0.0	0.5	0.0	0.7	91.2	31.5	37.2	93.5	81.4
Baboon	0.1	0.5	0.1	0.9	91.2	31.6	24.1	91.2	91.0
Bridge	0.4	1.2	0.5	1.6	83.7	32.9	28.4	91.8	91.5
Trucks	0.2	0.8	0.4	1.4	87.4	31.2	33.3	92.8	90.1
Ship	0.2	0.8	0.4	2.0	89.0	31.3	37.5	92.3	85.0
San Diego	0.0	0.4	0.0	0.6	83.4	30.0	24.2	91.3	90.2
Average	0.2	0.7	0.2	1.2	86.0	31.4	30.8	92.2	88.2

Copy and Paste Attack

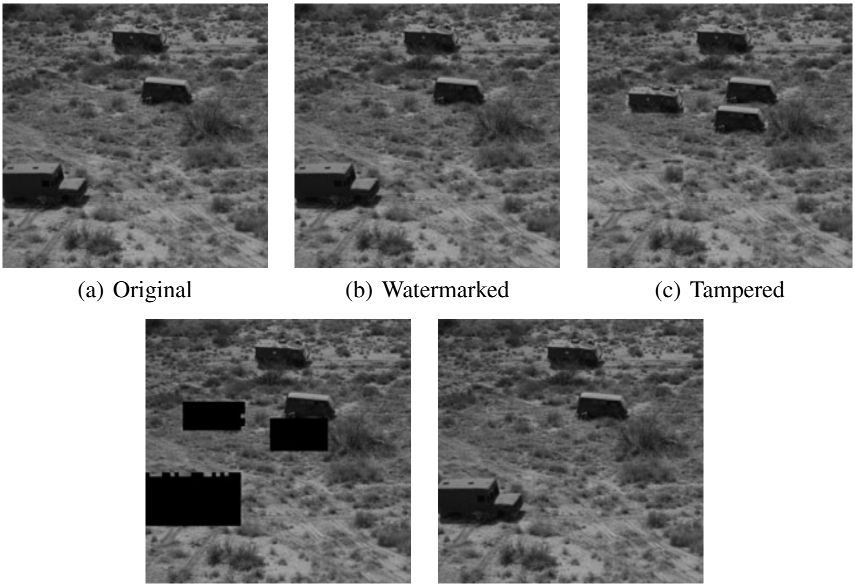
The copy and paste attack is utilized to compare the performance of detection rates SLT, DCT and PST for six grayscale test images (512 × 512) as given in Table 2. Three different tampering rates of 10%,20% and 30% will be applied to each test image to analyse the overall detection rate of the three transform methods. The tamper tests are performed with 100 random locations on each image. Consequently, 5400 test images are obtained based on this experimental setup. Table 2 shows the comparative performance of the three watermarking schemes against copy and paste attack with different amount of tampering. However, the results show that PST is the most sensitive method as it has the highest overall detection rate after experiencing all three tamper tests (10%, 20% and 30%) of all images. Figure 6(a-e), shows the original, watermarked, tampered, authenticated and restored images for the image *Trucks*, respectively.

Table 2. Comparative performance of the watermarking schemes against copy-paste attack

Test Image	**10% tamper**			**20% tamper**			**30% tamper**		
	SLT	PST	DCT	SLT	PST	DCT	SLT	PST	DCT
Lena	96.0	97.6	95.5	97.9	98.7	97.1	98.3	99.0	97.3
Baboon	96.7	97.3	96.3	98.1	98.8	96.5	98.5	99.0	97.0
Bridge	96.3	97.5	95.4	97.9	98.6	97.0	98.2	98.8	96.9
Trucks	95.7	97.6	95.1	97.6	98.7	96.7	98.3	98.8	96.9
Ship	96.1	97.6	95.2	97.7	98.8	96.2	98.3	98.8	96.7
San Diego	96.6	97.6	95.4	98.1	98.8	96.6	98.6	99.0	97.5
Average	96.2	97.5	95.5	97.9	98.7	96.7	98.4	98.9	97.0

JPEG Compression + Copy and Paste Attack

In Table 3, the six watermarked images (512 × 512) are compressed with three different JPEG compression rates QF of 85, 75 and 65. Th experimental setup is similar to the previous copy and paste attack with 100 random locations for



(a) Original (b) Watermarked (c) Tampered

(d) Authenticated (e) Restored

Fig. 6. Demonstration of the image *Trucks* in SLT semi-fragile watermarking scheme

Table 3. Comparative performance of the watermarking schemes against copy-paste attack (20$\%$ tampering) followed by JPEG compression with varying quality factor

Test Image	**QF = 85**			**QF = 75**			**QF = 65**		
	SLT	**PST**	**DCT**	**SLT**	**PST**	**DCT**	**SLT**	**PST**	**DCT**
Lena	92.1	94.9	91.6	91.6	51.4	77.0	75.3	50.7	54.0
Baboon	92.1	94.7	92.1	92.3	56.1	77.0	81.8	51.8	49.9
Bridge	91.9	94.0	91.6	92.3	55.3	76.6	79.5	51.6	49.5
Trucks	92.0	94.5	91.2	91.4	53.5	76.9	76.7	51.0	50.4
Ship	91.5	93.7	91.9	91.4	52.7	76.9	75.7	50.8	52.7
San Diego	93.1	94.5	92.2	92.4	55.4	77.6	81.6	51.5	49.7
Average	92.1	94.4	91.8	91.9	54.1	77.0	78.4	51.2	51.0

tampered areas. Overall, the PST achieves a relatively higher detection rate than DCT and SLT after experiencing QF of 85. However, for detection, it is worse at 54.1% with $QF = 75$. In comparison, SLT has the highest overall detection rate as 91.9% at $QF = 75$ and 65. From the analysis, SLT is showed to achieve a more accurate detection result than PST and DCT. On the whole, the result indicates that the best overall detection rate among the three methods is SLT, which has 91.9% detection rate with $QF = 75$. However, all the attacked images could not be recovered by any of the three transform schemes after applying JPEG compression attack.

This is duo to the fact that the restoration technique is based on LSB embedding in the spatial domain of the watermarked image which is fragile. As such, it can be easily removed by JPEG compression.

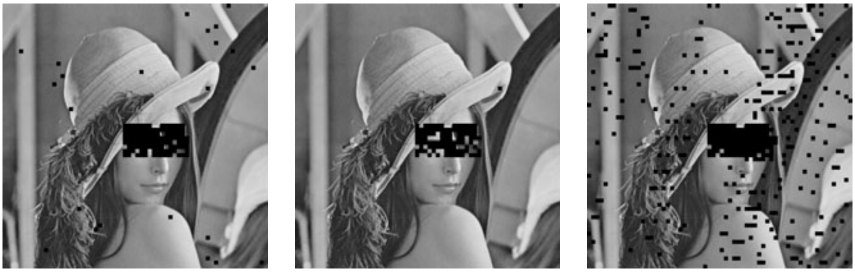
2.4 Summary

In this section, we reviewed a number of different fragile and semi-fragile watermarking schemes. Our proposed SLT semi-fragile watermarking scheme was discussed in detail. The performance of the SLT based semi-fragile scheme was compared with the PST and DCT based schemes by using average detection rate calculated from false positive and false negative detection rates. The comparative studies showed that the SLT-domain watermarking scheme performed better against JPEG compression, copy-paste attack, as well as combined JPEG compression and copy-paste attacks than the PST and DCT-domain watermarking schemes.

3 Image Forensics Technique - Benford's Law for Semi-fragile Watermarking

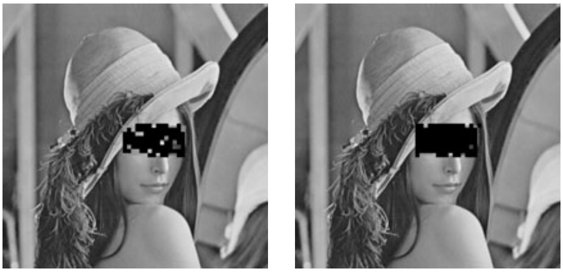
As mentioned in Section 1, semi-fragile watermarking scheme has been used to authenticate and localise malicious tampering of image content, while permitting some non-malicious or unintentional manipulations. These manipulations can include some mild signal processing operations such as those caused by transmission and storage of JPEG images. However, watermarked images could be compressed by unknown JPEG QFs. As a result, in order to authenticate the images, these algorithms have to set a pre-determined threshold that could allow them to tolerate different QF values when extracting the watermarks.

Figures 7 and 8 illustrate the overall relationship between the threshold, false positive and false negative detection rates. The watermarked image *Lena* has been tampered with a rectangular block and JPEG compressed at $QF = 75$. Figure 7(a) shows the pre-determined threshold $T = 0.5$ used for authentication. The authenticated image shows that the proposed semi-fragile watermarking scheme can localise the tampered region with reasonable accuracy, but with some false positive detection errors. In Figures 7(b) and 7(c), the lower and upper thresholds $T = 0.3$ and $T = 0.7$ were used for comparison, respectively. Figure 7(b) shows that the false positive rate has decreased whilst the false negative rate has increased in the authenticated image. Figure 7(c) shows the image has a lower false negative rate but with a higher false positive rate. From this comparison, $T = 0.5$ was chosen for JPEG compression at $QF = 75$. However, if $QF = 95$, then $T = 0.5$ may not be adequate as shown in Figure 8(a). The false negative rate is higher than Figure 8(b) with $T = 0.9$. Therefore, it would be advantageous to be able to estimate the QF of JPEG compression, so that an adaptive threshold can be applied for increasing the authentication accuracy. In this section, we discuss our proposed method [40] to utilise the generalised Benford's Law,as an image forensics



(a) $T = 0.5$ (b) $T = 0.3$ (c) $T = 0.7$

Fig. 7. Different thresholds for $QF = 75$



(a) $T = 0.5$ (b) $T = 0.9$

Fig. 8. Different thresholds for $QF = 95$

technique to estimate the QF for semi-fragile watermarked images. The background of Benford's Law, generalised Benford's Law and their relationship with the watermarked image, JPEG compressed watermarked image are also described.

3.1 Benford's Law and Generalised Benford's Law

Benford's Law was introduced by Frank Benford in 1938 [41] and was developed by Hill [42] for analysis of the probability distribution of the first digit $(1-9)$ of the number from natural data in statistics. Benford's Law has also been applied to accounting forensics [43, 44]. The DCT coefficients of a digital image was forward to obey Benford's Law, it has recently attracted a significant amount of research interests in image processing and image forensics [19, 45, 46]. The basic principle of Benford's Law is given as follows:

$$p\left(x\right) = log_{10}\left(1+\frac{1}{x}\right), x = 1, 2, ...9 \tag{4}$$

where x is the first digit of the number and $p(x)$ is the probability distribution of x. In contrast to digital image watermarking which is an *active* approach by embedding bits into an image for authentication, image forensics is essentially a *passive* approach of analysing the image statistically to determine whether it has been tampered with. Fu et al. [19] proposed a generalised Benford's Law, used for estimating the QF of the JPEG compressed image, as shown in equation 5.

$$p\left(x\right) = Nlog_{10}\left(1 + \frac{1}{s + x^{q}}\right), x = 1, 2, ...9 \tag{5}$$

where N is a normalisation, and s and q are model parameters [19]. Their research indicated that the probability distribution of the 1^{st} digit of the JPEG coefficients obey generalised Benford's Law after the quantisation. Moreover, the probability distributions were not following the generalized Benford's Law if the image had been compressed twice with different quality factors. Thus, by utilizing this property, the QF of the image can be estimated.

Figures 9 to 11 illustrate the comparisons between the probability distribution of Benford's Law, generalized Benford's Law and the mean distributions of the 1st digits of block JPEG coefficients of the watermarked images compressed at $QF = 100, 75, 50$, respectively. Throughout this section we adhere to the same terminology as used in [19], where *JPEG coefficients* refers to the 8×8 block-DCT coefficients after the quantisation. These results based on 1338 images from [47] indicate a good fitting between generalized Benford's Law and watermarked images compressed with different QFs. The results indicate that the probability distributions of the 1^{st} digits of JPEG coefficients of the watermarked images, as shown in Figures 9 to 11, obey the generalised Benford's Law model proposed by Fu et al. [19], in equation 5. Hence, we could employ their model to estimate the unknown QF of test images to adjust the threshold for authentication. The improved authentication process is described in the next section.

3.2 The Improved Authentication Method

In order to improve the detection rate in semi-fragile authentication process, the test image is first used for detecting the QF by the quality factor estimation process. This process works by firstly classifying the test image as compressed or uncompressed by adapting from [19]. If the test image has been compressed, the test image is then recompressed with the largest QF, from $QF = 100$ to $QF = 50$, in decreasing steps of 5. We decrease in steps of 5 as this gives us the most frequently used quality factors for JPEG compressed images (i.e. 95, 90, 85 etc.). For each compressed test image, the probability distribution of the 1^{st} digits of JPEG coefficients is obtained. Each set of values are then analysed by employing the generalized Benford's Law equation and using the best curve-fitting to plot the data. In order to obtain the goodness of fit, we calculate the sum of squares due to error (SSE) of the recompressed images. We can detect the QF of the test image by iteratively calculating the SSE for all QFs (starting at $QF = 100$, and decreasing in steps of 5), and as soon as $SSE < 10^{-6}$, we have reached the estimated QF for the test image. The threshold 10^{-6}, was reported in [19], has been set to allow us to detect the QF of the test image, and has also been verified by the results in our experiment.

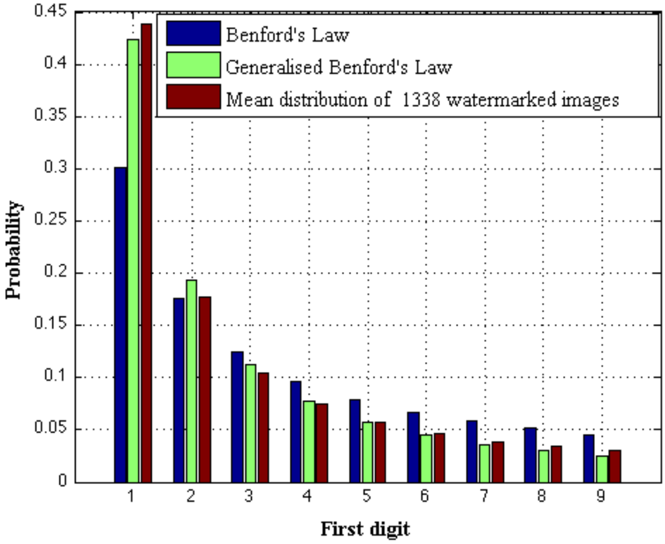


Fig. 9. 1^{st} digit of JPEG coefficients ($QF = 100$)

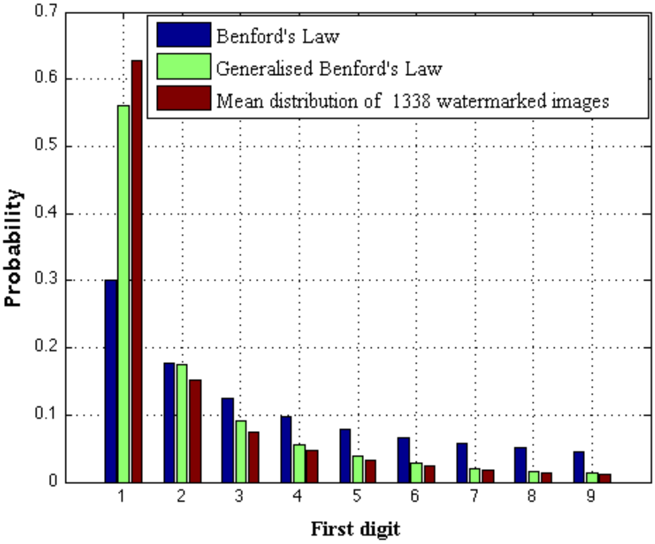


Fig. 10. 1^{st} digit of JPEG coefficients ($QF = 75$)

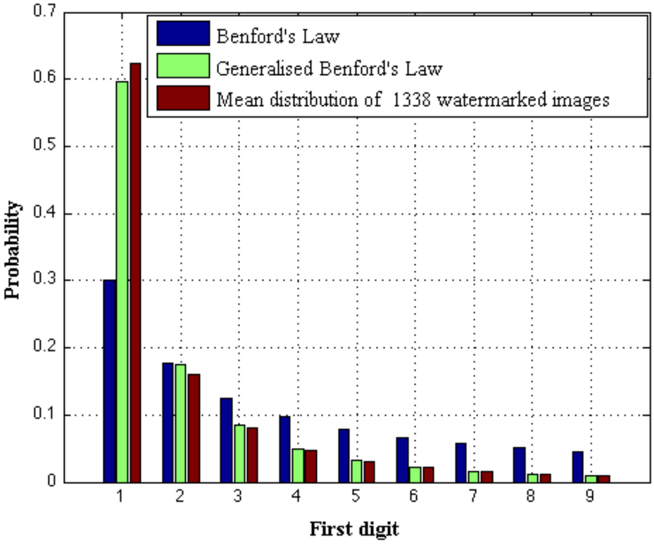


Fig. 11. 1^{st} digit of JPEG coefficients ($QF = 50$)

Figure 12 illustrates the results of estimating the QF for a test image that has previously been compressed with $QF = 70$. Three curves have been drawn in order to fit the three probability distribution data sets: generalized Benford's Law for $QF = 70$, the test image recompressed with $QF = 70$, and separately recompressed at $QF = 90$. The distribution of $QF = 90$ shows the worst fit and is considerably fluctuated, while the distribution of $QF = 70$ is a generally decreasing curve, which also follows the trend of generalized Benford's Law. These results indicate that if the test image has been double compressed without the same quality factor, the probability distribution would not obey the generalised Benford's Law.

Once the QF is estimated, the threshold T can be adapted according to different estimated QFs, based on the following conditions in Equation 6. Finally, the correlation coefficient between original watermarks and extracted watermarks for each block is compared using the attuned threshold T to authenticate, in order to determine whether any blocks have been tampered with.

$$T = \begin{cases} 0.9 & QF \geq 90 \\ 0.7 & 90 < QF < 75 \\ 0.5 & QF \leq 75 \end{cases} \tag{6}$$

3.3 Results and Evaluation

The watermarked images are generated based on a simple DCT domain based semi-fragile watermark embedding scheme by using the 1338 test images from [47].

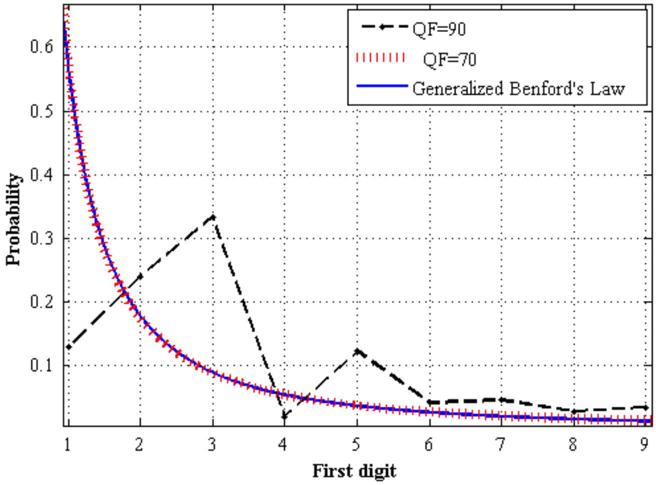


Fig. 12. Estimating the QF of a watermarked image

In order to achieve a fair comparison, different embedding parameters are randomised for each image such as the watermarks location, watermark string and watermark bits. For our analysis, four types of test images with and without attacks are considered as shown in Figure 13.

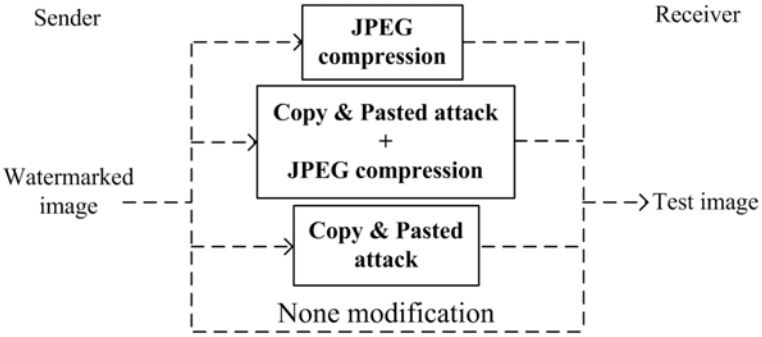


Fig. 13. Four types of test images with and without attacks

Table 4 summaries the results obtained for test images that have been JPEG compressed only. To evaluate the accuracy of the quality factor estimation process, each test image has been blind compressed from $QF = 100$ to $QF = 50$ in decreasing steps of 5. For each JPEG compression, the quality factor estimation process was

Table 4. QF estimation for watermarked images (JPEG compression only)

Actual QF	Estimated QF	P_{de}	T	P_{de2}
100	98.16	65.7%	0.9	98.8%
95	94.87	97.3%		
90	90.06	98.2%		
85	84.20	91.4%	0.7	99.1%
80	79.77	97.5%		
75	75.35	97.0%	0.5	99.4%
70	69.77	98.8%		
65	64.42	93.7%		
60	62.42	38.6%		
55	55.15	94.1%		
50	54.25	18.2%		

used to determine the QF. The mean estimated QFs for all 1338 test images and each correctly identified detection accuracy rate P_{de} for each JPEG compression quality factor are shown in Table 4, based on equation 7.

$$P_{de} = \frac{\partial}{\beta} \times 100\% \tag{7}$$

where ∂ is the number of correctly detected QF and β is the number of images tested. The mean estimated QF results indicate the QFs can be estimated with high accuracy. The only exceptions for lower correct detection rates, P_{de}, were obtained for $QF = 50$, $QF = 60$, and $QF = 100$. In the case of $QF = 50$, P_{de} was very low at approximately 18.2%, meaning that the process was probably detecting QFs close to $QF = 55$. For $QF = 60$, and $QF = 100$, the detection rates were slightly better at 38.6% and 65.7%, respectively. For comparison, both the mean estimated QF value and correct detection rate were used for each result to estimate the actual QF for the images. The QFs were then grouped into three different ranges: $QF \geq 90$, $90 < QF < 75$ and $QF \leq 75$. The grouping into three QF ranges did not have an overall effect on the authentication process. Results obtained for P_{de2} also showed the correct detection accuracy rates in these QF ranges were on average at 99%.

Table 5 summaries the results obtained for test images that have been attacked via copy-paste and then JPEG compressed. Each watermarked image has been tampered randomly in different regions by applying a copy-paste attack to 5% of the watermarked image (9830 pixels in 384512 pixels image), and also compressed with different QF values. The results showed that the quality factor estimation process was highly accurate even under these attacks. From Table 5, the lowest correct detection rates were obtained for $QF = 50$, $QF = 60$, and $QF = 100$. Two other experiments were performed with the test image subjected to only the copy and paste attack and with the test image without any modification. The detected QFs achieved for both experiments were approximately 99, and fit well in the upper range of $QF \geq 90$. Similarly, the results of P_{de2} also showed the correct

Table 5. QF estimation for watermarked images (Copy and paste attack + JPEG compression)

Actual QF	**Estimated QF**	P_{de}	T	P_{de2}
100	98.60	72%	0.9	99.1%
95	95.00	100%		
90	90.14	98.6%		
85	84.83	97.9%	0.7	99.3%
80	79.95	99.6%		
75	75.22	99.1%	0.5	99.2%
70	69.87	99.5%		
65	64.46	98.7%		
60	61.54	63.9%		
55	54.93	96.6%		
50	53.32	20.4%		

detection rates in the three ranges were highly accurate with an overall average of 99%. As such, the threshold can be adapted into the three QF ranges according to the estimated QF of each test image as described in Section 3.2.

3.4 Summary

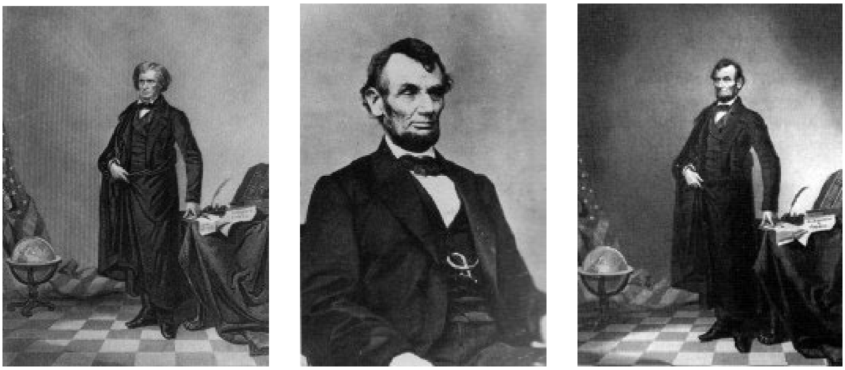
In this section, we presented the relationship between QF and threshold, and proposed a framework incorporating the generalised Benford's Law as an image forensics technique to accurately detect unknown JPEG compression levels in semi-fragile watermarked images. We discussed the limitations of using predetermined thresholds in semi-fragile watermarking algorithm. In our improved semi-fragile watermarking method, the test image was first analysed to detect its previously unknown quality factor for JPEG compression by using generalised Benford's Law model, before proceeding with the semi-fragile authentication process. The results showed that QFs can be accurately detected for most unknown JPEG compressions. In particular, the average QF detection rate was as high as 96% for watermarked images compressed with QFs between $95-65$, and 99% when the image was subjected to tampering of 5% pixels of the image and compressed with QFs between $95-65$. The threshold was adapted into three specific ranges according to the estimated QF of each test image.

4 Image Forensics

Recently, an interest has developed in identifying reliable techniques that are capable of accurately proving the authenticity of an image, without the requirement of actively inserting a digital watermark or signature into the data. Whilst the watermarking schemes discussed in section 2 have been shown to be useful for protecting the integrity of the image, there always exists the underlying risk that the watermark data might be forcibly or accidentally removed. When this happens, the image is effectively stripped of its identity, and its integrity is extremely difficult to prove. Forensic techniques aspire to achieve similar objectives but do not rely on the strength of embedded data. Instead, the ambition is to prove the authenticity of an image based solely on the data provided.

The two main areas of focus within the field of image forensics are *camera identification* and *forgery detection*. Camera identification is the task of successfully linking suspect images to the source camera that captured the image, in order to provide evidence that the origin of the image is as claimed. For example, a claim might be made by Person A that they captured an image of a compelling real-life event in order to gain acclamation. However, it is possible that Person B makes the same claim, and suggests that it was taken from their camera which happens to be a different make or model. A scrutinised forensic evaluation would attempt to review the properties of both cameras' image acquisition process, and determine the correct source for the image. This exercise might be relatively trivial if both camera's are vastly different, but what happens if Person A and Person B both own the same make and model camera? The forensic expert must then locate features in the image acquisition process of both cameras that differ. It should be possible to locate this feature within the data of the image in question, and therefore conclude which device captured the image. Forgery detection, on the other hand, is the practise of ensuring that the content of the image has not been manipulated. One of the most typical forms of content manipulation is *splicing*, which involves removing content from one image and overwriting it with something similar from another image to form a composite. This type of modification dates back over 150 years; a famous example of which is the Abraham Lincoln portrait [48]. In this example, a portrait of John Calhoun was manipulated such that it appeared as if the portrait was of Abraham Lincoln. In fact, Lincoln never posed for the portrait, and the image was actually constructed by flipping and resizing Lincoln's head from a head-shot photograph taken by Mathew Brady such that it resembled the same proportions as the Calhoun portrait. Calhoun's face was then replaced by Lincoln's face to produce a composite image.

Part of the challenge for image forensics lies in the fact that it is rarely immediately obvious whether or not an image has been manipulated. If a good job has been made of doctoring the image, it will look completely legitimate in plain sight. Therefore a distinction must be made between *clean* images that have not been altered in any way, and *dirty* images that are no longer true to their original form. Clean images are typically those that have come directly from the source that created them, without having been subjected to any external post-processing. However, it is often extremely rare to locate an image as clean as this, as most photographers (even at an amateur level) are likely to enhance their images through image editing software to provide better visual clarity, even though the content itself will remain true. To what extent such enhancements constitute a manipulation remains uncertain at this point. For this chapter we define clean images as those extracted straight from the camera that captured them, and dirty images as any image that has been manipulated in any way, including enhancements. By classifying the images according



(a) John Calhoun portrait. (b) Lincoln head-shot. (c) Composite image.

Fig. 14. The Lincoln composite

to these terms, we are effectively suggesting that a clean image accurately represents the exact scene from which the image was captured, and also inherits only the characteristic properties marked into the image data by camera processing.

Figure 15 presents a diagram of the two main areas of research in image forensics. The diagram shows that a given image can either be captured by a digital camera (in which case, the task is to identify anomalies in the camera processes that are also found in the image data), or the image will have been edited by software (in which case, anomalies are found in the image data that reflect manipulations). A suspect image is usually intercepted after either or both of these processes have been instantiated, and it is the job of the forensic specialist to establish the origin of the image.

In this section, we begin by explaining the most significant techniques that have been developed for the camera identification and forgery detection areas. In Section 4.2 we focus purely on camera identification, and present a novel approach to identifying anomalies within image data, before discussing the results of this work in Section 4.3. We then provide a concluding summary in Section 4.4.

4.1 Literature Survey

Camera Identification

One of the earliest reported approaches for digital camera identification characterised the imaging sensor from the device [49]. The imaging sensor is arguably the most important component of the image acquisition process, as it captures the light intensity of the scene on a pixel-by-pixel basis, and converts it into an electrical signal. From here, the signal will pass through a Colour Filter Array (CFA), which interpolates the colours for each pixel and the image is effectively born. However, it is possible that the imaging sensor operates with an element of noise, caused by *hot* or *dead* pixels. Errors such as this can often be seen in the final image, even if

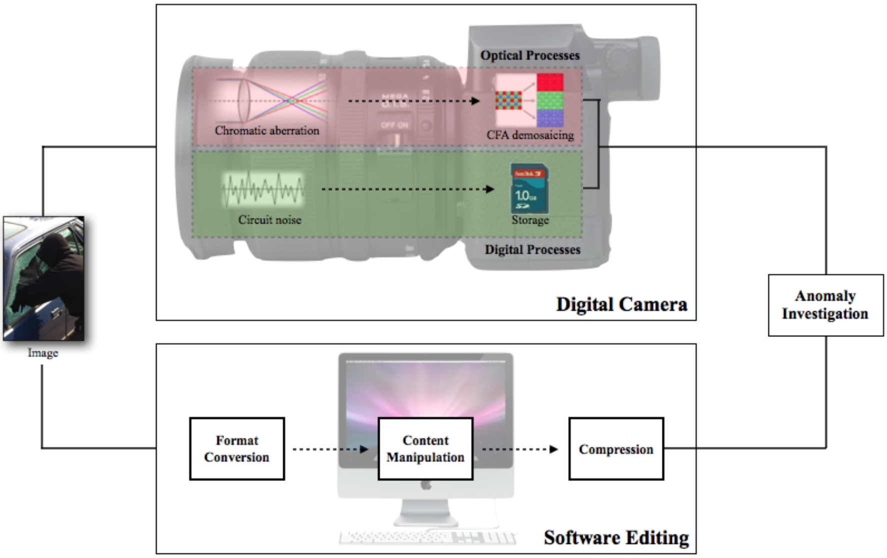


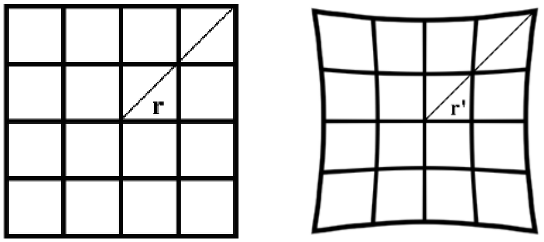
Fig. 15. Examples of camera-based and software-based image manipulations

the image has been lossy compressed. As the error is likely to be slightly different for several devices, the technique is useful for reliably linking images to the source sensor - and therefore the source camera - that captured the image. However, most modern cameras are able to detect deficiencies in the processing such as this, and often remove the hot or dead pixels altogether. As the scheme relies on the existence of such pixels, it can only be targeted towards cameras that do not correct errors such as these.

In 2006, research by K. S. Choi et al. led to the discovery that the camera lens produces aberrations in images, due to the design and manufacturing process [50]. Lens radial distortion was found to be quite a common property for inexpensive wide-angle lenses, and it causes straight lines to render as curved lines on the camera sensor. A camera lens has various focal lengths and magnifications in different areas, and when the transverse magnification M_T increases with the off-axis image distance r, a *barrel* distortion presents itself, as shown in Fig 16.

By calculating the precise radial distortion for a given device, as well as the relative radial distortion witnessed from a suspect image, it is possible to infer whether or not the image originated from that device. The technique acts as an excellent feature for providing a successful classification, but is likely to be insufficient in isolation. Instead, this feature will need to be used in conjunction with several other similar techniques in order to make a more informed and justified classification.

Arguably the most prominent research in the camera identification area, is that proposed by J. Lukáš et al. [20, 51], and verified by N. Khanna et al. in 2009 [22]. The technique relies on *pattern noise*, which is a deterministic component



(a) An undistorted rectangular grid.

(b) Grid with barrel distortion.

Fig. 16. Barrel distortion of a rectangular grid [50]

that remains consistent for all images that the sensor captures. Pattern noise can be sub-divided into two categories: *fixed pattern noise (FPN)* and *photo-response non-uniformity noise (PRNU)*. The FPN is an additive noise that is supressed to varying standards by many camera manufacturers, and is relative to exposure and temperature [20]. For these reasons, it is not reliable for camera identification purposes as it is inconsistent. PRNU, on the other hand, is a multiplicative noise and contains a property refered to as *pixel non-uniformity (PNU)*, which is defined as the sensitivity differences to light at each pixel. The PNU is a direct result of the manufacturing process and is therefore not influenced by exposure and light. Indeed, the PNU noise remains the same for each image that is taken, meaning this component is extremely useful for determining the source camera that captured an image. To complete the classification, a reference pattern for the camera must first be identified. This is achieved by using a denoising filter F and averaging the noise residuals $n^{(k)}$ from multiple images $p^{(k)}$.

$$n^{(k)} = p^{(k)} - F(p^{(k)}). \tag{8}$$

Selected regions from image p are then checked for the existence of the pattern noise from camera C by calculating the correlation P_C between the noise residual $n = p - F(p)$ with the camera reference pattern P_C, as shown in Equation (9).

$$P_C(p) = corr(n, P_C) = \frac{(n - \overline{n}) \cdot (P_C - \overline{P}_C)}{\|n - \overline{n}\| \|P_C - \overline{P}_C\|}. \tag{9}$$

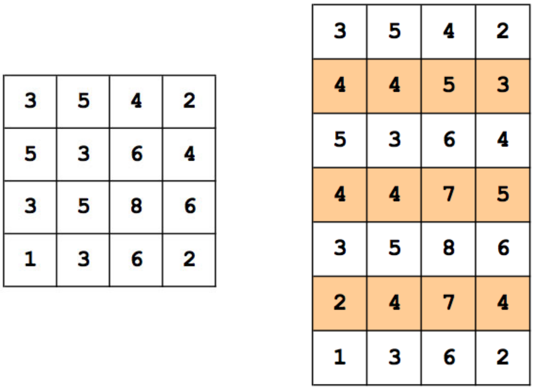
where the bar above a symbol denotes the mean value [52].

The pattern noise obtained from a suspect image can now be compared with the pattern noise obtained from the device itself. If the correlation is identical, then there can be little doubt that the image originated from the device, as the chances of two camera's producing the same pattern noise are extremely remote.

Forgery Detection

Significant progress has also been made in the forgery detection research area for authenticating image content, such as the splicing example discussed at the beginning of this section. In fact, the sensor pattern noise technique introduced by J. Lukáš et al. can easily be adapted to authenticate images. As the complete pattern noise exists for every pixel in an image, a manipulated image can be derived when the pattern noise is not present at a particular region of interest. It is important to note, however, that the PRNU noise will not be present in highly saturated areas of clean images, and is also highly supressed in dark areas, as the noise is multiplicative. Therefore, a region that does not contain the pattern noise should be checked to ensure that neither of these two properties hold true before classifying the image as tampered. Further details of how this can be achieved statistically are discussed in [52].

Whilst much research is concentrated on calculating anomalies in the image acquisition process of digital cameras, and then locating marks of those anomalies in the image data, some researchers have taken a different approach and are considering how ”fingerprints” of software manipulation also exist in the image data. The most prolifent work from this angle is lead by H. Farid’s research group at Dartmouth college. Specifically, they have reviewed how certain image manipulation operations such as resizing, alter the underlying pattern of pixels in a distinct way [53]. When creating a composite from two or more images, parts of an image are often enlarged (up-sampled), and when this happens, extra pixels are formed. Figure 17 shows what happens when a small 4x4 pixel patch is stretched to produce a 4x7 pixel patch. The numbers contained within the original 4x4 block shown in 17(a), correspond to the brightness at each location. The highlighted rows in 17(b) indicate added information, which is calculated by averaging the values of the immediate neighbours.



3	5	4	2
5	3	6	4
3	5	8	6
1	3	6	2

(a) A 4x4 pixel patch.

3	5	4	2
4	4	5	3
5	3	6	4
4	4	7	5
3	5	8	6
2	4	7	4
1	3	6	2

(b) Extra pixels added when enlarging.

Fig. 17. Enlarging a 4x4 pixel patch [53]

When images are enlarged in this manner, there exists a perfect correlation between neighbouring pixels, which is a rare property to find in natural images. Therefore, whenever this property is detected by a forensics specialist, they can derive a probability that the image has been manipulated.

In other work, H. Farid's research group also found that composite images can also be identified by studying the light reflected into the subjects eyes. The positioning of white dots (caused by flash photography) indicate the direction of the light when the image was captured [53]. When images are spliced together, these issues are often overlooked. For a clean image, when several people all appear in the scene the correlation of the light direction will match almost exactly. However, when a person has been spliced into the image from another image, the direction of light on the subjects eyes will not match. By studying the light pattern, it is often a fairly trivial process to determine whether the image is genuine or not. Similarly, the author discusses how lighting observations can be applied more generally to images in [21]. In this work, the author explains how the light striking a surface is dependant on the position of the light source. An estimate of the direction of the light source can be derived from an image by reviewing a given object's 2-D surface contour, such as a human jawline and chin. The lighting of this object can ultimately be compared against that of other objects in the photo, and if there exists a mismatch in lighting direction, then the image is likely faked.

As described in this section, there have been significant advances made in the fields of camera identification and forgery detection in recent years. For the remainder of this chapter, we concentrate solely on the camera identification area, and present a novel technique for locating anomalies in image data.

4.2 Statistical Process Control

At present, much research for camera identification has been based around identifying anomalies in a camera's image acquisition process, and then hoping to find a "fingerprint" of these properties in the image data. Whilst this research has produced some promising results, it is never easy to generalise the image acquisition process for a wide range of digital cameras, as each process can be quite vastly different from manufacturer to manufacturer. Instead, it is desirable to create a model such that the anomalies for any type of digital camera can be quickly and easily identified. In this section, we discuss how Statistical Process Control (SPC) can be used for such a purpose, and how it fits into the camera identification model as shown in Figure 18.

Introducing Statistical Process Control

The theory of SPC was developed in the late 1920's by Dr. Walter Shewhart, a physicist and statistician at the AT&T Bell Laboratories, USA, and was designed in an effort to acknowledge quality control and improvement for the manufacture of goods [54]. Shewhart recognised that products built to a high standard with good quality components, often produced better results in the field. In 1931, Shewhart

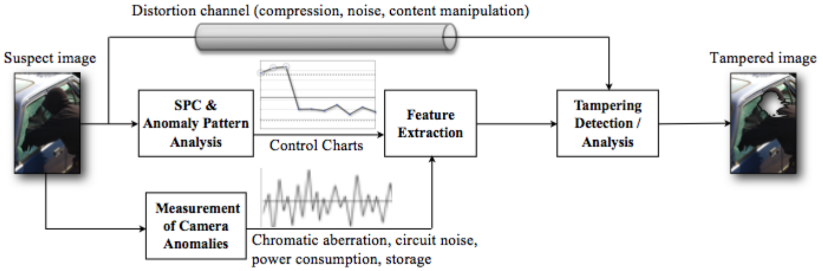

Distortion channel (compression, noise, content manipulation)
Suspect image
SPC & Anomaly Pattern Analysis
Control Charts
Feature Extraction
Tampering Detection / Analysis
Tampered image
Measurement of Camera Anomalies
Chromatic aberration, circuit noise, power consumption, storage

Fig. 18. SPC alternative to the camera identification process

released a study of his work that outlined a statistical approach for detecting the degree of control within processes over time [55]. The aim of Shewhart's work was to eliminate unexpected sources of variation that cause the process to operate with less accuracy. These variations were refered to as *special-cause*, and are caused by irregular events or circumstances that have an obvious impact on the process. Any variation that could be explained, was refered to as *common-cause* variation. In a perfect world, each measurement taken over time would produce the exact same result. However, in the real world, there are often external influences that affect the performance of processes.

SPC has been successfully applied to many areas of manufacture to maximise the efficiency of production processes to deliver high quality products. It was first applied to automobile manufacture by several Japanese manufacturers, and such was the success of its use on the end product, the Ford Motor Company soon followed [56]. It has since also been applied to industrial applications such as the pulp and paper industry [57–59], and has even been considered for improving healthcare processes [60].

The use of SPC can easily be adapted for use in image processing by substituting the measurements with image data taken from a digital camera. The quality of the complete image acquisition process for the camera can be infered, and a study of any widely varying images can lead to the discovery of a unique feature of the device that can act as a ”fingerprint” for camera identification.

Control Charts

A key tool of SPC for reviewing process variation, are *control charts*, which are used to graphically display the variation shifts from each measurement. Typically, two control charts are required to expose the data obtained from the process in its entirety: one to display the shifts in the process mean, and one to display the shifts or changes in the amount of process availability [55]. There are several types of control chart, each calculated in different ways, and chosen according to the best fit for the application. Our initial work is focused on *individuals* charts (commonly referred to as X charts) to display the process mean, and *moving range* charts (referred to as R_m charts) to display an estimate of the common-cause variability of the process. X

and R_m charts are suited to instances where individual measurements are obtained, and will therefore be useful for representing data collected from multiple images taken by several digital cameras.

Both control charts are comprised of a *centreline* CL (which is the mean value obtained from all measurements), an *upper control limit* UCL, and a *lower control limit* LCL, as well as the physical data obtained from each measurement X. The UCL and LCL are calculated at around ± 3 standard deviations above and below CL, respectively, to obtain results with a false-positive margin of approximately 0.27% [55]. If any measurement falls outside of these control limits, then the measurement is considered *out-of-control*.

Constructing the Control Charts

The construction of the control charts is based completely on the data measurements obtained for X. Traditionally, the R_m chart is plotted first, as these charts provide information on the overall process variability. The first step is to calculate the differences between neighbouring values in X to produce R_m. The CL is simply the mean of all measurements of R_m, denoted as $\overline{R_m}$, and is therefore calculated according to Equation (10).

$$\overline{R}_m = \frac{\sum_{i=1}^{k} R_{mi}}{k}. \tag{10}$$

where k refers to the total number of elements in R_m. A table of constants (Table 6) is then used to calculate the UCL and LCL control limits.

Table 6. Constants for Calculating Control Limits [54]

Observations in Sample	d_2	A_2	D_3	D_4
2	1.128	1.880	0	3.267
3	1.693	1.023	0	2.575
4	2.059	0.729	0	2.282
5	2.326	0.577	0	2.115
6	2.534	0.483	0	2.004
7	2.704	0.419	0.076	1.924
8	2.847	0.373	0.136	1.864
9	2.970	0.337	0.184	1.816
10	3.078	0.308	0.223	1.777
15	3.472	0.223	0.348	1.652
20	3.735	0.180	0.414	1.586

The UCL and LCL values are calculated by using Equation (11), where D_3 and D_4 are obtained when *observations in sample* $n = 2$.

$$\begin{aligned} UCL_{R_m} &= D_4\overline{R}_m \\ LCL_{R_m} &= D_3\overline{R}_m. \end{aligned} \tag{11}$$

When constructing X charts, the CL refers to the mean value of all measurements in X, and is denoted as $\overline{X}$. The UCL and LCL values are calculated by adding or subtracting 3 standard deviations from this value, where an estimate of the standard deviation is obtained from Equation (12).

$$\hat{\sigma}_X = \frac{\overline{R}_m}{d_2}. \tag{12}$$

where d_2 is taken from the table of constants, again when *observations in sample* $n = 2$. The UCL and LCL control limits can then be calculated from Equation (13).

$$\begin{aligned} UCL_X &= \overline{X} + 3\hat{\sigma}_X \\ LCL_X &= \overline{X} - 3\hat{\sigma}_X. \end{aligned} \tag{13}$$

Using Statistical Process Control for Image Forensics

SPC can be used to identify anomalies in the image acquisition process of a digital camera by collecting a series of identical images from the device, and using the mean pixel data (across each colour plane) as the measurements for X. As mentioned previously, the aspiration is that the anomaly investigation process will lead to the uncovering of unique ”fingerprints” for camera identification. According to *Flickr* (a popular image and video hosting site), the current most commonly used cameraphone device on their website is the Apple iPhone.

For our experiments, we therefore use four Apple iPhone 3G devices as primary devices. As cameraphones by definition are not primarily engineered for photography, inexpensive components are typically used, meaning the expectancy of witnessing a poor statistical control is increased. The Apple iPhone devices contain a 2 MegaPixel CMOS sensor and do not process any user settings or a zoom of any kind, meaning the exposure settings, focus, ISO settings, aperture, etc. are all automatically defined, if indeed they exist at all. The results obtained from the Apple iPhone 3G devices are later compared with those obtained from similar devices such as a Sony Ericsson W810i, and two Nokia N97 devices.

When acquiring the image data, it is important to nullify any environmental issues that could affect the data negatively. If the external conditions remain uncontrolled, it is likely that each device produces quite contrasting results, not necessarily because their image acquisition processes are different, but because, for example, the temperature or lighting conditions suddenly change. In our initial work in [61], the images are acquired from a room that is not subjected to outside lighting as this constantly changes. Instead, the test scene was lit via fluorescent lighting. In addition, the room was air-conditioned to a constant temperature so as to reduce the influence of temperature changes on the image acquisition process. It is worth noting, however, that the SPC model can be applied when these external influences do exist, so long as they are taken into consideration when reviewing the data. For instance, if the temperature increases as each image is taken, this will have an affect on the image acquisition process. As such, if the process becomes less and less controlled over time, then the temperature is a likely cause.

The scene itself comprises a white bowl of colourful confectionaries (Figure 19, where the bright colours maximise the load on the CFA. A location reference point is then set up to determine the position for each device, and a series of 10 images are collected one after the other for each device.

Fig. 19. Example image obtained from test scene

4.3 Results and Evaluation

In this section, we present the results from our initial implementation of SPC for image forensics, as reported in [61]. We also evaluate the significance of these results, and consider a refined implementation of a similar model based on images captured from a controlled environment. Comparisons are then made between both models, and a critique of how SPC can aid camera identification is provided.

The mean pixel values obtained from 10 images for all four Apple iPhone 3G devices is shown below in Table 7. By taking the 10 values for each device as X, control charts can be plotted to display the degree of control about the process mean $\overline{X}$. First, the R_m values are determined by calculating the difference between neighbouring values of X. Table 8 shows the X and R_m data for iPhone A.

Using this data, the CL for the R_m control chart is calculated as 1.141. The table of constants from Table 6 is used, where $n = 2$ to obtain $D_3 = 0$ and $D_4 = 3.27$. Subsequently, the UCL and LCL are then calculated as follows:

$$\begin{aligned} UCL_{R_m} &= D_4\overline{R}_m = (3.27)(1.141) = 3.729 \\ LCL_{R_m} &= D_3\overline{R}_m = (0.0)(0.294) = 0.0 \end{aligned} \tag{14}$$

Table 7. Mean pixel values obtained for all Apple iPhone 3G devices [61]

Shot No.	iPhone A	iPhone B	iPhone C	iPhone D
1	110.2097	105.4888	104.6182	104.4518
2	109.6045	106.5222	104.6503	97.4483
3	109.1334	106.7678	105.4275	97.1251
4	109.1161	98.0294	105.4614	97.0542
5	109.2108	98.064	105.676	97.0346
6	101.3616	97.7303	105.1278	97.0085
7	101.7246	98.9264	105.4571	97.4346
8	101.2875	96.9707	105.5588	97.4245
9	101.2724	98.4508	105.5459	97.0113
10	101.6922	97.4999	105.4956	97.0198

Table 8. X and R_m values obtained for iPhone A

Shot No.	X	R_m
1	110.2097	
2	109.6045	0.605
3	109.1334	0.471
4	109.1161	0.017
5	109.2108	0.095
6	101.3616	7.849
7	101.7246	0.363
8	101.2875	0.437
9	101.2724	0.015
10	101.6922	0.420

Similarly, the CL for the X chart, $\overline{X}$ is calculated as 105.461. The UCL and LCL control limits are calculated according to Equation (15).

$$\begin{aligned} UCL_X &= \overline{X} + 3\hat{\sigma}_X = 105.461 + (3)(1.011) = 108.497 \\ LCL_X &= \overline{X} - 3\hat{\sigma}_X = 105.461 - (3)(1.011) = 102.426. \end{aligned} \tag{15}$$

The final control charts for iPhone A are shown in Figure 20, where circled nodes indicate that the measurements are out-of-control as they fall outside the control limits.

The X chart (top) shows that every measurement taken from iPhone A is out-of-control. This indicates that the complete image acquisition process is statistically unstable. The R_m chart shows that sample 5 is out-of-control. When mapped to the X chart, this corresponds to the 5^{th} and 6^{th} images. The values of these two images are quite vastly different. The value for image 5 is 109.2108 compared to image 6 which yields the value 101.3616. By reviewing these two images further, it is possible to see a significant change in brightness between these images.

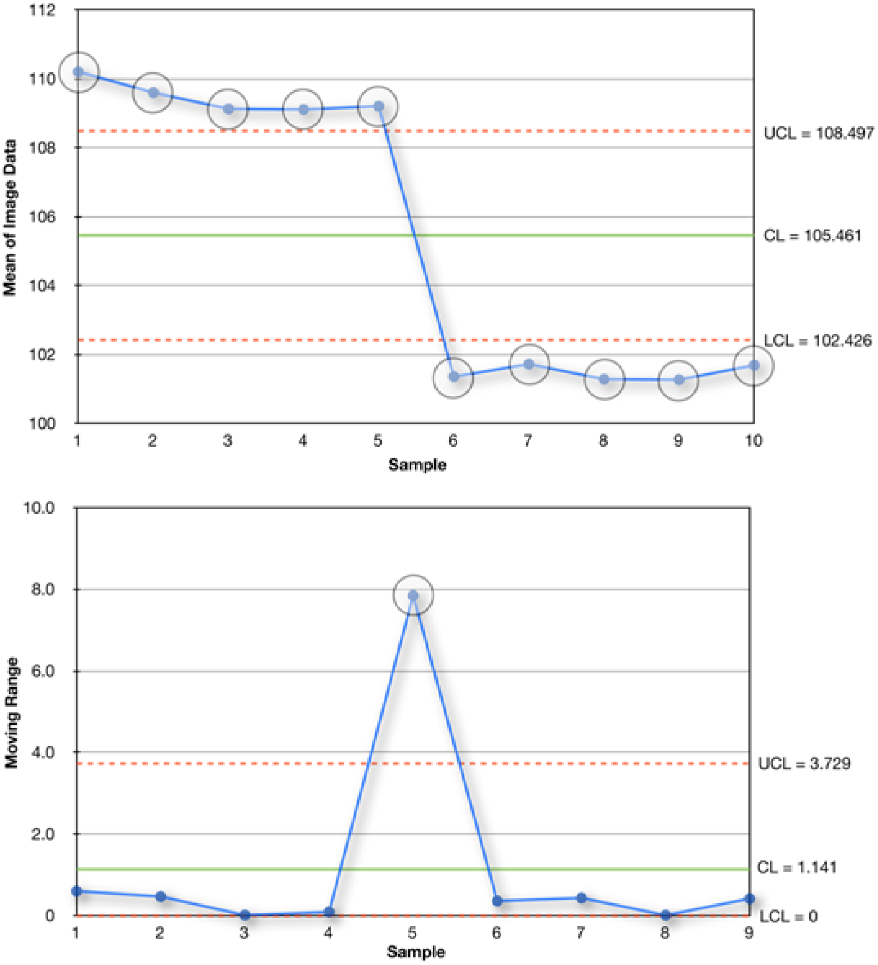


Fig. 20. X and R_m control charts for iPhone A

It is now worth evaluating how the results for iPhone A compare with the data obtained from the other iPhone devices. The control charts for iPhone B are shown in Figure 21.

Whilst the X chart is undoubtedly far more controlled than the X chart for iPhone A, the same significant drop in values can be seen, this time between image 3 and image 4. Again, when reviewing these two images, a large shift in brightness is observed.

Figure 22 illustrates the X and R_m control charts for iPhone C. For this device, there appears to be no significant shift in measurements, and in fact, the two

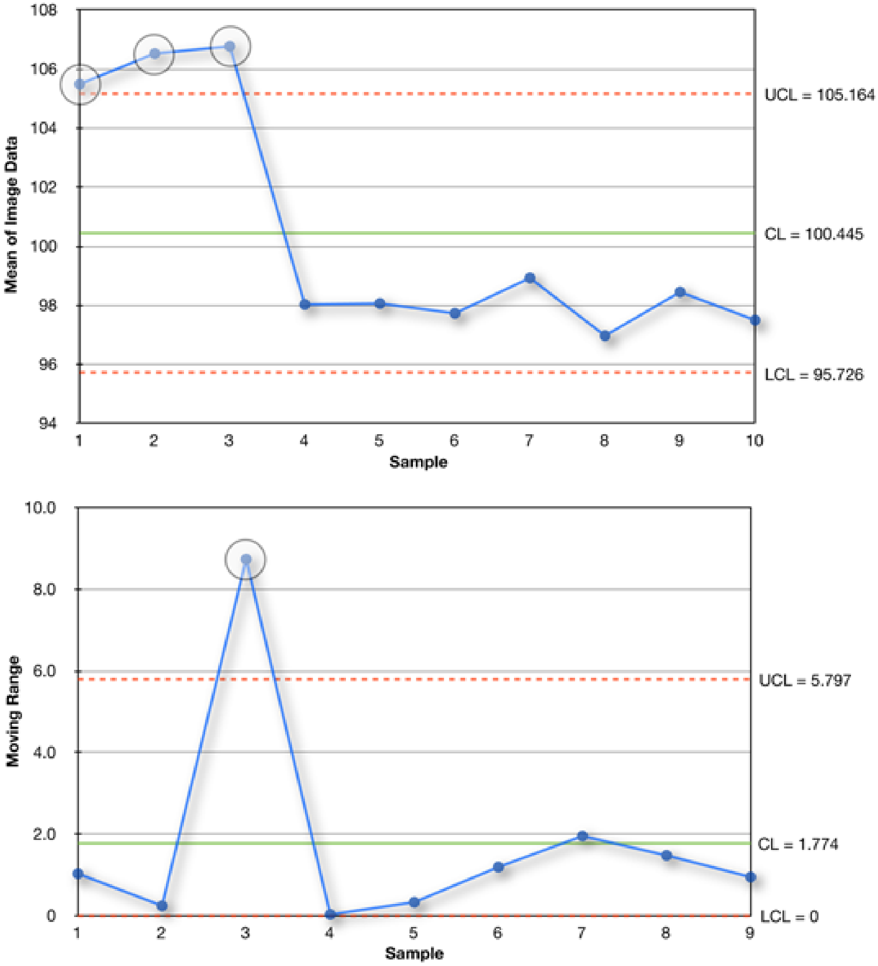


Fig. 21. X and R_m control charts for iPhone B

out-of-control measurements are only marginally outside the control limits. This indicates that this device was far more controlled than the previous two, at least for the 10 observations reviewed.

To complete the exercise for the iPhone devices, the X and R_m control charts are plotted for iPhone D, as shown in Figure 23. The same property of significant shifts in measurement values that was observed for iPhone A and iPhone B, also exists for iPhone D between the first and second images. Again, when reviewing both these images, a change in brightness value can be observed.

For comparison purposes, the same experiment was performed with the Sony Ericsson W810i cameraphone and standalone Samsung NV3 camera. We would

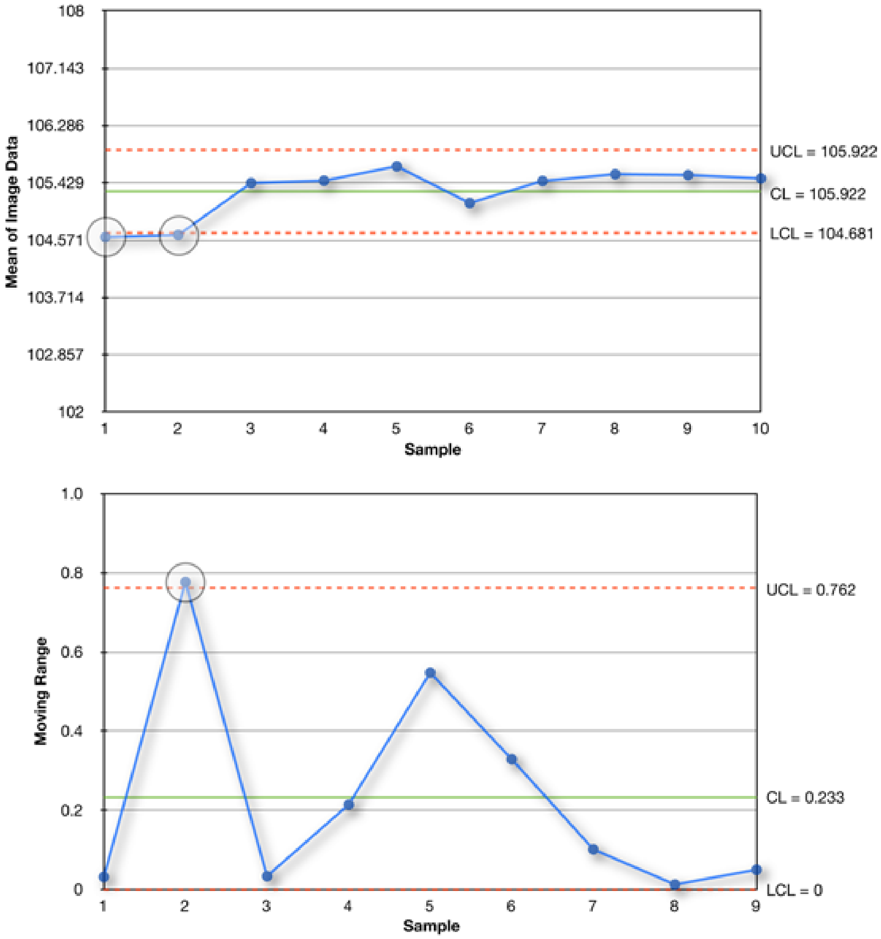


Fig. 22. X and R_m control charts for iPhone C

expect that the quality of the image acquisition process should be greatly improved for the Samsung NV3, as it is likely to use higher quality components, and more care is likely to have been taken to ensure errors are corrected in the pixel data. The Sony Ericsson W810i on the hand should be more comparable to the iPhone 3G devices, and if it does not posess the same brightness problem, it may be more controlled - although not as controlled as the Samsung NV3 is expected to be. The control charts for the Sony Ericsson W810i cameraphone are shown in Figure 24.

This device contains no out-of-control measurements, meaning that the image acquisition process is far more controlled compared to that of the iPhone 3G devices. The Samsung NV3 device is even more controlled, with the difference between the highest value measurement (132.04) and the lowest value measurement (131.89) only 0.15.

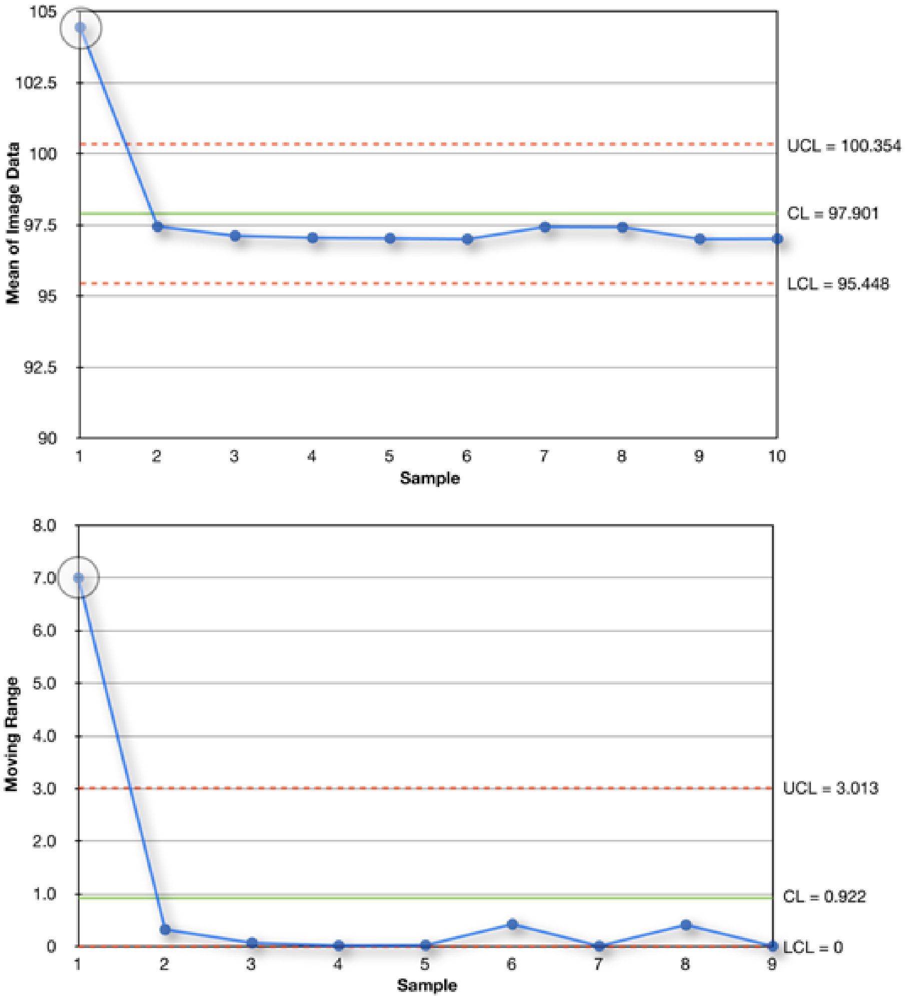


Fig. 23. X and R_m control charts for iPhone D.

As we only have access to one Sony Ericsson W810i and one Samsung NV3, we cannot collect enough data to make an informed review of any errors found within the image data. However, it is clear that the SPC model is reacting to the quality of each device.

Based on these observations it is obvious that there is some aspect of the iPhone image acquisition process that is affecting the brightness. The measurements taken before and after the decrease in data values are actually quite consistent, but the change in brightness is so vast that it renders much of the process out-of-control. It also appears from the SPC experiment as though the change in brightness is time dependant. Whilst iPhone C did not display any signs of this characteristic,

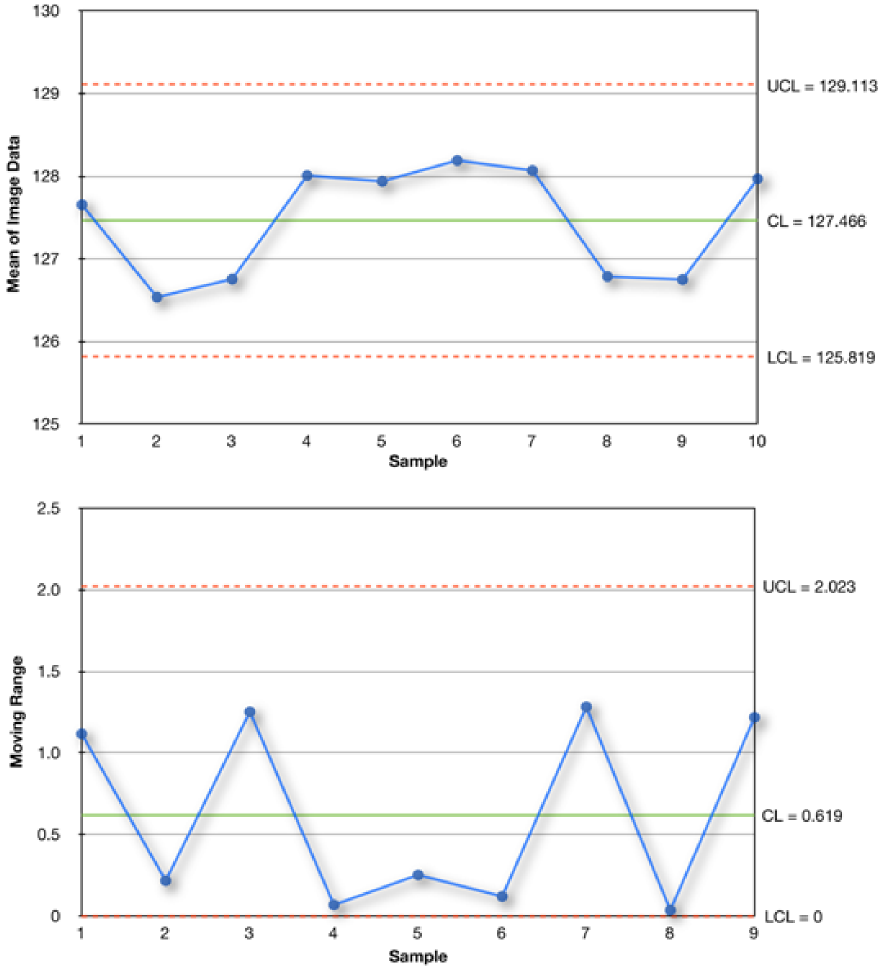


Fig. 24. X and R_m control charts for a Sony Ericsson W810i cameraphone

it might have showed itself if we took more than 10 measurements. The initial work has therefore highlighted a feature of the image acquisition process for iPhone 3G devices, that could be analysed in more detail to potentially create a unique "fingerprint" for identifying images captured from these devices.

In our most recent work, we have studied the effect that the lighting conditions have on the image acquisition process. As fluorescent lighting is known to flicker, the overall intensity of light could vary from shot to shot depending from when the image was captured. The environment was therefore adapted such that the fluorescent lighting was replaced by a flicker-free task lamp. The bulb itself emitted a true representation of daylight based on the Spectral Power Distribution (SPD). It is important to simulate daylight conditions as this ensures the digital camera's are still

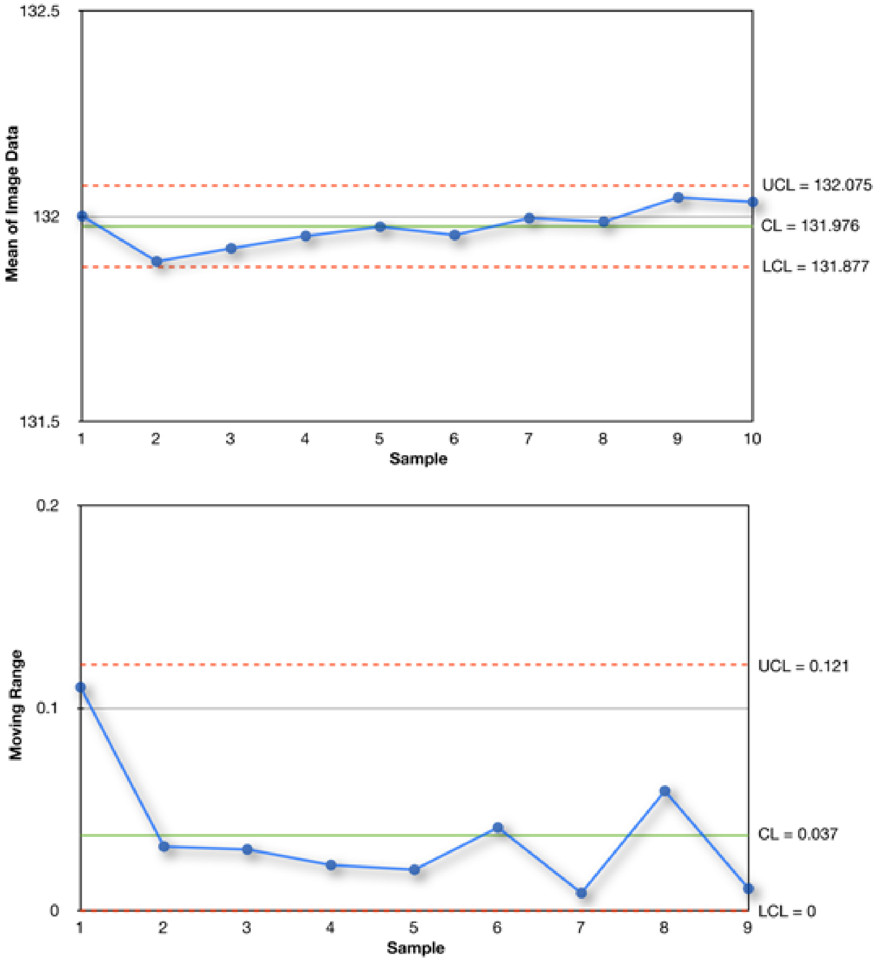


Fig. 25. X and R_m control charts for a Samsung NV3

processing the images based on real-world lighting. Some lamps will emit light that is faded or distorted, which would not provide useful results for our experiments.

The scene itself was also modified such that we use a light tent to ensure no external light sources filter onto the object. The bowl of confectionaries was also replaced with an *X-Rite ColorChecker®* chart, as this comprises 24 carefully selected colour squares, that each represent real-world colours (i.e. skin, sky, and landscape tones). The chart is specially designed such that each colour is reflected just as it is in the real-world. The colours within the chart are also defined in terms of their exact RGB reference values which means it will be possible to review exactly how each device is interpreting the colours if necessary. Figure 26 shows a sample image that was captured from this revised environment.

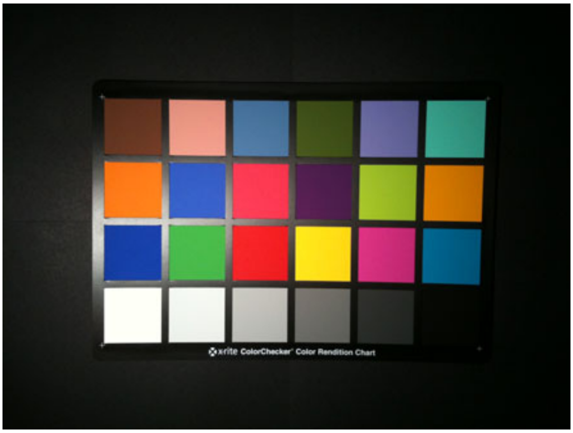


Fig. 26. Sample image taken from the modified test scene

Finally, the number of measurements taken from each device is increased from 10 to 30 to provide us with a more complete representation of the processing. The calculations involved for constructing the control charts, however, remain the same.

Figure 27 shows the X control chart obtained from iPhone A. Again, the same variation shifts that appeared in the earlier experiment can be noted. At each of these points, the brightness of the images shifted quite significantly. It can also be observed that the fluorescent lighting conditions from the first experiment was not the cause for the error scene in the camera processing, as the difference between the highest and lowest measurements for both experiments is approximately equal with that of this controlled experiment.

Whilst carrying out the experiment, an updated iPhone 3G model was released by Apple called the iPhone 3GS. The updated model carries a 3.2 MegaPixel camera, and allows the user to define the focal point of the image. To identify the significance of these improvements, we ran the SPC experiment on the new model. The results of which are expressed as an X chart in Figure 28.

This chart shows that the image acquisition process is far more controlled than that of its predecessor. Each of the 30 measurements taken are under perfect control, and there are no significant shifts in variation as seen for the iPhone 3G devices. A more diligent review of the new 'focus' setting on the iPhone 3GS shows that the exposure of the image is also defined when the focus is set. The exposure then remains the same until the camera is moved, or the conditions of the environment change drastically. This backs up our assumption that the brightness issues witnessed for the iPhone 3G devices are due to an exposure calculation error, which is why the brightness flicks between dark and bright over time.

The experiment was also performed against two Nokia N97 devices. The camera on the Nokia N97 devices contains a 5 MegaPixel resolution, and allows the user to define white balance settings, ISO settings, and also zoom. To form the most suitable comparison with the results obtained from the iPhone devices, the resolution was set

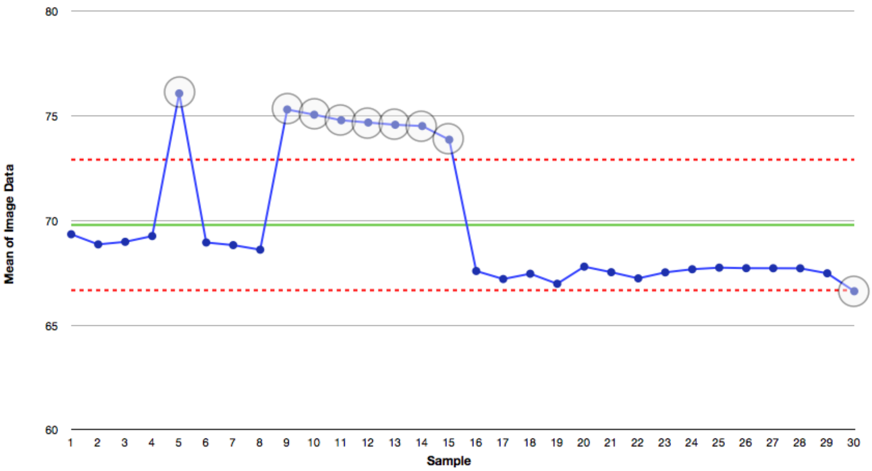


Fig. 27. X control chart acquired from iPhone A

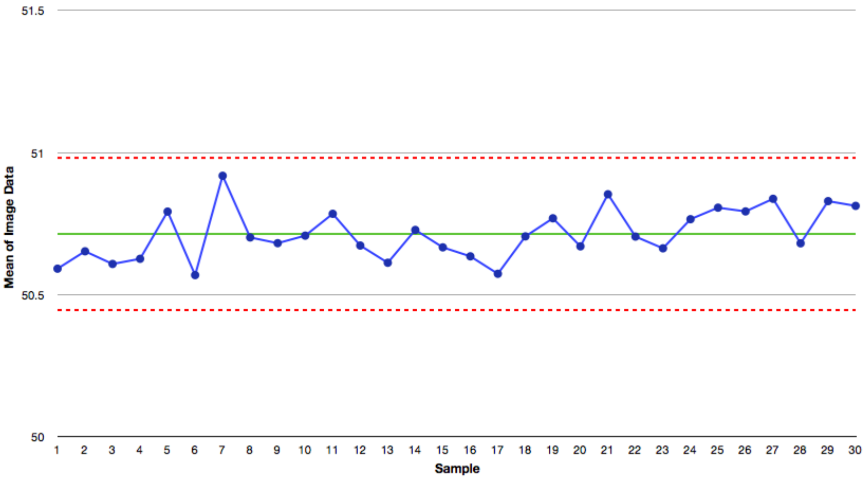


Fig. 28. X control chart acquired from an iPhone 3GS

to 2 MegaPixels, and all other settings were disabled where possible, or otherwise set to ”Automatic”. The X control charts for the two Nokia N97 devices are shown in Figures 29 and 30.

The control charts for the N97 devices show that each measurement - whilst more closely centred around $\overline{X}$ - is again not under complete statistical control. The second N97 device shows even less control than the first. By analysing

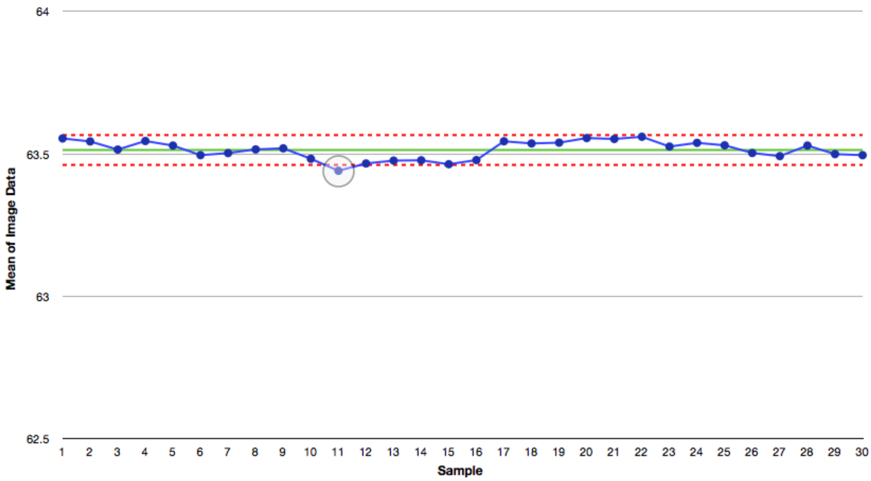


Fig. 29. X control chart acquired from Nokia N97 A

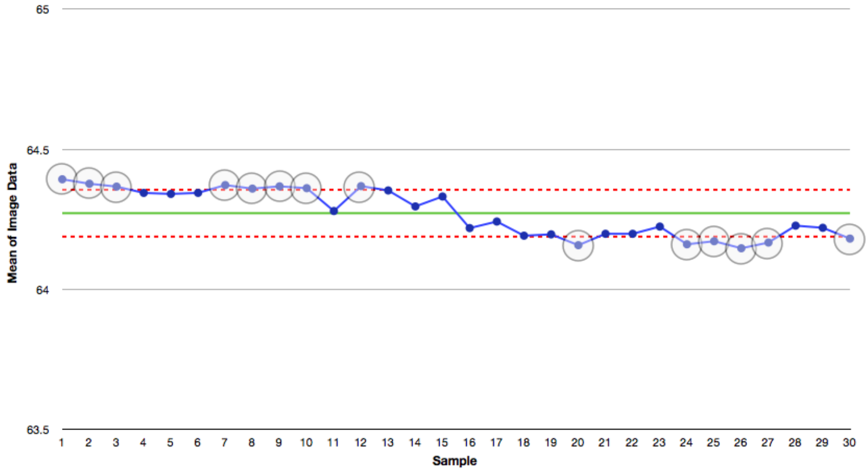


Fig. 30. X control chart acquired from Nokia N97 B

the out-of-control measurements in greater detail, (or indeed any contrasting measurements) and comparing them with the more controlled images, it is likely that the variation can be explained and a unique “fingerprint” uncovered.

4.4 Summary

In this section, we have introduced image forensics, and outlined the most prolifent research in the field - specifically, the latest research for camera identification and forgery detection have been introduced. In addition, we have demonstrated the benefits of using Statistical Process Control for analysing image data on a range of different digital cameras. Based on our initial research in [61], an anomaly in the camera processing elements was identified for iPhone 3G devices, whereby the brightness of the images was fluctuating. By analysing the latest iPhone 3GS model under the same conditions, we have been able to prove that the newer model does not contain this property, meaning there is promise for the reliable detection of images obtained from iPhone 3G devices, and images captured from the iPhone 3GS.

We have also proved through both experiments, that SPC is useful for modelling the overall control of the image acquisition process for a particular camera. Figure 31 illustrates the degree of variability obtained from the initial experiment in [61] for 6 devices. It is clear from this illustration that the iPhone 3G devices all operate with a similar degree of variation (approximately 21%). The Sony Ericsson W810i is far more controlled, and outputs a degree of variance of approximately 1%. The Samsung NV3 standalone digital camera is even further controlled, and offers a variation of only 0.5%.

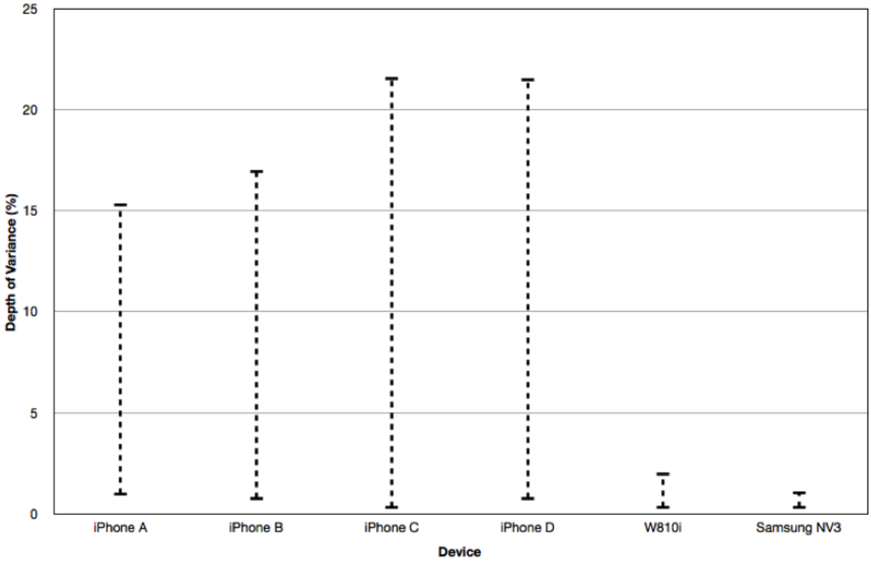


Fig. 31. Depth of variation for all devices

5 Conclusion and Future Work

Our future work in the image forensics domain will be concentrated on identifying more benefits of the SPC framework for camera identification. Further control charts can be examined, such as the *Exponentially Weighted Moving Average (EWMA)* chart, to determine whether there is an even more descriptive tool that can replace the X and R_m control charts. In SPC control charts, the formation of the measurements along CL can be used to derive common-cause and special-cause variation. Therefore, a scrutinised analysis of this content might be useful for isolating unique ”fingerprints” for digital cameras. Similarly, Pareto charts and Cause & Effect diagrams have also been proved to be useful for identifying the cause of variation for a range of processes. These techniques could be adapted for use in the image forensics domain for identifying anomalies in the image data.

References

[1] Vrusias, B., Tariq, M., Handy, C., et al.: Forensic Photography. In Technical Report, University of Surrey, Computing Dept. (2001)

[2] Ho, A.T.S.: Semi-fragile Watermarking and Authentication for Law Enforcement Applications. In: Innovative Computing, Information and Control (ICICIC 2007), pp. 286–286 (2007)

[3] Zheng, D., Liu, Y., Zhao, J., et al.: A survey of RST invariant image watermarking algorithms. ACM Computing Surveys 39(2) (2007)

[4] Cox, I.J., Miller, M.L., Bloom, J.A., et al.: Digital watermarking and Steganography, 2nd edn. Morgan Kaufmann, USA (2008), ISBN: 0123725852

[5] Li, C.T., Si, H.: Wavelet-based Fragile Watermarking Scheme for Image Authentication. Journal of Electronic Imaging 16, 130091–130099 (2007)

[6] Li, C.T., Yuan, Y.: Digital Watermarking Scheme Exploiting Non-deterministic Dependence for Image Authentication. Optical Engineering 45(12), 127001–127001 (2006)

[7] Ho, A.T.S., Zhu, X., Guan, Y.: Image content authentication using pinned sine transform. EURASIP Journal on Applied Signal Processing, 2174–2184 (2004)

[8] Lin, C.H., Su, T.S., Hsieh, W.S.: Semi-fragile watermarking Scheme for authentication of JPEG Images. Tamkang Journal of Science and Engineering 10(1), 57–66 (2007)

[9] Lin, H.Y.S., Liao, H.Y.M., Lu, C.H., et al.: Fragile watermarking for authenticating 3-D polygonal meshes. IEEE Trans. Multimedia 7(6), 997–1006 (2005)

[10] Rey, C., Dugelay, J.L.: A survey of watermarking algorithms for image authentication. EURASIP Journal on Applied Signal Processing (6), 613–621 (2002)

[11] Bartolini, F., Tefas, A., Barni, M., et al.: Image authentication techniques for surveillance applications. Proc. of IEEE 89(10), 1403–1418 (2001)

[12] Lin, C.Y., Chang, S.F.: Semi-Fragile Watermarking for Authenticating JPEG Visual Content. In: Proc. SPIE Security and Watermarking of Multimedia Contents II EI 2000, pp. 140–151 (2000)

[13] Lin, E.T., Podilchuk, C.I., Delp, J.: Detection of Image Alterations using semi-fragile watermarks. In: Proc. SPIE International Conference on Security and Watermarking of Multimedia Contents II, vol. 3971, pp. 152–163 (2000)

[14] Zou, D., Shi, Y.Q., Ni, Z., et al.: A Semi-Fragile Lossless Digital Watermarking Scheme Based on Integer Wavelet Transform. IEEE Trans. Circuits and Systems for Video Technology 16(10), 1294–1300 (2006)

[15] Zhu, X.Z., Ho, A.T.S., Marziliano, P.: A new semi-fragile image watermarking with robust tampering restoration using irregular sampling. Elsevier Signal Processing: Image Communication 22(5), 515–528 (2007)

[16] Zhu, Y., Li, C.T., Zhao, H.J.: Structural digital signature and semi-fragile fingerprinting for image authentication in wavelet domain. In: Proc. Third International Symposium on Information Assurance and Security, pp. 478–483 (2007)

[17] Yu, G.J., Lu, C.S., Liao, H.Y.M., et al.: Mean quantization blind watermarking for image authentication. In: Proc. IEEE International Conference on Image Processing, vol. 3, pp. 706–709 (2000)

[18] Kundur, D., Hatzinakos, D.: Digital watermarking for telltale tamper proofing and authentication. Proc. IEEE 87(7), 1167–1180 (1999)

[19] Fu, D., Shi, Y.Q., Su, Q.: A generalized Benford's law for JPEG coefficients and its applications in image forensics. In: Proc. SPIE Security, Steganography, and Watermarking of Multimedia Contents IX, vol. 6505 (2007)

[20] Lukáš, J., Fridrich, J., Goljan, M.: Digital Camera Identification From Sensor Pattern Noise. IEEE Trans. on Information Security and Forensics 1(2), 205–214 (2006)

[21] Johnson, M.K., Farid, H.: Detecting photographic composites of people. In: Shi, Y.Q., Kim, H.-J., Katzenbeisser, S. (eds.) IWDW 2007. LNCS, vol. 5041, pp. 19–33. Springer, Heidelberg (2008)

[22] Khanna, N., Mikkilineni, A.K., Delp, E.J.: Forensic Camera Classification: Verification of Sensor Pattern Noise Approach. Forensic Science Communications (FSC) 11(1) (2009)

[23] Fridrich, J.: Security of fragile authentication watermarks with localization. In: Proc. SPIE, Security and Watermarking of Multimedia Contents IV, vol. 4675, pp. 691–700 (2002)

[24] Wong, P.: A Watermark for Image Integrity and Ownership Verification. In: Proc. IS and T PIC (1998)

[25] Zhang, X., Wang, S.: Statistical Fragile Watermarking Capable of Locating Individual Tampered Pixels. IEEE Signal Processing Letters 14, 727–730 (2007)

[26] Zhang, X., Wang, S.: Fragile Watermarking With Error-Free Restoration Capability. IEEE Transactions on Multimedia 10, 1490–1499 (2008)

[27] He, H.J., Zhang, J.S., Chen, F.: Adjacent-block based statistical detection method for self-embedding watermarking techniques. IEEE Signal Processing 89(8), 1557–1566 (2009)

[28] Aslantas, V., Ozer, S., Ozturk, S.: Improving the performance of DCT-based fragile watermarking using intelligent optimization algorithms. Opt. Commun. 282(14), 2806–2817 (2009)

[29] Yeh, F., Lee, G.: Pyramid-structure-based reversible fragile watermarking. Optical Engineering 48(4), 470011–4700111 (2009)

[30] Ni, Z., Shi, Y.Q., Ansari, N., et al.: Robust Lossless Image Data Hiding Designed for Semi-Fragile Image Authentication. IEEE Trans. on Circuits and Systems for Video Technology (4), 497–509 (2008)

[31] Lin, E.T., Podilchuk, C.I., Delp, E.J.: Detection of image alterations using semi-fragile watermarks. In: Proc. of Security and Watermarking of Multimedia Contents, vol. 3971, pp. 152–163. SPIE, CA (2000)

[32] Lee, H., Rhee, K.: Reversible data embedding for tamper-proof watermarks. In: Innovative Computing, Information and Control (ICICIC 2006), vol. 3, pp. 487–490 (2006)

[33] Fridrich, J., Goljan, M.: Images with self-correcting capabilities. In: IEEE International Conference on Image Processing, vol. 3, pp. 792–796 (1999)

[34] Maeno, K., Sun, Q., Chang, S., et al.: New Semi-fragile image authentication watermarking techniques using random bias and nonuniform quantization. IEEE Trans. Multimedia 8(1), 32–45 (2006)

[35] Ding, K., He, C., Jiang, L.G., et al.: Wavelet-Based Semi-Fragile Watermarking with Tamper Detection. IEICE Transactions on Fundamentals of Electronics 88(3), 787–790 (2005)

[36] Zhao, X., Ho, A.T.S., Treharne, H., et al.: A Novel Semi-Fragile Image Watermarking, Authentication and Self-Restoration Technique Using the Slant Transform. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP 2007), vol. 1, pp. 283–286 (2007)

[37] Pratt, W., Chen, W.H., Welch, L.: Slant transform image coding. IEEE Trans. on Communications 22(8), 1075–1093 (1974)

[38] Zhu, X., Ho, A.T.S.: A slant transform watermarking for copyright protection of satellite images. In: Fourth Pacific Rim Conference on Multimedia, vol. 2, pp. 1178–1181 (2003)

[39] Hou, Z., Xu, N., Chen, H., et al.: Fast slant transform with sequency increment and its application in image compression. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics, vol. 7, pp. 4085–4089 (2004)

[40] Zhao, X., Ho, A.T.S., Shi, Y.Q.: Image forensics using generalised Benford's Law for accurate detection of unknown JPEG compression in watermarked images. In: 16th International Conference on Digital Signal Processing, DSP 2009 (2009), doi:10.1109/ICDSP.2009.5201261

[41] Benford, F.: The law of anomalous numbers. Proc. American Philosophical Society 78, 551–572 (1938)

[42] Hill, T.P.: The significant-Digit Phenomenon. American Mathematical Monthly 102, 322–327 (1995)

[43] Durtschi, C., Hillison, W., Pacini, C.: The effective use of Benfords Law to assist in detecting fraud in accounting data. Journal of Forensic Accounting v, 17–34 (2004)

[44] Nigrini, M.J.: Ive got your number. Journal of Accountancy (1999)

[45] Jolion, J.M.: Images and Benfords Law. Journal of Mathematical Imaging and Vision 14, 73–81 (2001)

[46] Perez-Gonzalez, F., Heileman, G.L., Abdallah, C.T.: Benford's Law in Image Processing. In: Proc. IEEE International Conference on Image Processing, vol. 1, pp. 405–408 (2007)

[47] Schaefer, G., Stich, M.: UCID - An Uncompressed Colour Image Database. In: Proc. SPIE, Storage and Retrieval Methods and Applications for Multimedia, pp. 472–480 (2004)

[48] Mitchell, M.J.: The Reconfigured Eye, pp. 204–208. MIT Press, Cambridge (1992), ISBN: 978-0262631600

[49] Gearadts, Z.J., Bijhold, J., Kieft, M., et al.: Methods for Identification of Images Acquired with Digital Cameras, vol. 4232, pp. 505–512. SPIE Press, CA (2001)

[50] Choi, K.S., Lam, E.Y., Wong, K.K.Y.: Source Camera Identification Using Footprints From Lens Aberration. In: Proc. SPIE Electronic Imaging, vol. 6069, pp. 172–179 (2006)

[51] Lukáš, J., Fridrich, J., Goljan, M.: Digital "Bullet Scratches" for Images. In: IEEE International Conference on Image Processing, vol. 3, pp. 65–68 (2005)

[52] Lukáš, J., Fridrich, J., Goljan, M.: Detecting Digital Image Forgeries using Sensor Pattern Noise. In: Proc. SPIE Electronic Imaging, vol. 6072, pp. 362–372 (2006)

[53] Farid, H.: Seeing Is Not Believing. IEEE Spectrum 46, 42–47 (2009)

[54] DeVor, R.E., Chang, T., Sutherland, J.: Statistical Quality Design and Control: Contemporary Concepts and Methods. Prentice Hall, Englewood Cliffs (1992), ISBN: 978-0023291807

[55] Shewhart, A.W.: Economic Control of Quality of Manufactured Product. American Society for Quality (1931), ISBN: 978-0873890762

[56] Ford Motor Company, Continuing Process Control and Process Capability Improvement. Manual Published by Statistical Methods Office (1984)

[57] Ho, A.T.S., Henriksson, C.: Improving Product Quality in a Pulp Mill Using Statistical Process Control (SPC). In: IEEE Canadian Conference on Electrical and Computer Engineering, pp. 1139–1143 (1993)

[58] Ho, A.T.S., Henriksson, C.: Improvement of Product Uniformity Using Statistical Process Control (SPC) in a BCTMP Mill. Pulp and Paper Canada 95, 37–40 (1994)

[59] Guillory, A.L.: Statistical Process Control in a Paper Mill. Chemical Engineering Progress 84, 52–57 (1988)

[60] Benneyan, J.C., Lloyd, R.C., Plsek, P.E.: Statistical Process Control as a Tool for Research and Healthcare Improvement. Qual. Saf. Health Care 12, 458–464 (2003)

[61] Bateman, P., Ho, A.T.S., Woodward, A.: Camera Identification using Statistical Process Control Techniques for Anomaly Detection. Accepted for International Conference on Information, Communications, and Signal Processing (2009)

Blind Measurement of Image Blur for Vision-Based Applications

Shiqian Wu[1], Shoulie Xie[1], and Weisi Lin[2]

[1] Institute for Infocomm Research
(A*STAR) Agency for Science, Technology and Research
1 Fusionopolis Way, #21-01 Connexis, Singapore 138632
{shiqian,slxie}@i2r.a-star.edu.sg
[2] School of Computer Engineering
Nanyang Technological University, Singapore 639798
wslin@ntu.edu.sg

Abstract. This chapter presents a novel metric for image quality assessment from a single image. The key idea is to estimate the point spread function (PSF) from the line spread function (LSF), whereas the LSF is constructed from edge information. It is proven that an edge point corresponds to the local maximal gradient in a blurred image, and therefore edges can be extracted from blurred images by conventional edge detectors. To achieve high accuracy, local Radon transform is implemented and a number of LSFs are extracted from each edge. The experimental results on a variety of synthetic and real blurred images validate the proposed method. To improve the system efficiency, a criterion for edge sharpness is further proposed and only the edge points from sharp edges are selected for extracting the LSF without using all edge information. The effects of nearby edges on the selected edge feature and the resultant LSF are analyzed, and two constrains are proposed to determine appropriate LSFs. The experimental results demonstrate the accuracy and efficiency of the proposed paradigm. This scheme has fast speed and can be served in blind image quality evaluation for real-time automatic machine-vision-based applications.

1 Introduction

Vision-based systems have been used in a wide spectrum of applications, ranging from product inspection, production control to vehicle navigation, target recognition and surveillance. To achieve consistently high performance under varying environments and accommodate uncertainties, one of the crucial steps is to evaluate the quality of the images acquired, based on which the usability of the images or the confidence of the decision-making is determined. As these systems are automatically operated, it is necessary to develop objective quality evaluations (Wolin et al. 1998) or called blind image quality assessments (Li 2002) which are reliable and repeatable from acquired images instead of human observation.

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 185–215.
springerlink.com

However, the determination of image quality is a formidable task, and no well accepted, standardized method of accomplishing this has been developed because "quality" of images is a subjective notion (Wolin et al. 1998, Li 2002, Keelan 2002). An image can be defined to be of good quality in one imaging system, but does not necessarily rank similarly in another one. From the vision application point of view, an image can be defined to be of good quality if it fulfils its intended task well. In this case, image quality is determined by the relevance of the information presented by the image to the task we seek to accomplish using the image. In other words, image quality is task-dependent (Keelan 2002).

From the perspective of blind image quality assessment, i.e., evaluating from one image, image quality can be characterized by a large number of attributes , e.g. contrast, brightness, noise variance, sharpness, radiometric resolution, point spread function (PSF), modulation and contrast transfer function (MTF, CTF), resolving power, etc. Some of these attributes are objective and absolute, and are subject to rigorous measurement. Others are highly dependent upon the intended use of the image. Generally, the specific metrics used to characterize an image are not necessarily the same for each application (Wolin et al. 1998, Li 2002, Keelan 2002). Li (Li 2002) exploited the appropriate mathematical tools to characterize most important aspects of image quality and proposed to appraise the image quality by three objective measures: 1) edge sharpness level; 2) random noise level and 3) structural noise level. It was indicated by Li (Li 2002) that the three measures jointly provide a heuristic approach of characterizing the most important aspects of visual quality, and edges are the most important features in images.

Sharpness is an important attribute of image quality. The lack of sharpness is mainly caused by the physical phenomena of blur and low-pass filtering during image processing. Frequently, the blur in real vision applications comes from two sources, i.e., out-of-focus blur and motion blur. Blur images inherently have less information than sharp images and lead to difficulty in image analysis and scene interpretation. Consequently, the imaging systems should have the functionality of blur detection.

Generally, digital imaging systems use active methods, for instance, with infrared or ultrasonic sensors to avoid blur. This affects the compactness and cost of cameras. Moreover, the active operation is not always possible in automatic vision systems. To satisfy the requirements of practical applications, several passive measure techniques, such as measurements based on gradient magnitude, histogram of local variance, high frequency components of Fourier Transform and so on have been proposed (Subbarao and Tyan 1995). These methods generally work on a sequence of images containing the identical content with different blur extents, and the sharpest image is then selected. Accordingly, these methods depend on image contents, and the criteria may vary from one image to another. In other words, these paradigms do not provide an explicit expression of blur measurements which can be applied to all images.

Some blind approaches based on one image have been developed, and Kundur & Hatzinakos (1996) provided a good survey in the related research. In this category of approaches, the most frequently used method is the frequency domain zeros (Rom 1975, Cannon 1976, Fabian & Malah (1991)). The idea is based on the fact that the frequency responses of some particular PSFs have regular zero crossings, which determine the types and extent of the PSF. The pattern of the zero crossings can be analyzed by using cepstrum (Rom 1975) or spectrum (Cannon 1976) to identify the type of blurs. A major drawback of this method is its high sensitivity to noise. More robust techniques (Fabian & Malah 1991, Chang et al. 1991) have been proposed to improve the accuracy of identifying the spectral nulls in the presence of noise. However, such methods cannot identify PSFs that do not possess any spectral nulls, such as the Gaussian PSF.

On the other hand, techniques working on spatial domain have received great attention because they are simple and have low computational requirements without complex Fourier transform. The general idea is to derive the PSFs using local image characteristics, such as point-source, line or edge information wherein edges are the frequently used information. Marziliano et al. (Marziliano et al. 2004) used only the vertical and horizontal edges for blur estimation. Kim et al. (Kim et al. 1998) proposed a method to identify PSF from vertical and horizontal edges as well as the 45^0 edges. Wu et al. (Wu et al. 2005, Wu et al. 2009) proposed to use edges in arbitrary directions. This alleviates the dependence of image content. Different from the ad hoc methods (Marziliano et al. 2004, Kim et al. 1998), some exact metrics have been proposed for PSF estimation (Kayargadde & Martens 1996, Wu et al. 2005, 2009).

In the meantime, several methods simultaneously incorporating blur identification with restoration have been reported, among which, the ARMA (autoregressive moving average) parameter estimation using maximum-likelihood (ML) method is the most popular one (Lagendijk et al. 1990a, 1990b). It is indicated by Lagendijk et al. that this paradigm can be effective when no information about the distortion (blurring and noise) is known. Normally, it is robust to noise, and can identify a variety of blurs including those that do not have zero crossings. Another technique under the same framework is the general cross-validation (GCV) method proposed by Reeves & Mersereau (Reeves & Mersereau 1992). It is demonstrated that the GCV method is superior over the ML one in the context of regularization parameter estimation, and is also more robust (Fortier et al. 1993). However, the performance of these methods largely depends on the determination of initial PSF, especially the order (support) of the PSF (Chen & Yap 2006). Moreover, these methods are iterative and intensive in computation.

This chapter presents the complete scheme of blind spatial-domain blur measurement based upon estimating the LSF in a single image. This work focuses on oft-encountered machine vision problems instead of astronomical (Luxen & Forstner 2002), remote sensing (Bones et al. 2000) or biological (Lehr et al. 1998) applications. The out-of-focus is our main concern, without considering the motion blur because the movement involved is usually slow in such applications. The blind blur measurement based on single image is presented in the following section. Implementing details are described in Section 3. Experimental results are given in Section 4. Some issues are further discussed in Section 5. A computational efficient method is presented in Section 6, followed by conclusion drawn in Section 7.

2 Metric for Blur Measurement

Considering the image degradation as a linear space-invariant model, a blurred image $g(x,y)$ can be expressed as follows (Kundur & Hatzinakos 1996):

$$g(x,y) = h(x,y) \otimes f(x,y) + n(x,y) \tag{1}$$

where $f(x,y)$ is the original image, $h(x,y)$ is the PSF, $n(x,y)$ is additive noise, and $\otimes$ represents the convolution. It has been indicated by Kundur & Hatzinakos (Kundur & Hatzinakos 1996) that this linear model can represent the degradation of images in many practical image processing applications.

In most cases, the out-of-focus blur can be modeled as a uniform disk with radius R:

$$h(x,y) = \begin{cases} 0 & \sqrt{x^2 + y^2} > R \\ 1/(\pi R^2) & \sqrt{x^2 + y^2} \le R \end{cases} \tag{2}$$

This is a simple geometric model with circular aperture ignoring diffraction in which the radius R reflects the degree of blur. It has been shown (Savakis & Trussell 1993, Sezan et al. 1991) that a more accurate and complex physical model does not perform significantly better.

Although the PSF is expressed in a simple form, the estimation is not easy because it is with two dimensions. In optics, the PSF can be determined by the image of a point source. Similarly, the line spread function (LSF) is also used to characterize a blurred line in an image. For simplicity, it is assumed hereinafter that the LSF is parallel to the horizontal (x-) axis. This does not affect the generality of the derivation that follows, because the PSF in Equation (2) is rotationally invariant, and the LSFs at different orientations are identical. The relation of the PSF and the LSF can be expressed as follows (Andrews & Hunt, 1977):

$$L(y) = \int_{-\infty}^{\infty} h(x,y)dx \tag{3}$$

where $L(y)$ denotes the LSF. Equation (3) illustrates that the PSF can be inferred from the LSF, while the LSF is simpler than the PSF because it is with one dimension.

In a typical image, there exist plenty of edges in heterogeneous space. Let $h_e(y)$ be a unit step edge parallel to the x-axis in a blurred image, and is represented by a unit step function $U(y)$:

$$h_e(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x',y')U(y-y')dx'dy' \tag{4}$$

Taking partial derivative of both sides with respect to y :

$$\begin{aligned}\frac{\partial h_e(y)}{\partial y} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x',y')\frac{\partial}{\partial y}U(y-y')dx'dy' \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x',y')\delta(y-y')dx'dy' \\ &= \int_{-\infty}^{\infty} h(x,y)dx = L(y)\end{aligned} \tag{5}$$

where $\delta(y-y')$ is the impulse function. Equation (5) reveals that the derivative of a blurred unit step edge is actually its correspondent LSF in the image. Since Equation (2) implies that the PSF of out-of-focus blur is circularly symmetric, Equation (5) is held for a unit step edge in any direction.

According to Equation (3), the LSF is the integral of the correspondent PSF in the direction of a line. Thus for the case of out-of-focus blur, its LSF can be expressed as follows:

$$L(y) = \int_{-\infty}^{\infty} h(x,y)dx = \int_{-\sqrt{R^2-y^2}}^{\sqrt{R^2-y^2}} \frac{1}{\pi R^2}dx = \frac{2}{\pi R^2}\sqrt{R^2-y^2} \tag{6}$$

The maximum of the LSF can be obtained from Equation (6):

$$L^{\max} = L(y)\,|_{y=0} = \frac{2}{\pi R} \tag{7}$$

Then, we have:

$$R = \frac{2}{\pi L^{\max}} \tag{8}$$

This is the metric to estimate the blur parameter *R* from an image, wherein the measurement of $L^{\max}$ can be done from any edge in an image.

3 Implementation

Since the proposed method relies on step edges, it is assumed that there is at least one sharp edge in the underlying image. To estimate the blur parameter, the identification mainly comprises four steps as shown in Fig. 1.

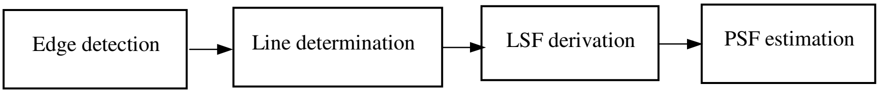
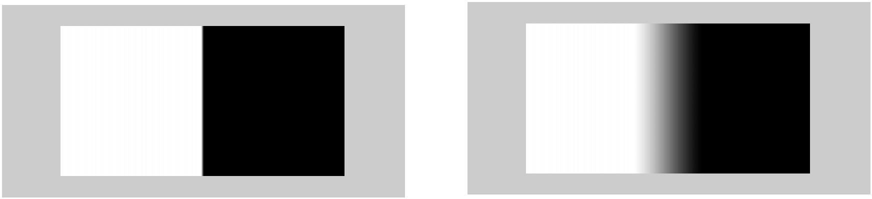


Fig. 1. Flowchart of the proposed method

3.1 Edge Location in Blurred Images

We have shown in Section 2 that it is possible to construct the LSF from edges, i.e. the derivative of unit step edge. However, if the image is blurred as shown in Fig. 2(b), it is not obvious where the edges are located, but the edge localization is crucial in extracting information from the edges.



(a) A sharp image with a unit step edge (b) The corresponding blurred image with *R =15*

Fig. 2. Sharp image with a step edge and the corresponding blurred image

Theorem: *If* $P(x_0, y_0)$ *is a real edge point in a sharp image f(x, y), it is with the local maximal gradient in a blurred image g(x, y).*

Proof: If $P(x_0, y_0)$ is an edge point, the first-order differential at $P(x_0, y_0)$ has the following property:

$$Df(x_0, y_0) = \max(Df(x, y)), \qquad P(x, y) \in \Omega \tag{9}$$

where Ω is the domain containing the edge point $P(x_0, y_0)$, and

$$D^2 f(x_0, y_0) = 0 \tag{10}$$

According to differential rule of convolution, we obtain from Equation (1) without considering the noise $n(x,y)$:

$$D(h(x, y) \otimes f(x, y)) = D(h(x, y)) \otimes f(x, y) = h(x, y) \otimes D(f(x, y)) \tag{11}$$

$$D^2(h(x, y) \otimes f(x, y)) = h(x, y) \otimes D^2(f(x, y)) \tag{12}$$

Then

$$D^2 g(x_0, y_0) = h(x_0, y_0) \otimes D^2(f(x_0, y_0)) = 0 \tag{13}$$

This proves that the real edges in blurred images still have local maximal gradients, and consequently they can be detected by conventional gradient operators.

As an example, take differentiation in the horizontal direction in the blurred image shown in Fig. 2 (b), we obtain the LSF demonstrated in Fig. 3. It is shown that the greatest horizontal gradient occur on actual edge. The LSF in Fig. 3 is identical to the result obtained from Equation (7).

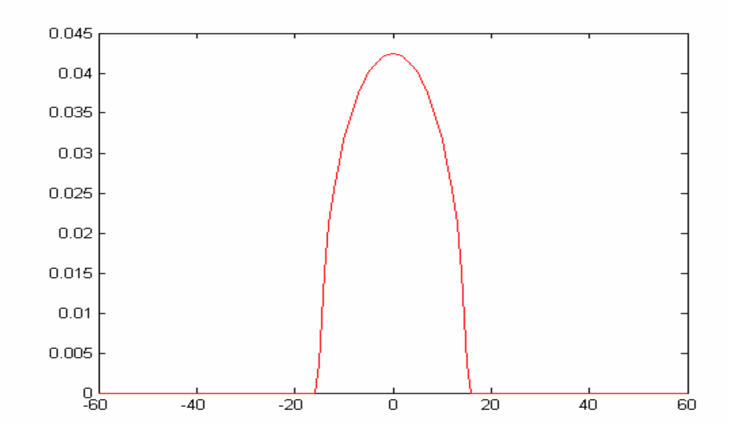


Fig. 3. The LSF extracted from image Fig. 2(b)

3.2 Edge Detection

Currently, many edge detectors have been proposed, among which Sobel and Canny detectors are the most commonly used two. Sobel operator works fast, but tends to amplify the noise as well. Canny method on the other hand, is sensitive to change of intensity and detects weak edges or even artifacts. Moreover, Canny detector involves with more computation. In the following analysis, Sobel detector is used for its efficiency. More discussion for edge detection will be given in Section 5.

Denote the intensity image as $I \in \mathbb{R}^{M \times N}$, and we obtain the gradient information of each pixel as follows:

$$G(i,j) = \sqrt{G_x^{\,2}(i,j) + G_y^{\,2}(i,j)} \tag{14}$$

$$\alpha(i,j) = \tan^{-1}(G_y(i,j) \,/\, G_x(i,j)) \tag{15}$$

where $G(i, j)$, $\alpha(i,j)$ are the gradient magnitude and direction respectively at location (i, j). Then the image I is converted to a binary image $B \in \mathbb{R}^{M \times N}$ according to the following:

$$B(i,j) = \begin{cases} 1 & if\ G(i,j) \geq \xi \\ 0 & if\ G(i,j) < \xi \end{cases} \quad i = 1,2\cdots, M,\ j = 1,2\cdots N \tag{16}$$

where

$$\xi = \frac{2}{MN}\sum_{j=1}^{N}\sum_{i=1}^{M} G(i,j) \tag{17}$$

3.3 Edge Selection and Location

To deduce the LSFs, it is necessary to locate the straight edges from the binary image *B*. In our proposed method, Radon transform is used to locate line features.

Radon transform (RT) of $f(x,y)$ is the line integral along a specific direction (Jain 1989):

$$R(s,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(x\cos\theta + y\sin\theta - s)dxdy \tag{18}$$

In the rotated coordinate system (*s*, *t*), where

$$s = x\cos\theta + y\sin\theta \tag{19}$$

$$t = -x\sin\theta + y\cos\theta \tag{20}$$

Equations (19) and (20) can also be represented as

$$R(s,\theta) = \int_{-\infty}^{\infty} f(s\cos\theta - t\sin\theta, s\sin\theta + t\cos\theta)dt, \quad -\infty < s < \infty, 0 \le \theta < \pi \tag{21}$$

Therefore, the RT maps the spatial domain (*x*, *y*) to the domain (s,θ), where a line in the spatial domain corresponds to a high value in the (s,θ) domain.

Set the angle interval be $\frac{\pi}{180}$, and the RT of image *B* be $R(s,\theta)$. Then the line candidates are determined as follows:

$$L(s_i,\theta_i) = \begin{cases} R(s_i,\theta_i) & if\ R(s_i,\theta_i) \ge \delta\max(R(s,\theta)) \\ 0 & if\ R(s_i,\theta_i) < \delta\max(R(s,\theta)) \end{cases}$$
$$\theta_i = 0, \frac{\pi}{180}, \cdots, \frac{179\pi}{180}, \ s_i = 0, 1, \cdots\sqrt{M^2+N^2} \tag{22}$$

where $0 < \delta < 1$ being a parameter to determine the number of line candidates.

The accuracy of the line direction is very important for LSF extraction. It should be noted that the RT is a global transform, which implies that the detection of a local line is affected by other lines. On the other hand, the choice of δ may miss some real lines and includes some artificial ones. To alleviate the effect of the random choice of δ, and accurately detect real lines with different image contents, local processing of the RT is proposed.

First, label the binary image *B* as follows:

$$B_L(i,j) = \begin{cases} k & if\ B(i,j) = 1\ and\, B(i,j) \in \Omega(k) \\ 0 & others \end{cases} \tag{23}$$

where $B(i,j)$ is 8-connection to *k*th edge $\Omega(k)$. If an edge segment is short, say, contains less than 10 pixels, it is then excluded.

Assume B_L have *u* edge segments. For each edge segment *k* ($k = 1,2,\cdots u$), search

$$J = \{(i,j) \mid B_L(i,j) = k\} \tag{24}$$

find the points:

$$K = \{J \mid\mid \alpha(J) - median(\alpha(J)) \mid\leq \sigma\} \tag{25}$$

where σ is a threshold. If there are enough connected points with similar gradient feature, an edge segment is located.

For each detected line, the RT is performed, and the line parameters s_l, θ_l are determined as follows:

$$s_l, \theta_l = \left\{s_l, \theta_l \mid R(s_l\ \theta_l) = \max\ R(s\ \theta)\right\} \tag{26}$$

Assume that a line AB shown in Fig. 4 represented by $(s_l,\ \theta_l)$ be detected with the RT, then the position of point *P* which is perpendicular to AB and crosses the origin is

$$\begin{aligned} x_P &= s_l \cos\theta_l \\ y_P &= s_l \sin\theta_l \end{aligned} \tag{27}$$

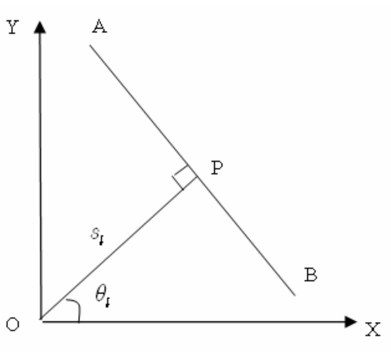


Fig. 4. A line represented by $(s_l,\ \theta_l)$

The segment AB is delineated as follows:

$$y = y_P + (x - x_P)\tan(\theta_l + \frac{\pi}{2}) \tag{28}$$

3.4 Pixel Interpolation and LSF Extraction

As illustrated in Section 2, the LSF is the derivative along the direction orthogonal to the edge in a given direction θ. In a discrete domain, the LSF is characterized by a sequence of points in 1-D orthogonal to the edge interested.

After obtaining the location and slope of the edge, for example AB as shown in Fig. 4, the sequence of points orthogonal to the edge and crossing *P* is:

$$\begin{aligned} x_i &= x_P + i \times T \times \cos\theta_l \\ y_i &= y_P + i \times T \times \sin\theta_l \end{aligned} \qquad i = 0,\ 1,\ 2, \cdots\ r \tag{29}$$

where *T* is the interval of the sequent points and is chosen as 1 pixel; *r* is the number of points to be extracted. As can be seen from Equation (29), the position (x_i, y_i) cannot be accurately located at the pixel position if θ_l is not equal to 0 or $\frac{\pi}{2}$. Interpolation is therefore necessary for the actual position.

Assume that one point *P* locate within its 4 neighbor pixels $g(i,j)$, $g(i,j+1)$, $g(i+1,j)$, and $g(i+1,j+1)$, which has distances d_x, d_y from $g(i,j)$ in x- and y- axes respectively, as shown in Fig. 5. First, we interpolate the intensities of points P_k ($k = 1,2,3,4$), which cross the point *P* as follows:

$$\begin{aligned} g_{P_1} &= g(i,j)(1-d_y) + d_y g(i+1,j) \\ g_{P_2} &= g(i,j)(1-d_x) + d_x g(i,j+1) \\ g_{P_3} &= g(i,j+1)(1-d_y) + d_y g(i+1,j+1) \\ g_{P_4} &= g(i+1,j)(1-d_x) + d_x g(i+1,j+1) \end{aligned} \tag{30}$$

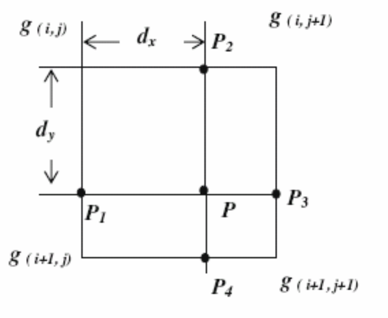


Fig. 5. Interpolation of point *P* which is not on the pixel

Then, the intensity of point *P* is computed as follows:

$$g_P = [g_{P_1}(1-d_x) + g_{P_3}d_x + g_{P_2}(1-d_y) + g_{P_4}d_x]/2 \tag{31}$$

After interpolation, we extract the sequence data $g(x_i, y_i)$ $(i = 1, 2, \cdots r)$ in the direction perpendicular to the edge.

Generally, the underlying edges in the image may not be a unit step edge. It is necessary to normalize them to unit step edges. Let $e(x)$ be an edge parallel to x-axis. Then

$$e(y) = Kh_e(y) \tag{32}$$

where *K* is the step value of the edge. We have

$$\int_{-\infty}^{\infty} \frac{\partial e(y)}{\partial y} dx = K \int_{-\infty}^{\infty} \frac{\partial h_e(y)}{\partial y} dx = K \int_{-\infty}^{\infty} L(y) dx = K \tag{33}$$

This indicates that the integration of edge derivative furnishes us the step value of this edge, and we can normalize an edge to the unit step one by division with *K*.

Consequently, the estimated LSF is obtained by taking differential of the sequence data $g(x_i, y_i)$ $(i = 1, 2, \cdots r)$ as follows:

$$\hat{L}(i) = D(g(x_i, y_i) / K) \qquad i = 1, 2, \cdots r \tag{34}$$

where *K* is determined by the maximum of $g(x_i, y_i)$, $i = 1, 2, \cdots r$.

3.5 LSF Alignment and Determination

It is obvious that the accuracy of the extracted LSF is affected by image noise and the pattern outside the edge region. To reduce this effect, one effective solution is to extract more LSFs from different locations of one edge and from different edges as well.

Denote $\hat{L}_{l_i}^{n} \in \mathbb{R}^{n \times r}$ be *n* LSFs, and $\hat{L}_{l_i}^{k}(j)$ (*j* = 1, 2 · · · *r*, *k* = 1, 2, · · · *n*) be the *j*th element of the *k*th LSF extracted from the *i*th edge. To average the LSFs, we have to know the position that each datum corresponds to, i.e. it is needed to align each LSF to the right position. As illustrated in Section 3.1, the largest gradient occurs at the actual edge. Accordingly, the peak gradient is the only basis to align the LSFs.

For each LSF sequence $\hat{L}_{l_i}^{k} \in \mathbb{R}^{r}$, find

$$J = \arg \max_{j=1,\ 2,\cdots r} (\hat{L}_{l_i}^{k}(j)) \tag{35}$$

then circularly shift each LSF sequences to keep the peak element at the central position. After alignment, the missing data will be padded as 0s, and the LSF $\hat{L}_{l_i}^{\max}$ from the edge is finally determined as follows:

$$\hat{L}_{l_i}^{\max} = median(\hat{L}_{l_i}^{1}(J), \hat{L}_{l_i}^{2}(J) \cdots \hat{L}_{l_i}^{n}(J)) \tag{36}$$

Fig. 6 demonstrates two LSF sequences (after taking derivative) extracted from an edge. The data in the circles are the peaks and correspond to the actual edges.

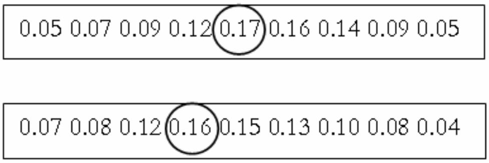


Fig. 6. Two LSF sequences and their alignment

Assume that there are q LSFs $\hat{L}_{l_1}^{\max}, \hat{L}_{l_2}^{\max} \ldots \hat{L}_{l_q}^{\max}$ extracted from different edges in the image, then the final LSF $\hat{L}_{l}^{\max}$ is determined by:

$$\hat{L}_{l}^{\max} = \max(\hat{L}_{l_1}^{\max}, \hat{L}_{l_2}^{\max} \ldots \hat{L}_{l_q}^{\max}) \tag{37}$$

3.6 Determination of PSF Parameter

After the maximum of LSF, $\hat{L}_{l}^{\max}$, is extracted, the PSF parameter *R* is calculated using Equation (8). As $\hat{L}_{l}^{\max}$ is determined by the maximum of all $\hat{L}_{l_i}^{\max}$ ($i = 1, 2, \cdots q$) shown in Equation (37), this implies that the blur extent is evaluated by the sharpest edge. Once R is known, we can use it as a criterion for measuring the blur degree, construct the PSF according to Equation (3) and perform image restoration as demonstrated in the next section.

4 Experimental Results

In order to verify the effectiveness of the proposed method, we have conducted experiments using a variety of synthetic and real blurred images in Matlab environment. The parameters predefined are as follows: the threshold $\sigma = 3$, the LSF length *r* = 21. The experimental results are demonstrated in following subsections.

4.1 Simple Images

Fig. 7 shows a simple and sharp image of size 900×900. To demonstrate the proposed method, the original image is blurred with different parameters *R* (*R* =1, 2, … , 11). After the blur parameter is estimated, the blurred image is restored using Lucy-Richardson algorithm. Note that the minimal unit in processing is 1 pixel, the PSF parameter is therefore rounded to an integer. The restoration performance is evaluated in terms of the identified parameter $\hat{R}$ and the improvement in SNR (ISNR) which is given by:

$$ISNR = 10\log_{10}\frac{\sum_{i,j}[f(i,j)-g(i,j)]^2}{\sum_{i,j}[f(i,j)-\hat{f}(i,j)]^2} \tag{38}$$

where $f(i,j)$ is the original image, $g(i,j)$ is the blurred image and $\hat{f}(i,j)$ is the corresponding restored image. If the ISNR is calculated using the identified parameter rounded to the nearest integer, the ISNR results are tabulated in Table 1.

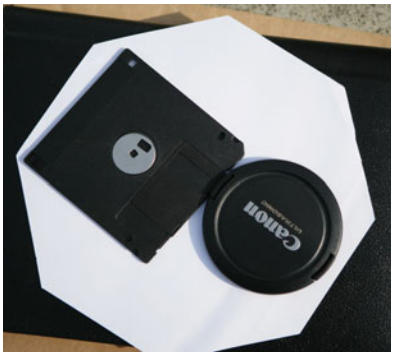
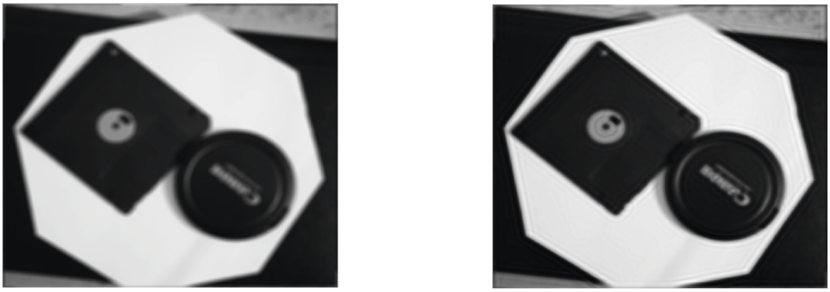


Fig. 7. Original simple image

It is seen from Table 1 that the error of identification is less than one pixel for all cases, and the fact that all ISNRs are positive reveals that all the restored images are improved. As an example, Fig. 8 shows a blurred image (*R* = 11) and the restored image.



(a) Synthetic blurred image (*R* = 11)

(b) Restored image via Lucy-Richardson method

Fig. 8. Synthetic blurred image and restored image

Table 1. Identified results for simple image

R (pixels)	1	3	5	7	9	11
$\hat{R}$ (pixels)	1.42	3.43	5.48	7.13	9.54	11.48
$\lvert R-\hat{R} \rvert$	0.42	0.43	0.48	0.13	0.54	0.48
ISNR	5.97	3.31	3.13	3.14	2.02	3.06

It is found from experiments that the sharp edges occur on the octagon. Table 2 demonstrates the accuracy of edge localization and the performances based on edges in different orientations when *R* is 5. As the internal angle at each vertex is 135, it is seen that the maximum error of the edge orientations is 3 degrees. The results illustrate that the performance is not dependent on edge orientation.

Table 2. Performance on edge localization and orientation

Edge	Angle (degree)	$\hat{R}$ (pixel)
1	10	5.60
2	11	5.82
3	59	5.48
4	57	5.66
5	103	5.50
6	103	5.74
7	148	5.61
8	148	5.62

It is also observed from Table 1 that even the error of estimated blur parameter is the same (for example, $\lvert R-\hat{R} \rvert = 0.48$), the ISNRs are not identical in different blur extents (ISNR = 3.13, 3.06 in *R* = 5, 11 respectively). The more blur the image is, the smaller the ISNR is when the estimated error is identical.

In order to illustrate the effect of rounding error on ISNRs, choose the blur radius to be one pixel away from the ideal parameter, the ISNRs under different blur extents are demonstrated in Fig. 9. It is found that the ISNRs with underestimation are always higher than those with overestimation, but the difference becomes smaller with the increase of blur extents. This implies that the error caused by truncation is small for heavily blurred images. Accordingly, we round the identified parameter to the nearest integer towards zero in the experimental results.

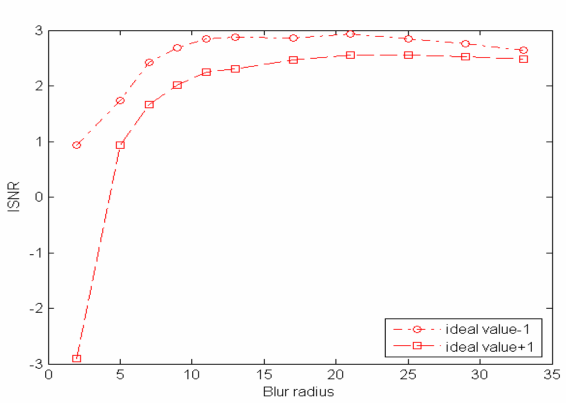


Fig. 9. Rounding error vs ISNR

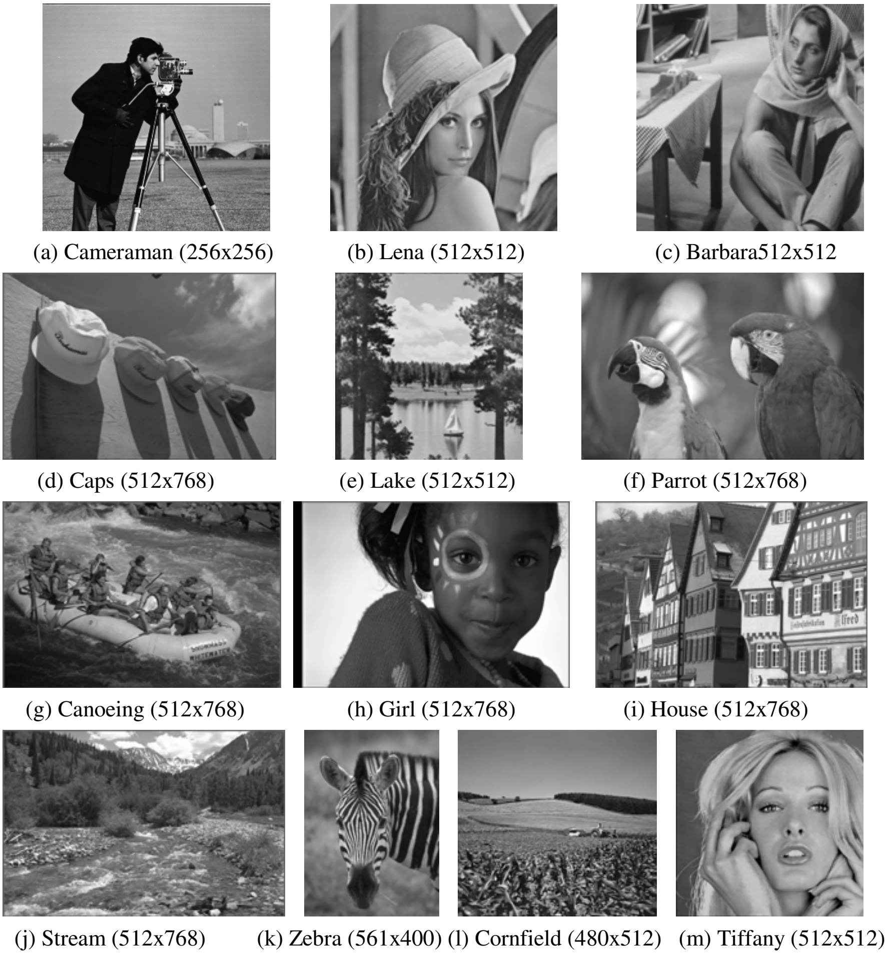
The two-order moment was proposed by Wu et al. (Wu et al. 2005) to measure the blur degree. The experimental results in the image shown in Fig. 7 are tabulated in Table 3. We see that the identified errors are comparable to the proposed metric (less than one pixel) when R = 3 ~ 9. But the estimated errors are large when the image is sharp (R < 3) or very blur (R >10). Therefore, the proposed method is capable of handling wider scopes of applications.

Table 3. Identified results for simple image using two-order moment

R (pixels)	1	3	5	7	9	11
$\hat{R}$ (pixels)	4.2	3.28	4.96	6.57	8.32	9.22
ISNR	-15.23	3.31	3.13	3.14	2.69	1.87

4.2 Standard Test Images

The 13 standard real-world images shown in Fig. 10 are used, with different picture sizes and visual content. The original images are blurred to different extents. The parameters identified are listed in Table 4. The ISNRs for three typical images under different blur conditions are illustrated in Fig. 11.



(a) Cameraman (256x256) (b) Lena (512x512) (c) Barbara512x512

(d) Caps (512x768) (e) Lake (512x512) (f) Parrot (512x768)

(g) Canoeing (512x768) (h) Girl (512x768) (i) House (512x768)

(j) Stream (512x768) (k) Zebra (561x400) (l) Cornfield (480x512) (m) Tiffany (512x512)

Fig. 10. Representative images for experiments

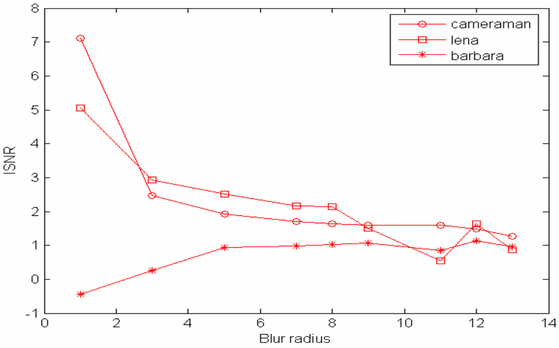


Fig. 11. The ISNRs and blur extents for representative images

Table 4. Identified results for representative images

R (pixels)								
Image	1	3	5	7	9	11	13	15
Cameraman	1.2	3.2	5.3	7.4	9.6	11.5	12.1	13.5
Lana	1.8	3.4	5.4	7.0	8.7	10.8	12.0	13.3
Barbara	2.1	4.1	5.4	7.5	9.6	11.9	13.3	12.7
Caps	1.9	4.1	5.7	7.2	8.7	10.3	11.6	12.6
Lake	2.4	4.4	5.6	5.8	9.5	13.7	14.2	14.9
Parrot	2.3	4.2	7.6	8.6	10.6	12.5	14.3	17.5
Canoeing	1.8	4.2	6.2	7.7	12.0	12.5	13.7	15.7
Girl	1.3	3.0	5.2	7.1	9.2	10.8	11.4	12.6
House	1.5	3.4	5.7	6.9	8.1	10.9	13.1	14.4
Stream	2.8	4.4	6.0	7.4	9.9	12.1	14.0	17.9
Zebra	2.6	3.9	5.8	6.1	7.0	7.9	8.9	9.2
Cornfield	2.2	3.5	5.6	7.6	10.1	11.9	12.7	12.2
Tiffany	1.9	3.2	5.2	7.0	9.5	11.6	14.2	15.1

As it is observed from Table 4 and Fig. 11, the ISNRs for different images with same blur extent are different even the PSFs are correctly estimated (for example, R = 5 ~ 8). This implies that ISNR depends on not only accuracy of PSF estimation but also image contents.

4.3 Blurred Images with Noise

To further show the effect of noise on the performance, Gaussian white noise is injected into the "Cameraman" image. The noisy image is measured by SNR as follows:

$$SNR = 10\log_{10}\frac{\text{var } of \;\; d(i,j)\otimes f(i,j)}{\text{var } of \;\; n(i,j)} \tag{39}$$

The results are shown in Table 5. Our results show that the proposed method is robust up to at least 30dB in SNR.

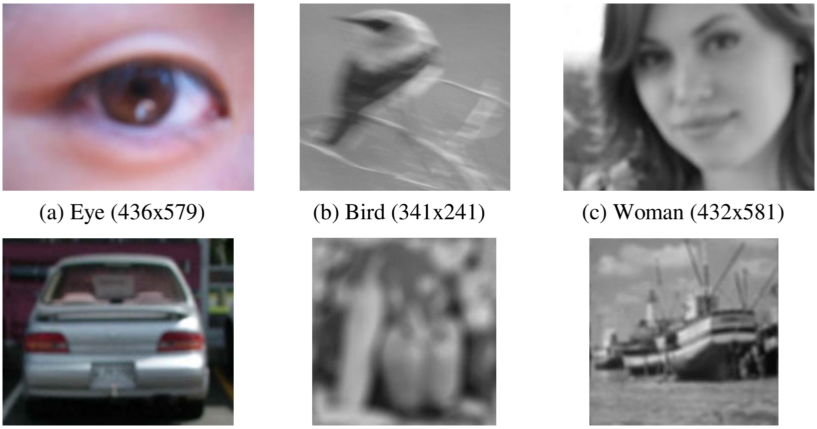
Table 5. Identified results under different noise levels

SNR (dB) *R* (pixels)	40	30	20	10
1	1.2	1.3	1.4	N/A*
3	3.2	3.0	3.0	N/A
5	5.5	5.1	4.3	N/A
7	7.6	6.9	N/A	N/A
9	9.4	8.3	N/A	N/A
11	10.6	N/A	N/A	N/A
13	N/A	N/A	N/A	N/A

* no edges are detected

4.4 Real Blurred Images

Then, some real defocus-blurred images downloaded from Internet as shown in Fig. 12 are used to test the proposed approach. In this case, we have no knowledge about the blur information, and the "true value" of *R* is obtained in the following way: restoring the blurred images by setting PSF = 1, 2, … 20, choose the right *R* which corresponds to the best restored image perceptually. The blurred parameters identified by the proposed method are shown in Table 6. It is observed from Table 6 that the identified results are quite good.



(a) Eye (436x579) (b) Bird (341x241) (c) Woman (432x581)

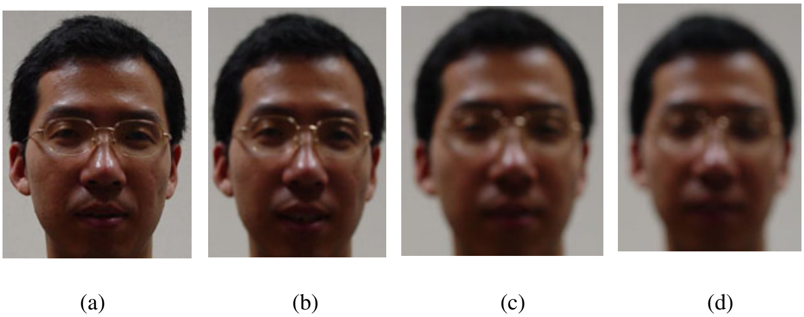
(d) Car (303x404) (e) Pepper (199x199) (f) Fishingboat (197x199)

Fig. 12. Real blurred images

Table 6. Estimated blur parameters for real blurred images

Image	$\hat{R}$ (pixels)	Estimated true value
Eye	9.42	9
Bird	6.47	5
Woman	8.07	9
Car	3.96	5
Pepper	8.57	9
Fishing boat	3.65	3

As a blind method, the proposed algorithm can serve as a tool for image quality evaluation used in robot navigation, face recognition, surveillance system and so on. For instance, blur detection can be used a pre-processor before object recognition; if an input image fails in this quality check, the vision-based system can trigger the camera to re-take the image. For the images which are available, the blur extents provide weightings for decision-making in face recognition.



(a) (b) (c) (d)

Fig. 13. A sequence of facial images

Fig. 13 shows four human facial images of size 494×373 applied to a facial recognition system. The first one is the best focused, and the last one is the most blurred. The test results for these images are tabulated in Table 7.

Table 7. Estimated blur radius for facial images

Image	A	B	C	D
$\hat{R}$ (pixels)	1.3	4.0	6.6	7.7

5 Further Analysis

In this section, the Sobel and Canny operators are compared, and the performance of the proposed method is thoroughly analyzed.

5.1 Sobel Operator vs Canny Detector

Canny edge detector is frequently used in image processing due to its good detection, good localization and only one response to a single edge. Table 8 shows the identified results using the representative images shown in Fig. 10 by Canny detector.

Table 8. Identified results using Canny operator

R (pixels) / Image	1	3	5	7	9	11	13	15
Cameraman	1.2	3.3	5.4	7.4	9.6	11.3	13.1	13.7
Lana	1.9	3.6	5.8	7.3	9.0	10.6	12.0	13.8
Barbara	1.8	4.2	6.2	8.4	10.6	11.9	13.1	14.0
Caps	1.7	4.2	6.7	7.4	10.3	10.6	12.1	13.7
Lake	1.9	4.1	6.2	7.8	8.5	10.5	12.5	13.7
Parrot	2.3	4.3	6.5	8.6	10.9	12.6	14.3	16.8
Canoeing	1.5	4.1	6.5	7.5	10.4	10.9	13.7	15.8
Girl	1.5	3.1	5.5	8.3	10.1	10.7	12.7	14.1
House	1.2	3.4	5.7	6.2	8.1	9.7	11.5	13.9
Stream	1.9	3.7	6.0	7.2	9.4	12.0	13.3	14.6
Zebra	2.0	3.5	4.7	5.7	7.2	8.4	9.6	10.0
Cornfield	1.5	3.6	5.8	7.6	9.9	11.9	13.3	14.9
Tiffany	1.8	3.3	5.1	7.5	9.5	11.8	13.9	16.4

It is observed from Tables 4 and 8 that Canny method generally outperforms Sobel detector as it provides higher quality edges than Sobel detector. The edges extracted by Canny method are continuous and smooth, especially for seriously blurred images so that is easy to detect lines when using local Radon transform. However, the computation of Canny detector is much more expensive than Sobel operator, and more predefined parameters are needed that it is not convenient to use. Sobel detector is chosen in this work because of its simplicity and fast speed. More importantly, although the accuracy of edge localization by Sobel detector is not as good as the Canny detector, the alignment of LSFs illustrated in Section 3.5

abates the effect of edge detectors, and enables Sobel detector to provide comparable performance.

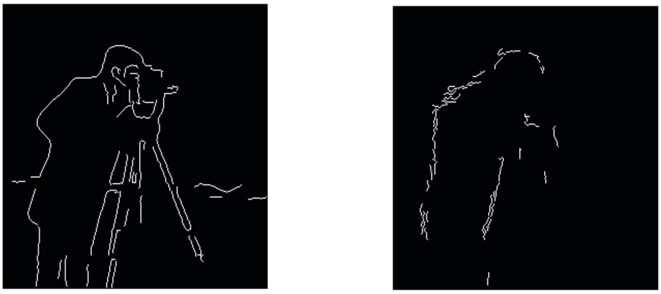
5.2 Error Source

The estimation accuracy of the proposed method mainly depends on the following factors: 1) how sharp the step edges occur in the image; 2) whether the LSF is derived from the sharpest edge; 3) how accurate the line orientations are detected; and 4) image content.

Due to imperfection in the imaging system and capturing process, the recorded image represents a degraded version of the original scene. This reveals that the spatially continuous PSF cannot be modeled as a Dirac delta function. Accordingly, it is impossible to find an ideal step edge shown in Fig. 2(a). This situation results in error of LSF extracted from unsharp edges in the underlying image as indicated in Equations (5) and (6). This explains why the estimated parameters shown in Tables 1, 4 and 8 are always bigger than theoretical values when R is small, although the edges can be accurately detected by Sobel operator and both the edge location and orientation are also accurately estimated in this case. On the other hand, a seriously blurred image will decrease the sharpness of edges due to large amount of spread.

It is proven in Section 3.1 that a real edge is with the local maximum of gradient in a blurred image, and we use Sobel operator for edge detection. However, Sobel operator, unlike the Canny edge detector, does not perform non-maximum suppression and hysteresis thresholding. This leads to wide ridges around the true edges and produces disconnected edges. Such situation becomes worse when the image is seriously blurred as demonstrated in Fig. 14. As indicated by Ziou (Ziou 2001), the gradient magnitude of the non-radial edge detectors is affected by edge translation and orientation. This implies that a sharpest edge in the original image may be not the sharpest one in the blurred image. For the "Cameraman" image, the back of the man's body (as shown in Fig. 14(a)) is the sharpest edge which derives the right PSF when the blur parameter R is no more than 13. This edge is not good for blur estimation when R is bigger than 13.

The detection of edge orientation is crucial for LSF extraction as it is the derivative along the direction orthogonal to the edge. It was indicated by Ziou (Ziou 2001) that the accuracy of edge orientation depends on detector scale, edge translation and edge orientation. Frequently, the edges detected by Sobel operator from blurred images yield parallel shifting of the actual edge and are broken into small pieces, as shown in Fig. 14(b). Although the thinning processing is performed before the Radon transform， the orientation error may be big for seriously blurred images ($R > 11$), as demonstrated in Fig. 15.



(a) Blurred radius R =3 (b) Blurred radius R = 13

Fig. 14. Edges detected in image Cameraman by Sobel operator

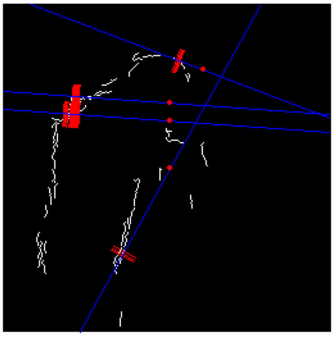


Fig. 15. Orientation error of edges (R =13)

6 An Efficient Solution

The aforementioned sections provide a general and accurate solution for blind blur measurement. It should be highlighted that this paradigm is lack of efficiency. The algorithm uses Radon transform for detection of edge directions, and leads to intensive computation. On another hand, it is seen from Equations (8) and (37) that the blur extent is determined by the sharpest edges in the underlying image, from which the LSF is extracted. To find the sharp edges, every edge in the underlying image is extracted for evaluation in the above solution. If we develop a criterion to measure the edge sharpness, only some points from sharp edges are required. This is the motivation to develop an efficient solution for blind blur assessment so that it can be deployed in real-time vision-based applications. In practice, there is always intrinsically blur regions in an image due to the limited depth of field for a camera. Use of the sharpest edge also leads to the measurement oriented to the region-of-interest in applications.

6.1 Edge Model and Sharpness

It is seen in Section 2 that Equation (8) is derived based on step edge, which is in general form as below:

$$E(x,y;b,c) = cU(x,y) + b \tag{40}$$

where *U*(*x*,*y*) is the unit step function, *c* is the contrast of the edge and *b* is the function value at the base of the edge. However, infinitely sharp edges such as step edges do not exist in practical images, due to the band-limiting characteristics of the optical acquisition system. An edge in an image is normally corresponding to a discontinuity in the scene function *f* (*i*, *j*), distorted by a PSF *h(i*, *j*) as shown in Equation (1). Specifically, a Gaussian function $h(x,y;\sigma)$, with σ as the smoothing parameter, is frequently used to represent the PSF of an optical system:

$$h(x,y;\sigma) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \tag{41}$$

Therefore, a step edge in Equation (40) convolved by a Gaussian PSF results in a scaled error function $s(x,y;\, b,c,w,\theta)$:

$$s(x,y;b,c,w,\theta) = b + \frac{c}{2}(1 + erf(\frac{x\cos\theta + y\sin\theta}{w\sqrt{2}})) \tag{42}$$

with *erf(.)* as the error function, *w* as the parameter controlling the width of the edge, and θ being the angle with respect to the x-axis. The 1D edge model is depicted in Fig. 16.

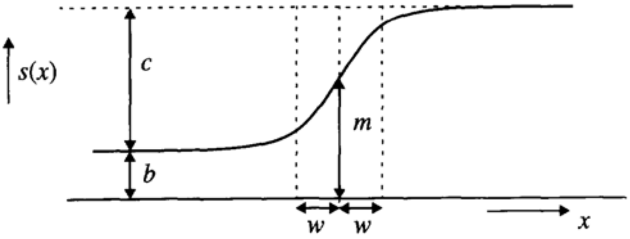


Fig. 16. Edge model in one-dimension

Assume that an edge is detected, say by Sobel or Canny detector, and based on the aforementioned edge model, the model parameters are derived using multi-point estimation as follows (Beck, 1995):

$$w^2 = \frac{a^2}{\ln(d_1^2 / d_2 d_3)} - \sigma \tag{43}$$

$$c = d_1 \sqrt{\frac{2\pi a^2}{\ln(d_1^2 / d_2 d_3)}} (\frac{d_2}{d_3})^{1/4a} \tag{44}$$

where *a* is the distance of the selected two points to the edge, d_i $(i = 1,2,3)$ is the response of the edge to the first derivative of a Gaussian filter at multiple points as follows:

$$d^x(x, y, c, w, \sigma, \theta) = s(x, y, c, w, \sigma, \theta) * \frac{\partial}{\partial x}(h(x, y; \sigma)) \tag{45}$$

$$d^y(x, y, c, w, \sigma, \theta) = s(x, y, c, w, \sigma, \theta) * \frac{\partial}{\partial y}(h(x, y; \sigma)) \tag{46}$$

and

$$d_i = \sqrt{(d_i^x)^2 + (d_i^y)^2} \qquad i = 1,2,3 \tag{47}$$

Furthermore, the direction of the smoothed gradient is equal to the direction of the edge:

$$\theta = \tan^{-1}(\frac{d^y(x, y; c, w, \sigma, \theta)}{d^x(x, y; c, w, \sigma, \theta)}) \tag{48}$$

We see from Fig. 16 that there are three parameters in the edge model: 1) contrast *c*; 2) intensity at the base of the edge *b*, or alternatively, intensity at the edge center *m* = *b*+*c/2*; 3) spatial width of the edge *w*. It is obvious that the parameters *c* and *b* are mainly determined by the scene characteristics without imaging information. The edges with large *c* alleviate the effect of noise and edge detectors. However, parameter *w* is dependent not only on intrinsic scene sharpness, but also the smoothness introduced by image formation process.

Considering the factors of contrast as well as width, we define the following criterion for edge sharpness measure:

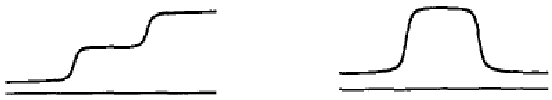
$$S = \frac{c}{w} \tag{49}$$

It is easy to understand from Fig. 16 that *S* represents the slope of transition. Big *S* reveals that the underlying edge is sharp. Therefore, we can select the sharp edges according to *S* for LSF extraction.

6.2 Effect of Nearby Edges on LSF

In the aforementioned analysis, only isolated edges have been discussed. Naturally, plenty of edges exist in an image, and the filter response of a particular edge is affected by a nearby edge, especially when the scale of the filter is large in comparison to the distance between these two edges. Not only the peak value of the response can change, but also the peak location can shift because of the influence of a nearby edge.

It has been analyzed by Beck (Beck, 1995) that two types of interaction between two nearby edges occur. If the contrasts of the two edges have the same sign, the edges together form a staircase edge; otherwise, the edges together form a pulse edge as shown in Fig. 17. For the pulse edge, the two edges move further apart as the filter scale increase, while for the staircase edge, the locations move closer together as the filter scale increase. The error of the peak values is determined by the first derivative of a Gaussian function. Refer to Beck's work (Beck, 1995) for detailed analysis.



(1) Staircase edge

(2) Pulse edge

Fig. 17. Effect of nearby edges

From the LSF point of view, it is highlighted that the LSFs extracted from two close edges are always contaminated by the diffraction of light energy. Fig. 18 demonstrates the interaction of two nearby LSFs (R = 15). The blur spreads a step edge's energy to a distant as far as R , and two nearby edges (for example, the distance is 20 pixels in the figure) will increase the peak value of the LSFs and results in smaller PSF.

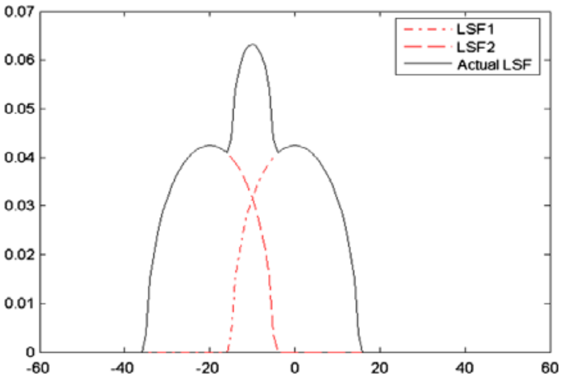


Fig. 18. Interaction of two nearby LSFs

Considering the effects of nearby edges, the following two constrains are necessary in extracting the LSFs:

***Constrain (1) Space constrain*:** suppose $P(i,j)$ to be an edge point whose gradient direction is θ calculated by Equation (48). If no edge is found in the gradient direction within distance *D*, the point $P(i,j)$ is selected for LSF calculation.

Constrain (2) Peak shift of LSFs: assume that $L_k(i,j)$ is an extracted LSF. If $| L_k^{\max}(i,j) - P(i,j) |> \delta$, where δ is a predefined threshold, then the $L_k^{\max}(i,j)$ is distorted and discarded.

After edge detection, we obtain the edge locations and directions, the LSFs can be extracted from the sharp edges according to the criterion of edge sharpness. It should be noted that the pixel interpolation is also necessary, which is identical as shown in Section 3.4. As the edge points extracted may be from different edges, LSF alignment is not necessary before LSF extraction. The following procedure summarizes the algorithm:

```
**Procedure:** estimate the blur extent of an image
Input: an image
Output: blur parameter $\hat{R}$
**Begin**
  1) Determine edge location via Canny edge detection.
  2) Find the appropriate edges according to space constrain.
  3) Find edge directions according to Equation (48)
  4) Compute sharpness S according to Equations (43), (44) and (49).
  5) Extract n edge points corresponding to the n largest S.
  6) Pixel interpolation according to Equations (30) and (31)
  7) Extract LSF according to Equation (34)
  8) Discard the LSFs which are not satisfied with the constrain of peak shift
  9) Find the maximal LSF $L^{\max} = \max(L_1^{\max},\ L_2^{\max} \dots\ L_n^{\max})$
  10) Compute the blur parameter $\hat{R}$ according to Equation (8)
**End**
```

6.3 Experimental Results

To verify the effectiveness of the proposed method, we conducted experiments using the above images. The parameters predefined are as follows: the threshold $\sigma = 1$, the space *D* = 10, distance *a* = 2, LSF length *r* = 21, and number of edge points *n* = 200.

For the standard real-world images shown in Fig. 10, the parameters identified by the efficient method are listed in Table 9.

Table 9. Results for representative images with the efficient method

R Image	1	3	5	7	9	11	13
Cameraman	1.1	3.2	5.4	6.8	8.6	10.4	11.3
Lana	1.6	3.7	5.7	8.2	10.7	12.6	15.4
Barbara	1.8	3.9	6.2	8.5	10.3	12.7	14.1
Lake	1.7	3.9	5.3	7.4	10.8	11.8	14.4
Canoeing	1.7	4.0	5.9	7.9	9.8	12.6	17.5
Girl	1.4	3.6	5.7	7.6	9.7	11.6	13.1
House	1.3	3.6	5.7	7.5	9.7	12.6	14.6
Zebra	1.9	4.0	5.8	7.7	9.3	10.6	12.7
Cornfield	1.3	3.3	5.5	7.5	10.2	12.5	14.4
Tiffany	2.1	3.3	5.1	6.8	9.6	13.8	17.2

Table 10. Results for real blurred images with the efficient method

Image	$\hat{R}$ (pixels)	Estimated true value
Eye	7.2	9
Bird	5.2	5
Car	3.2	5
Fishing boat	3.5	3

For the real defocus-blurred images shown in Fig.12, the blurred parameters identified by the efficient method are shown in Table 10. Table 11 demonstrates the results estimated by the efficient method for the facial images shown in Fig. 13.

Table 11. Results for facial images with the efficient method

Image	A	B	C	D
$\hat{R}$ (pixels)	1.1	3.5	5.7	6.5

As discussed in Section 5, the accuracy of the proposed method mainly depends on the following factors: 1) whether the LSF is derived from the sharpest edge; 2) how accurate the LSF is extracted. In Section 6, we used S to assess the sharpness of edges. To check the validity of sharpness criterion, the LSFs extracted corresponding to the first 400 biggest *S* in the image Lena are plotted in Fig. 19. It is observed that the LSF curve generally decreases, which implies that the sharpness decreases. However, the curve is not strictly monotonous, i.e., the location of the best result (maximal LSF) is not the first one, but occurs at point 8. Table 12 shows the locations of maximal LSFs in other images. It is seen that the optimal value is within 200 samples.

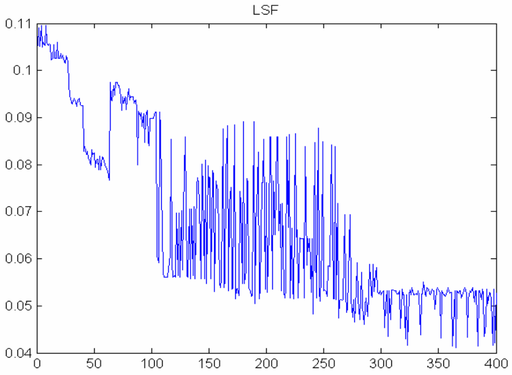


Fig. 19. The LSFs corresponding to the first 400 biggest *S* (*R* = 5)

Table 12. Location of maximal LSF (*R* = 5)

	Cameraman	Barbara	Lake	Canoeing	Girl
Location	70	125	20	135	130

It is noticed from Tables 9~ 11 with Tables 4, 6 and 7 that the performance by the efficient method is almost the same as the one presented in Sections 2 and 3, where all edges are extracted via Radon transform, a number of LSFs are extracted from one edge and their average LSF represents the result of the edge, and the final LSF is determined by LSFs from all edges. The solution proposed in this section estimates the blur extent by limited LSFs from only sharp edges. The edge direction is obtained from gradient angle of the underlying point without complex Radon transform. Nevertheless, such point-wise method is easily affected by image noise, and results in error of edge directions and LSF extraction.

However, it is highlighted that if the images are not serious blurred, say, *R* < 9, both methods yield similar results. This is because serious blurred images results in great error in edge location and direction, while the method in Sections 2 and 3 alleviates such error by full use of edge information in the underlying image. Fig. 20 shows the blurred image when *R* = 9. We see that the image is very blur, and usually images with *R*>9 are discarded in vision-based applications.

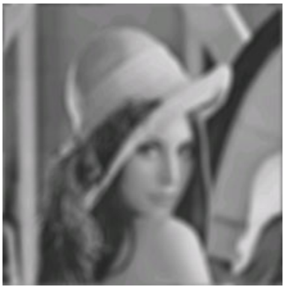


Fig. 20. A blurred image with $R = 9$

It is also noted that the proposed efficient approach achieves better results than the one in Sections 2 and 3 for the “zebra” image. This is because the image has sharp but very close edges that the interaction of edges is very strong. This verifies that the two constrains in extracting LSFs to eliminate the interaction of nearby edges in this section is effective in selecting the correct LSFs.

Although the method in Sections 2 and 3 is accurate, the computation is much more expensive. For an image with size of $M \times M$, the computational complexity of edge detection is approximately $O(M^2)$, and is independent of image content. The computational complexity of Radon transform is $O(M^2 \log M)$ and the computational complexity of LSF extraction is $O(rnq)$, where *r* is the number of points representing an LSF, *n* is the number of LSFs extracted from one line and *q* is the number of total lines found in the underlying image. Therefore, the operational time of LSF extraction is content-dependent.

On another hand, it is noted that the line determination of the efficient method needs computation of $O(\sum_{i=1}^{q} n_i)$ instead of $O(M^2 \log M)$, where n_i is the pixel number of the *i-th* edge, and *q* is the total number of lines. The computational complexity of LSF extraction is reduced to $O(rn)$ instead of $O(rnq)$ (note: $n = \sum_{i=1}^{q} n_i$) when the sharpest edge is detected. The comparison of operational time on Dell Precision T7400 (3.2GHz) in Matlab environment is tabulated in Table 13. It is observed that the efficient method has much faster speed than the scheme by Wu et al. (Wu et al., 2009) and can be implemented in real-time applications.

Table 13. Operational time

Method	Cameraman	Lena	Barbara
Method by Wu et al. (2009)	0.74	6.60	4.6
The efficient method	0.19	0.98	0.97

7 Conclusions

This chapter presents a new metric for image quality assessment in terms of image sharpness since the sharpness is considered as characteristics of blurring. The essential idea is to estimate the blur parameter from LSFs, whereas the LSF is constructed from edge information. To achieve high efficiency, a criterion for edge sharpness is proposed and only the sharpest edge is selected for LSF extraction without using each edge in the underlying image. The influences of nearby edges on LSF estimation are analyzed, and constrains on edge space as well as peak shift of LSFs are put forward to select appropriate LSFs. The experimental results demonstrate that the proposed method is accurate if the blur is not serious. This paradigm has fast speed and can serve as blind image quality evaluation for automatic machine-vision-based applications.

References

Andrews, H.C., Hunt, B.R.: Digital Image Restoration. Prentice Hall, Englewood Cliffs (1977)

Beck, P.V.: Edge-based Image Representation and Coding. Ph.D. dissertation, Delft University of Technology, The Netherlands (1995)

Bones, P.J., Bretschneider, T., Forne, C.J., Millane, R.P., McNeill, S.J.: Tomographic blur identification using image edges. In: Proc. SPIE, vol. 4123, pp. 133–141 (2000)

Cannon, G.O.M.: Blind deconvolution of spatially invariant image blurs with phase. IEEE Trans. Acoustics, Speech and Signal Processing 24, 58–63 (1976)

Chang, M.M., Tekalp, A.M., Erdem, A.T.: Blur identification using bispectrum. IEEE Trans. Signal Processing 39, 2323–2325 (1991)

Chen, L., Yap, K.-H.: Efficient discrete spatial techniques for blur support identification in blind image deconvolution. IEEE Trans. Signal Processing 54, 1557–1562 (2006)

Fabian, R., Malah, D.: Robust identification of motion and out-of focus blur parameters from blur and noisy images. CVGIP: Graphical Models and Image Processing 53, 403–412 (1991)

Fortier, N., Demoment, G., Goussard, Y.: GCV and ML methods of determining parameters in image restoration by regularization: fast computation in the spatial domain and experimental comparison. J. Visual Commun. Imag. Rep. 4, 157–170 (1993)

Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall, Englewood Cliffs (1989)

Kayargadde, V., Martens, J.-B.: Estimation of perceived image blur using edge features. International Journal of Imaging Systems and Technology 7, 102–109 (1996)

Keelan, B.W.: Handbook of Image Quality. Marcel Dekker, New York (2002)

Kim, S.K., Park, S.R., Paik, J.K.: Simultaneous out-of-focus blur estimation and restoration for digital auto-focusing system. IEEE Trans. Consumer Electronics 44, 1071–1075 (1998)

Kundur, D., Hatzinakos, D.: Blind image deconvolutions. IEEE Signal Processing Magazine 13, 43–63 (1996)

Lagendijk, R.D.L., Takalp, A.M., Biemond, J.: Maximum likelihood image and blur identification: a unifying approach. Optical Engineering 29, 422–435 (1990a)

Lagendijk, R.D.L., Biemond, J., Boekee, D.E.: Identification and restoration of noisy blurred image using the expectation-maximization algorithm. IEEE Trans. Acoust, Speech, Signal Processing 38, 1180–1191 (1990b)

Lehr, J., Sibarita, J.-B., Chassery, J.-M.: Image restoration in X-Ray microscopy: PSF determination and biological applications. IEEE Trans. Image Processing 7, 258–263 (1998)

Li, X.: Blind image quality assessment. In: Proc. IEEE Int. Conf. Image Processing, NY, pp. I-449–452 (2002)

Luxen, M., Forstner, W.: Characterizing image quality: blind estimation of the point spread function from a single image. In: Proc. Photogrammetric Computer Vision, Austria, pp. A205–A210 (2002)

Marziliano, P., Winkler, S., Dufaux, F., Ebrahimi, T.: Perceptual blur and ringing metrics: application to JPEG2000. Signal Processing: Image Commun. 19, 163–172 (2004)

Reeves, S.J., Mersereau, R.M.: Blur identification by the method of generalized crossvalidation. IEEE Trans. Image Processing 1, 301–311 (1992)

Rom, R.: On the cepstrum of two-dimensional functions. IEEE Trans. Inform. Theory IT 21, 214–217 (1975)

Savakis, E., Trussell, H.J.: On the accuracy of PSF representation in image restoration. IEEE Trans. Image Processing 2, 252–259 (1993)

Sezan, M.I., Pavlovic, G., Tekalp, A.M., Erdem, A.T.: On modeling the focus blur in image restoration. In: Proc. IEEE Int. Conf Acoust., Speech, and Signal Proc., Toronto, pp. 2485–2488 (1991)

Subbarao, M., Tyan, J.-K.: The optimal focus measure for passive autofocusing and depth-from-focus. In: Proc. SPIE Int. Symp., Videometrics V, vol. 2598, pp. 89–99 (1995)

Wolin, D., Johnson, K., Kipman, Y.: The importance of objective analysis in image quality evaluation. In: Proc. Int. Conf. Digital Printing Technologies, pp. 603–606 (1998)

Wu, S., Chen, L., Lin, W., Jiang, L., Xiong, W., Ong, S.H.: An objective out-of-focus blur measurement. In: Proc. 5th Int. Conf. Inform., Commun. & Signal Processing, Thailand, pp. 334–338 (2005)

Wu, S., Lin, W., Xie, S., Lu, Z., Ong, E., Yao, S.: Blind blur assessment for vision-based applications. J. Visual Commun. Imag. Rep. 20, 231–241 (2009)

Ziou, D.: The influence of edge direction on the estimation of edge contrast and orientation. Patt. Recog. 34, 855–863 (2001)

A Unified Tensor Level Set Method for Image Segmentation

Xinbo Gao[1], Bin Wang[1], Dacheng Tao[2], and Xuelong Li[3]

[1] VIPS Lab, School of Electronic Engineering, Xidian University,
Xi'an 710071, P.R. China
xbgao@mail.xidian.edu.cn, bwang.xd@gmail.com

[2] School of Computer Engineering, Nanyang Technological University,
50 Nanyang Avenue, Blk N4, 639798, Singapore
dacheng.tao@gmail.com

[3] State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of
Optics and Precision Mechanics, Chinese Academy of Sciences,
Xi'an 710119, P.R. China
xuelong_li@opt.ac.cn

Summary. This paper presents a new unified level set model for multiple regional image segmentation. This model builds a unified tensor representation for comprehensively depicting each pixel in the image to be segmented, by which the image aligns itself with a tensor field composed of the elements in form of high order tensor. Then the multi-phase level set functions are evolved in this tensor field by introducing a new weighted distance function. When the evolution converges, the tensor field is partitioned, and meanwhile the image is segmented. The proposed model has following main advantages. Firstly, the unified tensor representation integrates the information from Gaussian smoothed image, which results the model is robust against noise, especially the salt and pepper noise. Secondly, the local geometric features involved into the unified representation increase the weight of boundaries in energy functional, which makes the model more easily to detect the edges in the image and obtain better performance on non-homogenous images. Thirdly, the model offers a general formula for energy functional which can deal with the data type varying from scalar to vector then to tensor, and this formula also unifies single and multi-phase level set methods. We applied the proposed method to synthetic, medical and natural images respectively and obtained promising performance.

Keywords: Gabor filter bank, geometric active contour, tensor subspace analysis, image segmentation, level set method and partial differential equation.

1 Introduction

Image segmentation, a process of subdividing an image into a series of non-intersected regions with approximately similar properties, is a fundamental

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 217–238.
springerlink.com

step in process of automatic image analysis, since it helps to identify, describe and understand different interested objects in images.

The field of image segmentation kept continuously developing for almost forty years and a number of algorithms came forth steadily in each year [38]. In recent years, active contours models are increasingly and widely used in image processing. These methods all need to initialize a closed curve, also called the evolving curve [3] [4] [20] [27], in the image to be segmented, and then evolve it by a partial differential equation (PDE) until the evolving curve converges. According to the representation of the evolving curve, active contours models can be classified as the explicit [5] [6] [10] [36] and the implicit [1] [2] [3] [9] [12] [17] [19] [20] [22] [23] [26] [27] categories. Snake [5] [10] [36], a typical explicit active contours model, uses parametric equations to explicitly represent the evolving curve. However, implicit active contours models, also called level set methods, implicitly represent the evolving curve by using the zero level set of a signed distance map defined in higher dimensional space. Compared with explicit active contours, implicit active contours have many advantages of which the most important is the capability to handle the topological changing in an easier way than explicit active contours models do [26].

Based on the way to model image, level set methods can be further categorized into either edge-based [1] [2] [9] [12] [17] [19] [22] or region-based [3] [4] [34] [35] algorithms. The former has to design an edge indicator to locate the edges in the image, but these edges are not always keeping closed and also do not always correspond to the boundaries of objects. Meanwhile, it's hard to prevent the evolving curve from leaking at some weak boundaries because of the presence of noise. The latter seems to be a better choice and indeed it attracts more and more research interests in recent years. Essentially, it regards the image domain as the composition of some regions with similarity properties, and the image is segmented by finding these similarity properties. Chan and Vese [3] proposed a region-based level set method, called CV method, which incorporates Mumford-Shah [15] [16] functional into level set framework to give a cartoon representation to the image. The main idea of this method is to find an optimum cartoon representation for the image. The evolving curve is driven by an energy functional incorporating a "fitting" item. This "fitting" item defines the extent that the cartoon representation approaches the given image. The CV method [3] is well extended in two ways generally as follows. Chan *et al.* [4] extended it to segment multi-channel images, and Wang and Vemuri [34] [35] extended it to segment tensor diffusion MRI images. Although these methods can evolve the curve in vector or tensor data field, they just take into account the pixel density depicted by a scalar, and that is not adequate to represent all image information. On the other side, Vese and Chan extended it to multi-phase level set (MCV) method [33] which did not provide a practical formula to deal with the case more than two level set functions being used. Zhao *e*t al. [37] extended it to multiple level set functions by adding a constraint into the energy functional to ensure one pixel just belongs to one level set function. Lie *et al.* [14] shared the same idea with [33] except for using binary level set functions which has to be regularized by using another constraint. These two methods utilized an unnatural way to describe the partitioned regions, which results they had to append some regularization to the energy functional, and that is not computational.

Aiming at the above problems, we employ Gabor filter bank [7] [8] [11] [28] to extract the local geometrical features, *e.g.*, scale, orientation, and then integrate these Gabor features and the intensity of the image and the Gaussian smoothed image into a unified tensor representation. There are three main reasons to consider the Gabor filter bank: 1) being Gaussian functions with different deviations, the envelopes of Gabor functions can provide multiple scale edge information,2) the amplitudes of Gabor functions provide gradient information, 3) for the orientation selectivity of Gabor basis function, the Gabor filter bank can capture abundant orientation information of object boundaries. This unified tensor preserves more information than a scalar used in other level set method [3] [4] [33]. Based on this tensor representation, we propose a unified tensor level set method which can accept tensors as input for high quality image segmentation.

The proposed model has several advantages as follows. Firstly, this model is robust for noisy images since the unified tensor representation incorporates the information of the Gaussian smoothed image. Secondly, because the tensor representation contains Gabor features, the model has ability to make texture segmentation. Thirdly, the model takes into account the gradient information that intrinsically increases the weights of boundaries in the energy functional, which makes the evolving curve stop at the boundaries easier and help to segment objects in non-homogenous background. Finally, the model is capable of dealing with data type varying from scalar to vector and to high order tensor and provides a general energy functional formula for single and multi-phase level set functions. The proposed method has been applied to a set of synthetic, medical and natural images. The result proves that the proposed method is robust against salt and pepper type noise, deals with the non-homogenous image better, has the capacity of orientation selectivity, and detects object more accurately.

The organization of the reminder in this paper is as follows. In Section II, we firstly describe the background of the CV model [3], its extensions to multichannel images [3] and the extension to 2-order symmetrical tensor field [34] [35], and then introduce Gabor filter bank. Section III details the proposed unified tensor level set model, involving the construction of unified tensor field and the tensor level set method as well as its special cases. Section IV contains the implementation of the proposed model, including the regularization of the Heaviside step and Dirac delta function, evolution function, and its numerical scheme. Section V assesses the proposed model in comparison with the CV model [3] and WV method [34] [35] on synthetic and medical images. Section VI gives the concluding remarks and future work.

2 Backgrounds

Gabor filter bank and region-based level set methods are both the basis of the proposed method. In this section, we firstly describe the gradual progress in region-based level set methods, and then introduce the Gabor filter bank utilized to build the unified tensor representation for pixel.

2.1 Region-Based Level Set Methods

Compared with edge-based level set methods, region-based level set methods have many advantages [4], *e.g.*, more robustness against noise and insensitivity to initial position of the evolving curve. Since Chan and Vese proposed the scalar CV method [3], region-based level set methods have been sequentially extended from scalar to vector, and then to tensor. In this process the data structure became more and more complex, but the basic idea is similar. The rest of this section describes these methods respectively. Let $u_0 : \Omega \to R$ be a given image, where $\Omega \subset R^2$ and is the image domain. Let $C \subset \Omega$ be a closed curve implicitly represented by the zero level set of a Lipschitz function $\phi : \Omega \to R$. In scalar CV method [3], the energy functional $E(c_1, c_2, C)$ is defined by

$$\begin{aligned} E\left(c_1, c_2, C\right) = & \mu \text{Length}\left(C\right) + v \text{Area}\left(C\right) \\ & + \lambda_1 \int_{in(C)} \left|u_0\left(x, y\right) - c_1\right|^2 dxdy \\ & + \lambda_2 \int_{out(C)} \left|u_0\left(x, y\right) - c_2\right|^2 dxdy, \end{aligned} \tag{1}$$

where $\mu \geq 0$, $\nu \geq 0$, $\lambda_1 > 0$ and $\lambda_2 > 0$ are constant parameters controlling the influences of different energy items. c_1 and c_2 are the intensity averages of the regions inside and outside the evolving curve respectively. Length(C) is the length of evolving curve and Area(C) is the area of the region inside the evolving curve. Chan *et al.* [4] proposed a region-based level set method for multi-channel images, *e.g.*, color image. This method can be regarded as a kind of vector CV method [4]. The energy functional is defined by

$$\begin{aligned} E\left(\bar{c}^+, \bar{c}^-, C\right) = & \mu \text{Length}\left(C\right) \\ & + \int_{in(\Omega)} \frac{1}{N} \sum_{i=1}^{N} \lambda_i^+ \left|u_{0,i}\left(x, y\right) - c_i^+\right|^2 dxdy \\ & + \int_{out(\Omega)} \frac{1}{N} \sum_{i=1}^{N} \lambda_i^- \left|u_{0,i}\left(x, y\right) - c_i^-\right|^2 dxdy, \end{aligned} \tag{2}$$

where $\bar{c}^+$ and $\bar{c}^-$ are the averages vector of the regions inside and outside the evolving curve respectively, and $u_{0,i}(x, y)$ is the ith channel image. Wang and Vemuri [34] [35] proposed a level set method for segmenting DT-MRIs being a multi-channel image arranged in symmetrical tensor form. The energy functional is

$$E\left(T_1,T_2,C\right)=\mu\text{Length}\left(C\right)+\int_{in(\Omega)}\text{dist}^2\left(T\left(x,y\right),T_1\right)dxdy+\int_{out(\Omega)}\text{dist}^2\left(T\left(x,y\right),T_2\right)dxdy, \tag{3}$$

where T_1 and T_2 are the averages inside and outside regions respectively, and they are both symmetrical matrix, *i.e.*, 2-order symmetrical tensor.dist$(\cdot)$ is the Frobenius norm of matrices.

To sum up, the region-based level set methods extend from single image [3] to multi-channel images [4], and then to symmetrical tensor images [34] [35]. But they basically use a scalar for representing each pixel in the given image or single channel image, and they do not provide a full and comprehensive representation for images. In following section, we introduce Gabor filter bank, and describe the process using Gabor filter bank to extract the local geometric features of the image.

2.2 Gabor Filter Bank

Human visual system is similar to a filter bank.Marcelja [18] and Daugman [7] [8] used Gabor functions to model the responses of the visual cortex. Daugman [7] [8] further developed 2D Gabor functions used by Lee [11] and Tao *et al.* [28] [29] [30] [31] [13] to give images a Gabor-based image representation.

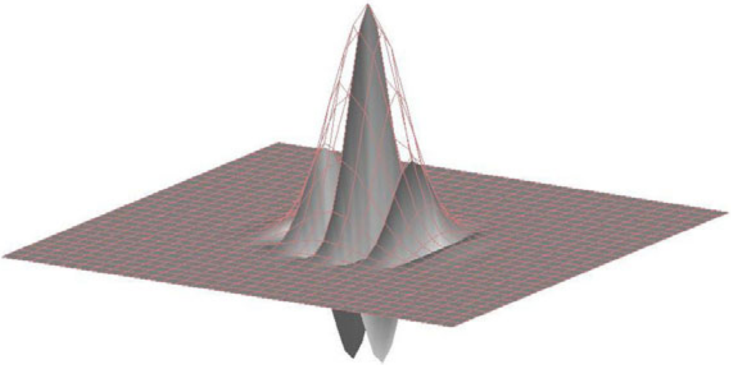


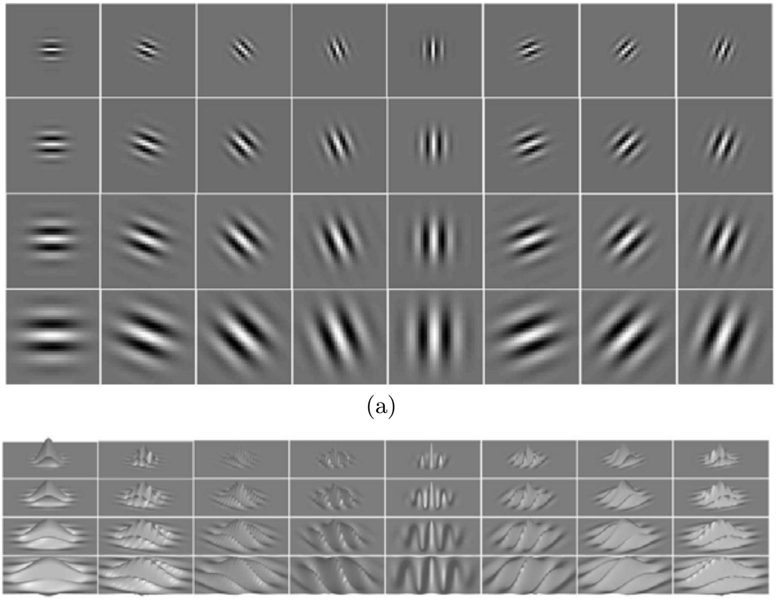
Fig. 1. The real part of a Gabor function with fixed direction and scale and its envelope represented by red grid line

2D Gabor function is defined as

$$G_{s,d}\left(\bar{x}\right)=\frac{||\bar{k}||}{\sigma^2}\cdot e^{-\frac{||\bar{k}||^2\cdot||\bar{x}||^2}{2\sigma^2}}\cdot\left[e^{i\bar{k}\cdot\bar{x}}-e^{-\frac{\sigma^2}{2}}\right], \tag{4}$$

where $\bar{k}=\frac{k_{max}}{f^s}e^{i\frac{\pi}{d_{max}}d}$ is the frequency vector determining the scales and directions of Gabor functions. $\bar{x}=(x,y)$ is the variable in a spatial domain.

σ is a parameter controlling the number of oscillations under the Gaussian envelope. s and d denote the scale and direction respectively. In fact, Gabor function is the product of a Gaussian function and a complex wave. As shown in Fig. 1, the real part of a 2D Gabor function with a fixed direction and scale has a Gaussian envelope represented by the red grid. When we choose different scales and directions, a series of Gabor functions are obtained. In our model, $k_{max} = \pi/2$, $f = \sqrt{2}$, $\sigma = 3\pi/2$, $d_{max} = 8$, $d \in \{0, 1, 2, 3, 4, 5, 6, 7\}$ and $s \in \{0, 1, 2, 3\}$, then a set of Gabor functions with four scales and eight directions are obtained and illustrated in Fig. 2.



(a)

(b)

Fig. 2. The real part of Gabor functions with four different scales and eight different directions. (a) Gabor functions in 2D. (b) Gabor functions in 3D. The scale from 0 to 3, stepping 1. And the orientation varies from 0 to 7, stepping 1.

From Fig. 2, we can visually find that these Gabor functions have great capacity for spatial localization and orientation selectivity. In our model, the Gabor-based image representation [7] [8] [11] [28] is obtained by convolving these functions with the image to be segmented. The convolution outputs contain gradient and orientation information which are incorporated into the unified tensor representation as a matrix component.

3 The Unified Tensor Level Set

This section firstly details the construction of the unified pixel tensor representation and the tensor field, and then describes the energy functional and gradient flow (*i.e.*, the evolution function) of the proposed unified tensor level set method. Finally the three special cases of the proposed model are discussed.

3.1 The Unified Tensor Representation

To segment an image more accurately, more overall information in the given image should be involved by segmentation algorithms, and more suitable representation for this information should be used to describe the image. The early region-based level set methods [3] [4] just use a scalar, *e.g.*, intensity, for representing pixel. Wang and Vemuri's method [34] [35] has ability to segment DT-MRIs being a multi-channel image arranged in symmetrical tensor. This method also can deal with the tensor data (being similar to DT-MRIs) created by the local structural tensor (LST) [32] [39] to segment texture images, but this method ignores an important feature, *i.e.*, the intensity of pixel. In short, none of these methods provides a relative comprehensive representation for images. Moreover, [4] [34] [35] mainly focus on multi-channel images or DT-MRIs, and not on an appropriate representation for a single image. By introducing Gabor features, we build a unified tensor representation for pixels. This tensor representation contains more in-formation (*e.g.* average intensity, gradient, and orientation.), and is relatively overall. As illustrated in left side of Fig. 3, the construction of unified tensor representation contains following steps.

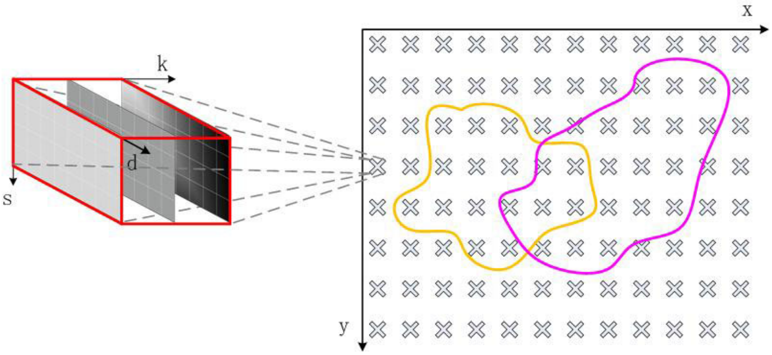


Fig. 3. The pixel tensor representation is zoomed out in left, and the two curves with different color evolve in the tensor field composed of the element in the form of the unified tensor representation

Step 1: To make our model more robust against noise, the initial image is smoothed with Gaussian filter bank, and then the smoothed image is included into the unified tensor representation as a matrix written as

$$\left[t_{x,y}^{s,d,k=1}\right]_{S\times D}=\frac{1}{\sqrt{SD}}\begin{bmatrix} G_{\sigma_1}\left(u_{x,y}^{0}\right) & \cdots & G_{\sigma_1}\left(u_{x,y}^{0}\right)\\ \vdots & \vdots & \vdots \\ G_{\sigma_S}\left(u_{x,y}^{0}\right) & \cdots & G_{\sigma_S}\left(u_{x,y}^{0}\right)\end{bmatrix}_{S\times D}, \tag{5}$$

where $t_{x,y}^{s,d,k}$ is an element in the high order tensor representation. s denotes the scale and its maximum number is S and $S=4$. d is the direction and its maximum number is D and $D=8$. $u_{x,y}^{0}$ is the initial image, and $G_{\sigma_{1,\ldots,S}}(\cdot)$ are the outputs generated by using the Gaussian functions with different standard deviations convolving with the initial image. The standard deviations correspond with the standard deviations used by Gabor filter bank.

Step 2: The intensity of each pixel in the initial image to be segmented is embedded into the unified tensor representation, and the process is formulated by

$$\left[t_{x,y}^{s,d,k=2}\right]_{S\times D}=\tfrac{1}{\sqrt{S\times D}}\left[u_{x,y}^{0}\right]_{S\times D}. \tag{6}$$

Step 3: The Gabor features are used to represent gradient and orientation of images. Having Gabor functions defined by (4) convolved with the initial image, the Gabor-based image representation in $R^{M\times N\times S\times D}$ is obtained. Thus, a rule of correspondence between each pixel of the initial image and a matrix in $R^{S\times D}$ is built as follows

$$\begin{array}{rl}\text{Gabor}\left(u_{x,y}^{0}\right) & =\left|u_{x,y}^{0}*G_{s,d}\left(x,y\right)\right| \\ \left[t_{x,y}^{s,d,k=3}\right]_{S\times D} & =\left[\text{Gabor}\left(u_{x,y}^{0}\right)\right]_{S\times D},\end{array} \tag{7}$$

where $G_{s,d}(\cdot)$ is the Gabor function defined by (4), and Gabor$(\cdot)$ is the outputs generated by convolving Gabor functions with the image respectively. By steps 1-3, an image is projected on a 5-order tensor in $R^{M\times N\times 4\times S\times D}$. The first two indices give the pixel location and the last three indices give the 3-order tensor representation. That is to say, the third index gives the value of the scale, and the fourth index gives the direction. Since the image varies from coarse to fine along the fifth index, we call this index the fineness. Thus, each pixel in the image is represented by a 3-order tensor in $R^{S\times D\times K}$ which contains the densities in different fineness, the gradient and orientation information extracted from the neighborhood of the pixel by using Gabor filter bank.

Step 4: To reduce the dimensionality of the unified tensor representation, the OTA [25], a kind of generalization of principle component analysis for tensors, is applied on the Gabor-based image representation, the dimensionality of the tensor representation is reduced as follows,

$$\left[t_{x,y}^{s,d,k=3}\right]_{S'\times D'}=\left[OTA\left(u_{x,y}^{0}*GT_{s,d}\left(x,y\right)\right)\right]_{S'\times D'}, \tag{8}$$

where $OTA(\cdot)$ denotes the output of offline tensor analysis (OTA). After OTA, $S\times D$ is reduced to $S'\times D'$, where $S'<S$ and $D'<D$. The unified tensor representation becomes the tensor in $R^{S'\times D'\times K}$, and the cost of computation is reduced from $O((S\times D\times K)\times N)$ to $O((S'\times D'\times K)\times N)$ per

time step, where N is the numbers of the pixels in the image. In a word, we utilize (5)-(8) to build a rule of correspondence between a pixel and a unified tensor. This tensor provides pixels a more comprehensive and flexible tensor description which results in a more accurate segmentation.

3.2 The Unified Tensor Level Set Method

In this section we propose the unified tensor level set method, and detail the energy functional, the associated evolution equation and a weighted distance function. Let us define a tensor T in $R^{M\times N\times S\times D\times K}$, and then unfold T along mode-1 and mod-2 simultaneously. Thus T becomes a tensor field [25] [21] with elements in form of 3-order tensor in $R^{S\times D\times K}$ and each element corresponds to a pixel in the image to be segmented. There are I evolving curves S_i in $\Omega \in R^{M\times N}$ that divide the field T into $J = 2^I$ regions. Representing these regions with their mean values respectively, we obtain a cartoon representation of the field T . The fitting error between this cartoon representation and the field T is E_e . Adding a regularization item, the energy functional is defined as

$$E\left(S\right) = E_g + E_e. \tag{9}$$

where E_g denotes the geometrical feature of the evolving curve, *i.e.*, the length of the curve. Accompanied with the decreasing of this energy, the fitting term is minimized and the segmentation result is obtained. This process is formulated by

$$\inf_{S_i} \left\{E\left(S_i\right)\right\}. \tag{10}$$

The evolving curve S_i in tensor field, illustrated in Fig.3, is implicitly represented by the zero set of the function $\phi_i : \Omega \to R$, and written as [3]

$$S_i = \{(x,y) : \phi_i\left(x,y\right) = 0\}. \tag{11}$$

Substituting (10) and (11) into (9), the energy functional is rewritten as

$$\begin{aligned} E\left(\Phi, C\right) &= \sum_{i=1}^{I} \mu_i \mathrm{Length}\left(\phi_i\right) + \sum_{j=1}^{J} E_j\left(c_j, \chi_j\right) \\ &= \sum_{i=1}^{I} \mu_i \int_\Omega \delta\left(\phi_i\left(x,y\right)\right)\left|\nabla\phi_i\left(x,y\right)\right| dxdy \\ &\quad + \sum_{j=1}^{J} \int_\Omega \mathrm{dist}^2_{x,y}\left(t^{s,d,k}_{x,y}, c^{s,d,k}_j\right)\chi_j\left(\Phi\right) dxdy, \end{aligned} \tag{12}$$

where $\Phi = \{\phi_1, \ldots, \phi_I\}$, $C = \{c_1, \ldots, c_J\}$, and $t^{s,d,k}_{x,y}$ is the element in the tensor field T . Length(ϕ_i) is the length of the evolving curve S_i , a regularization item, making the curve more smoothed, and it is computed by

$$\begin{aligned} \mathrm{Length}\left(\phi_i\right) &= \int_\Omega \left|\nabla H\left(\phi_i\left(x,y\right)\right)\right| dxdy \\ &= \int_\Omega \delta\left(\phi_i\left(x,y\right)\right)\left|\nabla\phi_i\left(x,y\right)\right| dxdy. \end{aligned} \tag{13}$$

$H(\cdot)$ is the Heaviside step function, *i.e.*,

$$H\left(x\right)=\begin{cases}1, & if\ x\geq 0\\ 0, & if\ x<0.\end{cases} \tag{14}$$

$\delta(\cdot)$ is the Dirac delta function, *i.e.*, the differential of Heaviside step function, defined as

$$\delta\left(x\right)=\frac{d}{dx}H\left(x\right). \tag{15}$$

χ_j in (12) denotes the sub-region and formulated as

$$\chi_j\left(\varPhi\right)=\prod_{l=1}^{I}\Big(\left(1-b_l\right)-\left(-1\right)^{b_l}H\left(\phi_l\right)\Big), \tag{16}$$

where $[b_l]=dec2binvec\left(j\right)$ and converts the denary number j into its binary notation. $c_j^{s,d,k}$ is the mean tensor inside the sub-region χ_j and defined as

$$c_j^{s,d,k}=\int_{\Omega}t_{x,y}^{s,d,k}\chi_j\left(\varPhi\right)dxdy\Big/\int_{\Omega}\chi_j\left(\varPhi\right)dxdy. \tag{17}$$

The distance function is defined as

$$\mathrm{dist}_{x,y}\left(T_{x,y}^{s,d,k},c_j^{s,d,k}\right)=\sqrt{\sum_{s=1}^{S}\alpha_s\sum_{d=1}^{D}\beta_d\sum_{k=1}^{K}\gamma_k\left(T_{x,y}^{s,d,k}-C_j^{s,d,k}\right)^2}, \tag{18}$$

where

$$\alpha_s\geq 0,\sum_{s=1}^{S}\alpha_s=1;\quad \beta_d\geq 0,\sum_{d=1}^{D}\beta_d=1;\quad \gamma_k\geq 0,\sum_{k=1}^{K}\gamma_k=1. \tag{19}$$

The definition of distance can be considered as a kind of weighted Frobenius norm or Hilbert-Schmidt norm. Replacing the energy in (10) by (12), (10) is rewritten as

$$\inf_{\varPhi,C}\left\{E\left(\varPhi,C\right)\right\}. \tag{20}$$

That means the energy functional, *i.e.*, (12), should be minimized accompanied with the evolution of the level set functions $\varPhi$, which is a problem of calculus of variations. In order to solve this problem, we firstly fix mean values C , and then compute the associate Euler-Lagrange equation for each unknown level set function ϕ_i . By adding an artificial time variable $t\geq 0$, the evolution equation is obtained as

$$\begin{aligned}\tfrac{\partial\phi_i}{\partial t}&=\mu\delta\left(\phi_i\right)div\left(\tfrac{\nabla\phi_i}{|\nabla\phi_i|}\right)\\ &\quad+\sum_{j=1}^{J}\tfrac{\partial\chi_j(\varPhi)}{\partial\phi_i}\sum_{s=1}^{S}\alpha_s\sum_{d=1}^{D}\beta_d\sum_{k=1}^{K}\gamma_k\left(t_{x,y}^{s,d,k}-c_j^{s,d,k}\right)^2,\end{aligned} \tag{21}$$

where $div(\cdot)$ is the mean curvature of each evolving curve,i $\partial\chi_j/\partial\phi_i$ is calculated by

$$\frac{\partial\chi_j(\Phi)}{\partial\phi_i} = -(-1)^{b_l}\delta(\phi_i)\prod_{l=1(l\neq i)}^{I}\Big((1-b_k)-(-1)^{b_l}H(\phi_l)\Big). \tag{22}$$

After computing the level set functions, the mean values are updated by (17).

3.3 Special Cases of the Proposed Model

The proposed unified tensor level set model is a generalized version of several level set method [3] [4] [33] [34] [35]. This section presents four special cases of the proposed model. By taking integral sign into the summation, the external force $E_e(\Phi, C)$ in (12) is rewritten as

$$\begin{aligned} E_e(\Phi, C) &= \sum_{j=1}^{J}\int_\Omega\left[\sum_{s=1}^{S}\alpha_s\sum_{d=1}^{D}\beta_d\sum_{k=1}^{K}\gamma_k\left(T_{x,y}^{s,d,k}-c_{+/-}^{s,d,k}\right)^2\right]\chi_j(\Phi)\,dxdy \\ &= \sum_{j=1}^{J}\sum_{s=1}^{S}\alpha_s\sum_{d=1}^{D}\beta_d\sum_{k=1}^{K}\gamma_k\int_\Omega\left(T_{x,y}^{s,d,k}-c_{+/-}^{s,d,k}\right)^2\chi_j(\Phi)\,dxdy \\ &= \sum_{j=1}^{J}\sum_{s=1}^{S}\alpha_s\sum_{d=1}^{D}\beta_d\sum_{k=1}^{K}\gamma_k E_j^{s,d,k}. \end{aligned} \tag{23}$$

Substituting (23) into (12), the energy functional is rewritten as

$$E(\Phi, C) = \sum_{i=1}^{I}\mu_i\mathrm{Length}(\phi_i) + \sum_{j=1}^{J}\sum_{s=1}^{S}\alpha_s\sum_{d=1}^{D}\beta_d\sum_{k=1}^{K}\gamma_k E_j^{s,d,k}, \tag{24}$$

where i denotes the i th level set function and I is the number of level set functions, j the index of sub-regions partitioned by level set functions and J the maximum number and $J = 2^I$. s denotes scale, d direction, and k fineness, and the maximum number of these parameters are $S = 4$, $D = 8$, and $K = 3$ respectively. Applying this method to the unified tensor field, we can reduce the unified tensor level sets into the following cases.

Case 1: When$I = 1$ and $J = 2^I = 2$, then there is only one level set function evolving in the field to be segmented. Meanwhile, if $S = D = K = 1$ and $\alpha_1 = \beta_1 = \gamma_1 = 1$, $T_{x,y}^{s,d,k}$ is simplified as a scalar. Thus the proposed model reduces into the scalar CV model [3].

Case 2: When $I = 1$ and $J = 2^I = 2$, then there is only one level set function evolving in the field to be segmented. Meanwhile, if $S = D = 1$ and $\alpha_1 = \beta_1 = 1$, $T_{x,y}^{s,d,k}$ is simplified as a vector. Thus the proposed model reduces into the vector CV model [4].

Case 3: When $I = 1$ and $J = 2^I = 2$, then there is only one level set function evolving in the field to be segmented. Meanwhile, if $K = 1$ and

$S = D = 2$, $T_{x,y}^{s,d,k}$ is simplified as 2-order symmetry tensor. Additionally if $\alpha = [1, 1, \cdots, 1]$, $\beta = [1, 1, \cdots, 1]$ and $T_{x,y} = \begin{bmatrix} \frac{\partial x}{\partial x} u_{x,y}^0 \cdot \frac{\partial}{\partial x} u_{x,y}^0 & \frac{\partial}{\partial x} u_{x,y}^0 \cdot \frac{\partial}{\partial y} u_{x,y}^0 \\ \frac{\partial}{\partial y} u_{x,y}^0 \cdot \frac{\partial}{\partial x} u_{x,y}^0 & \frac{\partial}{\partial y} u_{x,y}^0 \cdot \frac{\partial}{\partial y} u_{x,y}^0 \end{bmatrix}$ (*i.e.*, LST [32] [39]), the proposed model reduces into the model [34] [35] proposed by Wang and Vemuri.

Case 4: When $I = 2$ and $J = 2^I = 4$, then there are two level set functions evolving in the field to be segmented. Meanwhile, if $S = D = K = 1$ and $\alpha_1 = \beta_1 = \gamma_1 = 1$, $T_{x,y}^{s,d,k}$ is simplified as a scalar. Thus the proposed model reduces into the MCV model [33].

In brief, the proposed tensor level set method has capacity to deal with the data type varying from scalar to vector and then to tensor, and integrate the single and multi-phase level set functions into a general formula. Many methods [3] [4] [33] [34] can be obtained by simplifying the proposed model.

4 Implementation

The evolution of level set function is driven by the evolution equation, *i.e.*,(21). To solve this equation, we firstly regularize the Heaviside step function and the Dirac delta function, and then choose the numerical scheme for these PDEs for different unknown level set functions.

4.1 Regularization of the Heaviside and Dirac Functions

Heaviside step function and Dirac delta function are discontinuous functions. To minimize the energy functional defined by (12), the Heaviside step function and the Dirac delta function should be regularized to be continuous.

$$\begin{cases} H_{1,\epsilon} = \begin{cases} 1 & x > \epsilon, \\ 0 & x < -\epsilon, \\ \frac{1}{2}\left(1 + \frac{x}{\epsilon} + \frac{1}{\pi}\sin(\frac{\pi x}{\epsilon})\right) & |x| \leq \epsilon. \end{cases} \\ \delta_{1,\epsilon} = H'_{1,\epsilon} = \begin{cases} 1 & x > \epsilon, \\ 0 & x < -\epsilon, \\ \frac{1}{2\epsilon}\left(1 + \cos(\frac{\pi x}{\epsilon})\right) & |x| \leq \epsilon. \end{cases} \end{cases} \tag{25}$$

$$\begin{cases} H_{2,\epsilon} = \frac{1}{2}\left(1 + \frac{2}{\pi}\arctan(\frac{x}{\epsilon})\right) \\ \delta_{2,\epsilon} = H'_{2,\epsilon} = \frac{1}{\pi} \cdot \frac{\epsilon^2}{(x^2+\epsilon^2)}. \end{cases} \tag{26}$$

In practice, (25) and (26), illustrated in Fig. 4, are both applicable to the approximations of the Heaviside and Dirac functions. Although (25) looks more similar to the Heaviside and Dirac function than (26) does, our model employs (26) to replace the corresponding function, since (25) has small support interval which sometimes causes the whole algorithm more easily compute a local minimum value of the energy. For more details, please refer to [3].

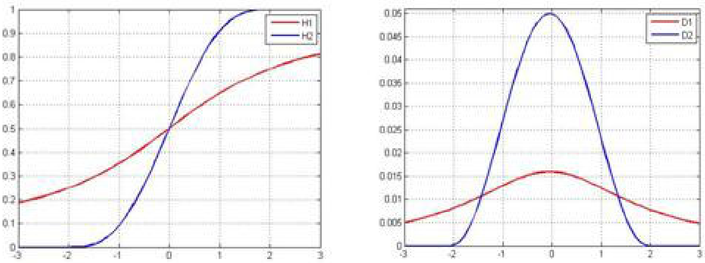


Fig. 4. Two different regularization of the Heaviside step and Dirac delta functions with $\varepsilon = 2.0$. The Heaviside step function is plotted with red line, and the Dirac delta function is plotted with blue line.

4.2 Numerical Scheme

There are many types of numerical scheme for solution of PDEs, *e.g.*, explicit scheme [1] [19] [20], implicit scheme [3]. Here, Since the semi-implicit scheme [4] [24] has the longer time step than explicit scheme does and it is more easily implemented than the implicit scheme, this kind of scheme is used in the proposed model. The semi-implicit numerical scheme for the evolution equation,*i.e.*, (17), is

$$\begin{aligned}\phi^{n+1}_{i,x,y} = \tfrac{1}{C}\Big[&m(C_1\phi^n_{i,x+1,y} + C_2\phi^n_{i,x-1,y} + C_3\phi^n_{i,x,y+1} + C_4\phi^n_{i,x,y-1}) \\ &-\Delta t\delta_\epsilon(\phi^n_{i,x,y})\sum_{j=1}^{J}\frac{\partial\chi_j(\Phi)}{\partial\phi_i}\sum_{s=1}^{S}\alpha_s\sum_{d=1}^{D}\beta_d\sum_{k=1}^{K}\gamma_k(t^{s,d,k}_{x,y} - c^{s,d,k}_j)^2\Big],\end{aligned} \tag{27}$$

where Δt is the time step,

$$m = \frac{\Delta t}{h^2}\delta_\varepsilon\left(\phi^n_{x,y}\right)\mu, \tag{28}$$

and $C = 1 + m\left(C_1 + C_2 + C_3 + C_4\right)$, where

$$\begin{aligned}C_1 &= \frac{1}{\sqrt{\left(\frac{\phi^n_{i,x+1,y} - \phi^n_{x,y}}{h}\right)^2 + \left(\frac{\phi^n_{i,x,y+1} - \phi^n_{i,x,y-1}}{2h}\right)^2}},\\ C_2 &= \frac{1}{\sqrt{\left(\frac{\phi^n_{i,x,y} - \phi^n_{i,x-1,y}}{h}\right)^2 + \left(\frac{\phi^n_{i,x-1,y+1} - \phi^n_{i,x-1,y-1}}{2h}\right)^2}},\\ C_3 &= \frac{1}{\sqrt{\left(\frac{\phi^n_{i,x+1,y} - \phi^n_{i,x-1,y}}{2h}\right)^2 + \left(\frac{\phi^n_{i,x,y+1} - \phi^n_{i,x,y}}{h}\right)^2}},\\ C_4 &= \frac{1}{\sqrt{\left(\frac{\phi^n_{i,x+1,y-1} - \phi^n_{i,x-1,y-1}}{2h}\right)^2 + \left(\frac{\phi^n_{i,x,y} - \phi^n_{i,x,y-1}}{h}\right)^2}}.\end{aligned} \tag{29}$$

The mean values are updated respectively with following regularized equation, *i.e.*,

$$c^{s,d,k}_j = \frac{\int_\Omega t^{s,d,k}_{x,y}\chi_j\left(\phi\right)dxdy}{\int_\Omega \chi_j\left(\phi\right)dxdy}. \tag{30}$$

5 Experimental Results

We conducted several experiments on synthetic, medical and natural images to illustrate the effectiveness of the proposed unified tensor level set method compared with the CV method [3] and the WV method [34].

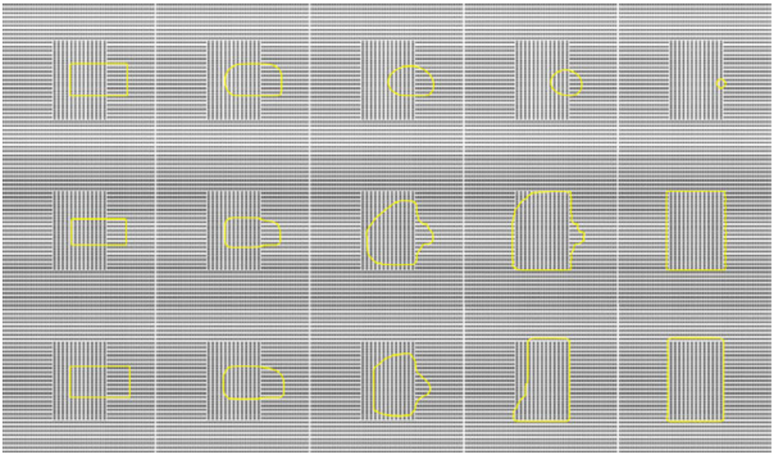


Fig. 5. The first to the final rows represent the evolution applying the CV method [3], the WV method [34] and the proposed method on a texture image with a rectangular object. The CV method [3] fails to segment objects correctly, but the WV method [34] and the proposed method do.

In experiment 1, the CV method [3], the WV method [34] and the proposed method were applied to a synthetic texture images. In this image, the only difference between object and background is the orientation, as shown in Fig. 5. According to [3], we replace intensities in the CV method [3] by $O(x, y) = \tan^{-1}(I'_y/I'_x)$ to represent the orientation, but the method does not work well for the image. Using LST [32] [39] for representing the local texture, the WV method [34] correctly segments the object. Due to the incorporation of the Gabor features, the proposed method is sensitive to specialized orientations, and also obtains the desired results.

Experiment 1 illustrates that the WV method [34] and the proposed method both have the feature of the orientation selectivity. Though, eight orientations are used in unified tensor representation, but this is not the maximum number that Gabor filter bank supports. If necessary, we can use more orientations to accurately extract the orientation information in images.

In Experiment 2, the image to be segmented is composed of three paper-cut snowflakes with different intensity. The number of objects is more than one, so the MCV method [33] is used to compare with the proposed method. The images are polluted by the salt and pepper noise. Fig. 6 shows that

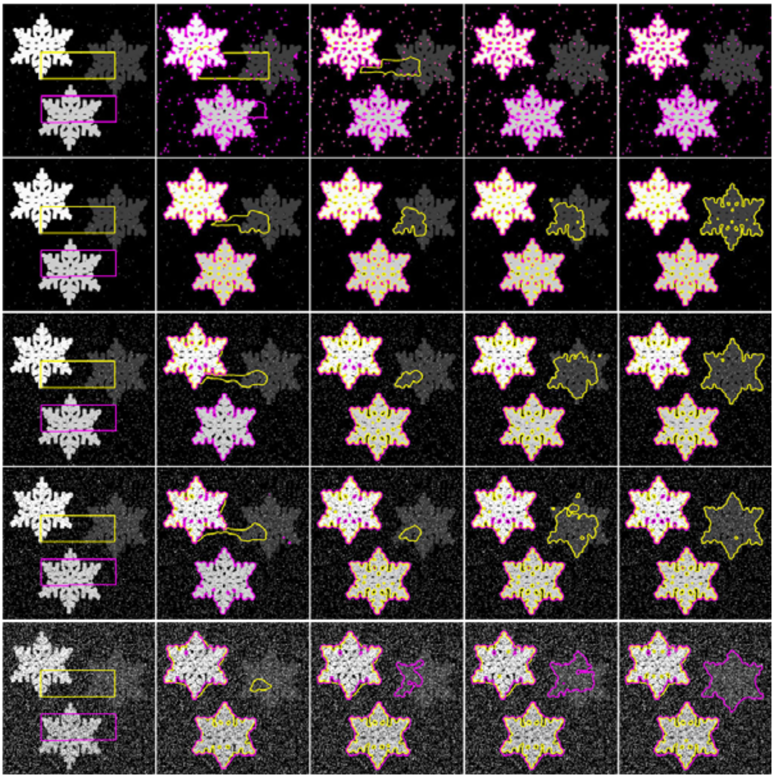


Fig. 6. The first row represents the evolution applying the MCV method [33] on the image with the noise density equaling 0.005, The second row to the fifth row represent the evolution applying the proposed method on the different images with the noise density 0.005, 0.05, 0.1 and 0.3 respectively. The noise adding into the image is the salt and pepper noise. The MCV method [33] is not robust against the noise even with the smallest noise density 0.005 in this experiment, but the proposed method is relatively robust against the noise. It can correctly segment the image, when the noise density increases to 0.3.

even the noise density is very small, *i.e.*, equals 0.005, the MCV method [33] still wrongly classifies the positive impulse points as the objects, and misses the snowflake with the darkest grayscale. The proposed method correctly segments all snowflakes from the background when the noise density increases from 0.005 to 0.05 and then to 0.1. Even when the noise density increases to 0.3, the proposed method still can demarcate all snowflakes.

Experiment 2 shows that the proposed method is more robust against the salt and pepper noise than the MCV method [33], since the proposed method reduces the influence from the salt and pepper noise by involving the intensity of the associated smoothed image.

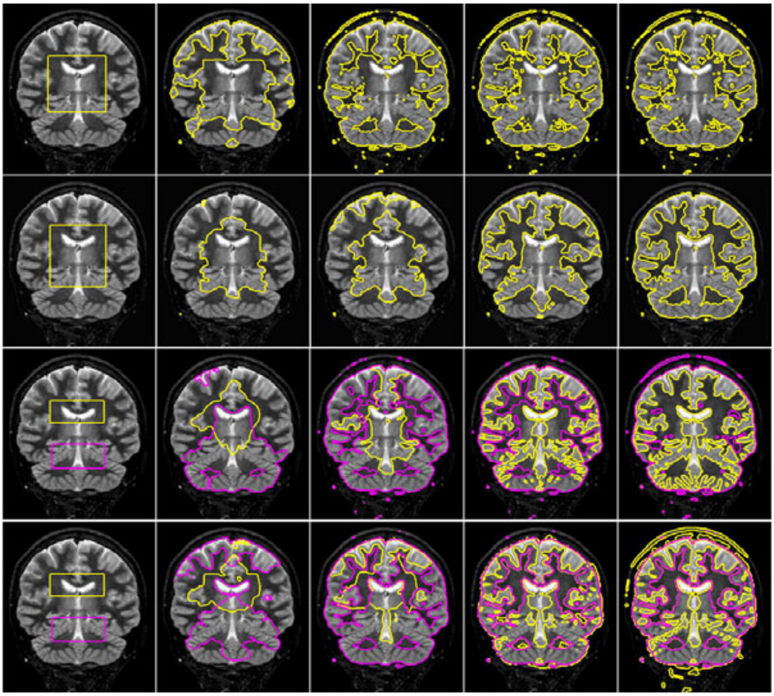


Fig. 7. The first and the second rows represent the evolution using the CV method [3] and the proposed method with one level set function, respectively. The third and fourth rows represent the evolution using MCV method [33] and the proposed method with two level set functions. In the two cases, the proposed method obtains better segmentation result.

Experiment 3 applied the CV method [3] , the MCV method [33] and the proposed method with different number of level set functions on a real Magnetic Resonance Image (MRI) of human brain, as shown in Fig. 7. The boundaries detected by CV method [3] are not smooth, and some real boundaries are missed. While the proposed method with one level set function correctly detects the boundaries and obtains better performance. The MCV method [33] obtains more accurate segmentation than CV method [3] does, but it still cannot correctly separate the cerebrospinal fluid from the grey matter, while the proposed method with two level set functions does. That is because the Gabor features contains gradient information, the weights of pixels on the boundaries in the energy functional (*i.e.*, (12)) are essentially increased so that the evolving curve can stop at boundaries easier.

Experiment 3 illustrates the proposed method is effective to segment real MRIs, and the segmentation result is more accurate than that of the CV method [3] and MCV method [33]. Additionally the result segmented by

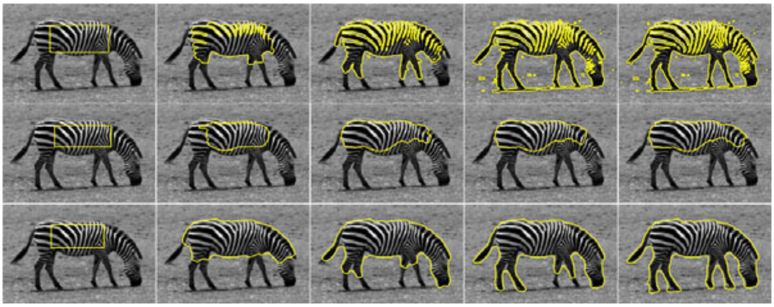


Fig. 8. The first to the last rows represent the evolution by using the CV method [3], the WV method [34] and the proposed method, respectively. The CV method [3] fails to segment the zebra, the WV method [34] also does not segment the zebra correctly, and the proposed approach can well segment the zebra from the grassland because the proposed unified tensor considers the local texture.

the proposed method looks more rational because the Gabor-based image representation coincides with human vision system.

In experiment 4, the CV method [3], the WV method [34] and the proposed method were applied to an image containing a zebra on the grassland respectively, as shown in Fig. 8. Because the proposed unified tensor contains the Gabor features, the image intensity and the Gaussian smoothed images, so it can duly separate the zebra from the grassland. However, because the CV method [3] only considers the intensity information, it cannot work as well as the proposed method. LST [32] [39] used by the WV method [34] lacks scale information, which results in the failure of the segmentation of zebra.

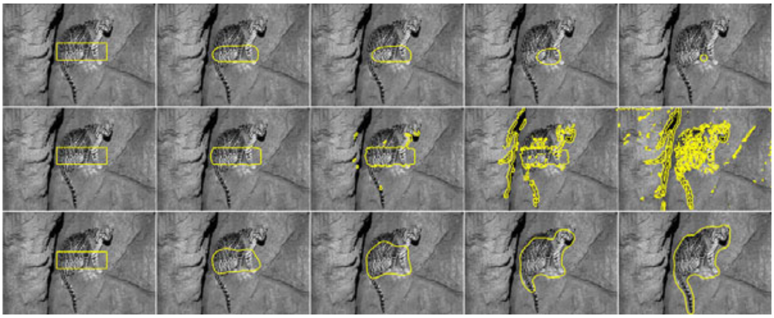


Fig. 9. The first to the last rows represent the evolution by using the CV method [3], the WV method [34] and the proposed method, respectively. The CV method [3] fails to segment the zebra, the WV method [34] also does not segment the zebra correctly, and the proposed approach can well segment the zebra from the grassland because the proposed unified tensor considers the local texture.

In experiment 5, the CV method [3] , the WV method [34] and the proposed method were applied on a natural image of a cat on rocks, as shown in Fig. 9. The CV method [3] does not correctly outline the cat, since it just considers the intensity of the pixels. However, the intensities are very close in this image. The WV method [34] also does not correctly demarcate the cat because of the weakness of the LST [32] [39]. Involving Gabor features, the proposed method makes the evolving curve stop at boundaries easier. Meanwhile, Gabor filter bank subtract the DC component of images to make the proposed method insensitive to the illumination. So the proposed method segments the image correctly.

In experiment 6, the CV method [3], the WV method [34] and the proposed method were applied to an image containing a leopard in the underbrush, as shown in Fig. 10. Because the proposed unified tensor contains the Gabor features, the image intensity and the Gaussian smoothed images, so it can duly separate the leopard from the underbrush. However, because the CV method [3] only considers the intensity information, it cannot work as well as the proposed method. The WV method [34] fails to segment the leopard because of the same reason as the above experiment.

Experiment 4, 5 and 6 suggest that the unified tensor representation is effective in segmenting the real texture image, *e.g.*, zebra and leopard images.

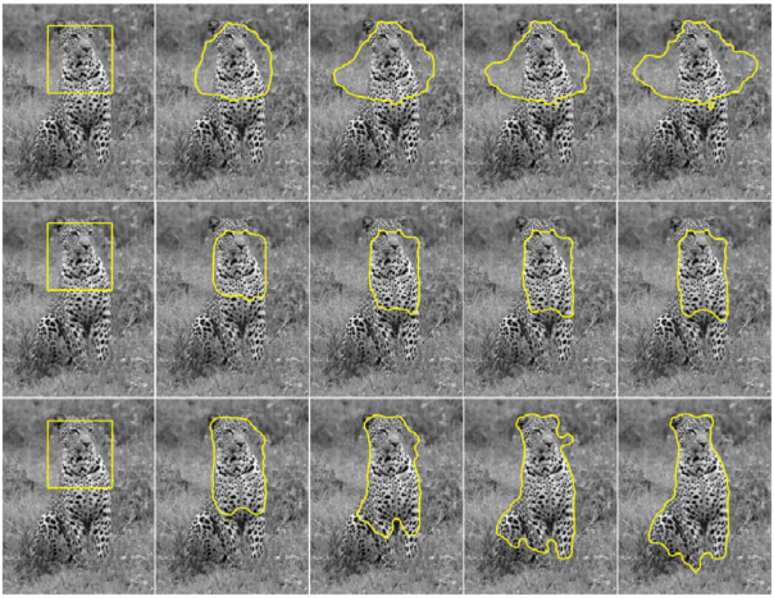


Fig. 10. The first to the last rows represent the evolution by using the CV method [3], the WV method [34] and the proposed method, respectively. The CV method [3] and the WV method [34] cannot separate the leopard from the grass land, but the proposed method almost does.

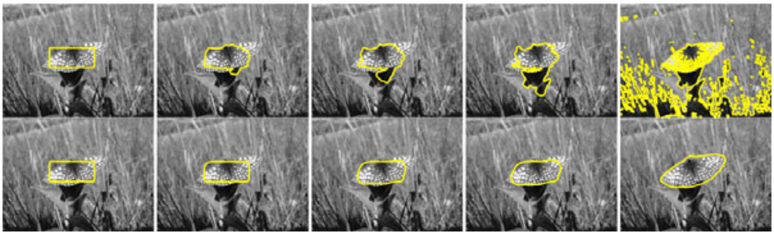


Fig. 11. The first and the second rows represent the evolution using the CV method [3] and the proposed method, respectively. The CV method [3] cannot separate the butterfly from the grass, but the proposed method almost does.

This is because Gabor features incorporating into the unified tensor describe the local texture better than LST [32] [39] used by the WV method [34].

Experiment 7 applied the CV method [3] and the proposed methods on a natural image with a butterfly settling on the one of big plant, as shown in Fig.11. The CV method [3] cannot separate the butterfly from the grassland, whereas the proposed basically does for the unified tensor representation being more comprehensive and containing more information than a scalar, *i.e.*, the intensity. This representation results in a correct segmentation performance.

Experiment 8 applied the CV method [3] and the proposed methods on a natural image with a horse running on the beach. The CV method [3] cannot completely separate the horse from the background, whereas the proposed basically does. This is because the proposed method uses a unified tensor representing each pixel in the image. This tensor contains information from the smoothed image which smoothes the weak boundaries in background, meanwhile the gradient and orientation extracted from the neighbor of pixels give a more accurate depiction to pixels. These all result in a natural result.

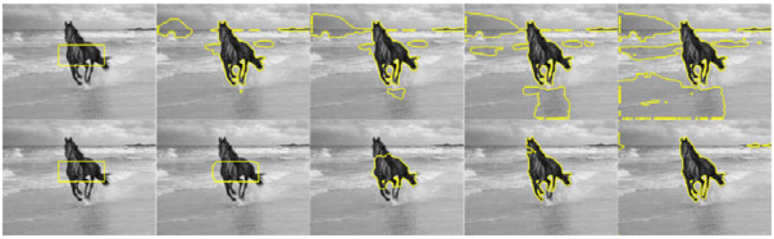


Fig. 12. The first and the second rows represent the evolution using the CV method [3] and the proposed method, respectively. The CV method [3] cannot separate the horse from the ocean wave and the beach, but the proposed method almost does.

Experiment 7 and 8 show that the proposed method is effective in the segmenting natural images, especially for the natural images with the complex background.

6 Conclusions

In this paper, we built a unified tensor representation for pixels to comprehensively depict the information of the image, and then provide a unified tensor level set method by proposing a weighted tensor distance definition. This method can deal with the data type varying from scalar to vector then to tensor, and integrates the single and multi-phase level set methods into a unified framework. By involving Gabor features into the unified tensor representation, our model has the capacity of orientation selectivity and better sensitivity to gradient. Meanwhile, by incorporating intensity in different fineness into the tensor pixel representation, the proposed method is more robust against noise, especially against the salt and pepper type noise. For the future work, more images will be used to check the effectiveness of the proposed method.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (60771068, 60702061, 60832005), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) in China and the National Laboratory of Automatic Target Recognition, Shenzhen University, China, the Program for Changjiang Scholars and innovative Research Team in University of China (IRT0645), and 100 Talents Program of The Chinese Academy of Sciences.

References

1. Adalsteinsson, D., Sethian, J.A.: A fast level set method for propagating interfaces. J. Computational Physics 118(2), 269–277 (1995)
2. Chopp, D.L.: Computing minimal surfaces via level set curvature flow. J. Computational Physics 106(1), 77–91 (1993)
3. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Processing 10(2), 266–277 (2001)
4. Chan, T.F., Sandberg, B.Y., Vese, L.A.: Active contours without edges for vector-valued images. J. Visual Communication and Image Representation 11(2), 130–141 (2000)
5. Chesnaud, C., Refregier, P., Boulet, V.: Statistical region snake-based segmentation adapted to different noise models. IEEE Trans. Pattern Analysis and Machine Intelligence 21(11), 1145–1157 (1999)
6. Cohen, L.D.: On active contour models and balloons. CVGIP: Image Understanding 53(2), 211–218 (1991)
7. Daugman, J.G.: Two-dimensional spectral analysis of cortical receptive field profiles. Vision Research 20(10), 847–856 (1980)
8. Daugman, J.G.: Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters. J. Optical Soc. Am. 2(7), 1160–1169 (1985)

9. Han, X., Xu, C., Prince, J.L.: A topology preserving level set method for geometric deformable models. IEEE Trans. Pattern Analysis and Machine Intelligence 25(6), 755–768 (2003)
10. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. Int'l J. Computer Vision 1(4), 321–331 (1988)
11. Lee, T.S.: Image representation using 2D Gabor wavelets. IEEE Trans. Pattern Analysis and Machine Intelligence 18(10), 959–971 (2003)
12. Li, C., Xu, C., Gui, C., Fox, M.D.: Level set evolution without reinitialization: a new variational formulation. In: IEEE Conf. Computer Vision Pattern Recognition, vol. 1, pp. 430–436 (2005)
13. Li, X., et al.: Discriminant locally linear embedding with high-order tensor data. IEEE Trans. on Systems, Man, and Cybernetics, Part B 38(2), 342–352 (2008)
14. Lie, J., Lysaker, M., Tai, X.: A Binary Level Set Model and Some Applications to Mumford-Shah Image Segmentation. IEEE Trans. Image Process. 15(5), 1171–1181 (2006)
15. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Comm. Pure Applied Mathematics 42(5), 577–685 (1989)
16. Mumford, D., Shah, J.: Boundary detection by minimizing functionals. In: Proc. IEEE Conf. Computer Vision Pattern Recognition, pp. 22–26 (1985)
17. Malladi, R., Sethian, J.A., Vemuri, B.C.: Shape modeling with front propagation: a level set approach. IEEE Trans. Pattern Analysis and Machine Intelligence 17(2), 158–175 (1995)
18. Marcelja, S.: Mathematical description of the responses of simple cortical cells. J. Optical Soc. Am. 70(11), 1297–1300 (1980)
19. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulation. J. Computational Physics 79(1), 12–49 (1988)
20. Osher, S., Fedkiw, R.: Level Set Methods and Dynamics Implicit Surfaces. Springer, Heidelberg (2003)
21. Park, S.W., Savvides, M.: A computationally efficient tensor framework without requiring mode factorization. IEEE Trans. on Systems, Man, and Cybernetics, Part B 37(5), 1156–1166 (2007)
22. Peng, D., Merriman, B., Osher, S., Zhao, H.K., Kang, M.: A PDE based fast local level set method. J. Computational Physics 155(2), 410–438 (1999)
23. Sethian, J.A.: Level Set Methods and Fast Marching Methods. Cambridge University, Cambridge (1999)
24. Smereka, P.: Semi-implicit level set methods for curvature and surface diffusion motion. J. Scientific Computing 19(1-3), 439–456 (2003)
25. Sun, J., Tao, D., Papadimitriou, S., Yu, P.S., Faloutsos, C.: Incremental Tensor Analysis:Theory and Applications: Theory and Applications. ACM Trans. Knowl. Disc. from Data 2(3), 11:1-11:37 (2008)
26. Suri, J.S., Liu, K., Singh, S., Laxminarayan, S.N., Zeng, X., Reden, L.: Shape recovery algorithms using level sets in 2D/3D medical imagery: a state-of-the-art review. IEEE Trans. Information Technology in Biomedicine 6(1), 8–28 (2002)
27. Sapiro, G.: Geometric Partial Differential Equations and Image Analysis. Cambridge University Press, Cambridge (2001)

28. Tao, D., Li, X., Wu, X., Maybank, S.J.: General tensor discriminant analysis and gabor features for gait recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 29(10), 1700–1715 (2007)
29. Tao, D., Song, M., Li, X., Shen, J., Sun, J., Wu, X., Faloutsos, C., Maybank, S.J.: Bayesian Tensor Approach for 3-D Face Modelling. IEEE Trans. Circuits and Systems for Video Technology 18(10), 1397–1410 (2008)
30. Tao, D., Li, X., Wu, X., Maybank, J.S.: Tensor Rank One Discriminant Analysis - A Convergent Method for Discriminative Multilinear Subspace Selection. Neurocomputing 71(12), 1866–1882 (2008)
31. Tao, W., Jin, H., Zhang, Y.: Color image segmentation based on mean shift and normalized cuts. IEEE Trans. on Systems, Man, and Cybernetics, Part B 37(5), 1382–1389 (2007)
32. Tschumperle, D., Deriche, R.: Diffusion PDEs on vector-valued images. IEEE Signal Processing Magazine 19(5), 16–25 (2002)
33. Vese, L.A., Chan, T.F.: A Multiphase Level Set Framework for Image Segmentation Using the Mumford and Shah Model. Int. J. Comput. Vision 50(3), 271–293 (2002)
34. Wang, Z., Vemuri, B.C.: Tensor field segmentation using region based active contour model. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 304–315. Springer, Heidelberg (2004)
35. Wang, Z., Vemuri, B.C.: DTI segmentation using an information theoretic tensor dissimilarity measure. IEEE Trans. Medical Imaging 24(10), 1267–1277 (2005)
36. Xu, C., Prince, J.L.: Snake, shapes, and Gradient vector flow. IEEE Trans. Image Processing 7(3), 359–369 (1998)
37. Zhao, H.K., Chan, T.F., Merrian, B., Osher, S.: A Variational Level Set Approach to Multiphase Motion. J. Comput. phys. 127, 179–195 (1996)
38. Zhang, Y.: Advances in Image and Video Segmentation. In: IRM (2006)
39. Zenzo, S.D.: A note on the gradient of a multi-image. J. Computer Vision, Graphic, and Image Processing 33(1), 116–125 (1986)

Recognition of Sketches in Photos

Bing Xiao[1], Xinbo Gao[1], Dacheng Tao[2], and Xuelong Li[3]

[1] VIPS Lab, School of Electronic Engineering, Xidian University,
Xi'an 710071, P.R. China
{bxiao,xbgao}@mail.xidian.edu.cn

[2] School of Computer Engineering, Nanyang Technological University,
50 Nanyang Avenue, Singapore 639798, Singapore
dctao@ntu.edu.sg

[3] State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of
Optics and Precision Mechanics, Chinese Academy of Sciences,
Xi'an 710119, P.R. China
xuelong_li@opt.ac.cn

Summary. Face recognition by sketches in photos makes an important complement to face photo recognition. It is challenging because sketches and photos have geometrical deformations and texture difference. Aiming to achieve better performance in mixture pattern recognition, we reduce difference between sketches and photos by synthesizing sketches from photos, and vice versa, and then transform the sketch-photo recognition to photo-photo/sketch-sketch recognition. Pseudo-sketch/pseudo-photo patches are synthesized with embedded hidden Markov model and integrated to derive pseudo-sketch/pseudo-photo. Experiments are carried out to demonstrate that the proposed methods are effective to produce pseudo-sketch/pseudo-photo with high quality and achieve promising recognition results.

Keywords: sketch-photo recognition, pseudo-sketch, pseudo-photo, image quilting, averaging overlapping areas.

1 Introduction

In the research of pattern recognition, face recognition has attracted great deal of attention. Automatic face recognition not only saves many person-hours but also lessens the subjective assessment [25], [41] and has played an important role in many application areas. Usually, we can only acquire witnesses' verbal description of the person in question instead of his photos, and consequently, simulated sketches have to been produced by artists or by combining interchangeable templates of local facial features with the help of computer [4], according to some described features. Face recognition is transformed into sketch-photo recognition which determines the person's identity from photo database based on simulated sketches automatically, but different mechanism for generating and expressing sketches and photos leads

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 239–262.
springerlink.com

to that most of the existing significant achievements for face recognition are not active for sketch-photo recognition. To be more specific, in producing sketches, artists make use of shadow texture to convey light and shade information, and sketch the contours subjectively, while photos are obtained with optical imaging equipments or other sensors objectively. Sketches and photos for a person have great geometrical deformations and large difference of texture. So, sketch-photo recognition becomes a challenging research focus of face recognition and deserves further research.

There has been considerable research on the problem of identifying a person by searching for the existing image most similar to his photo, but in contrast less research on the face recognition without photos of the person to be recognized has been done since the first automatic sketch-photo recognition algorithm was proposed in 1994 [27]. In the existing sketch-photo recognition algorithms, the research focus is transforming photos and sketches into the same modality to reduce their difference, so as to perform face recognition by sketches in pseudo-sketches or by pseudo-photos in photos with classical face recognition approaches [1], [35]. Face sketches referred to in sketch-photo recognition may be line-drawing sketch or complex sketch. Line-drawing sketches present the face information by nothing more than lines, and photos in database have to be converted into sketches, with which the recognition is performed. Complex sketches have more information depicted by lighting and shading besides lines, and photos in database and the sketch in question can be converted to each other; accordingly our aim is the recognition of complex sketch in photo database. Photos are converted into pseudo-sketches by Tang et al. firstly. They proposed a method based on principal components analysis (PCA) [31], [32], [33], a linear model, and then introduced manifold in the pseudo-sketch synthesis [18] in which nonlinearity between photos and sketches is approximated by local linearity. Gao et al. synthesize sketches [9], [42] based on embedded hidden Markov model (EHMM) so that complex nonlinear relationship between photos and sketches is learnt exactly. Furthermore, we proposed the algorithm [8], in which local strategy is introduced because local features are more specific and in favor of state estimation of EHMM. Facial pseudo-photo was obtained for the first time based on basic pixels [26], [27], low-level feature, and later, hybrid subspace method [17] and statistical methods [19] were proposed. The first two kinds of methods are based on assumption that there is linear mapping from sketch to photo, while statistical methods are nonlinear mappings. The drawback of the method in [19] is that it requires a great many of training samples which cannot be provided because of the high cost of sketch acquisition, so we proposed the method in [39] , a sketch-photo pair is sufficient for synthesizing a pseudo-photo in a way. In sketch-photo synthesis and recognition algorithms [8], [39], pseudo-sketch patches and pseudo-photo patches are combined by averaging the overlapping areas of adjacent patches, which may introduce blurring effect in synthesized sketches and photos. In the stage of recognition in sketches or photos, only eigenfaces based method is adopted.

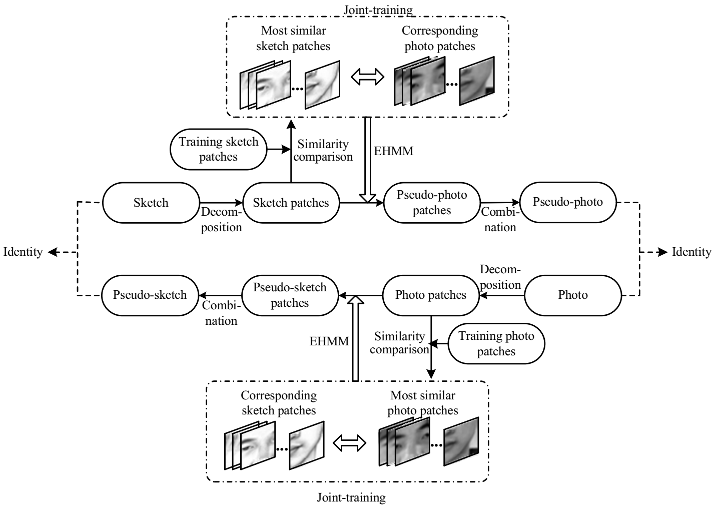


Fig. 1. The framework of the proposed methods

Aiming at these problems, we propose a novel face sketch-photo recognition algorithm [40] in which pseudo-sketch/pseudo-photo patches are combined into a pseudo-sketch/pseudo-photo with image quilting [7].

In our methods [8], [39], [40] whose framework is shown in Fig. 1, photos and sketches are divided into overlapping patches and sketch-photo patch pairs in training set are selected according to the similarity of testing and training photo/sketch patches. The nonlinear relationship of each selected sketch-photo patch pair is learnt by a pair of EHMMs. After pseudo-sketch patches/pseudo-photo patches are derived based on the learnt EHMM pairs, they are combined into a pseudo-sketch/pseudo-photo whose idea will be presented in detail in Section 3. After the photos and sketches are transformed into the same modality, many subspace learning based methods are applied into sketch-sketch recognition and photo-photo recognition. It can be proved that the proposed methods lead to pseudo-sketches/pseudo-photos with high quality and high recognition rate of sketch-photo recognition.

The remainder of this article is organized as follows. In section 2 and section 3, the synthesis and assembling of pseudo-sketch patches/pseudo-photo patches are explained in detail. The recognition of sketches/photos is performed with subspace learning based methods in section 4. The experimental results are presented in section 5, and the final section gives the conclusions.

2 Pseudo-sketch Patch/Pseudo-photo Patch Synthesis

EHMM [22] extracts two-dimensional facial features with a moderate computational complexity and has been used for the face-to-face transform [20]. It is employed to model the nonlinear relationship of the sketch patch and photo patch located at the same position in a sketch-photo pair. Given a training set with M photo-sketch pairs (P_i, S_i), where $i = 1, 2, \cdots, M$, and a photo P to be transformed, they are evenly divided into N overlapping patches whose size is $B \times B$ and overlapping degree is D. If a patch of the photo P is p and photo patches in training set are $\{p_{trj}\}$, where $j = 1, 2, \cdots, M \times N$, K training photo patches, most similar with the patch p, are hunted. Corresponding to the K photo patches, K sketch patches are selected to form pairs of photo patches and sketch patches. K EHMM pairs for them are constructed and K intermediate pseudo-sketch patches are derived based on these models. These pseudo-sketch patches are fused to result in the expected pseudo-sketch patch. The specific steps are given as below. Pseudo-photo patch synthesis can be performed similarly by exchanging the roles of photos and sketches.

Step 1: Searching for training photo patch and sketch patch pairs

The similarity degree of patches p and p_{trj} is measured with the doubly embedded Viterbi algorithm. K training photo patches with the greatest similarity degree w_l are hunted and denoted with p_{chol}, where $l = 1, 2, \cdots, K$, according to which K training sketch patches $\{s_{chol}\}$ are selected.

Step 2: Constructing EHMMs $\lambda_p = (\mathit{\Pi}_p, A_p, \mathit{\Lambda}_p, N_p)$ and $\lambda_s = (\mathit{\Pi}_s, A_s, \mathit{\Lambda}_s, N_s)$ for each pair of training photo patch and sketch patch $\{p_{chol}, s_{chol}\}$, $l = 1, 2, \cdots, K$ according to the Fig. 2.

• Observation vector sets O_p and O_s are extracted. O_p is the observation sequence for the training photo patch p_{chol} and denoted as $O_p = \{O_{yp}, 1 \leq y \leq Y\}$, where $O_{yp} = \{o_{yxp}, 1 \leq x \leq X\}$. O_{yp} is observation sequence of the y-th line,o_{yxp} is the observation vector of the x-th column in the y-th line, and X and Y are the numbers of observation vectors in horizontal and vertical directions. The observation sequence O_s of the sketch patch s_{chol} is similar to O_p . The observation vector at each pixel in p_{chol} and s_{chol} consists of

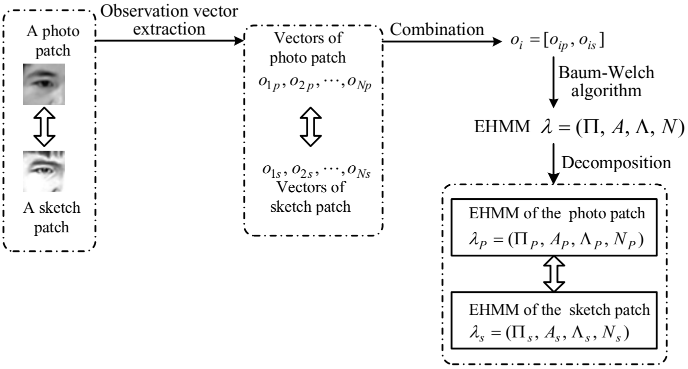

A photo patch
Observation vector extraction
Vectors of photo patch
$o_{1p}, o_{2p}, \cdots, o_{Np}$
$o_{1s}, o_{2s}, \cdots, o_{Ns}$
Vectors of sketch patch
A sketch patch
Combination
$o_i = [o_{ip}, o_{is}]$
Baum-Welch algorithm
EHMM $\lambda = (\Pi, A, \Lambda, N)$
Decomposition
EHMM of the photo patch
$\lambda_P = (\Pi_P, A_P, \Lambda_P, N_P)$
EHMM of the sketch patch
$\lambda_s = (\Pi_s, A_s, \Lambda_s, N_s)$

Fig. 2. The procedure of EHMM construction

the pixel gray value, Gaussian, Laplacian, horizontal derivative and vertical derivative operators, in which pixel gray value and Gaussian operator are related to the low-frequency information and average intensity respectively while other three operators are used for characterizing the high-frequency information;

- Observation vectors o_{yxp} and o_{yxs} for the x-th column and y-th line pixel in p_{chol} and s_{chol} are combined into a vector $o_{yx} = [o_{yxp}, o_{yxs}]$ so as to form the combined observation vectors $O = \{O_y, 1 \leq y \leq Y\}$, where $O_y = \{o_{yx}, 1 \leq x \leq X\}$;
- The number of super-states, embedded-states in each super-state and mixture components in each embedded-state are specified beforehand. Then the observation sequence O is segmented uniformly according to the number of states, and observation vectors within an embedded-state are clustered as many clusters as the number of mixture components in the embedded-state. The model $\lambda = (\mathit{\Pi}, A, \mathit{\Lambda}, N)$ of the photo patch and sketch patch p_{chol} and s_{chol} is joint-trained according to the segmented observation sequence O with the help of Baum-Welch algorithm [2], [23]. The algorithm is equivalent to the idea of expectation-maximization (EM) algorithm [3], [5]. With all parameters in λ initialized according to the sequence O, these parameters are modified iteratively based on the idea of EM algorithm until $P(O|\lambda)$, which is the similarity evaluation of observation vectors and the model λ , is convergent. EM algorithm is a two-step iteration:

In the E-step, $P(O|\lambda)$ is evaluated with forward algorithm [21]. Corresponding to the observation sequence, the state sequence S is $S = \{S_y, 1 \leq y \leq Y\}$, where $S_y = \{s_{yx}, 1 \leq x \leq X\}$. S_y is state sequence of the y-th line, s_{yx} is the state index of the x-th column in the y-th line. The forward and backward variables for the observation sequence O_y are defined as

$$\begin{aligned} \alpha^i_{yx}(k) &= P(o_{y1}, \cdots, o_{yx}, s_{yx} = \mathit{\Lambda}^{(i)}_k | S_y = \mathit{\Lambda}^{(i)}, \lambda) \\ \beta^i_{yx}(k) &= P(o_{y,x+1}, \cdots, o_{yX} | s_{yx} = \mathit{\Lambda}^{(i)}_k, S_y = \mathit{\Lambda}^{(i)}, \lambda) \end{aligned} \qquad (1)$$

which are computed by one-dimensional HMM forward-backward algorithms according to formulas (2) and (3):

$$\alpha^{(i)}_{y1}(k) = \pi^{(i)}_k b^i_k(o_1) \ \mathit{and} \ \alpha^i_{y(x+1)}(k) = [\sum_{l=1}^{N_e^{(i)}} \alpha^i_{yx}(l) a^{(i)}_{lk}] b^{(i)}_k(o_{y(x+1)}) \qquad (2)$$

$$\beta^i_{yX}(k) = 1 \ \mathit{and} \ \beta^i_{yx}(k) = \sum_{l=1}^{N_s^{(i)}} a^{(i)}_{kl} b^{(i)}_l(o_{y(x+1)}) \beta^i_{y(x+1)}(l) \qquad (3)$$

Based on these two variables, we can compute P^i_y and

$$P^i_y = P(O_y | S_y = \mathit{\Lambda}^i, \lambda) = \textstyle\sum_{k=1}^{N_e^{(i)}} \alpha^i_{yx}(k) \beta^i_{yx}(k) \ .$$

The forward variable for the observation sequence $O_1, O_2, \cdots, O_y$ is defined as

$$\eta_y(i) = P(O_1, \cdots, O_y, S_y = \Lambda^{(i)}|\lambda).$$

With these definitions in hand, the forward algorithm is carried through with the initialization of

$$\eta_1(i) = \Pi_i P_1^i$$

and

$$\eta_{y+1}(i) = [\textstyle\sum_{j=1}^{N} \eta_y(j) a_{ji}] P_y^i$$

is computed recursively until

$$\eta_Y(i) \text{ and } P(O|\lambda) = \textstyle\sum_{i=1}^{N} \eta_Y(i)$$

are obtained. Although the the form of $P(O|\lambda)$ is too intricate to be given straightforwardly, it is derived iteratively based on the known quantities such as GMMs, state transition, and so on.

State sequence and mixture index sequence corresponding to O is reestimated with the doubly embedded Viterbi algorithm [15], [37]. It starts with

$$\delta_1(i) = \textstyle\prod_i Q_1^i,$$

and

$$\delta_{y+1}(i) = max_{j \in [1, N^s]} [\delta_y(j) a_{ji}] Q_y^i$$

is computed recursively until $max_{i \in [1, N^s]} \delta_Y(i)$ is acquired, where

$$Q_y^i = max_{s_{y1}, s_{y2}, \cdots, s_{yX}} P(O_y, s_{y1}, s_{y2}, \cdots, s_{yX} | S_y = \Lambda^i, \lambda)$$

and it is processed with the help of one-dimensional HMM Viterbi algorithm: $\vartheta_{yx}^i(k)$ is initialized as

$$\vartheta_{y1}^i(k) = \pi_k^{(i)} b_k^{(i)}(O_1)$$

and induced according to

$$\vartheta_{y(x+1)}^i(k) = max_{1 \le l \le N_e^{(i)}} [\vartheta_{yx}^i(l) a_{lk}^{(i)}] b_k^{(i)}(o_{y(x+1)})$$

until

$$\vartheta_{yX}^i(k) \text{ and } Q_y^i = max_{1 \le k \le N_e^{(i)}} [\vartheta_{yX}^i(k)]$$

are got. When $max_{i \in [1, N^s]} \delta_Y(i)$ is computed, there is an array to keep track of the arguments that maximize $\delta_y(i)$ and $\vartheta_{yx}^i(k)$ in each iteration so that the best state sequence and mixture index sequence S_b and M_b are tracked back finally.

In the M-step, observation sequence O is segmented with the reestimated state and mixture index sequences S_b and M_b, and then the joint-trained

EHMM is updated according to the segmented observation sequence. The re-estimation of EHMM parameters are performed with formulas (4)-(8):

$$\hat{\Pi}_i = \frac{P(S_1 = \Lambda^{(i)}|\lambda, O)}{\sum_{i=1}^{N_s} P(S_1 = \Lambda^{(i)}|\lambda, O)}, \tag{4}$$

$$\hat{a}_{ij} = \frac{\sum_{y=1}^{Y} P(S_{y-1} = \Lambda^{(i)}, S_y = \Lambda^{(j)}|\lambda, O)}{\sum_{y=1}^{Y} P(S_{y-1} = \Lambda^{(i)}|\lambda, O)}, \tag{5}$$

$$\hat{\pi}_j^{(i)} = \frac{\sum_{y=1}^{Y} P(s_{y1} = \Lambda_j^{(i)}, S_y = \Lambda^{(i)}|\lambda, O)}{\sum_{y=1}^{Y} P(S_y = \Lambda^{(i)}|\lambda, O)}, \tag{6}$$

$$\hat{a}_{jl}^{(i)} = \frac{\sum_{y=1}^{Y}\sum_{x=1}^{X} P(s_{y(x-1)} = \Lambda_j^{(i)}, s_{yx} = \Lambda_l^{(i)}, S_y = \Lambda^{(i)}|\lambda, O)}{\sum_{y=1}^{Y}\sum_{x=1}^{X} P(s_{y(x-1)} = \Lambda_j^{(i)}, S_y = \Lambda^{(i)}|\lambda, O)}, \tag{7}$$

$$\hat{b}_j^{(i)}(k) = \frac{\sum_{y=1}^{Y}\sum_{x=1, s.t. o_{yx}=v_k}^{X} P(s_{yx} = \Lambda_j^{(i)}, S_y = \Lambda^{(i)}|\lambda, O)}{\sum_{y=1}^{Y}\sum_{x=1}^{X} P(s_{yx} = \Lambda_j^{(i)}, S_y = \Lambda^{(i)}|\lambda, O)}, \tag{8}$$

where $V = \{v_1, v_2, \cdots, v_K\}$ reserves distinct observation vectors and K is the length of it. If $P(O|\lambda)$ is convergent, EM algorithm is completed, otherwise it comes back to E-step.

• The model λ derived in the previous step is decomposed into $\lambda_p = (\Pi_p, A_p, \Lambda_p, N_p)$ for the photo patch p_{chol} and $\lambda_s = (\Pi_s, A_s, \Lambda_s, N_s)$ for the sketch patch s_{chol}. As already stated, λ_p and λ_s only have different GMMs in each pair of corresponding embedded-states and the decomposition of λ includes dividing the mean vector and covariance matrix of every mixture component in each embedded-state. We suppose that $\sum_{iml}^{(k)}$ is a diagonal matrix and $l = 0, 1$, where 0 and 1 represent photo patch and sketch patch respectively. Consequently, for the m-th mixture component in i-th embedded-state and k-th super-state, the division is performed as below:

$$\mu_{im}^{(k)} = [\mu_{im0}^{(k)}, \mu_{im1}^{(k)}] \text{ and } \sum\nolimits_{im}^{(k)} = \begin{bmatrix} \sum_{im0}^{(k)} & 0 \\ 0 & \sum_{im1}^{(k)} \end{bmatrix}.$$

Step 3: Synthesizing a pseudo-sketch patch s for the photo patch p of P

K intermediate pseudo-sketch patches s_{pseul} is derived by decoding based on each λ_p and reconstructing based on corresponding λ_s. The weighted average of these intermediate pseudo-sketch patches is the expected pseudo-sketch patch s, that is,

$$s = \sum_{j=1}^{K} w_j \times s_{pseuj} \tag{9}$$

3 Combination of Pseudo-sketch Patches/ Pseudo-photo Patches

With the pseudo-sketch/pseudo-photo patches, having overlapping regions, in hand, we assemble them into a pseudo-sketch/pseudo-photo by averaging the overlapping areas between patches. Given two neighboring patches horizontally, there is an overlapping area between them along their vertical edge, which is shown with shadow region in Fig.3. The overlapping area in the left patch is denoted as L_{ov} and that in the right patch is R_{ov} , which are of the same size $r \times c$. In [8], [39], as illustrated in Fig. 4, the pixel values of the overlapping area are computed by averaging the corresponding elements in L_{ov} and R_{ov} with Eq. (10).

$$V_{ov} = \frac{1}{2}(L_{ov} + R_{ov}) \tag{10}$$

Although this method is simple, it leads to blurring effect in the synthesized sketches and photos. We make use of the idea of image quilting [7], whose idea

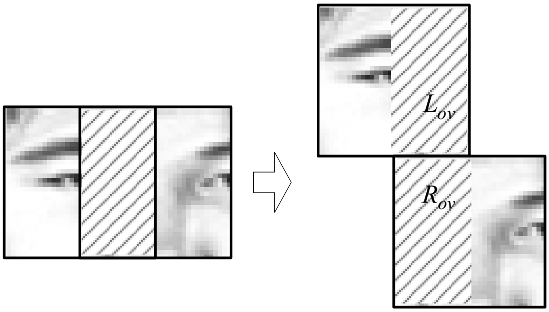


Fig. 3. Two overlapping patches

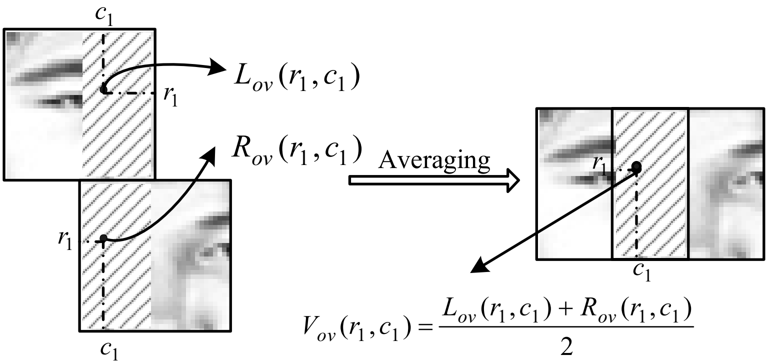


Fig. 4. Averaging two overlapping patches

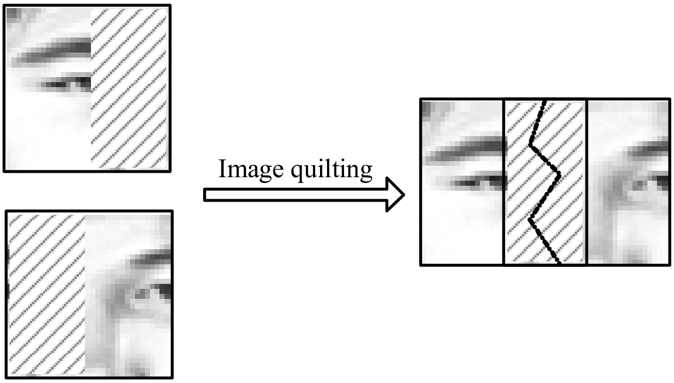


Fig. 5. Image quilting

is illustrated in Fig. 5. For the overlapping area shown in Fig. 3, the value of corresponding elements in L_{ov} and R_{ov} may be different and we have to determine whether the element value in L_{ov} or that in R_{ov} should be used for each pixel in overlapping area, which equals to finding an edge for combining these two patches smoothly. The value of pixels on the left side of the edge is given according to L_{ov} and that of pixels on the right side is according to R_{ov} . The solution is searching for an optimal edge

$$E^* = \{(1, y_1), (2, y_2), \cdots, (r, y_r)\},$$

where (i, y_i) means i-th row and y_i-th column in L_{ov} and R_{ov}, such that,

$$E^* = argmin_E \left(\sum_{(i,y_i)\in E} |L_{ov}(i, y_i) - R_{ov}(i, y_i)|^2 \right) \tag{11}$$

If

$$p_{i,y_i} = |L_{ov}(i, y_i) - R_{ov}(i, y_i)|^2, \tag{12}$$

the difference between $L_{ov}(i, y_i)$ and $R_{ov}(i, y_i)$, is regarded as the cost of traversing (i, y_i), the edge E^* is determined by searching for the minimum

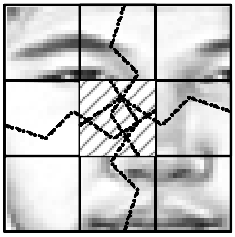
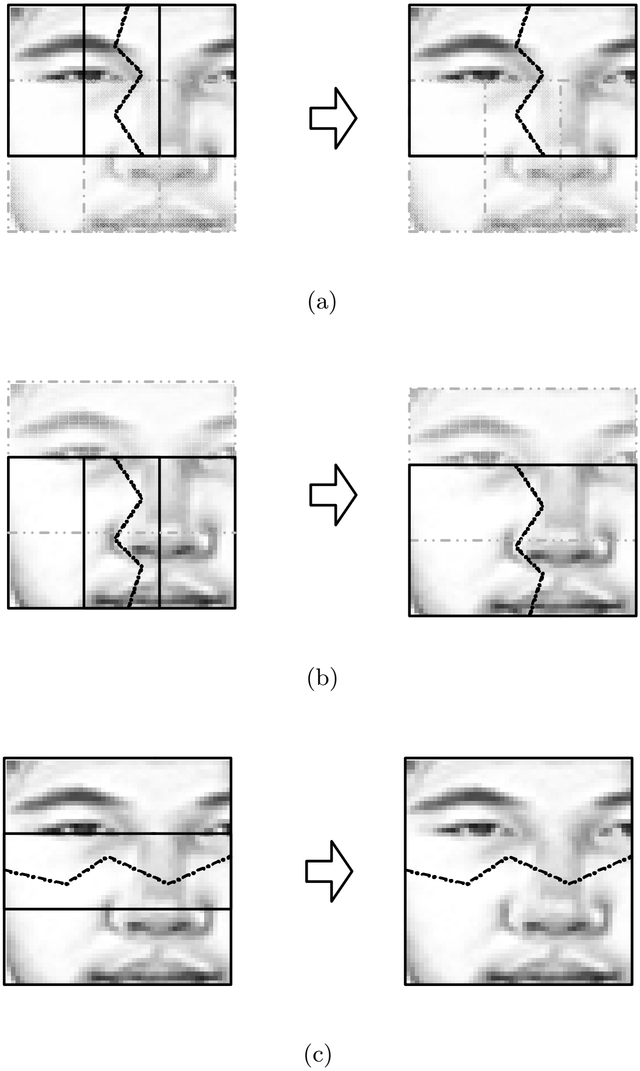


Fig. 6. Image quilting for four patches



(a)

(b)

(c)

Fig. 7. Scheme for stitching four patches

cost path [6] through the overlapping area from the top down, as Fig. 5. The elements belonging to the optimal edge E^* are located row by row in the overlapping area. When the element in the i-th row is determined, whose index of column is y_i, elements at $(i{+}1, y_i{-}1)$, $(i{+}1, y_i)$ and $(i{+}1, y_i{+}1)$ of L_{ov} and R_{ov} are compared and position of the element corresponding to the least difference is added into E^*. The steps are given as below. For two vertically neighboring patches, the minimum cost path through the overlapping area is founded horizontally.

Step 1: E^* is initialized as NULL and $p_{1,1}, p_{1,2}, \cdots, p_{1,c}$ are computed for the first row of L_{ov} and R_{ov} according to formula (12).

Step 2: The minimum among $p_{1,1}, p_{1,2}, \cdots, p_{1,c}$ is selected and preserved in p_{1,y_1}, and its row and column index $(1, y_1)$ is added into the set E^*.

Step 3: If the pixel traversed by E^* in the *(i-1)*-th row is not in the leftmost or the rightmost column of the overlapping area, $p_{i,y_{i-1}-1}, p_{i,y_{i-1}}$ and $p_{i,y_{i-1}+1}$ are calculated and the minimum p_{i,y_i} is selected. If it is in the leftmost or the rightmost column, $p_{i,y_{i-1}}$ and $p_{i,y_{i-1}+1}$, or $p_{i,y_{i-1}-1}$ and $p_{i,y_{i-1}}$, are compared, minimum among which is preserved in p_{i,y_i}. (i, y_i) is added into the set E^*.

Step 4: *Step* 3 is repeated from the second to the last row until the optimal edge E^* is acquired.

For four overlapping patches shown in Fig. 6, there are overlapping areas for each pair of neighboring patches and an overlapping area for four patches, shown as shadow region. We find the minimum cost path for each pair of neighboring patches and we will obtain four paths, but the path in the overlapping area for four patches is confusing. Consequently, we combine them according to the scheme shown in Fig. 7. In Fig. 7(a) and Fig. 7(b), the patches in the first line and in the second line are respectively combined into two large patches with image quilting algorithm shown in Fig. 5. These two large patches are combined in Fig. 7(c) and image quilting of four patches is completed.

4 Recognition of Sketches/Photos with Subspace Learning Based Methods

After the training photos are transformed into pseudo-sketches or the sketch to be identified is transformed into a pseudo-photo, sketch-photo recognition is performed by recognizing sketch in pseudo-sketches or pseudo-photo in photos. Training and testing samples of recognition are in the same modality and several subspace learning based face recognition algorithms can be applied straightforwardly. For an image set, subspace learning methods find a basis space constituted by a set of basis images which account for latent variation making images in the set distinct. By projecting an input face image into the basis space, it is encoded in a reduced dimensional feature space.

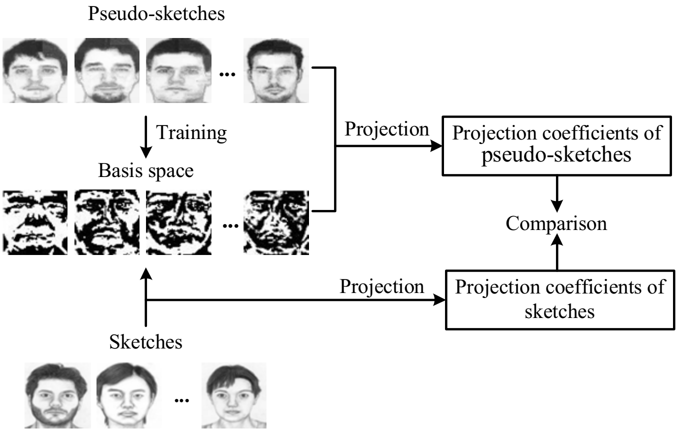

Pseudo-sketches
Training
Basis space
Projection
Projection coefficients of pseudo-sketches
Comparison
Projection
Projection coefficients of sketches
Sketches

Fig. 8. Sketch-sketch recognition based on subspace learning methods

For a sketch S and a pseudo-sketch set $\{S_i\}$,$i = 1, 2, \cdots, M$, the recognition of S in $\{S_i\}$ based on subspace learning methods is illustrated in Fig. 8 and this method is available for photos recognition by substituting pseudo-photo for the sketch and photo set for the pseudo-sketch set.

Step 1: $M - 1$ basis images $W = \{w_k\}$ are derived according to the pseudo-sketches $\{S_i\}$ to constitute basis space, $k = 1, 2, \cdots, M - 1$.

Step 2: All pseudo-sketches are project into the basis space W to derive the vectors of projection coefficients c_i according to $S_i = c_i \times W$.

Step 3: The sketch S is also project into the space W and the projection coefficients vector c_s is computed according to $S = c_s \times W$.

Step 4: Weight vector c_s is compared with all c_i, that is $d_i = \|c_s - c_i\|$, and the pseudo-sketch corresponding to the minimum values in $\{d_i\}$ is adopted to identify the sketch S.

The key of the above procedure is seeking for the basis images of pseudo-sketch set $\{S_i\}$, $i = 1, 2, \cdots, M$. In this paper, we employ several subspace learning methods including PCA [24], independent component analysis (ICA) [1], [13], kernel principal component analysis (KPCA) [14], [28], locality preserving projection (LPP) [11] and offline tensor analysis (OTA) [30]. In the future, we will consider some discriminative methods [43] [46]. In these algorithms, pseudo-sketches in the set $\{S_i\}$ are arranged as a matrix X, in which each column corresponds to a pseudo-sketch.

The basis images extracted out of the image set by PCA is eigenfaces which are orthogonal. The recognition method based on eigenfaces [36] is regarded as the first successful facial recognition method and surpasses other face recognition methods at its speed and efficiency. It works especially well when the faces are captured in frontal view and under similar lighting, which are satisfied by the experimental data in this paper. Average pseudo-sketch of $\{S_i\}$ is computed as

$$MS = \frac{1}{M}\sum_{i=1}^{M} X(:,i), \tag{13}$$

and the covariance matrix is

$$C = \frac{1}{M}\sum_{i=1}^{M}(X(:,i) - MS)(X(:,i) - MS)^T. \tag{14}$$

Eigenvalues and eigenvectors of C can be computed and eigenvectors corresponding to maximum eigenvalues are eigenfaces.

ICA is generalized on the basis of PCA to explore the high-order relationships among pixels. The basis images extracted by ICA are independent components which are assumed nongaussian and mutually independent. The correlation between pseudo-sketches is given as the rows of a mixing matrix A. Each basis image is a row of source matrix S which is extracted according to the formula

$$X^{'} = A \cdot S. \tag{15}$$

KPCA is integration of PCA with kernel methods and is a nonlinear form of PCA. PCA derives principal components of input images, while KPCA focuses on those of variables which are nonlinearly related to the input images. With a kernel, the input images are projected into a high dimensional feature space so that high order correlations between input images are explored. Dot product is computed by a kernel function k for each pair of the pseudo-sketch vectors to form the dot product matrix K:

$$K_{ij} = \left(k(X(:,i), X(:,j))\right)_{ij}. \tag{16}$$

Eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_M$ and eigenvectors $v_1, v_2, \cdots, v_M$ of K can be computed, and then eigenvectors are normalized such that

$$\lambda_m(v_m \cdot v_m) = 1. \tag{17}$$

The normalized eigenvectors are the basis images derived by KPCA.

Different from PCA preserving global structure of the image space, LPP specializes in holding local relationship which is in favor of classification. The basis images extracted LPP is called Laplacianface [12] and it has more discrimination than Eigenfaces. A similarity matrix M is defined for measuring the similarity of pairwise pseudo-sketches, based on which a diagonal matrix D is computed according to

$$D_{ii} = \sum_j M_{ji}. \tag{18}$$

Laplacian matrix is $L = D - M$. The generalized eigenvectors of $XLX^Tw = \lambda XDX^Tw$ are basis images.

Furthermore, LPP can be carried out in two directions of images. Each pseudo-sketch S_i in the pseudo-sketch set, $i = 1, 2, \cdots, M$, is of size $n_1 \times n_2$ and it is reshaped into a vector of $n_1 \times n_2$ elements in the above methods. Tensor subspace analysis (TSA) [10] attaches importance to the relationship between the rows of the pseudo-sketch matrix and that between the columns. The pseudo-sketch S_i can be represented as the second order tensor in the tensor space $R^{n_1} \otimes R^{n_2}$. Basis images are extracted in row- and column-directions respectively in the virtue of LPP.

Inspired by the idea of [34] [44] [45], each pixel of pseudo-sketch S_i is decomposed into 5-scales and 8-directions by Gabor filters and the pseudo-sketch S_i is represented as a fourth-order tensor $\chi_i \in R^{n_1 \times n_2 \times 5 \times 8}$, $i = 1, 2, \cdots, M$, that is to say, the second order tensor is extended into a fourth-order one. OTA [16] can accept a high order tensor and output its basis images along each dimension. Basis images matrix of each dimension is initialized as $U^{(1)} \in R^{n_1 \times n_1}$, $U^{(2)} \in R^{n_2 \times n_2}$, $U^{(3)} \in R^{5 \times 5}$ and $U^{(4)} \in R^{8 \times 8}$. The covariance matrix for the first dimension is defined as

$$C = \sum_{m=1}^{M} z_m z_m^T, \tag{19}$$

where $z_m = \chi_m \times_2 U^{(2)} \times_3 U^{(3)} \times_4 U^{(4)}$ and z_m is transformed into a matrix of size $n_1 \times (n_2 \times 5 \times 8)$. Basis images matrix for the first dimension $U^{(1)}$ is updated to eigenvectors of the matrix C. Basis images for other dimensions are computed similarly based on the updated $U^{(1)}$, $U^{(2)}$, $U^{(3)}$ and $U^{(4)}$.

5 Experimental Results and Analysis

In this section, we conduct experiments on the color face photo-sketch database and gray photo-sketch database to evaluate the proposed sketch-photo recognition algorithms from two aspects that are the quality of synthesized pseudo-sketches/pseudo-photos and recognition performance of sketches in photo set, with the direct sketch-photo recognition method as reference. These two databases are provided by the Multimedia Lab of the Chinese University of Hong Kong, and the instances of experimental data are shown in Fig. 9 and Fig.10, in which photos and the corresponding sketches are shown in the first and the second line respectively. In the experiments, all experimental data are resized into 64×64 and divided into patches of size 32×32 pixels. The overlapping degree of neighboring patches is 75%. As mentioned above, sketches and photos are transformed into the same modality by two proposed schemes that photos are transformed into pseudo-sketches and sketches into pseudo-photos, and then recognition is performed in sketches or photos. The transformation is based on EHMM which has 3 super-states from top to bottom, 6 embedded-states from left to right in each super-state and 12 mixture components in each embedded-state.

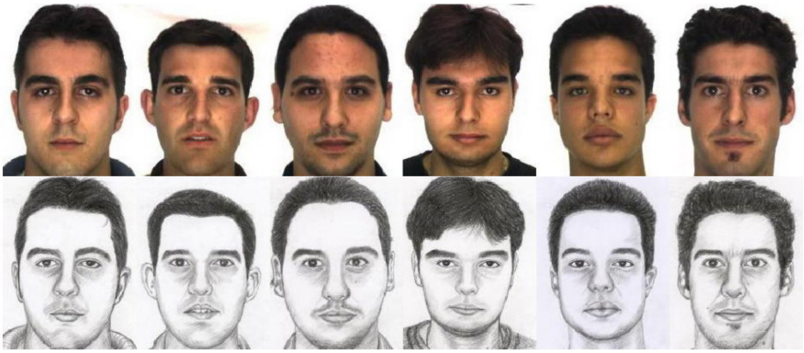


Fig. 9. Instances of color photo-sketch pairs provided by the Multimedia Lab of the Chinese University of Hong Kong

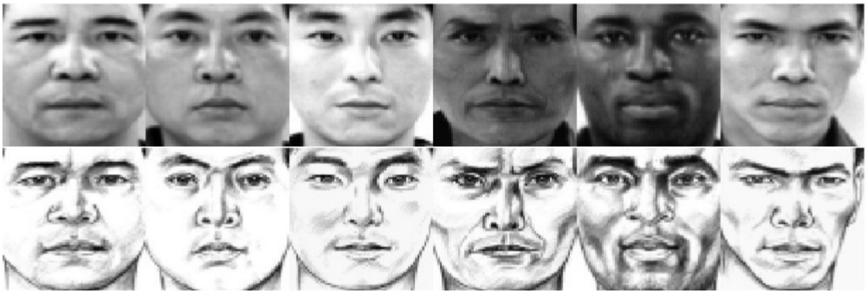


Fig. 10. Instances of gray photo-sketch pairs provided by the Multimedia Lab of the Chinese University of Hong Kong

5.1 Synthesis and Recognition of Pseudo-sketches

When we perform pseudo-sketch synthesis in the colorful photo-sketch database, leave-one-out strategy [29] is adopted. For the database, a photo is left as the testing sample in turn while other photo-sketch pairs are training samples until all photos in the database are transformed into pseudo-sketches. Fig. 11 shows examples of sketch synthesis results of the proposed methods [8], [40]. From the first column to the last one, examples of original photos, original sketches, pseudo-sketches obtained with the method in [8], pseudo-sketches obtained with the method in [40] are listed. The method in [40] produces pseudo-sketches not only with less blurring and blocking effect, but also more like original sketches.

In order to quantify the quality of synthesized pseudo-sketches, we make use of universal image quality index (UIQI) [38]. It evaluates the quality of testing image based on the reference one from three aspects. In the future,

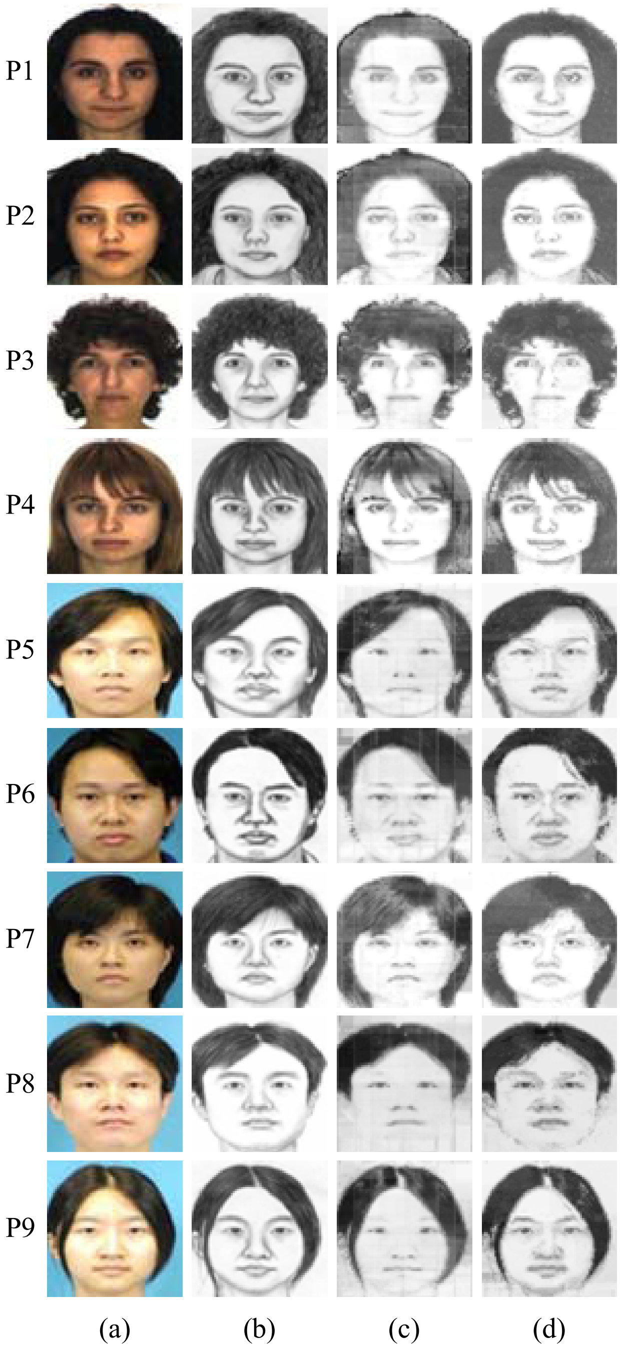


Fig. 11. Examples of pseudo-sketches obtained with two methods. (a) original photos (b) original sketches (c) pseudo-sketches obtained with the proposed method in [8], (d) pseudo-sketches obtained with the proposed method in [40].

we will consider more phosificated method for quality assessment [47]. First, their correlation degree is measured linearly according to Eq. (20),

$$C = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{20}$$

where σ_x and σ_y are the variance of reference image and that of testing one respectively, and σ_{xy} is their covariance. Secondly, luminance distortion of the reference image is reflected by the similarity of mean luminance $\bar{x}$ and $\bar{y}$ of reference and testing images, that is,

$$L = \frac{2\bar{x}\bar{y}}{(\bar{x})^2 + (\bar{y})^2}. \tag{21}$$

Thirdly, the similarity of contrast σ_x and σ_y is computed as

$$T = \frac{2\sigma_x \sigma_y}{(\sigma_x)^2 + (\sigma_y)^2}. \tag{22}$$

The UIQI is defined by integrating these three measurements, that is,

$$U = C \times L \times T \tag{23}$$

If σ_x and $\bar{x}$ are more close to σ_y and $\bar{y}$, C, L and T are getting nearer to one, otherwise, C is getting nearer to -1, and L and T are to zero. So, higher UIQI value indicates higher quality of testing images.

The pseudo-sketches and original photos are treated as testing images and original sketches are reference images. The UIQI value of pseudo-sketches and photos in Fig. 11, and average UIQI value of all pseudo-sketches are shown in Table 1. It can be found that UIQI value corresponding to pseudo-sketches is higher than that of photos and pseudo-sketches obtained with the method in [40] has higher UIQI value than those obtained with the method in [8]. The method in [40] leads to the pseudo-sketches having the highest quality.

Table 1. The UIQI of pseudo-sketches obtained with different methods and photos

	P1	P2	P3	P4	P5	P6	P7	P8	P9	Means
Fig. 11 (a)	0.6065	0.7792	0.7329	0.7613	0.8608	0.7043	0.7758	0.8627	0.8750	0.7797
Fig. 11 (c)	0.8346	0.8564	0.8858	0.8298	0.8780	0.8349	0.9176	0.8930	0.9069	0.8927
Fig. 11 (d)	0.8882	0.9044	0.8984	0.8758	0.9249	0.8639	0.9455	0.9287	0.9426	0.9064

After photos in training set are transformed into pseudo-sketches, the sketch to be identified is compared with pseudo-sketches by subspace learning based methods. In our experiment, all sketches in database are testing images and pseudo-sketches or photos are training samples. Six subspace learning based methods such as eigenfaces, ICA, KPCA, Laplacianfaces, TSA and

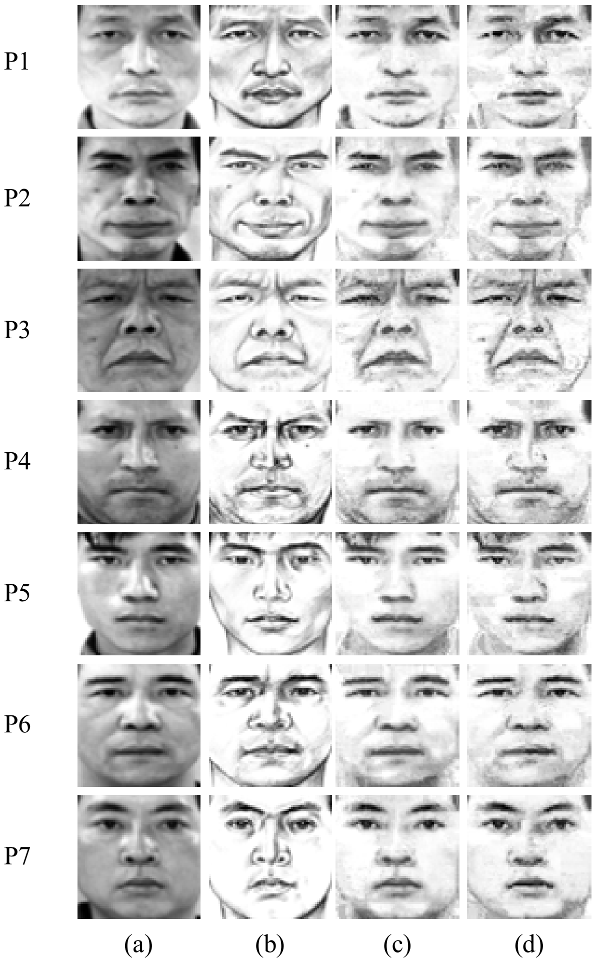


Fig. 12. Examples of pseudo-sketches obtained with two methods. (a) original photos, (b) original sketches, (c) pseudo-sketches obtained with the method in [8], (d) pseudo-sketches obtained with the method in [40].

Table 2. Pseudo-sketch recognition rate of three methods

Training set	Eigenfaces	ICA	KPCA	Laplacianfaces	TSA	OTA
Fig. 11 (a)	36.7%	78.89%	27.78%	15.56%	76.67%	67.78%
Fig. 11 (c)	94.4%	91.11%	94.44%	64.44%	78.89%	92.22%
Fig. 11 (d)	100%	93.33%	100%	81.11%	86.67%	97.78%

Table 3. The UIQI of pseudo-sketches obtained with different methods and photos

	P1	P2	P3	P4	P5	P6	P7	Means
Fig. 12 (a)	0.4661	0.4226	0.5189	0.5455	0.4995	0.6592	0.5551	0.5311
Fig. 12 (c)	0.5915	0.6068	0.6491	0.6464	0.6161	0.7376	0.7347	0.6887
Fig. 12 (d)	0.6340	0.6400	0.6551	0.6971	0.6485	0.7753	0.7866	0.6893

Table 4. Pseudo-sketch recognition rate of three methods

Training set	Eigenfaces	ICA	KPCA	Laplacianfaces	TSA	OTA
Fig. 12 (a)	19.05%	71.43%	14.29%	14.29%	57.14%	76.19%
Fig. 12 (c)	85.71%	85.71%	85.71%	66.67%	61.90%	95.24%
Fig. 12 (d)	90.48%	85.71%	90.48%	76.19%	47.62%	100%

OTA based approaches are used to match the testing and training images. The recognition rate of sketches is shown in Table 2 and the method in [40] obtains the highest recognition rate in any case.

We perform experiments in the gray sketch-photo database, the sketch synthesis results of the proposed methods [8], [40] are given in Fig. 12. The proposed method in [40] produces pseudo-sketches with less blocking effect and more like original sketches. Table 3 shows the UIQI of pseudo-sketches in Fig.12 and average UIQI value of all pseudo-sketches. Conclusion coincident with that of colorful sketch-photo database can be achieved. The result of sketch-photo recognition is shown in Table 4, and our method has the best recognition performance except that TSA is adopted in recognition.

5.2 Synthesis and Recognition of Pseudo-photos

Firstly, pseudo-photo synthesis and photo-photo recognition are conducted in the colorful photo-sketch database. The examples derived by the proposed methods in [39], [40] are illustrated in Fig. 13. Pseudo-photos synthesized by the proposed method in [40] are less blurred and reserve more texture information in favor of face recognition. The UIQI value of pseudo-photos is shown in Table 5 with original photos as reference images. All sketches in database or their corresponding pseudo-photos are testing images while photos are training samples for recognition, and recognition results are shown in Table 6. The method in [40] not only synthesizes pseudo-photos with the highest quality but also have the most favorable recognition results.

Similar experiments are conducted in the gray photo-sketch pairs. The photo synthesis results of the proposed methods in [39], [40] are compared in Fig. 14 and the recognition rate of sketches is shown in Table 7. For gray sketches and photos, the pseudo-photos resulted by our method in [40] are less blurred too and more like original photos. The recognition results demonstrate that our method achieves best results based on most of the subspace methods.

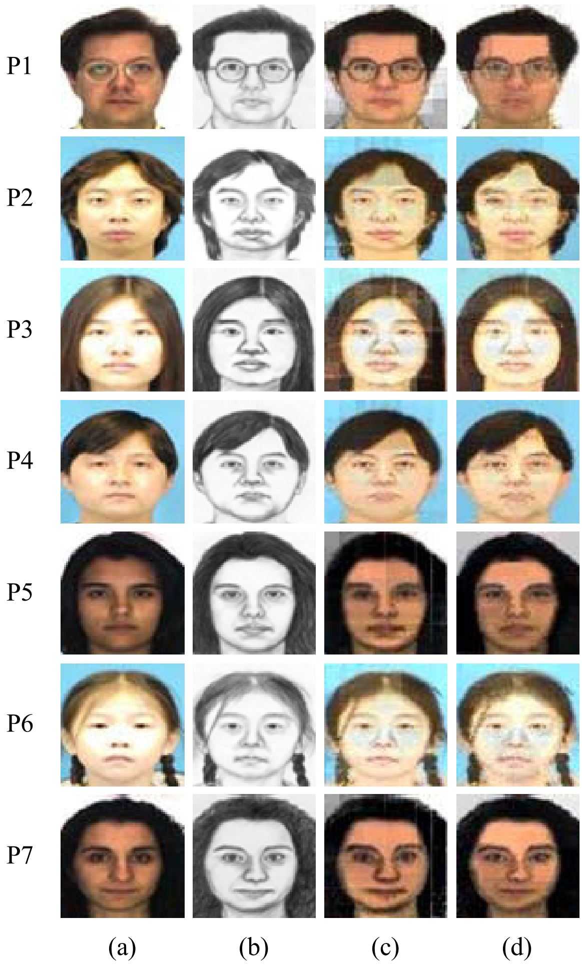


Fig. 13. Examples of pseudo-photos obtained with two methods. (a) original photos, (b) original sketches, (c) pseudo-photos obtained with the method in [39], (d) pseudo-photos obtained with the method in [40].

Table 5. The UIQI of pseudo-photos obtained with different methods and sketches

	P1	P2	P3	P4	P5	P6	P7	Means
Fig. 13 (b)	0.6895	0.8696	0.8726	0.8709	0.6655	0.8203	0.6065	0.7797
Fig. 13 (c)	0.8503	0.9105	0.9020	0.9522	0.8374	0.8600	0.7180	0.8865
Fig. 13 (d)	0.9232	0.9264	0.9124	0.9525	0.8979	0.8768	0.8302	0.9088

Table 6. Pseudo-photo recognition rate of three methods

Testing set	Eigenfaces	ICA	KPCA	Laplacianfaces	TSA	OTA
Fig. 13 (b)	36.7%	78.89%	27.78%	15.56%	76.67%	67.78%
Fig. 13 (c)	95.6%	78.89%	94.44%	70%	73.33%	98.89%
Fig. 13 (d)	100%	78.89%	100%	80%	80%	100%

Table 7. Pseudo-photo recognition rate of three methods

Testing set	Eigenfaces	ICA	KPCA	Laplacianfaces	TSA	OTA
Fig. 14 (b)	19.05%	61.9%	14.29%	14.29%	57.14%	76.19%
Fig. 14 (c)	71.43%	71.43%	80.95%	61.90%	33.33%	71.43%
Fig. 14 (d)	76.19%	61.9%	80.95%	71.43%	38.1%	80.95%

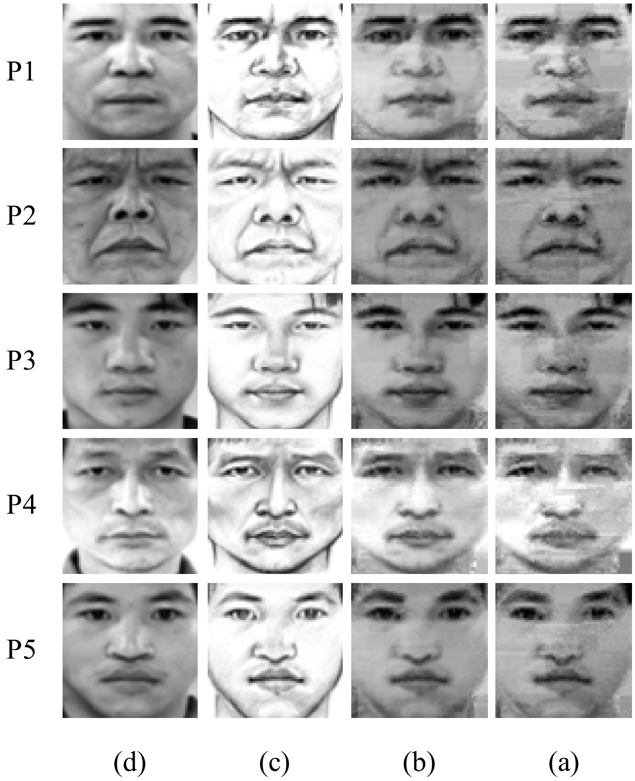


Fig. 14. Examples of pseudo-photos obtained with two methods. (a) original photos, (b) original sketches, (c) pseudo-photos obtained with the method in [39], (d) pseudo-photos obtained with the proposed method in [40]

6 Conclusion

With the aim of modeling the nonlinear relationship between sketch and photo with less training samples, we propose pseudo-sketch and pseudo-photo synthesis algorithms based on EHMM. Besides that, local features reflect more specific information than the whole face, so local strategy is adopted, that is to say, sketches and photos are divided into patches firstly, and then pseudo-sketch patches and pseudo-photo patches are integrated into the pseudo-sketch and pseudo-photo after they are derived with the help of EHMMs. After synthesizing pseudo-sketch and pseudo-photo, sketch-photo recognition is transformed into sketch-sketch or photo-photo recognition which is conducted with several subspace learning based methods. These methods achieve optimistic performance for synthesizing pseudo-sketch and pseudo-photo, and leads to higher sketch-photo recognition rate compared with the direct sketch-photo recognition.

Acknowledgments. The authors are grateful to the helpful comments and suggestions from the anonymous reviewers. Thanks must be expressed to the Multimedia Lab of the Chinese University of Hong Kong for providing us with the face photo-sketch images database. This research was supported partially by the National Natural Science Foundation of China under Grant 60832005, the Ph. D. Programs Foundation of Ministry of Education of China under Grant 20090203110002, the Key Science and Technology Program of Shaanxi Province of China under Grant. 2010K06-12, and the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2009JM8004.

References

1. Bartlett, M., Movellan, J., Sejnowski, T.: Face recognition by independent component analysis. IEEE Trans. Neural Netw. 13(6), 1450–1464 (2002)
2. Baum, L., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. 41(1), 164–171 (1970)
3. Bilmes, J.: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, Technical Report, ICSI TR-97-021, International computer science institute, University of California, Berkeley, USA (1998)
4. Davies, G.M., Willie, P.V.D., Morrison, L.J.: Facial composite production: a comparison of mechanical and computer driven systems. J. Appl. Psychol. 85(1), 119–124 (2000)
5. Dempster, A., Laird, N., Rubin, D.: Likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39(1), 1–38 (1977)
6. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numer. Math. 1(1), 269–271 (1959)
7. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proc. ACM Conference on Computer Graphics and Interactive Techniques, pp. 341–346 (2001)

8. Gao, X., Zhong, J., Tao, D., Li, X.: Local face sketch synthesis learning. Neurocomputing 71(10-12), 1921–1930 (2008)
9. Gao, X., Zhong, J., Tian, C.: Face sketch synthesis algorithm based on machine learning. IEEE Trans. Circuits Syst. Video Technol. 18(4), 487–496 (2008)
10. He, X., Cai, D., Niyogi, P.: Tensor subspace analysis. In: Proc. 19th Annual Conference on Neural Information Processing Systems (2005)
11. He, X., Niyogi, P.: Locality preserving projections. In: Proc. 17th Annual Conference on Neural Information Processing Systems, pp. 153–160 (2003)
12. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face recognition using laplacianfaces. IEEE Trans. Pattern Anal. Mach. Intell. 27(3), 1–13 (2005)
13. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks 13(4-5), 411–430 (2000)
14. Kim, K.I., Jung, K., Kim, H.J.: Face recognition using kernel principal component analysis. IEEE Signal Processing Letters 9(2), 40–42 (2002)
15. Kuo, S., Agazzi, O.: Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov models. IEEE Trans. Pattern Anal. Mach. Intell. 16(8), 842–848 (1994)
16. Lathauwer, L.D., DeMoor, B., Vandewalle, J.: On the best rank-1 and rank-(r1,r2. rn) approximation of higher-order tensors. SIAM J. Matrix Anal. A. 21(4), 1324–1342 (2000)
17. Li, Y., Savvides, M., Bhagavatula, V.: Illumination tolerant face recognition using a novel face from sketch synthesis approach and advanced correlation filters. In: Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing, pp. 57–360 (2006)
18. Liu, Q., Tang, X., Jin, H., Lu, H., Ma, S.: A nonlinear approach for face sketch synthesis and recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1005–1010 (2005)
19. Liu, W., Tang, X., Liu, J.: Bayesian tensor inference for sketch-based facial photo hallucination. In: Proc. Int'l Joint Conf. on Artificial Intelligence, pp. 2141–2146 (2007)
20. Nagai, T., Nguyen, T.: Appearance model based face-to-face transform. In: Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing, pp. 749–752 (2004)
21. Nefian, A.: A hidden markov model-based approach for face detection and recognition, Ph.D., Georgia Institute of Technology (1999)
22. Nefian, A., Hayes, M.H.: Face recognition using an embedded HMM. In: Proc. Int'l Conf. on Audio- and Video-based Biometric Person Authentication, pp. 19–24 (1999)
23. Nefian, A., Hayes III, M.H.: Maximum likelihood training of the embedded HMM for face detection and recognition. In: Proc. IEEE Int'l Conf. on Image Processing, pp. 33–36 (2000)
24. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philos. Mag. 2(6), 559–572 (1901)
25. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Hoffman, K.J.C., Marques, J., Worek, W.J.M.: Overview of the face recognition grand challenge. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 947–954 (2005)
26. Robert, G.U.J., de Van Lobo, N.: A framework for recognizing a facial image from a police sketch. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 586–593 (1996)

27. Robert, G.U.J., de Lobo, N., Van Kwon, Y.H.: Recognizing a facial image from a police sketch. In: Proc. 2nd IEEE workshop on applications of computer vision, pp. 129–137 (1994)
28. Scholkopf, B., Smola, A., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10(5), 1299–1319 (1998)
29. Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. B 36(2), 111–147 (1974)
30. Sun, J., Tao, D., Papadimitriou, S., Yu, P.S., Faloutsos, C.: Incremental tensor analysis: theory and applications. ACM Trans. on Knowledge Discovery from Data 2(3), 1–37 (2008)
31. Tang, X., Wang, X.: Face photo recognition using sketch. In: Proc. IEEE Int'l Conf. on Image Processing, pp. 257–260 (2002)
32. Tang, X., Wang, X.: Face sketch synthesis and recognition. In: Proc. IEEE Int'l Conf. on Computer Vision, pp. 687–694 (2003)
33. Tang, X., Wang, X.: Face sketch recognition. IEEE Trans. Circuits Syst.Video Technol. 14(1), 50–57 (2004)
34. Tao, D., Li, X., Wu, X., Maybank, S.: General tensor discriminant analysis and gabor features for gait recognition. IEEE Trans. Pattern Anal. Mach. Intell. 29(10), 1700–1715 (2007)
35. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 586–591 (1991)
36. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neurosci. 3(1), 71–86 (1991)
37. Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theory 13(2), 260–269 (1967)
38. Wang, Z., Bovik, A.: A universal image quality index. Signal Processing Lett. 9(3), 81–84 (2002)
39. Xiao, B., Gao, X., Tao, D., Li, X.: A new approach for face recognition by sketches in photos. Signal Process. 89(8), 1576–1588 (2009)
40. Xiao, B., Gao, X., Tao, D., Li, X., Li, J.: Photo-sketch synthesis and recognition based on subspace learning. Neurocomputing (2009) (submitted)
41. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.: Face recognition: a literature survey. ACM Comput. Surv. 35(4), 399–458 (2003)
42. Zhong, J., Gao, X., Tian, C.: Face sketch synthesis using E-HMM and selective ensemble. In: Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing, pp. 485–488 (2007)
43. Tao, D., Li, X., Wu, X., Maybank, S.: Geometric mean for subspace selection. IEEE Trans. Pattern Anal. Mach. Intell. 31(2), 260–274 (2009)
44. Tao, D., Li, X., Wu, X., Hu, X., Maybank, S.: Supervised tensor learning. Knowl. Inf. Syst. 13(1), 1–42 (2007)
45. Tao, D., Li, X., Wu, X., Maybank, S.: Tensor rank one discriminant analysis – a convergent method for discriminative multilinear subspace selection. Neurocomputing 71(10-12), 1866–1882 (2008)
46. Zhang, T., Tao, D., Yang, J.: Discriminative locality alignment. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 725–738. Springer, Heidelberg (2008)
47. Gao, X., Lu, W., Tao, D., Li, X.: Image quality assessment based on multiscale geometric analysis. IEEE Trans. Image Processing 18(7), 1608–1622 (2009)

Part II
Video Processing and Analysis

Object Detection and Tracking for Intelligent Video Surveillance

Kyungnam Kim and Larry S. Davis

[1] HRL Laboratories, LLC., Malibu CA, USA
kkim@hrl.com
[2] Computer Science Dept. University of Maryland, College Park, MD, USA
lsd@cs.umd.edu

Abstract. As CCTV/IP cameras and network infrastructure become cheaper and more affordable, today's video surveillance solutions are more effective than ever before, providing new surveillance technology that's applicable to a wide range end-users in retail sectors, schools, homes, office campuses, industrial /transportation systems, and government sectors. Vision-based object detection and tracking, especially for video surveillance applications, is studied from algorithms to performance evaluation. This chapter is composed of three topics: (1) background modeling and detection, (2) performance evaluation of sensitive target detection, and (3) multi-camera segmentation and tracking of people.

Keywords: video surveillance, object detection and tracking, background subtraction, performance evaluation, multi-view people tracking, CCTV/IP cameras.

Overview

This book chapter describes vision-based object detection and tracking for video surveillance application. It is organized into three sections. In Section 1, we describe a codebook-based background subtraction (BGS) algorithm used for foreground detection. We show that the method is suitable for both stationary and moving backgrounds in different types of scenes, and applicable to compressed videos such as MPEG. Important improvements to the above algorithm are presented - automatic parameter estimation, layered modeling/detection and adaptive codebook updating. In Section 2, we describe a performance evaluation technique, named PDR analysis. It measures the sensitivity of a BGS algorithm without assuming knowledge of the actual foreground distribution. Then PDR evaluation results for four different background subtraction algorithms are presented along with some discussions. In Section 3, a multi-view multi-target multi-hypothesis tracker is proposed. It segments and tracks people on a ground plane. Human appearance models are used to segment foreground pixels obtained from background subtraction. We developed a method to effectively integrate segmented blobs across views on a top-view reconstruction, with a help of ground plane homography. The multi-view tracker is extended efficiently to a multi-hypothesis framework (M^{3}Tracker) using particle filtering.

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 265–288.
springerlink.com

1 Background Modeling and Foreground Detection

Background subtraction algorithm

The codebook (CB) background subtraction algorithm we describe in this section adopts a quantization/clustering technique [5], to construct a background model (see [28] for more details). Samples at each pixel are clustered into a set of codewords. The background is encoded on a pixel by pixel basis.

Let $\mathcal{X}$ be a training sequence for a single pixel consisting of N RGB-vectors: $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$. Let $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_L\}$ represent the codebook for the pixel consisting of L codewords. Each pixel has a different codebook size based on its sample variation. Each codeword $\mathbf{c}_i$, $i = 1 \ldots L$, consists of an RGB vector $\mathbf{v}_i = (\bar{R}_i, \bar{G}_i, \bar{B}_i)$ and a 6-tuple $\mathbf{aux}_i = \langle \check{I}_i, \hat{I}_i, f_i, \lambda_i, p_i, q_i \rangle$. The tuple $\mathbf{aux}_i$ contains intensity (brightness) values and temporal variables described below.

$\check{I}, \hat{I}$: the *min* and *max* brightness, respectively,
that the codeword accepted;
f : the *frequency* with which the codeword has occurred;
λ : the *maximum negative run-length* (MNRL)
defined as the longest interval during the
training period that the codeword has NOT recurred;
p, q : the *first* and *last* access times, respectively,
that the codeword has occurred.

In the training period, each value, $\mathbf{x}_t$, sampled at time t is compared to the current codebook to determine which codeword $\mathbf{c}_m$ (if any) it matches (m is the matching codeword's index). We use the matched codeword as the sample's encoding approximation. To determine which codeword will be the best match, we employ a color distortion measure and brightness bounds. The detailed pseudo algorithm is given below.

```
Algorithm for Codebook Construction

I. $L \leftarrow$ ($\leftarrow$ means assignment), $\mathcal{C} \leftarrow \emptyset$ (empty set)
II. for $t$=1 to $N$ do
    i. $\mathbf{x}_t = (R, G, B)$, $I \leftarrow R + G + B$
    ii. Find the codeword $\mathbf{c}_m$ in $\mathcal{C} = \{\mathbf{c}_i | 1 \leq i \leq L\}$ matching to $\mathbf{x}_t$ based on two
        conditions (a) and (b).
            (a) $colordist(\mathbf{x}_t, \mathbf{v}_m) \leq \epsilon_1$
            (b) $brightness(I, \langle \check{I}_m, \hat{I}_m \rangle) =$ true
    iii. If $\mathcal{C} = \emptyset$ or there is no match, then $L \leftarrow L + 1$. Create a new codeword $\mathbf{c}_L$ by
        setting
        ■ $\mathbf{v}_L \leftarrow (R, G, B)$
        ■ $\mathbf{aux}_L \leftarrow \langle I, I, 1, t-1, t, t \rangle$.
    iv. Otherwise, update the matched codeword $\mathbf{c}_m$, consisting of $\mathbf{v}_m = (\bar{R}_m, \bar{G}_m, \bar{B}_m)$
        and $\mathbf{aux}_m = \langle \check{I}_m, \hat{I}_m, f_m, \lambda_m, p_m,$
        $q_m \rangle$, by setting
        ■ $\mathbf{v}_m \leftarrow (\frac{f_m\bar{R}_m+R}{f_m+1}, \frac{f_m\bar{G}_m+G}{f_m+1}, \frac{f_m\bar{B}_m+B}{f_m+1})$
        ■ $\mathbf{aux}_m \leftarrow \langle\ min\{I, \check{I}_m\},\ max\{I, \hat{I}_m\},\ f_m + 1,$
          $max\{\lambda_m, t - q_m\},\ p_m,\ t\ \rangle.$
    end for
III. For each codeword $\mathbf{c}_i$, $i = 1 \dots L$, wrap around $\lambda_i$ by setting $\lambda_i \leftarrow max\{\lambda_i, (N - q_i + p_i - 1)\}$.
```

The two conditions (a) and (b) are satisfied when the pure colors of $\mathbf{x}_t$ and $\mathbf{c}_m$ are close enough and the brightness of $\mathbf{x}_t$ lies between the acceptable brightness bounds of $\mathbf{c}_m$. Instead of finding the nearest neighbor, we just find the first codeword to satisfy these two conditions. ϵ_1 is the sampling threshold (bandwidth).

We refer to the codebook obtained from the previous step as the *fat* codebook. In the temporal filtering step, we refine the fat codebook by separating the codewords that might contain moving foreground objects from the true background codewords, thus allowing moving foreground objects during the initial training period. The true background, which includes both static pixels and moving background pixels, usually is quasi-periodic (values recur in a bounded period). This motivates the temporal criterion of MNRL (λ), which is defined as the maximum interval of time that the codeword has not recurred during the training period.

Let $\mathcal{M}$ denote the background model (a new codebook after temporal filtering):

$$\mathcal{M} = \{\mathbf{c}_m | \mathbf{c}_m \in \mathcal{C} \ \wedge \ \lambda_m \leq T_{\mathcal{M}}\}. \tag{1}$$

Usually, a threshold $T_{\mathcal{M}}$ is set equal to half the number of training frames, $\frac{N}{2}$.

To cope with the problem of illumination changes such as shading and highlights, we utilize a color model [28] separating the color and brightness components. When we consider an input pixel $\mathbf{x}_t = (R, G, B)$ and a codeword $\mathbf{c}_i$ where $\mathbf{v}_i = (\bar{R}_i, \bar{G}_i, \bar{B}_i)$, we have $\|\mathbf{x}_t\|^2 = R^2 + G^2 + B^2$, $\|\mathbf{v}_i\|^2 = \bar{R}_i^2 + \bar{G}_i^2 + \bar{B}_i^2$, $\langle \mathbf{x}_t, \mathbf{v}_i \rangle^2 = (\bar{R}_i R + \bar{G}_i G + \bar{B}_i B)^2$.

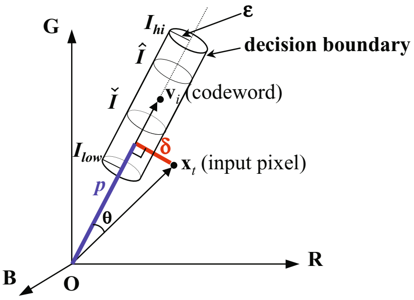


Fig. 1. The proposed color model - separate evaluation of color distortion and brightness distortion

The color distortion δ can be calculated by

$$\begin{aligned} p^2 = \|\mathbf{x}_t\|^2 \cos^2\theta = \tfrac{\langle \mathbf{x}_t, \mathbf{v}_i\rangle^2}{\|\mathbf{v}_i\|^2} \\ colordist(\mathbf{x}_t, \mathbf{v}_i) = \delta = \sqrt{\|\mathbf{x}_t\|^2 - p^2}. \end{aligned} \tag{2}$$

The logical brightness function is defined as

$$brightness(I, \langle \check{I}, \hat{I} \rangle) = \begin{cases} \textbf{true} & \text{if } I_{low} \leq \|\mathbf{x}_t\| \leq I_{hi} \\ \textbf{false} & \text{otherwise.} \end{cases} \tag{3}$$

Subtracting the current image from the background model is straightforward. Unlike Mixture-of-Gaussians (MOG) [2] or Kernel method [4] which compute probabilities using costly floating point operations, our method does not involve probability calculation. Indeed, the probability estimate in [4] is dominated by the nearby training samples. We simply compute the distance of the sample from the nearest cluster mean. This is very fast and shows little difference in detection compared with the probability estimate. The subtraction operation $BGS(\mathbf{x})$for an incoming pixel value $\mathbf{x}$ in the test set is defined as:

```
Algorithm for Background Subtraction (Foreground Detection)

  I. $\mathbf{x} = (R, G, B), \quad I \leftarrow R + G + B$
 II. For all codewords in $\mathcal{M}$ in Eq.1, find the codeword $\mathbf{c}_m$ matching to $\mathbf{x}$ based on two
     conditions:
       ■ $colordist(\mathbf{x}, \mathbf{v}_m) \leq \epsilon_2$
       ■ $brightness(I, \langle \check{I}_m, \hat{I}_m \rangle) = \textbf{true}$
III. $BGS(\mathbf{x}) = \begin{cases} \textbf{foreground} & \text{if there is no match} \\ \textbf{background} & \text{otherwise.} \end{cases}$
```

ϵ_2 is the detection threshold.

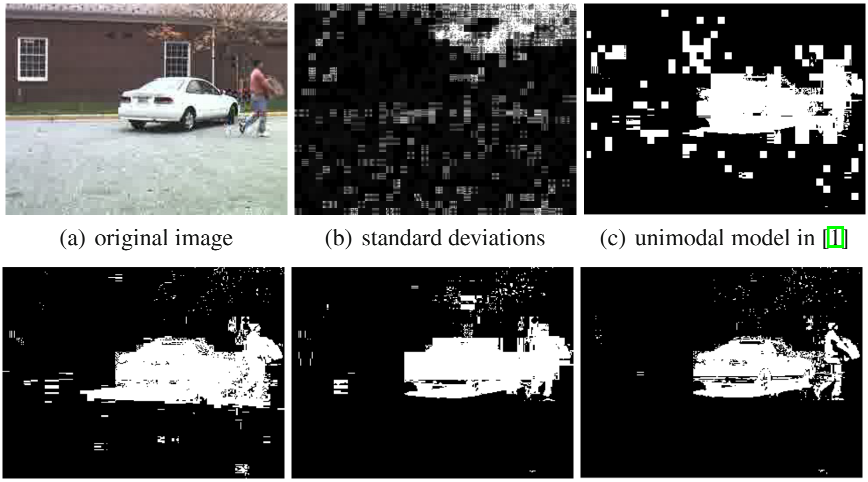
Detection Results and Comparison

This section demonstrates the performance of the proposed algorithm compared with MOG [2] and Kernel [4].

Fig.2(a) is an image extracted from the MPEG video encoded at 70 kbits/sec. Fig.2(b) depicts a 20-times scaled image of the standard deviations of green(G)-channel values in the training set. The distribution of pixel values has been affected by the blocking effects of MPEG. The unimodal model in Fig.2(c) suffers from these effects. For compressed videos having very abnormal distributions, CB eliminates most compression artifacts - see Fig.2(c)-2(f).

To test unconstrained training, we applied the algorithms to a video in which people are almost always moving in and out a building (see Fig.3(a)-3(d)). By λ-filtering, CB was able to obtain the most complete background model.

Multiple backgrounds moving over a long period of time cannot be well trained with techniques having limited memory constraints. A sequence of 1000 frames recorded at 30 frames per second (fps) was trained. It contains trees moving irregularly over that period. The number of Gaussians allowed for MOG was 20. A sample of size 300 was used to represent the background for Kernel. Fig.4(a)-4(d) shows that CB captures most multiple background events. This is due to a compact background model represented by quantized codewords. The implementation of our approach is straightforward and it is faster than MOG and Kernel.


(a) original image (b) standard deviations (c) unimodal model in [1]

(d) MOG (e) Kernel (f) CB

Fig. 2. Detection results on a compressed video

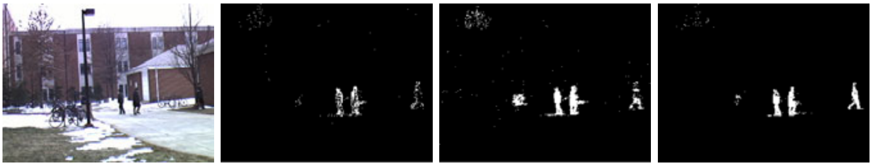

(a) original image (b) MOG (c) Kernel (d) CB

Fig. 3. Detection results on training of non-clean backgrounds

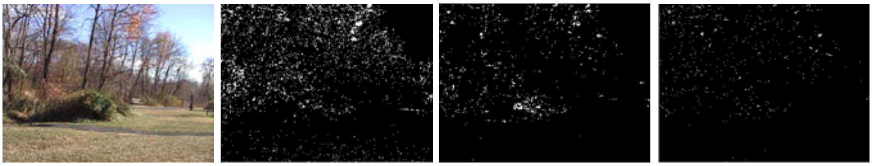

(a) original image (b) MOG (c) Kernel (d) CB

Fig. 4. Detection results on very long-time backgrounds

Automatic Parameter Estimation - ϵ_1 and ϵ_2

Automatic parameter selection is an important goal for visual surveillance systems as addressed in [7]. Two of our parameters, ϵ_1 and ϵ_2, are automatically determined. Their values depend on variation within a single background distribution, and are closely related to false alarm rates. First, we find a robust measure of background variation computed over a sequence of frames (of at least 90 consecutive frames, about 3 seconds of video data). In order to obtain this robust measure, we calculate the median color consecutive-frame difference over pixels. Then we calculate Θ (median color frame difference) which is the median over time of these median differences over space. For example, suppose we have a sequence of N images. We consider the first pair of frames, and calculate the color difference at each pixel, and take the median over space. We do this for all $N-1$ consecutive pairs, until we have $N-1$ medians. Then, Θ is the median of the $N-1$ values. In fact, an over-space median of medians over time is almost the same as Θ, while Θ is much easier to calculate with limited memory. Θ will be proportional to the within class variance of a single background. In addition, it will be a robust estimate, which is insensitive to the presence of relatively small areas of moving foreground objects. The same color difference metric should be used as in the background modeling and subtraction.

Finally, we multiply a constant k by this measure to obtain $\epsilon_1(= k\Theta)$. The default value of k is 4.5 which corresponds approximately to a false alarm rate of detection between .0001 - .002. ϵ_2 can be set to $k'\Theta$, where $(k-1) < k' < (k+1)$ but usually $k' = k$. Experiments on many videos show that these automatically chosen threshold parameters ϵ_1 and ϵ_2 are sufficient. However, they are not always acceptable, especially for highly compressed videos where we cannot always measure the robust median accurately.

Layered modeling and detection - Model maintenance

The scene can change after initial training, for example, by parked cars, displaced books, etc. These changes should be used to update the background model. We achieve this by defining an additional model $\mathcal{H}$ called a *cache* and three parameters described below:

- $T_{\mathcal{H}}$: the threshold for MNRL of the codewords in $\mathcal{H}$;
- T_{add}: the minimum time period required for addition, during which the codeword must reappear;
- T_{delete}: a codeword is deleted if it has not been accessed for a period of this long.

The periodicity of an incoming pixel value is filtered by $T_{\mathcal{H}}$, as we did in the background modeling. The values re-appearing for a certain amount of time (T_{add}) are added to the background model as short-term background. Some parts of a scene may remain in the foreground unnecessarily long if adaptation is slow, but other parts will disappear too rapidly into the background if adaptation if fast. Neither approach is inherently better than the other. The choice of this adaptation speed is problem dependent.

We assume that the background obtained during the initial background modeling is long-term. This assumption is not necessarily true, e.g., a chair can be moved after the initial training, but, in general, most long-term backgrounds are obtainable during training. Background values not accessed for a long time (T_{delete}) are deleted from the background model. Optimally, the long-term codewords are augmented with permanent flags indicating they are not to be deleted*. The permanent flags can be applied otherwise depending on specific application needs.

Thus, a pixel can be classified into four subclasses - (1) background found in the long-term background model, (2) background found in the short-term background model, (3) foreground found in the cache, and (4) foreground not found in any of them. The overview of the approach is illustrated in Fig.5. This adaptive modeling capability allows us to capture changes to the background scene.

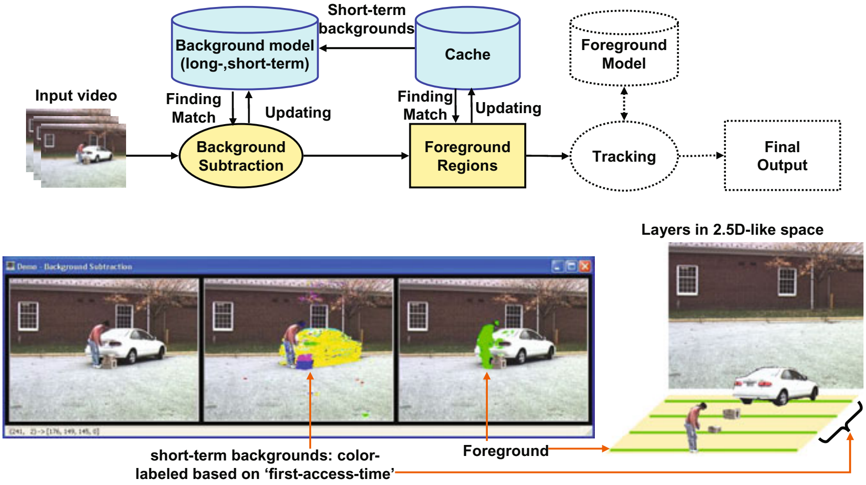


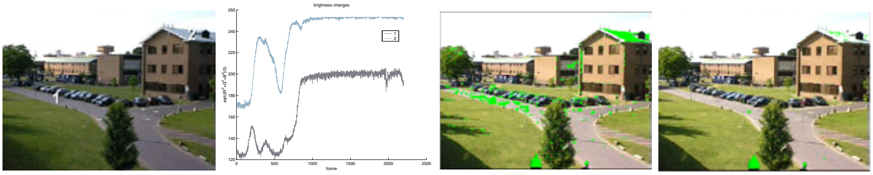
Fig. 5. The overview of our approach with short-term background layers: the foreground and the short-term backgrounds can be interpreted in a different temporal order. The diagram items in dotted line, such as Tracking, are added to complete a video surveillance system.

Adaptive codebook updating - detection under global illumination changes

Global illumination changes (for example, due to moving clouds) make it difficult to conduct background subtraction in outdoor scenes. They cause over-detection, false alarms, or low sensitivity to true targets. Good detection requires equivalent false alarm rates over time and space. We discovered from experiments that variations of pixel values are different (1) at different surfaces (shiny or muddy), and (2) under different levels of illumination (dark or bright). Codewords should be adaptively updated during illumination changes. Exponential smoothing of codeword vector and variance with suitable learning rates is efficient in dealing with illumination changes. It can be done by replacing the updating formula of $\mathbf{v}_m$ with $\mathbf{v}_m \leftarrow \gamma \mathbf{x}_t +$

$(1-\gamma)\mathbf{v}_m$ and appending $\sigma_m^2 \leftarrow \rho\delta^2 + (1-\rho)\sigma_m^2$ to Step II-iv of the algorithm for codebook construction. γ and ρ are learning rates. Here, σ_m^2 is the overall variance of color distortion in the color model, not the variance of RGB. σ_m is initialized when the algorithm starts. Finally the function $colordist()$ in Eq.2 is modified to $colordist(\mathbf{x}_t, \mathbf{v}_i) = \frac{\delta}{\sigma_i}$.

We tested a PETS'2001[1] sequence which is challenging in terms of multiple targets and significant lighting variation. Fig.6(a) shows two sample points (labelled 1 and 2) which are significantly affected by illumination changes and Fig.6(b) shows the brightness changes of those two points. As shown in Fig.6(d), adaptive codebook updating eliminates the false detection which occurs on the roof and road in Fig.6(c).



(a) original image - frame 1 (b) brightness changes - blue on roof, gray on road (c) before adaptive updating (d) after adaptive updating

Fig. 6. Results of adaptive codebook updating for detection under global illumination changes. Detected foregrounds on the frame 1105 are labelled with green color.

2 Performance Evaluation of Sensitive Target Detection

In this section, we propose a methodology, called Perturbation Detection Rate (PDR) Analysis [6], for measuring performance of BGS algorithms, which is an alternative to the common method of ROC analysis. The purpose of PDR analysis is to measure the detection sensitivity of a BGS algorithm without assuming knowledge of the actual foreground distribution. In PDR, we do not need to know exactly what the distributions are. The basic assumption made is that the shape of the foreground distribution is locally similar to that of the background distribution; however, foreground distribution of small ("just-noticeable") contrast will be a shifted or perturbed version of the background distribution. This assumption is fairly reasonable because, in modeling video, any object with its color could be either background or foreground, e.g., a parked car could be considered as a background in some cases; in other cases, it could be considered a foreground target. Furthermore, by varying algorithm parameters we determine not a pair of error rates but a relation among the false alarm and detection rates and the distance between the distributions.

Given the parameters to achieve a certain fixed FA-rate, the analysis is performed by shifting or perturbing the entire BG distributions by vectors in uniformly random

[1] IEEE International Workshop on Performance Evaluation of Tracking and Surveillance 2001 at http://www.visualsurveillance.org/PETS2001

directions of RGB space with fixed magnitude Δ, computing an average detection rate as a function of contrast Δ. It amounts to simulating possible foregrounds at certain color distances. In the PDR curve, we plot the detection rate as a function of the perturbation magnitude Δ given a particular FA-rate.

First, we train each BGS algorithm on N training background frames, adjusting parameters as best we can to achieve a target FA-rate which would be practical in processing the video. Typically this will range from $.01\%$ to 1% depending on video image quality. To obtain a test foreground at color contrast Δ, we pass through the N background frames again. For each frame, we perturb a random sample of M pixel values (R_i, G_i, B_i) by a magnitude Δ in uniformly random directions.

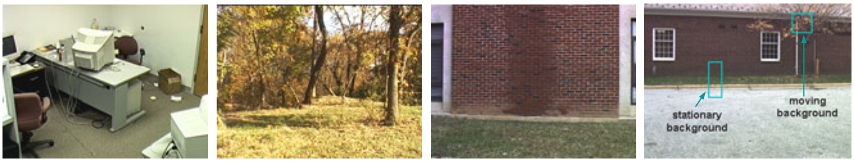
The perturbed, foreground color vectors (R', G', B') are obtained by generating points randomly distributed on the color sphere with radius Δ. Then we test the BGS algorithms on these perturbed, foreground pixels and compute the detection rate for the Δ. By varying the foreground contrast Δ, we obtain an monotone increasing PDR graph of detection rates. In some cases, one algorithm will have a graph which dominates that of another algorithm for all Δ. In other cases, one algorithm may be more sensitive only in some ranges of Δ. Most algorithms perform very well for a large contrast Δ, so we are often concerned with small contrasts ($\Delta < 40$) where differences in detection rates may be large.

In this study, we compare four algorithms shown in Table 1. Since the algorithm in [4] accepts normalized colors (KER) or RGB colors (KER.RGB) as inputs, it has two separate graphs. Figure 2 shows the representative images from four test videos.

To generate PDR curves, we collected 100 empty consecutive frames from each video. 1000 points are randomly selected at each frame. That is, for each Δ, (100)×(1000) perturbations and detection tests were performed. Those 100 empty frames are also used for training background models. During testing, no updating of the background model is allowed. For the non-parametric model in KER and KER.RGB, a sample size 50 was used to represent the background. The maximum

Table 1. Four algorithms used in PDR performance evaluation.

Name	Background subtraction algorithm
CB	codebook-based method described in Section 2
MOG	mixture of Gaussians described in [2]
KER and **KER.RGB**	non-parametric method using kernels described in [4]
UNI	unimodal background modeling described in [1]



(a) indoor office (b) outdoor woods (c) red-brick wall (d) parking lot

Fig. 7. The sample empty-frames of the videos used for the experiments

number of Gaussians allowed in MOG is 4 for the video having stationary backgrounds and 10 for moving backgrounds. We do not use a fixed FA-rate for all four videos. The FA-rate for each video is determined by these three factors - video quality, whether it is indoor or outdoor, and good real foreground detection results for most algorithms. The FA-rate chosen this way is practically useful for each video. The threshold value for each algorithm has been set to produce a given FA-rate. In the case of MOG, the learning rate, α, was fixed to 0.01 and the minimum portion of the data for the background, T, was adjusted to give the desired FA-rate. Also, the cluster match test statistic was set to 2 standard deviations. Unless noted otherwise, the above settings are used for the PDR analysis.

Evaluation Results

Figures 9(a) and 9(b) show the PDR graphs for the videos in Figures 7(a) and 7(b) respectively. For the indoor office video, consisting almost entirely of stationary backgrounds, CB and UNI perform better than the others. UNI, designed for unimodal backgrounds, has good sensitivity as expected. KER performs intermediately. MOG and KER.RGB do not perform as well for small contrast foreground Δ, probably because, unlike the other algorithms, they use original RGB variables and don't separately model brightness and color. MOG currently does not model covariances which are often large and caused by variation in brightness. It is probably best to explicitly model brightness. MOG's sensitivity is consistently poor in all our test videos, probably for this reason.

For the outdoor video, all algorithms perform somewhat worse even though the FA-rate has been increased to 1% from .01%. CB and KER, both of which model mixed backgrounds and separate color/brightness, are most sensitive, while, as expected, UNI does not perform well as in the indoor case. KER.RGB and MOG are also less sensitive outdoors, as before indoors.

Figure 2 depicts a real example of foreground detection, showing real differences in detection sensitivity for two algorithms. These real differences reflect performance shown in the PDR graph in Figure 9(c). The video image in Figure 8(a) shows someone with a red sweater standing in front of a brick wall of somewhat different reddish color. There are detection holes through the sweater (and face) in the MOG result (Figure 8(b)) . The CB result in Figure 8(c) is much better for this

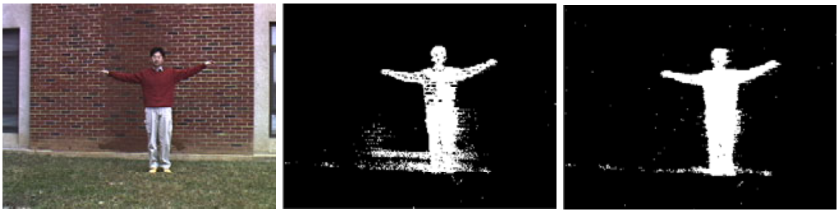

(a) original frame of a person in a red sweater
(b) detection using MOG
(c) detection using CB

Fig. 8. Sensitive detection at small delta

small contrast. After inspection of the image, the magnitude of contrast Δ was determined to be about 16 in missing spots. This was due to difference in color balance and not overall brightness. Figure 9(c) shows a large difference in detection for this contrast, as indicated by the vertical line.

Figures 9(d), 9(e), 9(f) show how sensitively the algorithms detect foregrounds against a scene containing moving backgrounds (trees) as well as stationary surfaces. In order to sample enough moving background events, 300 frames are allowed for training. As for previous videos, a PDR graph for the ‘parking lot’ video is given in Figure 9(d). Two windows are placed to represent ‘stationary’ and ‘moving backgrounds’ as shown in Figure 7(d). PDR analysis is performed on each window with the FA-rate obtained only within the window - a ‘window’ false alarm rate (instead of ‘frame’ false alarm rate).

Since most of the frame is stationary background, as expected, the PDR graph (Figure 9(e)) for the stationary background window is very close to that for the entire frame. On the other hand, the PDR graph (Figure 9(f)) for the moving background window is generally shifted right, indicating reduced sensitivity of all algorithms for moving backgrounds. Also, it shows differences in performance among algorithms, with CB and KER performing best. These results are qualitatively similar those for the earlier example of outdoor video shown in Figure 5. We can offer the same explanation as before: CB and KER were designed to handle mixed backgrounds, and they separately model brightness and color. In this video experiment, we had to increase the background sample size of KER to 270 frames from 50 in order to achieve the target FA-rate in the case of the moving background window. It should be noted that CB, like MOG, usually models background events over a longer period than KER.

3 Multi-camera Segmentation and Tracking of People

A multi-view multi-hypothesis approach, named M^{3}Tracker, to segmenting and tracking multiple (possibly occluded) persons on a ground plane is presented. During tracking, several iterations of segmentation are performed using information from human appearance models and ground plane homography. The full algorithm description is available in [30].

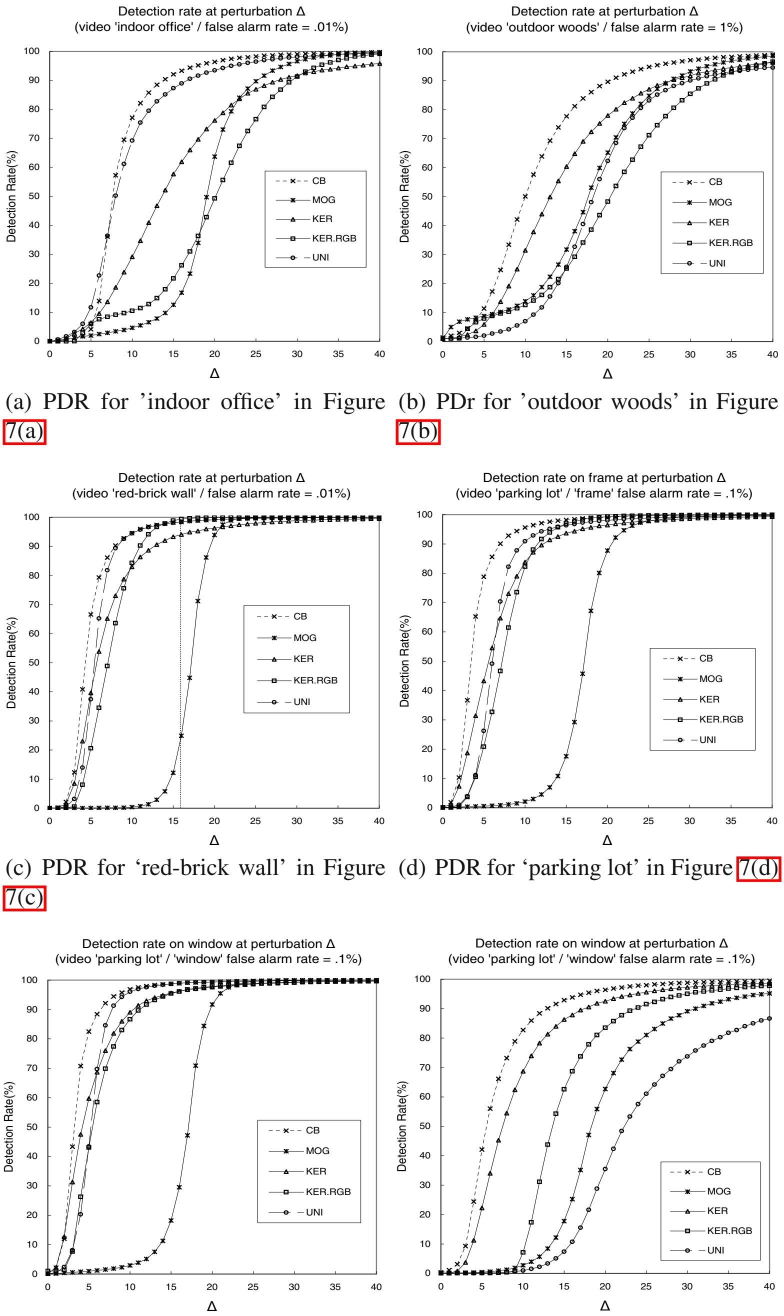
Survey on people tracking techniques

Table 2[2] lists different single-camera and multi-camera algorithms for people tracking along with their characteristics.

Human appearance model

First, we describe an appearance color model as a function of height that assumes that people are standing upright and are dressed, generally, so that consistently colored or textured color regions are aligned vertically. Each body part has its own color

[2] MCMC: Markov chain Monte Carlo, KLT: Kanade-Lucas-Tomasi, JPDAF: Joint Probabilistic Data Association Filter, Tsai’s: [22], I: indoor, O: outdoor, n/a: not applicable.



(a) PDR for 'indoor office' in Figure 7(a)

(b) PDr for 'outdoor woods' in Figure 7(b)

(c) PDR for 'red-brick wall' in Figure 7(c)

(d) PDR for 'parking lot' in Figure 7(d)

(e) PDR for window on stationary background (Figure 7(d))

(f) PDR for window on moving background (Figure 7(d))

Fig. 9. PDR graphs

Table 2. Characteristics of people tracking algorithms

Comparison chart 1

	Algorithm	Tracking	Segmentation	Occlusion Analysis	Human Appearance Model
single	Haritaoglu [8]	Heuristic	Yes	No	Temporal template
	Elgammal [9]	n/a	Yes	Yes	Kernel density est.
	Zhao [10]	MCMC	No	Yes	Shape, histogram
	Rabaud [11]	KLT tracker	No	No	Feature-based
multi-camera	Yang [12]	No (counting)	No	Yes	None
	Khan [13]	Look-ahead	No	Yes	None
	Kang [14]	JPDAF, Kalman	No	Yes	Polar color distrib.
	Javed [15]	Voting-based	No	No	Dynamic color histo.
	Mittal [16]	Kalman	Yes	Yes	Kernel density
	Eshel [17]	Score-based	No	Yes	None
	Jin [18]	Kalman	No	Yes	Color histogram
	Black [19]	Kalman	No	Yes	Unknown
	Xu [20]	Kalman	No	No	Histogram intersect.
	Fleuret [21]	Dynamic prog.	No	Yes	Color distrib.
	Ours	Particle filtering	Yes	Yes	Kernel density

Comparison chart 2

	Algorithm	Calibration	Area	Sensors	Background Subtraction	Initialization
single	Haritaoglu [8]	n/a	O	B/W	Yes	Auto
	Elgammal [9]	n/a	I	Color	Yes	Manual
	Zhao [10]	Yes	O	Color	Yes	Auto
	Rabaud [11]	No	O	Color	No	Auto
multi-camera	Yang [12]	Yes	I	Color	Yes	Auto
	Khan [13]	Homography	O	Color	Yes	n/a
	Kang [14]	Homography	O	Color	Yes	Unknown
	Javed [15]	No	I,O	Color	Yes	Manual
	Mittal [16]	Stereo	I	Color	Yes	Auto
	Eshel [17]	Homography	I,O	B/W	Yes	Unknown
	Jin [18]	Homography	I	Color, IR	Yes	Manual
	Black [19]	Tsai’s	O	Color	Yes	Auto
	Xu [20]	Tsai’s	O	Color	Yes	Auto
	Fleuret [21]	Homography	I,O	Color	Yes	Unknown
	Ours	Homography	I,O	Color	Yes	Auto

model represented by a color distribution. To allow multimodal densities inside each part, we use kernel density estimation.

Let $M = \{\mathbf{c}_i\}_{i=1...N_M}$ be a set of pixels from a body part with colors $\mathbf{c}_i$. Using Gaussian kernels and an independence assumption between d color channels, the probability that an input pixel $\mathbf{c} = \{c_1, ..., c_d\}$ is from the model M is estimated as

$$p_M(\mathbf{c}) = \frac{1}{N_M} \sum_{i=1}^{N_M} \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{c_j - c_{i,j}}{\sigma_j}\right)^2} \tag{4}$$

In order to handle illumination changes, we use normalized color $(r = \frac{R}{R+G+B}, g = \frac{G}{R+G+B}, s = \frac{R+G+B}{3})$ or Hue-Saturation-Value (HSV) color space with a wider kernel for 's' and 'V' to cope with the higher variability of these lightness variables. We used both the normalized color and HSV spaces in our experiments and observed similar performances.

Viewpoint-independent models can be obtained by viewing people from different perspectives using multiple cameras. A related calibration issue was addressed in [24, 26] since each camera output of the same scene point taken at the same time or different time may vary slightly depending on camera types and parameters.

Multi-camera Multi-person Segmentation and Tracking

Foreground segmentation. Given image sequences from multiple overlapping views including people to track, we start by performing detection using background subtraction to obtain the foreground maps in each view. The codebook-based background subtraction algorithm is used.

Each foreground pixel in each view is labelled as the best matching person (i.e., the most likely class) by Bayesian pixel classification as in [16]. The posterior probability that an observed pixel **x** (containing both color **c** and image position (x, y) information) comes from person k is given by

$$P(k|\mathbf{x}) = \frac{P(k)P(\mathbf{x}|k)}{P(\mathbf{x})} \tag{5}$$

We use the color model in Eq.4 for the conditional probability $P(\mathbf{x}|k)$. The color model of the person's body part to be evaluated is determined by the information of **x**'s position as well as the person's ground point and full-body height in the camera view (See Fig.10(a)). The ground point and height are determined initially by the method defined subsequently in Sec.3.

The prior reflects the probability that person k occupies pixel **x**. Given the ground point and full-body height of the person, we can measure **x**'s height from the ground and its distance to the person's center vertical axis. The occupancy probability is then defined by

$$O_k(h_k(\mathbf{x}), w_k(\mathbf{x})) = P[w_k(\mathbf{x}) < W(h_k(\mathbf{x}))] = 1 - \mathrm{cdf}_{W(h_k(\mathbf{x}))}(w_k(\mathbf{x})) \quad (6)$$

where $h_k(\mathbf{x})$ and $w_k(\mathbf{x})$ are the height and width of $\mathbf{x}$ relative to the person k. h_k and w_k are measured relative to the full height of the person. $W(h_k(\mathbf{x}))$ is the person's height-dependent width and $\mathrm{cdf}_W(.)$ is the cumulative density function for W. If $\mathbf{x}$ is located at distance $W(h_k(\mathbf{x}))$ from the person's center at a distance W, the occupancy probability is designed so that it will be exactly 0.5 (while it increases or decreases as $\mathbf{x}$ move towards or move away from the center).

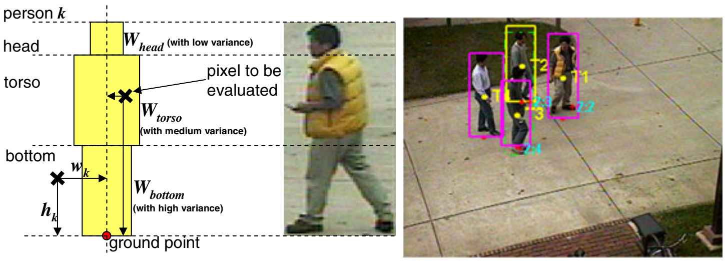
The prior must also incorporate possible occlusion. Suppose that some person l has a lower ground point than a person k in some view. Then the probability that l occludes k depends on their relative positions and l's (probabilistic) width. Hence, the prior probability $P(k)$ that a pixel $\mathbf{x}$ is the image of person k, based on this occlusion model, is

$$P(k) = O_k(h_k, w_k) \prod_{g_y(k) < g_y(l)} (1 - O_l(h_l, w_l)) \quad (7)$$

where $g_y(k)$ is the y-location of the ground point of k and $\mathbf{x}$ is omitted for simplicity (i.e., $h_k = h_k(\mathbf{x})$ and $w_k = w_k(\mathbf{x})$).

The best class k^* is determined by maximum a posteriori (MAP) estimation: $k^* = \arg\max_k P(k)P(\mathbf{x}|k)$. Finally, the foreground maps are segmented into the best matching persons based on their appearance models and occlusion information.

Model initialization and update. Full automatic tracking is enabled by initializing the human appearance model when a person is detected in a view by searching for isolated foreground blobs (See Fig. 10(b)). In order to get a bounding box of a person from the foreground map, we used the object detection technique in [25]. The bounding boxes in the figure were created when the blobs are isolated before. For the case when a person does not constitute an isolated blob, a manual selection is employed. The full-body height of a person is initialized upon model creation


person k
head
torso
bottom
W_{head} (with low variance)
pixel to be evaluated
W_{torso} (with medium variance)
w_k
h_k
W_{bottom} (with high variance)
ground point

(a) (b)

Fig. 10. (a) Illustration of appearance model, (b) Bounding box detection and is updated during segmentation. When the unclassified pixels (those having a probability in Eq.4 lower than a given threshold) constitute a connected component of non-negligible size, a new appearance model should be created.

Multi-view integration. Ground plane homography: The segmented blobs across views are integrated to obtain the ground plane locations of people. The correspondence of a human across multiple cameras is established by the geometric constraints of planar homographies. For N_V camera views, $N_V(N_V - 1)$ homography matrices can possibly be calculated for correspondence; but in order to reduce the computational complexity we instead reconstruct the top-view of the ground plane on which the hypotheses of peoples' locations are generated.

Integration by vertical axes: Given the pixel classification results from Sec.3, a ground point of a person could be simply obtained by detecting the lowest point of the person's blob. However those ground points are not reliable due to the errors from background subtraction and segmentation.

We, instead, develop a localization algorithm that employs the center vertical axis of a human body, which can be estimated more robustly even with poor background subtraction [29]. Ideally, a person's body pixels are arranged more of less symmetrically about a person's central vertical axis. An estimate of this axis can be obtained by Least Mean Squares of the perpendicular distance between the body pixel and the axis as in ③ in Fig.11. Alternatively, the Least *Median* Squares could be used since it is more robust to outliers.

The homographic images of all the vertical axes of a person across different views intersect at (or are very close to) a single point (the location of that person on the ground) when mapped to the top-view (See [29], [30]). In fact, even when the ground point of a person from some view is occluded, the top-view ground point integrated from all the views is obtainable if the vertical axis is estimated correctly. This intersection point can be calculated by minimizing the perpendicular distances to the axes. Fig.11 depicts an example of reliable detection of the ground point from the segmented blobs of a person. The N_v vertical axes are mapped to the top-view and transferred back to each image view.

Let each axis L_i be parameterized by two points $\{(x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2})\}_{i=1\dots N_V}$. When mapped to the top-view by homography as in ④ in Fig.11, we obtain $\{(\hat{x}_{i,1}, \hat{y}_{i,1}), (\hat{x}_{i,2}, \hat{y}_{i,2})\}_{i=1\dots N_V}$. The distance of a ground point (x, y) to the axis is written as $d\left((x, y), L_i\right) = \frac{|a_i x + b_i y + c_i|}{\sqrt{a_i^2 + b_i^2}}$ where $a_i = \hat{y}_{i,1} - \hat{y}_{i,2}$, $b_i = \hat{x}_{i,2} - \hat{x}_{i,1}$, and $c_i = \hat{x}_{i,1}\hat{y}_{i,2} - \hat{x}_{i,2}\hat{y}_{i,1}$. The solution is the point that minimizes a weighted sum of square distances:

$$(x^*, y^*) = \arg\min_{(x,y)} \sum_{i=1}^{N_V} w_i^2 d^2((x, y), L_i) \tag{8}$$

The weight w_i is determined by the segmentation quality (confidence level) of the body blob of L_i (We used the pixel classification score in Eq.5).

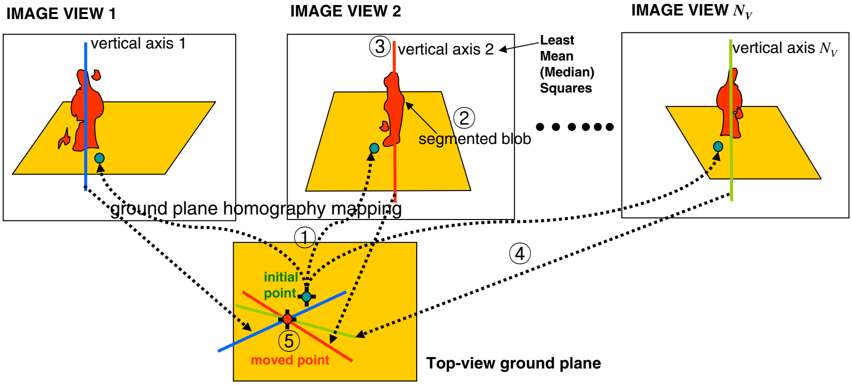


Fig. 11. All vertical axes of a person across views intersect at (or are very close to) a single point when mapped to the top-view

If a person is occluded severely by others in a view (i.e., the axis information is unreliable), the corresponding body axis from that view will not contribute heavily to the calculation in Eq.8. When only one axis is found reliably, then the lowest body point along the axis is chosen.

To obtain a better ground point and segmentation result, we can iterate the segmentation and ground-point integration process until the ground point converges to a fixed location within a certain bound ϵ. That is, given a set of initial ground-point hypotheses of people as in ① in Fig.11, segmentation in Sec.3 is performed (②), and then newly moved ground points are obtained based on multi-view integration (④ and ⑤). These new ground points are an input to the next iteration. 2-3 iterations gave satisfactory results for our data sets.

There are several advantages of our approach. Even though a person's ground point is invisible or there are segmentation and background subtraction errors, the robust final ground point is obtainable once at least two vertical axes are correctly detected. When total occlusion occurs from one view, robust tracking is possible using the other views' information if available; visibility of a person can be maximized if cameras are placed at proper angles. Since the good views for each tracked person are changing over time, our algorithm maximizes the effective usage of all available information across views. By iterating the multi-view integration process, a ground point moves to the optimal position that explains the segmentation results of all views. This nice property is used, in the next section, for a small number of hypotheses to explore in a large state space that incorporates multiple persons and multiple views.

Extension to Multi-hypothesis Tracker

Next, we extend our single-hypothesis tracker to one with multiple hypotheses. A single hypothesis tracker, while computationally efficient, can be easily distracted by occlusion or nearby similarly colored objects.

The iterative segmentation-searching presented in Sec.3 is naturally incorporated with a particle filtering framework. There are two advantages - (1) By searching for a person's ground point from a segmentation, a set of a few good particles can be identified, resulting in low computational costs, (2) Even if all the particles are away from the true ground point, some of them will move towards the true one as long as they are initially located nearby. This does not happen generally with particle filters, which need to wait until the target “comes to" the particles.

Our final M^{3}Tracker algorithm of segmentation and tracking is presented with a particle filter overview and our state space, dynamics, and observation model.

Overview of particle filter, state space, and dynamics. The key idea of particle filtering is to approximate a probability distribution by a weighted sample set $S = \{(\mathbf{s}^{(n)}, \pi^{(n)}) | n = 1...N\}$. Each sample, $\mathbf{s}$, represents one hypothetical state of the object, with a corresponding discrete sampling probability π, where $\sum_{n=1}^{N} \pi^{(n)} = 1$. Each element of the set is then weighted in terms of the observations and N samples are drawn with replacement, by choosing a particular sample with probability $\pi_t^{(n)} = P(\mathbf{z}_t | \mathbf{x}_t = \mathbf{s}_t^{(n)})$.

In our particle filtering framework, each sample of the distribution is simply given as $s = (x, y)$ where x, y specify the ground location of the object in the *top-view*. For multi-person tracking, a state $\mathbf{s}_t = (\mathbf{s}_{1,t}, ..., \mathbf{s}_{N_p,t})$ is defined as a combination of N_p single-person states. Our state transition dynamic model is a random walk where a new predicted single-person state is acquired by adding a zero mean Gaussian with a covariance $\mathbf{\Sigma}$ to the previous state. Alternatively, the velocity $\dot{x}, \dot{y}$ or the size variable $height$ and $width$ can be added to the state space and then a more complex dynamic model can be applied if relevant.

Observation. Each person is associated with a reference color model $\mathbf{q}^\star$ which is obtained by histogram techniques [27]. The histograms are produced using a function $b(\mathbf{c}_i) \in \{1, ..., N_b\}$ that assigns the color vector $\mathbf{c}_i$ to its corresponding bin. We used the color model defined in Sec.3 to construct the histogram of the reference model in the normalized color or HSV space using N_b (e.g., $10 \times 10 \times 5$) bins to make the observation less sensitive to lighting conditions.

The histogram $\mathbf{q}(C) = \{q(u; C)\}_{u=1...N_b}$ of the color distribution of the sample set C is given by

$$q(u; C) = \eta \sum_{i=1}^{N_C} \delta[b(\mathbf{c}_i) - u] \tag{9}$$

where u is the bin index, δ is the Kronecker delta function, and η is a normalizing constant ensuring $\sum_{u=1}^{N_b} q(u; C) = 1$. This model associates a probability to each of the N_b color bins.

If we denote $\mathbf{q}^\star$ as the reference color model and $\mathbf{q}$ as a candidate color model, $\mathbf{q}^\star$ is obtained from the stored samples of person k's appearance model as mentioned before while $\mathbf{q}$ is specified by a particle $\mathbf{s}_{k,t} = (x, y)$. The sample set C in Eq.9 is replaced with the sample set specified by $\mathbf{s}_{k,t}$. The top-view point (x, y) is transformed to an image ground point for a certain camera view v, $H_v(\mathbf{s}_{k,t})$, where H_v is a homography mapping the top-view to the view v. Based on the ground point, a region to be compared with the reference model is determined. The pixel values inside the region are drawn to construct $\mathbf{q}$. Note that the region can be constrained from the prior probability in Eq.7, including the occupancy and occlusion information (i.e., by picking pixels such that $P(k) > Threshold$, typically 0.5). In addition, as done in pixel classification, the color histograms are separately defined for each body part to incorporate the spatial layout of the color distribution. Therefore, we apply the likelihood as the sum of the histograms associated with each body part.

Then, we need to measure the data likelihood between $\mathbf{q}^\star$ and $\mathbf{q}$. The Bhattacharyya similarity coefficient is used to define a distance d on color histograms: $d[\mathbf{q}\star, \mathbf{q}(\mathbf{s})] = \left[1 - \sum_{u=1}^{N_b} \sqrt{q\star(u)q(u;\mathbf{s})}\right]^{\frac{1}{2}}$. Thus, the likelihood $(\pi_{v,k,t})$ of person k consisting of N_r body parts at view v, the actual view-integrated likelihood $(\pi_{k,t})$ of a person $\mathbf{s}_{k,t}$, and the final weight of the particle $(\pi_{k,t})$ of a concatenation of N_p person states are respectively given by:

$$\pi_{v,k,t} \propto e^{\sum_{r=1}^{N_r} -\lambda d^2[\mathbf{q}_r^\star, \mathbf{q}_r(H_v(\mathbf{s}_{k,t}))]}, \quad \pi_{k,t} = \Pi_{v=1}^{N_V} \pi_{v,k,t}, \quad \pi_t = \Pi_{k=1}^{N_p} \pi_{k,t} \tag{10}$$

where λ is a constant which can be experimentally determined.

The M^{3}Tracker algorithm. Iteration of segmentation and multi-view integration moves a predicted particle to an a better position on which all the segmentation results of the person agree. The transformed particle is re-sampled for processing of the next frames.

Algorithm for Multi-view Multi-target Multi-hypothesis tracking

```
I.   From the "old" sample set $S_{t-1} = \{\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)}\}_{n=1,...,N}$ at time $t-1$, construct the
     new samples as follows:
II.  **Prediction**: for $n = 1, ..., N$, draw $\tilde{\mathbf{s}}_t^{(n)}$ from the dynamics. **Iterate** Step III to IV for
     each particle $\tilde{\mathbf{s}}_t^{(n)}$.
III. **Segmentation & Search**
     $\tilde{\mathbf{s}}_t = \{\tilde{\mathbf{s}}_{k,t}\}_{k=1...N_p}$ contains all persons' states. The superscript $(n)$ is omitted
     through the Observation step.
      i. **for** $v \leftarrow 1$ to $N_V$ **do**
             (a) For each person $k$, $(k = 1...N_p)$, transform the top-view point $\tilde{\mathbf{s}}_{k,t}$ into
                 the ground point in view $v$ by homography, $H_v(\tilde{\mathbf{s}}_{k,t})$
             (b) perform segmentation on the foreground map in view $v$ with the occlusion
                 information according to Sec.5.
         **end for**
     ii. For each person $k$, obtain the center vertical axes of the person across views, then
         integrate them on the top-view to obtain a newly moved point $\tilde{\mathbf{s}}_{k,t}^*$ as in Sec.3.
    iii. For all persons, if $\|\tilde{\mathbf{s}}_{k,t} - \tilde{\mathbf{s}}_{k,t}^*\| < \varepsilon$, then go to the next step. Otherwise, set
         $\tilde{\mathbf{s}}_{k,t} \leftarrow \tilde{\mathbf{s}}_{k,t}^*$ and go to Step III-i.
```

```
IV. **Observation**
    i. **for** $v \leftarrow 1$ to $N_V$ **do**
           For each person $k$, estimate the likelihood $\pi_{v,k,t}$ in view $v$ according to Eq.10. $\tilde{\mathbf{s}}_{k,t}$ needs to be transferred to view $v$ by mapping through $H_v$ for evaluation. Note that $\mathbf{q}_r(H_v(\tilde{\mathbf{s}}_{k,t}))$ is constructed only from the non-occluded body region.
       **end for**
    ii. For each person $k$, obtain the person likelihood $\pi_{k,t}$ by Eq.10.
    iii. Set $\pi_t \leftarrow \Pi_{k=1}^{N_p} \pi_{k,t}$ as the final weight for the multi-person state $\tilde{\mathbf{s}}_t$.
V. **Selection**: Normalize $\{\pi_t^{(n)}\}_i$ so that $\sum_{n=1}^{N} \pi_t^{(n)} = 1$.
    For $i = n...N$, sample index $a(n)$ from discrete probability $\{\pi_t^{(n)}\}_i$ over $\{1...N\}$, and set $\mathbf{s}_t^{(n)} \leftarrow \tilde{\mathbf{s}}_t^{a(n)}$.
VI. **Estimation**: the mean top-view position of person $k$ is $\sum_{n=1}^{N} \pi_t^{(n)} \mathbf{s}_{k,t}^{(n)}$.
```

People Tracking Results

The results on the indoor sequences are depicted in Fig.12. The bottom-most row shows how the persons' vertical axes are intersecting on the top-view to obtain their ground points. Small orange box markers are overlaid on the images of frame 198 for determination of the camera orientations. Note that, in the figures of 'vertical axes', the axis of a severely occluded person does not contribute to localization of the ground point. When occlusion occurs, the ground points being tracked are displaced a little from their correct positions but are restored to the correct positions quickly. Only 5 particles (one particle is a combination of 4 single-person states) was used for robust tracking. Those indoor cameras could be easily placed properly in order to maximize the effectiveness of our multi-view integration and the visibility of the people.

Fig.13(a) depicts the graph of the total distance error of people's tracked ground points to the ground truth points. It shows the advantage of multiple views for tracking of people under severe occlusion.

Fig.13(b) visualizes the homographic top-view images of possible vertical axes. A vertical axis in each indoor image view can range from 1 to each maximum image width. 7 transformed vertical axes for each view are depicted for visualization. It helps to understand how the vertical axis location obtained from segmentation affects ground point (intersection) errors on the top-view. When angular separation is close to 180 degrees (although visibility is maximized), the intersection point of two vertical axes transformed to top-view may not be reliable because a small amount of angular perturbation make the intersection point move dramatically.

The outdoor sequences (3 views, 4 persons) are challenging in that three people are wearing similarly-colored clothes and the illumination conditions change over time, making segmentation difficult. In order to demonstrate the advantage of our approach, single hypothesis (deterministic search only) tracker, general particle filter, and particle filter with deterministic search by segmentation (our proposed method) are compared in Fig.14. The number of particles used is 15.

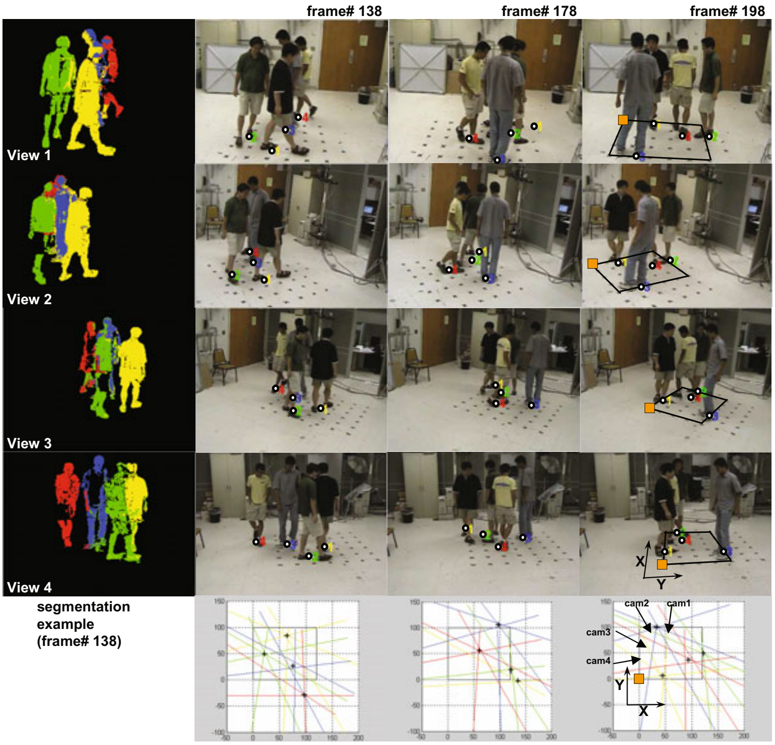
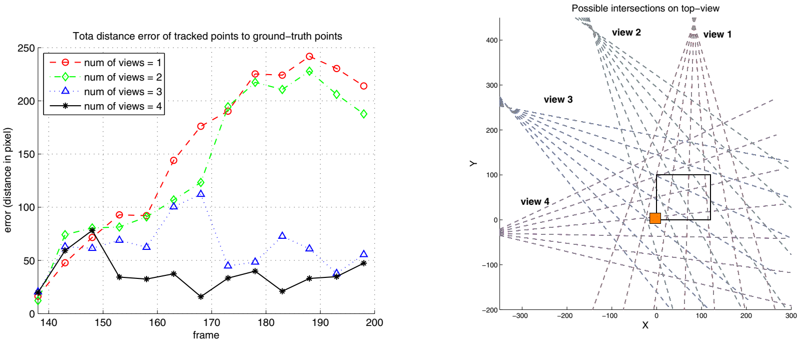


Fig. 12. The tracking results of 4-view indoor sequences from Frame 138 to 198 are shown with the segmentation result of Frame 138



(a) Total distance error of persons' tracked ground points to the ground truth points

(b) Homographic images all different vertical axes

Fig. 13. Graphs for indoor 4 camera views

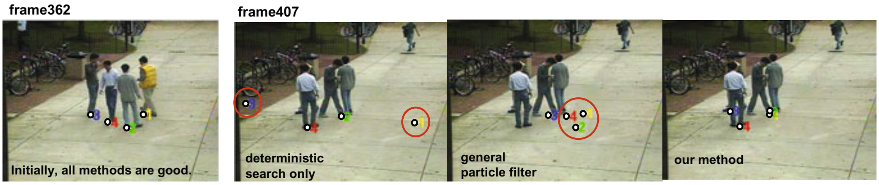


Fig. 14. Comparison on three methods: While the deterministic search with a single hypothesis (persons 2 and 4 are good, cannot recover lost tracks) and the general particle filter (only person 3 is good, insufficient observations during occlusion) fail in tracking all the persons correctly, our proposed method succeeds with a minor error. The view 2 was only shown here. The proposed system tracks the ground positions of people afterwards over nearly 1000 frames.

Conclusions

All the topics described in the book chapter are all closely related and geared toward intelligent video surveillance.

Our adaptive background subtraction algorithm, which is able to model a background from a long training sequence with limited memory, works well on moving backgrounds, illumination changes (using our color distortion measures), and compressed videos having irregular intensity distributions.

We presented a perturbation method for measuring sensitivity of BGS algorithms. PDR analysis has two advantages over the commonly used ROC analysis: (1) It does not depend on knowing foreground distributions, (2) It does not need the presence of foreground targets in the video in order to perform the analysis, while this is required in the ROC analysis. Because of these considerations, PDR analysis provides practical general information about the sensitivity of algorithms applied to a given video scene over a range of parameters and FA-rates.

A framework to segment and track people on a ground plane is presented. The multi-view tracker is extended efficiently to a multi-hypothesis framework (M^3 Tracker) using particle filtering. To tackle with the explosive state space due to multiple targets and views, the iterative segmentation-searching is incorporated with a particle filtering framework. By searching the ground point from segmentation, a set of a few good particles can be identified, resulting in low computational costs.

References

1. Horprasert, T., Harwood, D., Davis, L.S.: A statistical approach for real-time robust background subtraction and shadow detection. In: IEEE Frame-Rate Applications Workshop, Kerkyra, Greece (1999)
2. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Int. Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 246–252 (1999)
3. Harville, M.: A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 543–560. Springer, Heidelberg (2002)

4. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000)
5. Kohonen, T.: Learning vector quantization. Neural Networks 1, 3–16 (1988)
6. Chalidabhongse, T.H., Kim, K., Harwood, D., Davis, L.: A Perturbation Method for Evaluating Background Subtraction Algorithms. In: Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, VS-PETS (2003)
7. Scotti, G., Marcenaro, L., Regazzoni, C.: A S.O.M. based algorithm for video surveillance system parameter optimal selection. In: IEEE Conference on Advanced Video and Signal Based Surveillance (2003)
8. Haritaoglu, I., Harwood, D., Davis, L.S.: W^4: real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 809–830 (2000)
9. Elgammal, A., Davis, L.S.: Probabilistic Framework for Segmenting People Under Occlusion. In: IEEE International Conference on Computer Vision, Vancouver, Canada, July 9-12 (2001)
10. Zhao, T., Nevatia, R.: Tracking Multiple Humans in Complex Situations. IEEE Trans. Pattern Analysis Machine Intell. 26(9) (September 2004)
11. Rabaud, V., Belongie, S.: Counting Crowded Moving Objects. In: IEEE Conf. on Comp. Vis. and Pat. Rec. (2006)
12. Yang, D., Gonzalez-Banos, H., Guibas, L.: Counting People in Crowds with a Real-Time Network of Image Sensors. In: IEEE ICCV (2003)
13. Khan, S.M., Shah, M.: A Multiview Approach to Tracking People in Crowded Scenes Using a Planar Homography Constraint. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 133–146. Springer, Heidelberg (2006)
14. Kang, J., Cohen, I., Medioni, G.: Multi-Views Tracking Within and Across Uncalibrated Camera Streams. In: Proceedings of the ACM SIGMM 2003 Workshop on Video Surveillance (2003)
15. Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking Across Multiple Cameras With Disjoint Views. In: The Ninth IEEE International Conference on Computer Vision, Nice, France (2003)
16. Mittal, A., Davis, L.S.: M_2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene. International Journal of Computer Vision 51(3) (February/March 2003)
17. Eshel, R., Moses, Y.: Homography Based Multiple Camera Detection and Tracking of People in a Dense Crowd. In: Computer Vision and Pattern Recognition, CVPR (2008)
18. Jin, H., Qian, G., Birchfield, D.: Real-Time Multi-View Object Tracking in Mediated Environments. In: ACM Multimedia Modeling Conference (2008)
19. Black, J., Ellis, T.: Multi Camera Image Tracking. In: 2nd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (2001)
20. Xu, M., Orwell, J., Jones, G.A.: Tracking football players with multiple cameras. In: ICIP 2004 (2004)
21. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multi-Camera People Tracking with a Probabilistic Occupancy Map. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2), 267–282 (2008)
22. Tsai, R.Y.: An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision. In: IEEE Conference on Computer Vision and Pattern Recognition (1986)
23. Tu, Z., Zhu, S.-C.: Image segmentation by data-driven Markov chain Monte Carlo. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5), 657–673 (2002)

24. Javed, O., Shafique, K., Shah, M.: Appearance Modeling for Tracking in Multiple Non-overlapping Cameras. In: IEEE CVPR 2005, San Diego, June 20-26 (2005)
25. Senior, A.W.: Tracking with Probabilistic Appearance Models. In: Proceedings ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems, June 1, pp. 48–55 (2002)
26. Chang, T.H., Gong, S., Ong, E.J.: Tracking Multiple People Under Occlusion Using Multiple Cameras. In: BMVC (2000)
27. Perez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 661–675. Springer, Heidelberg (2002)
28. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Real-time foreground-background segmentation using codebook model. Real-Time Imaging 11(3), 172–185 (2005)
29. Hu, M., Lou, J., Hu, W., Tan, T.: Multicamera correspondence based on principal axis of human body. In: International Conference on Image Processing (2004)
30. Kim, K., Davis, L.S.: Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 98–109. Springer, Heidelberg (2006)

Filtering Surveillance Image Streams by Interactive Machine Learning

Cristina Versino[1] and Paolo Lombardi[1,2,*]

[1] European Commission
Joint Research Centre
Nuclear Security Unit,
Via Fermi 2749, 21027 Ispra, Italy
Cristina.versino@jrc.ec.europa.eu,
paolo.lombardi.vision@gmail.com
[2] Computer Vision and Media Lab, DIS,
University of Pavia, via Ferrata 1, 27100 Pavia, Italy

Abstract. As surveillance cameras become widespread, filters are needed to process large streams of image data and spot interesting events. Programmed image filters generally result in low to medium performing solutions. Data-derived filters perform better in that they tap on selected image features, but require a per-sensor effort by an analyst or a machine learning expert. This contribution addresses filter shaping as a data-driven process that is 'placed in the hands of many end-users' with extensive domain knowledge but no expertise in machine learning. The focus is on interactive machine learning technologies as a means to achieve self-programming and specialization of image filters that learn to search images by their content, sequential order, and temporal attributes. We describe and assess the performance of two interactive algorithms designed and implemented for a real case study in process monitoring for nuclear safeguards. Experiments show that interactive machine learning helps detect safeguards relevant event while significantly reducing the number of false positives.

Keywords: Interactive machine learning, Hidden Markov models, Decision trees.

1 Overview

As camera technology becomes widespread (for example surveillance cameras), *filters* are needed to process large streams of image data and spot relevant events as seen from the camera point-of-view in the world.

A-priori programming of image filters generally results in low to medium performing solutions, because the programming attitude is - by definition - generalist, largely assuming, and non-situated. By contrast, *data-derived filters* perform better in that they tap on specific image features, but require a per-sensor effort by an analyst or machine learning (ML) expert whose availability is scarce. To overcome this limitation, we address filter shaping as a data-driven process that is *placed in the hands of many end-users* with extensive domain knowledge but no expertise in ML.

[*] Corresponding author.

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 289–325.
springerlink.com

The contribution arises from a case study in *process monitoring* for nuclear safeguards where a *review tool for streams of surveillance images* is designed and developed. This combines: a scene change detection algorithm, pattern classification in a relevance feedback context, automatic selection of image features, and suitable visualization and interaction loops.

The focus of this contribution is on *interactive machine learning* (IML) technologies as a means to achieve self-programming and specialization of filters that learn to search images by their content, sequential order, and temporal attributes to speed up the detection of relevant images in large surveillance streams. The ultimate objective is to design a self-improving tool to assist nuclear inspectors in image review tasks. In a nutshell, filters learn to detect relevant images based on examples labeled by the nuclear inspectors online, in an iterative way. At each iteration, new image filters are induced and swiftly used to re-classify the image stream for inspection.

For IML to become viable, we acknowledge the need to provide the user with a meaningful experience, given that the review tool changes its behavior over time due to its learning capabilities. Further, a number of practical conditions exist for IML be deployable to everyday users. First, *all ML technicalities have to be 'hidden'*: our target user is neither a scientist nor an expert of ML, nor a person who has the attitude and time to work with ML configuration parameters. Second, the *quality of learning should be manifest to the user* for him/her to decide whether to accept or not the selection of images presented as 'relevant' by the tool. Are these *all* the relevant images s/he needs to see out of a large set? Is the tool capturing what s/he is looking for or is it just over-fitting few example images? Is the learning tool 'confused'? Third, since users label examples online, *computation time becomes a central issue*: it needs to stay within a human acceptable level from one training iteration to the next.

In this Chapter we discuss interactive machine learning from an applicative standpoint. Sections 2, 3 and 4 are introductory: we review state-of-the-art approaches in IML, establish the difference between uninformed and informed image filters, and describe the context of surveillance for nuclear safeguards. Section 5 explains the significance of using ML in this context: it provides continuity of knowledge between image reviews which otherwise lacks. Sections 6 to 8 deal with the filters themselves: Section 6 presents an uninformed filtering technique, whereas Sections 7 and 8 describe the core contributions of this Chapter, namely two IML filters for image streams based on decision trees and Markov models. In Section 9 we present tests conducted on the use of these filters on real safeguards image streams. Section 10 wraps up and concludes.

2 Interactive Machine Learning

Classical machine learning (ML) assumes a sufficiently large set of labeled data is available to estimate the ML model parameters and test its generalization capability. Computation time, due to model inference and testing, is generally a non issue. Predictive accuracy of the model is the main goal of ML.

By contrast, interactive machine learning (IML) applies to situations where few labeled examples become available over time. The training set grows iteratively with a user labeling examples online -mostly to correct wrong classifications or to resolve uncertainties that appear when the model is tested. In this setting, computation time *is* the user's time, and becomes a central issue.

Conceptually, IML approaches may be framed in two groups depending on *who* drives the examples' labeling process: we distinguish between (i) *user centered methods* and (ii) *active learning techniques.*

User centered methods leave entirely to the user the decision as to which data points to label. *The underlying rationale is to take highest advantage of the user's understanding of the domain area from which the data originated.* This also includes the user's expectations on the IML model's performance on a given task. For example, in most surveillance tasks it is *not* necessary that an image review tool classifies as relevant *all* images related to a specific event: it is sufficient that the tool detects a 'large enough' set of images around the event so as to focus the user's attention to that part of the image stream [1]. Although fuzzy, this performance criterion makes sense for humans. However, it hardly translates in a clear cut optimization criterion needed for machine learning, and is very application dependent.

In the case of *active learning*, an algorithm selects data points which should receive a label by the user. This decision is driven, for instance, by the *uncertainty associated to data classification as measured by the algorithm.* For example, data points that fall close to decision boundaries of a classifier are the most uncertain from a classification point of view, and would be presented with priority for the user to label them.

In short, user centered approaches appear to be more domain- and task-related, while active learning techniques are more generic and driven by the idea of maximizing the predictive accuracy of the IML model.

In what follows user centered approaches and active learning are illustrated by selected sample works drawn from the state-of the-art. Though not mutually exclusive, it appears that either one approach or the other is adopted in each IML setting we have surveyed.

2.1 User Centered Methods

User centered techniques originated within the human-computer interaction research community for the design of perceptual user interfaces and interactive sensors as enabling technologies for ubiquitous computing.

As an example, [2, 3, 4] describe 'Crayons', a tool to design user interfaces for camera-based interaction. By this tool, interface designers can turn any surface visible to cameras into interactive widgets (e.g., buttons, sliders). Widgets are then triggered by the presence of hands on their area. To implement the widgets, the core part of the tool is a facility to train a 'skin' classifier by IML. The designer *paints* portions of training images with 'skin' or a 'background' crayons, thus generating training data in few stokes. The labeled data is used to induce automatically a decision tree based classifier [5, 6]. The classifier is then run on test images for the user to decide if the skin detection performance is appropriate. To this goal, the tool interface augments the original images with two semi-transparent layers, one showing the original training data (if any), the other showing the image classification.

Several key aspects of this set-up are recurrent in user centered IML scenarios dealing with image data. *First*, ordinary users have no familiarity with image processing, feature selection, and machine learning to build an image classifier. They do, however, have a good understanding of what it means to locate an object of interest on the image plane. *Second*, to describe the image content, a *generic IML toolkit* needs to embed a *large repertoire of image features* (e.g. color, shape, texture, motion-related) and gradually s*elect those describing the content of interest to the user*. The Crayon prototype includes 175 features per pixel. Although feature selection has a computational cost, it is largely paid off by the fact that testing a compact classifier is *much faster* than running one built on a fixed, large set of (mostly irrelevant) features. Feature selection speeds-up interaction with the user when working with data sets made of thousands of images. *Third*, special care is to be put on the *design of the interface for the user to generate the training examples*. In the case of Crayons, the ‘paint’ action generates hundreds of examples in one go, thus reducing that number of ‘user iterations’ needed to generate an accurate a classifiers. *Fourth*, the authors conclude that having a fast training algorithm is more important than strong induction. By using the Crayons toolkit, interface designers were able to created skin classifiers *in few minutes*, despite no specific knowledge in image processing or machine learning. Though data overfitting may occur, this can be readily perceived by the designer and corrected by the addition of new examples on ‘problem areas’. Hence, *fifth*, a powerful *output interface is needed for the user to decide on the quality* of learning.

Other works [7, 8] on user centered approaches have studied the role of *rich user feedback* as means to improve a classifier’s accuracy over the longer term; researched ways to *incorporate user feedback* into machine learning algorithms besides the generation of additional training examples, and emphasized the value of IML systems that can *explain their results in a user understandable way*.

Finally, [9] explored techniques that let users build directly classifiers (decision trees) by visual means.

2.2 Active Learning

Active learning is based on the premise that the learner has access to a *pool* of unlabeled data and can ask labels for some of these. The issue is to find a strategy that limits the number of label requests while inducing an accurate classifier [10].

A work with good theoretical foundations in pool-based active learning is presented in [11, 12] using support vector machines (SVMs) [13] as classifier. Benchmark results are provided for text classification [11] and image retrieval [12].

For illustrative purposes, a simple linear SVM for binary classification can be visualized as the hyperplane separating (in a high-dimensional feature space) the (projected) ‘relevant’ and ‘irrelevant’ training data by a maximal margin (Fig. 1, the training data are the empty circles and the crosses), the margin of the SVM being the distance between the closest training data points of both classes to the hyperplane. The maximal margin requirement is to optimize the predictive accuracy on test data [13]. Given a training set, SVMs can be induced by established learning techniques [13].

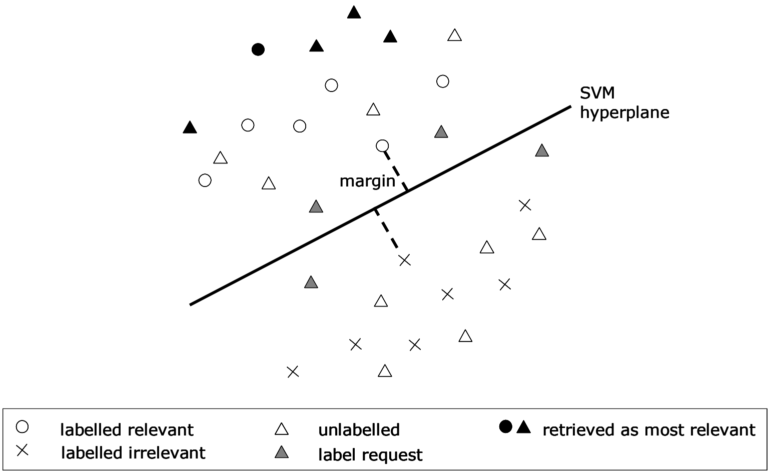


Fig. 1. A linear SVM trained by active learning for information retrieval

The active learning setting requires defining a search strategy to select the next data point from the pool to be labeled by the user (Fig. 1, all triangles). Noting that each labeled instance restricts the space of possible SVMs consistent[1] with the extended training set, the goal is to select a data point which, when labeled, reduces the SVMs solution space as quickly as possible. In view of this, a good search strategy is to choose a data point that halves the set of SVMs solution space [11].

Because it is unpractical to compute the size of the SVM solution spaces obtained by labeling each possible data point as positive or negative, approximations are provided based on the concept of margin [11]. From a computational view, the cheapest one is the *simple margin* rule. It states that, given a SVM, the next points to label are the data points closest to the SVM's hyperplane (Fig. 1, gray triangles). The rule reflects the intuition that the most uncertain predictions fall close to the boundary of the current classifier. Hence, labeling these points mostly helps disambiguating the classification problem.

For a symmetric argument, data points which fall far away from the SVM hyperplane have a higher prediction confidence. This observation leads to a natural implementation of *content-based image retrieval* (see also Section 7) with SVMs in an active learning context [12]. Given a pool of images, the retrieval system is seeded by the user labeling a random set of images. On this, a first SVM is induced and an iterative process starts. At each iteration, the user is asked to label the images closest to the SVM boundary and a new SVM is induced. Given a final

[1] The consistent space is the space of SVMs that separate perfectly the training set extended by the newly labeled data point.

SVM, the images retrieved by the system are those classified ‘relevant’ and falling the farthest away from the SVM hyperplane (Fig. 1, black triangles and black circles).

Experimental results [12] show that the *active* SVM achieves higher search accuracy in image retrieval tasks than traditional query refinement schemes. In particular, it outperforms the *passive* SVM where the examples to be labeled by the user are selected at random.

On the negative side, it has been observed that the performance of SVMs in image retrieval trained by active learning can be affected by several factors [14]: (i) small size training sets may lead to SVMs instabilities; (ii) SVMs are sensitive to the presence of an unbalanced number of positive and negative feedbacks; (iii) overfitting may occur due to an unbalanced number of image features and examples in the training set. To mitigate these problems, *an asymmetric bagging and random subspace technique* has recently been proposed and validated on benchmark image sets [14]. Briefly, bagging addresses the SVM instability problem by creating multiple classifiers then aggregated by majority voting. Each classifier is trained on a balanced number of positive and negative samples due to a procedure which bootstraps only the negatives examples (asymmetric bagging). Finally, this is coupled with a random subspace method, which bootstraps -this time- in feature space to re-balance the number of features with the number of training examples.

In summary, active learning is a lively area of research and very relevant for image retrieval contexts. Still, it appears that the use of these techniques requires a level of expertise in machine learning technologies to set-up a working system that, today, surpasses the ordinary user’s capabilities.

2.3 User Centered Methods or Active Learning?

The groups of techniques presented earlier in this Section appear to be researched as mutually exclusive, but they address the same problem – interactive machine learning (IML).

Interestingly, the two approaches to IML emerged from different communities of researchers – the Human Computer Interaction (HCI) and the Machine Learning (ML) communities. It is not surprising that HCI emphasizes the user’s role in IML, while active ML is driven by the classifier’s point of view on the IML problem.

Noting that it would be valuable to research and combine the two approaches in a single one, the remainder of this Chapter focuses on ML filters for surveillance and presents a user centered approach to IML for the review of surveillance images in a nuclear safeguards context.

3 Filters and Learning Techniques for Surveillance Image Streams

In a typical surveillance scenario, one or more cameras look over a scene and the outgoing image streams are transmitted to a set of displays for online inspection and to storage for offline (delayed) inspection. Cameras are usually installed to view large areas so as to minimize the number of cameras and data streams.

The task of inspecting image streams can require the application of different techniques, very much depending on the nature of the area under surveillance and on the installation pattern of cameras. For example, in a scenario of surveillance of an airport or station where the police is looking for a person wearing known clothes, a human supervisor may screen the images for people wearing colors matching the description. The same approach can be used by an automatic system: a filter may select images containing blobs of the given colors describing a human shape.

This is what we call an *informed search*: the supervisor (human or image filter) checks the presence of data that is known to correspond to a meaningful event. This technique can be used only when the supervisor possesses a precise expectation on the target of her search. The more precise the information, the better the selection of meaningful image streams segments.

When the supervisor possesses only general information on the type of events, *uninformed search* techniques must be employed. In the same example as above, if the police are not searching any specific person, attention may be focused to violation of prohibited areas or counter-crowd movements. These events are not always meaningful, but show at least a correlation with events of interest.

These concepts of informed and uninformed search can be applied also to image streams review tools. In short, informed search techniques consist in matching a model of a significant event with the image data. They require that the image review software incorporates such a model. They are typical of a top-down approach, i.e. they start from knowledge and expectations and verify these on the data. Uninformed search techniques are based on the detection of features that show a correlation with significant events. Uninformed search is by nature a bottom-up approach, starting from evidence of stimuli to build knowledge.

In the literature, uninformed techniques are also known as *novelty detection* techniques [15, 16]. Novelty detection is the identification of new signals that are different from a reference. Two fundamental components in these systems are: (i) the *reference signal* to which new data are compared, and (ii) the *conditional test* that triggers the identification of 'novel' data. The reference signal is an image feature that can be *learned* from images assumed to be 'normal'. In video surveillance, the most common form of uniformed technique is background maintenance for motion detection [17]. In this context, machine learning has been proposed to bootstrap and maintain the background model. Methods being applied include statistical models [18], kernel-based classifiers [19], Wiener filters [20], Kalman filter [21], and mean shift [22]. In Section 6 we illustrate the state-of-the-art filter used today in nuclear safeguards, which relies on simple form of background model.

Informed techniques in video analysis are numerous; those pertinent to our discussion fall in two categories: (i) behavior analysis and (ii) video annotation. Learning for video surveillance has been focused prevalently around *behavior analysis*. The goal is to learn patterns of 'normal' behavior in surveillance video and to alert guards if an abnormal situation is detected. The concept is that trajectories of tracked objects are selectively clustered and labeled by machine learning to add semantics. Methods of representation and clustering include polynomial fitting, multi-resolution quantization, Hidden Markov Models, kernel methods, neural networks, and k-means [23]. The currently accepted reference model for these systems has been introduced in [24]: machine learning algorithms build a graph representation of an image in which nodes represent points of interest (e.g. entry/exit points, stop points) and arcs represent activity paths (e.g. motion, change of activity). Groupings of points of interest and activity paths can be labeled for higher-level semantics (e.g. “car enters”, “car moves” and “car stops” may be labeled as “parking”).

To our knowledge, all behavior-analysis systems assume the availability of *motion tracking data*, typically extracted with the uninformed techniques reviewed above [25, 26]. This in turn requires either a high visual correlation (i.e. similar appearance) or a high temporal correlation (i.e., reasonable frame rate) of object occurrences in consecutive images. Let us defer a deeper discussion on this to Section 8. Here we want to underline that current research on behavior analysis is focused on real-time, video-rate image streams. Review and search of image streams stored at a far lower frame rate (e.g. one image per minute) because of, for instance, scarce storage capabilities or features dictated by industrial process, cannot be treated with the same methods.

Video annotation is rarely used in video surveillance; however the filters we discuss in Sections 7 and 8 are highly related to this topic. A typical learning-based video annotation system takes a video segmented into short units and extracts low-level features from each unit to describe its content. Given a concept in a set of predefined concepts, each unit is then manually annotated to be ‘positive’ or ‘negative’ according to whether it is associated with this concept [27]. Examples abound in the literature, spanning rule-based systems [28], multi-cue statistical learning [29, 30], support vector machines [27], graph matching [31], among the others. We are interested in cases where the standard video segmentation techniques are not applicable. In these cases standard learning algorithms must be adapted accordingly, and interactivity must be enhanced to make profitable use of supervised training.

In later Sections we discuss the application of uninformed and informed filters using interactive machine learning similar to that developed for video annotation to the field of surveillance image streams analysis, with novel contributions especially focused on the application setting of nuclear safeguards.

4 The Nuclear Safeguards Surveillance Context

Nuclear safeguards verify that a State’s nuclear material is not diverted to build weapons or explosive devices. In the European Union, more than 1000 nuclear sites are verified by about 200 inspectors of the EURATOM safeguards authority.

Camera surveillance in nuclear facilities helps to attain safeguards at a reasonable cost without interfering with a facility’s operations (Fig. 2).

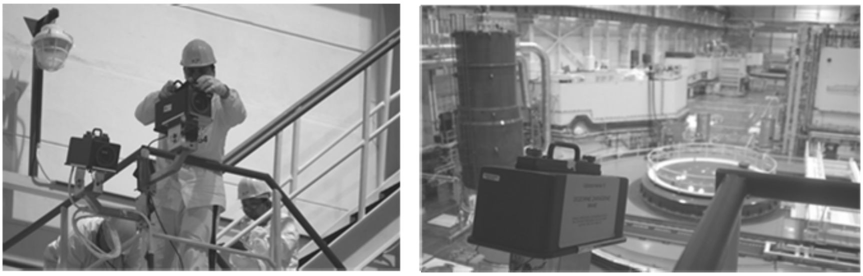
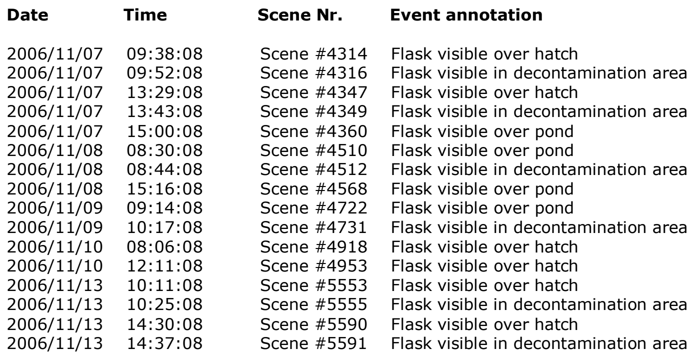


Fig. 2. Inspectors setting-up cameras in a nuclear plant (left). A camera's view (right) (© D. Calma/IAEA).

The specific scenario of nuclear plants poses several challenges to state-of-the-art image filters. *First*, the field of view covers all the locations where important processing takes place (Fig. 2, right). These locations are many meters apart, thus the appearance of a *flask* of nuclear material, which is the object of interest in safeguards image reviews, significantly changes during the process. *Second*, the image acquisition rate is very low -one frame every several minutes. This frame rate is designed to match the pace of the process to be supervised. It also guarantees that all acquired images can be stored on the camera's local memory, a need stemming from the fact that most cameras operate on-site as stand-alone systems. *Third*, the flask is visible only in a small subset of images. Typically, each installed surveillance camera acquires several thousands of images (10 to 100 thousands) before these need be reviewed[2]. Of these images, less than 1% are expected to relate to safeguards-relevant events. The remaining either present no change between consecutive frames or contain events not involving flask movements (e.g. moving cranes, trolleys, illumination changes).



Date	Time	Scene Nr.	Event annotation
2006/11/07	09:38:08	Scene #4314	Flask visible over hatch
2006/11/07	09:52:08	Scene #4316	Flask visible in decontamination area
2006/11/07	13:29:08	Scene #4347	Flask visible over hatch
2006/11/07	13:43:08	Scene #4349	Flask visible in decontamination area
2006/11/07	15:00:08	Scene #4360	Flask visible over pond
2006/11/08	08:30:08	Scene #4510	Flask visible over pond
2006/11/08	08:44:08	Scene #4512	Flask visible in decontamination area
2006/11/08	15:16:08	Scene #4568	Flask visible over pond
2006/11/09	09:14:08	Scene #4722	Flask visible over pond
2006/11/09	10:17:08	Scene #4731	Flask visible in decontamination area
2006/11/10	08:06:08	Scene #4918	Flask visible over hatch
2006/11/10	12:11:08	Scene #4953	Flask visible over hatch
2006/11/13	10:11:08	Scene #5553	Flask visible over hatch
2006/11/13	10:25:08	Scene #5555	Flask visible in decontamination area
2006/11/13	14:30:08	Scene #5590	Flask visible over hatch
2006/11/13	14:37:08	Scene #5591	Flask visible in decontamination area

Fig. 3. Report resulting from an image review

[2] The frequency of the reviews is determined for each specific plant to ensure that these are *timely*, i.e. they allow the verification that no abrupt diversion of nuclear material has occurred in the plant.

Inspectors eliminate the no-change images by applying an uninformed scene change detection filter [32] (Section 6). This operation may reduce the image set to 10% of the original size. The latter, reduced set is reviewed by inspectors on a frame-by-frame basis, and safeguards-relevant images are annotated to produce a *review report* (Fig. 3), i.e. a list of time-stamped, chronologically ordered images, each one labeled by the class of the event recognized by the inspector.

5 Combining Uninformed and Informed Search Strategies

In this context, we have developed tools based on interactive machine learning, collectively named 'Safeguards Review Station' (SRS), to assist nuclear inspectors in the review of surveillance images. The approach followed is that *SRS tools learn to improve their detection performance by tapping on information available both from past reviews and from the on-line compilation of a review report.*

Figure 4 highlights how the state-of-the-art review flow is augmented in the SRS. In the state-of-the-art review flow (block arrows), image data is generated over time continuously, but it is broken down in batches to ensure the timeliness of reviews. Given a batch of images, a filter is applied to extract *events*. This step is in large part automatic and relies on scene change detection (SCD). The *main role of the inspector is in the annotation of the SCD events*. As noted in Section 4, many of these SCD events turn out to be false positives. The few safeguards-relevant events are labeled by the event class and become part of the review report (Fig. 3).

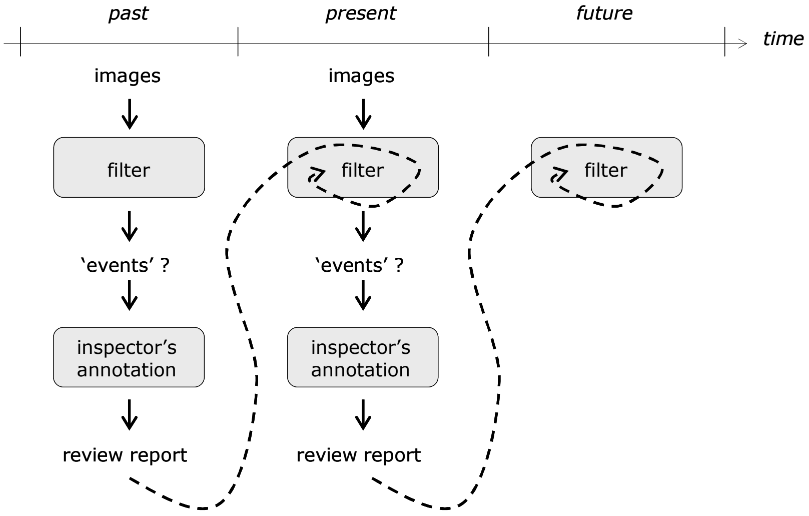


Fig. 4. The state-of-the-art review flow (solid arrows) and the augmented flow in the Safeguards Review Station (dashed arrows)

In a traditional review flow, when a new batch of images becomes available, the same review process is *repeated unchanged*, this meaning that the SCD filter which assists the inspector in the detection of events *does not learn from the results of past reviews*.

By contrast, the SRS concept is to 're-connect' the sequence of reviews *because the stream of images to be reviewed over time is generated by the same nuclear process*. By adopting this view, *we design filters that take advantage of past results to support the present and future reviews*.

For the SRS, a departure point is the archive of review reports produced by inspectors over time for a given plant and camera view. Table 1 lists plant- and camera-specific information that can be derived from these reports. Besides highlighting the typical *classes of safeguards-relevant events* annotated by inspectors (P1), they establish an association between these classes of events and their *visual appearance* exemplified by the corresponding time-stamped image files (P2). The annotated *sequence of events* provides information on the stages of the processing of flasks of material within the plant (P3), and the event time-stamps give an indication on the *duration* of each stage of the processing (P4).

Table 1. Plant properties that can be derived from review reports

Property ID	Property description
P1	Classes of typical events taking place in the plant as per the events' annotations (e.g. 'flask over hatch', 'flask in decontamination area', 'flask over pond').
P2	Examples of the visual appearance of safeguards-relevant events as seen from a specific camera (by retrieving the corresponding image files).
P3	Sequence of the events (e.g. a 'hatch' event is followed by a 'decontamination' event, then by a 'pond' event).
P4	Duration of events (by computing the time interval between events).

On these properties, we can build *informed filters* to detect the typical classes of safeguards-relevant events with higher precision: possibly all relevant events with less false positives.

Specifically, the SRS includes two novel filters, each used in cascade to a state-of-the-art SCD filter. The novel filters are based on: (i) decision trees (DT) and (ii) Markov models (MM). Filter DT, described in Section 7, relies on the *visual appearance of events* (properties P1 and P2 in Table 1). By contrast MM (Section 8) performs a 'meta-classification' of the *sequence of events extracted* by SCD, and hence it is trained on statistics about P1, P2 and P4 derived from past review reports.

6 Searching Image Streams by Scene Change Detection

Uninformed event detection arises from monitoring some quantities derived from image streams over time. An event is declared if a quantity alters its value with respect to a given condition. The conditional test triggering the event can be as simple as an individual threshold, to as complex as frequency component analysis or case-based reasoning [15, 18, 19]. The crucial point is that the parameters of the conditional check *do not relate to information specific to the event type being searched*. Rather, they are set a priori so as to trigger the detection for a broader family of events that contains the target event.

As an example of uninformed technique, we illustrate the scene change detection (SCD) filter in use in official nuclear inspection software [32]. The technique is a *two-frame differencing* based on the average intensity value of pixels inside one or more area of interest (AOI). Before starting the algorithm, the user draws the AOIs on a reference image (Fig 5) around image regions including locations of interest. For an image at a given time, SCD computes the average intensity of pixels belonging to an AOI. This is compared to the same value computed on the previous image, so as to derive a measure of the relative change. If the change in intensity breaks a threshold, an SCD event is marked for that image and that AOI. The same computation is repeated separately for every AOI, possibly with different thresholds. As a result, for each image the SCD filter may detect a number of events up to the number of AOIs defined by the inspector. By altering the thresholds, the user can increase or decrease the number of events detected by SCD.

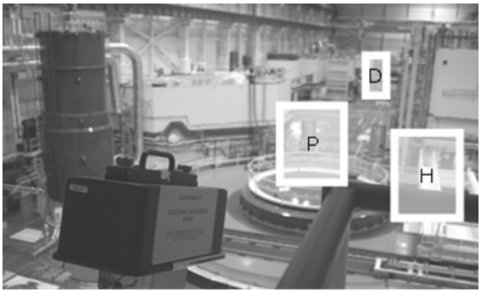


Fig. 5. AOIs are drawn on the camera's image plane

Although they may change in some installations, there are three locations of interest in a typical nuclear plant that are usually assigned an AOI:

1. the *hatch* (H),
2. the *decontamination area* (D),
3. the *pond* (P).

In a normal operation process (Fig. 6) a flask of nuclear fuel enters the hatch and reaches the decontamination area, whence it is moved to the pond. From the pond, the flask moves back to the decontamination area and then exits the scene through the hatch. In SCD, each AOI can be assigned to a label. In plants with the three locations listed above, the label set is {h, d, p}. All images for which a motion event has been detected are associated with one or more of these labels. For instance, if an image *t* is labeled [h, p], this means that the AOIs of hatch and pond exhibited sufficient change from *t*-1 to *t*. Later in this Chapter, Figure 8 shows an example of how the output is communicated in the user interface. Inspectors browse the labeled batch and annotate only the relevant images with the appropriate event class.

Due to the loose correlation between the target event and the signal being monitored, filters that use uninformed techniques like SCD typically are employed in the first stage of an image stream analysis to eliminate obvious false negatives. In general, the parameters that regulate the conditional test must be set as to attain a true positive rate of 100%. As a trade-off, the number of false positives is generally very high.

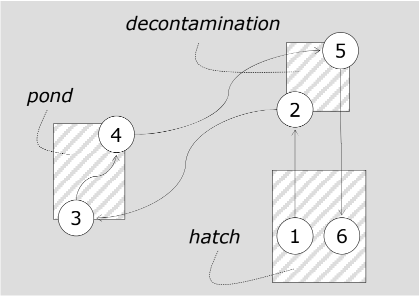


Fig. 6. Schematic flow of the movements of a flask of nuclear material within a plant

In our studies of image surveillance for nuclear safeguards, most parts of the image sequences contain little or no activity. Sequences of about 20,000 images must be reviewed offline by nuclear inspectors operating under time pressure. The use of SCD filter can reduce the amount of images to be inspected on average by 90%, from 20,000 to 2,000. Reviewing 2000 images one-by-one is still a hard job.

Actually inspectors are interested only in motion events that include the movement of nuclear materials, and not in movements due to persons or auxiliary machines. Typically, out of 2,000 images associated with SCD, only 3-30 images are meaningful events. Thus the rate of false positives of this technique is generally close to 99%. It is then reasonable to follow SCD with informed techniques that contain a model of the meaningful events to highlight SCD events of potential interest.

Why do inspectors employ this uninformed SCD? The answer has to do mainly with the degree of *programmability* offered by uninformed techniques. Nuclear inspectors, like most users of image review software for surveillance, are not expert of image processing or automatic pattern recognition. On the other hand, they do understand event detection by setting a threshold. This *understanding* and the total *degree of control* given by changing one or two simple parameters give the inspector confidence of mastering the tool.

7 Searching Image Streams by Their Content

The image review technique described in this Section is based on the premise that image *content* can be described in a *suitable feature space and be searched on.* Typically, this is a vector space on which the images are projected, and where the relevant images are expected to be discriminated from the irrelevant ones by an appropriate classifier. A central issue in search by content techniques is then to *find image representations that enhance image data relevant to the search and reduce the remaining aspects* [33].

7.1 Image Features

Sophisticated search by content systems use image processing techniques to first segment the image in regions of interest and extract objects silhouettes [34]. Then, for each region, a set of features is computed. Features can be related to color, texture, and shape, and they can be binary-valued, discrete or continuous. Finally, images are stored in a database as vectors of features for each region of interest. Other systems simply compute features on the global image such as the mean intensity value or the color distribution histogram [35].

A popular example of search by content technique is *image retrieval by similarity* [33, 34, 35, 36, 37]. The core idea is that the user may query a database for visually similar images by pointing an example image together with a set of features to describe the image at hand. In a simple *nearest-neighbor* approach [38], the image retrieval program computes for each image in the database its distance to the example in the selected feature space. Images within a small distance to the example are considered similar to it, and retrieved for inspection by the user.

A weak point in this procedure is that the result of the query depends strongly on the 'right' choice of features, a step which requires expert knowledge in the definition of features. This knowledge is usually not available to the end-user nor of interest to her.

Not only that. The 'right' features are relative to the specific image set on which the query is run. As an example, suppose that the images of interest to the user show a red can, and that this happens to be the only red object in the database. In this case, a 'red color' feature is optimal to represent and search the images. By contrast in a database with many different red objects, searching by the 'red feature' would retrieve many irrelevant images.

7.2 Image Classifiers

Classification techniques also depend on a number of parameters that make them non user-friendly. A classifier can be built according to a number of construction options that are specific to the classification technique at hand and that must be somehow decided. For instance, if the intended classifier is a neural network [39], one must specify the network architecture (e.g. number of neurons and how they relate to each other).

As pointed out in Section 2.1, for search by content techniques to become work tools for everyday users, it is important to 'hide' to the largest possible extent all parameters on which these techniques depend. This is the purpose of *relevance feedback* mechanisms designed for user centered interactive machine learning (IML) [3, 4, 7, 8, 9]. The underlying principle is that the user should concentrate on issues where she is expert and 'ignore the rest'. In our case study, this means that inspectors should concentrate on the labeling of example images (say which are safeguards-relevant and which are irrelevant, their core expertise) and ignore how the image content is described, or how the classifier is built.

In Sections 7.3 and 7.4 we describe DT, a tool that implements a user centered relevance feedback mechanism for the classification of safeguards-relevant images. DT is based on image descriptors commonly used by image retrieval software and uses *decision trees* [5] as classifier. Both the image descriptors and the decision tree that performs the classification are selected automatically and remain hidden to the end-user. The only input required from her is the labeling of the example images (the *training set*) that occurs in an iterative way, implicitly guiding the tool to the discovery of suitable images descriptors and decision tree parameters.

Briefly, a decision tree is a classifier that takes as input the vector of features describing an image and outputs a 'relevant/irrelevant' decision (Fig. 7).

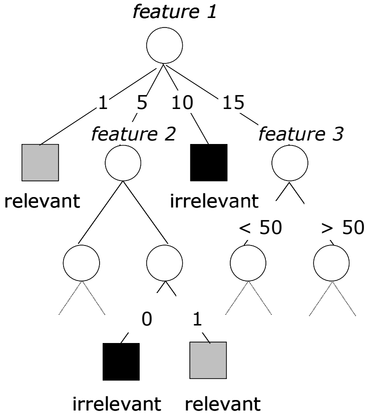


Fig. 7. A decision tree

Each node in the tree is labeled by a feature; arches are labeled by possible values of features or tests, and leaves by 'relevant/irrelevant' decisions. Given an input image described by a feature vector, the image is classified by navigating the tree top-down from the root until a leaf is reached.

The decision tree itself is induced on the basis of a training set of examples by the C4.5 algorithm [5] or its extension See5 [6]. To build a tree, these algorithms tests different features in order of decreasing importance. A feature is considered important if, by using it, one is able to discriminate most relevant and irrelevant examples of the training set. Using this heuristic, the decision tree is built recursively. First, the most important feature is selected. Then, a new decision tree building problem is addressed with one feature less and some examples less -those examples that are classified by the first feature are not considered anymore.

Why decision trees? The motivation for using them in our application is manifold and reflects issues highlighted in Section 2.1 on user centered methods for interactive machine learning.

First, decision trees deal in a natural way with features of different nature: binary, discrete-valued and continuous. This provides flexibility in the choice of the initial repertoire of image features implemented in an image review tool which should work well for many different environments. On the contrary, other families of classifiers are targeted to either discrete-valued or continuous features. For example, neural networks are particularly suited for the processing of continuous variables.

Table 2. Coarse comparison of advantages and drawback of different classifiers

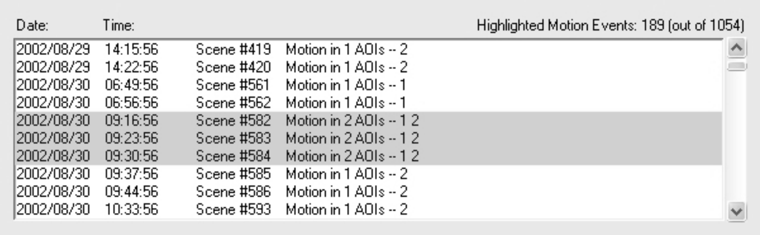
Classifier	Evaluation	Comment
neural nets	+	map complex patterns (non-linear relationships)
	+	learn incrementally
	+/-	mostly suited to deal with continuous data
	-	learning depends on parameters
decision trees	+	learning is parameter –free
	+	mix naturally binary, discrete, continous-valued data
	-	do not learn incrementally
nearest-neighbor	+	very simple learning algorithm
	+/-	mostly suited to deal with continuous data
	-	learning depends on parameters
	-	very sensitive to the presence of irrelevant features
naïve Bayes	+	very simple learning algorithm
	+	learning is parameter –free
	-	features need to be independent from each other

Second, the learning algorithms that are used to induce decision trees come close to the ‘parameter-free ideal’. Basically, one needs to have a set of positive and negative examples on which to run the induction algorithm. In other words, the result little depends on settings of other parameters as it is needed for other classifiers. For example, in a nearest-neighbor classifier one has to decide the number of centroids; in a neural network the number of neurons, the network architecture, the learning rate, etc. Because our aim is to have the IML module working in a ‘silent way’, we should avoid parameters setting as much as possible. Table 2 summarizes advantages and drawbacks of different classifiers *at a very coarse level*. The Table highlights that, when choosing a classifier, one has to consider a number of dimensions and tradeoffs.

Third, the algorithm that induces the decision trees also performs feature selection. The resulting classifier will thus be *compact and fast* to be run in the testing phase, a property required for online learning. An additional standard argument in favor of feature selection is that it 'overcomes the curse of dimensionality' related to an unbalanced number of features and training examples, which is typical of IML settings[3]. Feature selection is then meant to increase the predictive accuracy of the classifier. On this, it must be reported that, quite surprisingly, recent benchmarks revealed that feature selection is seldom needed to enhance the prediction accuracy of classifiers [40], and indicate other reasons to limit the number of features, such as computational and storage requirements.

7.3 The DT Image Review Tool

The DT tool is designed to filter scene change detection (SCD) events by decision tree classifiers. When an image triggers SCD, it is submitted for further opinion to the decision trees. If these classify the image as relevant (this time based on an informed analysis of its content), then the corresponding SCD event is 'highlighted' in the DT interface, i.e. the image is recommended for review by the inspector. In this way, the initial list of SCD events is profiled by DT to a reduced list of highlights. In the example shown in Figure 8, DT has highlighted 189 events out of the original 1,054 SCD events triggered on an image set of 16,000 images.



```
Date:        Time:                                      Highlighted Motion Events: 189 (out of 1054)
2002/08/29   14:15:56    Scene #419   Motion in 1 AOIs -- 2
2002/08/29   14:22:56    Scene #420   Motion in 1 AOIs -- 2
2002/08/30   06:49:56    Scene #561   Motion in 1 AOIs -- 1
2002/08/30   06:56:56    Scene #562   Motion in 1 AOIs -- 1
2002/08/30   09:16:56    Scene #582   Motion in 2 AOIs -- 1 2
2002/08/30   09:23:56    Scene #583   Motion in 2 AOIs -- 1 2
2002/08/30   09:30:56    Scene #584   Motion in 2 AOIs -- 1 2
2002/08/30   09:37:56    Scene #585   Motion in 1 AOIs -- 2
2002/08/30   09:44:56    Scene #586   Motion in 1 AOIs -- 2
2002/08/30   10:33:56    Scene #593   Motion in 1 AOIs -- 2
```

Fig. 8. SCD ('Motion') events hightlighted by DT

Since SCD events may be triggered by changes occurring in different areas of interest (AOIs), as explained Section 6, DT can associate a decision tree to each AOI defined by the inspector for SCD. For example, the list of highlighted events in Figure 8 results from the activity of two trees associated, respectively, to AOI nr. 1 defined over the hatch area and AOI nr. 2 defined over the pond. A SCD event is highlighted if *at least one* decision tree classifies the image as relevant. A tree takes its decision based only on the image part covered by the AOI –i.e. DT computes image features within each AOI. In this way, each tree specializes in the detection of a specific safeguards-relevant event as it is perceived by the camera in the corresponding AOI.

[3] As noted in Section 2.2, this motivated part of the work presented in [14] for active learning with support vectors machine.

7.4 Training Decision Trees with DT

The general idea of user centered IML is to let the user label in an *iterative and interactive* way a number of example images (training set) on which a classifier is induced automatically. Initially, the she labels just few relevant and irrelevant images by browsing the image stream. On this basis, a first classifier is built automatically and run to classify the images in the database. The result of the classification is presented to the user for relevance feedback: for as many images as she wants, she can point out correct and wrong classifications. In this way, the initial training set grows and a more refined classifier is constructed on the basis of this extended examples' set. Again, the new and improved classifier is used to reclassify the database of images and the result of the classification is displayed for user feedback. The convergence point of the process is a classifier that retrieves images of interest to the user with some false positives.

For the tree training phase, DT needs a review report to have been compiled for an image set, and that this contains a number of events of the type for which the inspector would like to train a tree. In other words, the training session happens *off* the review context. As described in Section 5, the trained trees will then become available to filter scene change detection events at *next* review time (cfr. Fig. 4).

Assume that a review report has become available on a given image set to train decision trees for two selected AOIs. Typically, these would be the hatch and the pond, because the events taking place in these areas are the most significant from a safeguards standpoint.

The user interface for the training is made of several panels (Fig. 9). In what follows, we illustrate the information displayed in these panels at some mid point during the interactive training session.

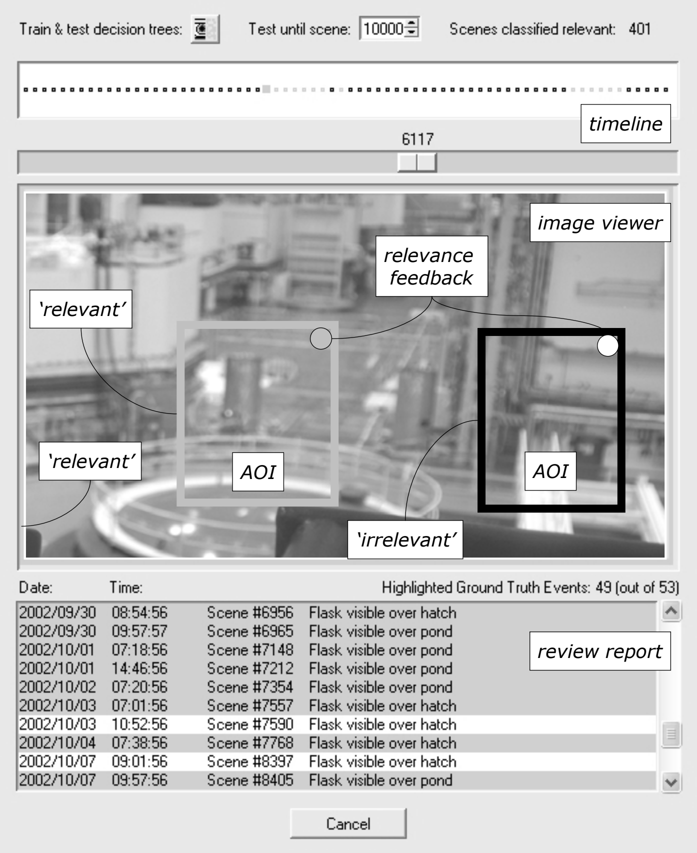
The bottom panel shows the *review report,* i.e the ground truth on the whole image set. It lists and counts the events in the report -in this case 53 in total.

The mid panel is a *viewer* to display the images in the set, focusing on one image at a time. It also shows the AOIs.

The top panel is a *timeline*, where each square represents an image in the set. During training, a function of the timeline is to provide an overview on the classification of a part of the image stream: gray squares represent images classified ‘relevant’ by (at least) a decision tree, black squares indicate images classified as ‘irrelevant’ by both trees.

The same color convention is adopted within the other two panels. When an image is displayed on the viewer, the surrounding frame shows its classification by the color (i.e., gray for relevant, black for irrelevant). The frames of the AOIs are also colored to reflect how each decision tree has classified the image at hand. In this way, the user knows which tree voted the image as relevant, if any. On the review report panel, events classified as relevant by the trees are highlighted by a gray background. A counter indicates the number of ‘highlighted ground truth events’, i.e. how many report events are currently correctly classified by the trees as relevant out of the total true events. Note that the ratio between these numbers is the *recall* performance index used in *information retrieval* ([33], see Section 9).

The information displayed in the three panels is inter-connected. Clicking on an event of the review report loads the corresponding image on the viewer and scrolls the timeline to make visible a segment of the image set which includes the event. Any other image in the set can be viewed either by the scrollbar on top of the viewer or by moving the mouse over the timeline, the current image being indicated by a bold square. This latter navigation modality provides a very fast way to browse the classification of individual images and to support the *relevance feedback* process.



Train & test decision trees:

Test until scene: 10000

Scenes classified relevant: 401

Highlighted Ground Truth Events: 49 (out of 53)

Date:	Time:		
2002/09/30	08:54:56	Scene #6956	Flask visible over hatch
2002/09/30	09:57:57	Scene #6965	Flask visible over pond
2002/10/01	07:18:56	Scene #7148	Flask visible over pond
2002/10/01	14:46:56	Scene #7212	Flask visible over pond
2002/10/02	07:20:56	Scene #7354	Flask visible over pond
2002/10/03	07:01:56	Scene #7557	Flask visible over hatch
2002/10/03	10:52:56	Scene #7590	Flask visible over hatch
2002/10/04	07:38:56	Scene #7768	Flask visible over hatch
2002/10/07	09:01:56	Scene #8397	Flask visible over hatch
2002/10/07	09:57:56	Scene #8405	Flask visible over pond

Fig. 9. DT interface to train decision trees

The 'input device' to provide feedback is the circle shown in each AOI's top-right corner. By default, its color is 'white', meaning that no information has been provided by the inspector on the target classification for the image. Clicking on the circle by <mouse-left> and <mouse-right> changes the feedback from neutral (white) to relevant (gray) and irrelevant (black), respectively. In this way, DT collects examples of relevant and irrelevant images associated to each AOI. These are used to train the associated decision trees.

Finally, the user can control the number of images used to test the classification performance of the decision trees. This is to support in an effective way the iterative nature of training. Note that the cost of each cycle is dominated by the number of images used to test, and not by tree induction algorithm given that the number of training examples remains limited. At early iterations, the parameter ‘Test until scene’ must be set to a few thousands (say, 2000). This is because the first trees induced on the basis of very few examples are not expected to be accurate, but are used to bootstrap the example labeling process on the ‘classification problem areas’ that appear evident on the timeline (Fig. 10). At later iterations, increasing the number of images for testing gives the user a feeling on the generalization capability of the trees, i.e. their ability to classify correctly images that are not part of the training set, a sort of cross-validation. An indication on the quality of learning is also given by the number of ‘scenes classified relevant’ displayed on the top-right corner. The ratio between the ‘highlighted ground truth events’ and the number of scenes classified relevant is the *precision* performance index used in information retrieval ([33], see Section 9). The general goal of learning is to maximize both *recall* and *precision* on both the training and the test set. However, for surveillance applications, the *distribution* of the detected events also matters. As noted in Section 2, in surveillance it is not necessary to classify as relevant *all* images related an event, but a sufficiently large number for the event to ‘appear’ on the timeline (Fig. 9).

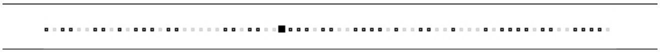


Fig. 10. A ‘confused’ timeline is a sign of bad classification

The ‘rules of thumb’ for the inspector to select the training examples are as follows.

Train one tree at a time. DT gives the possibility to train one tree at a time by excluding the other AOIs. By doing this, the classification shown on the timeline and on the other panels just reflects the activity of one tree. This can simplify the user’s understanding on the quality of the learning.

Separate well training and test examples. Generally an image set contains various cycles of the processing of flasks of nuclear material. It is recommended to focus the training examples on a consecutive number of cycles and leave the remaining cycles ‘untouched’ for testing. This approach makes sense in our context, because the process supervised by the cameras is *stationary*, does not evolve over time. This is generally true for industrial processes whose nature is essentially repetitive.

Focus relevance feedback on ‘classification problem areas’. A strong point of IML is that the user can provide training examples where these are *needed*. This is in opposition to traditional machine learning (ML) where a large batch of training examples is compiled and mined all together. With IML, the user naturally focuses her feedback on misclassifications, exactly providing the information that the classifier has missed at the previous iteration.

Balance positive and negative examples. In our application context the number of irrelevant images largely dominates the number of the relevant ones. During interactive training, attention must be paid to balance positive and negative examples in the training set to avoid trees degenerate to the 'all-is-irrelevant' classifier. To counteract this, a first strategy is to increase the number of positive examples by labeling as relevant images *correlated* to safeguards-relevant events. For example, the event 'flask over hatch' is correlated with images showing a crane's cables pull out the flask from the hatch. Temporally, these images just precede the hatch entry event. By labeling the cables images as relevant, the user creates a stronger detector for the hatch event. Visually, the effect is that one sees on the timeline a longer segment of images classified as relevant, a clear signature for a hatch event. A second possibility (which is hard-coded in DT) is to induce decision trees by *differential misclassification costs* and not by the total number of classification errors. In this framework, the cost associated with a classification error depends on the predicted and true class of the misclassified example. Because missing a safeguards-relevant is a more severe error that classifying as relevant an irrelevant image, DT sets a much higher cost for the misclassification of the examples marked relevant in the training set.

A general remark to the IML approach is that this process is worth if: (i) it converges after a reasonable number of feedback iterations; (ii) the gain obtained by filtering the scene change detection events by the trees is significant; (iii) the filter can be re-used over time without requiring re-training. On (i), it can be reported that in our benchmarks on safeguards images the training sets contained at most 100 examples. This seems to be a reasonable effort for the user to pay given the *workload reductio*n provided by the DT: this is illustrated by the benchmark presented in Section 9. Concerning the re-usability of DT filters over time, this is favored by the fact that nuclear plants need to operate under stable conditions. In case of a degrading review performance, DT foresees the possibility re-train the decision trees. The original training set is extended by new examples defined on a new image set and review report. Although decision trees do not learn incrementally, the cost of re-inducing the trees on an extended training set is negligible for training sets of limited size like those used in IML. As a final remark on the use of IML versus ML, we can report that we ran a classical ML training phase by batch learning over the review report and randomly selected irrelevant images. The retrieval performance was inferior to that of IML, probably because by such a procedure one misses the opportunity of balancing the positive and negative examples in an intelligent and 'human-controlled' way by indicating as relevant images correlated to the safeguards relevant events.

8 Searching Image Streams by Sequence and Time Attributes

Video *content* analysis is by far the most explored topic in image streams analysis. Conversely the *temporal structure* of image streams is generally more the focus of compression algorithms. These algorithms exploit the high correlation between a frame and its successors/predecessors without depending of models of event sequences and timings (uninformed techniques) [41].

An informed approach is applicable when image streams are expected to contain *regular motion or event patterns*. For instance, in video-based human-machine interaction a user's hands or face movements are checked for model gestures that correspond to commands [42]. Similar regular patterns can be observed in surveillance applications in some constrained scenarios, like for instance home and ambient intelligence applications [43]. Also process monitoring in a manufacturing plant can benefit from model-based temporal analysis. Manufacturing processes follow standard paths, and deviations can indicate errors and anticipate defects in the final product.

To build the *model* underlying an informed technique for temporal structure analysis, image streams must exhibit at least one of the following properties:

1. traceable objects;
2. regular sequence of events;
3. regular timing of events.

The first property has to do with *continuity of information*, whereas the second and third properties have to do with *recurrence of information*. The presence of one of these properties alone is sufficient to enable some kind of analysis, independently of the other properties.

The case of *traceable objects* corresponds to situations where object movements can be tracked. This is the assumption underlying gesture recognition and most surveillance and behavior analysis systems (e.g. tracking people and animals). Objects are if they exhibit a video signature with *high self-correlation* and low cross-correlation with other objects or the background. In turn, this requires either highly characterized visual features (e.g. color, motion, shape, texture, etc) – for visual self-correlation – or a high acquisition *frame rate* of the system cameras – for temporal self-correlation –, or, in most scenarios, both elements.

Image streams with *regular sequences of events* contain events that happen with regularity one after the other. Consider the example of monitoring a manufacturing process where the product should be visible by surveillance cameras after each processing inside machines: if it does not appear after a step, there has clearly been a malfunction.

Similarly *regular timing of events* may constitute *per se* a valid model, even without conserving visual appearance or logical sequence. These situations present regular *durations* and/or *intervals* between events. An example is monitoring abandoned objects in metro stations: if an image region is occupied by a foreground object for more than the expected 'regular' time, the software detects an event [44]. To use durations, frames must be accurately time stamped especially when multiple cameras are involved.

The first case, the one of traceable objects, is by far the most commonly studied in computer vision [25, 45]. When the conditions of traceability are not met, video mining algorithms must rely on the other two properties, if present. Our contribution in this Chapter is to describe an informed technique based on a Markov Model representation of events that does not use traceability of objects.

The context of nuclear plant monitoring is favourable to this study because nuclear flasks are hardly traceable: they have a weak visual signature when in the decontamination areas, where in some settings they occupy <20 pixels; and the acquisition rate is typically very low, ranging from 1 fps to 0.001 fps. Hence self-correlation of neither visual nor temporal features is sufficient to enable visual *tracking of a flask*. However, the process a flask undergoes is *structured* and *recurrent*. Therefore tracking can be performed by modeling the *process's regularities*, instead of the flask's.

8.1 Hidden Semi-markov Models

To better understand the contribution of *sequence and timing* of events to video mining, we have isolated the filtering on image content from the filtering on these temporal properties. We use scene change detection to transform the sequence of images into a sequence of symbols. As said in Section 6, the outcome is a *sequence of symbols* corresponding to the active areas of interest, and the associated image *timestamps*. This approach allows us to study the use in this context of the most common models of time series, i.e. Markov Models [46].

Hidden Markov Models are powerful and well-known tools that address all of the following aspects: i) a discrete, symbolic observation space; ii) a discrete, finite state space; iii) known machine learning algorithms [47]. *Hidden models* are justified in contexts where objects are not traceable. In our scenario flasks have no distinctive identifier and even a trained supervisor cannot discern one flask from another. Also, the low frame rate confuses the perception of motion direction: deciding whether a flask is entering or exiting a location requires more reasoning than simply inspecting the visual content of the previous and present frames. In probabilistic terms, these characteristics make the process 'hidden'.

A further step is to include explicit duration models. *Hidden semi-Markov models* (HSMM) are a family of probabilistic models useful for representing stochastic stationary processes with explicit state durations that can be observed only indirectly [47]. HSMMs introduce state occupancy distributions to represent sojourn times in non-absorbing states. The distribution of sojourn times is not constrained to the geometric distribution as in hidden Markov models.

A semi-Markov model is composed of an embedded first-order Markov chain X, and of discrete distributions of sojourn times S. The embedded chain is described by (T, χ_o), where χ_o is the initial state distribution and T is the transition matrix, such that $T_{ij} = P(X_{t+1}=j \mid X_t=i)$. For a semi-Markov model, $T_{ii}= 0, \forall i$. The sojourn time distributions are a set of discrete distributions depending only on the current state, $S = \{S_i, \forall i\}$. The model is hidden if the relation between the state and the observation is probabilistic. The emission distributions for every state are summarized in the emission matrix E, $E_{is} = P(O_t=s \mid X_t=i)$, s being an emitted symbol (Fig. 11).

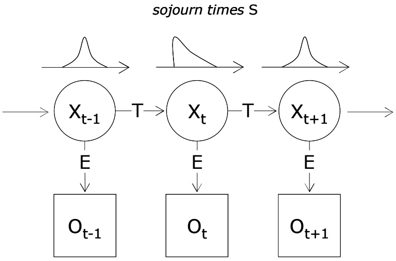


Fig. 11. A hidden semi-Markov model represents events with generic durations

The model building process is made by training the HSMM model on timings and durations on hand-labeled image streams, i.e. in our case on the official reports (cfr. Fig. 3). The algorithm for training an HSMM is an extension of the Baum-Welch algorithm and is detailed in [48].

8.2 Application to Nuclear Plants

In this Section let us develop the HSMM for a nuclear plant to exemplify the technique. We define a *plant* by the parameters (F, K, N). F is the maximum number of flask processes supported in parallel by the plant. K is the number of available cranes to move the flasks around ($K \leq F$). N is the number of processing stages that make up the nuclear process. In the example of Figure 6, N = 6. In realistic plants, $1 \leq F \leq 3$, K=1, and $3 \leq N \leq 10$.

For building the state space, let us first define the *real states* of the plant as F-arrays of labels indicating the progression stage of each flask in its individual process. This space includes all possible permutations (with repetition) of flask positions. The *real states* of a plant with F = 3 and N = 6 are [1 1 1], [1 1 2], [1 2 2], [6 6 6]. State [2 3 5] means that one flask is in stage 2, a second is in stage 3, and a third is in stage 5. We alter the number of real states for two reasons:

1. flasks are *indistinguishable*: flasks cannot be distinguished from one another, and so we can reduce the state space to the number of *combinations with repetitions*. For instance [122], [212], and [221] are equivalent.
2. emissions *depend on the transition*: by design, we associate the emission of a symbol to the event of a flask *entering* a state. So the starting state in a transition determines which symbol is emitted when the landing state is reached. For instance, consider a plant with F=2. Its real state [2 3] can be entered from [u 3], u≠2, thus triggering an emission linked to a flask entering stage 2. Or it can be entered from [2 v], v≠3, thus triggering an emission typical of a flask entering stage 3. To allow for multiple emission distributions, each *real state* is designed to be represented by *F virtual states*.

Given these premises, the size of the state space for a plant (F, K, N) is M:

$$M = F\frac{(N+F-1)!}{F!(N-1)!} = \frac{(N+F-1)!}{(F-1)!(N-1)!} \tag{1}$$

The *transition matrix* is a *M*x*M* square matrix. Even though *M* is independent of the number of cranes *K*, *K* constrains the transition probabilities so that T_{ij} is 0 if two states *i* and *j* differ for more than *K* single-process states. If *K*=1, only one process at a time can change state, hence for instance the transition [1 1]→[2 2] has 0 probability. All virtual states referring to the same real state have the same transition probabilities towards other real states (equal rows in T).

In our instantiation the distributions of *sojourn times* S are referred to single-flask processes, so that a real state has *F* associated distributions. We have found this approach to produce more accurate predictions than assigning one single distribution to each real state. We use parametric distributions of Gamma shape for single-flask stage durations. Gamma distributions are attractive because they can flexibly assume a shape ranging from an exponential to a bell Gaussian-like shape.

Because the analysis is performed on the labels outputted by scene change detection (SCD), we define the *emission alphabet* of size *A* as the set of SCD labels associated to the areas of interest drawn by inspectors. In the example used throughout this Chapter the emission alphabet is {H, D, P}, and *A* = 3. The matrix E is of size *M*x*A*. For the use as filter, the emission probabilities must be set to 1 for the symbol expected for that transition, and 0 otherwise. This will constrain the model to admit only the correct symbols for each state and to filter out the false alarms.

A trivial example of T for a plant with *F*=2, *K*=1, and *N*=2 is given in (2). Note that, with *N*=2, the only possible events are either a flask going from stage 1 to stage 2 or vice versa (T_{ii} being null by definition).

$$\mathrm{T} = \left(\begin{array}{cccccc} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{array}\right) \begin{array}{c} \textit{Real states} \\ [1\ 1] \\ [1\ 1] \\ [1\ 2] \\ [1\ 2] \\ [2\ 2] \\ [2\ 2] \end{array} \qquad \mathrm{E} = \left(\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{array}\right) \tag{2}$$

8.3 The MM Image Review Tool

The general idea is that, given the *history of annotations* produced by the inspector during the review, the (T, E, χ_o, S) model can highlight the next most likely relevant scene change detection (SCD) image. The inspector can decide whether to annotate the proposed image with a label corresponding to an admitted event, e.g. H D or P, or to reject it. The interaction is repeated with the next image until the end of the review (Fig. 12).

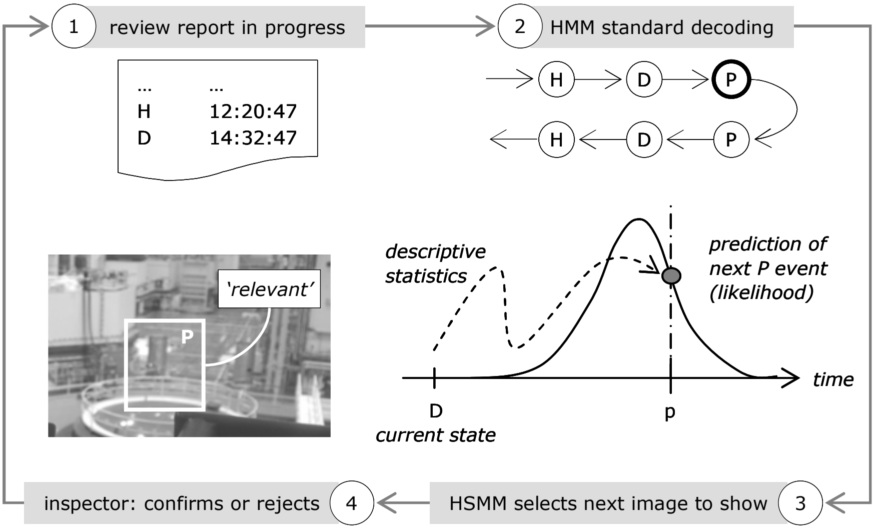

1
review report in progress
...
...
H
12:20:47
D
14:32:47
2
HMM standard decoding
H
D
P
H
D
P
descriptive statistics
prediction of next P event (likelihood)
'relevant'
P
time
D
current state
p
3
HSMM selects next image to show
4
inspector: confirms or rejects

Fig. 12. Using the HSMM as a filter for assisting a review

Let us consider a mid-point during the review, when the inspector's past annotations form a sequence ν. In the framework of an official review, we are authorized to consider the inspector as a fully reliable knowledge source, so that ν is true with probability 1. This implies that the uncertainty on the timing of events and on the value of symbols in ν is null. HSMM decoding [48] becomes superfluous: we are allowed to use simple HMM decoding [47] to retrieve the *current state* distribution χ, with an immediate advantage in terms of computational complexity.

The next image to be inspected is selected by computing the likelihood of every future image selected by SCD given χ and S. This likelihood is given by summing the S distributions of all stages associated to a state i with uniform probability ($1/F$) and then weighting this sum by the probability of the state χ_i, $\forall i$. The image exhibiting Maximum Likelihood (MAXL) is selected. If the inspector rejects this candidate, its likelihood is set to zero and the second MAXL is proposed, etc (Fig. 13). When an image is accepted and annotated, its symbol is added to ν and the procedure restarts.

It is reasonable to project the likelihood of duration models for one step in the future only, when the predictions are reliable. Predictions for more than one step would involve resorting on the convolution of duration models, which quickly flattens the joint probability distribution and hence deteriorates the filtering effect. For instance, in the case of Gaussian duration models, the convolution of two Gaussian distributions – corresponding to two prediction steps – is a Gaussian with a variance that is sum of the two original variances. The resulting distribution has a much lower peak than individual distributions and brings less information on when the second event will take place.

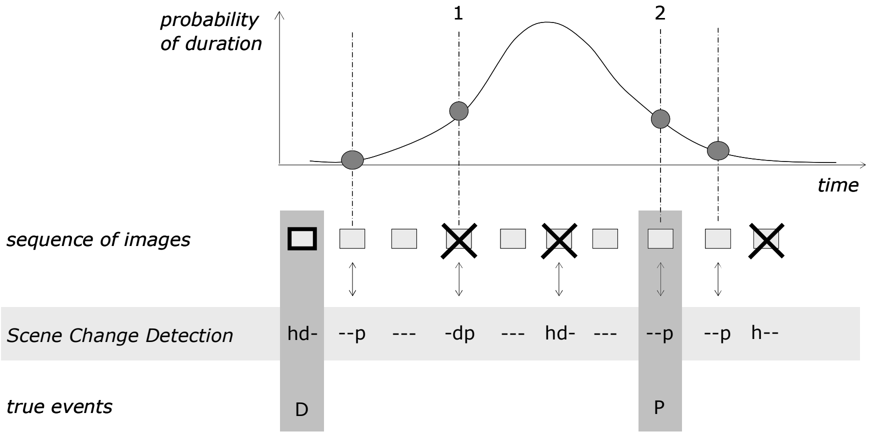


Fig. 13. How HSMM selects the next image to show. Starting from the last annotated image (first, bold square, with annotation D), the model skips all images associated to unexpected symbols (last two crossed squares). It then shows the image with Maximum Likelihood given the sojourn time pdf (circle nr. 1). The inspector rejects it (cross on the image), and so the model shows the second most likely (circle nr. 2). This reveals to be the correct one and is annotated by the inspector as a 'P' event.

In case of a missed detection, which can happen due to the Bayesian framework of predictions, ν can be either *inconsistent* or *consistent* given (T, E, χ_o). In the former case – easily detected because χ is zero –, confirmed symbols are suppressed in turn, with the heuristic of "last-confirmed suppressed first", and the remaining sequence is re-decoded until a consistent sequence is found. The prediction procedure described above is then applied from that point.

The case of a consistent sequence is treated as correct in all senses. Thus an error (missed detection) giving birth to a consistent sequence is not discovered unless, as the review progresses, the sequence becomes inconsistent. This problem unfortunately nullifies any guarantee of null false detection rate of the HSMM. In our software solution we have implemented case-based automatic reasoning to avoid false negatives, but the details are out of the scope of the current discussion. The point here, as we will see in Section 9, is the high rate of filtering provided by this time-based informed technique.

The guarantee of detecting a missed event given an inconsistent sequence instead exists. The duration distributions can be chosen so that all images are assigned a non-null probability of being selected (distributions with support on the whole set). Thus, in the worst case the whole set will be proposed for inspection, image by image, until consistency is re-established. Inspection of the whole set corresponds to the procedure currently in use.

MM has been embedded in a software tool with a user interface similar to Figure 9. The image selected for review is shown in the main window and the corresponding data are highlighted in the image list below.

From a user's viewpoint a MM-based review has a peculiar aspect: images are presented in a likelihood-driven order on the basis of duration models and not in a chronological order. If the MAXL image is refused, the second most likely image may happen to refer to a previous instant, and so the inspector sees a backward leap in the image stream. To avoid confusion, a visualization of the back and forth jumps is suggested. In any case, given that duration likelihoods are projected for one step only in the future, these jumps are localized in time around the segment of the image stream currently under review.

Because MM uses a model of structured processing, it does not discover images of *irregular behavior*. MM is useful as a way to eliminate regular events so that irregular events can be more easily isolated and inspected.

As a conclusive remark, employing the MM filter gives a strong advantage in terms of *workload reduction* (Section 9) and is not invasive for the inspector. The filter runs in the background and can be deactivated at anytime. Furthermore, its computational load is completely manageable by a standard laptop processor, thanks in particular to the choice of substituting HMM-like decoding in place of the more complex HSMM decoding.

8.4 Training with MM

Training the MM filter is substantially different than training the decision trees (Section 7.4), even though the concepts of interactivity and iteration are still applied. In MM, the user must specify the following parameters:

1. the plant's parameters F, K, and N (Section 8.2);
2. the expected (standard) sequence of processing steps of an individual flasks.

These parameters are plant-specific and need to be defined only the first time an inspector reviews image streams from a given plant. Typically they are set by an expert at system setup.

Before the first use, MM must be trained to set the HSMM matrixes and duration models. In our scenario, the training set is provided by the reports redacted in the past about processing on the current plant. They contain all the necessary information to train both the symbolic and duration models, i.e. meaningful events and relative time stamps (Fig. 3). The reports are double-checked for precision and are plant-specific, two desirable aspects for a training set. MM automatically scans all reports regarding a given plant and uses the training algorithm in [48] to set the HSMM parameters. This procedure can be computationally intensive but it is done offline, outside review time.

When MM starts training for the first time, the system manager must also specify the state in which the plant was before the first event in the first report used for training. This specification may be trivially that no flask is present in the plant, but still it is necessary. In the following, when MM is used in a review, the last state of the last available report is employed as the starting state for decoding. This state is computed and stored automatically by MM during training or use. If use of MM is discontinued, in the sense that inspectors do not run MM for a review, the starting state must be specified again before employing MM the next time.

During a review the model is used as it is before starting, without any online alteration. After completing the review on a batch, inspectors double-check the resulting list of events and then approve it. Only at this point, the new report is added to the tail of the previous reports to form a continuous history of past reviews and MM models are retrained on this history.

Updating the model online does not deliver better results. In fact as we explained above, the MM filter may skip momentarily one image. If an update were done in such a situation, the resulting models may spoil the filtering performance for the rest of the review. Thus, interactivity in the case of MM is confined to the offline retraining phase.

9 Benchmarking Techniques to Search Streams of Images

In the following, we focus on experimental results obtained by running scene change detection (SCD), decision trees (DT) and Markov Models (MM) image review tools on grayscale images acquired by a GEMINI surveillance system [49] in two different plants, *A* and *B*.

A is a single-flask plant ($F = 1$), while *B* has the capacity to process two flasks ($F = 2$) with the constraint of a single crane ($K = 1$) (Section 8.2).

The tests aimed at measuring the performance of each tool in the detection of flask events: events of type H and P for DT, and events H, D, and P for MM. For DT, we skipped events of type D because, as noted in Section 8, these have an extremely weak visual signature.

9.1 Image Sets

Table 3 provides information about the image sets used for the tests. Sets A1, A2 and A3 were acquired in plant *A,* while B1 and B2 stem from plant *B.* Each image set spans over several months of plant activity. For each set, the number of target events to identify is shown: this ground truth information has been derived from inspectors' official review reports. As anticipated in Section 4, the number of safeguards-relevant events is very exiguous compared to the number of images in each set. Also, it is not unusual to have image sets where no safeguards-relevant activity needs to be reported by the inspectors as in A3.

Table 3. Image sets used in the benchmark. For each set, the Table lists the number of images in the set, and the number of events to be detected over the hatch (H), decontamination (D), and pond (P) areas.

Image set	Images	H	D	P
A1	20160	17	17	17
A2	15661	1	1	1
A3	16022	-	-	-
B1	16020	30	30	30
B2	15446	12	12	12

9.2 Performance Metrics

To evaluate the performance of the image review tools we adopt two measures.

The *first* is a standard evaluation performance for classifiers used in information retrieval, namely *recall* and *precision* [33]. For a given classification method M and benchmark containing R^* true events, the retrieval indexes are defined as:

$$recall_M = \frac{CR_M^*}{R^*} \tag{3}$$

$$precision_M = \frac{CR_M^*}{CR_M} \tag{4}$$

where:

- CR_M is the number of images classified by M as relevant,
- CR_M^* is the number of relevant images correctly classified by M.

The classification is optimal in terms of *recall* and *precision* when both indexes have value 1. *Recall* equals 1 when there are no false negatives, i.e. no relevant image is classified as irrelevant. *Precision* equals 1 when there are no false positives, these being defined as:

$$FP_M = CR_M - CR_M^* \tag{5}$$

The *second* performance index is application-oriented [50]. Given that SCD is the default filter used in safeguards reviews, we measure the user advantage to filter SCD events by a second classifier M (provided that $recall_{SCD}$ and $recall_M$ equal 1) by:

$$workload\_reduction_M = \frac{FP_{SCD} - FP_M}{FP_{SCD}} \tag{6}$$

If FP_M equals FP_{SCD}, M does not bring any advantage and the *workload reduction* is 0. If FP_M is 0, M brings a tremendous advantage and the *workload reduction* is 1.

9.3 Experimental Results

SCD has been parameterized in this benchmark to guarantee that *all* events are detected by optimal thresholds for each area of interest (AOI), i.e. the thresholds that provide $recall_{SCD}$ equal to 1 minimize FP_{SCD}. Further, because DT is tested only on H and P events, while MM is tested on the whole chain of events H-D-P, SCD was run twice on each image set. The first time on two AOIs only, to be followed by DT: the SCD performance is indicated by SCD-2. The second time SCD was run on three AOIs for MM, and it is referred to as SCD-3.

DT and MM were trained by the procedures described in Sections 7.4 and 8.4 on roughly the first half of the events contained in A1 and B1. Hence the remaining events test DT and MM generalization performance. For DT, the image content on each AOI was indexed by a grayscale histogram at 64 components.

Since both DT and MM detected all ground truth events (i.e. *recall* = 1 for both techniques), the evaluation is focused on *precision* (Fig. 14) and *workload_reduction* (Fig. 15).

Concerning *precision* (4), Figure 14 shows that SCD produces many false positives, even when it is parameterized to work at its best (striped bars). In particular, to detect all decontamination events, a very low detection threshold had to be set, and this accounts for the significant drop in *precision* between SCD-2 and SCD-3. With SCD-3, the number of SCD events to be reviewed by an inspector almost doubles. In this context, Figure 14 shows that the *use of informed techniques after SCD pays off*. Both DT and MM score a higher *precision* on all image sets. The use of search by content techniques like DT is advantageous for the events of type H and P, while MM copes well with the task of tracking the whole flask-processing chain, despite the very noisy stream of SCD-3 events.

The real advantage of using these techniques after SCD is measured *by the workload reduction* index (6) shown in Figure 15. The high peaks reached by DT and MM show that they manage to reduce the number of false positives generated by SCD by a very significant amount. From the user point of view, this reduction implies at least 60% of images less to be reviewed.

Note that, compared to the previous chart on *precision*, MM now scores in a comparable way w.r.t. DT, in that MM *workload reduction* is measured relatively to SCD-3.

Finally, as term of comparison for a popular search by content technique, namely image retrieval (IR), Figures 14-15 include the performance of IR as implemented in [1]. This is basically a nearest-neighbor classifier, as referred to in Section 7.1-7.2. The comparison between IR and DT is interesting, because a reason for choosing decision trees as classifiers was their ability to focus the classification automatically on the ‘relevant features’ of the image content (Table 2). In our experiments on real safeguards image streams, the lower *precision* by IR confirms that feature selection before classification is a necessary step. As a matter of fact, the trained decision trees include few features (always less than 5) of the original grayscale histograms, suggesting that the remaining components act as noise in the IR retrieval by similarity process. Also, being economic in terms of number of features, DT classifiers are *fast to be tested* on large images sets like ours, a key aspect for interactive machine learning to be acceptable to users.

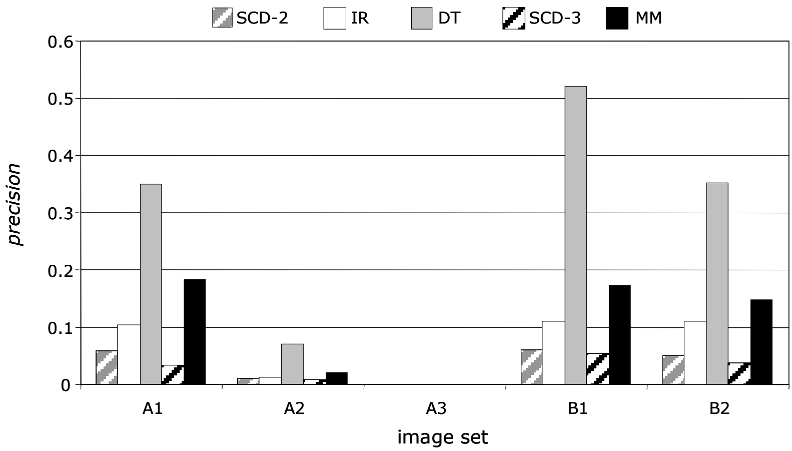


Fig. 14. *Precision* provided by SCD-2 (scene change detection on 2 AOIs), SCD-3 (SCD on 3 AOIs), IR (image retrieval by nearest neighbors), DT (decision trees), and MM (Markov models) measured on five image sets

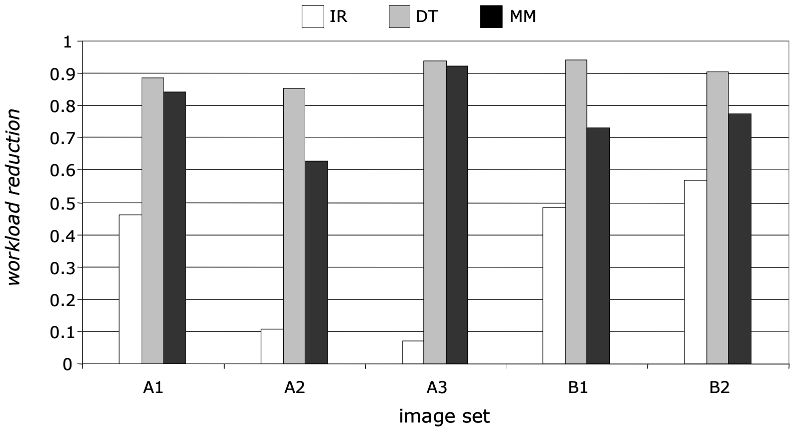


Fig. 15. *Workload reduction* provided by three filters IR (image retrieval by nearest neighbors), DT (decision trees), and MM (Markov models) measured on five images sets

10 Discussion and Wrap-Up

The application of *machine learning* (ML) to the analysis of image streams in fields related to video surveillance is a hot topic in the computer vision community. Until today, the penetration of automatic image-stream analysis systems in the market and real-world situations has been far lower than expectations. After the terrorist menace of the early years of this decade all observers had projected a dramatic increase in the demand of video analytics. However, projections of sector growth continue to be corrected downwards year after year [51, 52].

From our survey of a large set of online sources, the main reasons for this missed opportunity are two, but strictly interconnected: one is *reliability*, the other is the *complex setup of such systems*. Automatic image stream analysis works well under certain conditions of illumination, camera setup, clutter of scene, and so on. The setup needs to be done by highly trained individuals that perfectly understand the assumptions that must be respected for these systems to achieve the expected performance. When these assumptions are not met, reliability slips below acceptable standards. In this picture, machine learning will most probably play a crucial role in the future. For a real breakthrough in applied video analytics, ML methods will have to substitute technical experts of computer vision in setup phase and in self-reliability assessments.

The issue of a more natural setup and training of vision algorithms has been discussed and exemplified all over this Chapter. As for the reliability issue, the authors of [53] advocate for the advent of *autonomic computer vision systems*. The term ‘autonomic computing’ was inspired by natural self-governing systems, and in particular from the autonomic nervous system of mammals. In practical terms, the concept implies four capabilities: *self-monitoring* of internal parameters, *self-regulation* to ensure a preset quality of service, *self-repair* in case of failure, and *self-description* of internal state. All these should be achieved given high-level objectives from system administrators, *without intervention by highly skilled experts*. Programming these capabilities for all possible settings will be prohibitive; the same authors suggest using automated learning, like clustering of data, to identify the ‘normal’ state of operation.

Based on our experience, we believe that in the near-to-medium term interactive machine learning will preserve the upper hand over unsupervised as to quality of models learned, especially when the supervisor possesses highly specialized application-related knowledge. Notwithstanding the likely advances in unsupervised learning, the ‘user in the loop’ paradigm will still be more desirable in highly delicate fields (e.g. security, nuclear) where failure of video surveillance may have dramatic impact.

Our contribution in this Chapter has hence focused on *interactive machine learning* (IML) and on ways to capture field knowledge from users that have no training in computer vision or ML. We have addressed an application field, that of monitoring for nuclear safeguards, having the following aspects: (i) identifiable objects with strong visual features in some images, contrasted by low resolution of the same object in following images, (ii) low frame rate, and – as a consequence of the latter – (iii) high variability of appearance in the (often) only image that shows the target in a given location. All three characteristics contrast with the assumption of *traceability* underlying many computer vision studies.

Our contribution is not limited to nuclear safeguards, but can be extended to other video surveillance scenarios that have those same characteristics. The most compelling scenario is general industrial process monitoring, where the video system may not be set for video-surveillance standards. A surveillance system on a parking lot that delivers a frame every ten or twenty seconds is also a good example of an alternative scenario. Another example is browsing surveillance images ex-post, for instance in a university campus environment, that have been recorded at a lower rate than acquisition because of limited storage space.

As specific technical contributions, in this Chapter we presented a range of filters based on scene change detection (SCD), decision trees (DT) and Markov Models (MM) to search streams of surveillance images, together with a benchmark on process monitoring for nuclear safeguards.

Search techniques can be uninformed, as SCD, or informed, as DT and MM. A main difference is that informed techniques use priors on events derived from previous searches and annotations made by the users to facilitate the recognition of new, related events.

Priors may concern not only the image content, but also some image meta-data information, like the temporal one. These techniques can be applied when the typology of events to be searched is recurrent, and has a distinct visual or temporal signature.

Uninformed techniques should be applied when no prior exist about what is being searched. They are general purpose filters, and easily programmable. These two properties make them the default solution even in cases, like the safeguards process monitoring, where one could take advantage of past search results and annotations.

A barrier for informed search techniques to be applied more largely lies in the number of machine learning parameters on which these techniques depend and which everyday users do not understand. The interactive machine learning approach can alleviate the problem. In our contribution the ‘user in the loop’ philosophy took shape in two different ways.

In the DT image review tool, interactive machine learning is used to address the *shaping* of search by content filters. All ‘machine learning parameters’ for the image classifier are estimated implicitly on the basis of a relevance feedback loop by which the user provides a growing set of positive and negative examples of what she is looking for. During training, the iterative visualization of the classification result provides the user with intuitive information about the convergence of the learning process. This is crucial for her to decide if she can ‘trust’ the tool event detection capability.

In MM, the ‘user in the loop’ part is shifted on the usage side of the tool, i.e. to perform the actual image review. MM is a predictor of the temporal location of events based on sequence and duration priors. The estimation process of its model is fully automatic. Because long term predictions are inherently more inaccurate than the short term ones, MM is intelligently embedded in the image review cycle: it takes advantage of the online event annotations provided by the user to iteratively re-estimate the temporal position of the next relevant event. In this way, MM helps the user detecting relevant events while reducing the number of false positives.

A non technical contribution of this Chapter is a comparison between interactive machine learning techniques as proposed by different research communities: the *user centered methods*, put forward by the human computer interaction community, and *active learning* techniques which are closer to the machine learning research community. In abstract terms the problem addressed is the same: machine learning on the basis of a *limited number of examples labeled by the user online* (and, hence, under *limited computation time*). A fundamental difference between the two approaches lies in *who* drives the examples' labeling process. When the user drives it, emphasis is on informing machine learning by the user's knowledge in a goal directed way. By contrast, in active learning the process is driven by an algorithm's self analysis about its points of uncertainty on a given problem. Finding suitable ways to integrate these approaches can be a valuable future research area.

References

[1] Versino, C., Stringa, E., Gonçalves, J.G.M.: Review of Surveillance Images using Classification Techniques in a Relevance Feedback Context. In: Proc. Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS2003), pp. 46–53 (2003)

[2] Fails, J.A., Olsen, D.R.: Light Widgets: Interacting in Every-day Spaces. In: Intelligent User Interfaces, IUI 2002, pp. 63–69. ACM, New York (2002)

[3] Fails, J.A., Olsen, D.R.: Interactive Machine Learning. In: Intelligent User Interfaces, IUI 2003. ACM, New York (2003)

[4] Fails, J.A., Olsen, D.R.: A Design Tool for Camera-based Interaction. In: Human Factors in Computing Systems, CHI 2003, pp. 449–456. ACM, New York (2003)

[5] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)

[6] `http://www.rulequest.com/see5-info.html`

[7] Stumpf, S., Sullivan, E., Fitzhenry, E., Oberst, I., Wong, W.-K., Burnett, M.: Integrating rich user feedback into intelligent user interfaces. In: Proc. 13th International Conference on Intelligent user Interfaces, pp. 50–59 (2008)

[8] Stumpf, S., Rajaram, V., Li, L., Wong, W.K., Burnetta, M., Dietterich, T., Sullivan, E., Herlocker, J.: Interacting meaningfully with machine learning systems: Three experiments. International Journal of Human-Computer Studies 67(8), 639–662 (2009)

[9] Ware, M., Frank, E., Holmes, G., Hall, M., Witten, I.H.: Interactive machine learning: letting users build classifiers. International Journal of Human-Computer Studies 56(3), 281–292 (2002)

[10] Zhou, X., Huang, T.S.: Relevance feedback in image retrieval: A comprehensive review. Multimedia Systems 8(6), 1432–1882 (2003)

[11] Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In: 17th International Conference on Machine Learning, pp. 401–412 (2000)

[12] Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: Proc. of ACM Int.Conf. on Multimedia, pp. 107–118 (2001)

[13] Vapnik, V.: Statistical Learning Theory. Wiley, Chichester (1998)

[14] Tao, D., Tang, X., Li, X., Wu, X.: Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans. Pattern Recognition and Machine Intelligence 28(7), 1088–1099 (2006)

[15] Markou, M., Singh, S.: Novelty Detection: A Review – Part 1: Statistical Approaches. Signal Processing 83 (2003)

[16] Markou, M., Singh, S.: Novelty Detection: A Review – Part 2: Neural network based approaches. Signal Processing 83, 2499–2521 (2003)

[17] Mcivor, A.M.: Background subtraction techniques. In: Proc. of Image and Vision Computing, pp. 147–153 (2000)

[18] Stauffer, C., Grimson, E.: Learning Patterns of Activity Using Real-Time Tracking. IEEE Trans. Pattern Recognition and Machine Intelligence 22(8), 747–757 (2000)

[19] Ahmed Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.S.: Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance. Proceedings of the IEEE 90(7), 1151–1163 (2002)

[20] Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and practice of background maintenance. In: Proc. 7th IEEE Int. Conf. on Computer Vision, pp. 255–261 (1999)

[21] Messelodi, S., Modena, C.M., Zanin, M.: A Computer Vision System for the Detection and Classification of Vehicles at Urban Road Intersections. Pattern Analysis & Applications 8(1), 17–31 (2005)

[22] Liu, Y., Yao, H., Gao, W., Chen, X., Zhao, D.: Nonparametric Background Generation. In: Proc. 18th Int. Conf. on Pattern Recognition, vol. 4, pp. 916–919 (2006)

[23] Morris, B.T., Trivedi, M.M.: A Survey of Vision-Based Trajectory Learning and Analysis for Surveillance. IEEE Trans. Circuits and Systems for Video Technology 18(8), 1114–1127 (2008)

[24] Makris, D., Ellis, T.: Learning semantic scene models from observing activity in visual surveillance. IEEE Trans. Systems, Man, Cybernetics B 35(3), 397–408 (2005)

[25] Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM J. Computing Surveys 38(4), 13 (2006)

[26] Hu, W., Tan, T., Wang, L., Maybank, S.: A survey of visual surveillance of object motion and behaviors. IEEE Trans. System, Man, Cybernetics C 34(3), 334–352 (2004)

[27] Ewerth, R., Freisleben, B.: Semi-supervised learning for semantic video retrieval. In: Proc. 6th ACM Int. Conf. on Image and Video Retrieval, pp. 154–161 (2007)

[28] Dorado, A., Calic, J., Izquierdo, E.: A rule-based video annotation system. IEEE Trans. on Circuits and Systems for Video Technology 14(5), 622–633 (2004)

[29] Iyengar, G., Nock, H.J.: Discriminative model fusion for semantic concept detection and annotation in video. In: Proc. 11th ACM Int. Conf. on Multimedia, pp. 255–258 (2003)

[30] Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Mei, T., Zhang, H.-J.: Correlative Multi-Label Video Annotation. In: Proc. 15th Int. Conf. on Multimedia, pp. 17–26 (2007)

[31] Wang, M., Hua, X.-S., Hong, R., Tang, J., Qi, G.-J., Song, Y.: Unified Video Annotation via Multigraph Learning. IEEE Trans. Circuits and Systems for Video Technology 19(5), 733–746 (2009)

[32] GARS, `http://www.canberra.com/pdf/Products/Systems_pdf/GARS-SS.pdf`

[33] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)

[34] Vaccaro, R., Dellepiane, S., Nicchiotti, G.: Content based database system: A solution to multimedia data management. In: Proc. of the Image and Video Content-Based Retrieval Workshop, pp. 89–95 (1998)

[35] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petrovic, D., Steele, D., Yanker, P.: Query by image and video content: The QBIC system. IEEE Computer, Los Alamitos (1995)

[36] Brunelli, R., Mich, O.: Image retrieval by examples. IEEE Transactions on Multimedia 2, 164–171 (2000)

[37] Meilhac, C., Nastar, C.: Relevance feedback and category search in image databases. In: Proc IEEE International Conference on Multimedia Computing and Systems, vol. 1, pp. 512–517 (1999)

[38] Cover, T., Hart, P.: Nearest Neighbor Pattern Classification. IEEE Trans. on Information Theory 13, 21–27 (1967)

[39] Hertz, J., Krogh, A., Palmer, R.G.: Introduction to the Theory of Neural Computation. Addison Wesley, Reading (1991)

[40] Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (eds.): Feature Extraction, Foundations and Applications. Series Studies in Fuzziness and Soft Computing, Physica-Verlag. Physica-Verlag, Springer (2006)

[41] Sullivan, G.J., Wiegand, T.: Video Compression—From Concepts to the H.264/AVC Standard. Proc. IEEE 93(1), 18–31 (2005)

[42] Mitra, S., Acharya, T.: Gesture Recognition: A Survey. IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews 37(3), 311–324 (2007)

[43] Brdiczka, O., Langet, M., Maisonnasse, J., Crowley, J.L.: Detecting Human Behavior Models From Multimodal Observation in a Smart Home. IEEE Trans. on Automation Science and Engineering (2003)

[44] Ferrando, S., Gera, G., Massa, M., Regazzoni, C.: A New Method for Real Time Abandoned Object Detection and Owner Tracking. In: IEEE International Conference on Image Processing, pp. 3329–3332 (2006)

[45] Trucco, E., Plakas, K.: Video Tracking: A Concise Survey. IEEE Journal of Oceanic Engineering 31(2), 520–529 (2006)

[46] Lombardi, P., Versino, C.: Tracking Nuclear Material at Low Frame Rate and Numerous False Detections. In: Proceedings of ICVS 2007 Vision Systems in the Real World: Adaptation, Learning, Evaluation - Conference, Bielefeld, Germany (2007)

[47] Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77(2), 257–286 (1989)

[48] Guédon, Y.: Estimating hidden semi-Markov chains from discrete sequences. Journal Comput. Graphical Statistics 12(3), 604–639 (2003)

[49] http://canberraeurisys.com.cn/PDF/gemini_surv.pdf

[50] Lombardi, P., Versino, C., Gonçalves, J.G.M., Heppleston, M., Tourin, L.: MM: The Markov Model Tool for Image Reviews. In: Proc. of the Institute of Nuclear Materials Management (2008)

[51] IMS Research lowers video surveillance forecast but outlook still positive (March 2009), http://www.imsresearch.com/press_release_details.html&press_id=865

[52] http://www.securitysystemsnews.com/index.php?=p=article&id=ss200906FZVYeV

[53] Crowley, J.L., Hall, D., Emonet, R.: Autonomic Computer Vision Systems. In: Proc. 5th Int. Conf. On Computer Vision Systems (ICVS 2007), p. 182 (2007)

Automatic Video Activity Recognition

Haowei Liu[1], Ming-Ting Sun[1], and Rogerio Feris[2]

[1] University of Washington, Seattle, WA 98195
[2] IBM T.J Watson Research Center, Hawthorn, NY 10532

Abstract. With the demands on public security and the availability of large storage systems, an increasing number of video surveillance systems are being deployed all over the world to help people detect interesting target events. However, most of these systems require intensive human monitoring, or require human operators to review video footage corresponding to extended periods of time, only to find a few short clips that are of interest. The problem fosters a demand of an automatic computer surveillance system, which can assist the human operators in identifying possible interesting events. This challenge has attracted researchers from different domains, leading to a variety of proposed approaches, particularly in the field of human activity recognition. These approaches vary in the choice of representation and methodologies as well. This chapter gives a survey and reviews the state of the art approaches to automatic human activity recognition in videos.

1 Introduction

There is an increasing interest in developing video monitoring systems all over the world. For example, the retailers are interested in surveillance systems that can prevent fraud in self-checkout systems or detect thefts at the point of sale. The medical staff is interested in home care surveillance systems that could help monitoring the elders so that they can respond promptly in case of emergencies. The law enforcement officers are interested in traffic surveillance systems that can help detect possible traffic violation events. Thus, it is very desirable to have an automatic computer surveillance system that can help the human operators in identifying possible interesting events.

Owing to its great potential, the problem of how to automatically discover and recognize human activities from video data has become a popular topic and attracted many researchers in the computer vision community. A solution to this problem will not only facilitate an automatic video surveillance system [30] but will also improve applications, such as video retrieval or summary and human machine/robot communication [2]. In addition to its importance for many practical applications, it is also important in the context of machine learning, particularly on how video processing approaches could allow a high-level understanding of the data.

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 327–349.
springerlink.com Springer-Verlag Berlin Heidelberg 2011

The challenges faced by researchers include the following:

- Video quality: When the video sequence is captured, factors such as lighting changes or the photometric variability of the camera can affect how objects appear in the captured video frames. When it is dark, the objects might not even be easily visible.

- Occlusions: There might be multiple objects in the video, and the target object might be occluded by others, or even by itself.

- Camera motion: The camera could be moving as well, either regularly or irregularly, making the tracking or estimation of the target objects harder.

- Cluttered background: There might not be always a clean background in the video. Combined with the effect of camera motion, part of the background could be erroneously detected as an object.

- Event variations: Objects may be deformable and thus exhibit high variations even for the same activity. In general, an event could have many variations and exceptions, making the recognition difficult. We are interested in unusual events. However, it is often difficult to describe what is an unusual event due to the large possible variations and exceptions of usual and unusual events.

- Semantic gap: Many detection results depend on the semantics of the activities. It is often very difficult to detect high-level semantics of an activity from low-level features.

Therefore, the requirements of the video analysis techniques are manifold. They need to be capable of dealing with difficult situations such as cluttered background, poor lighting conditions, camera motion, occlusion, geometric and photometric variability, etc. In this chapter, we focus on single camera activity recognition. Also, due to its usefulness in typical real world surveillance settings, most works on activity recognition focus on human activities, i.e. recognizing how human body parts move, and also human motion patterns, i.e., the trajectories of human movement when the information about different parts are not available.

Numerous methods have been proposed to address the problem of action recognition and analysis in video sequences. They are different in various ways including models to use, features to extract, etc. In the subsequent sections, we will review the state of the art approaches in different major categories. We start with the template matching based approaches in Section 2, and elaborate on the trajectory based methods in Section 3. In Section 4, we talk about methods based on state-space models, and in Section 5, we introduce the popular methods based on the “Bag of Words” features. We introduce the publicly available datasets and report the performance comparison in Section 6, and finally conclude the chapter in Section 7.

2 Activity Recognition Using Template Matching

Template-based approaches use the straightforward principle of nearest neighbor search. They are typically used under the supervised learning framework with a labeled training set. When a new video sequence is given, these approaches compare it with a set of stored video templates in the training set. The new video sequence is then assigned the class label of the template in the training set that most resembles it. Typically, these approaches make use of spatial temporal features to match and identify specific actions in videos. They are usually applied when training a specific model for each activity of interest is infeasible, for example, under the situation where the size of the training set is small compared to the number of activities. The difficulties underlying these approaches are the applicability of the extracted spatial temporal template and a suitable choice of distance measures.

In [7], Bobick and Davis use motion history images - a.k.a. temporal templates - for action classification. A motion history image is one that encodes the motion information in recent frames. Let *I(x, y, t)* be a video sequence, and *D(x, y, t)* be the sequence of binary images indicating the regions of motion, which can be obtained by simply differencing the adjacent frames or using other background subtraction techniques. Then each pixel of the motion history image *H(x, y, t)* can be defined as follows:

$$H(x,y,t)=\begin{cases} \tau & \mathit{if}\ D(x,y,t)=1 \\ \max(0,H(x,y,t-1)-1) & \mathit{otherwise} \end{cases} \tag{1}$$

where *D(x, y, t)* = 1 indicates that the pixel is inside the region of motion for frame *t* and τ is an application dependent parameter that controls the temporal extent of the motion. That is, if a pixel at frame *t* is inside the region of motion, then its motion history is set to τ, which will be decremented if the pixel is stationary in the next frame. The intuition behind the motion history image is to aggregate the motion information of each pixel in the past τ frames, and as a result, the recently moving pixel would be brighter in the motion history image. The authors show that by using the temporal template, activities such as sitting down, hand waving, and crouching can be well distinguished. The approach provides a simple way to gather the motion statistics from video sequences and by matching these statistics against known action models, the class label of a new test video can be quickly determined. However, the approach is not suitable when applied to complex activities, especially those that could overwrite the motion history images. Representation of the time, the order, or the sequence of how an activity is performed would be needed to address this problem [52].

Efros et al. [8] introduce a spatial temporal descriptor that works well for human activity recognition on far-field videos where the whole object of interest, for example, a person, may only be 30 pixels tall. The descriptor is based on the computation of optical flow. Let $I_A(x, y, t)$ be sequence A, then for each frame t of sequence A, the descriptor first computes a horizontal and a vertical optical-flow field $F_{x,}$ and F_y, which are decomposed into $F_x = F_x^+ - F_x^-$, and $F_y = F_y^+ - F_y^-$ where the four fields F_x^+ ,F_x^- ,F_y^+, F_y^- are nonnegative and later smoothed and normalized. Note that $F_x^+ (i, j) = F_x (i, j)$ if $F_x (i, j) > 0$, and 0 otherwise, while $F_x^- (i, j) = F_x (i, j)$ if $F_x (i, j) < 0$, and 0 otherwise, similarly for F_y^+ and F_y^-. The similarity between frames from two sequences can be determined by comparing the computed optical-flow fields using measures such as the normalized correlation or the Sum of Absolute Difference (SAD). The sequence-level similarity can also be obtained by aggregating the frame-level similarities. With the motion descriptor, the authors show that it is possible to distinguish different sports activities, such as run left, run right, walk left, and walk right. However, this approach requires segmenting, tracking, stabilizing each human figure in the sequence, and then reliably extracting the bounding boxes, which might not always be possible in complex scenes.

Other approaches include the space-time shape volumes proposed by Blank et al. [9] for classification, a space time volumetric feature proposed by Ke et al. [1], and a similarity metric between video patches based on the intensity variation proposed by Shechtman and Irani [10]. In [39], Wang et al. attempt to cluster human activities for each video frame based on deformable matching.

A common drawback of the aforementioned template based methods is their inabilities to generalize from a collection of examples and create a single template which captures the intra class variability of an action.

More recently, Rodriguez et al. [11] address this problem using the Maximum Average Correlation Height filter, or MACH filter [12], which was originally proposed to solve target identification problems in static images. Given a set of training images I_1, I_2 …I_N, the MACH filter aims to create a template H in the frequency domain that gives the best representation of these images. The intuition is that if the filter H is viewed as a linear transform, then when correlated with the training images, it should produce a set of correlation planes that resemble each other and exhibit the least possible variations. It should also maximize the height of the main lobe on the correlation plane, while minimizing the magnitude of undesirable side lobes. Figure 1 shows a profile plot of a correlation example. A good H should try to make the plot resemble the Dirac Delta function by making the main lobe as narrow and as high as possible while keeping the magnitude and the number of side lobes small. Thus, H can be obtained by optimizing the aforementioned criterions. By using the Clifford Fourier Transform [13] to handle vector valued functions, Rodriguez et al. [11] successfully extend the MACH filter into the video domain where the features usually are not simply pixel values but multi-dimensional vectors, such as gradients in different directions or optical flow, and achieve better performances on standard datasets.

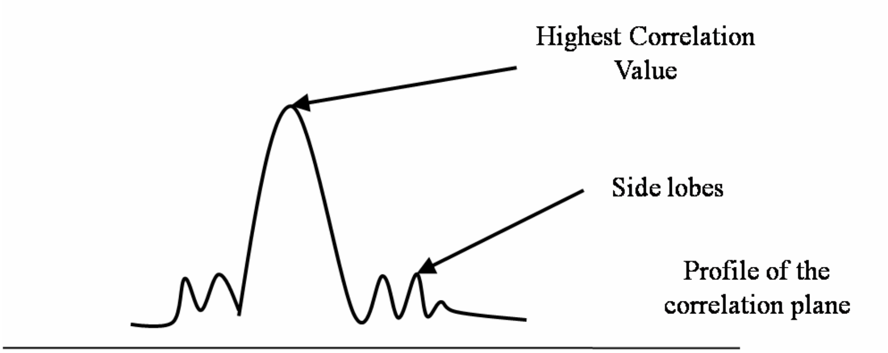


Fig. 1. A sample profile of the correlation with a peaked main lobe and a few side lobes

3 Activity Recognition Based on Tracking

Recognizing activities based on tracking is extensively studied and researched in the computer vision community. These approaches are usually applied in the settings where the bounding boxes of the moving objects can be reliably extracted. Later on, features such as object speed, trajectory, or minimum bounding box size can be computed for activity recognition. For example, in a traffic surveillance system, the trajectories of cars can be gathered to determine the anomalies, such as a car making a wrong turn.

3.1 Trajectory Clustering

In earlier works [14, 15, 16, 17], the moving objects, typically cars and pedestrians, are tracked and the trajectories are clustered in order to locate interesting locations in the scene, for example, regular paths, junctions, or exits, based on which, objects with trajectories that deviate from regular patterns can then be detected, for example, a blindly walking person or a speeding car.

More specifically, in [14], Makris et al. accumulate trajectory data over long time periods to learn a path model, which is later used to predict future pedestrian's locations and aid the recognition of unusual behavior identified as atypical motion. Tan et al. [15, 16] perform hierarchical clustering on the collected trajectory data in order to discover finer-grained motion patterns in a crossroad, e.g., traffic in the left lane versus traffic in the right lane. In [17], Junejo et al. propose a framework unifying automatic camera calibration, moving object tracking, and trajectory clustering. Aerial images are also incorporated and registered with the 2D images to uncover the scene structure.

The benefit of these approaches is that they provide a simple and intuitive way for a preliminary analysis of the video. The raw trajectories are collected using a tracking algorithm, and subsequently, the raw trajectory data are clustered to uncover interesting spatial patterns in the scene. The drawback is that modeling or detecting more complicated activities is difficult as more advanced models are needed to incorporate the temporal information. We review some of these advanced models next.

3.2 Activity Recognition Using Higher-Level Representations

Recently, people become interested in detecting more complicated activities, which requires a higher-level representation of the trajectory instead of sparse samples in the 2D space. One common approach is to represent the trajectories as strings. Then, more sophisticated models can subsequently be applied to these strings to uncover latent semantics.

As an example, in [18], a surveillance system is set up to detect various activities in a kitchen. The objects of interest, such as the stove, the fridge, or the tables are labeled and then the person is tracked to find out the sequence of his/her interactions with those objects of interest. For example, if the person first goes to the fridge, then the table, and finally the stove, after tracking, the activity can be represented by the string {fridge, table, stove}. Note that an activity can be composed of sub-activities. Consider, for example, the activity *a* = {1, 2, 3, 1, 2, 3}. Note that subsequence {1, 2} occurs with the same frequency as {1, 2, 3}. In other words, {1, 2} does not encode any extra information given the subsequence {1, 2, 3}, and therefore we can see that *a* is composed by the recurrent sub-activity {1, 2, 3}. In order to efficiently represent an activity and uncover possible sub-activities, the string representation is transformed into a suffix tree. A suffix tree for a string *S* is one such that each suffix of *S* corresponds to exactly one path from the tree's root to a leaf, which can be constructed in linear time using the approach in [24]. Take the string *S* = {1, 2, 3, 2, 3, 2, 4} for example, its suffix tree is shown in Figure 2(a) and the suffix {2, 3, 2, 3, 2, 4} corresponds to the path R->2. After a suffix tree *T* is constructed, a histogram of the constituent suffixes can be efficiently computed by traversing through *T* starting from the root-node. The count of a particular suffix is the number of times its corresponding edge is traversed.

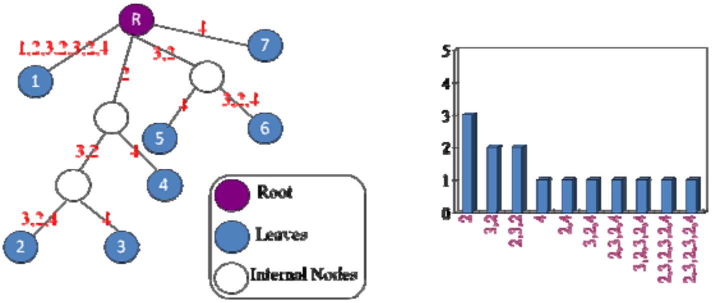


Fig. 2. Examples regenerated from [18]. (a) Constructed suffix tree for string S ={1, 2, 3, 2, 3, 2, 4} (b) Histogram of the subsequences of string S.

For example, to know the counts of subsequences {3, 2}, we simply count how many times edge {3, 2} is traversed when going through all the possible paths from the root to the leaves. In this case, it is two, since edge {3, 2} is traversed twice (Paths R->5 and R->6). The histogram representation for the string *S* is shown in Figure 2(b). With the histograms at hands, the similarity between two activities can be measured by comparing their corresponding histograms. The benefit of the Suffix-tree model is that how the activities are composed or structured can be easily seen from the suffix tree representation, and also by transforming the activities into a tree, standard graph-theoretic methods can be easily applied.

In [25], Zhang et al. are interested in analyzing the car trajectories extracted from a crossroad. Instead of looking at the spatial distribution of trajectories, they explore the temporal relationships and infer semantic meanings from the trajectories. They start out by dividing the entire scene into several semantic regions that could correspond to particular goals. All the cars are tracked and the collected trajectories are then segmented according to these regions. In each region, these segments are clustered in order to form atomic events in different regions. For each of these atomic events, a Hidden Markov Model (HMM) classifier can be built. Upon encountering an unseen trajectory, the system first segments it according to the predefined regions and then applies each classifier to find out the atomic events that constitute the trajectory. For example, a trajectory could be the result of the set of atomic events: {going in the left lane of the main road, turn left, going in the left lane of the side road}. With the HMM classifiers, any trajectory can be represented as a string. These strings are then used to train a stochastic grammar to uncover more complicated rules, such as, "Turn left from the side road to the main road". These grammar-based approaches benefit from the fact that a context-free grammar, similar to suffix-tree, is already studied extensively. Therefore, after defining and detecting the primitive events, the production rules that define higher-level activities can be learned efficiently.

There are several drawbacks to these trajectory-based approaches. First, the extracted features might be sensitive to different camera settings, which we will elaborate in the next section. Second, to model complex activities, primitive events usually need to be manually defined or labeled, but in many cases, it may not be clear which events should be considered primitives. Finally, the trajectories might not always be available due to occlusion or very few motion information exhibited by the target of interest.

3.3 Cross Scene Transfer

One common drawback of the tracking-based approaches is that, the extracted features would depend on the camera settings. For example, the object location is usually scene dependent as it will change if the camera is set up differently.

Therefore, most of the time when the camera is redeployed, the entire system will need to be retrained. In [19], Bose et al. study an extensive set of features for classifying people and cars in videos. They find out that some features are scene dependent, while others are scene invariant, i.e., these features can be used to train the system in one scene and test in another without degrading the performance much. Table 1 shows the partial lists for scene invariant and scene dependent features. Orientation is the direction of principal axis of the object. Note that it is a scene invariant feature since the vertical world direction projects to the vertical axis in the image for most camera setups, so the objects would have an almost constant orientation. Variation in area refers to the second derivative of the numbers of pixels as a function of time, normalized by the mean area. Percentage occupancy is the number of silhouette pixels divided by the area of the bounding box. The magnitude of object velocity is also mostly constant regardless of the camera settings. On the other hand, object area and aspect ratio, two commonly used features in vision systems are scene dependent, implying that most vision systems do not transfer well to new scenes. Based on this, they propose to first train a generic classifier using the scene invariant features only. When applied to a new scene, the classifier is adjusted using the scene dependent features. This way, the cost of retraining the system is reduced and the incorporation of the scene dependent features also boosts the performance of the classifier.

Table 1. Partial list of scene invariant and scene dependent features

Scene Invariant Features	Scene Dependent Features
Orientation	Area in pixels
Variation in area	Aspect ratio
Percentage Occupancy	X coordinate
Velocity magnitude	Y coordinate

3.4 Human Activity Recognition Based on Tracking

Other than the aforementioned surveillance systems, specifically for human action modeling, a variety of techniques that rely on tracking body parts (e.g., hands, arms, limbs, torsos, etc.) to classify human actions were also proposed [20, 21]. The classical review of [22] covers significant amount of work that falls into this category. Methods are also proposed [23] to distinguish activities under large viewpoint changes.

There are also research works focusing on hand activity recognition by tracking the human hands and estimating the hand poses [40, 41]. Although very promising results have been achieved, it is still hard to accurately detect and track body parts in complex environments due to cluttered background or self-occlusion. Also, the high degree of freedom of human bodies or hands makes the estimation of the model parameters difficult as well.

4 Activity Recognition Using State-Space Models

A range of methods based on state-space models now permeate the field of computer vision. These models have been widely applied for short-term action recognition and more complex behavior analysis, involving object interactions and activities at multiple levels of temporal granularity. Examples include Hidden Markov Model and its variations such as coupled HMMs [35], Layered HMMs [36], Stochastic Grammars [37], and Conditional Random Fields [38].

These models are typically used to model composite activities. Usually atomic activities are defined first and these models are applied over these atomic activities to classify more complicated ones. For example, as stated previously, Zhang et al. [25] first divide the scene from a traffic surveillance camera into different regions, based on which, vehicle trajectories are segmented. Then Stochastic Grammars are applied on the segmented trajectories in order to find the hidden semantic meanings.

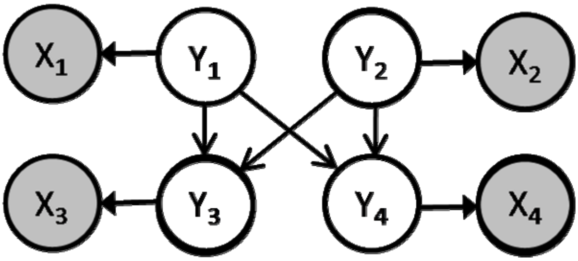


Fig. 3. Graphical representation of a coupled Hidden Markov Model. The shaded nodes represent observations, denoted by variables X_1 ~ X_4, while nodes Y_1 ~ Y_4 are state variables of interest to be inferred, for example, class labels.

In [35], Brand et al. propose a coupled HMM to model the causal interactions between activities. Conventionally, HMM is used to model a sequential activity. When there are two or more activities interacting with each other, HMM will not be able to capture the interaction. The reason is that it makes the assumption that the current state of one activity only depends on the previous state of its own, which implies independence between any two activities. Suppose we are to model two tennis players in game, HMM will not work well as the current state of one player is causally related with the previous state of the others. For example, in a tennis game, if you go up to the net, you will usually drive your opponent back and a weak serve will tend to bring your opponent forward. Figure 3 shows the graphical representation of coupled HMM. As we can see from the figure, the states of the two processes are intertwined and mutually dependent and the current state of one process not only depends on the previous state of its own but that of the other process. In [35], the coupled HMM is used to recognize different arm gestures in Tai Chi sports, where the arm movements are correlated. They report 94% accuracy compared to the 70% accuracy using HMM.

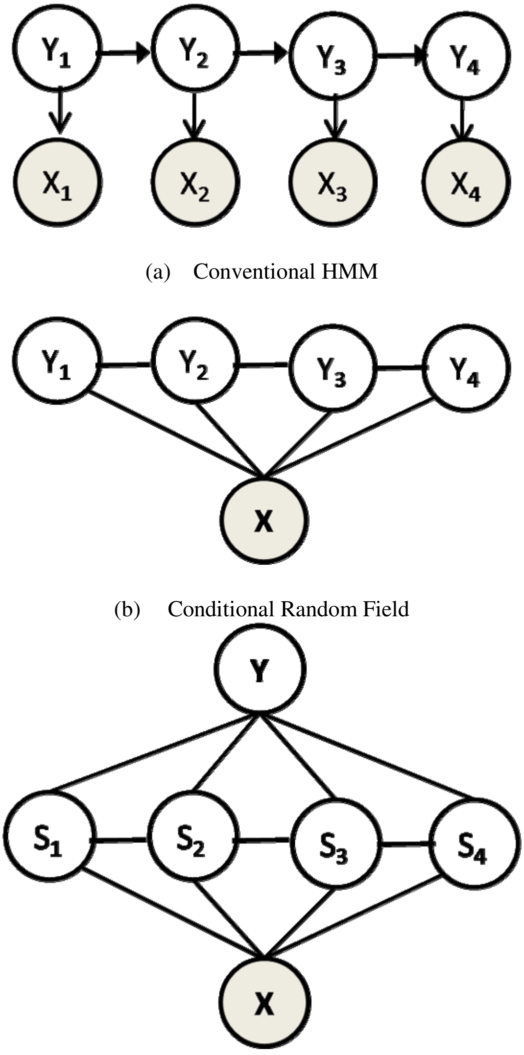
In [36], Oliver et al. try to identify different activities (phone conversation, presentation, face to face conversation, user present and engaged in some other activities, distant conversation, and nobody present). They use two layers of HMM's to recognize these activities. At the first layer, two banks of HMM's are trained. The first bank of HMM's detects different classes of sounds, such as human speech, phone ring, music, office noise, and ambient noise. The second bank of HMM's aims at detecting how many people there are in the office (nobody, one, or more). After classification, the HMM's at the first layer feed their classification results into the HMM on the second layer, which will make a final classification based on the sound and the video information and also the history of mouse and keyboard activities. The reported accuracy is 99.7% compared to 72.7% without the layering structure.

Similar to [25], Bobick et al. try to detect complex hand gestures using stochastic grammars. At the lower level, HMM's are trained to recognize atomic hand-movements, such as moving toward the right, the left, upward, or downward from both hands. Subsequently, the outputs of the HMM's are fused together to train a stochastic grammar in order to recognize a more complex activity, for example, right hand moving clockwise.

The aforementioned models are variants of HMM, a generative model assuming conditional independences between the observations, which makes modeling of long-range dependencies difficult. In [44], J. Lafferty et al. propose the Conditional Random Field (CRF), the discriminative counterpart of HMM, that directly maximizes the conditional distribution of the label variables, and thus, removes the independence assumptions. In [45], D. Vail et al. use CRF to identity which robot is the seeker, i.e., the one that moves the soccer to its closest teammate or the goal position, in a multi-robot soccer game using features such as relative positions and velocities. Similarly, C. Sminchisescu et al. [46] apply CRFs to classify human motion activities (i.e. walking, jumping, etc.) and their model can also discriminate subtle motion styles such as normal walk and wander walk.

In [47], A. Quattoni et al. propose the Hidden State Conditional Random Field (HCRF) model, an extension of CRF, by introducing a layer of hidden variables for object recognition, where the hidden variables model different parts of the object. The authors also apply this model to recognize different head gestures including head nodding, head shaking, and head junking and arm gestures as well [38, 48].

In [49], Y. Wang et al. apply HCRF to recognize human activities. Specifically, they combine HCRF with the constellation of local features, similar to [6] for human activities and achieve good performances on standard datasets. The limitation of HCRF is that the training involves summing over all the possible labeling of the latent variables, which could be hard if the hidden variables form



(a) Conventional HMM

(b) Conditional Random Field

(c) Hidden state Conditional Random Field

Fig. 4. Graphical representations of (a) HMM, (b) CRF, and (c) HCRF. The shaded nodes represent observation, denoted by variables X and X_i, while nodes Y and Y_i are class labels and S_i are latent variables. For CRF and HCRF, since there could be interactions between the observations, X_1 - X_4 are merged into a vector variable X. For HCRF, since the class label at each time stamp could be mutually dependent, variables Y_1-Y_4 are merged into a vector variable Y.

complicated structures. To tackle this problem, they propose Max-Margin HCRF [50] by incorporating structure learning during the training process in order to learn the optimal structures of the latent variables. Figure 4 shows the graphical representations of HMM, CRF, and HCRF respectively.

By modeling the complicated interactions among the observations and latent variables, these models are capable of recognizing complex activities. However, this does come with a price. These models typically fall under the category of undirected graphical models, which are harder to learn as the variables are coupled together. Besides, the majority of these methods are supervised, requiring manual labeling of video clips, and when the state space is large, the estimation of many parameters makes the learning process more difficult.

5 Activity Recognition Using Bag of Words Models

Different from previous methods based on trajectories or template matching, bag of words models have recently become very popular and have shown great promise in action recognition. The idea is rooted from natural language processing, which assumes a document can be suitably represented as a histogram of its constituent words. For example, suppose we are given a document and would like to find out what category it belongs to. As Figure 5 illustrates, we can first represent the document as a histogram of the words in it, then by looking at the word distribution, we can infer that this document is more likely an article about Meteorology than Machine Learning since the counts of the science related words, e.g., Earth and Moon, are high.

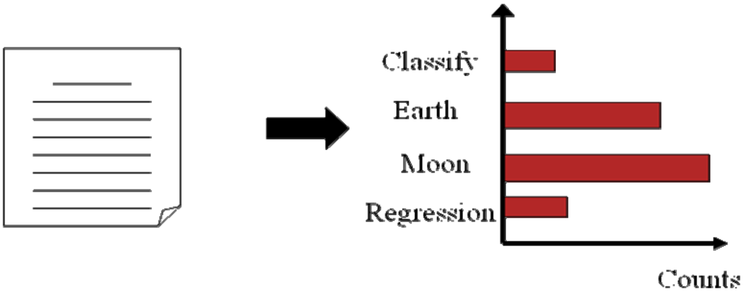


Fig. 5. We can represent a document as a histogram of its constituent words. By looking at the word distribution of the document, we can have a rough idea about which category the document belongs to.

5.1 Extraction of the "Bag of Words" Features

When applied to computer vision, these approaches in general require the extraction of sparse space-time interest points [3, 4] from the video sequences. In [4], Laptev and Lindeberg detect these interest points by extending the 2D Harris corner detector into the *t* dimension. Gradients are first found along *x*, *y*, and *time* axes and then the second moment matrices are computed at each pixel in order to determine if it is a spatial temporal corner.

In [3], Dollar et al. propose to detect denser interest points using a temporal Gabor filter. For a video sequence with pixel values *I(x, y, t)*, separable linear filters are applied to the video in order to obtain the response function as follows:

$$R=(I*g*h_{ev})^2+(I*g*h_{odd})^2 \tag{2}$$

where * indicates the convolution, $g(x, y, \sigma)$ is the 2D Gaussian smoothing kernel applied only along the spatial dimensions (x, y), and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as:

$$h_{ev}(t,\tau,f)=\cos(2\pi ft)e^{-t^2/\tau^2} \tag{3}$$

$$h_{od}(t,\tau,f)=\sin(2\pi ft)e^{-t^2/\tau^2} \tag{4}$$

The two parameters σ and *f* correspond to the spatial and temporal scales of the detector respectively. The frequency of the harmonic functions is given by *f*. A region undergoing a complex, non-translational motion induces a strong response [3]. At the points where strong motion occurs, spatial temporal volumes or cuboids can be extracted, which are also called visual words space-time interest points. Figure 6 shows the detected interest points using the approach in [3] for a few sequences from the KTH dataset [5]. Considering Figure 6(c) as an example, we can see that the interest points occur at places around the arms, where the periodic motion induces strong responses.

After the extraction of the interest points, application dependent features such as local optical flow or 3D gradient can be computed from these 3D volumes and then concatenated to form feature vectors. Usually Principal Component Analysis (PCA) is followed in order to reduce the feature dimension. Suppose the interest points with computed features are extracted from a set of video sequences. These extracted interest points can be used to build a codebook or dictionary using the K-means clustering algorithm. Later on, each video sequence can then be described by a histogram of words from the codebook, and finally, either discriminative or generative models can be applied for categorizing the activities.

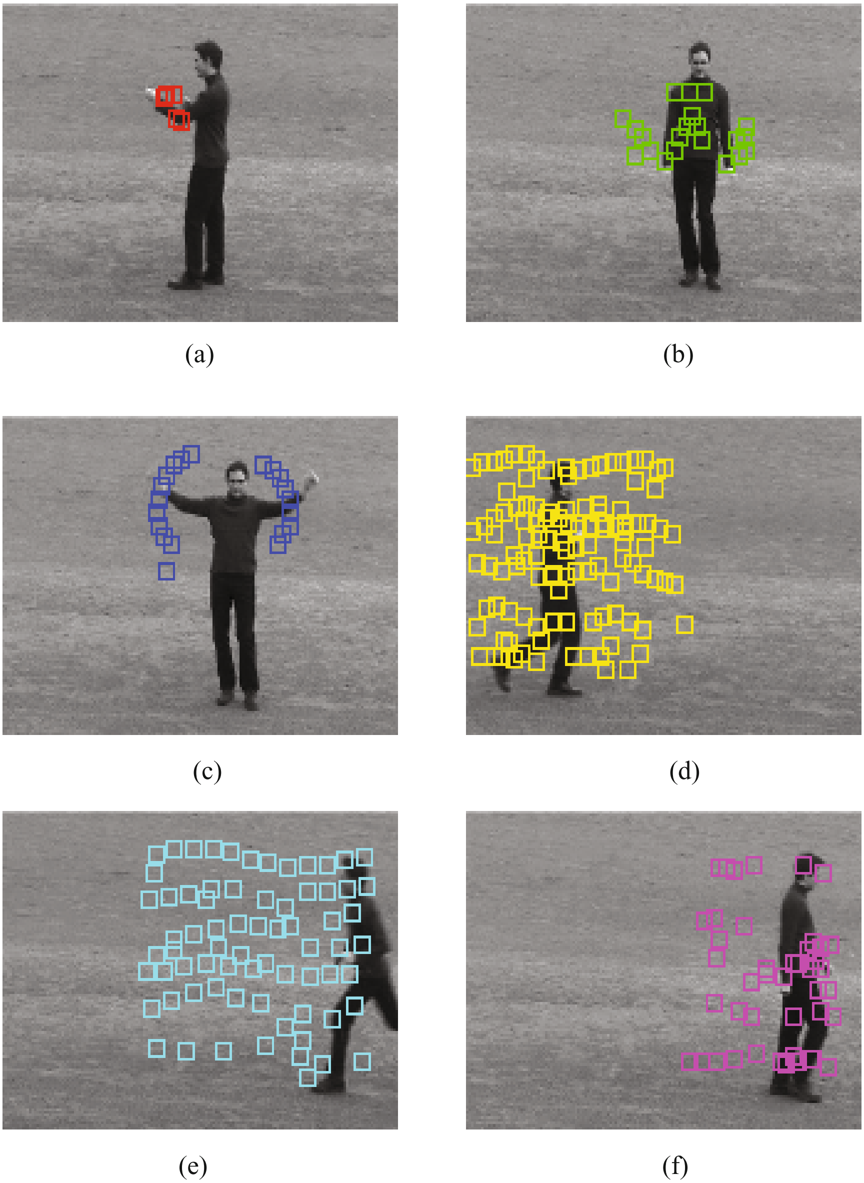


Fig. 6. Sample sequences with detected interest points for the KTH dataset. From (a) to (f), the activities are boxing, handclapping, hand-waving, jogging, running, and walking. The squares represent the detected interest points.

5.2 Activity Recognition Using the "Bag of Words" Features

By representing each video sequence as a histogram of visual words, conventional discriminative models can be applied in order to determine the class label of a testing video sequence. In this setting, the histograms are treated as the input vector variable *x*, and the class labels *y* is the output variable. The function $y = f(x)$ is to be inferred from a training set (x, y_i), $i=1...N$. In [3, 5], highly discriminative results are obtained using SVM classifiers based on these descriptors under a supervised learning framework.

Recently, Niebles et al. [6] enhanced this approach by proposing a novel model characterized as a constellation of the "Bag of words" image features, which encodes both the shape and appearance of the actor for recognizing human activities in each video frame. The proposed model is illustrated in Figure 7.

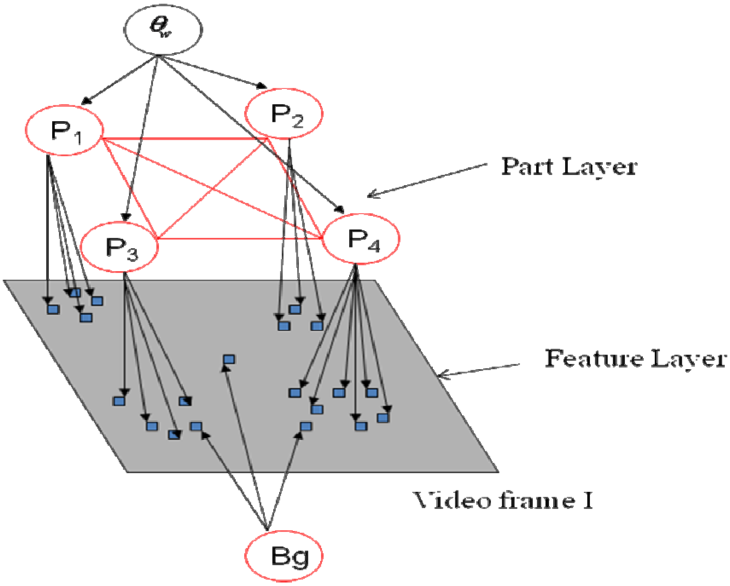


Fig. 7. The hierarchical model regenerated from [6]. The red nodes constitute the part layer, which is a higher layer while the feature layer is composed of the detected Bag of Words features.

Let the model parameter for a particular human activity be θ_w and assume the human body is composed of *P* parts, for example, $P = 4$ in Figure 7. Variable $P_1 \ldots P_p$ represent the locations of the parts of the body. Given a video frame *I*, with *W* detected visual words $\mathbf{w} = \{w_i\}$, $i=1\ldots W$, instead of directly using the visual words to recognize human activity, the model imposes a structural constraint on top of the feature-layer by introducing a part-layer composed of variables $P_1 \ldots P_p$, and a variable Bg representing the background. The joint distribution of

$P_1 \ldots P_p$ is modeled as a multivariate Gaussian, implying that the body parts should not deviate from each other too much. In addition, a *1* x *W* vector variable **m** is introduced to represent the assignment of each of the *W* detected visual words, that is, it is a vector of *W* elements, each taking integer values in the range [*0-P*]. Each visual word can be assigned to any of the *P* parts or the background. Suppose the *jth* detected visual word comes from the second body-part, then we can set **m**(*j*)=2. Learning is done using Expectation Maximization (EM) to infer the model parameters and the hidden variables.

We can consider the problem of recognizing human activities using the "Bag of words" features as one that assigns each of the visual words to one of the body parts. Hence, the benefit of this two-layer model over a naïve "Bag of words" model is that by introducing the structural constraint, it keeps in mind the domain knowledge, that is, how a human should look like, when solving the assignment problem. With this constraint, lots of invalid human body configurations can be filtered out quickly. Without the constraint, it is possible that the model would assign features induced by the hands to legs, and vice versa. The fact is also confirmed by the better performance reported in [6].

Other popular methodologies include using language models for unsupervised activity recognition. Due to the success in the image domain [26, 27] for object recognition, these methods combined with the bag of words model have also been proposed to solve the activity classification problems. They have become very popular as they could achieve excellent performance in standard datasets [28] and long surveillance videos [29, 30]. Generally, these unsupervised algorithms extract spatial temporal feature descriptors and then use document topic models such as Probabilistic Latent Semantic Analysis (pLSA) [31], Latent Dirichilet Allocation (LDA) [32], or HDP [33] to discover latent topics [34, 29, 30]. pLSA and LDA are two commonly used models for unsupervised learning. We briefly review them next.

Figure 8 illustrates the graphical models of pLSA and LDA respectively. They are both generative models. For pLSA, a word can be generated as follows: select a document d_i with probability P(d_i), pick a latent class z_k with probability P(z_k | d_i), and finally generate word w_j with probability P(w_j | z_k). The parameters, that is, how words are distributed given a topic *z*, can be learned using EM. pLSA assumes that one document contains only one topic, therefore, a document *d* is classified into topic *k* if P(z_k | *d*) is the largest. LDA assumes that one document can contain a mixture of topics. The mixture weight is controlled by the variable π whose distribution is determined by α. Given a topic *z*, the word distribution is controlled by β. Learning for LDA is harder but can still be done using sampling techniques.

In [34], Niebles et al. extract spatial-temporal interest points and use a generative model based on pLSA to cluster activities. In [30], Wang et al. try to recognize activities in a crossroad scene. The goal is to model the interactions between cars and pedestrians without relying on tracking. For each video, they first detect the moving pixels simply by computing the intensity difference between two successive frames on a pixel basis. If the difference at a pixel is above a threshold, the pixel is detected as a moving pixel. The motion direction at each moving pixel is obtained by computing the optical flow. These result in two features, position and direction, for each moving pixel. For each moving pixel, its position and direction are then quantized into different locations and directions. Then, they treat each video clip as a document and a moving pixel a word for word document analysis, and then use LDA to cluster the activities and find out interactions in surveillance videos. They report good performance in identifying different interactions between cars and pedestrians in the crossroad.

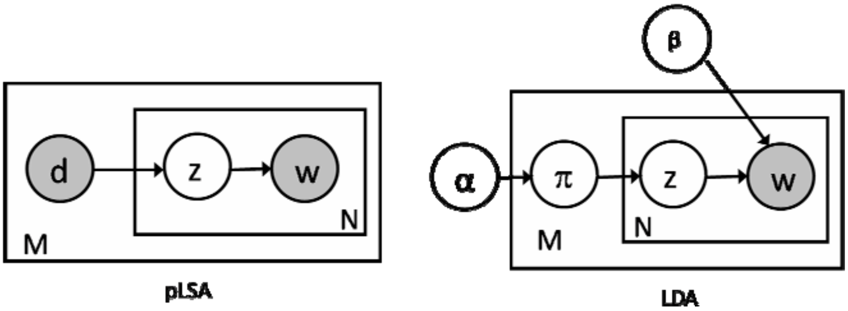


Fig. 8. The graphical model for pLSA and LDA. The shaded node represents observed variables while the white nodes are hidden variables. The enclosing rectangle represents multiple copies of the model, i.e, M documents, N dictionary words. For pLSA, w is the word variable, z is the topic variable, and d is the document variable. For LDA, w is the word variable, z is the topic variable, β controls how the words are distributed given a topic, π is the weight of each topic z and α controls how π is distributed.

The benefit of using the bag of words models is the easiness of representing the possible activities and the application of methods from the Natural Language Processing community. Although they achieve excellent results in real world video data, they could be further improved. Since they are borrowed from the language processing community, which usually represent documents as histograms of words, when applied to the image or video domain, they usually do not consider the spatial-temporal relationships among "visual words" unless the relationships are represented explicitly [29]. To address the problem, more recently, Savarese et al. [28] use spatial temporal correlations to encode flexible long-range temporal information into the local features. The other critic is the size of the codebook. As mentioned previously, these methods require vector-quantizing the local interest points into video words using the K-means algorithm, which clusters the cuboids based on their feature similarity. It has been noted that the size of the codebook affects the performance and that the optimal codebook size is often determined through experiments. In [51], J. Liu et al. tackle the two problems at once by using the information maximization clustering in order to both determine the optimal size of codebook and incorporate the correlogram to capture the spatial temporal relationships among the visual words. In [43], the authors also try to improve the performance by incorporating features extracted from static frames.

6 Datasets and Performance Comparison

In order to obtain a fair performance evaluation of different approaches, it is desirable to use a standard test dataset. However, although many approaches have been proposed, not everyone reports the performance on the same dataset. In this section, we introduce the currently available standard datasets and summarize the performance of some of the methods we reviewed.

6.1 Standard Datasets

Currently, two standard datasets for human activity recognition are widely used. The KTH dataset [5] is by far the largest dataset, containing six types of human activities (walking, jogging, running, boxing, hand waving, and hand clapping), which are performed by 25 actors in four different scenarios, resulting in 600 sequences, each with a spatial resolution of *160* x *120* pixels and a frame-rate of 25 frames per second. Each sequence is about 15 seconds long.

The Weizmann dataset [9] contains 10 types of activities (walking, running, jumping, gallop sideways, bending, one hand waving, two hand waving, jumping in place, jumping jack, and skipping), each performed by 9 actors, resulting in 90 video sequences, each with a spatial resolution of *180* x *144* pixels and a frame-rate of 50 frames per second. Each sequence is about three seconds long.

More recent efforts to publish new dataset include the TREC Video Retrieval Evaluation Project [42] that released real world surveillance videos taken from the Gatwick airport in London. Also, J. Liu et al. [43] collected 1168 video sequences from Youtube that cover 11 kinds of sports activities including basketball shooting, volleyball spiking, trampoline jumping, and so on. These videos are mostly low resolution and differ in many aspects. For example, some are taken with shaky cameras, while others are not. The object scale, viewpoint, background cleanness and lighting condition are all different across the videos. Each sequence is roughly two to three seconds long with a frame-rate of 25 frames per second. In [53], the authors collected a set of 12 activities from Hollywood movies to study how context information, i.e., where the activities occur, can help in recognizing the activities.

6.2 Performance Comparison

Here, we report the performance of a few approaches reviewed on the KTH and the Weizmann datasets. For the approaches that use spatial temporal interest points [3], the parameters for the detector are set as follows: $\sigma = 2$ and $\tau = 2.5$ for the KTH dataset [5], and $\sigma = 1.2$ and $\tau = 1.2$ for the Weizmann dataset [9]. In all cases, $f = 4/\tau$. PCA is used to reduce the dimension of the feature vectors to 100. Table 2 summarizes the performance of different methods.

Table 2. Performance reported for different methods on the KTH dataset [5] and the Weizmann dataset [9]. Numbers are reported on a per video sequence basis. * represents those reported on a per feature basis. ^{x} represents numbers on a per frame basis.

Methods	Type	Features	Models	KTH	Weizmann
J. Liu et al. [43]	supervised	ST [3]+image	Decision Tree	91.8%	
J. Liu et al. [51]	supervised	ST [3]+correlogram	SVM	94.1%	
Y. Wang et al. [50]	supervised	Motion features [8]	MMHCRF	92%, 78%*	100%, 93%*
Y. Wang et al. [49]	supervised	Motion features [8]	HCRF	87%, 67%*	97.2%, 90.3%*
Savarese et al. [28]	unsupervised	ST [3]+correlogram	pLSA	86.8%	N.A
Nieble et al. [34]	unsupervised	ST [3]	pLSA	83.3%	90%
Dollar et al. [3]	supervised	ST [3]	SVM	81.17%	N.A
Schuldt et al. [5]	supervised	ST [5]	SVM	71.72%	N.A
Ke et al. [1]	supervised	Template based	Binary classifier	63%	N.A
Blank et al. [9]	supervised	Template based	N.A	N.A	97.9%*
Nieble et al. [6]	supervised	Bag of words	Graphical Model	N.A	72.8% 55%^{x}

This table summarizes the methods and features a few state-of-the-art approaches adopted and the performance numbers they reported on both datasets. However, it is still hard to say how well these methods perform and what merits these approaches possess. The complication arises from the following facts. To start with, the models are trained differently and thus the training sets and the testing sets are generally different among different methods. For example, the training samples could be different depending on how many percentages of the data are withheld as training data.

If the difference in training data were the only factor, we might still have a rough feeling about how well each method is performed. A more significant factor that makes it hard to fairly evaluate these methods is the difference in how they are evaluated. For example, the number reported by Blank et al. [9] on the Weizmann dataset [9] is actually one reported on a per feature basis, different from others on a per sequence basis. In [9], the authors reported the performance on a per sequence basis, however, it is to evaluate the performance of clustering, which is different from others performing classification tasks. Also, Nieble et al. [6] reported a lower performance on a per frame basis.

To cite another example, the performance reported by Ke et al. [1] on the KTH dataset [5] is lower compared to those reported by others. However, what the authors are evaluating is the capability to search for a certain activity in a video sequence, which is harder than simply classifying a video sequence into one of the categories.

The third factor arises from the inadequacy of the standard datasets. Most activities in the two available standard datasets are stationary, that is, in most cases, the actors are standing still performing different activities, which makes it impossible to evaluate motion-based or tracking-based approaches using these datasets.

To summarize, although many methods have been proposed to solve the problem of automatic activity recognition, many of them report their performance on customized datasets. This makes it difficult to fairly and extensively evaluate the performance of different approaches. We hope that the problem will be alleviated in the future if more standard and representative datasets are published and more research work could report their performance on these standard datasets.

7 Conclusions

The concerns of public security coupled with the advance of video and storage technology has led to a great need for an automatic surveillance system, and therefore, automatic video activity recognition is becoming an important research topic. It extends the conventional object recognition problems in static images into the video domain by introducing the extra time dimension. With the extra dimension, more information is available. This chapter reviews what previous research regarding activity recognition has paid off and comments on the pros and cons of each genre.

To conclude, these approaches can be categorized into the following classes:

Template based approaches search for the best match of the testing sequence in the database. The benefit of these approaches is that no model training is required but the choice of extracted feature and similarity measure is crucial in achieving better performance.

More complicated activities or interactions among objects can be modeled using graphical models, such as Markov Random Field or Dynamic Bayesian Network or their variants. In general, Markov Random Field is able to handle more complicated activities but it is also more difficult to train as the variables cannot be easily decoupled.

Approaches based on tracking extract statistics such as the trajectories from the tracker for activity recognition. These trajectories can be clustered to find out regular path patterns. For modeling more complex activities, a higher-level representation is required to impose on these trajectories.

The “Bag of words” model is becoming more popular. Different supervised and unsupervised models are developed based on it and achieve good performance. The main critic is its ignorance of relationship among the extracted features but increasing research is on going to bring feature relationships into the model.

While the field has seen rapid progress, challenges still remain. It is still not clear how we can model the complex interactions among multiple objects or joint activities. In addition, occlusion handling is still hard but usually crucial in many applications. For instance, in hand activity recognition, it will be immensely useful to know the configuration of the hand, but in reality, the hand itself could be self-occluded, creating an impasse for correct estimation.

Finally, more work is needed to address these challenges and make the current approaches more robust in order to further the aim of an effective automatic surveillance system.

References

1. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: IEEE International Conference on Computer Vision (2005)
2. Krueger, V., Kragic, D., Ude, A., Geib, C.: Meaning of action: a review on action recognition and mapping. Int. Journal on Advanced Robotics, Special issue on Imitative Robotics 21, 1473–1501 (2007)
3. Dollar, P., Rabaud, V., Cottrellm, G., Belongie, S.: Behavior recognition via sparse spatiotemporal features. In: PETS (2005)
4. Laptev, I., Lindeberg, T.: Space-time interest points. In: IEEE International Conference on Computer Vision (2003)
5. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: Proceedings of the International Conference on Pattern Recognition (2004)
6. Niebles, J.C., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
7. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transaction on Pattern Analysis and Machine Intelligence 23, 647–666 (2001)
8. Efros, A., Berg, E., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision (2003)
9. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space time shapes. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
10. Shechtman, E., Irani, M.: Space-time behavior based correlation. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
11. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach: a spatiotemporal maximum average correlation height filter for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
12. Mahalanobis, A., Vijaya Kumar, B.V.K., Sims, S.R.F., Epperson, J.: Unconstrained correlation filters. Applied Optics 33, 3751–3759 (1994)
13. Ebling, J., Scheuermann, G.: Clifford Fourier transform on vector fields. IEEE Transactions on Visualization and Computer Graphics 11(4), 469–479 (2005)
14. Makris, D., Ellis, T.: Path detection in video surveillance. In: Image and Visual Computing (2002)
15. Hu, W., Xie, D., Tan, T.: A Hierarchical Self-Organizing Approach for Learning the Patterns of Motion Trajectories. IEEE Transaction on Neural Network 15(1) (January 2004)
16. Fu, Z., Hu, W., Tan, T.: Similarity based vehicle trajectory clustering and anomaly detection. In: IEEE Int. Conf. Image Processing, vol. 2, pp. 602–605 (2005)
17. Junejo, I.N., Foroosh, H.: Trajectory Rectification and Path Modeling for Video Surveillance. In: International Conference on Computer Vision (2007)
18. Hamid, R., Maddi, S., Bobick, A., Essa, I.: Structure from Statistics - Un-supervised Activity Analysis using Suffix Trees. In: International Conference on Computer Vision (2007)
19. Bose, B., Grimson, E.: Improving Object Classification in Far-Field Video. In: IEEE Conference on Computer Vision and Pattern Recognition (2004)
20. Ramanan, D., Forsyth, D.A.: Automatic annotation of everyday movements. In: Neural Information Processing Systems (2003)
21. Fanti, C., Zelnik-Manor, L., Perona, P.: Hybrid models for human motion recognition. In: International Conference on Computer Vision (2005)

22. Gavrila, D.: The visual analysis of human movement: a survey. Computer Vision and Image Understanding 73, 82–98 (1999)
23. Ikizler, N., Forsyth, D.: Searching video for complex activities with finite state models. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
24. Ukkonen, E.: Constructing suffix trees on-line in linear time. In: Proc. Information Processing 92. IFIP Transactions A-12, vol. 1, pp. 484–492 (1994)
25. Zhang, Z., Huang, K., Tan, T., Wang, L.: Trajectory Series Analysis based Event Rule Induction for Visual Surveillance. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
26. Li, L.-J., Fei-Fei, L.: What, where and who? Classifying event by scene and object recog-nition. In: IEEE International Conference on Computer Vision (2007)
27. Cao, L., Fei-Fei, L.: Spatially coherent latent topic model for concurrent object segmentation and classification. In: IEEE International Conference on Computer Vision (2007)
28. Savarese, S., Pozo, A.D., Niebles, J., Fei-Fei, L.: Spatial-temporal correlations for unsupervised action classification. In: IEEE Workshop on Motion and Video Computing (2008)
29. Wang, X., Ma, K.T., Ng, G.W., Grimson, E.: Trajectory analysis and semantic region modeling using a nonparametric Bayesian model. In: IEEE Conference on Computer Vision and Patter Recognition (2008)
30. Wang, X., Ma, X., Grimson, E.: Unsupervised activity perception by hierarchical Bayesian models. In: IEEE Conference on Computer Vision and Patter Recognition (2007)
31. Hofmann, T.: Probabilistic latent semantic analysis. In: Conference on Uncertainty in Artificial Intelligence (1999)
32. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
33. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical dirichlet process. Journal of the American Statistical Association (2006)
34. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision 79, 299–318 (2008)
35. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (1997)
36. Olivera, N., Gargb, A., Horvitz, E.: Layered representations for learning and inferring office activity from multiple sensory channels. Computer Vision and Image Understanding 96, 163–180 (2004)
37. Bobick, A.F., Ivanov, Y.A.: Action recognition using probabilistic parsing. In: IEEE Conference on Computer Vision and Pattern Recognition (1998)
38. Quattoni, A., Wang, S., Morency, L.-P., Collins, M., Darrell, T.: Hidden-state conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 1848–1852 (2007)
39. Wang, Y., Jiang, H., Drew, M.S., Li, Z., Mori, G.: Unsupervised discovery of action classes. In: IEEE Conference on Computer Vision and Pattern Recognition (2006)
40. Athitsos, V., Sclaroff, S.: Estimating 3D Hand Pose from a Cluttered Image. In: IEEE Conference on Computer Vision and Pattern Recognition (2003)
41. Sudderth, E.B., Mandel, M.I., Freeman, W.T., Willsky, A.S.: Visual Hand Tracking Using Nonparametric Belief Propagation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop on Generative Model Based Vision (2004)

42. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, MIR (2006)
43. Liu, J., Luo, J., Shah, M.: Recognizing Realistic Actions from Videos in the Wild. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
44. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: IEEE International Conference on Machine Learning (2001)
45. Vail, D., Veloso, M., Lafferty, J.: Conditional Random Fields for Activity Recognition. In: ACM International Conference on Autonomous Agents and Multiagent Systems (2007)
46. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Conditional models for contextual human motion recognition. In: IEEE International Conference on Computer Vision (2005)
47. Quattoni, A., Collins, M., Darrell, T.: Conditional random fields for object recognition. In: Advances in Neural Information Processing Systems (2005)
48. Wang, S., Quattoni, A., Morency, L.-P., Demirdjian, D., Darrell, T.: Hidden Conditional Random Fields for Gesture Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2006)
49. Wang, Y., Mori, G.: Learning a discriminative hidden part model for human action recognition. In: Advances in Neural Information Processing Systems (2008)
50. Wang, Y., Mori, G.: Max-Margin Hidden Conditional Random Fields for Human Action Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
51. Liu, J., Shah, M.: Learning Human Actions via Information Maximization. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
52. Bobick, A.F.: Movement, activity, and action: The role of knowledge in the perception of motion. Philosoph. Trans. Roy. Soc. Lond. B 352, 1257–1265 (1997)
53. Marszalek, M., Laptev, I., Schmid, C.: Actions in Context. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
54. Liu, J., Luo, J., Shah, M.: Recognizing Realistic Actions from Videos in the Wild. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)

Robust Spatio-Temporal Features for Human Action Recognition

Riccardo Mattivi[1] and Ling Shao[2]

[1] The University of Trento
rmattivi@disi.unitn.it
[2] The University of Sheffield
ling.shao@sheffield.ac.uk

Abstract. In this chapter we describe and evaluate two recent feature detectors and descriptors used in the context of action recognition: 3D SIFT and 3D SURF. We first give an introduction to the algorithms in the 2D domain, named SIFT and SURF. For each method, an explanation of the theory upon which they are based is given and a comparison of the different approaches is shown. Then, we describe the extension of the 2D methods SIFT and SURF into the temporal domain, known as 3D SIFT and 3D SURF. The similarities and differences for both methods are emphasized. As a comparison of the 3D methods, we evaluate the performance of 3D SURF and 3D SIFT in the field of Human Action Recognition. Our results have shown similar accuracy performance, but a greater efficiency for 3D SURF approach compared with 3D SIFT.

1 Introduction

In the 2D domain, the search for interest points in images has many applications, such as image registration, camera calibration, image retrieval, object recognition, image stitching etc. The task of extracting such 'interest points' can be divided into two main parts: feature detection and feature description.

In the feature detection part, the interest points are detected at distinctive locations in the image, such as at edges, corners, blobs or other locations. The detector of such interest points should be able to find the same interest point under different viewing conditions, such as different viewpoints, illumination changes, contrast etc. This characteristic is known as repeatability.

In the feature description part, the area around the interest point is described as a feature vector. The descriptor has to be distinctive, discriminative and robust against noise, geometric and photometric deformations. In applications such as camera calibration and image registration, one single feature should be correctly matched with high probability against a large number of features.

In literature, a high number of detectors and descriptors in the 2D domain has been proposed, e.g. [18, 21, 20], and a detailed comparison and evaluation have been done [23, 22].

In this chapter we also focus the attention on the 2D feature detectors and descriptors SIFT (Scale-Invariant Feature Transform, [18]) and SURF (Speeded-Up Robust

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 351–367.
springerlink.com

Features, [3]). SIFT approach is currently widely used and it has been proved to give better performances compared with other extraction and description methods [22]; The SURF algorithm approximates or slightly outperform SIFT and it is shown to be computationally more efficient [3]. The applications of SIFT range from object recognition [18] to image stitching and recognizing panoramas [5], from robotic mapping and navigation [27] to augmented reality [9]. SURF's applications are similar with SIFT, such as object recognition [3, 24], 3D reconstruction [4] and, being faster than SIFT, the SURF descriptor can also be used for real time application as visual SLAM [30].

The theoretic background in the 2D domain for SIFT and SURF is required to better explain and understand the extensions of these descriptors into the temporal domain: 3D SIFT [26] and 3D SURF [28]. The aim of these methods is to detect and describe 'space-time interest point' (STIP) in video sequences. STIPs are, at present, used for action recognition [14, 7, 15, 28, 26], content based video copy detection [29] and also for biomedical applications [6, 1].

The aim of this chapter is (1) to explain the theoretic background for both SIFT and SURF detectors and descriptors in both 2D and 3D domain, (2) to compare and evaluate 3D SIFT and 3D SURF in the field of Human Action Recognition.

In order to clarify between the names used in the detection part and description part of each method, in the following we refer to the 2D SIFT detection part as Difference of Gaussian (DoG) and to the SURF detection part as 2D Hessian. In the 3D domain, the detection part of 3D SIFT is called Space-Time DoG and the name 3D Hessian stands for the detection part of 3D SURF. We refer as SIFT, SURF, 3D SIFT and 3D SURF for the description part of each method. In the paper of Willems et al. [28], the developed method is named 'An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector'; we are here calling this method as 3D SURF because of simplicity and to enphasize the evolution of SURF into the 3D domain.

The chapter is structured as follow: in Section 2 the feature detection part for both 2D and 3D approaches is explained, while the feature description of these techniques is shown in Section 3. In Section 4 we explain the methodology adopted for our experiments in human action recognition and in Section 5 the results are shown. Section 6 concludes the chapter.

2 Feature Detection

The detection of interest points has been widely investigated in literature. Harris [10] proposed a corner detector, based on the eigenvalues of the second moment matrix. Later, Lindeberg [17] has introduced the concept of automatic scale selection: the interest points are detected at their characteristic scale. Mikolajczyk et al. [20] further developed interest point detection algorithms with Harris-Laplace and Hessian-Laplace feature detectors. Kadir and Brady [12] proposed another feature detector based on the maximization of the entropy and Matas et al. [19] detect interest points on maximally stable extremal regions. Lowe [18] introduced SIFT detector (and also descriptor) which is based on local maxima or minima of Difference of Gaussians (DoG) filter. Recently, Bay et al. [3] developed another interest point detector which relies on a Hessian matrix-based measure and on the approximation of second order Gaussian derivatives with integral images and box filters.

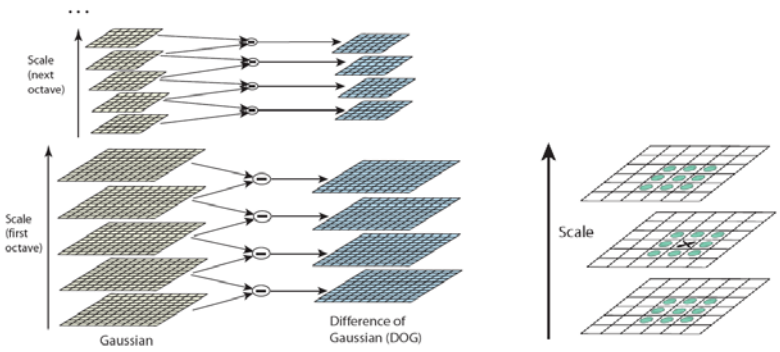
In this sub-section, we explain the theoretic concepts for finding interest points in an image using SIFT and SURF detection methods, called DoG and 2D Hessian, respectively. We also show here the evolution of these methods into the temporal domain and we named Space-Time DoG the detection part of 3D SIFT and 3D Hessian the detection part of 3D SURF.

2.1 DoG

Lowe [18] developed a method for detection and description of distinctive scale-invariant features named Scale Invariant Feature Transform (SIFT). These features are shown to be invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint. This fact implies that a single feature can be correctly matched with high probability against a large database of features (e.g. in object and scene recognition applications [18, 22]).

The major stages of computation used to generate the set of image features are (1) scale-space extrema detection, (2) interest point localization, (3) orientation assignment, (4) interest point description. We here take into consideration only the first steps of finding such features: the computation of a Difference of Gaussian (DoG) and the detection of local maxima and minima. Later steps will be explained in sub-section 3.1.

The first stage in the algorithm is to detect scales and locations that are repeatable under different viewing conditions, geometric deformations, addition to noise, etc. Lindeberg [16] has shown that the only possible scale-space kernel is the Gaussian function. Due to this concept, the scale space of an image is defined as a function $L(x,y,\sigma)$, which is the convolution of a variable-scale Gaussian, $G(x,y,\sigma)$, with an input image, $I(x,y)$:



(a) Difference of Gaussian for various octaves and scales

(b) Search for local maxima and minima [18]

Fig. 1. SIFT detection part

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \tag{1}$$

where $*$ is the convolution operation in x and y, and

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{2}$$

In order to detect stable interest point locations in the scale-space, Lowe [18] proposed to use scale-space extrema of the Difference of Gaussian function convolved with the image, $D(x,y,\sigma)$, which is the result from the difference of two nearby scales separated by a constant multiplicative factor k:

$$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y) \tag{3}$$

$$= L(x,y,k\sigma) - L(x,y,\sigma) \tag{4}$$

The Difference of Gaussian function is a close approximation to the scale-normalized Laplacian of Gaussian, $\sigma^2\nabla^2 G$, studied by Lindeberg [16]. Moreover, Mikolajczyk et al. [23] found that the maxima and minima of $\sigma^2\nabla^2 G$ produce the most stable image features compared to a range of other possible image functions. The relationship between D and $\sigma^2\nabla^2 G$ can be understood as

$$\sigma\nabla^2 G = \frac{\partial G}{\partial\sigma} \approx \frac{G(x,y,k\sigma) - G(x,y,\sigma)}{k\sigma - \sigma} \tag{5}$$

$$G(x,y,k\sigma) - G(x,y,\sigma) \approx (k-1)\sigma^2\nabla^2 G. \tag{6}$$

When the scales of DoG function differ by a constant factor k, the DoG already incorporates the σ^2 scale normalization, which is required for the scale-invariant Laplacian. The factor $(k-1)$ in the equation is a constant over all scales and it does not influence extrema location; the approximation error goes to zero as k goes to 1.

The efficient approach used by Lowe in the construction of $D(x,y,\sigma)$ is shown in Figure 1a. As can be seen, the scale spaces are implemented as an image pyramid (see also Figure 2a), where the images are repeatedly smoothed with a Gaussian filter and subsampled in order to achieve a higher level of the pyramid. In this implementation, each image is incrementally convolved with Gaussians to produce images separated by a constant factor k in scale space. Lowe chose to divide each octave of scale space into an integer number s of intervals, so that $k = 2^{1/s}$. The method produces $s+3$ images in the stack of blurred images for each octave so that the final extrema detection covers a complete octave. DoG images are computed subtracting adjacent image scales (see the right side of Figure 1a). Once a complete octave has been processed, the 2 images at the top of the stack are downsize by a factor of 2. The process is repeated for all octaves.

The local maxima and minima of $D(x,y,\sigma)$ are searched comparing the sample point to its eight neighbors in the current image and nine neighbors in the scale above and below, as illustrated in Figure 1b. A point is selected only if it is larger or smaller than all of these neighbors. Most of the sample points will be discarded during the first few checks and this helps in speeding up the algorithm. For more details about Lowe's implementation, please refer to [18].

2.2 2D Hessian

Speeded Up Robust Features (SURF) is a scale and rotation invariant interest point detector and descriptor used in the 2D domain and it has been developed by Bay et al. [3]. As shown by the authors, SURF approximates or slightly outperforms SIFT algorithm in term of repeatability, distinctiveness and robustness. Moreover, SURF detector and descriptor can be computed much faster. This performance is achieved by relying on integral images for image convolutions, on a fast Hessian detector and on a distribution-based descriptor.

We here describe the detector part of SURF, which relyes on a basic approximation of the Hessian-matrix, as the DoG detector used in SIFT is also an approximation of the Laplacian-based detector. Here, the SURF detector pushes the approximation even further: integral images and box type filters are used for convolution, greatly decreasing the computational time required. The Hessian matrix is used for two purposes: its determinant is a measure for detecting blob-like structures and for the selection of the characterstic scale.

Given a point in an image I at location (x,y), the Hessian matrix $H(x,y,\sigma)$ at scale σ is defined as follows

$$H(x,y,\sigma) = \begin{pmatrix} L_{xx}(x,y,\sigma) & L_{xy}(x,y,\sigma) \\ L_{xy}(x,y,\sigma) & L_{yy}(x,y,\sigma) \end{pmatrix} \tag{7}$$

where $L_{xx}(x,y,\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2}g(\sigma)$ with the image I at location (x,y) and similarly for $L_{xy}(x,y,\sigma)$ and $L_{yy}(x,y,\sigma)$.

As previously said in the DoG sub-section, Gaussians filters are optimal for scale-space analysis, but in practice the real filters have to be discretised and cropped. The authors pushed the approximation for the Hessian matrix further with box filters. These approximate the second order Gaussian derivatives, denoted as D_{xx}, D_{yy} and D_{xy}, and can be evaluated at a very low computational cost using integral images; the calculation time is independent of the filter size. As a recall, an integral image at a location (x,y) holds the sum of all pixels of the original image in the rectangular region spanned by $(0,0)-(x,y)$.

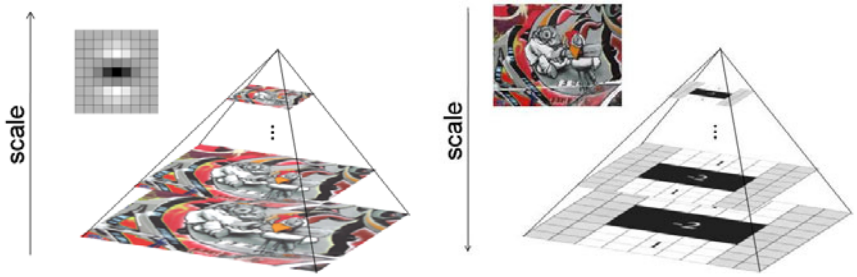
The determinant of the Hessian matrix is approximated as

$$det(H_{approx}) = D_{xx}D_{yy} - (wD_{xy})^2 \tag{8}$$

where the relative weight w is set to 0.9. The approximated determinant of the Hessian represents the blob response in the image at location (x,y) and these responses are stored in a response map over different scales since the interest points are detected at different scales.

The scale spaces are implemented as an image pyramid and divided into octaves, as it has been done in the SIFT detector (see Figure 1a). In Lowe's approach, the filter size is kept constant and the images are sequentially convolved and subsampled (as explained in sub-section 2.1); in Bay's approach, the use of box filters and integral images permits not to apply the same filter to the output of a previously filtered layer, but box filters of different sizes can be applied at exactly the same speed directly on the original image and subsampled images. The faster implementation is done by up-scaling the filter size instead of iteratively reducing the image size. In Figure 2b the approach used by SURF is shown, compared with the SIFT approach in Figure 2a.

Interest points are expressed as maxima of the determinant of the Hessian matrix and then localized in the image and over scales using a non-maximum suppression algorithm in a 3×3×3 neighborhood, as it is done in the SIFT algorithm (see Figure 1b).



(a) SIFT **(b)** SURF

Fig. 2. SIFT Vs. SURF image filtering

2.3 Space-Time DoG (3D)

Space-Time Difference of Gaussians is an extension of the SIFT feature detector into the 3D domain. The purpose of the detector is to find stable interest point locations not only in scale-space, but also in time and such features are called Space-Time Interest Points (STIP). This method has been applied for biomedical image processing purposes [6, 1], where the third dimension is other layers of 3D human Magnetic Resonance Imaging (MRI). For our purpose of human action recognition, the third dimension is represented by the time. The implementation and the theoretic concepts are very similar to the 2D case (see sub-section 2.1).

The scale-space-time of a video sequence is defined as a function $L(x,y,t,\sigma)$, which is the convolution of a variable-scale Gaussian, $G(x,y,t,\sigma)$, with an input video sequence, $I(x,y,t)$ as

$$L(x,y,t,\sigma) = G(x,y,t,\sigma) * I(x,y,t) \tag{9}$$

where $*$ is the 3D-convolution operation in x, y and t and

$$G(x,y,t,\sigma) = \frac{1}{(2\pi\sigma)^{\frac{3}{2}}} e^{-(x^2+y^2+t^2)/2\sigma^2} \tag{10}$$

The Space-time DoG function, $D(x,y,t,\sigma)$, convolved with the video sequence is defined as

$$D(x,y,t,\sigma) = (G(x,y,t,k\sigma) - G(x,y,t,\sigma)) * I(x,y,t) \quad (11)$$

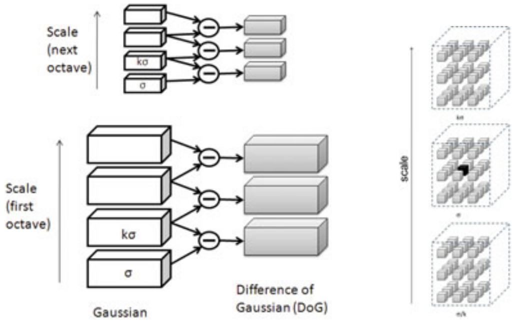
$$= L(x,y,t,k\sigma) - L(x,y,t,\sigma) \quad (12)$$

The initial video sequence is incrementally convolved with a Gaussian to produce volumes separated by a constant factor k in scale-space-time, as shown in the left side of Figure 3a. Adjacent volume scales are subtracted to produce the Space-Time DoG volumes shown on the right side of Figure 3a.

In order to find local extrema, each sample point is compared to its 26 neighbors at time t, t-1 and t+1 in the current volume and to its 27 neighbors at time t, t-1 and t+1 in the scale above and below, as shown in Figure 3b.

The sample space-time interest point is selected only if it is larger than all of these neighbors. Most of the sample points are discarded during the first checks. After the detection of scale-space extrema, the edge-like features are discarded [1].

The drawback of this method, compared with its 2D version, is the computational time, as 3D convolutions are very computational demanding. In the following section, this step is speeded up with box filters and integral video approach.



(a) Space-time Difference of Gaussian for various octaves and scales **(b)** Search for local maxima and minima

Fig. 3. 3D SIFT detection part

2.4 3D Hessian

3D Hessian in an extension of the original SURF detector into the temporal domain developed by Willems et al. [28]. This method has some common properties with Space-Time DoG at a much lower computational cost. The two main advantages of this approach are: (1) features can be localized both in the spatio-temporal domain and over scales simultaneously using the determinant of the Hessian as saliency measure; (2) an efficient implementation of the detector is built by approximating all 3D convolutions using box-filters and integral video representation.

For spatio-temporal feature detection, the authors proposed the use of the Hessian matrix defined as

$$H(x,y,t;\sigma,\tau) = \begin{pmatrix} L_{xx}(x,y,\sigma,\tau) & L_{xy}(x,y,\sigma,\tau) & L_{xt}(x,y,\sigma,\tau) \\ L_{yx}(x,y,\sigma,\tau) & L_{yy}(x,y,\sigma,\tau) & L_{yt}(x,y,\sigma,\tau) \\ L_{tx}(x,y,\sigma,\tau) & L_{ty}(x,y,\sigma,\tau) & L_{tt}(x,y,\sigma,\tau) \end{pmatrix} \tag{13}$$

where $L_{xx}(x,y,\sigma,\tau)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2}g(\sigma,\tau)$ with the video I at location (x,y,t) and similarly for $L_{xy}(x,y,\sigma,\tau)$ and $L_{yy}(x,y,\sigma,\tau)$.

The strength of each space-time interest point at a certain scale is computed by its determinant as follow

$$S = |det(H)| \tag{14}$$

The scale selection is realized with the scale-normalized determinant of the Hessian matrix. Using this approach, a single scale-invariant measure is obtained and it is used for the localization and for the selection of the spatial and temporal scales. It is interesting to notice that this implementation does not require an iterative method as it must be used in another well known approach proposed by Laptev. For more details about it, please refer to [14].

In the implementation, the use of integral video structure and box-filters reduce greatly the computational time. As a first step, the video is converted into an integral video structure, where an entry at location (x,y,t) holds the sum of all pixels in the rectangular region spanned by $(0,0)-(x,y)$, summed over all frames $[0,t]$. The use of integral videos permits to obtain the sum of values within any rectangular volume with only 8 additions, independently of the volume size. The Gaussian second-order derivatives are roughly appoximated with box-filter, as it was done in the 2D SURF detector. In total, 6 different second order derivatives in the spatio-temporal domain are needed and are denoted as D_{xx},D_{yy}, D_{tt}, D_{xy}, D_{tx} and D_{ty}. They can be computed using rotated versions of the two box-filters shown in Figure 4.

The scale spaces do not have to be computed hierarchically, as it is done in the Space-Time DoG method (see sub-section 2.3), but can be efficiently implemented by upscaling the box-filters and keeping the video volume in its original size. Each octave is divided into 5 scales, with a ratio between subsequent scales in the range 1.2-1.5 for the inner 3 scales. The determinant of the Hessian is computed over several octaves of both the spatial and temporal scale.

As a final step, a non-maximum suppression algorithm is used to obtain all extrema within the obtained 5 dimensional search-space (x,y,t,σ,τ).

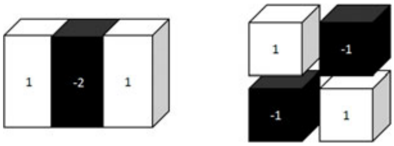


Fig. 4. Box filter approximations for the Gaussian second order partial derivatives used in 3D Hessian

3 Feature Description

Once an interest point in an image (or a space-time interest point in a video) is detected, the content around it has to be described. Several methods have been proposed in the 2D domain to described the content as a feature vector, e.g. moment invariants local derivatives [13], steerable filters [8], complex filters [2], SIFT [18], SURF [3], GLOH [22], CS-LBP [11]. A comparison of several 2D descriptor in shown in the work of Mikolajczyk et al. [22].

In this section, we explain the approaches used by SIFT and SURF description methods for finding interest points in an image. We also explain here the evolution of these methods into the temporal domain, known as 3D SIFT and 3D SURF.

3.1 2D SIFT

Scale Invariant Feature Transform (SIFT) descriptor [18] is based on the gradient distribution in a detected image region. As a first step in the algorithm, the image gradient magnitudes and orientations are computed in a square region centered on the interest point location, and the scale of the interest point is used to select the level of Gaussian blur for the image. For each image sample, $L(x,y)$, at scale σ, the gradient magnitude, $m(x,y)$, and orientation, $\theta(x,y)$ are computed using pixel differences as

$$m_{2D}(x,y) = \sqrt{L_x^2 + L_y^2} \tag{15}$$

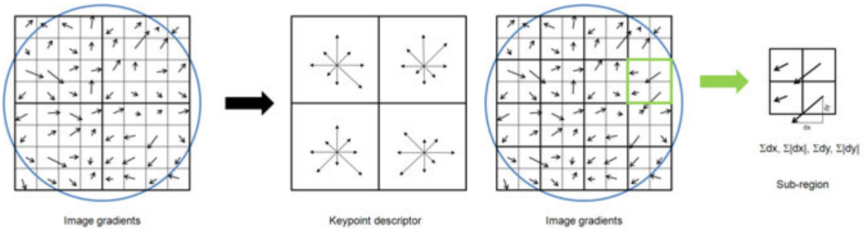
$$\theta(x,y) = tan^{-1}(\frac{L_y}{L_x}) \tag{16}$$

where L_x and L_y are respectively computed using finite difference approximations: $L_x = L(x+1,y) - L(x-1,y)$ and $L_y = L(x,y+1) - L(x,y-1)$.

Orientation invariance is achieved by rotating the coordinates of the descriptor and the gradient orientations relative to the main orientation of the interest point detected in a previous step [18]. In order to avoid sudden changes in the descriptor due to small changes in the position of the square window, a Gaussian weighting function is used to give more enphasis to the gradients that are close to the interest point location and less to the gradients that are far from it (a weight is therefore assigned to the magnitude $m_{2D}(x,y)$ of each sample point). The Gaussian function has σ equal to one half the width of the descriptor window and it is shown on the left side of Figure 5a with a circular window .

The descriptor is built by histogramming the gradient orientations over a $M \times M$ sample regions, as show on the right side of Figure 5a. The histogram is calculated as n orientations and the length of each arrow on the right side of Figure 5a corresponds to the magnitude value of that histogram entry. The final descriptor is then obtained concatenating the $M \times M$ array of histogram with n orientation bins in each. As a last step, the final vector is normalized to unit length to be invariant to illumination changes.

As an example, the common SIFT descriptor is divided into 16 subregions and the histogram contains 8 orientation bins. This parameters give a descriptor whose feature vector is 4×4×8=128 dimensions lenght. In Figure 5a only 4 subregions are shown.



(a) SIFT descriptor **(b)** SURF descriptor

Fig. 5. SIFT and SURF descriptors

3.2 2D SURF

The Speeded Up Robust Features (SURF) descriptor [3] is based on the description of the nature of the image intensity pattern in a detected region, unlike SIFT descriptor which is based on a histogram approach. The first step of SURF algorithm consists of constructing a square window centered around the interest point location; in order to be rotational invariant, this region has to be oriented along the main orientation selected in a previous step [3]. The size of the window is 20s, where s is the scale at which the interest point was detected. The main window is then divided regularly into smaller M×M square subregions (as it was done in SIFT) and these divisions of the principal window helps in keeping into consideration the spatial information. For each sub-region, few simple features at 5×5 regularly spaced sample points are computed using Haar wavelet. The Haar wavelet response in horizontal direction is denoted as d_x and d_y is the Haar wavelet response in the vertical direction. The filter has size equal to 2s. As it was done in the SIFT descriptor, a Gaussian weighting function (with $\sigma = 3.3s$) is used to weight the dx and dy responses to increase the robustness towards geometric deformations and localisation errors.

For every sub-region, the wavelet responses dx and dy are summed up. The sum of the absolute values of the responses, $|dx|$ and $|dy|$, are also computed and stored, in order to take into account the polarity of the intensity changes. For each sub-region, the descriptor vector has therefore four dimension as $v = (\sum dx, \sum dy, \sum|dx|, \sum|dy|)$. The illumination invariance is achieved as the wavelet responses are invariant to a bias in illumination and the invariance to contrast is obtained by normalizing the descriptor into a unit vector (as the final step in SIFT descriptor). Given the standard parameter $M = 4$, the descriptor vector is 4×4×4=64 dimension length.

Compared with SIFT descriptor, SURF algorithm can describe the area around an interest point much faster and this efficiency is achieved by relying on integral images, Haar wavelets and on simple summation of wavelet responses. Moreover, SURF descriptors integrates the gradient information within a subregion, while SIFT depends on the orientation of individual gradients. SURF's descriptor is half the size of SIFT, property that permits a faster computation for a following matching step between feature vectors (e.g. in image stitching applications).

3.3 3D SIFT

The 3-Dimensional Scale-Invariant Feature Transform (3D SIFT) descriptor is an extension of SIFT into the 3-Dimensional space and it has been developed by Scovanner el at. [26]. The gradient magnitude and orientation is computed in a similar manner as in 2D SIFT. In a 3-Dimensional space, the magnitude and orientations are given by

$$m_{3D}(x,y,t) = \sqrt{L_x^2 + L_y^2 + L_t^2} \tag{17}$$

$$\theta(x,y,t) = tan^{-1}(\frac{L_y}{L_x}) \tag{18}$$

$$\phi(x,y,t) = tan^{-1}(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}) \tag{19}$$

where L_x, L_y and L_t are respectively computed using finite difference approximations: $L_x = L(x+1,y,t) - L(x-1,y,t)$, $L_y = L(x,y+1,t) - L(x,y-1,t)$ and $L_t = L(x,y,t+1) - L(x,y,t-1)$. In this manner, each pixel has two values which represent the direction of the gradient in three dimensions: θ encodes the angle in the 2D gradient direction, while ϕ encodes the angle away from the 2D gradient direction. The angle ϕ is always in the range $(-\frac{\pi}{2}, \frac{\pi}{2})$ since $\sqrt{L_x^2 + L_y^2}$ is positive. Every angle is therefore represented by a single unique (θ, ϕ) pair, which encodes the direction of the gradient in three dimensions. A weighted histogram is then computed dividing θ and ϕ into equally sized bins; this step can be seen as dividing the sphere into meridians and parallels. For each 3D sub-region the orientations are accumulated into a histogram and the final descriptor is a vectorization of the sub-histograms

$$hist(i_\theta, i_\phi) += \frac{1}{\omega} m_{3D}(x', y', t') exp(\frac{-((x-x')^2 + (y-y')^2 + (t-t')^2)}{2\sigma^2}) \tag{20}$$

where (x,y,t) represents the location of the interest points, (x',y',t') represents the location of the pixel being added to the orientation histogram and ω is the solid angle of the sphere. In order to be rotational invariant, the surrounding 3D neighborhood should be rotate in the direction of the peaks of this histogram, but this is not the case for our tests in Section 4.

To create the sub-histograms, the sub-regions surrounding the interest point are sampled, as shown in Figure 6, where each pixel contains a single magnitude value m_{3D} and two orientation values θ and ϕ. For each 3D sub-region the orientations are accumulated into a histogram and the final descriptor is a vectorization of all sub-region histograms.

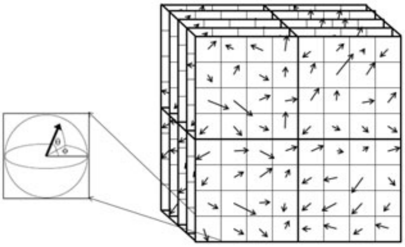


Fig. 6. 3D SIFT descriptor

3.4 3D SURF

3D SURF descriptor has been developed by Willems et al. [28] and is a direct extension of SURF.

A rectangular video patch is extracted around each interest point and its dimension is $s\sigma \times s\sigma \times s\tau$ where σ is the spatial scale and τ is the temporal scale. A typical value for s is set to 3 according to the authors. The volume is divided into $M \times M \times N$ subvolumes, where M and N are the number of division in the spatial and temporal direction respectively.

For each sub-volume, simple features are computed using the 3 axis-aligned Haar-wavelets shown in Figure 7. The Haar wavelet response in horizontal direction is denoted as d_x, and d_y and d_t is the Haar wavelet response in the vertical and in the time direction, respectively. Within each sub-regions, the wavelet responses are weighted by a Gaussian function and are summed up; the vector $v = (\sum d_x, \sum d_y, \sum d_t)$ is obtained for each sub-volumes. The sums over the absolute values could be taken into consideration, as it was done in the 2D case (sub-section 4), doubling the descriptor's size.

Compared with 3D SIFT, 3D SURF is computationally faster, relyes on integral video representation and Haar wavelet responses. As it was in 2D SURF, 3D SURF descriptor integrates the gradient information within each subregions, instead of computing 3-Dimensional gradient orientations.

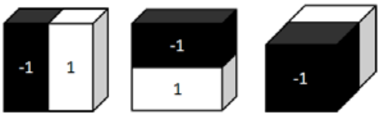


Fig. 7. Haar wavelet filters d_x, d_y and d_t

4 Experimental Setup

In order to evaluate and compare the described 3D detection and description methods, we build a classification framework for human action recognition. The methodology we adopt is a Bag of Words classication model [7]. As a first step, space-time interest points are detected using Space-Time DoG or 3D Hessian feature detection method and small video patches are extracted from each interest point. They represent the local information used to learn and recognize the different human actions. Each video patch is then described using 3D SIFT or 3D SURF feature description method, respectively. The result is a sparse representation of the video sequence as small video patch descriptors.

Having obtained all these data for the training set, a visual vocabulary is built by clustering using the k-means algorithm. The center of each cluster is defined as a spatial-temporal 'word' of which length depends on the length of the descriptor adopted. Each feature descriptor is successively assigned to the closest (using Euclidean distance) vocabulary word and a histogram of spatial-temporal word occurrence in the entire video is computed. Thus, each video is represented as a collection of spatial-temporal words from the codebook in the form of a histogram.

For classification, we use Support Vector Machines (SVM) with rbf kernel. As the algorithm has random components, such as the clustering phase, any experimental result reported is averaged over 20 runs. The database used is the standard KTH human action database [25]. This database contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. Each action class is performed several times by 25 subjects in different scenarios of outdoor and indoor environment. The camera is not static and the videos contain small scale changes. In total, the dataset contains 600 sequences. We divide the dataset into two parts: 16 people for training and 9 people for testing, as it has been done in [25, 15]. We limit the length of all video sequences to 300 frames.

5 Results

In our simulations, we used Scovanner's publicly available code[1] for the 3D SIFT description part (sub-section 3.3). We adopted the suggested parameters, which are slightly changed from what is described in the original paper [26]. For the detection part of 3D SIFT, named Space-Time DoG (sub-section 2.3), we implemented our own code. Regarding the 3D Hessian detection method (sub-section 2.4) and 3D SURF description method (sub-section 3.4), we run the program publicly available on Willems's website[2]. During the clustering phase, a codebook of $k = 1000$ visual words is built and 100 features are extracted from each video sequence. These parameters have been chosen due to previous tests.

In the previous sections, we described separately the detection part and description part for each method in order to better explain the different feature detection

[1] http://www.cs.ucf.edu/~pscovann/

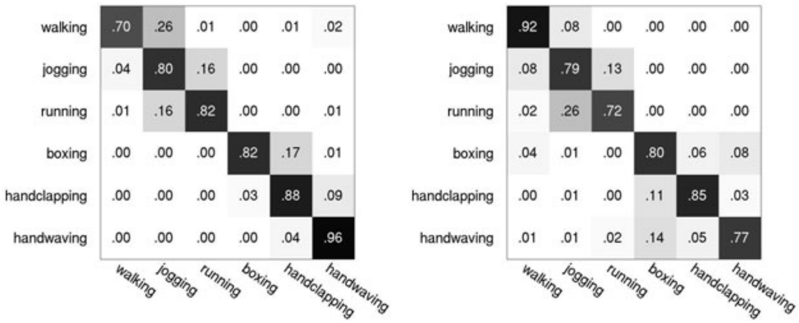
[2] http://homes.esat.kuleuven.be/~gwillems/research/Hes-STIP/

and feature description methods. Here we are testing the entire 3D SIFT and 3D SURF approaches.

In Table 1 the performance for both methods is shown. As it can be seen, the performance is quite similar: 3D SURF obtains 80.86% of accuracy, while the accuracy of 3D SIFT is 82.72%, gaining about 2%. The confusion matrices for both methods are shown in Figure 8.

Table 1. Performances using 3D SIFT and 3D SURF

Feature detector	Feature descriptor	Feature dimensions	Accuracy
Space-Time DoG	3D SIFT	640	82.72 %
3D Hessian	3D SURF	288	80.86 %



	walking	jogging	running	boxing	handclapping	handwaving
walking	.70	.26	.01	.00	.01	.02
jogging	.04	.80	.16	.00	.00	.00
running	.01	.16	.82	.00	.00	.01
boxing	.00	.00	.00	.82	.17	.01
handclapping	.00	.00	.00	.03	.88	.09
handwaving	.00	.00	.00	.00	.04	.96

(a) 3D SIFT, accuracy 82.72 %

	walking	jogging	running	boxing	handclapping	handwaving
walking	.92	.08	.00	.00	.00	.00
jogging	.08	.79	.13	.00	.00	.00
running	.02	.26	.72	.00	.00	.00
boxing	.04	.01	.00	.80	.06	.08
handclapping	.00	.01	.00	.11	.85	.03
handwaving	.01	.01	.02	.14	.05	.77

(b) 3D SURF, accuracy 80.86 %

Fig. 8. Confusion matrices for 3D SIFT and 3D SURF

In the 2D domain, SURF has been proved to perform similar to, or slighlty outperform, SIFT [4]. Hovewer, in our tests, the combination of the 3D SIFT detection and description methods is performing better than the combination of the 3D SURF detection and description methods. This could be explained as the 3D Hessian is computing approximations of the Gaussian filters used in the 3D convolution. 3D SURF descriptor is an accumulation of gradient information within each subregions, while 3D SIFT descriptor is an histogram representation of the 3-Dimensional gradient orientations and this could be another reason for the noticed differences in the field of human action recognition. The accuracy of these methods could also be affected by the descriptor's dimension: 3D SIFT generates a 640 feature vector length, while 3D SURF descriptor is a vector length of 288 dimensions.

The computational time for both approaches is shown in Table 2. The time is measured on a computer equipped with an AMD Opteron 252 running at 2.6 Ghz with 8 Gb RAM and computed as an average of 10 runs on different video sequences. 3D SURF is outperforming 3D SIFT by approximately 35 times faster. This is explained as 3D SURF relies on box filter approximation and integral video representation, which greatly speed up the performance. Moreover, in the 3D SIFT approach, the feature detection part takes the majority of the time, 306.54 seconds, while the descriptor part requires 47.4 seconds for the description of 100 local video patches.

Table 2. Computational time for detection and description of 100 features using 3D SIFT and 3D SURF

Feature detector	Feature descriptor	Environment	Computational time (s)
Space-Time DoG	3D SIFT	Matlab	353.94
3D Hessian	3D SURF	C	10.35

6 Conclusion

In this chapter we have described SIFT and SURF in order to detect and to describe interest points in an image. We have explained the theory upon which they are based and we have shown their similarities and differences. From the 2D detection and description methods, we explained the evolution into the 3D domain for both 3D SIFT and 3D SURF respectively. Moreover, we highlighted their similarities and differences.

As a comparison of their performance, we evaluated both the 3D approaches in the field of human action recognition on the standard KTH human action database. In our tests, the performance for both techniques is very close to each other: 3D SIFT is performing slightly better (82.72 % of accuracy) than 3D SURF (80.86 % of accuracy). However, 3D SURF is much more efficient than 3D SIFT, almost 35 times faster. Due to the greater efficiency and to the similar performance, 3D SURF method would be our choice for future development of human action recognition on a more realistic and challenging database, such as the Hollywood Human Action database [15].

References

[1] Allaire, S., Kim, J.J., Breen, S.L., Jaffray, D.A., Pekar, V.: Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis. In: Computer Vision and Pattern Recognition Workshops (January 2008)

[2] Baumberg, A.: Reliable feature matching across widely separated views. vol. 1, pp. 774–781 (2000)

[3] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)

[4] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., Van Gool, L.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110(3), 346–359 (2008), ISSN 1077-3142

[5] Brown, M., Lowe, D.G.: Recognising panoramas. In: IEEE International Conference on Computer Vision, vol. 2, pp. 12–18 (2003)

[6] Cheung, W., Hamarneh, G.: N-sift: N-dimensional scale invariant feature transform for matching medical images. In: 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2007, pp. 720–723 (2007)

[7] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features, pp. 65–72 (2005)

[8] Freeman, W.T., Adelson, E.H.: The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 891–906 (1991)

[9] Gordon, I., Lowe, D.G.: What and where: 3D object recognition with accurate pose. In: Ponce, J., Hebert, M., Schmid, C., Zisserman, A. (eds.) Toward Category-Level Object Recognition. LNCS, vol. 4170, pp. 67–82. Springer, Heidelberg (2006)

[10] Harris, C., Stephens, M.: A combined corner and edge detection. In: Proceedings of The Fourth Alvey Vision Conference, pp. 147–151 (1988)

[11] Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern Recogn. 42(3), 425–436 (2009), ISSN 0031-3203

[12] Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision 45(2), 83–105 (2001)

[13] Koenderink, J.J., van Doom, A.J.: Representation of local geometry in the visual system. Biol. Cybern. 55(6), 367–375 (1987), ISSN 0340-1200

[14] Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV, pp. 432–439 (2003)

[15] Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies, pp. 1–8 (June 2008)

[16] Lindeberg, T.: Scale-space theory: A basic tool for analysing structures at different scales. J. of Applied Statistics 21(2), 224–270 (1994)

[17] Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision 30, 79–116 (1998)

[18] Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004), ISSN 0920-5691

[19] Matas, J., Chum, O., Martin, U., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of British Machine Vision Conference, London, vol. 1, pp. 384–393 (2002)

[20] Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Proc. ICCV, pp. 525–531 (2001)

[21] Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 128–142. Springer, Heidelberg (2002)

[22] Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell 27(10), 1615–1630 (2005), ISSN 0162-8828

[23] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. Int. J. Comput. Vision 65(1-2), 43–72 (2005), ISSN 0920-5691

[24] Quack, T., Bay, H., Van Gool, L.: Object recognition for the internet of things. In: Floerkemeier, C., Langheinrich, M., Fleisch, E., Mattern, F., Sarma, S.E. (eds.) IOT 2008. LNCS, vol. 4952, pp. 230–246. Springer, Heidelberg (2008)

[25] Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR 2004), vol. 3, pp. 32–36. IEEE Computer Society Press, Washington, DC (2004)

[26] Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: MULTIMEDIA 2007: Proceedings of the 15th International Conference on Multimedia, pp. 357–360. ACM, New York (2007), ISBN 978-1-59593-702-5

[27] Se, S., Lowe, D., Little, J.: Vision-based mobile robot localization and mapping using scale-invariant features. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 2051–2058 (2001)

[28] Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)

[29] Willems, G., Tuytelaars, T., Van Gool, L.: Spatio-temporal features for robust content-based video copy detection. In: MIR 2008: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 283–290. ACM Press, New York (2008)

[30] Zhang, Z., Huang, Y., Li, C., Kang, Y.: Monocular vision simultaneous localization and mapping using surf, pp. 1651–1656 (June 2008)

Human Action Recognition Based on Radon Transform

Yan Chen, Qiang Wu, and Xiangjian He

Centre for Innovation in IT Services and Applications (iNext)
University of Technology, Sydney, Australia
{jade,wuq,sean}@it.uts.edu.au

Abstract. A new feature description is used for human action representation and recognition. Features are extracted from the Radon transforms of silhouette images. Using the features, key postures are selected. Key postures are combined to construct an action template for each action sequence. Linear Discriminant Analysis (LDA) is applied to obtain low dimensional feature vectors. Different classification methods are used for human action recognition. Experiments are carried out based on a publicly available human action database.

1 Introduction

Human action recognition is a fundamental topic in computer vision and has become an active research area in recent years. It has many applications in the areas of smart surveillance, user interface for control command and so on.

Human postures play an important role in human action. Therefore, human posture analysis is one of the most important steps toward successful human action recognition. Blank *et al.* [4] represented actions by using the shapes of silhouettes and recognized the action based on poisson transform. Efros *et al.* [10] used optical flow to characterize motion in very low resolution video sequences. Their method relied on previously aligned video clips of human actions. Singh, Mandal and Basu [24] used Radon transform for pose recognition. But, their work was restricted to hand or feet positions. This work is inspired by Boulgouris [5] who used Radon transform for gait recognition. Wang *et al.* used R transform for action recognition in [31]. R transform is based on Radon transform. In [31], they claimed that R transform was translation invariant, rotation invariant and scaling invariant. In fact, rotation invariance is not good for action recognition and rotation invariance may sometimes even degrade action recognition result. For example, a feature which is rotation invariant cannot be used to distinguish the postures corresponding to standing from those of lying. Like R-transform, Radon transform is also scaling invariant and translation invariant based on the following facts. Scaling invariance can be easily achieved through alignment

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 369–389.
springerlink.com

of the objects (*i.e.*, the human silhouettes in this paper) of different scales and translation invariance can be achieved by moving the centers of the objects to the origin before applying a Radon transform. In one word, Radon transform is better for action recognition than R-transform when we request the recognition to be scaling and translation invariant but not rotation invariant.

Key postures are selected for the representation of an action sequence because a typical human action contains only a few important postures which are significantly different from each other. Toyama and Black [28] used exemplar frames to achieve human action tracking. Their method was based on learning and was not straightforward. Lim and Thalmann [16] proposed a method to extract key postures using curve simplification. The limitation is that it requires specific 3D motion capture device, which is costly and is not practical for some applications. Lv and Nevaita [17] adopted motion energy for automatic extraction of 3D key postures, which also required 3D motion data. Chen . proposed a method using entropy to select key postures from a video in [7]. But, they ignored the local features of a frame. Key postures selected according to shapes or other vision features are different from key frames which are selected using the video compression information as shown in [6].

In order to obtain key postures in this paper, the action sequences are extracted from the silhouette using Radon transforms. An unsupervised clustering method is applied to identify the key postures in each sequence. Then, the key postures are used in the subsequent training and testing steps.

To further optimize the extracted action features inside the key postures, the Linear Discriminant Analysis (LDA) is adopted to reduce the dimension of the feature vectors. Several benchmark classifiers, including BayesNet [20], C4.5 or Decision Trees [22], and the Sequential Minimal Optimization (SMO) algorithms [21], are used in this work for action learning and classification.

A contribution of this paper is that it proposes a method that uses key postures to achieve action recognition. Using key postures is computation efficient. This paper applies a Radon transform to represent key postures. It will demonstrate that Radon transform is a good descriptor for human posture representation.

Although motion extraction and tracking are also important in action recognition, they are beyond the scope of this paper. It is supposed that all the moving objects have been extracted. The inputs of the proposed method are the silhouettes of the moving objects.

The remaining sections are organized as follows. Section 2 introduces the Radon transform and posture representation using Radon transforms. Section 3 presents the key posture extraction based on Radon Transform. Section 4 illustrates the classification of the action sequences. The experiments can be found in Section 5 and the conclusions are made in Section 6.

2 Human Posture Representation Using Radon Transforms

In this section, the Radon transform is introduced first and then we will discuss how the human postures are represented using Radon transforms.

2.1 Radon Transform

Radon transform is named after J. Radon who showed how to describe a function in terms of integral projections in 1917 [23]. The integral of a function over a line is the Radon transform. Radon transform is well known for its wide range of applications in various areas, such as radar imaging, geophysical imaging and medical imaging. Radon transform has various definitions, we use the one in [1] to illustrate it.

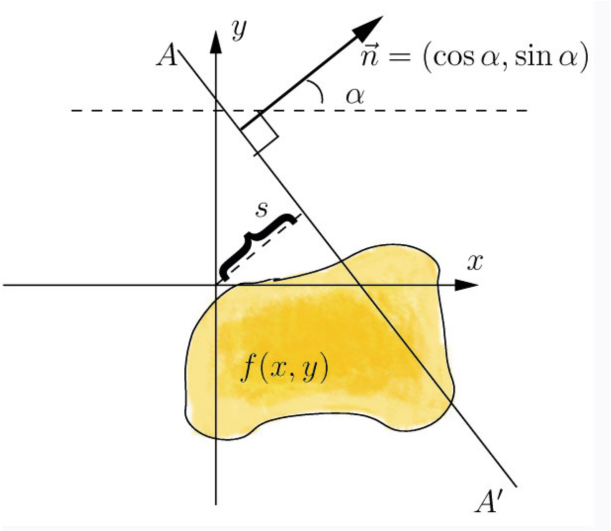


Fig. 1. The Radon transform computation [1]

As shown in Figure 1, let f be a continuous function vanishing outside some interested regions in the Euclidean plane R^2. The Radon transform, denoted by R_f is a function defined on the space of lines L (AA' in Figure 1) in R^2 by

$$R_f(L) = \int_L f(x) d\sigma(x), \tag{1}$$

where the integration is performed with respect to the arc length measure $d\sigma(x)$ on L. Concretely, any straight line L can be parameterized by

$$(x(t), y(t)) = t(\sin\alpha, -\cos\alpha) + s(\cos\alpha, \sin\alpha), \tag{2}$$

where s is the distance of L from the origin and α is the angle L makes with the x axis. Thus the quantities (s, α) are coordinates on the space of all lines in R^2, and the Radon transform can be expressed in these coordinates by

$$R_f(\alpha, s) = \int_{-\infty}^{\infty} f(x(t), y(t))dt \tag{3}$$

$$= \int_{-\infty}^{\infty} f(t(\sin\alpha, -\cos\alpha) + s(\cos\alpha, \sin\alpha))dt. \tag{4}$$

Figure 1 shows the computation of Radon transforms. Figure 2 shows an example of Radon transform. Two parallel lines are in the image (see Figure 2 (a)). Its Radon transforms have two intensive highlights with the same angle α but different distances from the original s (see Figure 2 (b)).

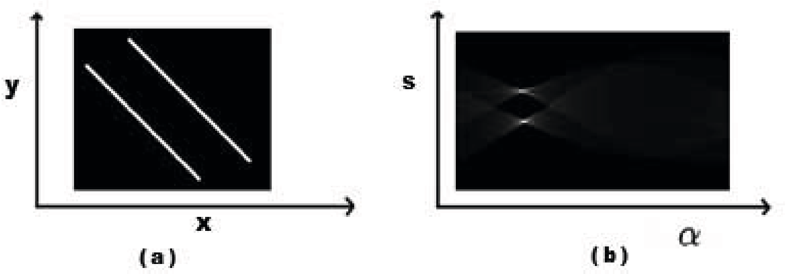


Fig. 2. Radon transform for two parallel lines

2.2 Human Posture Representation Using Radon Transform

Radon transform has several useful properties. Some of them are relevant to human posture representation [9].

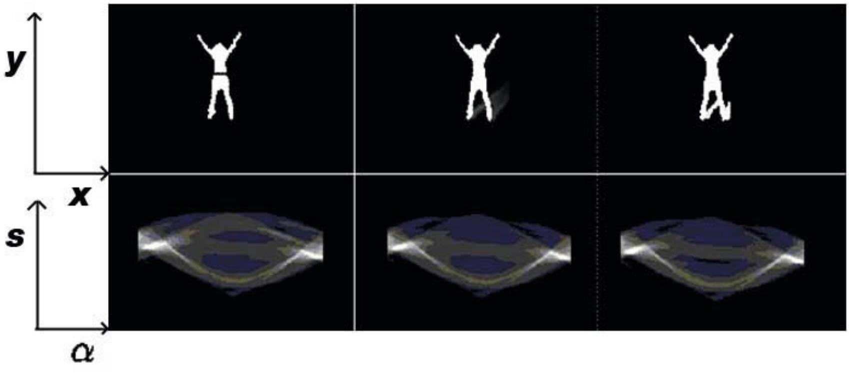


Fig. 3. Human images and their corresponding Radon transform

Subsequent to the extraction of human images, Radon transform on the human silhouette images are used to represent the corresponding human postures. Figure 3 shows some human images and their corresponding Radon transforms. The human images are in normal $x - y$ coordinate system. The Radon transform images are in $\alpha - s$ coordinate system. These human images are different because of shadow and noise. But their Radon transforms look much closer and all have two similar bright parts.

3 Key Posture Selection

Key postures are used to represent action sequences because actions are continuous. Key postures are the most significant postures in an action sequence. Once the key postures are selected, the other postures in the sequences can be clustered into one of these key postures.

3.1 Affinity Propagation Clustering

Clustering data based on a measure of similarity is a critical step in pattern recognition and image processing. It classifies data into different subsets so that the data in a subset share common characteristics. The measure used for the clustering method is the sum of squared errors between the data points and their corresponding data center. k-means [25], fuzzy clustering [3], and affinity propagation clustering [11] are some examples of the existing clustering methods. Some clustering methods, *e.g.*, k-means, begin with an initial set of randomly selected exemplars and iteratively refine this set so as to decrease the sum of squared errors. Other clustering methods, *e.g.*, affinity propagation, do not require initially selected exemplars explicitly.

The input of affinity propagation clustering is the similarity of data points and user's preference. The similarity $s(i, k)$ indicates how well the data point with index k is suited to be the exemplar for data point i. Similarity can be obtained in many ways. For example, the similarity can be measured using Euclidean distance, Manhanlanobis distance and so on. In order to achieve minimum squared error for clustering, similarity is set to a negative squared error. For points x_i and x_k, the similarity between these two points is $s(i, k) = -\left\|x_i - x_k\right\|^2$. Instead of taking the number of clusters as input, affinity propagation takes the user preference as an input. The user's preference is an $N \times 1$ matrix p. $p(i)$ indicates the preference to choose point i as a cluster center. If all data points are suitable as exemplars, the preference is to use a common value. The preference is set to the median of the similarity for this work.

There are two kinds of messages exchanged between data points. Each takes into account a different kind of competition. The responsibility message, $r(i, k)$, sent from data point i to candidate exemplar point k, reflects the accumulated evidence for how well-suited point k is to serve as the exemplar for point i. The availability message, $a(i, k)$, sent from candidate exemplar point k to point i, reflects the accumulated evidence for how appropriate it would be for point i to choose point k as its exemplar.

Figure 4 is the procedure of affinity propagation clustering. At beginning, the availabilities are initialized to zero: $a(i,k) = 0$. Then, the iteration begins. The responsibilities are computed using Equation 5, where $s(i,k)$ is the element of similarity matrix, and $a(i,k)$ is the element of availability matrix. The update of the responsibilities updates all candidate exemplars competing for ownership of a data point. The availabilities are computed using Equation 6.

$$r\left(i,k\right) = s\left(i,k\right) - max_{k \neq k\text{'}}\left\{a(i,k\text{'}) + s(i,k\text{'}\right\}. \qquad (5)$$

$$a\left(i,k\right) = min\left\{0, r(k,k) + \sum_{i\text{'} \notin i,k} max\left\{0, r(i\text{'},k)\right\}\right\}. \qquad (6)$$

The availability $a(i,k)$ is set to the self responsibility $r(k,k)$ plus the sum of the positive responsibilities that candidate exemplar k receives from other points. The availabilities and responsibilities are combined to identify exemplars. For point i, the value of k that maximizes $a(i,k) + r(i,k)$ either identifies point i as an exemplar if $i = k$, or identifies the data point that is the exemplar for point i. The message passing procedure is terminated in the following situations:

- a preset number of iterations has been reached,
- changes in the messages fall below a threshold, or
- the local decisions stay constant for some number of iterations.

There are two reasons to choose affinity propagation clustering for key posture selection in this paper. Firstly, it speeds up the convergence time compared with other methods [11]. Secondly, it does not require initially chosen exemplars. By considering all data points as candidate centers and gradually identifying clusters, affinity propagation is able to avoid many of the poor solutions caused by unlucky initialization.

3.2 Key Posture Identification

Radon transforms are computed frame by frame for an video action. Affinity propagation clustering is applied to Radon transform to obtain key postures. The inputs for affinity propagation clustering are the similarity matrix and the preference as exemplar for each frame. The similarity matrix is computed to measure between frame's Radon transforms. In this work, each frame is equally regarded as an exemplar. Therefore, the preference acting as exemplar is set at the same value for each frame. The median of all Radon transforms is used as the preference. The outputs of the affinity propagation clustering are the cluster centers which are the key postures for an action.

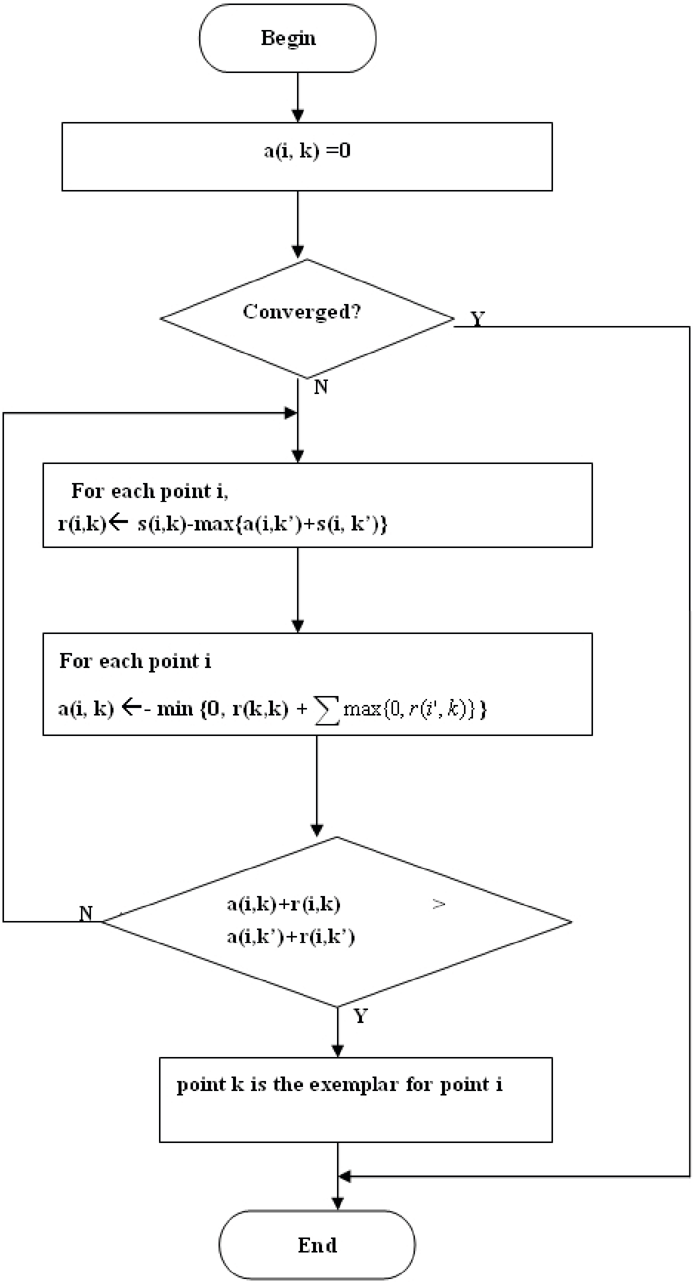

Begin
a(i, k) =0
Converged?
Y
N
For each point i,
r(i,k)← s(i,k)-max{a(i,k')+s(i, k')}
For each point i
a(i, k) ←- min {0, r(k,k) + ∑max{0, r(i', k)}}
a(i,k)+r(i,k) >
a(i,k')+r(i,k')
N
Y
point k is the exemplar for point i
End

Fig. 4. Affinity propagation clustering procedure

4 Action Recognition

One of the challenges for action recognition is to obtain the action template from the given information of an action. In the following, we describe our method to obtain templates of human actions using key postures, and the method for learning or classification based on the templates.

4.1 Action Template Creation

Different people perform similar actions in different styles. Therefore, there are different action sequences for one action no matter whether the action is performed by the same person or by different people. However, from observation, each action has similar key postures although the key postures are not exactly the same. The proposed method uses the combination of all key postures as the feature descriptor for an action. The advantage of using the combination of key postures is that it does not require starting posture alignment.

Suppose that there is an action sequence which has N frames denoted in the set F by

$$F = \{F_1, F_2, F_3, \cdots, F_i, \cdots | 1 \leq i \leq N\} . \tag{7}$$

Their corresponding Radon transforms are denoted in the set R by

$$R = \{R_1, R_2, R_3, \cdots, R_i, \cdots | 1 \leq i \leq N\} . \tag{8}$$

The Radon transform of a frame $R_j (1 \leq j \leq N)$ is a matrix with large amount of data. The key postures in an action sequence selected by the method shown in Section 3 are named as $K_1, K_2, \cdots, K_j$ $(1 \leq j \leq N)$.

Let $R_{K_1}, R_{K_2}, \cdots, R_{K_j}$ $(1 \leq j \leq N)$ be the corresponding Radon transforms of the selected key postures. For each action, the template TP is calculated by

$$TP = R_{K_1} + R_{K_2} + \cdots + R_{K_j} \qquad (1 \leq j \leq N). \tag{9}$$

The dimension of TP and R is very high. Dealing with such high dimensional data will cause poor recognition rate and high computing complexity. Therefore, feature extraction is needed to extract most important information from TR for the classification purpose. During feature extraction, the dimension of data is reduced. Besides that, a good feature extraction will enhance those features of input data that achieve better classification results. Typical methods for extracting the most expressive features and reducing the feature dimension include Principal Component Analysis (PCA) [12][15][33], Independent Component Analysis (ICA) [13] and their variances. In addition to PCA and ICA, Linear Discriminant Analysis (LDA) [2][12][34] or Fisher's Linear Discriminant (FLD) [8][19] are used to discriminate different patterns. PCA and LDA are two widely used, conventional tools for dimension reduction and feature extraction [18][27]. However, there is a tendency for the preferred use of LDA over PCA because LDA deals directly with discrimination between classes. In contract with LDA, PCA deals with the data in its entirety for the principal component analysis without paying any attention to the underlying class structures [18]. LDA requires category information in order to compute a vector which best discriminates between classes. Therefore, in this paper, LDA is chosen for feature extraction from template TP.

For a given dataset with c classes, LDA aims to find the best $c-1$ features in the underlying data that best discriminate among classes. LDA defines two measures:

1. within-class scatter matrix, as represented by

$$S_w = \sum_{j=1}^{c}\sum_{i=1}^{N_j}\left(x_i^j - \mu_j\right)\left(x_i^j - \mu_j\right)^T , \tag{10}$$

   where x_i^j is the i-th sample of class j, μ_j is the mean of class j, c is the number of classes, and N_j is the number of samples in class j.
2. between-class scatter matrix, as represented by

$$S_b = \sum_{j=1}^{c}\left(\mu_j - \mu\right)\left(\mu_j - \mu\right)^T , \tag{11}$$

   where μ represents the mean of all classes.

Then, LDA tries to find a best $(c-1)$ feature space W_{pro} that maximizes the ratio of the between-class scatter matrix to the within-class scatter matrix, *i.e.*, maximizing the ratio $\frac{S_b}{S_w}$.

The best feature space W_{pro} is defined as follows:

$$W_{pro} = \frac{W^T S_b W}{W^T S_w W} = [W_1, W_2, \cdots, W_{c-1}]. \tag{12}$$

Accordingly, when given a dataset denoted by X, its selected feature set denoted by X_{LDA} can be obtained by projecting X onto the $(c-1)$ feature space as follows:

$$X_{LDA} = X \cdot W_{pro}. \tag{13}$$

In order to obtain the the best features to discriminate the actions, the action templates TP and their corresponding action classes are inputed for LDA computation. Equation 13 is applied to obtain the most significant features of TP. After using LDA, the dimension of TP has been decreased dramatically.

4.2 Learning and Classification Procedure

The classifiers used for this part of work are the Bayesian-based classifier BayesNet [12][20], C4.5 or Decision Trees [22], and the Sequential Minimal

Optimization (SMO) algorithm [29][32]. These three classifiers are available in the WEKA package, a publicly available toolbox for automatic classification [32].

BayesNet enables the use of a Bayesian Network learning using various search algorithms and quality measures. C4.5 is a classifier for generating a pruned or unpruned C4.5 decision tree. C4.5 is a supervised symbolic classifier based on the notion of entropy since its output, a decision tree, can be easily understood and interpreted by human. SMO is an algorithm for training a support vector classifier.

During the training stage, the action templates of the training samples are inputed into a classifier. The classifier learns from the input samples and stores information for recognition task.

For recognition, the action descriptions of the testing samples are calculated as described in the previous sections. The descriptions are input into the classifier. The classifier determines the action according to the information it learns from the sample action.

5 Experiments

The experiments were based on Weizmann Institute of Science's human action database [4]. To our best knowledge, this database is one of the few reasonable sized available public databases for human action recognition. It contains 90 action videos (9 subjects, each subject performing 10 natural actions). The actions include walking (walk), running (run), bending (bend), gallop-sideways (side), one-hand-waving (wave1), two-hand-waving (wave2), jumping-forward-on-two-legs (jump), jumping-in-place (pjump), skipping (skip) and jumping-jacking (jack). The actions include both periodic actions (*e.g.*, walk, run) and non-periodic actions (*e.g.*, bend). For the periodic actions, a subject performed the same action multiple time (two or three times). The resolution of the video is 180×144. The video sequence was taken around the speed of 25 frames per second (FPS). Human silhouettes are provided by the database. The quality of these silhouettes is generally good although there are some defects. Morphological operations including dilation and erosion are applied to repair these defects. Figure 5 is an example of the database performing jack action. The heads of some humans are missing.

5.1 Experiments 1

Leaving-one-out cross-validation is used for the research in this experiment because it avoids any possible bias introduced when relying on a particular division of the sample into test and training components. In experiment 1, leaving-one-out was leaving one subject out (LOSO). That means 8 subjects' 80 action videos were used for training while the remaining 1 subject's 10

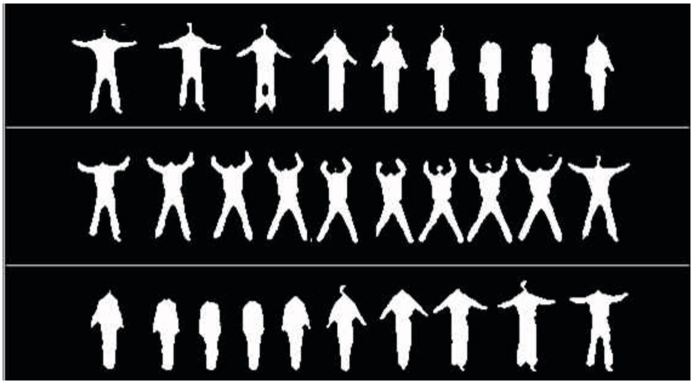


Fig. 5. An Example of Database (Jack)

action videos were used for testing. The experiment repeated 9 times until all of these 9 subjects' actions were used for testing.

All of the action sequences had their corresponding key postures according to Section 3. After the key postures were obtained, the action templates were created according to Subsection 4.1. The learning and testing processes were conducted according to Subsection 4.2.

During the training process, the action templates of the training samples (8 subjects, 80 actions videos) were inputed into a classifier. The classifier learned from these training input action templates and stored information for future recognition task. During the testing process, the testing template was fed into the classifier. Then, the classifier classified the template based on the knowledge that it learned from the training templates.

The confusion matrix are shown in Table 1, Table 3 and Table 5 using difference classifiers. Table 1 shows the results using the SMO classifier. Each row represents the nine actions classified and their actual results. For example, at row 5, there are nine running actions to be classified. The results for the running action show that eight out of the nine actions are correctly classified as 'run' and one of them is classified as 'skip'. At row 7, there are nine testing samples for skipping being classified. The results show that seven out of the nine samples are correctly classified as 'skip', while one of them is classified as 'wave2' and the other one is classified as 'side'. Table 2 shows the accuracies for this experiment using key postures based on SMO classifier. It examines true positive rate, false positive rate, false negative rate, precision rate and recall rate. All of these rates are evaluated for a recognition system. Table 3 shows the results using BayesNet classifier. Table 4 shows the accuracies for this experiment using key postures based on BayesNet classifier. The results obtained using Decision Tree/C4.5 classifier is illustrated in Table 5. Table 6 shows the accuracies for this experiment using key postures based on C4.5 classifier.

Table 1. Confusion matrix for LOSO cross validation (SMO)

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	9	0	0	0	0	0	0	0	0	0
jace	0	9	0	0	0	0	0	0	0	0
jump	0	0	6	3	0	0	0	0	0	0
pjump	0	0	3	6	0	0	0	0	0	0
run	0	0	0	0	8	0	1	0	0	0
side	0	0	0	0	0	9	0	0	0	0
skip	0	0	0	0	0	1	7	0	0	1
walk	0	0	0	0	0	0	0	9	0	0
wave1	0	0	0	1	0	0	0	0	7	1
wave2	0	0	0	0	0	0	0	0	0	9

Table 2. Accuracies for LOSO cross validation(SMO)

	TP rate	FP rate	FN rate	Precision	Recall
bend	1.000	0.000	0.000	1.000	1.000
jack	1.000	0.000	0.000	1.000	1.000
jump	0.667	0.037	0.333	0.947	0.667
pjump	0.667	0.049	0.333	0.931	0.667
run	0.889	0.000	0.111	1.000	0.889
side	1.000	0.012	0.000	0.988	1.000
skip	0.778	0.012	0.222	0.984	0.778
walk	1.000	0.000	0.000	1.000	1.000
wave1	0.778	0.000	0.222	1.000	0.778
wave2	1.000	0.025	0.000	0.976	1.000

Table 3. Confusion matrix for LOSO cross validation (BayesNet)

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	9	0	0	0	0	0	0	0	0	0
jack	0	8	0	0	0	0	0	0	0	1
jump	0	0	7	1	1	0	0	0	0	0
pjump	0	0	3	5	1	0	0	0	0	0
run	0	0	0	0	8	0	1	0	0	0
side	0	0	0	0	0	9	0	0	0	0
skip	0	0	0	0	0	1	7	0	1	0
walk	0	0	0	0	1	0	0	8	0	0
wave1	0	0	0	1	0	0	0	0	7	1
wave2	0	0	0	0	0	0	0	0	0	9

Table 4. Accuracies for LOSO cross validation(BayesNet)

	TP rate	FP rate	FN rate	Precision	Recall
bend	1.000	0.000	0.000	1.000	1.000
jack	0.889	0.000	0.111	1.000	0.889
jump	0.778	0.037	0.222	0.955	0.778
pjump	0.556	0.025	0.444	0.957	0.556
run	0.889	0.037	0.111	0.960	0.889
side	1.000	0.012	0.000	0.988	1.000
skip	0.778	0.012	0.222	0.984	0.778
walk	0.889	0.000	0.111	1.000	0.889
wave1	0.778	0.012	0.222	0.984	0.778
wave2	1.000	0.012	0.000	0.988	1.000

Table 5. Confusion matrix for LOSO cross validation (C4.5)

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	8	0	0	0	0	1	0	0	0	0
jack	0	8	0	0	0	0	0	0	0	1
jump	1	0	7	1	0	0	0	0	0	0
pjump	0	0	3	5	0	0	1	0	0	0
run	0	0	0	0	8	0	1	0	0	0
side	0	0	0	0	0	9	0	0	0	0
skip	0	0	0	0	0	1	7	1	0	0
walk	0	0	0	1	0	0	0	8	0	0
wave1	0	0	0	1	0	0	0	0	6	2
wave2	0	0	0	0	0	0	0	0	0	9

Table 6. Accuracies for LOSO cross validation(C4.5)

	TP rate	FP rate	FN rate	Precision	Recall
bend	0.889	0.012	0.111	0.986	0.889
jack	0.889	0.000	0.000	1.000	1.000
jump	0.778	0.037	0.222	0.955	0.778
pjump	0.556	0.037	0.444	0.938	0.556
run	0.889	0.000	0.111	1.000	0.889
side	1.000	0.025	0.000	0.976	1.000
skip	0.778	0.025	0.222	0.969	0.778
walk	0.889	0.012	0.111	0.986	0.889
wave1	0.667	0.000	0.333	1.000	0.667
wave2	1.000	0.037	0.000	0.964	1.000

5.2 Experiments 2

In this part of experiment, leaving-one-sample-out (LOO) cross validation was used because it is easier to compare our method with others work. Since we had 90 samples in our dataset, we had 89 samples for training, and the remaining one for testing. The same experiment was repeated 90 times until all of the 90 samples were used as testing samples. The overall accuracy was estimated and it was the average of the result obtained from the repeated experiments.

After the key postures were obtained for each video, the action templates were created for each action video. The learning and testing processes were conducted according to Section 4.2. The recognition rate for SMO and Bayesnet achieve 100%. However, the recognition rates for C4.5 is 92.222%. Table 7, Table 9 and Table 11 show the confusion matrices using SMO, Bayesnet and C4.5 respectively. Table 8 and Table 10 are the accuracies for this experiment using SMO classifier and BayesNet classifier respectively. Table 12 shows the accuracies of the experiment using C4.5 classifier.

Table 13 compares the best accuracy of our approach with the results of related studies which used the same data set. Compared with leaving one sample out, our leaving one subject is more strict because we leave one subject's 10 actions for testing. All the actions performed by this subject are not in the training data. The information of the testing subject is totally unknown from the training process. As shown in our result, our approach has still achieved comparative results although our test condition has been more strict. With a loose constraint, our leaving-one-sample out has achieved 100% recognition rate (see Table 8 and Table 10).

Table 7. Confusion matrix for leave-one-out cross validation (SMO)

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	9	0	0	0	0	0	0	0	0	0
jack	0	9	0	0	0	0	0	0	0	0
jump	0	0	9	0	0	0	0	0	0	0
pjump	0	0	0	9	0	0	0	0	0	0
run	0	0	0	0	9	0	0	0	0	0
side	0	0	0	0	0	9	0	0	0	0
skip	0	0	0	0	0	0	9	0	0	0
walk	0	0	0	0	0	0	0	9	0	0
wave1	0	0	0	0	0	0	0	0	9	0
wave2	0	0	0	0	0	0	0	0	0	9

Table 8. Detail accuracy for leave-one-out cross validation(SMO)

	TP rate	FP rate	FN rate	Precision	Recall
bend	1	0	0	1	1
jack	1	0	0	1	1
jump	1	0	0	1	1
pjump	1	0	0	1	1
run	1	0	0	1	1
side	1	0	0	1	1
skip	1	0	0	1	1
walk	1	0	0	1	1
wave1	1	0	0	1	1
wave2	1	0	0	1	1

Table 9. Confusion Matrix for leave-one-out cross validation (BayesNet)

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	9	0	0	0	0	0	0	0	0	0
jack	0	9	0	0	0	0	0	0	0	0
jump	0	0	9	0	0	0	0	0	0	0
pjump	0	0	0	9	0	0	0	0	0	0
run	0	0	0	0	9	0	0	0	0	0
side	0	0	0	0	0	9	0	0	0	0
skip	0	0	0	0	0	0	9	0	0	0
walk	0	0	0	0	0	0	0	9	0	0
wave1	0	0	0	0	0	0	0	0	9	0
wave2	0	0	0	0	0	0	0	0	0	9

Table 10. Detail Accuracy for leave-one-out cross validation(BayesNet)

	TP rate	FP rate	FN rate	Precision	Recall
bend	1	0	0	1	1
jack	1	0	0	1	1
jump	1	0	0	1	1
pjump	1	0	0	1	1
run	1	0	0	1	1
side	1	0	0	1	1
skip	1	0	0	1	1
walk	1	0	0	1	1
wave1	1	0	0	1	1
wave2	1	0	0	1	1

Table 11. Confusion matrix for leave-one-out cross validation (C4.5)

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	9	0	0	0	0	0	0	0	0	0
jack	0	9	0	0	0	0	0	0	0	0
jump	0	0	8	0	0	0	0	1	0	0
pjump	0	1	0	8	0	0	0	0	0	0
run	0	0	1	0	8	0	0	0	0	0
side	0	0	0	0	0	8	1	0	0	0
skip	0	0	0	0	1	0	8	0	0	0
walk	0	0	0	0	0	0	0	8	1	0
wave1	0	1	0	0	0	0	0	0	8	0
wave2	0	0	0	0	0	0	0	0	0	9

Table 12. Detail accuracy for leave-one-out cross validation(C4.5)

	TP rate	FP rate	FN rate	Precision	Recall
bend	1.000	0.000	0.111	1.000	0.900
jack	1.000	0.025	0.000	0.818	1.000
jump	0.889	0.012	0.111	0.889	0.889
pjump	0.889	0.000	0.111	1.000	0.889
run	0.889	0.012	0.111	0.889	0.889
side	0.889	0.000	0.111	1.000	0.889
skip	0.889	0.012	0.111	0.889	0.889
walk	1.000	0.012	0.111	0.900	0.900
wave1	0.889	0.012	0.111	0.889	0.889
wave2	1.000	0.000	0.000	1.000	1.000

Table 13. Comparison with related studies

Matching Method	Brief Comments of Methods	Test Dataset	Best Accuracy
Ikizler [14]	‘bag-of-rectangles’ SVM	9 actions (no skip)	100%
Blank et al. [4]	Poisson equation space time shape	9 actions (no skip)	99.61%
Thurau [26]	no background subtraction	10 actions	87%
Wang & Suter [30]	Dynamic shape LPP	10 actions	100%
Our Approach (leave one subject out)	Radon transform SVM	10 actions	87.78%
Our Approach (leave one sample out)	Radon transform SVM	10 actions	100%

6 Conclusions

In this paper, one action recognition approach based on key posture sequence matching has been proposed. The main points are listed here:

1. Radon transform is used to represent the human postures because it has suitable features for human posture representation. The Radon transforms of the extracted human silhouettes are calculated for further process.
2. Affinity propagation clustering is applied to the Radon transforms of the human postures to extract key postures from the action video. The reasons for using affinity propagation cluster are that it does not require initial chosen exemplar and it has short convergence time. Affinity propagation clustering clusters the similar postures to an exemplar. The exemplars are the key postures of the action.
3. Summarization of all Radon transforms of the key postures is used for the description of an action. Because of high dimensions of the template, LDA is used to reduce the dimension of the description.
4. The conventional classifiers, including SMO, BayesNet and C4.5 are employed for training and testing. Experiments are carried out using leaving one subject out and leaving one sample out cross validation.

The benefit of using key postures for human action recognition is the reduced computation complexity. The advantages of using this method for human action recognition are listed below.

1. The computation complexity has been reduced dramatically by using key postures. This is because key postures can characterize the action well. The key postures representation for human action can be used not only for human action recognition but also for action retrieval.
2. The method does not require alignment between sequences. The selection of the starting posture is not a problem any more using this approach because we use one single template for matching.

References

1. Radon transform, http://en.wikipedia.org/wiki/Radon_transform
2. Altman, E.I., Marco, G., Varetto, F.: Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). New York University Salomon Center, Leonard N. Stern School of Business (1993)
3. Baraldi, A., Blonda, P.: A survey of fuzzy clustering algorithms for pattern recognition. II. IEEE Transactions on Systems, Man, and Cybernetics, Part B 29(6), 786–801 (1999)
4. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2 (2005)

5. Boulgouris, N.V., Hatzinakos, D., Plataniotis, K.N.: Gait recognition: a challenging signal processing technology for biometric identification. IEEE Signal Processing Magazine 22(6), 78–90 (2005)
6. Calic, J., Izuierdo, E.: Efficient key-frame extraction and video analysis. In: Proceedings of International Conference on Information Technology: Coding and Computing, pp. 28–33 (2002)
7. Chen, D.Y., Liao, H.Y.M., Tyan, H.R., Lin, C.W.: Automatic Key Posture Selection for Human Behavior Analysis (2005)
8. Cooke, T.: Two Variations on Fisher's Linear Discriminant for Pattern Recognition. IEEE Transactions On Pattern Analysis And Machine Intelligence, 268–273 (2002)
9. Deans, S.R.: The Radon Transform and Some of Its Applications. A Wiley-Interscience Publication, New York (1983)
10. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 726–733 (2003)
11. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science 315(5814), 972 (2007)
12. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press, London (1990)
13. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural networks 13(4-5), 411–430 (2000)
14. İkizler, N., Duygulu, P.: Human action recognition using distribution of oriented rectangular patches. In: Elgammal, A., Rosenhahn, B., Klette, R. (eds.) Human Motion 2007. LNCS, vol. 4814, pp. 271–284. Springer, Heidelberg (2007)
15. Jolliffe, I.T.: Principal component analysis. Springer, New York (2002)
16. Lim, I.S., Thalmann, D.: Swiss Federal Inst Of Technology Lausanne (Switzerland). Key-posture extraction out of human motion data by curve simplification (2001)
17. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: IEEE CVPR, pp. 1–8 (2007)
18. Martinez, A.M., Kak, A.C.: Pca versus lda. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(2), 228–233 (2001)
19. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.R.: Fisher discriminant analysis with kernels. In: Proceedings of the 1999 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing IX, pp. 41–48 (1999)
20. Pavlovic, V., Garg, A., Kasif, S.: A Bayesian framework for combining gene predictions*, pp. 19–27 (2002)
21. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines. Advances in Kernel Methods-Support Vector Learning, 208 (1999)
22. Quinlan, J.R.: Induction of decision trees. Machine learning 1(1), 81–106 (1986)
23. Radon, J.: Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. Berichte Sächsische Akademie der Wissenschaften, Leipzig, Mathematisch-Physikalische Klasse 69, 262–277 (1917)
24. Singh, M., Mandal, M., Basu, A.: Pose recognition using the Radon transform. In: 48th Midwest Symposium on Circuits and Systems, pp. 1091–1094 (2005)
25. Theodoridis, S., Koutroumbas, K.: Pattern recognition. Academic Press, London (2006)

26. Thurau, C.: Behavior histograms for action recognition and human detection. In: Elgammal, A., Rosenhahn, B., Klette, R. (eds.) Human Motion 2007. LNCS, vol. 4814, pp. 299–312. Springer, Heidelberg (2007)
27. Tominaga, Y.: Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. Chemometrics and Intelligent Laboratory Systems 49(1), 105–115 (1999)
28. Toyama, K., Blake, A.: Probabilistic tracking with exemplars in a metric space. International Journal of Computer Vision 48(1), 9–19 (2002)
29. Vapnik, V.: Estimation of dependences based on empirical data. Springer, Heidelberg (2006)
30. Wang, L., Suter, D.: Learning and matching of dynamic shape manifolds for human action recognition. IEEE Transactions on Image Processing 16(6), 1646–1661 (2007)
31. Wang, Y., Huang, K., Tan, T.: Human activity recognition based on r transform. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8 (2007)
32. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with Java implementations. ACM SIGMOD Record 31(1), 76–77 (2002)
33. Yeung, K.Y., Ruzzo, W.L.: Principal component analysis for clustering gene expression data, pp. 763–774 (2001)
34. Yu, H., Yang, J.: A direct LDA algorithm for high-dimensional data with application to face recognition. Pattern Recognition 34(10), 2067–2070 (2001)

2D and 3D Pose Recovery from a Single Uncalibrated Video

A View and Activity Independent Framework

Jean-Christophe Nebel, Paul Kuo, and Dimitrios Makris

Digital Imaging Research Centre, Kingston University

1 Introduction

Human pose recovery from video sequences is an important task in computer vision since a set of reconstructed body postures provides essential information for the analysis of human behaviour and activity. Although systems have been proposed, they all rely on either controlled environments involving several and, generally, calibrated cameras or motion models learned for specific scenarios. Unfortunately, these constrains are not suitable for most real-life applications such as the study of athletes' performances during competition, human computer interfaces for nomadic devices, video retrieval or the detection of antisocial behaviours from images captured from a closed-circuit television (CCTV) camera. Therefore, pose recovery remains a major challenge for the computer vision community.

The goal of pose recovery is to localise a person's joints and limbs in either an image plan (2D recovery) or a world space (3D recovery). The procedure usually results in the reconstruction of a human skeleton. A practical system should only be based on data recorded by a single uncalibrated camera. Moreover, in order to build a robust pose recovery framework, one has to address both the complexity of human poses, which includes a large posture space and self-occlusions, and the diversity in character appearance, which varies with individuals, clothing and viewpoints. The aim of this chapter is to present such a system. After a comprehensive literature review of the field of posture reconstruction from video data, the proposed pose recovery framework is outlined. It is divided into two essential modules. The first one focuses on 2D pose recovery. Not only is it an important step towards 3D pose recovery, but a 2D skeleton can also be used directly for applications such as linear gait analysis [57] and body part tracking [32]. The second module deals with the problem of 3D posture estimation. Since this task is particularly challenging when neither activity nor calibration parameters are known, some human biomechanics constrains need to be integrated in the framework.

2 Related Work

Human pose recovery has become a very active research topic in computer vision. Although the usage of optical markers led to the development of commercial

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 391–411.
springerlink.com

motion capture systems (mocap) [63], they impose constrains which are not acceptable in most computer vision applications. Similarly, methods based on data captured from several cameras have limited practicality [3, 5, 19, 34, 66]. Therefore, they will not be covered in the following review.

The ability to extract body configurations simply from a single video sequence has the potential to allow subtle analysis of human motion and even body language, which have applications in a wide range of domains including visual surveillance, human computer interaction, image retrieval and sports science. Although such studies can be performed using 2D human body cues alone when conducted in a controlled environment [57], the recovery of 3D postures provides much richer information. Since 2D pose recovery can be an essential step towards 3D pose recovery, these areas of research are covered separately in this state-of-art section. Finally, this review is completed with a section on evaluation of pose estimation methods.

2.1 2D Pose Recovery

Techniques aiming at 2D pose recovery are usually divided into two main categories [12]. Bottom-up approaches detect individual body parts and combine them to build a full body, whereas top-down ones start by identifying the whole human body shape before breaking it down into its compositional elements. Bottom-up methods exploit a variety of low level image cues to discover body parts. They can be detected by taking advantage of intrinsic partitions present in an image defined by edges [45, 50, 58]. An implementation of normalised graph cut (NCut) based on salient edges produced image pieces which were used to reconstruct human shapes following parsing rules [58]. Since limbs are often delimited by parallel edges, body part candidates can be identified by paring parallel lines according to anthropometric constraints [45, 50]. Ren et al. relied on refined edge maps where edges were, first, divided in linear segments and, then, constrained Delaunay triangulation was applied [50]. Parallel constraints were also included in an NCut based framework where body pieces were assembled using a dynamic programming approach [37]. Other cues, specific to individual body parts, have been combined to edge information to improve pose recovery: they include face/head, skin and limb detectors generated from training data [17, 55, 65]. For example, 2D poses were inferred by a data driven belief propagation Monte Carlo algorithm using a variety of images cues [17]. Alternatively, these cues were used to build a set of weak classifiers, which were combined within a more powerful meta-classifier using the Adaboost algorithm [65].

Despite the success of these bottom-up approaches, the absence of an explicit 2D body model may lead to the production of impossible human postures when the body part pairing process fails. Consequently, alongside these techniques, many top-down methods have been developed. They rely on 2D articulated body models that are collection of parts arranged in a deformable configuration. Initially, poses were estimated by simply minimising a cost function consisting of individual body parts and part paring [9]. This scheme was refined by, first, adding constraints of symmetry and colour homogeneity in body parts [47] and, secondly, tackling self-occlusion problem by using an extended body model containing occlusion likelihoods [56]. Such approach has also been used for body part tracking to initialise opportunistically trackers when a stylized pose is detected [46]. Although top-down techniques have proved efficient at recovering general poses, they tend not to provide accurate underlying body part segmentation. Consequently, a hybrid top-down/bottom-up approach, such as the one proposed in this chapter, should allow the production of plausible 2D poses with well defined body elements.

2.2 3D Pose Recovery

The process of 3D pose recovery from a single video sequence derives a 3D skeleton from high level 2D information, i.e. a 2D posture [23, 27, 36, 37, 38, 49, 59] or a silhouette [15, 24, 25, 26, 42, 43]. 3D pose reconstruction methods can be classified according to their reliance on data collected for specific motions. Activity-specific approaches focus on learning prior models of motions directly from carefully selected training data provided by motion capture systems. Among them, example based approaches explicitly sample the entire space of possible solutions and store the extracted 2D features with their corresponding 3D poses. In such a framework, the recovery of a 3D posture is performed by interpolating between a set of 3D poses whose 2D features match the most the input query. This method was applied successfully by Poppe where silhouettes collected from various viewpoints were used as high level 2D information [43]. The main limitation of these techniques is that a very large training set is required to provide satisfactory accuracy and generalisation properties. In contrast, dimensionality reduction techniques learn an informative and compact representation of the training set which is then used to recover 3D poses with the usage of mapping functions. Complexity of human 3D motion requires nonlinear dimensionality reduction techniques to recover low dimension embeddings of original data. Mapping between image and 3D posture spaces can be intrinsic to a dimensionality reduction technique, e.g. Gaussian process latent variable models (GPLVM) [48], or learned in an additional step in so-called spectral methods [2, 52, 60]. The elegant framework provided by GPLVM has made it recently extremely popular in the field [14, 24, 26]. However, because of its expensive computational cost, practical applications have been limited to small training data sets which compromise accuracy and generalisation properties. Since the complexity of spectral methods is proportional to the size of the training set, they are a suitable alternative to GPLVM if appropriate mapping between embedded and initial data space can be calculated. Lee and Elgammal learned a nonlinear mapping through generalized radial basis function mapping between the silhouette space and the torus manifold they used to represent walking motion [25, 26]. Since they were able to generate and handle a very large synthetic training data set, they managed to perform very accurate 3D body pose tracking.

In many practical applications of 3D pose recovery such as visual surveillance, there is no limitation in the type of activities characters can be involved in. Moreover, there is often a particular interest in detecting uncommon behaviours. Therefore, in this context, activity-specific approaches are clearly not suitable. Since geometric camera calibration reveals the relationship between the 3D space and its projection on the image plane, this is a line of research which has potential for the reconstruction of a 3D articulated structure in an activity independent framework. In a typical visual surveillance scenario where a CCTV network is exploited, cameras cannot be calibrated manually. Consequently, calibration parameters have to be estimated automatically. In order to simplify this task, Taylor offered a pose recovery method based an orthographic projection model that assumes 3D objects are far away from the camera and thus the depth of their surface points is almost constant [59]. Although this approach has been widely used [36, 38, 49], quantitative evaluation of their pose estimates revealed accuracy is seriously compromised by such a strong assumption [23]. Another approach proposes to compute 3D pose using inverse kinematics based on prior knowledge about pose distribution. Lee and Cohen presented a data-driven iterative approach, where pose candidates are generated in Markov chain Monte Carlo search guided by image observations [27]. Despite a high computational cost, accuracy of posture estimates provided by this system is too low for many applications. This review suggests the most promising way of predicting 3D postures without activity constrain is to exploit geometric camera calibration if parameters can be estimated automatically without making unrealistic assumptions.

2.3 Validation of Pose Recovery Algorithms

Although some evaluation of pose estimates can be done qualitatively, quantitative evaluation against ground truth data is eventually required. While 2D joint positions can be generated painstakingly using ground truth authoring tools such as ViPER [31], it is not possible to produce manually 3D ground truth from video sequences. Fortunately, a few research groups have made available to the pose recovery research community data sets including ground truth data. The most popular of these data sets is called HumanEva (HE) [54] and is now a de facto benchmark for pose recovery. It consists of a set of videos consisting of around 100,000 frames associated with motion capture data which were collected synchronously. Therefore, mocap data provides the 3D ground truth of human poses. Moreover, since video cameras were calibrated, 3D data points can be projected onto the image plane so that joint 2D locations are available for evaluation of 2D pose recovery algorithms. In addition, a standard set of error metrics has been defined so that results can be compared between research teams. For example, the error, *E*, between a pose estimate *X'* and the ground truth *X*, where a body configuration is defined by a set of *n* joints $X=\{x_1,x_2,\ldots,x_n\}$, is expressed by the average of the absolute distances between recovered joints,

$$E(X, X') = \sum_{i=1}^{n} \frac{\delta_i \left\| x_i - x'_i \right\|}{\sum_{j=1}^{n} \delta_j} \tag{1}$$

where $\delta_i=1$ if the joint *i* can be estimated by the evaluated method.

HumanEva video sequences show indoor scenes where a variety of human subjects (males and females) perform different types of motions including walking, jogging, balancing, jumping and boxing. Since the walking sequences are performed around a rectangular carpet, they provide a particularly interesting testing environment where walking cycles are seen from a large variety of viewpoints and distances. Although not as complete as the HumanEva one, other publicly accessible data sets prove very useful for pose estimation. They include outdoor walking sequences produced by H. Sidenbladh [53], the MuHAVi data set of primitive actions [39] and the CMU motion capture library [7].

In this chapter, presented algorithms are illustrated and quantitatively evaluated using HumanEva sequences.

3 Pose Recovery Framework

Although many optical markerless systems have been proposed for 2D/3D pose recovery, they all rely on either data captured in a controlled environment or the assumption that characters are involved in specific activities. Consequently, those schemes are not suitable for many practical applications such as visual surveillance where images are usually produced by a single uncalibrated camera and human motion is unrestricted. In order to deal with such a challenging environment, a novel pose recovery framework is proposed. In this approach, constraints are limited to those which can be justified as being legitimate within visual surveillance scenarios.

The structure of the suggested system (see Fig. 1) can be divided in four main modules. Initially, the sole input is a video sequence captured by a single uncalibrated camera. From this, 2D postures are automatically extracted and 2D skeletons are generated. Then, key poses are used to calculate camera parameters. Finally, this information is exploited to generate a set of 3D pose estimates for each video frame. The last part of the methodology is concerned with the allocation of a unique 3D pose to each image. This is achieved by selecting among a set of estimates the posture which optimises some cost function inferred from prior knowledge.

This information may involve general 3D physics-based human kinematics models [4], which would enforce physically plausible postures and realistic dynamics properties between poses, or pose libraries built from the capture of body configurations representing human motion space [7]. Since this chapter is focused on presenting a non-learning based approach for the production of 3D pose estimates, pose selection will not be addressed. Interested readers should refer to the following review for more details on using prior knowledge to reduce 3D pose space [44].

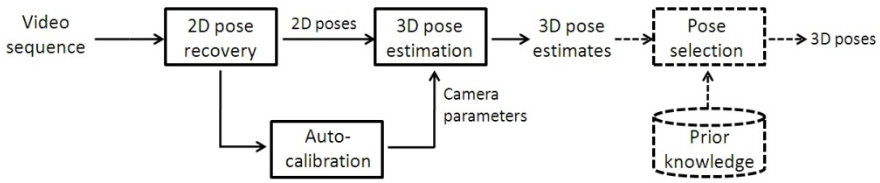


Fig. 1. Pose recovery framework (elements in dotted lines are outside the scope of this chapter)

4 2D Pose Recovery

The goal of 2D pose recovery is to localise a character's joints and limbs in the image plane to generate a human skeleton. However, since occlusions may prevent to access some or any visual information regarding the location of a specific body part, it is essential for a robust pose recovery system to provide a confidence measure with every pose estimate. The suggested approach relies on a probabilistic framework which clusters foreground pixels to identify body parts. This partition process is supported by the introduction of a human body model to impose some anthropometric constraints.

The flow diagram of the pose recovery algorithm is shown in Fig. 2. The only input is a video sequence from which image cues are extracted and associated to every foreground pixel. Next, these pixels are partitioned into a set of clusters corresponding to the expected number of body parts. In order to label these clusters and ensure the production of a plausible posture, an adjustable 2D human body model is fitted on the foreground partition. Then, pixel body part labels are integrated into the pixel feature vector. Finally, the clustering and model fitting processes iterate until a stable body configuration is found. A 2D pose is then generated with its associated confidence measure.

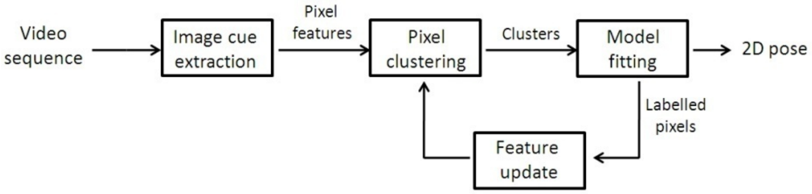


Fig. 2. 2D Pose recovery algorithm

4.1 Image Feature Extraction and Clustering

A robust algorithm aiming at recovering postures from a single camera must rely on as many relevant image features as possible to be able to deal with the largest variety of views, positions and character appearances. Initially, foreground pixels are selected using an advanced motion segmentation algorithm dealing with shadow removal [35] as illustrated in Fig. 3(b). Then, each of these pixels is associated to a feature vector describing its location, individual motion, orientation and value, i.e. grey level or colour. These cues were selected because they tend to exhibit homogeneity within a body part and help discriminate between different limbs.

Location and individual motion express the property that body parts are made out of a continuous set of pixels moving in a continuous manner, except when occlusions occur. Whereas pixel position is given, their motion, i.e. speed and direction, is computed using a standard optical flow algorithm providing dense motion information [28]. To deal with noisy data, the motion map is smoothed using a moving-average temporal filter.

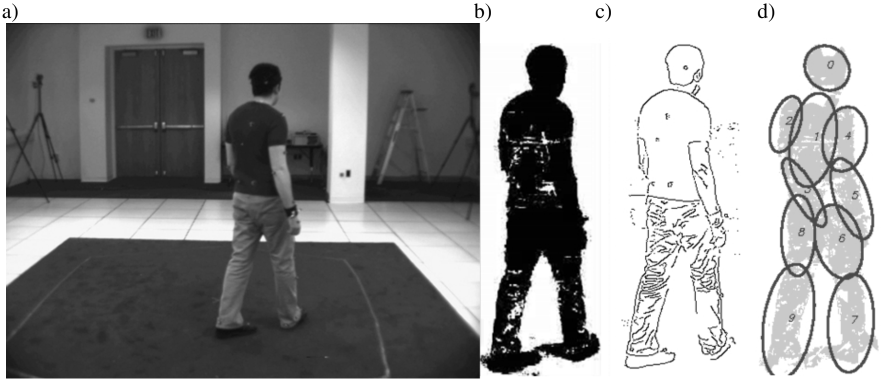


Fig. 3. a) Input image from sequence S1_Walking_C1 of HumanEva data set, b) foreground segmentation, c) edge detection and d) partition of pixel features in 10 clusters, where a 2-standard deviation boundary is used to represent each cluster as an ellipse.

Orientation reflects the fact that limbs can be modelled using parallel lines corresponding to the underlying skeleton [50]. In order to associate a direction to each foreground pixel, edges are detected using a Canny Edge detector, see Fig. 3(c), then they are converted to line segments via Hough transform and, finally, their orientation is interpolated to all foreground pixels.

Finally, pixel value is used to discriminate between body parts since each of them can usually be modelled by either homogenous colour/grey value or a low number of colour/grey value patterns [47]. In the case of colour values, since experiments showed that the colour space choice does not affect results in body part detection, colour cues can be simply expressed by their RGB values.

Clustering is performed using Gaussian Mixture Models (GMMs) in the high dimensional space or ‘cue space’ defined by the pixel features previously mentioned and pixel associations to each body part as evaluated during the model fitting process. These labels constrain pixel partition towards clusters which display a topology compatible with a human body configuration.

The choice of GMM clustering is dictated, first, by its ability to deal with clusters that have different sizes and may be correlated, and, secondly, by its probabilistic nature that allows the estimation of a confidence measure for pose recovery (detailed description of that measure is provided later). GMM clustering partitions foreground pixels in the cue space into as many clusters as there are body parts, *n*. Therefore, a set of *n* probabilities, $P(p_i|C_j),where\ \ j\in\ [0..\ n]\ and\ \Sigma_i P(p_i|C_j)=1$, is produced for each foreground pixel p_i, indicating the likelihood of a pixel belonging to each of the *n* clusters, C_j. Fig. 3(d) illustrates some partition results.

GMM clustering is usually initialised by K-means clustering [13]. However, since this algorithm does not always produce optimal partitions, it must be initialised either many times with random seeds or with reasonable estimates of cluster centres. Given the prior knowledge the foreground pixels belong to a human body, the overlapping of a 2D articulated model on these pixels and the projection of the body part centres in the cue space provide reasonable seeds for K-means clustering.

4.2 2D Body Model Fitting

The aim of this process is to impose anthropometric constraints to the clustering process and to eventually produce a 2D pose estimate defined by a 2D skeleton. This hierarchical procedure optimises some overlapping costs by fitting a 2D articulated model on the clustered pixels.

As commonly used in top-down pose recovery approaches, the 2D generic human body model (*M*) consists of 10 body pieces, $M=\{ m_{head,} m_{torso,} m_{lua}, m_{lla}, m_{rua}, m_{rla}, m_{rul}, m_{rll}, m_{lul}, m_{lll}\}$[1] represented by basic shapes [45, 54, 66]: a circle for the head, a rectangle for the torso and ellipses for the eight limbs, see Fig. 4(a). Initially, the model is constructed using standard body part ratios [8] and its scale is estimated using the height of the segmented foreground. However, during the iterative clustering and model fitting process, the size of each body part is adjusted to accommodate any viewpoint and posture. Lengths of limbs are calculated from the produced clusters as the Euclidean distance between the joints of adjacent clusters. The position of a joint, $J_{j\text{-}k}$, between two clusters C_j and C_k is estimated using the conditional probabilities produced by GMM clustering:

$$J_{j-k} = \arg\max_i \left\{ P(p_i \mid C_j) + P(p_i \mid C_k) - \left| P(p_i \mid C_j) - P(p_i \mid C_k) \right| \right\} \tag{2}$$

where p_i are the foreground pixels.

The algorithm performing the fitting of the 2D model onto the clustered pixels starts by detecting the most reliable body parts, i.e. head and then torso, before dealing with the limbs. The location of the head is estimated using the omega head detection algorithm [67] (see Fig. 4(b)). This method searches for an Ω-shaped model, which represents head and shoulders, in the image by minimising the Chamfer distance between the model and image edges. Then, the torso is detected by modelling the values of its pixels using GMM as proposed by Mckenna et al. [33]. This relies on using a sample region of the torso which can be inferred using the locations of the head and foreground pixels which are already known (see Fig. 4(c)). After discarding pixels which, statistically, appear to be outliers [10], the GMM is trained using the pixel values drawn from this sampling region. Then, torso pixels are detected from the segmented foreground by the trained GMM, as shown in Fig. 4(d). Finally, the torso region is approximated by a rectangular torso model, see Fig. 4(e), whose position, orientation, scale and height/width ratio are optimised according to the overlap between pixels belonging to the model and detected torso pixel

$$Overlap = \frac{A_{m_{torso}} \cap A_{pixel_{torso}}}{\sqrt{A_{m_{torso}} A_{pixel_{torso}}}} \tag{3}$$

[1] Apart from head and torso pieces, parts' names are abbreviated by 3 letters denoting: "left" or "right", "upper" or "lower" and "arm" or "leg".

where $A_{m_{torso}}$ and $A_{pixel_{torso}}$ denote the pixel area of the rectangle model and the number of detected torso pixels respectively.

The final stage of 2D model fitting recovers limb configuration. Each limb is translated and rotated to maximise the overlapping costs of the body partitions, i.e. clustered pixels. This is achieved by maximising the joint probabilities, as defined in Equation (6), between limbs and partitions. The expression of joint probabilities is detailed in the next section. Fig. 4(f) illustrates the result of fitting the model on the foreground.

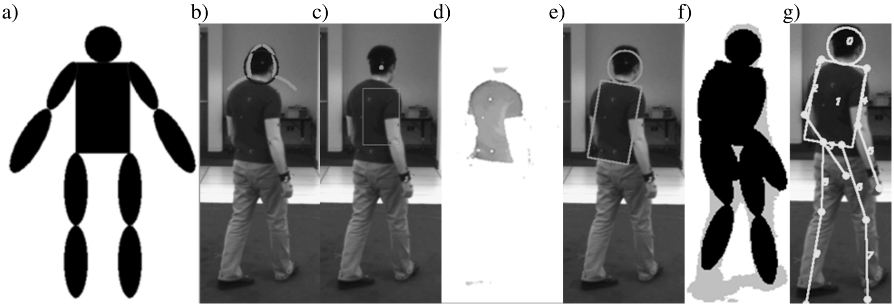


Fig. 4. 2D model fitting; a) 2D generic human body model, b) omega head detection, c) torso sample region, d) detected torso pixels, e) fitted head and torso f) fitted model and g) 2D skeleton.

4.3 Production of 2D Pose Estimates

Once the iterative processes of clustering and model fitting converge, a skeleton, as shown in Fig. 4(g), is extracted from the final body model where clusters' boundaries along principal axes define body joints. Fig. 5 shows 2D pose estimates produced for various views and activities, i.e. walking, running and balancing sequences.

In addition to a pose estimate, a confidence measure is calculated to rate its accuracy. It is formulated as the probability that a pose is recovered successfully, *P(pose)*. If one assumes this is determined by the success of recovering all body parts and their associated recovery probabilities are independent, it can be expressed by:

$$P(pose) = \prod_{j} P(X_j) \tag{4}$$

where $P(X_j)$ denotes the probability of body part, X_j, to be recovered successfully. This is evaluated by extending the definition of the overlap measure (Equation 3) to other body parts.

$$P(X_j) \sim Overlap\ (m_j, C_j) = \frac{A_{m_j} \cap A_{C_j}}{\sqrt{A_{m_j} A_{C_j}}} \tag{5}$$

where A_{m_j} and A_{C_j} denote the pixel areas of the model part m_j and cluster C_j respectively. Since the cluster C_j is defined by GMM clustering over the entire foreground pixels, F, its pixel area is conceptually equivalent to the sum of the probabilities of foreground pixels $p_i \in F$ belonging to that cluster:

$$A_{C_j} \sim \sum_{p_i \in F} P(p_i / C_j) \tag{6}$$

Similarly, $A_{m_j} \cap A_{C_j}$ corresponds to the sum of the probabilities of model pixels $p_i \in m_j$ belonging to that cluster:

$$A_{m_j} \cap A_{C_j} \sim \sum_{p_i \in m_j} P(p_i / C_j) \tag{7}$$

Therefore, $P(X_j)$ is expressed by

$$P(X_j) \sim \frac{\sum_{p_i \in m_j} P(p_i / C_j)}{\sqrt{A_{m_j} \sum_{p_i \in F} P(p_i / C_j)}} \tag{8}$$

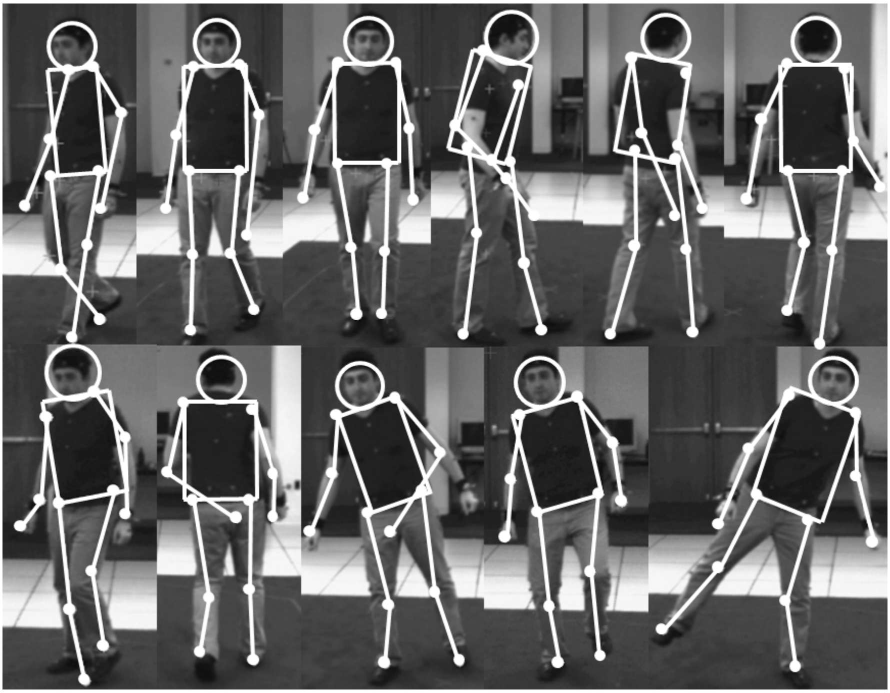


Fig. 5. 2D pose estimates produced for various views and activities: walking (1st raw) and running and balancing (2nd raw) sequences

Experiments conducted on a set of HumanEva sequences show this activity independent algorithm achieves an average accuracy of 25 pixels per joint [22]. Moreover, results show good correlation (r=0.7) between accuracy and confidence measure. Therefore, selection of estimated poses according to this metric allows significant increase of precision, e.g. poses whose confidence values are among the top 10% have an accuracy below 20 pixels.

It is important to note this confidence value is not only useful for pose evaluation but also for many applications built upon pose recovery. For example, body part tracking using either Kalman or Particle filter requires a prior probability to estimate how much an observation can be trusted [32].

The proposed method, which does not require any training, produces quantitatively and qualitatively convincing results. Although recent learning based approaches [16, 18, 25, 43] achieve 5-15-pixel error for HumanEva data, they are limited to the estimation of poses which are present in their training set. Consequently, they are not suitable for most realistic scenarios. Although other activity independent approaches have been suggested [45, 50, 58], only qualitative results were reported, which do not allow objective comparisons. Finally, activity independent pose tracking was proposed and tested on HumanEva sequences [32]. They achieved an average error of 13 pixels for this easier task which relies on manual initialisation.

5 3D Pose Recovery

The aim of human 3D pose recovery is the production of a 3D skeleton representing angles between adjacent joints. Given that the process has to be suitable for applications such as visual surveillance where a character's activity is largely unknown, the proposed methodology is based on transforming 2D image points into a set of 3D poses in real world using pinhole projection. However, since this problem is ill-constrained, multiple postures are generated. Consequently, a pose selection mechanism is then required to extract the correct posture.

A conceptual flow diagram of the 3D pose recovery algorithm is shown in Fig. 6. From a sequence of 2D skeletons, key poses as defined by a human biomechanics constrain are extracted. Then, the camera is automatically calibrated at these instants. Subsequently, joint 3D positions are calculated for those key frames using camera parameters and pinhole projection. Finally, other 3D poses are estimated by propagating key posture 3D information using another human bipedal motion constraint.

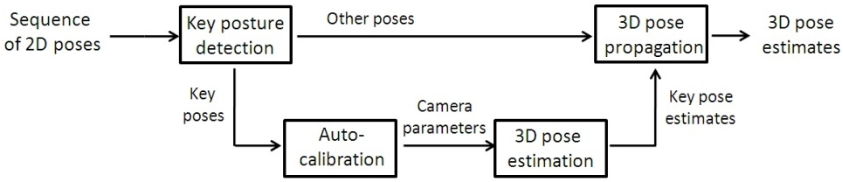


Fig. 6. Conceptual framework for production of 3D posture estimates

5.1 Camera Auto-calibration from Mid-stance Position

In order to perform 3D pose reconstruction based on pin-hole projection, camera calibration parameters need to be estimated. The most commonly used technique was proposed by Tsai [61] and is based on known correspondences between 3D points and their 2D image plane projections. In practice, this method requires capturing images of a calibration object of known geometry. Consequently, it does not appear suitable for many applications where frequent and physical access to camera is not possible. As a consequence, research effort has been invested in developing auto-calibration techniques. Solutions have been proposed exploiting camera motion [1, 29, 41]. For visual surveillance scenario, where the camera is usually fixed, research focus has been on taking advantage of the observed activity within a scene. Although many methods have been proposed [20, 30, 51], they all impose rather strong constraints on either pedestrian activity or 3D scene geometry.

In order to deal with realistic environments, a more general approach has to be developed. The technique presented here proposes to estimate camera calibration parameters by only using a generic property learned from human biomechanics. First, the camera model is simplified using common assumptions: the principal axis goes through the centre of the image and there is neither lens distortion nor skew. The required projection parameters, i.e. the focal length and the camera's relative position to the object of interest, are then computed using Tsai's coplanar calibration model [61]. This requires a set of correspondences between 3D coplanar points and their projected locations on the image plane. Study of human biomechanics reveals that, during a cycle of human bipedal motion, there is an instant, which is called the mid-stance position (Fig. 7), when shoulders and hips become coplanar. As a result, shoulder and hip joint locations are suitable for coplanar calibration if mid-stance postures are detected.

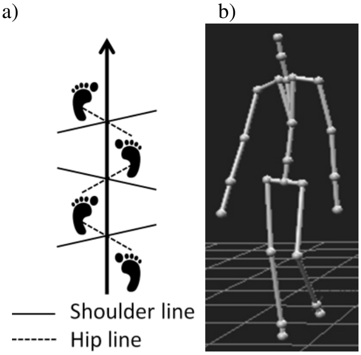


Fig. 7. a) Variation of hip-shoulder angle during a walking cycle, b) mid-stance position

The detection of these postures within a video sequence is achieved through estimation of the calibration parameters following Tsai's two step algorithm using a set of 5 coplanar point candidates. During the first stage, these points permit the evaluation of 5 extrinsic parameters, i.e. 3D rotation and image centre coordinates. These values are then exploited in the second stage to resolve the ambiguity between the last extrinsic parameter, i.e. depth, and the effective focal length. This relies on solving an over-determined linear system, which produces 10 estimates of the parameter pairs. Since experiments have shown that there is correlation between the variability of these estimates and the coplanarity quality of the 5 coplanar point candidates [21], this calculation is used to jointly detect mid-stance frames and evaluate the size of the associated 3D models and, eventually, to produce the camera parameters required for 3D reconstruction.

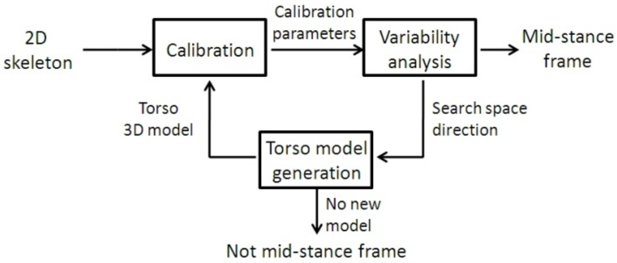


Fig. 8. Algorithm allowing detection of mid-stance position from a 2D skeleton

Fig. 8 describes the main steps of the camera auto-calibration algorithm which is applied to every frame. From each 2D skeleton, 5 torso points, e.g. shoulder and hip joints and mi-hip point, are used as coplanar point candidates. Using a generic torso 3D model, calibration parameters are calculated using Tsai's algorithm. Then the variability of the 10 focal length estimates is analysed. If their standard deviation is below a given threshold, the frame is labelled as a mid-stance frame whose associated calibration parameters and torso 3D model will be used for 3D pose reconstruction. Alternatively, variability is used to guide the model search within the torso space. If no new model candidate can be produced, the process is stopped and the frame will not be associated to calibration parameters. Otherwise a new 3D model is produced, camera parameters are re-calculated and the process iterates. Finally, once all frames have been processed, a set of mid-stance frames is available. Their associated focal lengths and torso 3D models are averaged and a generic 3D full body model is adjusted to fit the torso model requirements.

Experiments were conducted with a variety of walking and running sequences seen from many different camera angles and lenses [21]. They reveal that, if 2D joint positions are evaluated accurately, the camera focal length is usually predicted within 2% of its actual value.

5.2 Keypose Reconstruction Using Projection Model

Key poses have been identified and they are associated with camera parameters and a 3D articulated model. Therefore, their 2D joint positions can be projected in the world 3D coordinate system using the pin-hole projection model. This method relies on constructing projections lines from the camera position, i.e. optical centre through each image point as illustrated in Fig. 9(a). Then, joint 3D coordinates are localised on these lines using the camera-object distance and the articulated model as constraints.

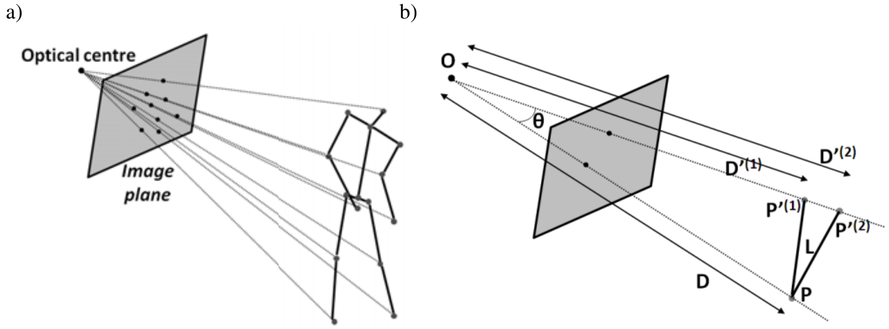


Fig. 9. a) Projected from image plane to world 3D coordinate system using the pin-hole projection model, b) Joint 3D reconstruction based on known coordinates of adjacent joint.

Since the 3D positions of shoulder and hip points were estimated in the calibration process – they form the required 3D coplanar model -, only limbs (upper/lower arms/legs) and head need to be reconstructed. Their positions, *P'*, are estimated in a hierarchical manner using known joint locations, *P*, as starting points. As illustrated on Fig. 9(b), the distance, *D'*, between the optical centre, *O*, and the joint position to estimate can be expressed using simple trigonometry by the distance, *D*, between the optical centre and a known joint position, the angle, θ, between their projection lines and the length of the limb they define, *L*:

$$D'^2 - 2D'D\cos\theta + (D^2 - L^2) = 0 \tag{9}$$

Since projection from image plane to 3D world coordinate is an ill-posed problem, a quadric equation is obtained. Consequently, each estimation of joint position generates up to two valid solutions, which leads to the production of up to2^9 pose estimates for a given frame. As mentioned earlier, pose selection among those estimates can be achieved using, for example, 3D physics-based human kinematics models or pose libraries. In the following sections, it is assumed the problem of pose selection has been solved.

5.3 3D Pose Estimation Based on Support Foot Propagation

The general idea is to also apply equation (5) for 3D reconstruction of frames which are not defined as displaying mi-stance positions. However, since it requires the knowledge of the 3D coordinates of at least one point of the 3D skeleton and calibration is not possible, some new constraint has to be introduced.

Further study of human biomechanics reveals that most human bipedal motions such as walking, loitering, balancing and dancing, rely on a series of 'steps' where one leg 'swings' around a 'support' leg which holds the body weight [11] (see Fig. 10). Since the 'support' foot stays in contact with the ground during the whole 'step', in such motion, there is permanently a point whose 3D position is identical between two successive frames. Therefore, once a posture has been reconstructed such as a key pose the following and previous frames can also be processed by starting their reconstruction from the position of the support foot. This propagation process can be refined by combining support foot estimates coming from the previous and next key frames.

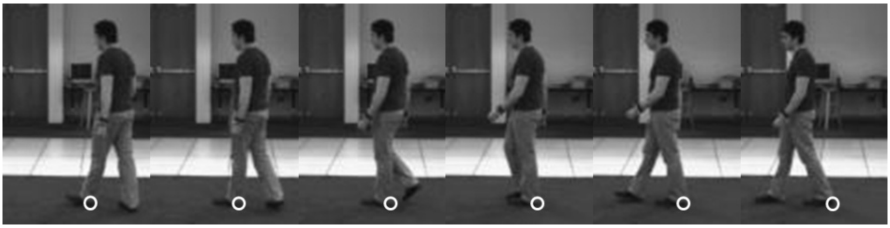


Fig. 10. Right leg swinging around the left leg whose foot is static during the whole step

Since a sequence of 2D skeleton is available at the start of the 3D reconstruction process, detection of the support foot is straight forward: it only relies in comparing foot positions between consecutive frames. Although this constrain does not fully hold for some activities such as running, when for a brief instant both feet leave the ground, this can be addressed by either interpolating between poses which meet the constrain requirements or by constructing the 3D pose from the foot which has the lowest velocity and is assumed to be immobile. Although, whatever the selected strategy, reconstruction accuracy suffers, the error is not propagated for long since reconstruction is reinitialised with the detection of the next mid-stance position.

5.4 3D Pose Recovery for a Walking Sequence

Fig. 11 illustrates the performance of the proposed 3D pose recovery methodology applied to a sequence of "walking in a circle" - S2 Walking (C1) - from the HumanEva data set [54]. Here, it is assumed the pose selection process successfully identified the most relevant posture from the set of estimates. Since motion capture data were collected synchronously with video data and cameras were calibrated, 2D locations of key body points in the sequences were extracted

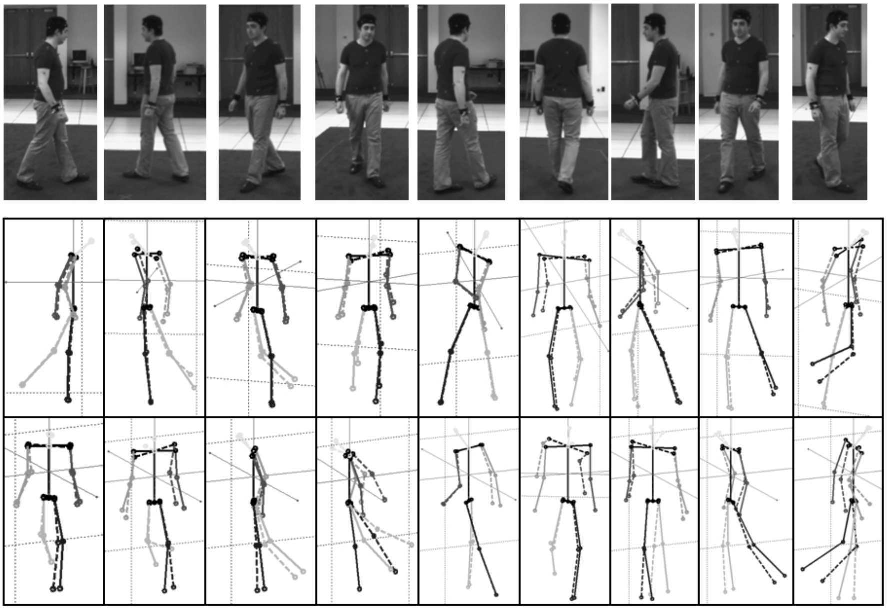


Fig. 11. Reconstruction results; first row: walking sequence images; second and third row: reconstructed (solid) and ground truth (dotted) postures observed from the original viewpoint and a novel viewpoint respectively. The first 4 images are detected as mid-stance frames.

automatically by projecting mocap 3D points on the image plane. In Fig. 11, the first 4 columns show images which were detected as mid-stance frames and their 3D reconstruction is compared to the 3D ground truth which corresponds to the mocap data. The other columns demonstrate the ability of reconstructing other frames using support foot propagation.

In order to quantitatively evaluate the proposed algorithm, performance are provided using, first, the camera calibration parameters provided with the data set and, secondly, the auto-calibration scheme. Table 1 display these results along those achieved by other state of art 3D pose recovery techniques applied on the HumanEva data set. The 12-mm precision obtained by Cheng [6] using multiple manually calibrated cameras can be seen as the maximal accuracy that a single camera computer vision system could achieve on this data set.

Most methods are activity specific and recover poses with a 3-4-cm accuracy [25, 40, 43, 62, 64]. However, as Poppes's results show, the selection of training data is essential: in their approach reconstruction error doubles if the tested subject is not present in the training set. 3D reconstruction without any training data is a more difficult task which has been investigated by fewer groups. Fig. 12 provides a frame base accuracy comparison between activity independent methods, i.e. orthographic projection [59] and the proposed approach with auto-calibration or

Table 1. Performance of 3D pose recovery algorithms

Algorithm	Error (mm)	Constraints	Training
Calibrated camera	21	Manual calibration	No
Proposed	80	Bipedal motion	No
Taylor [59]	236	None	No
Lee [25]	31	Activity specific & cyclic	Yes
Urtasun [62]	33	Activity specific	Yes
Vondrak [64]	34	Activity specific	Yes
Okad [40]	38	Activity specific	Yes
Poppe [43]	40	Activity specific Subject in training set	Yes
Poppe [43]	80	Activity specific	Yes
Cheng [6]	12	Multiple calibrated cameras Manual initialisation	No

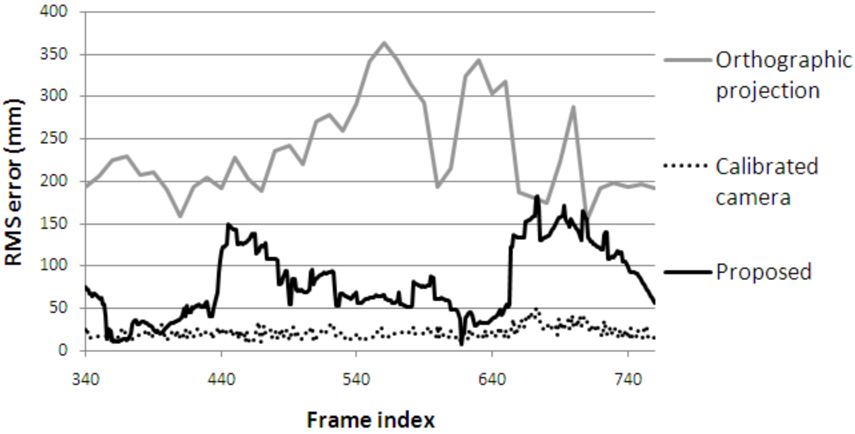


Fig. 12. Accuracy comparison between activity independent methods

calibrated camera. The orthographic projection model offers only a 24-cm accuracy, the proposed approach achieves a 8-cm accuracy with auto-calibration; this accuracy is usually sufficient to label poses for action recognition applications [23]. Though training based approaches perform better, their applicability is much more limited. Moreover, results with a calibrated camera show, camera model based approaches have the potential to outperform training based ones. This is consistent with the fact that pose recovery processes based on training data tend to smooth out stylistic variations of a specific motion.

6 Conclusions

This chapter describes a framework for 2D and 3D pose recovery from a single uncalibrated video. Since this methodology relies on generic human biomechanics constrains, which are valid in most bipedal motions, it can be considered as view and activity independent. Therefore, this approach is suitable for many real life applications, such as visual surveillance, where neither scenario nor environment can be controlled.

Many extensions of this framework can be envisaged. The probabilistic nature of the 2D pose recovery process makes it particularly suited to be incorporated within body part trackers which, in addition to initialisation, usually require prior probabilities about observations. Such scheme would undoubtedly improve the accuracy of 2D pose estimates since predictions provided by the tracker could be integrated in the feature vectors used for identifying body parts.

The proposed method for 3D posture recovery could also be enhanced. First of all, a pose selection module needs to be developed exploiting human kinematics, pose consistency and any available prior knowledge to reduce 3D pose space. Furthermore, since 3D motion analysis based on extracted 3D postures could infer character's activity, poses could be refined using relevant learned based activity models if available. Similarly, linear motion could be detected. Then, calibration methods relying on this constrain could be applied to extend the number of frames where camera parameters could be estimated.

References

[1] Armstrong, M., Zisserman, A., et al.: Euclidean Reconstructing from Image Triplets. In: Buxton, B.F., Cipolla, R. (eds.) ECCV 1996. LNCS, vol. 1064, pp. 3–16. Springer, Heidelberg (1996)

[2] Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 15(6), 1373–1396 (2003)

[3] Bhatia, S., Sigal, L., et al.: 3D Human Limb Detection using Space Carving and Multi-View Eigen Models. In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 17–24 (2004)

[4] Brubaker, M., Fleet, D., et al.: Physics-Based Human Pose Tracking. In: Proc. NIPS Workshop on Evaluation of Articulated Human Motion and Pose Estimation (2006)

[5] Caillette, F., Howard, T.: Real-Time Markerless Human Body Tracking with Multi-View 3-D Voxel Reconstruction. In: Proc. British Machine Vision Conf. (2004)

[6] Cheng, S., Trivedi, M.: Articulated Body Pose Estimation from Voxel Reconstructions using Kinematically Constrained Gaussian Mixture Models: Algorithm and Evaluation. In: Proc. Workshop on Evaluation of Articulated Human Motion and Pose Estimation (2007)

[7] Carnegie Mellon University, Motion Capture Library, `http://mocap.cs.cmu.edu` (last accessed January 2010)

[8] Da Vinci, L.: Description of Vitruvian Man (1492)

[9] Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. Int. J. Computer Vision 61(1), 55–79 (2005)

[10] Fritsch, J., Kleinehagenbrock, M., et al.: Audiovisual person tracking with a mobile robot. In: Proc. Int. Conf. on Intelligent Autonomous Systems, pp. 898–906 (2004)

[11] Fryer, C.: Biomechanics of the lower extremity. Instruct Course Lect. 20, 124–130 (1971)

[12] Gavrila, D.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding 73(1), 82–98 (1999)

[13] Hartigan, J., Wong, M.: A K-means clustering algorithm. Applied Statistics 28(1), 100–108 (1979)

[14] Hou, S., Galata, A., et al.: Real-time Body Tracking Using a Gaussian Process Latent Variable Model. In: Proc. Int. Conf. on Computer Vision (2007)

[15] Howe, N.: Silhouette lookup for monocular 3D pose tracking. Image and Vision Computing 25(3), 331–341 (2007)

[16] Howe, N.: Recognition-Based Motion Capture and the HumanEva II Test Data. In: Proc. Workshop on Evaluation of Articulated Human Motion and Pose Estimation (2007)

[17] Hua, G., Yang, M., et al.: Learning to estimate human poses with data driven belief propagation. In: Proc. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 747–754 (2005)

[18] Husz, Z., Wallace, A., Green, P.: Evaluation of a Hierarchical Partitioned Particle Filter with Action Primitives. In: Proc. Workshop on Evaluation of Articulated Human Motion and Pose Estimation (2007)

[19] Izo, T., Grimson, W.: Simultaneous pose recovery and camera registration from multiple views of a walking person. J. Image and Vision Computing 25(3), 342–351 (2007)

[20] Krahnstoever, N., Mendonca, P.: Bayesian autocalibration for surveillance. In: Proc. Int. Conf. on Computer Vision (2005)

[21] Kuo, P., Nebel, J.-C., et al.: Camera Auto-Calibration from Articulated Motion. In: Proc. Advanced Video and Signal Based Surveillance, pp. 135–140 (2007)

[22] Kuo, P., Makris, D., et al.: Integration of Local Image Cues for Probabilistic 2D Pose Recovery. In: Proc. Int. Symp. on Visual Computing (2008)

[23] Kuo, P., Thibault, A.: Exploiting Human Bipedal Motion Constraints for 3D Pose Recovery from a Single Uncalibrated Camera. In: Proc. Int. Conf. on Computer Vision theory and Applications (2009)

[24] Lawrence, N., Carl Henrik, E., et al.: Gaussian Process Latent Variable Models for Human Pose Estimation. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 132–143. Springer, Heidelberg (2008)

[25] Lee, C.-S., Elgammal, A.: Body Pose Tracking From Uncalibrated Camera Using Supervised Manifold Learning. In: Proc. NIPS Workshop on Evaluation of Articulated Human Motion and Pose Estimation (2006)

[26] Lee, C.-S., Elgammal, A.: Nonlinear manifold learning for dynamic shape and dynamic appearance. Computer Vision and Image Understanding 106(1), 31–46 (2007)

[27] Lee, M., Cohen, I.: A Model-Based Approach for Estimating Human 3D Poses in Static Images. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(6), 905–916 (2006)

[28] Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proc. Imaging Understanding Workshop, pp. 121–130 (1981)

[29] Luong, Q., Faugeras, O.: Self-Calibration of a Moving Camera from Point correspondences and Fundamental Matrices. Int. J. Computer Vision 22(3), 261–289 (1997)

[30] Lv, F., Zhao, T., et al.: Self-Calibration of a Camera from Video of a Walking Human. In: Proc. Int. Conf. on Pattern Recognition (2002)

[31] Mariano, V., Min, J., et al.: Performance Evaluation of Object Detection Algorithms. In: Proc. Int. Conf. on Pattern Recognition, pp. 965–969 (2002)

[32] Martinez-del-Rincon, J., Nebel, J.-C., et al.: Tracking Human Body Parts Using Particle Filters Constrained by Human Biomechanics. In: Proc. British Machine Vision Conf.

[33] Mckenna, S., Raja, Y., et al.: Tracking colour objects using adaptive mixture models. Image and Vision Computing 17, 225–231 (1999)

[34] Menier, C., Boyer, E., et al.: 3D Skeleton-Based Body Pose Recovery. In: Proc. Int. Symp. on 3D Data Processing, Visualization and Transmission (2004)

[35] OpenCV 2.0 C Reference, `http://opencv.willowgarage.com/documentation/index.html` (last accessed January 2010)

[36] Mori, G., Malik, J.: Estimating Human Body Configurations Using Shape Context Matching. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 666–680. Springer, Heidelberg (2002)

[37] Mori, G., Ren, X., et al.: Recovering human body configurations: Combing segmentation and recognition. In: Proc. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 326–333 (2004)

[38] Mori, G., Malki, J.: Recovering 3D Human Body Configurations Using Shape Contexts. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(7), 1052–1062 (2006)

[39] MuHAVi: Multicamera Human Action Video Data, `http://dipersec.king.ac.uk/MuHAVi-MAS` (last accessed January 2010)

[40] Okada, R., Soatto, S.: Relevant Feature Selection for Human Pose Estimation and Localization in Cluttered Images. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 434–445. Springer, Heidelberg (2008)

[41] Pollefeys, M., Van Gool, L.: Stratified Self-Calibration with the Modulus Constraint. IEEE Trans. on Pattern Analysis and Machine Intelligence 21(8), 707–724 (1999)

[42] Poppe, R., Poel, M.: Comparison of silhouette shape descriptors for example-based human pose recovery. In: Proc. Int. Conf. on Automatic Face and Gesture Recognition, pp. 541–546 (2006)

[43] Poppe, R.: Evaluating Example-based Pose Estimation: Experiments on the HumanEva sets. In: Proc. Workshop on Evaluation of Articulated Human Motion and Pose Estimation (2007)

[44] Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding 108(1-2), 4–18 (2007)

[45] Ramanan, D., Forsyth, D.: Finding and traking people from the bottom up. In: Proc. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 467–474 (2003)

[46] Ramanan, D., Forsyth, D., et al.: Strike a pose: Tracking people by finding stylized poses. In: Proc. Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 271–278 (2005)

[47] Ramanan, D.: Learning to parse images of articulated bodies. In: Proc. Advanced in Neural Information Processing Systems, pp. 1129–1136 (2007)

[48] Rasmussen, C., Williams, G.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)

[49] Remondino, F., Roditakis, A.: 3D Reconstruction of Human Skeleton from Single Images or Monocular Video Sequences. In: Noltemeier, H. (ed.) WG 1980. LNCS, vol. 100, pp. 100–107. Springer, Heidelberg (1981)

[50] Ren, X., Berg, A., et al.: Recovering human body configurations using pairwise constraints. In: Proc. Int. Conf. on Computer Vision, pp. 824–831 (2005)

[51] Renno, J., Remagnino, P., et al.: Learning Surveillance Tracking Models from the Self-Calibrated Ground Plane. Acta Automatica Sinica 29(3), 381–392 (2003)

[52] Saul, L., Roweis, S.: Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326 (2000)

[53] Sidenbladh, H.: Image sequences provided,
`http://www.csc.kth.se/~hedvig/data.html`
(last accessed January 2010)

[54] Sigal, L., Black, M.: HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion, Technical Report CS-06-08, Brown University (2006)

[55] Sigal, L., Black, M.J.: Predicting 3D People from 2D Pictures. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2006. LNCS, vol. 4069, pp. 185–195. Springer, Heidelberg (2006)

[56] Sigal, L., Black, M.: Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In: Proc. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 2041–2048 (2006)

[57] Spencer, N., Carter, J.: Towards pose invariant gait reconstruction. In: Proc. Int. Conf. on Image Processing, vol. 2, pp. 261–264 (2005)

[58] Srinivasan, P., Shi, J.: Bottom-up recognition and parsing of the human body. In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 1–8 (2007)

[59] Taylor, C.: Reconstruction of Articulated Objects from Point Correspondences in a Single Image. In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 677–684 (2000)

[60] Tenenbaum, J.: Global Geometric Framework for Nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)

[61] Tsai, R.: A versatile camera calibration technique for high accuracy 3D machine vision metrology using off-the-shelf TV cameras and lensesp. IEEE J. Robotics and Automation RA 3, 323–343 (1987)

[62] Urtasun, R., Darrell, T.: Sparse Probabilistic Regression for Activity-independent Human Pose Inference. In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 1–8 (2008)

[63] Motion Capture Systems from Vicon, `http://www.vicon.com` (last accessed January 2010)

[64] Vondrak, M., Sigal, L., Jenkins, O.: Physical Simulation for Probabilistic Motion Tracking. In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 1–8 (2008)

[65] Wang, Y., Mori, G.: Boosted multiple deformable trees for parsing human poses. In: Proc. Workshop on Human Motion Understanding, Modeling, Capture and Animation, pp. 16–27 (2007)

[66] Yang, H., Lee, S.: Reconstructing 3D human body pose from stereo image sequences using hierarchical human body model learning. In: Proc. Int. Conf. on Pattern Recognition, vol. 3, pp. 1004–1007 (2006)

[67] Zhao, T., Nevatia, R.: Bayesian human segmentation in crowded situations. In: Proc. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 459–466 (2003)

A Comprehensive Study of Sports Video Analysis

Ming-Chun Tien[1], Ja-Ling Wu[1], and Wei-Ta Chu[2]

[1] Graduate Institute of Networking and Multimedia
National Taiwan University
No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan, 10617
{trimy,wjl}@cmlab.csie.ntu.edu.tw

[2] Department of Computer Science and Information Engineering
National Chung Cheng University
168 University Road, Minhsiung Township, Chiayi County, Taiwan, 62102
wtchu@cmlab.csie.ntu.edu.tw

Abstract. Commercial applications of video analysis are getting valuable with the development of digital television. People can easily record all kinds of programs and enjoy the videos in their leisure time. Among these programs, broadcasting sports videos are usually more tedious than others since they involve not only the main games, but also break time or commercials. And even main games comprise periods which are not splendid enough for the audience. Therefore, a considerable amount of research focuses on automatically annotating semantic concepts in sports videos, and providing a spellbinding way to browse videos. In this chapter, we briefly introduce related work of video analysis for different kinds of sports, and propose a generic framework for sports video annotation. We explicitly elaborate the state of the art techniques for sports videos analysis. Visual and audio information are utilized to extract mid-level features, and different models for semantic annotation are expounded with practical examples. We also expand on applications of sports video analysis from the viewpoints of the audience, professional athletes, and advertisers.

1 Related Work of Video Analysis for Different Sports

The advancement of digital video coding and transmission has caused a sharp rise in video amount. Consequently, automatic video semantic analysis becomes substantial for efficiently indexing, retrieval, and browsing of the data in digital libraries [21, 22]. Techniques of video segmentation [36], shot classification [18], and event detection [41] have been proposed to facilitate semantic understanding. Among all types of videos, sports videos involve much more regularities than others owing to two main reasons: 1) Sports games are held on specific playground with clear and definite rules. 2) A cameraman usually

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 413–441.
springerlink.com

adopts regular manners to capture meaningful events during the game for the audience. These regularities make it possible to extract semantics from the videos. However, there are too many kinds of sports in the world and to find a generic framework for analyzing all of them is nearly impossible. In this chapter, we focus on video analysis of mainstream sports such as soccer, tennis, basketball, football, billiards, baseball, etc.

There has been a proliferation of research on sports video analysis in the past ten years. Most of them focused on highlight extraction, structure analysis and semantic event annotation. *Gong et al.* [23] utilized object color and texture features to generate highlights in broadcast soccer videos. *Xu et al.* [51] and *Xie et al.* [49] detect plays/breaks in soccer games by frame view types and by motion/color features, respectively. *Li et al.* [31] summarized football video by play/break and slow-motion replay detection using both cinematic and object descriptors. *Rui et al.* [40] detected highlights using audio features alone without relying on expensively computing video features. Besides visual/audio features extracted from the video, *Babaguchi et al.* [4] combined text information from closed captions (CC) to seek for time spans in which events are likely to take place. With the aid of web-casting text information, *Xu et al.* [50] tried to annotate sports videos with semantic labels which not only cover general events, e.g. scoring/fouls in basketball, but also the semantics of events, e.g. names of players. Moreover, some works analyzed the superimposed caption to more accurately annotate the videos [14, 56].

Object trajectories also provide rich information for semantic understanding. *Assfalg et al.* [3] employed camera motion and locations of the players to detect events in soccer videos. *Tovinkere et al.* [45] utilized object trajectories to achieve semantic event detection in soccer video with a set of heuristic rules which are derived from a hierarchical entity-relationship model. *Intille et al.* [29] analyzed interactions in football videos based on object trajectories, which could be clues for play classification. Not surprisingly, [3, 45, 29] assumed that trajectory information is obtained in advance. To obtain the object trajectories automatically, *Pingali et al.* [38] proposed a real-time tracking system for tennis videos captured by a stationary camera. In [38], player trajectories are obtained by dynamically clustering tracks of local features, and ball segmentation/tracking is realized based on shape and color features of the ball. *Guéziec* [24] exploited the kinematic properties of the baseball's flight to track the ball during pitches in real-time. However, extensive prior knowledge such as camera locations and coverage have to be known for tracking the ball.

In this chapter, we present a generic framework for sports video analysis as illustrated in Fig. 1. Techniques for video/audio signals processing are dilated in Sect. 2 and Sect. 3, respectively. The extracted video/audio mid-level features are then utilized for semantic annotation based on three methodologies as expatiated in Sect. 4. Moreover, Sect. 5 introduces some applications of sports video analysis from different viewpoints and Sect. 6 concludes this chapter.

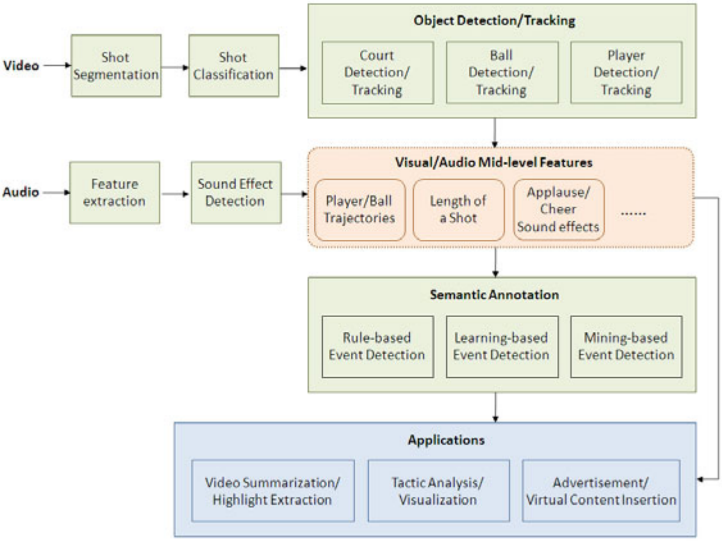

Video
Shot Segmentation
Shot Classification
Object Detection/Tracking
Court Detection/ Tracking
Ball Detection/ Tracking
Player Detection/ Tracking
Audio
Feature extraction
Sound Effect Detection
Visual/Audio Mid-level Features
Player/Ball Trajectories
Length of a Shot
Applause/ Cheer Sound effects
......
Semantic Annotation
Rule-based Event Detection
Learning-based Event Detection
Mining-based Event Detection
Applications
Video Summarization/ Highlight Extraction
Tactic Analysis/ Visualization
Advertisement/ Virtual Content Insertion

Fig. 1. A generic framework of sports video analysis. The video is first segmented into shots and a shot classification scheme is applied to each shot. Mid-level video/audio features are then extracted for semantic annotation. The mid-level features and the annotation results can be further employed to some applications of sports video analysis.

2 Analysis of Visual Information

The mission of a cameraman is to convey meaningful events to the audience. Since most athletic events take place in an area of specific colors and dimension, sports programs usually focus on the area during the game. Moreover, to perfectly capture the scene without loss of event concepts, a cameraman will control the camera in particular manners. Hence, sports video analysis significantly relies on visual information. Generic frameworks utilize low-level visual features (e.g. color, texture, and motion) to achieve video preprocessing steps including shot segmentation, shot classification, object detection, and object tracking. According to the results of video preprocessing, representative mid-level visual features are extracted to describe each video segment. These mid-level visual features play important roles in further analysis, and can be employed to many applications.

2.1 Shot Segmentation

A video stream can be segmented into several video shots according to production cues such as scene changes and shot boundaries. In sports videos, abrupt transitions (cuts) occur when the camera view changes from one scene to another, and gradual transitions (e.g. dissolve, fade-in/out, and wipe) occur in the beginning and the end of a reply. Hence, detecting abrupt/gradual transitions is an essential step of indexing, browsing, and searching sports videos.

Shot segmentation for uncompressed video

There is an extensive literature on shot boundary detection (SBD) algorithms [27, 33]. The SBD problem can be solved in the uncompressed domain with the aid of color/edge information. The basic idea of color-based SBD is that the color content remains almost consistent in the same shot while changes rapidly across shots. Hence, we can detect cut transitions as peaks in the time series of the differences between color histograms of contiguous frames. The color histograms difference (CHD) can be defined as

$$CHD_i = \frac{1}{N}\sum_{r=0}^{k}\sum_{g=0}^{k}\sum_{b=0}^{k}|p_i(r,g,b) - p_{i-1}(r,g,b)|, \tag{1}$$

where each color component is quantized to k different values, and $p_i(r,g,b)$ denotes the number of pixels of color (r,g,b) in frame i of N pixels. A cut transition is detected if a local maximum CHD_i exceeds a threshold Θ_{CHD}. Since scene change is a local activity in the temporal domain, the threshold Θ_{CHD} should be determined adaptively to reduce false detection caused by large object and camera motion [53].

The color-based SBD can be extended to a twin-comparison algorithm [57] to detect gradual transitions. The twin-comparison algorithm takes a two-phase detection strategy. As illustrated in Fig. 2, CHD values of all frames are calculated in the 1^{st} phase and a high threshold Th_h is set to detect cut transition. If the CHD_i is less than Th_h but higher than an additional low threshold Th_l, frame i is considered as a possible start frame (F_s) of a gradual transition. During a normal gradual transition, color histogram difference between the start frame and each of its consecutive frames (usually called the accumulated color histogram difference, ACHD) will increase gradually. Hence, in the 2^{nd} phase, we calculate ACHD values to determine the real start/end frame of a gradual transition. For each possible start frame F_s, frame i is detected as the end frame if the corresponding ACHD value has increased to a value larger than Th_h, and the value of CHD_{i+1} is less than Th_l. If the corresponding end frame is not found for a possible start frame, we just drop this start point and search for another gradual transition.

During a cut or dissolve, new intensity edges (entering edges) appear far from the location of old edges and old edges (exiting edges) disappear far from the location of new edges. Hence, the edge-based SBD method [54] employs the edge change ratio (ECR) to detect transitions. The ECR between frame $i-1$ and frame i can be defined as

$$ECR_i = \max(\frac{E_i^{in}}{E_i}, \frac{E_{i-1}^{out}}{E_{i-1}}), \tag{2}$$

where E_i denotes the number of edge pixels in frame i, E_i^{in} and E_{i-1}^{out} are the numbers of entering and exiting edge pixels in frame i and frame $i-1$, respectively. Hard cuts, fades, dissolves, and wipes exhibit characteristic patterns of ECR sequences. Thus we can detect shot boundaries and even classify different types of transitions by counting and analyzing the spatial distribution of the entering/exiting edge pixels.

Shot segmentation for compressed video

Considering the time efficiency of SBD, algorithms developed for compressed videos have been proposed in recent years since MPEG videos contain rich set of pre-computed features, e.g. DC coefficients and motion vectors. Methods using DC coefficients require significant decoding of the MPEG-compressed video and do not work well on both hard cuts and gradual transitions simultaneously. With the same idea of using compressed video data, *Haoran et al.* [28] proposed an SBD method based on the dissimilarity between I/P/B frames with respect to various types of macroblocks used for coding.

There are four types of macroblocks (MBs) in an MPEG coding scheme: intracoded (In), forward predicted (Fw), backward predicted (Bw), and bidirectionally predicted (Bi). The number of each type MB in a frame is relative

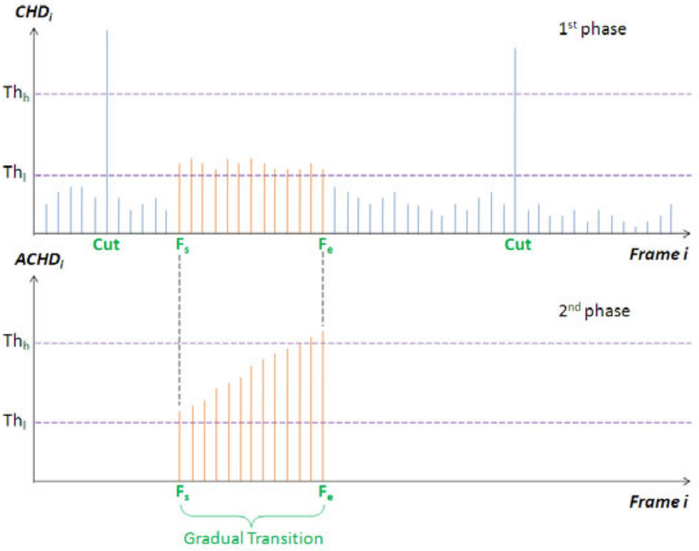


Fig. 2. The color-based twin-comparison algorithm. In the 1^{st} phase, a high threshold is set to detect cut transition, and a low threshold is set to obtain frame candidates of a gradual transition (F_{gt}). In the 2nd phase, accumulated color histogram difference is calculated to determine the start/end of the gradual transition from all $F_{gt}{}'s$.

to the dissimilarity between that frame and its neighboring frames. In [28], *Haoran et al.* defined the frame dissimilarity ratio (FDR) as

$$FDR_i = \begin{cases} \frac{Fw_{i-1}}{Bi_{i-1}} & \text{, for I/P frame.} \\ \max(\frac{Fw_i}{Bi_i}, \frac{Bw_i}{Bi_i}) & \text{, for B-frame.} \end{cases} \tag{3}$$

If a shot change occurs at a reference frame, most MBs in the previous B-frame are predicted from the previous reference frame. Therefore, Fw in the previous frame will be high and resulting in high FDR. Similarly, if a shot change occurs at a B-frame, all frames between the previous and the following reference frame will contain much more Fw/Bw MBs than Bi MBs, which also results in high FDRs. However, the latter case results in successive high FDRs rather than an exact FDR peak. To determine the exact shot boundary, a modified frame dissimilarity ratio (MFDR) is defined as

$$MFDR_i = FDR_i \times DMBC_i, \tag{4}$$

where $DMBC_i$ is the dominant MB change for frame i, which is defined by

$$DMBC_i = \begin{cases} 1 \text{ , for I/P frame.} \\ 0 \text{ , for B frame and } (Bw_i - Fw_i)(Bw_{i-1} - Fw_{i-1}) > 0. \\ 1 \text{ , for B frame and } (Bw_i - Fw_i)(Bw_{i-1} - Fw_{i-1}) \leq 0. \end{cases} \tag{5}$$

Take the frame structure : $I_1B_2B_3B_4P_5B_6B_7B_8P_9$ as an example, if a shot change takes place at B_3, FDRs of B_2, B_3 and B_4 will be high. On the other hand, since B_2 contains more forward predicted MBs while B_3 and B_4 contain more backward predicted MBs, $DMBC_1$, $DMBC_2$, $DMBC_3$ and $DMBC_4$ will be 1, 1, 1 and 0, respectively. The corresponding MFDRs are calculated and an exact shot boundary can be obtained by finding the first frame of successively high MFDR frames.

Haoran et al. also applied DMBC defined in Eq.(5) to detect gradual transitions and proposed a method to further divides the shots into subshots (e.g. pan, tilt, and zoom shots) by motion vector information in an MPEG stream. Hence, it is quite appropriate for analyzing videos involving complex shots and shooting manners, such as basketball and football videos.

2.2 Shot Type Classification

Shots can be categorized into a set of scene types for each kind of sports video based on cinematic features and object-based features [17], for example, *Liu et al.* [35] classified basketball video shots into six scene types including fast motion court-view, slow motion court-view, penalty, in-court medium, and bird-view. Considering time efficiency, we can omit extracting object-based features and simply classify shots into play or break by cinematic features [16]. However, for applications requiring more detailed semantics, object-based features are indispensable. To develop a generic framework for all kinds of

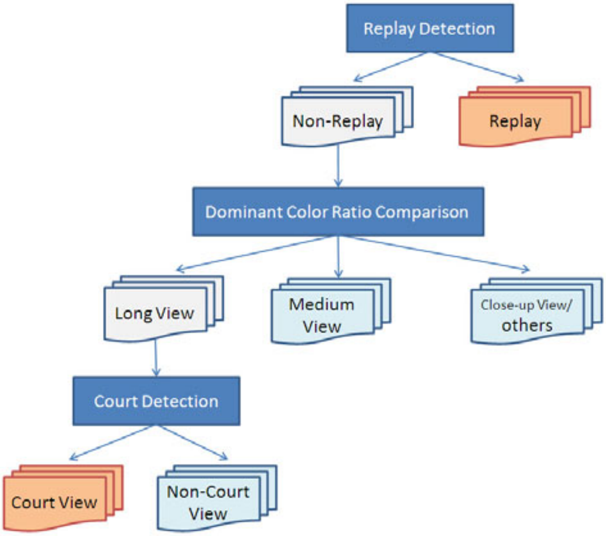


Fig. 3. The shot type classification scheme. Shots are classified into five categories by replay detection, dominant color ratio comparison and court detection.

sports videos, we introduce a hierarchical scheme for shot type classification. As shown in Fig. 3, replay detection, dominant color ratio comparison, and court detection are applied to each level, and shots are ultimately classified into five categories including replay, court view, non-court view (in long view shots), medium view, and close-up view/others. Among these five shot types, replays are produced after the occurrence of excellent play skills or exiting events, while other important game information is embedded in court view shots; therefore, obtaining replays and court view shots are quite useful for sports analysis. In this section, we will elaborate the techniques of replay detection and dominant color ratio comparison, while leave the court detection method to be explained in Sect. 2.3.

Replay Detection

In sports videos, a replay may contain slow motion or non-slow motion or both. Some works have been proposed to detect slow motion replays via analyzing motion model of the video sequences [37, 47]. However, these methods can not be applied to detect non-slow motion replays, and will fail when slow motion replays are generated by high speed camera. Since a replay is usually sandwiched in between two identical editing effects (i.e. logo transitions) to point out the beginning and the end of a replay, the replay detection problem can be converted to the logo detection one [6, 44].

Dominant Color Ratio Comparison

In most sports, the playing field is characterized by specific colors. Since the field often dominate large portion of the frame during a play, detecting dominant (field) color in the videos is useful for further analysis. The statistics of the dominant color are learnt in the HSI (hue-saturation-intensity) color space which depicts the characteristics of human visual perception better than other color spaces. Some sports fields may comprise more than one dominant color; hence, we accumulate the HSI histograms of the first K frames and model the histogram by a Gaussian Mixture Model (GMM) which consists of M Gaussian densities:

$$p(\xi|\lambda) = \textstyle\sum_{i=1}^{M} w_i b_i(\xi), \qquad w_1 + w_2 + \cdots + w_M = 1. \tag{6}$$

ξ is the color vector of a pixel and w_i is the weight of the i-th mixture component b_i. The parameters (w_i and b_i) are estimated by the expectation maximization (EM) algorithm and the dominant colors are determined by Algorithm 1. Moreover, to develop a robust algorithm adapting to light variations in the temporal domain, automatically updating the statistics of dominant colors is essential. The dominant colors are dynamically adjusted according to the newly decoded video frames.

Algorithm 1 (Dominant Colors Detection)

```
Given a set of mixture components $b_i's$ and their corresponding weights $w_i's$,
determine the dominant color set $\Phi$.
Sort the $M$ mixture components in the descending order according to
  their weights and push the corresponding mean color $\xi^i$ of each ordered
  component into a queue **Q:**$(\xi^1, \xi^2, \cdots \xi^M)$.
Set the dominant color set $\Phi$ as an empty set.
for $i = 1 : M$ do
    Compute the neighboring color set $\Psi$ of $\xi^i$.
    Add $\Psi$ into the set $\Phi$.
    if $p(\Phi|\lambda) > Th_{dominant}$
        break
    end if
end for
return $\Phi$
```

In Algorithm 1, the color distance between ξ^i and each element in the neigh-boring color set Ψ should be less than a threshold. Both the chromaticity and the achromaticity channels are taken into consideration to measure the distance between two colors by *robust cylindrical metric* [39]. As illustrated in Fig. 4, the *chroma distance* (D_{chroma}) of two colors C_1 and C_2 can be measured by

$$D_{chroma}(C_1, C_2) = \sqrt{(S_1)^2 + (S_2)^2 - 2S_1S_2\cos\theta}, \tag{7}$$

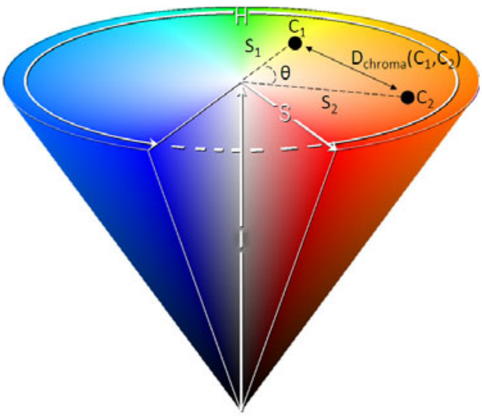


Fig. 4. HSI color cone. The chroma distance of two colors C_1 and C_2 can be measured by their saturation values and the angle between their hues.

where S_1 and S_2 are the corresponding saturation values of C_1 and C_2, and θ is the angle between hue values of the two colors. While the *intensity distance* ($D_{intensity}$) is measured by

$$D_{intensity}(C_1, C_2) = |I_1 - I_2|. \tag{8}$$

The color distance of two colors C_1 and C_2 is then determined by

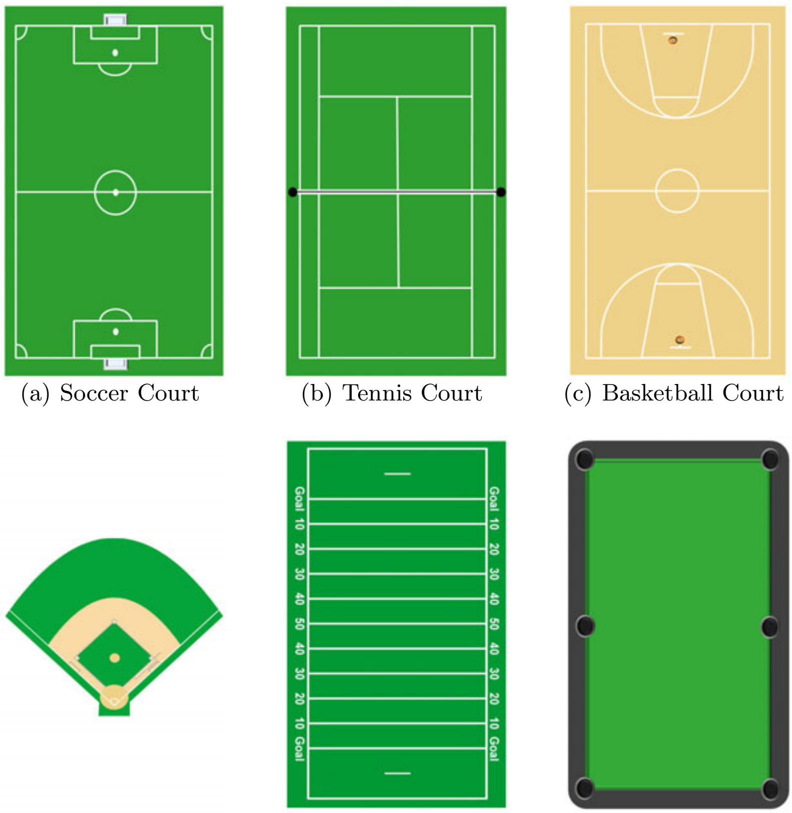
$$D_{cylindrical}(C_1, C_2) = \sqrt{D_{intensity}(C_1, C_2)^2 + D_{chroma}(C_1, C_2)^2}. \tag{9}$$

However, if the saturation and the intensity of the color lie in the achromatic region, only intensity distance is considered to measure the color distance in Eq.(9).

A shot type can be determined by a key frame or by a set of frames. For example, given a video shot, we can compare the ratio of dominant color pixels in each frame with two thresholds, say Th_{long} and Th_{med}, and label each frame with long view, medium view, or close-up view/others. Then, this shot is assigned to the major label of all the frames.

2.3 Object Detection and Tracking

For sports videos, detecting specific objects and even obtaining their trajectories makes it easier to understand the semantics of the videos. For example, a long time appearance of the court usually indicates play-time rather than break-time, and the ball/player trajectories can reveal occurrence of certain events. Here we focus on three object types: the court, the ball(s), and the player(s).


Goal 10 20 30 40 50 40 30 20 10 Goal
Goal 10 20 30 40 50 40 30 20 10 Goal

(a) Soccer Court (b) Tennis Court (c) Basketball Court

(d) Baseball Court (e) Football Court (f) Billiards Table

Fig. 5. Court models of different sports

Court Detection and Tracking

According to the rules of the competition, matches are played on natural or artificial surfaces with specification. Fig. 5 illustrates the court models of different sports. The court lines are usually painted with certain color (e.g. white in soccer/tennis/basketball/baseball court) or set with clear edge (e.g. the billiards table). Hence, we can detect the court with the aid of line or edge information. *Farin et al.* [19, 20] proposed a line-based camera calibration technique which not only detected the court in the video but also determined the calibration parameters. In [19, 20], possible court line pixels are first detected according to color information with constraint to exclude large white areas or fine textured areas. Based on the detected line pixels, a *RANSAC*-like algorithm [19] is applied to extract the dominant line in each frame, and the line segment boundaries are determined by least square approximation.

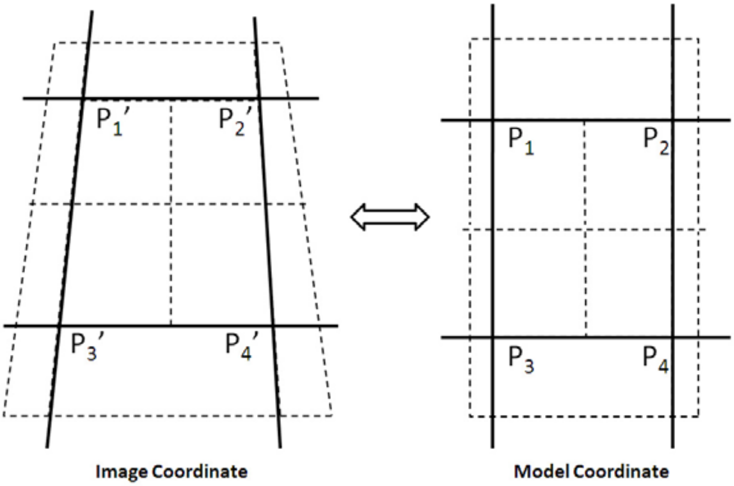


Fig. 6. Model fitting. We can use four intersection points to compute the geometric transformation between two planes, and the four points are extracted by iteratively configuring two horizontal and two vertical lines between the image and the court model.

To identify which line in the image corresponds to which line in the court model, a *model fitting* step is utilized to infer the geometric transformation between the two planes. The geometric transformation can be written as a 3×3 homography matrix **H** which maps a point $p = (x, y, w)^T$ in the court model coordinates to a point $p' = (x', y', w')^T$ in the image coordinates. As illustrated in Fig. 6, we can use four intersection points to compute the geometric transformation, and the four points are extracted by iteratively configuring two horizontal and two vertical lines between the image and the court model. Since the court position will not change too much in two successive frames of a shot, we can efficiently estimate and track the court positions in following frames when the initial court position has been located [20]. This camera calibration method can be applied to most mainstream sports videos with little variation. For example, the possible court line pixels in billiards videos are extracted by edge detection instead of the above-mentioned method [13].

Ball Detection and Tracking

Ball detection/tracking is challenging since the ball is a small object relative to the frame and it is often blurred because of its fast move. A general methodology for ball detection/tracking first detects all possible ball candidates according to color, motion, shape, and size information (Fig. 7 is an example of the ball candidate detection procedure), and then applies a tracking algorithm to identify the real ball positions. *Guéziec* [24] proposed a

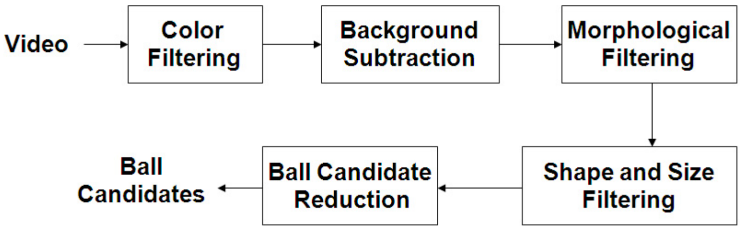


Fig. 7. A sample process of ball candidate detection. In addition to color, shape and size in-formation, background subtraction, morphological filtering and a designed reduction mechanism are utilized to detect ball candidates.

method to extract the trajectory of a baseball from video clips using *Kalman filter* [48] which incrementally tracks the ball in the scene based on prediction and pattern matching, while *Shum et al.* [42] performed a global search methods based on dynamic programming to find the trajectory of the ball.

Kalman filter is a well known algorithm commonly used for object tracking, removing measurement errors and estimating a system's variables. Linear equations must describe the system and measurement evolutions over time. *Kalman filter* provides optimal estimates of the system parameters, such as position and velocity, given measurements and knowledge of a system's behavior. In general, *Kalman filter* assumes that the following two relations can describe a system:

$$x_k = A_k x_{k-1} + w_k, \tag{10}$$

$$z_k = H_k x_k + v_k, \tag{11}$$

where x_k is the state vector, such as a position, velocity, acceleration or other parameters, while z_k is the measurement, such as a position. w_k (process noise, or process evolution), and v_k (measurement noise) are mutually uncorrelated white noise vectors. Eq.(10) determines the evolution of states over time, and Eq.(11) relates measurement and state. Once the system is established, a recursive algorithm is utilized to estimates x_k optimally.

Possible ball positions in different frames can be found utilizing the procedures illustrated in Fig. 7 However, for some sports videos like basketball videos, ball detection is much more difficult due to the complicated scene which results in plenty of ball candidates or ball occluded by players. In this case, ball tracking based on dynamic programming is much more suitable to find out the correct ball trajectory. Given two frames f_i and f_j $(i < j)$, the 2D velocity of the ball can be calculated by

$$Velocity_{i \to j} = \frac{\sqrt{(X_j - X_i)^2 + (Y_j - Y_i)^2}}{T_{i \to j}} \tag{12}$$

where (X_i, Y_i) and (X_j, Y_j) are respectively the positions of the ball candidates in f_i and f_j, and $T_{i \to j}$ is the time duration between f_i and f_j. For two

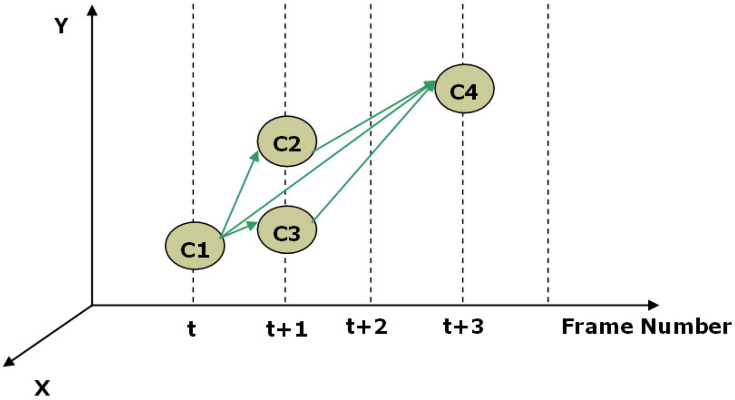


Fig. 8. Tracking process based on dynamic programming. The X and Y axes represent 2D coordinates of the ball, and the horizontal axis shows the frame number of the current candidate. Assume the ball candidates are represented as nodes, when the velocity of the ball calculated from candidates in f_i and f_j satisfying the velocity constraint, the nodes corresponding to these candidates will be connected by an edge. After connecting the candidates by edges, a complete route that represents the trajectory of the ball is searched recursively based on dynamic programming to maximize a predefined criterion.

nearby frames in a shot, the velocity of the ball will be within a certain range (velocity constraint). The tracking conception is described in Fig. 8, in which the X and Y axes represent 2D coordinates of the ball, and the horizontal axis shows the frame number of the current candidate. Assume the ball candidates are represented as nodes in Fig. 8, when the velocity of the ball calculated from candidates in f_i and f_j satisfying the velocity constraint, the nodes corresponding to these candidates will be connected by an edge. After connecting the candidates by edges, a complete route that represents the trajectory of the ball is searched recursively based on dynamic programming to maximize a predefined criterion [42].

Player Detection and Tracking

Information of players' positions is important for semantic analysis. How players move along the time axis further conveys significant cues for event detection and tactic analysis. The essential idea of player detection is to find the region with non-dominant-color pixels, which is surrounded by dominant-color area. Given a video frame, we can construct a binary map

$$B(x,y) = \begin{cases} 0, & \text{if } p(x,y) \text{ is in dominant color,} \\ 0, & \text{if } (x,y) \text{ is a court line pixel,} \\ 1, & \text{otherwise,} \end{cases} \tag{13}$$

where $B(x, y)$ represents the value of the binary map at a given point (x, y), and $p(x, y)$ represents the color values of this pixel. A region growing procedure is then used for player segmentation. If there are only few players appearing in the video and the players will not seriously occlude with each other, we can easily track players on the basis of the velocity constraint mentioned above. However, for basketball videos or football videos where a player often move very close to other players, only using velocity constraint usually incurs bad tracking results. Hence we can derive player trajectory based on another tracking technique such as *Continuously Adaptive Mean Shift Algorithm* (*CamShift*) [2], an adaptation of the *Mean Shift algorithm* for object tracking. The *CamShift* Algorithm is summarized in Algorithm 2.

Algorithm 2 (CamShift Algorithm)

```
Step1. Set the region of interest (ROI) of the probability distribution image
       to the entire image.
Step2. Select an initial location of the Mean Shift search window. The selected
       location is the target distribution to be tracked.
Step3. Calculate a color probability distribution of the region centred at the
       Mean Shift search window.
Step4. Iterate Mean Shift algorithm to find the centroid of the probability
       image. Store the $zero^{th}$ moment (distribution area) and the centroid
       location.
Step5. For the following frame, center the search window at the mean location
       found in Step 4 and set the window size to a function of the $zero^{th}$
       moment. Go to Step 3.
```

3 Analysis of Audio Information

The audio stream have strong hint for event detection in sports video. For example, the whistling occurs right after a foul play, and exited commentator speech usually follows a goal event. In this section, we will introduce techniques of detecting sound effects (whistling, applause, etc.) which are meaningful for the occurrence of sports events. Sound effects recognition has been widely studied in recent years [8, 12]. A general procedure of sound effect detection is depicted in Fig. 9. Audio signals exhibit strong context such that variables can be predicted from previous values. Hence, low-level audio features are extracted and Hidden Markov Model (HMM), a well known statistical model used for temporal pattern recognition, is applied to detect the specific sound effect.

3.1 Low-Level Audio Feature Extraction

An audio signal is first segmented into basic units (say frames of k milliseconds) for feature extraction. For each audio unit, several audio features are

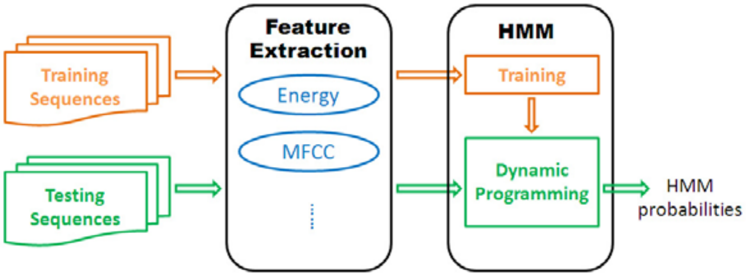


Fig. 9. The general procedure of audio analysis. Low-level audio features are extracted and Hidden Markov Model (HMM) is applied to detect the specific sound effect.

extracted for modeling, including energy, band energy ratio, zero-crossing rate, frequency centroid, bandwidth, mel-frequency cepstral coefficient (MFCC), and delta/acceleration which have been shown to be beneficial to sound effects recognition [8, 12, 35, 55]. We briefly introduce some of the above mentioned features as follows.

- **Energy.** Energy is defined by the logarithm (log) of the square sum about the amplitude of all audio samples within a basic unit, that is

$$Energy = \log \sum_{i=1}^{N} {S_i}^2. \tag{14}$$

- **Mel-frequency cepstral coefficient.** Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound. The frequency bands in MFC are equally spaced on the mel scale which closely approximates the human auditory system response. Mel scale can be calculated by Eq.(15), where f is the normal frequency scale.

$$Mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \tag{15}$$

  Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. As depicted in Eq.(16), MFCC are commonly computed from FFT power coefficients filtered by a triangular bandpass filter bank, where S_k is the output of the k-th filter bank and N is the number of samples in a basic unit.

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^{K} (\log S_k) \cos[n(k - 0.5)\frac{\pi}{k}], \quad n = 1, 2, \cdots N \tag{16}$$

- **Delta/Acceleration.** Delta (δ_n) and acceleration (ACC_n) are the first and the second order characteristics of MFCC which can be computed by Eq.(17) and Eq.(18), respectively.

$$\delta_n = C_n - C_{n-1} \tag{17}$$

$$ACC_n = \delta_n - \delta_{n-1} \tag{18}$$

3.2 Sound Effect Detection by HMM

Training data sets of k types of sound effects (e.g. ordinary sounds, commentator speech, excited commentator speech, applause/cheer, and racket hit sound) are collected to model each sound effect using HMM. Parameters of each model λ_k are adjusted based on the Baum-Welch algorithm [7]. After modeling, how likely an audio sequence belongs to an audio effect is evaluated. For each audio sequence $\mathbf{A} = \{x_1, x_2, \cdots x_n\}$ containing n audio basic units, the corresponding audio low-level features of each basic unit form an observation vector $\boldsymbol{O}$. The likelihood of each HMM is computed and the audio sequence $\boldsymbol{A}$ is recognized as sound effect i if $p(O|\lambda_i) = \max_k p(O|\lambda_k)$.

4 Semantic Annotation of Sports Video

With the proliferation of multimedia content, automatic annotation becomes important for users to retrieve videos. Humans generally use high-level semantic concepts to query videos, while computers only recognize low-level features such as color, motion, and texture. This semantic gap brings about challenges of video annotation. Generally speaking, sports video annotation can be done in the following three perspectives: structure-based annotation, event-based annotation, and ontology-based annotation. Since sports games follow specific structure rules, we can annotate sports videos with structure labels, e.g. a play, a break, a play in the second period of a basketball game. To a higher level of semantics, event-based annotation matches the user's intention more closely. Conventional works utilize visual/audio mid-level features to annotate events in sports videos. Ontology-based annotation approaches construct multimedia ontologies by manually assigning external knowledge to video content, hence it lacks of an automatic mechanism. In this section, we focus on event-based annotations and introduce three approaches including rule-based, learning-based, and mining-based event annotation.

In Sect. 2 and Sect. 3, we have respectively introduced how to derive mid-level visual and audio features from sports videos. From the classification point of view, mid-level features can be treated as the result of dimension reduction given video data with low-level features. These mid-level features more closely describe the way that human comprehend the video content than low-level features. An intuitive method for event detection is to use a rule-decision tree which combines both visual and audio mid-level features. With the aid of machine learning techniques, we can determine how each feature dominates each kind of event through training stage rather than using a pre-known decision tree. We can even detect frequent events of sports videos in the perspective of data mining. To convey the main ideas of these three approaches more explicitly, we take tennis video as an example, and one can apply these approaches to other sports videos with a little modification.

4.1 Events Definition and Mid-level Features Extraction

We first define events for each kind of sports videos, for example, five well-known events for tennis video are defined as follows.

- **Ace or unreturned serve.** A player successfully serves, and his/her opponent fails to return the ball. In the case of ace, the opponent is not able to touch the ball and therefore fails to return. In the case of unreturned serve, the opponent barely touches the ball but the returned ball touches net or is out-of-court.
- **Fault.** A player fails in his/her first serve, and the camera immediately switches out of the court view.
- **Double fault.** A player consecutively fails in two serves. In double fault, the camera doesn't switch out of the court view after the first failed serve, and the player successively fails the second serve.
- **Baseline rally.** A player successfully serves and his/her opponent successfully returns. They then strike around the baseline until one of them fails to return.
- **Net approach.** A player successfully serves and his/her opponent successfully returns. One or both of them once approach the net to stress his/her opponent.

To faithfully present the game and amuse the audience at the mean time, the producer usually uses court view to capture important events in playtime, while switches to other view type (e.g. replay or close-up view) between two events or in break-time. Hence, each court view shot carries one events averagely, which helps us to achieve event-based annotation using the court view shot as the basic unit. A tennis video is first segmented by the shot boundary detection method introduced in Sect. 2.1, and court view shots are extracted using the hierarchical shot classification scheme proposed in Sect. 2.2. Moreover, we utilize several representative visual/audio mid-level features to describe a court view shot for tennis video:

- **The relative position between the player and the court (D_r).** With the visual object detection technique (cf. Sect. 2.3), we can locate each player and each court line in video frames, and then find the relative position between the player and the court. This information is quite useful since many events are related to or defined by the player's position. Taking the bottom part of the tennis court as an instance, we partition the court into two regions by the court lines as depicted in Fig. 10. For a court view shot, if a player ever moves to the region one, this shot is likely to involve a net approach event. The top part of the court is partitioned symmetrically to the bottom part, and a binary mid-level feature D_r denotes whether a player steps into the region one.

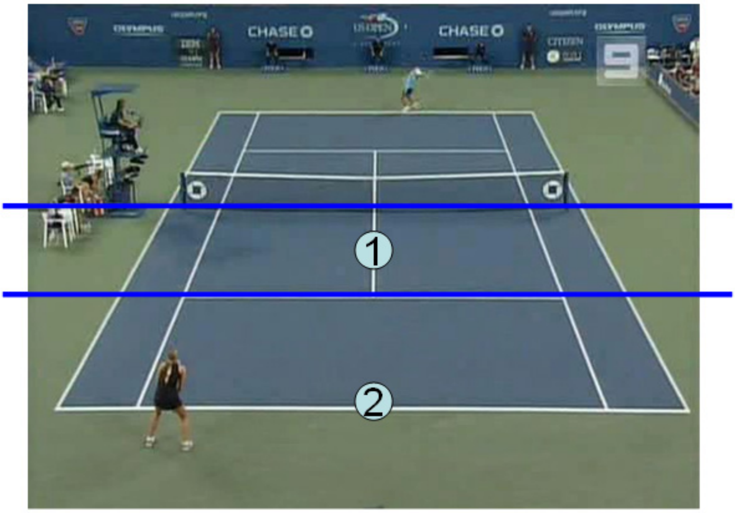


Fig. 10. The relative position between the player and the court. Taking the bottom part of the tennis court as an instance, we partition the court into two regions by the court lines.

$$D_r = \begin{cases} 1, & \text{if a player steps into the region one.} \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

- **Average moving distance of players (D_m).** Instead of using relative position, we can describe how the player exactly moves on the basis of statistics, e.g. the moving distance of players. For each court view shot, we accumulate the moving distance of each player (in the real world court coordinates), and then we can subtly examine whether this shot implies intense events or not. The mid-level feature D_m is calculated by averaging the moving distances of two players.
- **Time-length of the shot (D_t).** Different events will have different time-lengths, for example, the length a double fault is longer than that of a single fault. The length of a rally event is more likely longer than that of an ace. Therefore, we take the time-length, D_t, of a shot as a mid-level feature for event annotation.
- **Applause/cheer sounds effects (D_a).** Due to sports etiquette, the audiences only acclaim after some events. For example, in tennis matches, ace or unreturned serve always brings applause/cheer sounds while fault or double fault does not. Thus, the occurrence of applause or cheer sounds significantly provides clues for event annotation. With the applause/cheer detection technique (cf. Sect. 3), we can use a binary mid-level feature D_a to present whether this kind of audio effect occurs.

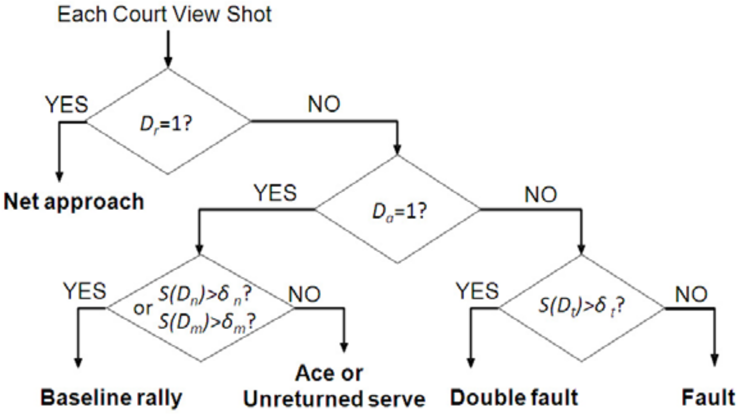


Fig. 11. The rule-based decision tree for tennis video event detection

$$D_a = \begin{cases} 1, & \text{if applause/cheer occurs.} \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

- **The number of racket hits in a shot (D_n).** For racket-sports, the number of hits is also an indication of event type, e.g. there are more racket hits in rallies than in aces or unreturned serves. Based on hit detection technique described in section Sect. 3, we use a mid-level feature D_n to denote the number of racket hits in a shot.

We can design different mid-level features which represent the real-world conditions for each kind of sports videos. These mid-level features are significantly different from low-level features used in conventional work and provide more meaningful information for event annotation.

4.2 Rule-Based Event Detection

According to the prior-knowledge of game regulations, broadcasting conventions, and inherent characteristics of events, we can design a rule-based decision tree for event detection based on mid-level features. Taking the tennis video as an example, we examine each court view shot with the detection process illustrated in Fig. 11. The tree first judges if any one of the players ever steps into the region one (the front of the net) according to D_r, and a net approach event is detected if $D_r = 1$. This decision is made by net approach's definition and intuition from observations. For court view shots where players never step to region one, the tree then checks whether applause/cheer sounds occur. If yes ($D_a = 1$), this shot should involve ace/unreturned serve or baseline rally; otherwise, double fault or fault event is carried in this shot.

Shots containing applause/cheer sounds are further classified into ace /unreturned serve and baseline rally according to the number of hit and the average moving distance of the players in a shot. Generally, there is only one racket hit in aces and at most two hits in unreturned serves. On the other hand, a baseline rally event means that players combat with each other for a longer time such that the average moving distance of players and the number of racket hits should be large. Similarly, shots without applause/cheer sounds can be further classified into double fault or fault according to the time-length of this play since a double fault takes longer time than a fault.

The decision boundaries δ_m , δ_n and δ_t are determined by the Bayesian decision theory [15]. For example, to decide δ_t, we first gather some double faults ($Event_1$) and faults ($Event_2$) respectively, and use a Gaussian distribution for each to model the time-length characteristics. A play with length δ_t belongs to a double fault if

$$\lambda_{21} p(Event_1|D_t) > \lambda_{12} p(Event_2|D_t), \tag{21}$$

where λ_{ij} is the cost incurred when $Event_j$ is wrongly classified to $Event_i$. By employing Bayes formula, we can replace the posterior probabilities in Eq.(21) by the product of the prior probabilities and conditional densities:

$$\lambda_{21} p(D_t|Event_1) p(Event_1) > \lambda_{12} p(D_t|Event_2) p(Event_2). \tag{22}$$

Rewrite Eq.(22) and decide the shot to be a double fault if

$$S(D_t) = \frac{p(D_t|Event_1)}{p(D_t|Event_2)} > \frac{\lambda_{12}}{\lambda_{21}} \frac{p(Event_2)}{p(Event_1)} = \delta_t. \tag{23}$$

and to be a fault otherwise. The conditional densities $p(D_t|Event_1)$ and $p(D_t|Event_2)$ are described by the Gaussian distributions mentioned above. The prior probabilities $p(Event_1)$ and $p(Event_2)$ are estimated according to the formal records reported by the Australian Open. Moreover, the costs λ_{12} and λ_{21} could be adjusted to show different preferences in detection.

4.3 Learning-Based Event Detection

The rule-based event detection sequentially makes "hard decision" by checking visual/audio mid-level features. However, some exceptions or court/player detection errors would incur erroneous event detection. For example, player detection/tracking is sometimes annoyed by the superimposed caption or the ball boys. Once the player's position is erroneously detected as in the region one, this play would be rudely detected as a net approach even if all other descriptors present significantly different opinions. Therefore, a learning-based method which jointly considers all visual/audio mid-level features can be applied to event detection.

There have been some works endeavoring to detect events based on learning methods. *Leonardi et al.* [30] extracted motion information and exploited HMM to detect "goal" event in soccer videos. With the aid of color-based features, *Bach et al.* [5] exploited multi-stream HMM to characterize baseball

Table 1. Data transformation: mapping between features and symbols

Features	Scale	Symbol
D_r	$D_r = 0$	m
	$D_r = 1$	n
D_m	*Short*	a
	Medium	b
	Long	c
D_t	*Short*	d
	Medium	e
	Long	f
D_a	$D_a = 0$	u
	$D_a = 1$	v
D_n	*Few*	x
	Moderate	y
	Plenty	z

events, such as homerun, fly out, and base hit. From the perspective of discriminative learning, some studies [46, 52] based on support vector machines (SVM) were proposed to construct classifiers for event detection. In the case of tennis event detection, we apply LIBSVM [9] to construct a multi-class classifier. All mid-level features are concatenated as a vector to describe a court view shot. In the training stage, all court view shots are manually labeled with the ground truth, and statistical characteristics of all events are then classified by the SVM classifier, which defines the decision boundaries between different classes. In the detection stage, each court view shot represented by a feature vector is evaluated with the trained classifier.

4.4 Mining-Based Event Detection

If we view the mid-level features as symbols to represent video sequences, we can observe that there are some frequent patterns in the symbol stream since some events frequently take place in sports videos. For example, there are often a large number of net approach events in a tennis match; therefore, the symbol which indicates players stepping into the front of the net will appear frequently. Hence, we can treat the event detection problem as a data mining problem and utilize mining techniques [25] to find frequent patterns that describe the characteristics of events [43].

Generating symbolic streams

We take each court view shot in the video as a time unit and transform the extracted features of each time unit into symbolic streams according to

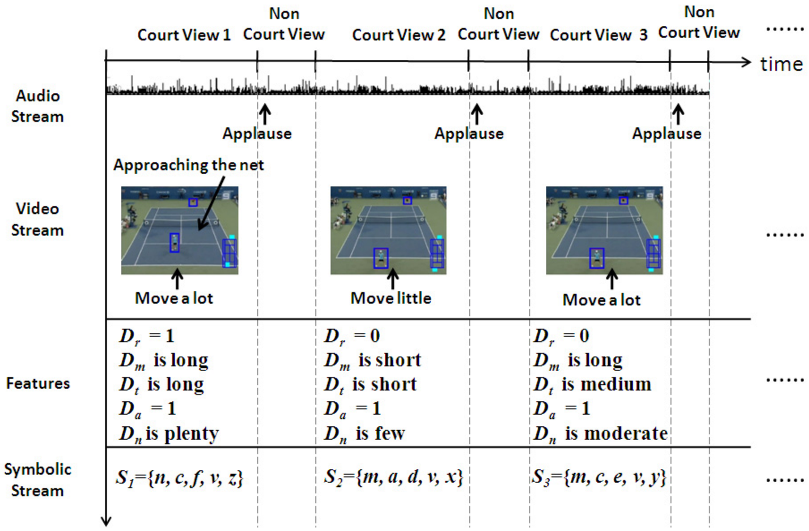


Fig. 12. The Examples of symbolic streams. Each court view shot is transformed to a corresponding symbolic stream Si according to the middle-level features. The mapping between features and symbols is listed in Table 1.

the mapping given in Table 1. As shown in Table 1, each feature D_* has its corresponding symbol set; for example, the symbol set of D_m is $\{a, b, c\}$. Fig. 12 shows some examples of symbolic streams. Let S_i be the symbolic stream representing the features derived from a particular time instant (court view shot) i. Given a video, a series of symbolic streams (denoted as $S = S_1, S_2, \cdots, S_n$, where n is the total number of court view shots) can be obtained.

Mining of frequent patterns

We define a *pattern* as $p = p_1 p_2 \cdots p_m$, where m is the number of symbols used to represent a symbolic stream S_i, and p_j is a subset of the underlying symbol set with respect to feature D_*. If p_j matches all the symbols in the underlying symbol set, we use the "don't care" character $*$ to denote p_j. Let $|p_j|$ be the number of "none don't care" (non-$*$) symbols in the set p_j. The length of a pattern p is defined as $\sum |p_j|$, and a pattern with length k is called a $k-pattern$. Moreover, we define *subpattern* of a pattern $p = p_1 p_2 \cdots p_m$ as a pattern $p' = {p_1}'{p_2}' \cdots {p_m}'$ such that ${p_j}' \subseteq p_j$ for every j where ${p_j}' \neq *$. Due to a strong correlation between frequencies of patterns and their subpatterns, the traditional Apriori-Algorithm [1] may reduce the search space in mining slowly. Consequently, the Max-subpattern Tree introduced in [26] is adopted to efficiently find frequent patterns in S.

Algorithm 3 (Event Mining Algorithm)

```
Step1. Scan $S$ once to find the set of frequent 1-patterns ($F_1$), by accumulating
       the frequent count for each 1-pattern and selecting among them whose
       frequent count is no less than the given threshold, $Th_1$. Form the
       candidate frequent max-pattern $C_{max}$ from $F_1$ and take $C_{max}$ as the
       root of the Max-subpattern Tree.
Step2. Scan $S$ once. For each symbolic stream $S_i$, insert $MS(S_i, C_{max})$ into
       the Max-subpattern Tree with its count=1 if it is not already there;
       otherwise, increase the count of $MS(S_i, C_{max})$ by one. The detail of
       the insertion algorithm can be found in [26].
Step3. Obtain the set of frequent k-patterns from the Max-subpattern Tree :
       for k=2 to length of $C_{max}$
       {
           Derive candidate patterns of length k from frequent patterns of
           length k-1.
           Scan the Max-subpattern Tree to find frequency count of these candidate
           patterns and eliminate the non-frequent ones. The frequency
           count of each node is calculated by summing the count values of the
           node itself and its ancestor in the Max-subpattern Tree. If the derived
           frequent k-pattern set is empty, return.
       }
```

Follow the definitions given in [26], let F_1 be the set of frequent 1-patterns. A *candidate frequent max-pattern*, C_{max}, is the maximal pattern which can be derived from F_1. For example, if the frequent 1-pattern set is $\{m * * * *, n * * * *, *a * **, * * *v*, * * * * z\}$, C_{max} will be $\{m, n\}a * vz$. The *maximal subpattern* of two patterns p^1 and p^2 is denoted by $MS(p^1, p^2)$ and defined as follows: $MS(p^1, p^2)$ is a common subpattern of both p^1 and p^2, in addition, none of other common subpattern has the length longer than $MS(p^1, p^2)$. For example, if $p^1 = \{m, n\}a * vz$ and $p^2 = maguz$, $MS(p^1, p^2)$ will be $ma * *z$. Based on the above-mentioned definitions, the mining algorithm is shown in Algorithm 3.

We go through each frequent pattern derived from the mining algorithm and manually map all frequent patterns to corresponding events. Frequent patterns mapped to the same event are merged into a set. Finally, we could categorize all frequent patterns into several sets and each set represents a specific event. According to the relationship between patterns and events, we can achieve event detection for the test videos. In contrast to the rule-based event detection which has to construct a decision tree based on specific domain knowledge, mining-based methodology is more general to be applied to various sports videos.

5 Applications

With the rise in the amount of sports video content, more and more applications in great demand have been proposed from the viewpoints of the audience, professional athletes, and advertisers. Applications for different purposes can be accomplished with the aid of different techniques used for sports video analysis. We introduce some interesting applications in this section.

5.1 Video Summarization and Highlight Extraction

It usually takes the audience hours to watch a whole sports video and acquire interesting or splendid events. However, valuable semantics generally occupy a small portion of the whole video content. A sports news TV program tries to make a summary of a match and to produce a sequence of highlights for the audience who have no time to enjoy the whole game. However, the summary and the highlight might not be the most representative for every viewer since they are both determined by few or even one producer. For example, in a basketball game, some viewers want to watch all dunk plays, but some viewers only want to watch the scoring plays of his/her favorite team. With techniques of sports video analysis introduced in previous sections, we can choose a more attractive way to summarize videos or extract highlights based on user preference. Furthermore, various multimedia-enabled receiving devices have different capability of storage and transmission speed, which emphasizes the importance of content adaptation. One perspective of content adaptation is to transmit different content with different semantics to various devices. Hence, designing a flexible mechanism of video summarization and highlight extraction becomes an essential matter.

In Sect. 2.1, we have described how to divide a sports video sequence into several video clips. For each video clip, we determine if a specific event takes place in it, as introduced in Sect. 4. The summaries can be generated by concatenating a set of video clips annotated with predefined events. Simply concatenating all these clips might result in a long summary, which is still copious for the viewer. A ranking list indicating the events importance can be defined according to the viewer's preference, such that the summary can select attractive events while leave out the less important ones for the viewer.

5.2 Tactic Analysis and Visualization

For professional players and coaches, collecting possible tactics taken by the opponent is quite essential since they can find the competitor's weaknesses and practice corresponding strategies before the match. Nowadays, human have to watch videos of several matches to conclude tactic information, which is time-consuming and exhausting. Thus, automatically analyzing sports videos and proving possible tactics has become a flourishing research topic. Trajectories of the ball and the players are the most useful cues in the video for tactic analysis. For example, *Zhu et al.* [58] exploited object trajectories and web-casting text to extracted tactic information from the goal events in broadcast soccer videos. *Chen et al.* [11] designed a physics-based algorithm to reconstruct 3D ball trajectory in basketball videos and then obtain shooting location statistics, which helps the defense team to infer which area has to be guarded with more attention. Moreover, how to visualize the extracted tactics is another interesting work. For each sport, we can classify tactics into several categories and take the statistics of each category as an indication that whether a team or a player is used to perform a specific tactic. Since tactics carried out before goal events provide significant information for the professional players/coaches, another fascinating way is to show the trajectories of the ball or a certain player in a period before an ongoing goal event [13]. If the scene in the video is calibrated with a 3D real-world model, one can even generate 3D cartoon of a moving object for visualization [32].

5.3 Advertisement/Virtual Content Insertion

The population of sports audience has been amazingly increasing, which leads the commercial industry to advertise their products along with the game. In addition to setting billboard around the field/stadium, or inserting commercials during break of the game, another choice is to automatically insert advertisement into the video content without switch to a stand-along commercial video. Such an advertisement insertion system automatically detects adequate insertion points in both temporal and spatial domains, and inserts the advertisement in a reasonable way. The insertion algorithm should consider not only if the advertisement can impress the audience, but also if the audience is seriously interrupted while watching the game. *Chang et al.* [10] took psychology, advertising theory, and computational aesthetics into account for improving the effectiveness of advertising in tennis videos. *Liu et al.* [34] proposed a generic system which can insert virtual content (not only advertisement, but also other message such as information of the game or the player) into videos based on visual attention analysis.

6 Conclusions

In this chapter, we comprehensively introduce the state of the art techniques for sports videos analysis, and propose a generic framework for sports video annotation. We explicitly elucidate how to extract visual/audio mid-level features, and practical examples are used for helping the reader easily grasp the main ideas of how these mid-level features work in different methodologies for event annotation. Moreover, applications from three viewpoints (i.e. the audience, professional athletes, and advertisers) are explained to strengthen the importance of sports video analysis.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. Int. Conf. on Very Large Data Bases, pp. 487–499 (1994)
2. Allen, J.G., Xu, R.Y.D., Jin, J.S.: Object tracking using CamShift algorithm and multiple quantized feature spaces. In: Proc. Pan-Sydney Area Workshop on Visual Information Processing (2004)
3. Assfalg, J., Bertini, M., Bimbo, A.D., et al.: Soccer highlights detection and recognition using HMMs. In: Proc. IEEE ICME, vol. 1, pp. 825–828 (2002)
4. Babaguchi, N., Kawai, Y., Kitahashi, T.: Event based indexing of broadcasted sports video by intermodal collaboration. IEEE T. Multimedia 4(1), 68–75 (2002)
5. Bach, N.H., Shinoda, K., Furui, S.: Robust highlight extraction using multi-stream hidden Markov models for baseball video. In: Proc. IEEE ICIP, vol. 3, pp. 173–176 (2005)
6. Bai, H., Hu, W., Wang, T., et al.: A novel sports video logo detector based on motion analysis. In: King, I., Wang, J., Chan, L.-W., Wang, D. (eds.) ICONIP 2006. LNCS, vol. 4233, pp. 448–457. Springer, Heidelberg (2006)
7. Bishop, C.M.: Pattern recognition and machine learning. Springer, Heidelberg (2006)
8. Cai, R., Lu, L., Zhang, H.-J., Cai, L.-H.: Highlight sound effects detection in audio stream. In: Proc. IEEE ICME, vol. 3, pp. 37–40 (2003)
9. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/
10. Chang, C.-H., Hsieh, K.-Y., Chung, M.-C., et al.: ViSA: virtual spotlighted advertising. In: Proc. ACM MM, pp. 837–840 (2008)
11. Chen, H.-T., Tien, M.-C., Chen, Y.-W., et al.: Physics-based ball tracking and 3D trajectory reconstruction with applications to shooting location estimation in basketball video. J. Vis. Commun. Image R. 20(3), 204–216 (2009)
12. Cheng, W.-H., Chu, W.-T., Wu, J.-L.: Semantic context detection based on hierar-chical audio models. In: Proc. ACM SIGMM Int. Workshop on Multimedia Information Retrieval, pp. 109–115 (2003)
13. Chou, C.-W., Tien, M.-C., Wu, J.-L.: Billiard Wizard: A tutoring system for broad-casting nine-ball billiards videos. In: Proc. IEEE ICASSP, pp. 1921–1924 (2009)
14. Chu, W.-T., Wu, J.-L.: Explicit semantic events detection and development of realistic applications for broadcasting baseball videos. Multimedia Tools and Applications 38(1), 27–50 (2008)
15. Duda, R.O., Hart, P.E., Stock, D.G.: Pattern classification. John Wiley and Sons, Chichester (2001)
16. Ekin, A., Tekalp, A.M.: Shot type classification by dominant color for sports video segmentation and summarization. In: Proc. ICASSP, vol. 3, pp. 173–176 (2003)
17. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and sum-marization. IEEE T. Image Process. 12(7), 796–807 (2003)
18. Fan, J., Elmagarmid, A.K., Zhu, X., et al.: ClassView: hierarchical video shot classification, indexing, and accessing. IEEE T. Multimedia 6(1), 70–86 (2004)
19. Farin, D., Han, J., With, P.H.N.: Fast camera calibration for the analysis of sport sequences. In: Proc. IEEE ICME, pp. 482–485 (2005)

20. Farin, D., Krabbe, S., With, P.H.N., et al.: Robust camera calibration for sport videos using court models. In: Storage and retrieval methods and applications for multimedia, vol. 5307, pp. 80–91. SPIE, CA (2004)
21. Gao, X., Yang, Y., Tao, D., et al.: Discriminative optical flow tensor for video semantic analysis. Computer Vision and Image Understanding (Elsevier) 113(3), 372–383 (2009)
22. Gao, X., Li, X., Feng, J., et al.: Shot-based video retrieval with optical flow tensor and HMMs. Pattern Recognition Letters (Elsevier) 30(2), 140–147 (2009)
23. Gong, Y., Sin, L.T., Chuan, C.H., et al.: Automatic parsing of TV soccer programs. In: Proc. IEEE Int. Conf. on Multi Comput. Syst., pp. 167–174 (1995)
24. Guéziec, A.: Tracking pitches for broadcast television. IEEE Computer 35(3), 38–43 (2002)
25. Han, J., Kamber, M.: Data Mning: Concepts and Techniques. Morgan Kaufmann, San Francisco (2005)
26. Han, J., Dong, G., Yin, Y.: Efficient mining of partial periodic patterns in time series database. In: Proc. Int. Conf. on Data Engineering, pp. 106–115 (1999)
27. Hanjalic, A.: Shot-boundary detection: unraveled and resolved? IEEE T. Circ. Syst. Vid. 12(2), 90–105 (2002)
28. Haoran, Y., Rajan, D., Tien, C.L.: A unified approach to detection of shot boundaries and subshots in compressed video. In: Proc. IEEE ICIP, vol. 2, pp. 1005–1008 (2003)
29. Intille, S.S., Bobick, A.F.: Recognizing planned, multi-person action. Comput. Vis. Image Und. 81, 414–445 (2001)
30. Leonardi, R., Migliorati, P., Prandini, M.: Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled Markov chains. IEEE T. Circ. Syst. Vid. 14(5), 634–643 (2004)
31. Li, B., Sezan, I.: Event detection and summarization in American football broadcast video. Storage and retrieval for media databases. In: SPIE, vol. 4676, pp. 202–213 (2002)
32. Liang, D., Liu, Y., Huang, Q., et al.: Video2Cartoon: generating 3D cartoon from broadcast soccer video. In: Proc. ACM MM, pp. 217–218 (2005)
33. Lienhart, R.: Comparison of automatic shot boundary detection algorithms. Storage and retrieval for image and video databases. In: SPIE, vol. 3656, pp. 290–301 (1999)
34. Liu, H., Jiang, S., Huang, Q., et al.: A generic virtual content insertion system based on visual attention analysis. In: Proc. ACM MM, pp. 379–388 (2008)
35. Liu, S., Xu, M., Yi, H., et al.: Multimodal semantic analysis and annotation for basketball video. EURASIP J. Appl. Si. Pr. 2006, 1–13 (2006)
36. Niu, Z., Gao, X., Tao, D., et al.: Semantic video shot segmentation based on color ratio feature and SVM. IEEE Cyberworlds, 157–162 (2008)
37. Pan, H., van Beek, P., Sezan, M.I.: Detection of slow-motion replay segments in sports video for highlights generation. In: Proc. IEEE ICASSP, vol. 3, pp. 1649–1652 (2001)
38. Pingali, G.S., Jean, Y., Carlbom, I.: Real time tracking for enhanced tennis broadcasts. In: Proc. IEEE CVPR, pp. 260–265 (1998)
39. Plataniotis, K.N., Venetsanopoulos, A.N.: Color Image Processing and Applications. Springer, Berlin (2000)
40. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for TV Baseball programs. In: Proc. ACM MM, pp. 105–115 (2000)

41. Shen, J., Tao, D., Li, X.: Modality mixture projections for semantic video event detection. IEEE T. Circ. Syst. Vid. 18(11), 1587–1596 (2008)
42. Shum, H., Komura, T.: A spatiotemporal approach to extract the 3D trajectory of the baseball from a single view video sequence. In: Proc. IEEE ICME, vol. 3, pp. 1583–1586 (2004)
43. Tien, M.-C., Wang, Y.-T., Chou, C.-W., et al.: Event detection in tennis matches based on video data mining. In: IEEE ICME, pp. 1477–1480 (2008)
44. Tong, X., Lu, H., Liu, Q., et al.: Replay detection in broadcasting sports video. In: Proc. IEEE ICIG, pp. 337–340 (2004)
45. Tovinkere, V., Qian, R.J.: Detecting semantic events in soccer games: Towards a complete solution. In: Proc. IEEE ICME, pp. 1040–1043 (2001)
46. Wang, J., Xu, C., Chng, E., et al.: Event detection based on non-broadcast sports video. In: Proc. IEEE ICIP, vol. 3, pp. 1637–1640 (2004)
47. Wang, L., Liu, X., Lin, S., et al.: Generic slow-motion replay detection in sports video. In: Proc. IEEE ICIP, vol. 3, pp. 1585–1588 (2004)
48. Welch, G., Bishop, G.: An introduction to the Kalman Filter, http://www.cs.unc.edu/~welch/kalman/kalmanIntro.html
49. Xie, L., Chang, S.-F., Divakaran, A., et al.: Structure analysis of soccer video with Hidden Markov Models. Pattern Recognition Letters, 767–775 (2002)
50. Xu, C., Wang, J., Lu, H., et al.: A novel framework for semantic annotation and personalized retrieval of sports video. IEEE T. Multimedia 10(3), 325–329 (2008)
51. Xu, P., Xie, L., Chang, S.-F., et al.: Algorithms and system for segmentation and structure analysis in soccer video. In: Proc. IEEE ICME, pp. 721–724 (2001)
52. Ye, Q., Huang, Q., Gao, W., et al.: Exciting event detection in broadcast soccer video with mid-level description and incremental learning. In: Proc. ACM MM, pp. 455–458 (2005)
53. Yeo, B.-L., Liu, B.: Rapid scene analysis on compressed video. IEEE T. Circ. Syst. Vid. 5(6), 533–544 (1995)
54. Zabih, R., Miller, J., Mai, K.: A feature-based algorithm for detecting and classifying scene breaks. In: Proc. ACM MM, pp. 189–200 (1995)
55. Zhang, B., Dou, W., Chen, L.: Ball hit detection in table tennis games based on audio analysis. In: Proc. IEEE ICPR, vol. 3, pp. 220–223 (2006)
56. Zhang, D., Chang, S.-F.: Event detection in baseball video using superimposed caption recognition. In: Proc. ACM MM, pp. 315–318 (2002)
57. Zhang, H., Kankanhalli, A., Smoliar, S.W.: Automatic partitioning of full-motion video. Multimedia Syst. 1(1), 10–28 (1993)
58. Zhu, G., Huang, Q., Xu, C., et al.: Trajectory based event tactics analysis in broadcast sports video. In: Proc. ACM MM, pp. 58–67 (2007)

Abbreviations

accumulated color histogram difference (ACHD)
closed captions (CC)
color histograms difference (CHD)
Continuously Adaptive Mean Shift (CamShift)
dominant MB change (DMBC)
edge change ratio (ECR)
expectation maximization (EM)
frame dissimilarity ratio (FDR)
Gaussian Mixture Model (GMM)
Hidden Markov Model (HMM)
macroblocks (MBs)
Mel-frequency cepstral coefficients (MFCCs)
modified frame dissimilarity ratio (MFDR)
region of interest (ROI)
shot boundary detection (SBD)
support vector machines (SVM)

Semantic Content Analysis of Video: Issues and Trends

Aparna Garg and Allan Ramsay

School of Computer Science, University of Manchester, U.K.
A.Garg@postgrad.manchester.ac.uk
allan.ramsay@manchester.ac.uk

Abstract. Key issues in bridging the semantic gap for content analysis of video include flexibility required from the software, real time implementation and cost effectiveness. In recent years industry has begun to take a more realistic view of what to expect from video content analysis systems in the near future. This chapter presents the state-of–the-art trends in semantic video analysis in industry. The key challenges in bridging the semantic gap are discussed. It also presents the research trends in *video analytics*.

Keywords: Semantic, video, understanding, analytics, trends, challenges, issues, semantic-gap.

1 Introduction

Semantic content analysis of video infers high-level concepts in videos from low-level visual information represented in pixels. A concept of interest could be presence of an object or certain behavior such as fighting in the video. A video has both the static and changing visual information. The analysis of static visual information is done using image analysis algorithms. In this chapter the focus is on issues and methods specific to video i.e. the changing visual information.

With the proliferation of video data, the need for automatic annotation of video for analysis as well as retrieval is being felt by industry. However content understanding of video is still rather under-developed and bridging the s*emantic gap* between low level features and high level semantic representation of video is an open challenge. Some of the main issues that limit the real world application of semantic analysis are limited flexibility of current content analysis software, hardware limitations for real time processing of video signal and cost effectiveness of solutions.

In this chapter the need for semantic content analysis and state-of-the-art methods for extraction of high-level semantic data from low-level features are presented. The problems and challenges in bridging the *semantic gap* are discussed with real world examples. The research trends in video analytics that have emerged over past few years are reviewed.

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 443–457.
springerlink.com

2 Need for Semantic Content Analysis in Real World Applications

In the last decade there has been an explosion in the use of videos, especially for security related tasks. Some of the sectors with huge repositories are the World Wide Web, broadcasting, security, retail, transport, government, medicine etc. It is not humanly possible to annotate or monitor these videos; hence most of this data never gets used because it is difficult to locate specific content. For example, following the Brixton bombing in 1999, police seized 1097 SVHS video tapes containing approximately 26,000 hours of CCTV footage. In order to locate the correct video clip, police officers had to manually search through 100 million separate frames of recorded video. According to a report only 30% CCTV footage remains accessible after 15 minutes. Similarly broadcast houses own millions of videos, but it is not possible to search for a news clip with Obama and Clinton together in it. For another example consider a security operator in a super-store trying to monitor live feed from 16-30 cameras for few hours. It is not humanly possible to keep track of what is happening in all the cameras. To be able to get the benefit of this multimedia data it is imperative that the data can be automatically analyzed and described in a way that is simple to understand by a layman. The analytics can be post-event or online. For post-event analysis it should be possible to quickly identify and retrieve the data which meets particular criteria, for example a black car that turned right. For online analysis events of interest should be automatically highlighted to the operator and video clips annotated for later retrieval, for example a person going from car to car in a car park.

The only data that is generally tagged to the video clip by current generation systems is camera id, location and time. The operator some-times manually tags other data of interest. There are well-established image processing algorithms, which extract color and shape of a moving object. However the low-level image processing-based methods (i.e. color and texture) exhibit practical drawbacks and most users would prefer to retrieve images and sequences on the basis of higher-level (i.e. semantic) features.

3 Semantic Analysis System Architecture

A typical video system architecture is shown below in figure 1. It comprises of video cameras which can be analog, digital or smart cameras. The smart cameras and intelligent edge devices have processors (DSP/embedded) and implement video processing at source (edge) itself. Intelligent edge devices/smart cameras shift the burden of processing from the servers and by transmitting processed data e.g. semantic description of video can reduce internet data volume. The video data, which may be processed or raw, is sent to servers which can be local or central. Here processed data means various levels of processing like compression, low level feature extraction or semantic feature extraction.

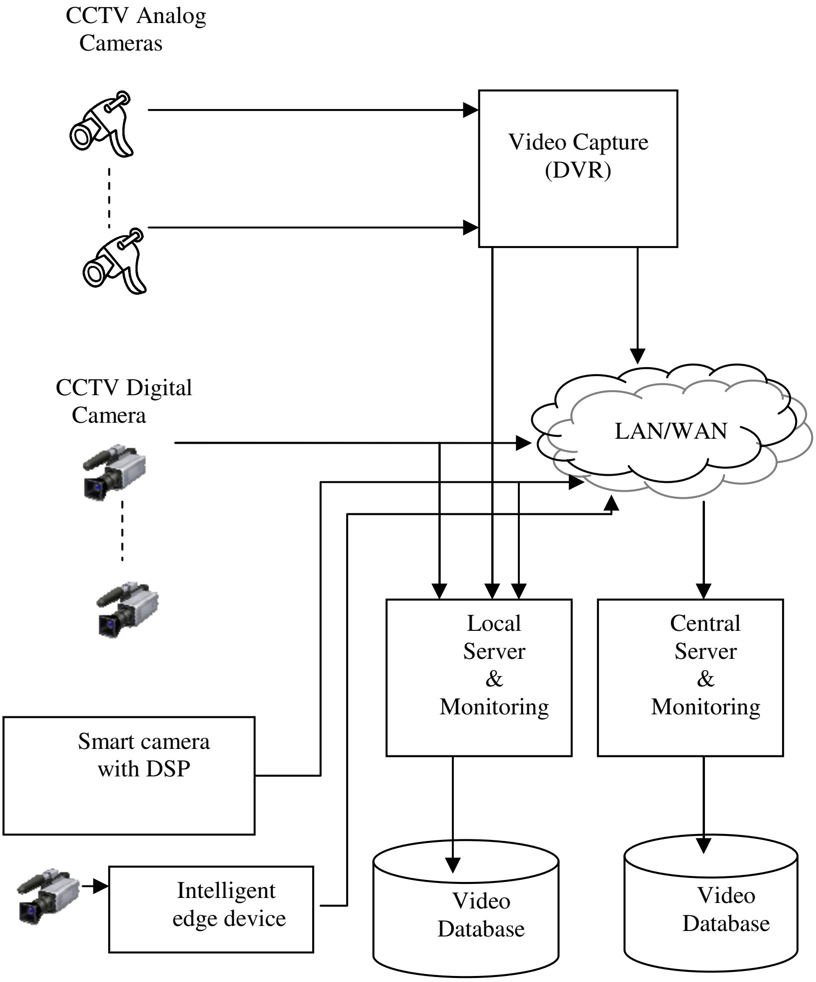


Fig. 1. Video system architecture

Figure 2 shows typical modules of a semantic video analysis system. The raw video from cameras is segregated into shots depending on user requirement. Object motion features in shots is detected and used for blob shape analysis and motion tracking. These low-level video features are used by semantic analysis system to extract high level human language like description of raw video data.

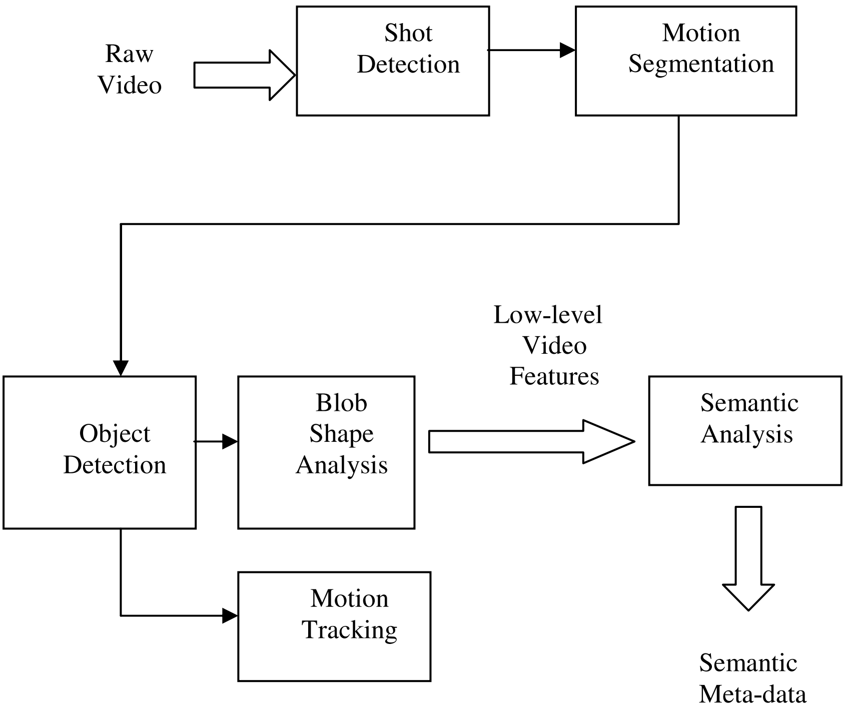


Fig. 2. Block diagram of semantic video analysis system

4 Problems and Issues in Semantic Analysis of Real World Applications

Video query/analysis by semantic keywords is one of the most difficult problems in multimedia data retrieval/analytics. The difficulty lies in the mapping between low-level video representation and high-level semantics. For instance the mapping of a human blob shape and trajectory to a high level concept such as dancing. This problem is referred in the literature as the *semantic gap*; current algorithms in computer vision cannot bridge the semantic gap. Some of the key issues in bridging the semantic gap are flexibility required from the software, real time processing power required, cost effectiveness and limited availability of concept labeled data. These issues are discussed in the following paragraphs.

4.1 Flexibility

Video Analytics requires the mapping software to be flexible in how it maps the low level features to high-level semantic concepts. For example [26] defines 'approach' as 'change of state from being far to being near'. Consider an example of a person approaching a store window as shown in figure 3 from the CAVIAR dataset [44], the arrow shows the direction of motion. This requires defining 'far' and 'near' in terms of pixels, so if the camera zooms or pans the definition is no longer valid. Suppose that the software can somehow automatically acquire homography mapping and account for change in pan and zoom without manual calibration; even so, a high level descriptor such as 'approach window' in case of another window in the same scene or different scene still needs to be redefined. Now consider that someone parked a wheel-chair in front of the shop window so that it is not possible to go so near the window; this necessitates re-defining of 'near window' state. It must be also noted that the area defined as 'near' has to be carefully planned, as this definition would miss all cases where a person is already near and moves nearer.

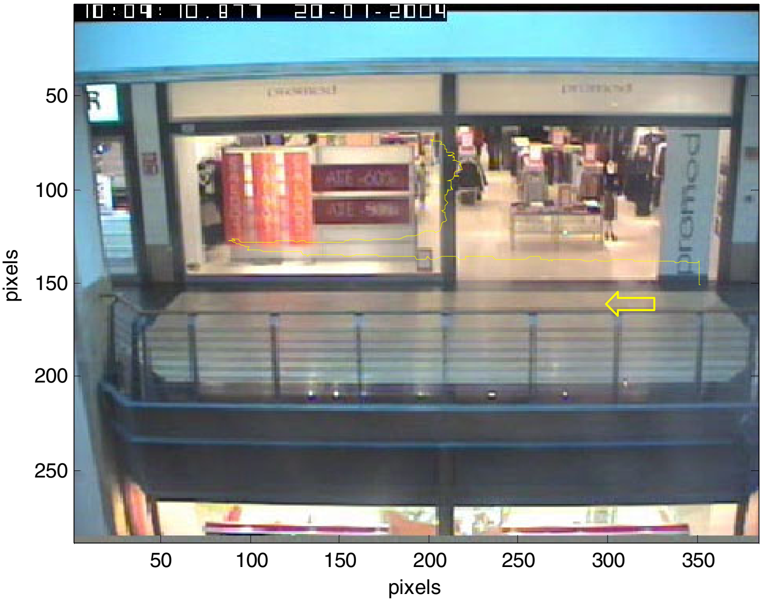


Fig 3. Trajectory of a person who approaches display window browses the window and then goes into the store

Even a simple event like approach requires more flexibility from the semantic video analysis system than can usually be provided. The software should be able to auto calibrate when the camera pans, tilts or zooms as well as automatically learn the far and near concept for approach behavior when the environment or context changes.

Consider applying this concept of 'near' to the slightly more complex event of people 'walking together'. 'Walking together' can be inferred from two concepts near each other at same time and moving in same general direction. Here the concept of ‘near’ would depend on the context, for example people walking in an open field would be ‘walking together’ even when they are not very near to each other. Hence the flexibility demanded from the concept ‘near’ is the flexibility in context of location. An inference algorithm for ‘walking together’ is given in [3] where a nearest neighbor classification scheme based on non-parametric feature vector classifies walk together events, however for a different location retraining with new data would be required.

From the example above it can be seen that even for inference of simple semantic concepts flexibility or adaptability in terms of location context, event context and change in environment is required.

4.2 Availability of Ground Truth Data

There are about 20-30 ground truth tagged video repositories which are freely available to bench-mark algorithms. [34, 36] emphasize that there is lack of ground truth annotated training data to build classifiers to map low level features to high level concepts. Lack of suitable ground truth data also restricts benchmarking of algorithms as they can’t be tested under different conditions and different locations. [46] lists most of the benchmarking video data repositories which are available. Manually annotating video data is time consuming and expensive. Also it is difficult to know in advance what ground truth information will be required for a future algorithm hence tailored ground truth data is not available. Moreover it is generally not easy to adapt existing ground truth data for testing different cognitive tasks. [36] mentions that it is not known what to model and where to get data to model it.

4.3 Real Time Implementation

Meta-data tagging, theft prevention, dangerous event analysis are some of the cases where video analytics has to be online and in real time. Video analytic solutions are computation intensive. In order to cope with this, low level algorithms are often implemented in real time digital signal processors and other smart embedded devices. The real time performance and hardware requirements for video analytic software is still an unexplored area. A few papers report real time performance of content analysis algorithms. [7] presents a metadata tagging algorithm for sports videos with real time performance. [4, 21] reports real time performance of a dsp based smart camera for video analytic computations. [7] mentions that because video analytic algorithms are still in their infancy, optimized real time implementations have only been recently considered. This paper also stresses the need for establishment of performance metrics for evaluation of real time video analytic algorithms.

4.4 Other Issues

Some of the other issues in deployment of video analytics in real world applications are high cost, robustness and accuracy.

The high cost of video analytic systems comes from both hardware cost for computation intensive analytics as well as cost of software development because of limited flexibility of analytic software. As video analytics can only detect a few types of simple events reliably, the high cost of analytic systems is not usually justified and prevents large scale deployment of video.

Robustness of a video analytic algorithm is its performance under adverse conditions. However according to [6] visual analysis of events is a complex task and success at one level cannot be predicated upon success at lower levels and definition of robustness is an open issue. In real world environments the performance of video analytic algorithms can deteriorate rapidly due to changes in weather, environmental changes etc. This results in an increased false positive detection rate for an event and limits successful application of video analytics.

Video Analytics is still an area with many open questions. It is not known to what extent the problem of the semantic gap can be solved by having a sufficient repository of data and to what extent the solutions will be limited by algorithm's inability to infer the change in environment and the impossibility of having a repository of all possible events. Further [6] mentions that while humans are good at certain cognition tasks, for instance detecting unusual behavior, they are poor in other tasks like looking for abandoned baggage in a crowded scene.

5 Real World Applications

So far there are very few applications in the real world which use video analytics. This is because of the difficulty in bridging the *semantic-gap*. Current video analytic solutions are not flexible enough for the real world environment, they are expensive to implement and real world performance of algorithms often deteriorates rapidly. The research trails behind what industry requires. However the video analytic solutions that industry has adopted are a good indication of what works in the real world and is cost effective. Some of the state of art solutions in application areas of broadcast, surveillance, and business intelligence are discussed below.

5.1 Broadcast

There is an urgent need for tools to annotate broadcast and internet videos but only semi-automatic tools to annotate videos are available. These tools provide automatic low level feature extraction support like shot segmentation and object detection. High-level concepts are then input manually like object recognition, event description etc.

5.2 Surveillance

The surveillance industry is now beginning to deploy automatic event analysis tools for some specific events like loitering, abandoned baggage, sterile zone monitoring, perimeter crossing, slip and fall events, counting people/vehicles, tailgating, point of sale fraud detection.

Even these few event analysis solutions that are available suffer from a high number of false alarms and high cost, and cannot cope with small changes in environment.

5.3 Business Intelligence

Video analytics in department stores are being used to collect data on shopping patterns. Low-level features like time spent in front of display, facial expression, gaze direction, shopper interaction with products, point-of-sale information are used to infer high level concepts. The higher level semantic concepts like a person looked at the promotion and made a purchase or did not make a purchase can be inferred.

6 Research Trends

This section presents a taxonomy of semantic content analysis peculiar to video i.e. analysis of change in the video image frames (motion). It also discusses the research trends based on the taxonomy. It is not a complete survey of research in this area. There are two main application areas for motion analytic algorithms – for humans and for vehicles. Human motion is complex where-as vehicle motion is more constrained. Hence semantic content analysis algorithms have been more successful and have matured further for vehicle motion than for human motion.

Approaches to activity analysis can be broadly classified into two – predefined model based approach where the mapping of predefined activity from low level image features is learnt or adopting a hierarchical divide-and-conquer approach where primitive activities are used to infer a complex activity. In the past few years the trend has been to build models of primitive actions and infer complex actions from primitive actions. A taxonomy of semantic event analysis based on approaches adopted is shown in figure 4.

6.1 Model Based Event Analysis

Model-based semantic event analysis is used to learn predefined activities of interest. Models are learnt using a mapping between low level spatial-temporal features and high level semantic concepts. Model based approaches have been most widely reported in research papers. Models infer and differentiate between simple activities like running, walking, meeting, cars turning at junctions, unusual/abnormal activity etc in a constrained environment. Most research has been reported on location specific environment. Supervised models are those which need classified training data. Unsupervised models use clustering approaches to detect unusual activity for unlabelled data.

6.1.1 Supervised Models

Supervised models use pattern recognition techniques and state space methods to extract semantic descriptions of behavior\action from video data. Crowd monitoring is another area where model based analysis has been reported in the last few years.

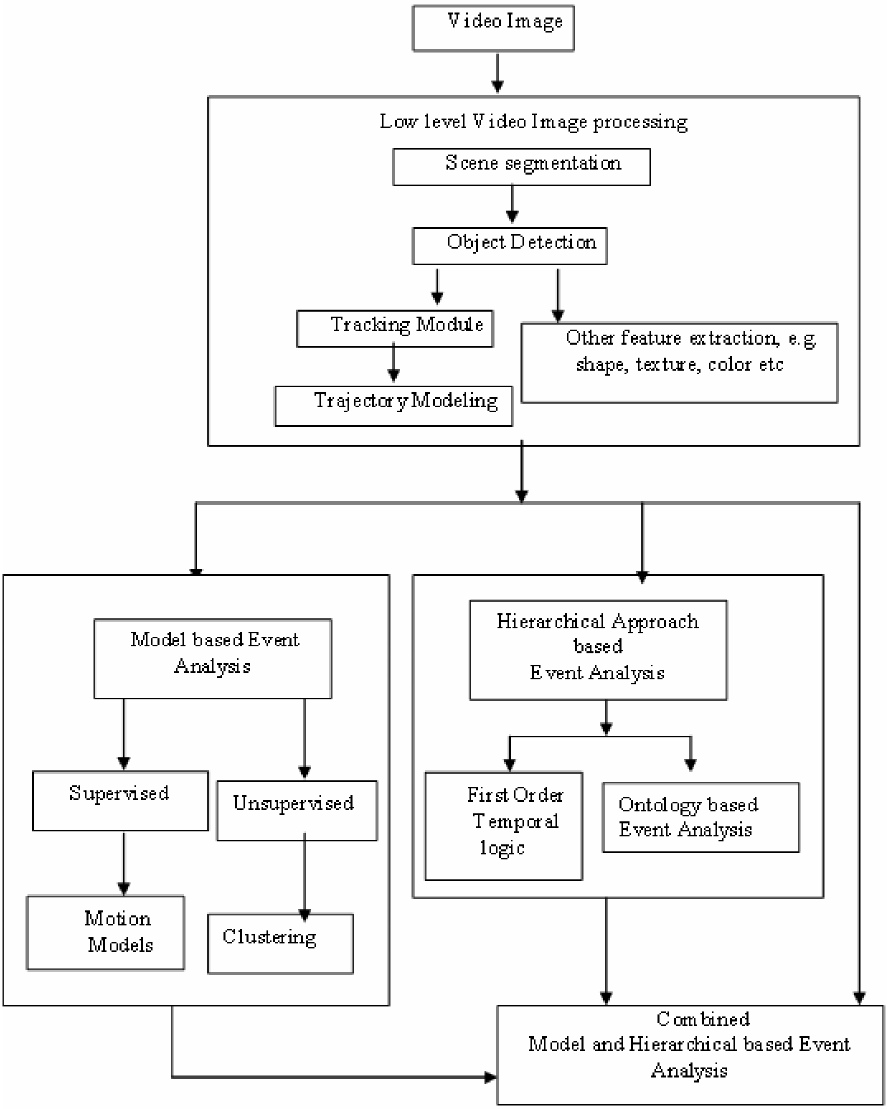

Video Image
Low level Video Image processing
Scene segmentation
Object Detection
Tracking Module
Trajectory Modeling
Other feature extraction, e.g shape, texture, color etc
Model based Event Analysis
Supervised
Unsupervised
Motion Models
Clustering
Hierarchical Approach based Event Analysis
First Order Temporal logic
Ontology based Event Analysis
Combined Model and Hierarchical based Event Analysis

Fig. 4. Taxonomy of semantic event analysis

Amongst state space models Hidden Markov Models are most popular. HMMs have been used for building state space models for prediction and cognition of simple or constrained complex actions. HMMs suited to representing complex behavior (states) evolving over time. HMM models whose internal states can represent semantic sub-events/primitive events have been reported. Actions in a room like washing dishes, reading, eating, going to a printer, printing, going to a computer, getting the phone have been inferred and explained in terms of sub-behavior in [5, 8, 27]. [11,30] annotates sports videos in terms of sub-behaviors. Nguyen et al. propose a hierarchical and therefore scalable Hidden Markov Memory model for recognizing complex high-level human behavior, which they tested on office activity data [27]. Brand et al. [5] show that observed activity can be organized into meaningful semantic states instead of being a black box by entropy minimization of joint distribution of an HMM internal state machine. The model was tested on office activity and traffic data. Duong et al. propose switching hidden semi-Markov model for activity classification and abnormality detection, the activities are modeled hierarchically [8]. The model was used to infer activities in a room like washing dishes, eating breakfast, reading news paper. Robertson and Reid report a methodology for action recognition using HMM to model behavior as a sequence of actions [30]. Actions are described by a feature vector comprising both trajectory information (position and velocity), and a set of local motion descriptors. Action recognition is achieved via probabilistic search of image feature databases representing previously seen actions. The model was used to annotate tennis videos. Gao et al. present a framework using HMM to recognize sports shots which are then used as input to a second HMM model to map high-level events [11][11]. Optical flow vectors are used to represent motion; dimension reduction of motion feature vector is achieved by applying general tensor discriminant analysis and linear discriminant analysis to optical flow tensors.

Interactions between people have been recognized by [3][3] and [28] using pattern recognition and state space methods respectively. Oliver et al. present and compare two state based models, HMM and Coupled HMM, for recognizing human behavior with a particular focus on human interactions [25]. Blunsden et al. apply nearest neighbor methods on non-parametric video feature representation to classify human interaction like walk together, meet, approach etc [3].

HMMs have also been used to automatically learn scene models. Makris et al. automatically learn scene semantic labels for fixed spatially related entities in the scene, semantic regions like entry/exit zones, paths/routes and junctions are learnt, using trajectory observations. HMMs interpret the observed activity in terms of spatially located scene features [24].

Foresti et al. in use long-term change detection algorithm to detect changes in the scene, which are then classified by neural networks for abandoned objects [10]. Also the trajectory data is used to learn a neural tree to classify usual and unusual human actions in a specific location. Nascimento et al. propose switched dynamical models to represent the human trajectories in a specific location. Activities like browsing display window, entering shop are then represented in terms of dynamical models [25]. [13] presents location independent method based on motion models to recognize specific activity such as depositing an object, exchanging bags, or removing an object.

Crowd motion modeling uses flow descriptors rather trajectory tracking to detect unusual activity in crowds and is an active area of research. Andrade et al. use optical flow methods instead of tracking statistics for single camera and dense crowd conditions to extract information from the crowd video data [2]. The optical flow features are encoded with Hidden Markov Models to detect emergency or abnormal events in the crowd. To increase the detection sensitivity a local modeling approach is used. The results were reported with simulated crowds. Ke et al. [19] use volumetric shape descriptors and flow descriptors of crowded scenes with distance matching methods to detect events of interest in a crowded scene. The method does not require figure/ground separation and was tested on events in cluttered scenes such as waving for a bus, pressing the elevator button, bend down to pick up an object. Hoogs et al. [15] propose using relational semantic predicates as constraints on graphs to perform spectral graph analysis to identify groups participating in an activity. The scheme was tested on crowd gathering and dispersal events; the approach puts no constraints on number of participants in the activity.

Biologically inspired pattern recognition models have been explored recently, largely for object and scene recognition. Serre et al. present a novel biologically inspired model for differentiating between scenes [31], and Huang et al. [14] and Song & Dacheng [37] have shown how the performance of this approach can be improved by concentrating on parts of the image that have interesting-looking content, and by using feedback to guide the search for models. Jhuang et al. extend a neurobiological model of motion processing in the visual cortex for action recognition[18]. The model was tested on human behaviour like walking, jogging, clapping etc and mice behaviour like drinking, eating, grooming etc. In general biologically inspired model approach requires a great deal of annotated training data, which as noted above is not generally available for video.

State space models need to re-learn if new action classes are added, this restricts their usability in real world domains. Also the state space models rely on video clips which have been appropriately segmented for detecting the event they have been trained for. Pattern recognition models are simple, can add new behavior classes easily and are practical to implement but they do not allow incorporation of semantic abstraction in terms of spatial-temporal variations easily.

6.1.2 Unsupervised Models

Clustering is one of the earliest approaches adopted to detect unusual data. Its advantage is that it does not require classified learning data and can learn from changes in the environment. Xiang et al. use unlabelled data to cluster similar behaviors and abnormal behaviors using clustering [43]. Low level feature vectors representing blobs and other shapes are used to infer behavior patterns. Izo et al. use a clustering model to detect anomalies in a busy urban scene [16]. Khalid and Naftel propose using Fourier transforms of motion trajectories to Self Organizing Maps to classify sign language trajectories and human motion trajectories [20].

6.2 Hierarchical Event Analysis

The hierarchical event analysis approach is based on the premise that complex events can be described in terms of a few primitive events. This approach is adopted with a view to making video event analytics more adaptable in regards to its ability to introduce new events which were not predefined at the time of system development. As video events are complex and varied, hierarchical event analysis for a large number of video events requires interaction between various development groups. Also for easy reuse the modules need to have standard definitions and interfaces. This has prompted researchers to develop ontologies for video events.

Ersoy et al. present a hierarchical framework based on first order logic, for using domain independent event primitives to model complex domain specific events [9]. Primitive events are mapped from low level trajectory based spatial-temporal features using hand-crafted rules. The results were tested on simulated car park data. Smith et al. present an Ontology framework to describe video events [35]. Vrusias et al. propose a framework to learn the domain ontology from text annotations, which is then applied to blob and trajectory data using statistical methods to automatically annotate unseen video footage with appropriate keywords that have been identified in human annotations of other videos [40].

6.3 Combined Model and Hierarchical Event Analysis

In recent years using a combination of model-based and hierarchical approaches has become a research trend. The primitive events models are learnt and complex events are defined/learnt in terms of primitive events. Various techniques like state-space, pattern–recognition, event calculus as well as approaches similar to those used for text mining have been reported in literature.

Bayesian Networks were used by [22, 29, 39, 41] to build hierarchical event models. In recent literature Bayesian networks have been increasingly proposed because of their ability to add inferred probabilities to a pre-defined model of the event space. Town used ontology to train structure and parameters of Bayesian Network for event Recognition [39]. The activities the model was learnt for were running walking, browsing, fighting etc. Park and Aggarwal use hierarchical action concept to recognize individual activities and interactions amongst entities [29]. Body pose is recognized using individual Bayesian network; at the next level Dynamic Bayesian Networks are used for recognizing actions and decision trees are used to recognize interactions at the highest level. Kwak and Han propose hierarchical model based event analysis using Dynamic Bayesian Networks [22]. Dynamic Bayesian networks provide a probabilistic framework to combine lower level events and temporal-logical relationships to infer higher level events. The framework was tested on ticket office transactions. Wang et al. use hierarchical Bayesian model in which atomic activities are modeled as distributions over low-level visual features, and interactions are modeled as distributions over atomic activities [41]. The framework extends existing language models Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) to model visual features and models interactions without supervision. Similar activities are clustered together. The activities in an aircraft ramp area were used for testing.

7 Conclusions

Semantic video analysis is still rather under-developed, but spurred by proliferation of video databases it has been an area of active research in recent years. A few applications have been used in the real world where environment can be highly constrained. However semantic video analysis environment is essentially unconstrained making semantic analysis an open challenge. Semantic video analysis from low level video features requires flexibility at each level of abstraction in terms of environment change, event description and location context. It is not known if this flexibility required from the system can be achieved. Other researchers have discussed lack of ground truth tagged video data repositories as well as lack of knowledge about what data to model. Also bench marking definitions for semantic video analytic algorithms are an open issue.

References

[1] Allen, J.F.: Maintaining knowledge about temporal interval. Communications of ACM 26(11), 832–843 (1983)

[2] Andrade Ernesto, L., Scott, B., Fisher Robert, B.: Hidden Markov Models for Optical Flow Analysis in Crowds. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 1, pp. 460–463 (2006)

[3] Scott, B., Ernesto, A., Robert, F.: Non Parametric Classification of Human Interaction. In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (eds.) IbPRIA 2007. LNCS, vol. 4478, pp. 347–354. Springer, Heidelberg (2007)

[4] Bramberger, M., Brunner, J., Rinner, B., Schwabach, H.: Real-time video analysis on an embedded smart camera for traffic surveillance. In: Proceedings of 10th IEEE Real-Time and Embedded Technology and Applications Symposium, RTAS 2004, pp. 174–181 (2004)

[5] Brand, M., Kettnaker, V.: Discovery and Segmentation of Activities in Video. IEEE Trans. Pattern Analysis and Machine Intelligence 22(8), 844–851 (2000)

[6] Dee Hannah, M., Velastin Sergio, A.: How close are we to solving the problem of automated visual surveillance? Machine Vision and Applications 19, 329–343 (2008)

[7] Desurmont, X., Wijnhoven, R.G.J.: Performance evaluation of real-time video content analysis systems in the CANDELA project. In: Proc. of the SPIE - Real-Time Imaging IX (2005)

[8] Duong, T.V., Bui, H.B., Phung, D.Q., Venkatesh, S.: Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. I, pp. 838–845 (2005)

[9] Ilker, E., Filiz, B., Subramanya, S.R.: A Framework for Trajectory Based Visual Event Retrieval. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2004), vol. 2, pp. 23–27 (2004)

[10] Luca, F.G., Lucio, M., Regazzoni Carlo, S.: Automatic Detection and Indexing of Video-Event Shots for Surveillance Application. IEEE Transactions On Multimedia 4(4), 459–471 (2002)

[11] Xinbo, G., Yimin, Y., Dacheng, T., Xuelong, L.: Discriminative Optical Flow Tensor for Video Semantic Analysis. In: Computer Vision and Image Understanding, vol. 113, pp. 372–383. Elsevier, Amsterdam (2009)

[12] Jungong, H., Dirk, F., Peter, H.N., Weilun, L.: Real-Time Video Content Analysis Tool for Consumer Media Storage System. IEEE Transactions on Consumer Electronics 52(3), 870–878 (2006)

[13] Ismail, H., David, H., Larry, S.: W4: Real-Time Surveillance of People and Their Activities. IEEE Transactions On Pattern Analysis And Machine Intelligence 22(8), 809–830 (2000)

[14] Yongzhen, H., Kaiqi, H., Liangsheng, W., Dacheng, T., Tieniu, T., Xuelong, L.: Enhanced Biologically Inspired Model. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)

[15] Hoogs, A., Bush, S., Brooksby, G., Perera, A.G.A., Dausch, M., Krahnstoever, N.: Detecting Semantic Group Activities Using Relational Clustering. In: IEEE Workshop on Motion and video Computing, WMVC 2008, January 8-9, pp. 1–8 (2008)

[16] Izo, T., Grimson, W.E.L.: Unsupervised Modeling of Object Tracks for Fast Anomaly Detection. In: IEEE International Conference on Image Processing, ICIP 2007, vol. 4, pp. 529–532 (2007)

[17] Jacobsen, C., Zscherpel, U., Perner, P.: A Comparison between Neural Networks and Decision Trees. In: Perner, P., Petrou, M. (eds.) MLDM 1999. LNCS (LNAI), vol. 1715, pp. 144–157. Springer, Heidelberg (1999)

[18] Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A Biologically Inspired System for Action Recognition. In: IEEE 11th International Conference on Computer Vision, vol. 14(21), pp. 1–8 (2007)

[19] Yan, K., Rahul, S., Martial, H.: Event Detection in Crowded Videos. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, October 14-21, pp. 1–8 (2007)

[20] Shehzad, K., Andrew, N.: Classifying spatiotemporal object trajec-tories using unsupervised learning of basis function coefficients. Multimedia Systems 12(3), 227–238 (2006)

[21] Richard, K., Anteneh, A., Ben, S., Alexander, D.: Camera mote with a high-performance parallel processor for real-time frame-based video processing. In: First ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2007, pp. 109–116 (2007)

[22] Suha, K., Hee, H.J.: Hierarchical Event Representation and Rec-ognition Method for Scalable Video Event Analysis. In: Tenth IEEE International Symposium on Multimedia, pp. 586–591 (2008)

[23] Li, J.Z., Ozsu, M.T., Szafron, D.: Modeling of moving objects in a video database. In: Proc. 4th Int. Conf. On Multimedia and Computing System, pp. 336–343 (1997)

[24] Makris, D., Ellis, T.J.: Automatic Learning of an Activity-Based Semantic Scene Model. In: IEEE International Conference on Advanced Video and Signal Based Surveillance, Miami, FL, USA, pp. 183–188 (2003)

[25] Nascimento, J.C., Figueiredo, M.A.T., Marques, J.S.: Segmentation and Classification of Human Activities. In: HAREM 2005: International Workshop on Human Activity Recognition and Modelling (2005)

[26] Nevatia, R., Hobbs Jand Bolles, B.: An Ontology for Video Event Representation. In: Proceedings of the 2004 IEEE Computer Society Conference on Com-puter Vision and Patter Recognition Workshops(CVPRW 2004) (2004)

[27] Nguyen, N.T., Bui, H.H., Venkatesh, S., West, G.: Recognizing and Monitor-ing High-Level Behaviors in Complex Spatial Environments. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 620–625 (2003)

[28] Oliver, N.M., Rosario, B., Pentland, A.P.: A Bayesian Computer Vision System for Modeling Human Interactions. IEEE Trans. Pattern Analysis and Machine Intelligence 22(8), 831–843 (2000)

[29] Park, S., Aggarwal, J.K.: Semantic-Level Understanding of Human Actions and Interactions Using Event Hierarchy. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop, pp. 2–12 (2004)

[30] Neil, R., Ian, R.: Behaviour understanding in video: a combined method. In: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV 2005), vol. 1, pp. 808–815 (2005)

[31] Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, pp. 994–1000 (2005)

[32] Shim, C.-B., Chang, J.-W., Kim, Y.-C.: Trajectory-Based Video Retrieval for Multimedia Information Systems. In: Yakhno, T. (ed.) ADVIS 2004. LNCS, vol. 3261, pp. 372–382. Springer, Heidelberg (2004)

[33] Siskind, J.: Grounding the Lexical Semantics of Verbs in Visual Perception using free Dynamics and Event logic. Artificial Intelligence Review 1(5), 31–90 (2001)

[34] Smeaton Alan, F.: Techniques used and open challenges to the analysis, indexing and retrieval of digital video. Information Systems 32, 545–559 (2007)

[35] Smith John, R.: VERL: An Ontology Framework for Representing and Annotating Video Events. IEEE MultiMedia, 76–86 (October-December 2005)

[36] Smith, J.R.: The Real Problem of Bridging the "Semantic Gap". In: Sebe, N., Liu, Y., Zhuang, Y.-t., Huang, T.S. (eds.) MCAM 2007. LNCS, vol. 4577, pp. 16–17. Springer, Heidelberg (2007)

[37] Dongjin, S., Dacheng, T.: C1 Units for Scene Classification. In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4 (2008)

[38] Tim, T., Wai, Y., Arbee, L., Chen, P.: Retrieving Video Data via Motion Tracks of Content Symbols. In: Proceedings of the Sixth International Conference on Information and Knowledge Management, pp. 105–112 (1997)

[39] Town, C.: Ontology-Driven Bayesian Networks for Dynamic Scene Understanding. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (2004)

[40] Bogdan, V., Dimitrios, M., John-Paul, R.: A Framework for Ontology Enriched Semantic Annotation of CCTV Video. In: Eight International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2007 (2007)

[41] Xiaogang, W., Xiaoxu, M., Grimson, E.: Unsupervised Activity Per-ception by Hierarchical Bayesian Models. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8 (2007)

[42] Tao, X., Shaogang, G.: Activity based surveillance video content modeling. Pattern Recognition 41(7), 2309–2326 (2008)

[43] Tao, X., Shaogang, G.: Video behaviour profiling and abnormality detection without manual labeling. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2(17-21), pp. 1238–1245 (2005)

[44] `http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1`

[45] `http://ngsim.fhwa.dot.gov/`

[46] `http://www.multitel.be/cantata/`

Semi-automatic Knowledge Extraction, Representation, and Context-Sensitive Intelligent Retrieval of Video Content Using Collateral Context Modelling with Scalable Ontological Networks

Atta Badii, Chattun Lallah, Meng Zhu, and Michael Crouch

Intelligent Media Systems and Services Research Centre (IMSS),
School of Systems Engineering, University of Reading
United Kingdom
e-mail: {atta.badii,c.lallah,meng.zhu,m.w.crouch}@reading.ac.uk

Abstract. There are still major challenges in the area of automatic indexing and retrieval of digital data. The main problem arises from the ever increasing mass of digital media and the lack of efficient methods for indexing and retrieval of such data based on the semantic content rather than keywords. To enable intelligent web interactions or even web filtering, we need to be capable of interpreting the information base in an intelligent manner. Research has been ongoing for several years in the field of ontological engineering with the aim of using ontologies to add knowledge to information. In this chapter we describe the architecture of a system designed to semi-automatically and intelligently index huge repositories of special effects video clips. The indexing is based on the semantic content of the video clips and uses a network of scalable ontologies to represent the semantic content to further enable intelligent retrieval.

1 Introduction

The advent of the Internet and digital media technologies has led to an enormous increase in the production and online availability of digital media assets as well as made the retrieval of particular media objects more challenging. Processing of digital data, such as text, image and video, has achieved great advances during the last few decades. However, as the well known 'semantic gap' [Sme2000] still exists between the low-level computational representation and the high-level conceptual understanding of the same information, more intelligent semantic-driven modelling, multi-modal indexing and retrieval for digital data are needed.

The DREAM (Dynamic REtrieval Analysis and semantic metadata Management) project aims at paving the way towards semi-automatic acquisition of knowledge from visual content. This is being undertaken in collaboration with Partners from the UK Film Industry, including Double Negative[1], The Foundry[2]

[1] Double Negative, http://www.dneg.com

[2] The Foundry, http://www.the-foundry.co.uk

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 459–474.
springerlink.com

and FilmLight[3]. Double Negative is the test partner who provided the test materials and user requirements and evaluated the system prototype. One of the main challenges for the users in this industry is the storage and management of huge repositories of multimedia data, in particular, video files, and, having to search through distributed repositories to find a particular video shot. For example, when Special Effects Designers need a category of clips containing "fire explosions" which they may wish to use in the making of a new special effect, it is a tedious and time consuming task for them to search for similar video clips which feature specific objects of interest. The first prototype of DREAM has been evaluated in this film post-production application domain and aims to resolve the existing problems in indexing and retrieving video clips.

This chapter presents the DREAM (Dynamic REtrieval Analysis and semantic metadata Management) research project which aims to prepare the way for the semi-automatic acquisition of knowledge from visual content, and thereby addressing the above mentioned problems. The chapter shows how Topic Map Technology [Pepp2009] [Garshol2002] has been used to model the semantic knowledge automatically extracted from video clips to enable efficient indexing and retrieval of those clips. The chapter also highlights some related work and presents how semi-automatic labelling and knowledge acquisition are carried out in the integrated framework. The chapter concludes with an analysis of the results of the evaluation of the prototype.

2 State of the Art – Multimedia Information Indexing and Retrieval

From time to time, researchers have sought to map low-level visual primitives to high-level semantic-related conceptual interpretations of a given media content. The objective has been to achieve enhanced image and video understanding which could further benefit in its subsequent indexing and retrieval. Early attempts in this research area have focussed on extracting high-level visual semantics from low-level image content. Typical examples include: discrimination between 'indoor' and 'outdoor' scenes [Szummer1998] [Paek1999], 'city' vs. 'landscape' [Gorkani94], 'natural' vs. 'manmade' [Bradshaw2000], etc. However, the granularity aspect of those research results is considered to be limited due to the fact that only the generic theme of the images can be identified.

Only recently researchers started to develop methods for automatically annotating images at object level. Mori et al [Mori1999] proposed a co-occurrence model which formulated the co-occurrency relationships between keywords and sub-blocks of images. The model had been further improved by Duygulu et al [Duygulu2002] using the Brown et al machine translational model [Brown1993] with the assumption that image annotation can be considered as a task of translating blobs into a vocabulary of keywords. Zhou and Huang [Zhou2000] explored how keywords and low-level content features can be unified for image retrieval.

[3] FilmLight, `http://www.filmlight.ltd.uk`

Westerveld [Westerveld2000] and Cascia et al [Cascia1998] sought to combine the visual cues of low-level visual features and textual cues of the collateral texts contained in the HTML documents of on-line newspaper archives with photos, to enhance the understanding of the visual content.

Li et al [Wang2003] developed a system for automatic linguistic indexing of pictures using a statistical modelling approach. The 2D multi-resolution Hidden Markov Model is used for profiling categories of images, each corresponding to a concept. A dictionary of concepts is built up, which is subsequently used as the linguistic indexing source. The system can automatically index images by firstly extracting multi-resolution block-based features, then selecting top *k* categories with the highest *log* likelihood for the given image to the category, and finally determining a small subset of key terms from the vocabulary of those selected categories stored in the concept dictionary as indexing terms.

The most recent effort in this area focussed on introducing a knowledge representation scheme, such as an ontology, into the process of semantic-based visual content tagging [Maillot2004] [Schober2004]. Athansiadis et al [Athanasisadis2007] proposed a framework for simultaneous image segmentation and object labelling. The work focused on a semantic analysis of images, which contributes to knowledge-assisted multimedia analysis and bridging the semantic gap. The possible semantic labels, formally represented as fuzzy sets, facilitate the decision making on handling image regions instead of the traditional visual features. Two well known image segmentation algorithms, i.e. watershed and recursive shortest spanning tree, were modified in order to stress the independence of the proposed method from a specific image segmentation approach. Meanwhile, an ontology-based visual context representation and analysis approach, blending global knowledge in interpreting each object locally, had been developed. Fuzzy algebra and ontological taxonomic knowledge representation had been combined in order to formulate the visual contextual information. The contextual knowledge had then been used to re-adjust the labelling of results of the semantic region growing, by means of fine-tuning the membership degrees of the detected concepts.

3 The DREAM Framework

The DREAM framework can be considered as a knowledge-assisted intelligent visual information indexing and retrieval system, which has been particularly tailored to serve the film post-production industry. The main challenge in this research work was to architect an indexing, retrieval and query support framework. The proposed framework exploits content, context and search-purpose knowledge as well as any other domain related knowledge in order to ensure robust and efficient semantic-based multimedia object labelling, indexing and retrieval. Our current framework provides for optional semi-automatic (man-in-the-loop) labelling; thus enabling semantically triggered human intervention to support optimal cooperation in the semi-automatic labelling of what is currently mostly a manual process. This framework is underpinned by a network of scalable ontologies, which grows alongside the ongoing incremental annotation of video content. To support these scalable ontologies, we deployed the Topic Map Technology, which also enables transparent and flexible multi-perspective access to the repository and pertinent knowledge.

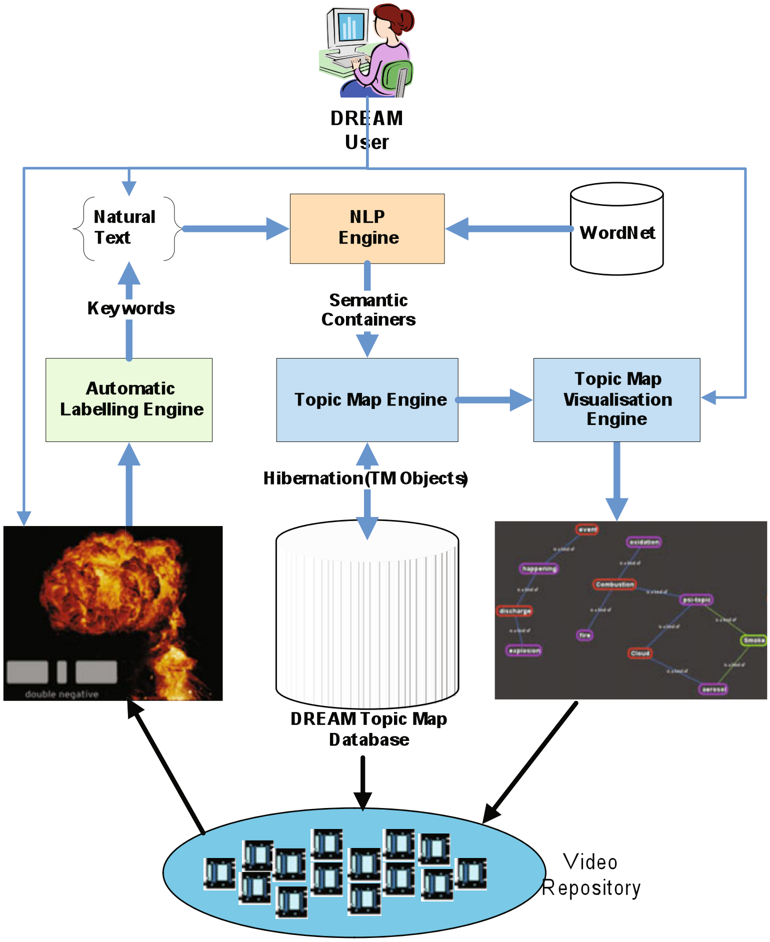


Fig. 1. DREAM Framework for film post-production domain

Figure 1 shows the architecture of the DREAM framework. In the film post-production domain, massive repositories of video clips are held in various locations. As the size of the repository grows, it becomes a nightmare for people trying to find a specific clip for a specific purpose at a specific time. The DREAM framework was developed to help overcome this bottleneck between the large volume of semantically disorganised data and the high-level demand in semantically enhanced efficient and robust indexing and retrieval of such data.

In deploying the DREAM framework, as illustrated in Figure 1, a User is able to query or navigate through the repository of video clips using the Topic Map Visualisation Engine which provides an interface to query the knowledge base or to navigate through the connections between the semantic concepts of the video clips. To support this query and navigate through the knowledge base, a Topic Map Engine has been designed and completed. In order to build the knowledge base, a Collaterally-Cued Automatic Labelling Module has been implemented. This reads in the video clips and extracts the main objects of interest, in terms of semantic keywords, from the clips. The User can then confirm those keywords and/or add more contextualised or domain-specific information so as to make their own viewpoint-specific set of keywords, which is then fed into the NLP Engine. The NLP Engine uses external knowledge such as WordNet to add meaning to the captured information which enhances the situated knowledge element. The situated knowledge element is then in a form that we term as "Semantic Containers", these are passed to the Topic Map Engine, which merges them with the existing knowledge found in the DREAM Topic Map Database. This enables intelligent querying and visualisation of the Semantic Network of concepts which indexes millions of video shots.

4 Knowledge Representation

The DREAM framework deploys Topic Map Technology to incorporate both content and context knowledge e.g. the knowledge of both episodic queries and their overall business context including users' dynamic role-based purposes, and, a priori higher level domain concepts (not just key words) that are so-to-speak "mind-mapped" to suit the users' relevant data-views ("i-schemas") for maximally transparent and flexible multi-perspective access to provide information retrieval that is context-sensitised, concept-oriented and a priori ontology-network-enabled.

Topic Maps, as defined in ISO/IEC 13250 [ISO2000], are an international standard for organising and representing information on the Web. A Topic Map can be represented by an XML document in which different element types are used to represent topics, occurrences of topics, and associations (or relationships) between topics. The Topic Map model provides a mechanism for representing large quantities of structured and unstructured information and organising it into "topics". The topics can be interrelated by an association and can have occurrences linked to them. A Topic Map can thus be referred to as a collection of topics and associations. The associations are used to navigate through the information in many different ways. The network can be extended as the size of the collection grows, or it is merged with other topic maps to provide additional paths through the information. The most important feature of the Topic Map is that it explicitly captures additional tacit information. It is this capability that has captured the interest of people working on knowledge management issues, identifying Topic Maps as a mechanism that can help to capture what could be considered "knowledge" from a set of information objects.

The real benefit of integrating Topic Maps within the DREAM Framework is the resulting retrieval gains that in turn confer high-scalability. Normally, the topic maps are linked to each other. It is very beneficial for a user to be able to incrementally develop semantically annotated subnets representing part of the media, and to build this up by linking it with representation of other parts. In this way, various contexts can be applied to the maps as they are linked together. During retrieval, it is natural for users to associate various topics in the way that is best suited to the formulation of the expected domain queries that serve the objectives of the domain process, in other words, the process logic e.g. editing goals that are typically assigned to be completed by the editors. The natural evolution of ideas amongst those engaged in the process may prompt them to re-visit certain associations, extend or refine some and add other new associations. Accordingly the facility for creating new associations amongst the topics exists. Hence, the topic maps are continuously evolving during the entire life of the repository, as new content is added, there are always chances of discovering new associations, both intra and inter content associations.

5 Knowledge Acquisition

This section describes the process of knowledge acquisition through the annotation of video content within the DREAM framework. Variation of the different annotation methods is supported, including automatic annotation and textual annotation, manual annotation, and, visual annotation as shown in Figure 3.

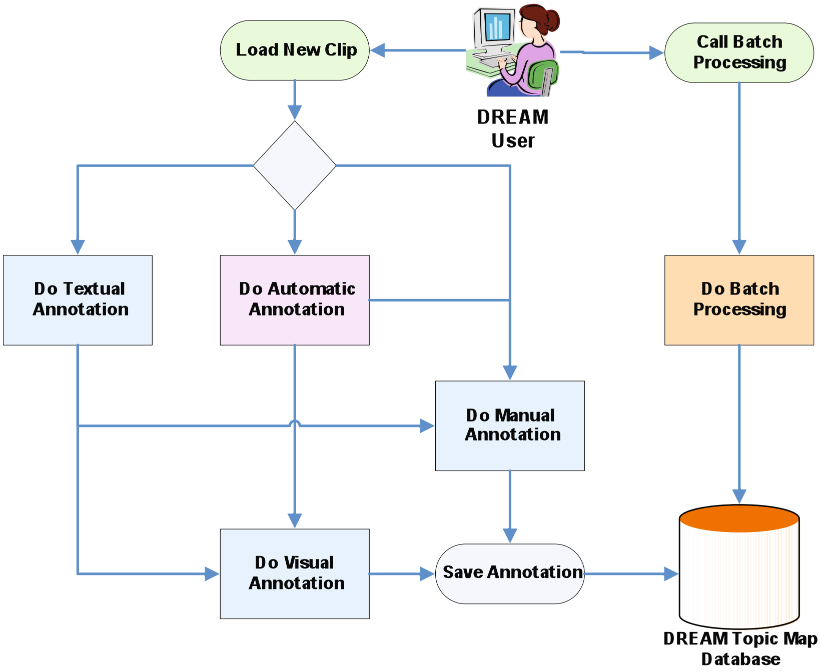


Fig. 2. Video Content Annotation Process in DREAM

For each new clip, the user decides which annotation method they wish to use. They may choose the automatic annotation process which uses the Automatic Labelling Engine to process the annotation or they may choose the textual annotation process which uses the NLP Engine to process text entered by the user. The output from the automatic annotation process or textual annotation process can further be refined by using the manual annotation and visual annotation, as illustrated in Figure 2. The framework also supports batch processing which allows automatic annotation of a range of video clips using the Automatic Labelling Module. The process of automatic annotation is further described in section 5.1 and in section 5.2 where we cover the process of semi-automatic annotation.

5.1 Automatic Video Annotation

The automatic video labelling module [Badii2009] is a fundamental component of the DREAM framework. It aims to automatically assign semantic keywords to objects of interest appearing in the video segment. The module had been implemented to label the raw video data in a fully automated manner. The typical user of the system is the video content library manager who will be enabled to use the system to facilitate the labelling and indexing of the video data. With this function, all the objects of interest including moving and still foreground objects will be labelled with linguistic keywords.

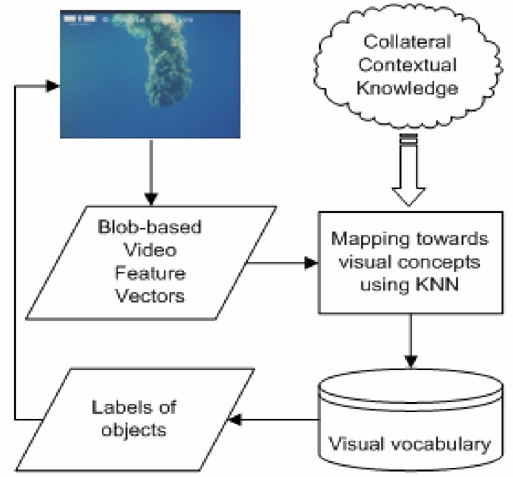

Collateral Contextual Knowledge
Blob-based Video Feature Vectors
Mapping towards visual concepts using KNN
Labels of objects
Visual vocabulary

Fig. 3. Workflow of the Automatic Labelling Module

Figure 3 shows the typical workflow of the Automatic Labelling Module. The module takes the raw video clips and the associative metadata, i.e. motion vectors and mattes, as input whereby the low-level blob-based visual features, i.e. colour, texture, shape, edge, motion activity, motion trajectory, can be extracted and encoded as feature vectors. It is those visual blobs that were compared against the visual concepts defined in the visual vocabulary of objects. These objects consist of a set of clusters of visual feature vectors of different types of special effect foreground objects such as blood, fire, explosion, smoke, water splashes, rain, clouds etc, to find the best matching visual concepts using the K-Nearest Neighbour algorithm. However, those are all traditional methodologies that would benefit from appropriate semantic enrichment by way of linkages to latent and collateral associations so as to empower the representation of higher-level context related visual concepts. Therefore, we introduced what we refer to as the collateral context knowledge, which was formalised by a probabilistic based visual keyword co-occurrence matrix, to bias (i.e. appropriately converge) the traditional matching process.

Table 1. Auto-Labelling accuracy based on the training/test ratio of 9:1, 7:3 and 5:5

Class label	**9:1**	**7:3**	**5:5**	**Class label**	**9:1**	**7:3**	**5:5**
blood and gore; blood	75%	94%	83%	misc; welding;	100%	100%	100%
blood and gore; gore;	0%	100%	100%	muzzle flash;	0%	100%	100%
bullet hits; sparks;	75%	100%	100%	sparks;	0%	25%	40%
crowds figures;	100%	100%	100%	water;bilge pumps;	100%	100%	100%
explosions fire smoke;explosion;	100%	92%	75%	water; bilge pumps;	100%	100%	80%
explosions fire smoke; fire;	100%	100%	90%	water; boat wakes;	100%	67%	50%
explosions fire smoke; fire burst;	100%	0%	100%	water; bubbles;	0%	0%	0%
explosions fire smoke; flak;	100%	100%	100%	water; cascading water;	0%	0%	50%
explosions fire smoke; sheet fire;	100%	100%	100%	water; drips;	100%	100%	50%
explosions fire smoke; smoke;	86%	90%	90%	water; interesting water surfaces;	0%	50%	67%
explosions fire smoke; steam;	100%	100%	100%	water; rain;	0%	75%	71%
falling paper;	100%	100%	100%	water; rivulets;	0%	100%	0%
Lens flares;	100%	86%	83%	water; splashes;	0%	100%	60%
Misc; car crash;	0%	0%	0%	weather; clouds;	100%	100%	100%
Misc; fecal-matter;	100%	89%	86%	weather; snow;	100%	100%	50%
Misc; washing line;	100%	100%	100%	**Average**	**66%**	**80%**	**75%**

Table 1 shows the labelling accuracy for the DREAM auto-labelling module, including both content and context related labels, based on the training-test ratio of 9:1, 7:3 and 5:5 respectively. Among the 3 different experimental setups, the training-test ratio of 7:3 achieved the superior performance with an average accuracy of 80% followed by 75% for 5:5 and 66% for 9:1. Despite poor performance for several categories, many other categories achieved a very high labelling accuracy percentage; including some at 100%.

5.2 Semi-automatic Video Annotation

In DREAM, we automatically construct Topic Maps for each new video clip. The output from the Automatic Labelling Engine is updated by the User and the resulting unstructured natural text is processed by the NLP Engine, as illustrated in Figure 4. The NLP Engine creates a structured description of the semantic content of the clips. These structured descriptions are termed as Semantic Containers and are further described in [Badii2008]. These semantic containers allow the representation of both simple sentences and complex sentences in terms of entities and actions, which are used by the DREAM Topic Map Engine to generate Topic Maps automatically. A semantic container may contain other semantic containers, enabling representation of complex sentences. This enables an entity (topic) to be linked to a group of entities (topics) [Badii2008]. As a result, complex sentences are represented by a single semantic container, with a number of "inner" semantic containers detailing the semantic information processed by the NLP Engine. The Topic Map Engine reads those containers and creates a list of topics and associations, which are merged with existing topics and associations in the DREAM Topic Map Database. The new topics are assigned occurrences which index the new clips. The semantic merging of new topics with existing topics creates a network of ontologies which eventually grows each time new clips are indexed.

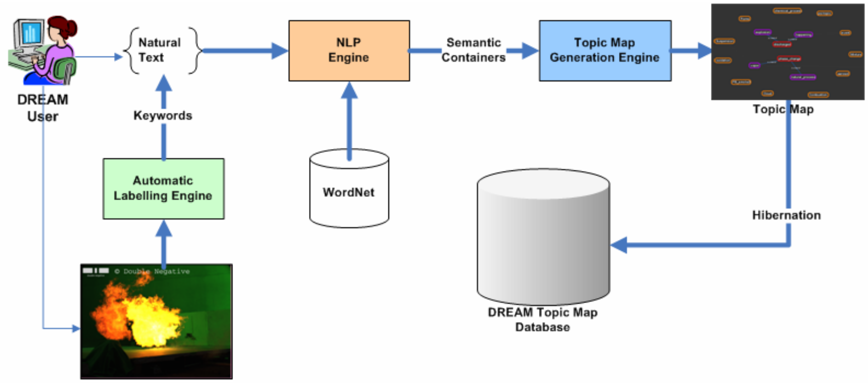


Fig. 4. Semi-automatic labelling process in the DREAM framework

The key to the evolution of the Knowledge Base, as new semantically-defined Topics are added, is the process of adding the synonyms and semantic ancestries, of the concept being added, utilising the WordNet lexical database. This defines the topic in a detailed and robust manner, enabling the linkage to the existing ontology, while ensuring that concepts are stored in the Knowledge Base by their meaning, rather than by word(s) representing the concept. This allows for seamless semantic merging of concepts, and a Knowledge Base that is well-tuned towards retrieval, and visual exploration.

6 Knowledge Retrieval

The Topic Map allows casual browsing of the knowledge base with a richly cross-linked structure over the repository content. Topic occurrences create 'sibling' relationships between repository objects and the video shots. A single resource may be the occurrence of one or more topics, each of which may have many other occurrences. When a user finds/browses to a given resource, this sibling relationship enables them to rapidly determine where the other related resources are to be found. Topic associations create 'lateral' relationships between subjects, the movie concepts – allowing a user to see which other concepts covered by the repository are related to the subject of current interest and to easily browse these concepts. Associative browsing allows an interested data consumer to wander through a repository in a directed manner. A user entering the repository via a query might also find associative browsing useful in increasing the chance of unforeseen discovery of relevant information. A DREAM Topic Map Visualisation, as shown in the Framework diagram [Figure 1] has been implemented to provide such interactive visual browsing of the DREAM Knowledge Base.

Efficient retrieval of desired media is the end goal of the DREAM framework. With the DREAM Knowledge Base built and ever-evolving with newly annotated media being merged into it, the requirement remains for interfaces to be able to achieve this goal. In DREAM, two such interfaces were developed, these being a Visualisation for exploring the Knowledge Base itself, through which media can be retrieved, and a Query Interface that allows directed querying of the Knowledge Base.

6.1 Retrieval Visualisation

The Retrieval Visualisation utilises the DREAM Visualisation Engine to create an interface enabling the user to explore the Knowledge base for concepts, and retrieving the Occurrences (Media) attached to them. This operation can range from being rather simple, to allowing for more complicated searching, so that dynamic visual semantic searches become a reality.

The initial entry-point into a search is a simple text search through the Knowledge Base, returning all concepts that match it. This includes searching through the synonyms of each concept, such that any Topic that contains the search string in its synonyms will also be returned, albeit with that specifically matched synonym shown as its display name, for convenience. With the simple search results presented, the user can then choose the path down which they would wish to explore the map further, by clicking on the appropriate Topic that represents the meaning of the word for which they were searching. This process disregards all other search results, and sets the focus upon the selected Topic, whereupon all related Topics are brought into the focus of the visualisation, thus connected to the Topic itself, with the association descriptions labelled on the edge between the nodes.

Colour-coding is used to visually distinguish the different types of association and Topic relevance. Topics with attached Occurrences are coloured orange, whereas empty Topics are coloured Yellow. Currently Selected Topics are enlarged, and coloured Red. Additionally, associations related to Topic definition (typically this 'is a kind of' relationship) are coloured Blue, whereas associations representing a tagged relationship between two Topics (such as "a car crashing into a wall") are coloured Green.

The criteria for the data to be visualised can be modified using a variety of filters, which can be deployed by the user to further direct their search. The first of these filters is a "distance filter", that allows the user to specify the distance from the selected Topic for as far as can be appropriated within the scope of the visualisation of nodes. For example, with a distance of 1, only the immediately associated Topics are shown, whereas, if the distance were 2, all Topics that are within two 'hops', "degrees of separation" from the selected Topic are visualised. This is useful to see an overview of the associations of the Topic as it is situated within the ontological structure, as well as to reduce on-screen Topic clutter, if the selected Topic is one that has a large number of associated Topics. Other filters let the user specify which types of associations they are interested in seeing in the visualisation, depending on the type of search that they may be conducting. For example, showing the homonyms of a Topic would just serve to provide unnecessary clutter, unless the user believes that they may be exploring the wrong route through the semantic associations of a word, thus wishing to have all possible associations depicted on the screen together. Enabling homonym associations then gives the user instant access to jump to a completely different area of the Knowledge Base, with ease, and in a visually relevant manner.

When selecting a Topic, while browsing through the Topic Map, the Occurrences attached to the selected Topic are presented to the user, (as can be seen in Figure 5), in a tree structure mapping out the properties of the selected Topic(s). The Occurrences are represented here by the clip name, which when clicked displays the key frame for that Clip. The key frames were originally used in the Automatic Labelling process, but also provide an easy visual trigger for user in browsing through.

Additional retrieval techniques can be utilised through this interface, by selecting multiple Topics, with only the common occurrences between selected Topics being returned. This allows for a much finer query into the Knowledge Base, to retrieve very specifically tagged clips. The hierarchical semantic structure of the Knowledge Base is also used to aid the retrieval, by exploring the children of a selected Topic as well, for occurrences. For example, if there was a clip featuring a

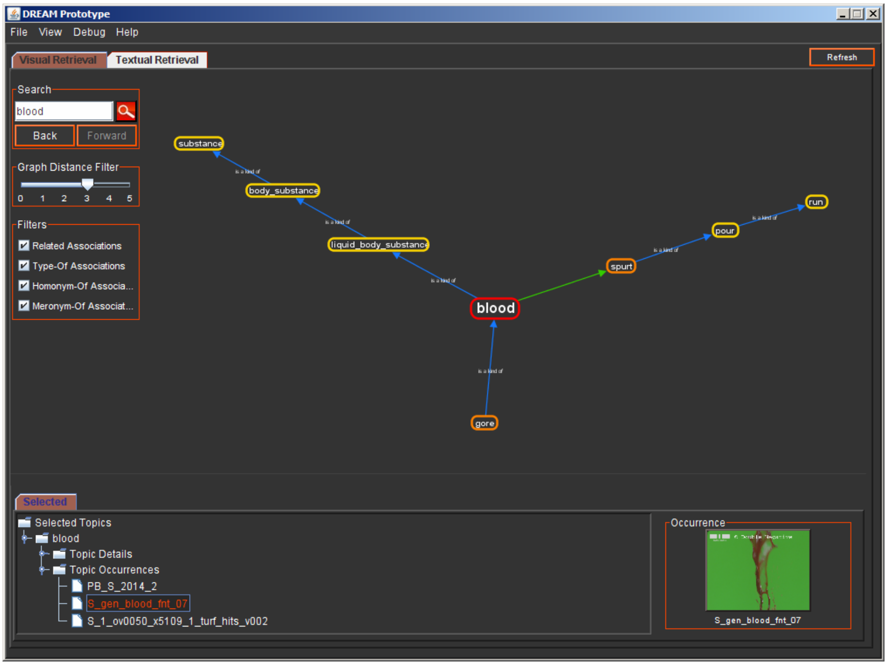


Fig. 5. Retrieval Visualisation Interface

"car" crashing into a "wall", the user could select the Topics "motor vehicle" and "wall", and the clip would be returned, as "car" is a child of "motor vehicle".

7 User Evaluation of the DREAM Framework

The evaluation of the performance of the DREAM System is a critical task, even more so when the annotations of the video clips are the output of a semi-automatic labelling framework, rather than the result of a concept sorting effort by a team of film post-production domain specialists. The user partner Double Negative examined 145 clips and for each clip the tester at Double Negative gave a score of relevancy from 1 to 5 for each tag as automatically generated by DREAM. Additionally the users commented on the various aspects of the functionality of DREAM and appeared to be able to use its features with ease after some initial training. The usability of the system was ranked as fairly high overall. The users found the visualisation engine with personalise-able colour coding of topics as particularly helpful in navigating the topics of interest from a viewpoint-specific basis. This feature enabled them to avoid cognitive overload when having to consider multi-faceted selection criteria.

Table 2. User Evaluation Results

Category Names	**Number of Samples**	**Avg. User Score (Low 1 – 5 High)**
Blood and gore; blood;	12	5
Blood and gore; gore;	4	1
Bullet hits; sparks;	4	3.3
crowds figures;	3	3
explosions fire smoke; explosion;	50	4.7
explosions fire smoke; fire;	16	3.6
explosions fire smoke; fire burst;	2	5
explosions fire smoke; smoke;	7	4.4
explosions fire smoke; steam;	5	5
lens flares;	7	4.2
misc; car crash;	1	5
misc; poo;	12	5
misc; washing line;	5	5
misc; welding;	1	1
muzzle flash;	4	1
sparks;	2	1
water; bilge pumps;	2	5
water; boat wakes;	6	5
water; cascading water;	3	5
water; drips;	3	4
water; interesting water surfaces;	5	5
water; rain;	6	5
water; rivulets;	1	5
water; splashes;	3	4.8
water; spray;	3	5
weather; clouds;	4	5
weather; snow;	5	5

Double Negative evaluated the results of processing a sub-set of their library through the DREAM system with the film editing staff (i.e. practitioner users) ranking the accuracy of the Topic Map generated for each clip. As Table 2 shows, the scores given by the practitioner users were generally very high, but with consistent low scores in certain categories, giving an average overall score of 4.4/5. These scores, along with other evaluation results, indicate the success of the system, but also highlight the areas where performance could be improved. For example, the system had difficulty in identifying clips with very brief exposure of a single indicative feature that constituted the main essence of the clip category for example as in video clips of sparks and flashes whereby the feature to be identified (i.e. the spark or flash) is shown very briefly in the clip.

8 Conclusion

In this chapter, we have presented the DREAM Framework and discussed how it semi-automatically indexes video clips and creates a network of semantic labels, exploiting the Topic Map Technology to enable efficient retrieval. The framework architecture, which has been presented in the context of film post-production, as a challenging proving ground, has responded well to the high expectations of users in this application domain which demands efficient semantic-cooperative retrieval. We have described how the DREAM framework has also leveraged the advances in NLP to perform the automatic creation and population of Topic Maps within a self-evolving semantic network for any media repositories, by defining the topics (concepts) and relationships between the topics. We also briefly discussed how this framework architecture handles collaborative labelling through its Automatic Labelling Engine. The first DREAM prototype has already been implemented and evaluated by its practitioner users in the film post-production application domain. The results confirm that the DREAM architecture and implementation has proven to be successful. Double Negative have the DREAM system trained with only 400 video clips and deployed within their routine film (post)production processes to achieve a satisfactory performance in labelling and retrieving clips from a repository holding a collection of several thousand clips.

The DREAM paradigm can in future be extended to further domains, including the Web, where any digital media can go through an offline process of (semi)automatic-labelling before being published. The publishing agent will use the semantic labels to create a network of connected concepts, using Topic Maps, and this can be merged with existing concepts on the web space. This will eventually enable more intelligent web interaction, information filtering and retrieval, using semantic concepts as the query parameters rather than keywords. The DREAM Project Team is currently engaged in extending the functionality of the prototype to increase its scalability and ensure a wider uptake across a whole range of application domains particularly in supporting collaborative creative processes in the film, media publishing, advertising, training and educational sectors through our test partners worldwide.

References

[Ahmed2000] Ahmed, K.: Topic maps for Repositories, http://www.gca.org/papers/xmleurope2000/papers/s29-04.html (last accessed January 2010)

[Athanasisadis2007] Athanasisadis, T., Mylonas, P., Avrithis, Y., Kollias, S.: Semantic image segmentation and object labelling. IEEE Trans. On Circuits and systems for Video Technology 17(3), 298–312 (2007)

[Badii2008] Badii, A., Lallah, C., Kolomiyets, O., Zhu, M., Crouch, M.: Semi-Automatic Annotation and Retrieval of Visual Content Using the Topic Map Technology. In: Proc of 1st Int. Conf. on Visualization, Imaging and Simulation, Romania, pp. 77–82 (November 2008)

[Badii20009] Badii, A., Zhu, M., Lallah, C., Crouch, M.: Semantic-driven Context-aware Visual Information Indexing and Retrieval: Applied in the Film Post-production Domain. In: Proc. IEEE Workshop on Computational Intelligence for Visual Intelligence 2009, US (March 2009)

[Bradshaw2000] Bradshaw, B.: Semantic based image retrieval: a prob-abilistic approach. In: Proc. of the Eighth ACM Int. Conf. on Multimedia, pp. 167–176 (2000)

[Brown1993] Brown, P., Pietra, S.D., Pietra, V.D., Mercer, R.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19(2), 263–311 (1993)

[Cascia1998] Cascia, M.L., Sethi, S., Sclaroff, S.: Combining Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web. In: Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries (1998)

[Duygulu2002] Duygulu, P., Barnard, K., Freitas, N., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)

[garshol2002] Garshol, L.: What are Topic Maps, `http://xml.com/pub/a/2002/09/11/topicmaps.html?page=1` (last accessed January 2010)

[Gorkani1994] Gorkani, M.M., Picard, R.W.: Texture orientation for sorting photos 'at a glance'. In: Proc. of the IEEE Int. Conf. on Pattern Recognition (October 1994)

[Haralick1973] Haralick, R.M., Shanmugam, K., Dinstein, I.: Texture features for image classification. IEEE Trans. On Sys. Man and Cyb. SMC-3(6), 610–621 (1973)

[ISO2000] ISO/IEC 13250:2000 Document description and processing languages – Topic Maps. International Organisation for Standardization ISO, Geneva (2000)

[Maillot2004] Maillot, N., Thonnat, M., Boucher, A.: Towards ontology based cognitive vision. Mach. Vis. Appl. 16(1), 33–40 (2004)

[Marques2002] Marques, O., Furht, B.: Content-Based Image and Video Retrieval. Kluwer Academic Publishers, Norwell (2002)

[Morris2004] Morris, T.: Computer Vision and Image Processing. Palgrave Macmillan Publishers, Ltd., New York (2004)

[Mori1999] Mori, Y., Takahashi, H.: Oka, R.: Image-to-word trans-formation based on dividing and vector quantizing images with words. In: MISRM 1999 First Int. Workshop on Multimedia Intelligent Storage and Retrieval Management (1999)

[Paek1999] Paek, S., Sable, C.L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B.H., Chang, S.F., McKeown, K.R.: Integration of visual and text based approaches for the content labelling and classification of Photographs. In: ACM SIGIR 1999 Workshop on Multimedia Indexing and Retrieval, Berkeley, CA, August 19 (1999)

[Pepp2009] Pepper, S.: The TAO of Topic Maps: finding the way in the age of infoglut, `http://www.ontopia.net/topicmaps/materials/tao.html` (last accessed January 2010)

[Ruil1999] Rui, Y., Huang, T.S., Chang, S.F.: Image Retrieval: current techniques, promising directions and open issues. Journal of Visual Communication and Image Representation (1999)

[Schober2004] Schober, J.P., Hermes, T., Herzog, O.: Content-based image retrieval by ontology-based object recognition. In: Proc. KI 2004 Workshop Appl. Descript. Logics (ADL 2004), Ulm, Germany, pp. 61–67 (September 2004)

[Sme2000] Smeulder, A.W.M., Worring, M., Anntini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12) (December 2000)

[Szummer1998] Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: IEEE Int. Workshop on Content-based Access of Image and Video Databases (1998)

[Wang2003] Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modelling approach. IEEE Trans. Pattern Analysis and Machine Intelligence 25(9), 10751088 (2003)

[Westerveld00] Westerveld, T.: Image Retrieval: Content Versus Context. In: Proceedings of Content-Based Multimedia Information Access, pp. 276–284 (2000)

[Zhou2000] Zhou, X.S., Huang, S.T.: Image Retrieval: Feature Primitives, Feature Representation, and Relevance Feedback. In: IEEE Workshop on Content-based Access of Image and Video Libraries (2000)

A Human-Centered Computing Framework to Enable Personalized News Video Recommendation

Hangzai Luo[1] and Jianping Fan[2]

[1] Software Engineering Institute, East China Normal University, Shanghai, China
hzluo@sei.ecnu.edu.cn
[2] Dept. of Computer Science, UNC-Charlott, NC 28223, USA
jfan@uncc.ed

Abstract. In this chapter, an interactive framework is developed to enable personalized news video recommendation and allow news seekers to access large-scale news videos more effectively. First, multiple information sources (audio, video and closed captions) are seamlessly integrated and synchronized to achieve more reliable news topic detection, and the inter-topic contextual relationships are extracted automatically for characterizing the interestingness of the news topics more effectively. Second, topic network (i.e., news topics and their inter-topic contextual relationships) and hyperbolic visualization are seamlessly integrated to achieve more effective navigation and exploration of large-scale news videos at the topic level, so that news seekers can have a good global overview of large-scale collections of news videos at the first glance. Through a hyperbolic approach for interactive topic network visualization and navigation, large amounts of news topics and their contextual relationships are visible on the display screen, and thus news seekers can obtain the *news topics of interest* interactively, build up their mental search models easily and make better search decisions by selecting the visible news topics directly. Our system can also capture the search intentions of news seekers implicitly and further recommend the most relevant news videos according to their importance and representativeness scores. Our experiments on large-scale news videos (10 TV news programs for more than 3 months) have provided very positive results.

1 Introduction

According to the CIA world factbook, there are more than 30,000 television stations in the world. These stations broadcast a large number of TV news programs (news videos) every day. Different organizations and individuals utilize these broadcast news videos for different purposes, such as presidential candidates' debat for public assessment, economic performance analysis and prediction, sports and crime reports. People watch the news videos (TV news programs) to understand what is happening now and predict what might happen in the near future, so that they can make better daily decisions.

Due to the large number of broadcast channels and TV news programs, finding news videos of interest is not a trivial task: (a) Most existing content-based video retrieval (CBVR) systems assume that news seekers can formulate their information

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 475–495.
springerlink.com

needs precisely either in terms of keywords or example videos. Unfortunately, news seekers may not be able to know what is happening now (i.e., if they know it, it is not a news), thus it is very hard for them to find the suitable keywords or example videos to formulate their news needs precisely without obtaining sufficient knowledge of the available news topics of interest. Thus there is an urgent need to develop new techniques for detecting news topics of interest from large-scale news videos to assist news seekers on finding news videos of interest more effectively. (b) Because the same news topic can be discussed in many TV channels and news programs, topic-based news search may return large amounts of news videos and thus simple news search via keyword matching of news topics may bring the serious problem of information overload to news seekers. (c) Most existing CBVR systems treat all the news seekers equally while completely ignoring the diversity and rapid change of their search interests. Besides the rapid growth of broadcast TV channels and news programs, we have also observed different scenarios of news needs from different people, thus it is very difficult to come up with a *one size fits all* approach for accessing large-scale news videos. (d) The keywords for news topic interpretation may not be expressive enough for describing the rich details of video content precisely and using only the keywords may not be able to capture the search intentions of news seekers effectively. Thus visualization is becoming a critical component of personalized news video recommendation system [1-2, 9-12]. (e) The objectives for personalized video recommendation and content-based video retrieval are very different, which make it unsuitable to directly apply the existing CBVR techniques for supporting personalized video recommendation. Thus supporting personalized news video recommendation is becoming one important feature of news services [3-4].

There are some existing approaches to support personalized video recommendation by using only the associated text terms such as the titles, tags, and comments [3-4], and the relevant videos are recommended according to the matching between the associated text terms for video content description and the users' profiles. Unfortunately, the text terms, which are associated with the videos, may not have exact correspondence with the underlying video content. In addition, a sufficient collection of users' profiles may not be available for recommendation purpose. Thus there is an urgent need to develop new frameworks for supporting personalized news video recommendation, which may not completely depend on the users' profiles and the associated texts for video content description.

Context between the news topics is also very important for people to make better search decisions, especially when they are not familiar with the available news topics and their search goals or ideas are still fuzzy. The inter-topic context can give a good approximation of the interestingness of the news topics (i.e., like PageRank for characterizing the importance of web pages [17]). Thus it is very attractive to integrate topic network (i.e., news topics and their inter-topic contextual relationships) for characterizing the interestingness of the news topics, assisting news seekers on making better search decisions and suggesting the future search directions.

To incorporate topic network for supporting user-adaptive topic recommendation, it is very important to develop new algorithm for large-scale topic network visualization, which is able to provide a good balance between the local detail and the global context. The local detail is used to help news seekers focus on the news topics of interest in current focus. The global context is needed to tell news seekers where the other news topics are and their contextual relationships with the news topics in current focus, such global context can effectively suggest the new search directions to news seekers. Thus supporting visualization and interactive navigation of the topic network is becoming a complementary and necessary component for personalized news video recommendation system and it may lead to the discovery of unexpected news videos and guide the future search directions effectively.

On the other hand, the search criteria are often poorly defined or depend on the personal preferences of news seekers. Thus supporting interactive visualization, exploration and assessment of the search results are very important for allowing news seekers to find the news videos of interest according to their personal preferences. Information retrieval community has also recognized that designing more intuitive system interface for search result display may have significant effects on assisting users to understand and assess the search results more effectively [13]. To incorporate visualization for improving news search, effective techniques for intelligemt news video analysis should be developed to discover the meaningful knowledge from large-scale news videos.

Several researchers have used the ontology (i.e., video concepts and their simple inter-topic contextual relationships such as “IS-A" and “part-of") to assistant visual content anslysis and retrieval [23-24]. Because the news content are highly dynamic, the inter-topic contextual relationships cannot simply be characterized by using “IS-A" or “part-of", which are used for ontology construction. Thus it is unacceptable to incorporate the ontology for supporting personalized news video recommendation. On the other hand, automatic video understanding is still an open problem for computer vision community [25-31].

In this chapter, an interactive approach is developed to enable personalized news video recommendation, and our approach has significant differences from other existing work: (a) Rather than performing semantic video classification for automatic news video understanding, we have integrated multiple information sources to achieve more reliable news topic detection. (b) The associations among the news topics (i.e., inter-topic contextual relationships) are determined automatically and an interestingness score is automatically assigned to each news topic via statistical analysis, and such interestingness scores are further used to select the news topics of interest and filter out the less interesting news topics automatically. (c) A hyperbolic visualization tool is incorporated to inform news seekers with a better global overview of large-scale news videos, so that they can make better search decisions and find the most relevant news videos more effectively. (d) A novel video ranking algorithm is developed for recommending the most relevant news videos according to their importance and representativeness scores.

The chapter is organized as follows. Section 2 briefly reviews some related work on news topic detection and personalized information recommendation; Section 3 introduces our work on integrating topic network and hyperbolic visualization to enable user-adaptive topic recommendation; Section 4 introduces our new scheme on news video ranking for supporting personalized news video recommendation; Section 5 summarizes our work on algorithm and system evaluation; We conclude in Section 6.

2 Related Work

To enable personalized news video recommendation, one of the most important problems is to extract *news topics of interest* automatically from large-scale news videos. This problem is becoming very critical because of the following reasons: (a) The amount of news topics could be very large; (b) Different news topics may have different importance and interestingness scores, such importance and interestingness scores may also depend on the personal preferences of news seekers. In this section, we have provided a brief review of some existing work which are critical for developing personalized news video recommendation system: (1) automatic news topic detection; (2) news visualization; (3) personalized video recommendation.

Topic extraction refers to the identification of individual stories or topics within a broadcast news video by detecting the boundaries where the topic of discussion changes. News topics may be of any length and consist of complete and cohesive news report on one particular topic. Each broadcast channel has its own peculiarities in terms of program structures and styles, which can be integrated for achieving more accurate detection of news topics and their boundaries [11-12]. News topics can also be detected by using some existing techniques for named-entity extraction [5-6].

There are two well-accepted approaches for supporting personalized information retrieval [20-22]: *content-based filtering* and *collaborative filtering*. Because the profiles for new users are not available, both the collaborative filtering approach and the content-based filtering approach cannot support new users effectively. Thus there is an urgent need to develop more effective frameworks for supporting personalized news video recommendation.

Visualization is widely used to help the users explore large amount of information and find interesting parts interactively [9-12]. Rather than recommending the most interesting news topics to news seekers, all of these existing news visualization systems disclose all the available news topics to them, and thus news seekers have to dig out the news topics of interest by themselves. When large-scale news collections come into view, the number of the available news topics could be very large and displaying all of them to news seekers may mislead them. Thus it is very important to develop new algorithms for characterizing the interestingness of news topics and reducing the number of news topics to enable more effective visualization and exploration of large-scale news videos.

3 User-Adaptive News Topic Recommendation

In this chapter, a novel scheme is developed by incorporating *topic network* and hyperbolic visualization to recommend the *news topics of interest* for assisting news seekers on accessing large-scale news videos more effectively. To do this, an automatic scheme is developed to construct the topic network for representing and interpreting large-scale news videos at the topic level. In addition, a hyperbolic visualization technique is developed to enable interactive topic network navigation and recommend the news topics of interest according to the personal preferences and timely observations of news seekers, so that they can make better search decisions.

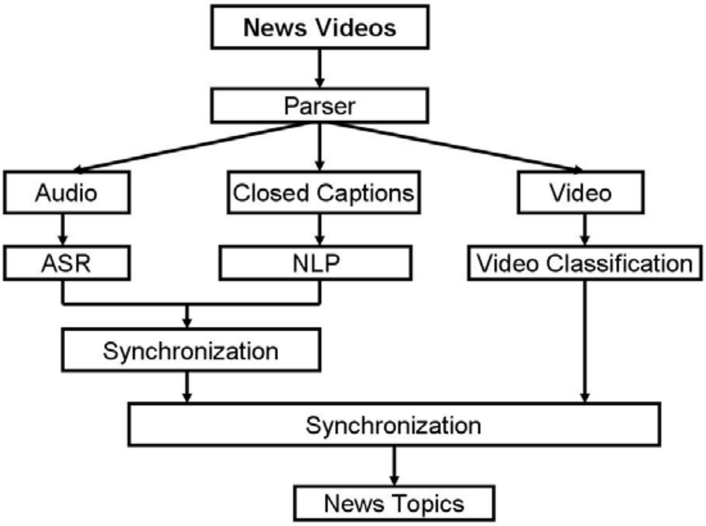


Fig. 1. The flowchart for synchronizing multiple sources for news topic detection, where automatic speech recognition (ASR), natural language processing (NLP), and semantic video classification are seamlessly integrated

3.1 News Topic Detection

For TV news programs, there are three major information sources (audio, video and closed captions) that can be integrated and synchronized to enable more reliable news topic detection. We have developed a new scheme for automatic news topic detection by taking the advantage of multiple information sources (cross-media) as shown in Fig. 1. First, automatic speech recognition (ASR), natural language processing (NLP), and semantic video classification are performed on these three information sources parallelly to determime the keywords for news topic description from both the audio channel and the closed captions and detect the video concepts from the video channel. Second, the audio channel is synchronized with the closed caption channel, and the video channel is further synchronized with the audio channel and the closed caption channel. Finally, the detection results of news topics from these three information sources are integrated to boost the performance of our news topic detection algorithm.

The closed captions of news videos can provide abundant information and such information can be used to detect the news topics of interest and their semantic interpretations with high accuracy. To do this, the closed captions are first segmented into a set of sentences, and each sentence is further segmented into a set of keywords. In news videos, some special text sentences, such as "*somebody*, CNN, *somewhere*" and "ABC's *somebody* reports from *somewhere*", need to be processed separately. The names for news reporters in those text sentences are generally not the content of news report. Therefore, they are not appropriate for news semantics interpretation and should be removed. Because there have some clear and fixed patterns for these specific sentences, we have designed a context-free syntax parser to detect and mark this information. By incorporating 10-15 syntax rules, our parser can detect and mark such specific sentences in high accuracy. Standard text processing techniques are used to remove the stop words automatically.

Most named entity detectors may fail in processing all-capital strings because initial capitalization is very important to achieve accurate named entity recognition. One way to resolve this problem is to train a detector with ground truth from the text documents of closed captions. However, it's very expensive to obtain the manually marked text material. Because English has relatively strict grammar, it's possible to parse the sentences and recover the most capital information by using part-of-speech (POS) and lemma information. TreeTagger [7] is used to perform the part-of-speech tagging. Capital information can be recovered automatically by using the TreeTagger parsing results.

After such specific sentences are marked and the capital information is recovered, an open source text analysis package LingPipe [8] is used to perform the named entity detection and resolve co-reference of the named entities. The named entities referring to the same entity are normalized to the most representative format to enable statistical analysis, where the news model of LingPipe is used and all the parameters are set to default value. Finally, the normalized results are parsed again by TreeTagger to extract the POS information and resolve the words to their original formats. For example, TreeTagger can resolve "better" to "well" or "good" according to its POS tag.

We have defined a set of over 4000 elemental topics, each keyword represents an elemental topic, and all these detected news stories that consist of one particular

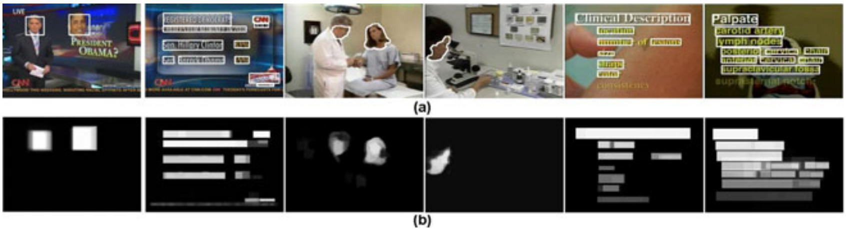


Fig. 2. Integrating confidence map for salient object detection: (a) original images and the detected salient objects; (b) confidence maps for the salient objects

keyword are assigned to the corresponding cluster of news topic. Our multi-task learning algorithm is performed to learn the topic detectors from a given corpus by exploiting the inter-topic correlation [25-27]. Once we have a set of topic detectors, they are used to determine the most topic-similar clusters for the new piece of news videos.

For TV news videos, the video shots are the basic units for video content representation, and thus they can be treated as one of the semantic items for news topic detection. Unlike the keywords in text documents, the re-appearance of video shots cannot be detected automatically via simple comparison of their visual properties. For news videos, video objects, such as text areas and human faces, may provide important clues about news stories of interest. Text lines and human faces in news videos can be detected automatically by using suitable computer vision techniques [28]. Obviously, these automatic detection functions may fail in some cases. Thus the results that are detected by using a single video frame may not be reliable. To address this problem, the detection results on all the video frames within the same video shot are integrated and the corresponding confidence maps for the detection results are calculated as shown in Fig. 2 [27]. The video concepts associated with the video shots can provide valuable information to enable more accurate news topic detection, and semantic video classification is one potential solution to detect such video concepts [27]. To detect the video concepts automatically, we have adopted our previous work reported in [25-28].

Unfortunately, the closed captions may not synchronize with the video channel accurately and have a delay of a few seconds in general. Thus the news topics that are detected from the closed captions cannot directly be synchronized with the video concepts that are detected from the news videos. On the other hand, the closed captions have good synchronization with the relevant audios. Therefore, they can be integrated to take advantage of cross-media to clarify the video content and remove the redundant information. Even the audio channel generally synchronizes very well with the video channel, the accuracy of most existing techniques for automatic speech recognition (ASR) is still low. By integrating the results for automatic speech recognition with the topic detection results from the closed captions, we can synchronize the closed captions with the video content in higher accuracy. After the closed captions are synchronized with the news videos, we can assign the video shots to the most relevant news topics that are accurately detected from the closed captions. Thus all the video shots, which locate between the start time and the end time of a given new topic that has been detected from the closed captions, are assigned to the given news topic automatically.

3.2 Topic Association Extraction

The contextual relationships among these significant news topics are obtained automatically, where both the semantic similarity and the co-occurrence probability for the relevant news topics are used to define a new measurement for determining the inter-topic associations effectively. The inter-topic association (i.e., inter-topic contextual relationship) $\phi(C_i, C_j)$ is determined by:

$$\phi(C_i, C_j) = -\alpha \cdot \log \frac{d(C_i, C_j)}{2L} + \beta \cdot \frac{\psi(C_i, C_j)}{\log \psi(C_i, C_j)}, \quad \alpha + \beta = 1 \tag{1}$$

where the first part denotes the semantic similarity between the news topics C_j and C_i, the second part indicates their co-occurrence probability, α and β are the weighting parameters, $d(C_i, C_j)$ is the length of the shortest path between the news topics C_i and C_j by searching the relevant keywords for news topic interpretation from WordNet [23], L is the maximum depth of WordNet, $\psi(C_i, C_j)$ is the co-ocurrence probability between the relevant news topics. The co-occurrence probability $\psi(C_i, C_j)$, between two news topics C_j and C_i, is obtained in the news topic detection process. Obviously, the value of the inter-topic association $\phi(C_i, C_j)$ increases with the strength of the contextual relationship between the news topics C_i and C_j.

Thus each news topic is automatically linked with multiple relevant news topics with the higher values of the associations $\phi(\cdot, \cdot)$. One portion of our large-scale topic network is given in Fig. 3, where the news topics are connected and organized according to the strength of their associations, $\phi(\cdot, \cdot)$. One can observe that such a topic network can provide a good global overview of large-scale news videos and can precisely characterize the interestingness of the relevant news topics, and thus it can be used to assist news seekers on making better search decisions.

To integrate the topic network for supporting user-adaptive topic recommendation, it is very attractive to achieve graphical representation and visualization of the topic network, so that news seekers can obtain a good global overview of large-scale news videos at the first glance and make better search decisions in the process of interactive topic network exploration and navigation. Unfortunately, visualizing large-scale topic network in a 2D system interface with a limited screen size is not a trivial task. To achieve more effective visualization of large-scale topic network, we have developed multiple innovative techniques: (a) highlighting the news topics according to their interestingness scores for allowing news seekers to obtain the most

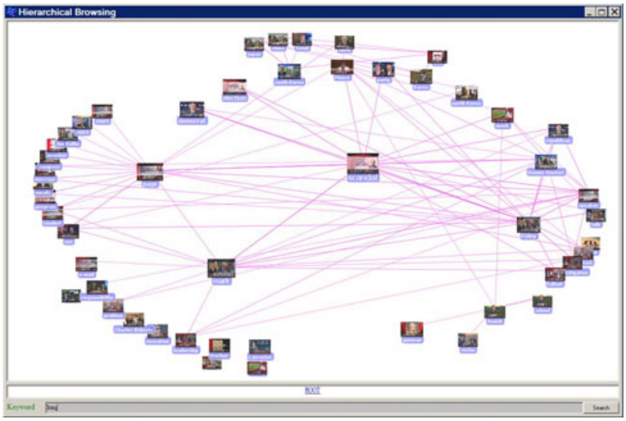


Fig. 3. One portion of our topic network for organizing large-scale news videos

important insights at the first glance; (b) integrating hyperbolic geometry to create more space for large-scale topic network visualization and exploration.

3.3 Interestingness Scores of News Topics

We have integrated both the popularity of the news topics and the importance of the news topics to determine their interestingness scores. The popularity of a given news topic is related to the number of TV channels or news programs which have discussed or reported the given news topic. If one news topic is discussed or reported by more TV channels or news programs, it tends to be more interesting. The importance of a given news topic is also related to its linkage structure with other news topics on the topic network. If one news topic is related to more news topics on the topic network, it tends to be more interesting [17]. For example, the news topic for “roadside bond in Iraq" may relate to the news topics of “gap price increase" and “stock decrease". Thus the interestingness score $\rho(C_i)$ for a given news topic C_i is defined as:

$$\rho(C) = \lambda \cdot \log(m(C_i) + \sqrt{m^2(C_i) + 1}) + \gamma \cdot \log(k(C_i) + \sqrt{k^2(C_i) + 1}), \lambda + \gamma = 1 \tag{2}$$

where $m(c_i)$ is the number of TV channels or news programs which have discussed or reported the given news topic C_i, $k(c_i)$ is the number of news topics linked with the given news topic C_i on the topic network. Thus the interestingness score for a given news topic increases adaptively with both the number of the relevant TV channels or news programs and the number of the linked news topics. Such interestingness scores can be used to highlight the most interesting news topics and eliminate the less interesting news topics for reducing the visual complexity for large-scale topic network visualization and exploration.

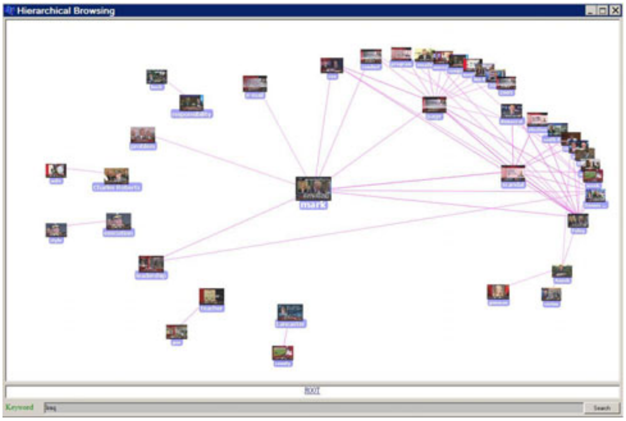


Fig. 4. One view of hyperbolic visualization of our topic network

3.4 Hyperbolic Topic Network Visualization

Supporting graphical representation and visualization of the topic network can provide an effective solution for exploring large-scale news videos at the topic level and recommend the *news topics of interest* interactively for assisting news seekers to make better search decisions. However, visualizing large-scale topic network in a 2D system interface with a limited screen size is a challenging task. We have investigated multiple solutions to tackle this challenge task: (a) A string-based approach is incorporated to visualize the topic network with a nested view, where each news topic node is displayed closely with the most relevant news topic nodes according to the values of their associations. The underlying inter-topic contextual relationships are represented as the linkage strings. (b) The geometric closeness of the news topic nodes is related to the strength of their inter-topic contextual relationships, so that such graphical representation of the topic network can reveal a great deal about how these news topics are connected. (c) Both geometric zooming and semantic zooming are integrated to adjust the levels of visible details automatically according to the discerning constraint on the number of news topic nodes that can be displayed per view.

Our approach for topic network visualization exploits hyperbolic geometry [14-16]. The hyperbolic geometry is particularly well suited for achieving graph-based layout of the topic network, and it has “more space" than Euclidean geometry. The essence of our approach is to project the topic network onto a hyperbolic plane according to the inter-topic contextual relationships, and layout the topic network by mapping the relevant news topic nodes onto a circular display region. Thus our topic network visualization scheme takes the following steps: (a) The news topic nodes on the topic network are projected onto a hyperbolic plane according to their inter-topic contextual relationships, and such projection can usually preserve the original contextual relationships between the news topic nodes. (b) After such context-preserving projection of the news topic nodes is obtained, Poincaré disk model [14-16] is used to map the news topic nodes on the hyperbolic plane to a

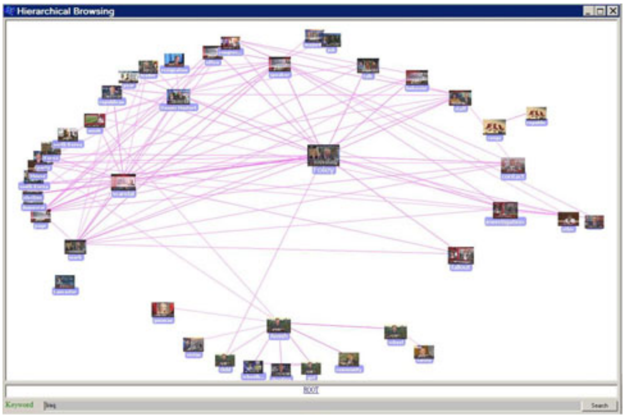


Fig. 5. Another view of hyperbolic visualization of our topic network

2D display coordinate. Poincaré disk model maps the entire hyperbolic space onto an open unit circle, and produces a non-uniform mapping of the news topic nodes to the 2D display coordinate.

Our approach for topic network visualization relies on the representation of the hyperbolic plane, rigid transformations of the hyperbolic plane and mappings of the news topic nodes from the hyperbolic plane to the unit disk. Internally, each news topic node on the graph is assigned a location $z = (x, y)$ within the unit disk, which represents its Poincaré coordinates. By treating the location of the news topic node as a complex number, we can define such a mapping as the linear fractional transformation [14-16]:

$$z_t = \frac{\theta z + P}{1 + \bar{P}\theta z} \tag{4}$$

where P and θ are the complex numbers, $|P| < 1$ and $|\theta| = 1$, and $\bar{P}$ is the complex conjugate of P. This transformation indicates a rotation by θ around the origin following by moving the origin to P (and $-P$ to the origin).

To incorporate such transformation for topic network visualization, the layout routine is structured as a recursion that takes a news topic node and a wedge in which to lay out the news topic node and its relevant news topic nodes. It places the news topic node at the vertex of the wedge, computes a wedge for each relevant news topic node and recursively calls itself on each relevant news topic node. The relevant news topic nodes are placed in the middle of their subwedges at a distance computed by the formula:

$$d = \sqrt{\left(\frac{(1-s^2)sin(a)}{2s}\right)^2 + 1} - \frac{(1-s^2)sin(a)}{2s} \tag{5}$$

where a is the angle between the midline and the edge of the subwedge and s is the desired distance between a relevant news topic node and the edge of its subwedge. In our current implementation, we set $s = 0.18$. The result, d, is the necessary distance from current news topic node to its relevant news topic node. If the value of d is less than that of s, we set d to s for maintaining a minimum space between the current news topic node and the relevant news topic node. Both s and d are represented as the hyperbolic tangent of the distance in the hyperbolic plane.

3.5 Personalized Topic Network Generation

After the hyperbolic visualization of the topic network is available, it can be used to enable interactive exploration and navigation of large-scale news videos at the topic level via *change of focus*. The *change of focus* is implemented by changing the mapping of the news topic nodes from the hyperbolic plane to the unit disk for display, and the positions of the news topic nodes in the hyperbolic plane need not to be altered during the focus manipulation. As shown in Fig. 4 and Fig. 5, news seekers can change their focuses of the news topics by clicking on any visible news topic node to bring it into the focus at the screen center, or by dragging any visible news topic node interactively to any other screen location without losing the

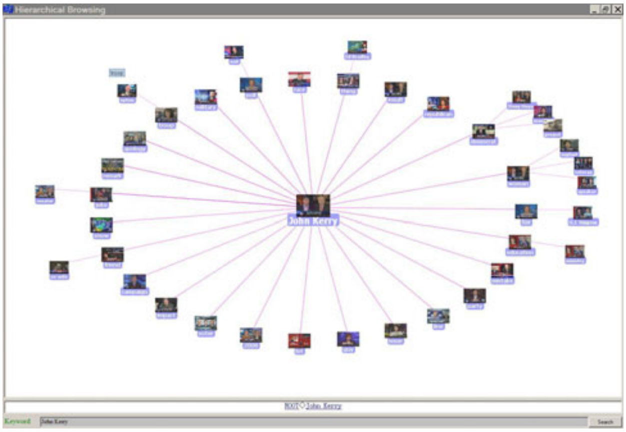


Fig. 6. The most relevant news topics for interestingness propagation

contextual relationships between the news topics, where the rest of the layout of the topic network transforms appropriately. In such interactive topic network navigation and exploration process, news seekers can obtain the *news topics of interest* interactively, build up their mental search models easily and make better search decision effectively by selecting the visible news topics directly. Because our hyperbolic visualization framework can assign more spaces for the news topic node in current focus and ignore the details for the residual news topic nodes on the topic network, it can theoretically avoid the overlapping problem by supporting change of focus and thus it can spporting large-scale topic network visualization and navigation.

On the other hand, such interactive topic network exploration process has also provided a novel approach for capturing the search interests of news seekers automatically. We have developed a new algorithm for generating personalized topic network automatically from the current search actions of news seeker while the new seeker navigates the topic network. Thus the *personalized interestingness score* for a given news topic C_i on the topic network is defined as:

$$\rho(C_i) = \rho^{org}(C_i) + \rho^{org}(C_i)\left\{\alpha\frac{e^{v(C_i)} - e^{-v(C_i)}}{e^{v(C_i)} + e^{-v(C_i)}} + \beta\frac{e^{s(C_i)} - e^{-s(C_i)}}{e^{s(C_i)} + e^{-s(C_i)}} + \delta\frac{e^{d(C_i)} - e^{-d(C_i)}}{e^{d(C_i)} + e^{-d(C_i)}}\right\} \tag{6}$$

where $\alpha + \beta + \delta = 1$, $\rho^{org}(C_i)$ is the original interestingness score for the given news topic C_i as defined in Eq. (2), $v(C_i)$ is the visiting times of the given news topic C_i from the particular news seeker, $s(C_i)$ is the staying seconds on the given news topic C_i from the particular news seeker, $d(C_i)$ is the interaction depth for the particular user to interact with the news topic C_i and the relevant news videos which are relevant to the given news topic C_i, α, β and δ are the weighting factors. The visiting times $v(C_i)$, the staying seconds $s(C_i)$, and the interaction depth $d(C_i)$ can be captured automatically in the user-system interaction procedure. Thus the personalized interestingness scores of the news topics are determined immediately when such user-system interaction happens, and they will increase adaptively with the visiting times $v(C_i)$, the staying seconds $s(C_i)$, and the interaction depth $d(C_i)$.

After the personalized interestingness scores for all these news topics are learned from the current search actions of news seeker, they can further be used to weight the news topics for generating a *personalized topic network* to represent the user profiles (i.e., search preferences of news seeker) precisely. Thus the news topics with smaller values of the personalized interestingness scores can be eliminated automatically from the topic network, so that each news seeker can be informed by the most interesting news topics according to his/her personal preferences.

The search interests of news seeker may be changed according to his/her timely observations of news topics, and one major problem for integrating the user's profiles for topic recommendation is that the user's profiles may over-specify the search interests of news seeker and thus they may hinder news seeker to search other interesting news topics on the topic network. Based on this observation, we have developed a novel algorithm for propagating the search preferences of news seeker over other relevant news topics on the topic network, i.e., the news topics which have stronger correlations with the news topics which have been accessed by the particular news seeker. Thus the personalized interestingness score $v(C_j)$ for the news topic C_j to be propagated is determined as:

$$\chi(C_j) = \rho(C_j)\bar{\phi}(C_j), \quad \bar{\phi}(C_j) = \sum_{l \in \Omega} \phi(C_l, C_j) \tag{7}$$

where Ω is the set of the accessed news topics linked with the news topic C_j to be propagated as shown in Fig. 6 and Fig. 7, $\phi(C_l, C_j)$ is the inter-topic association between the news topic C_j and the news topic C_l which is linked with C_j and has

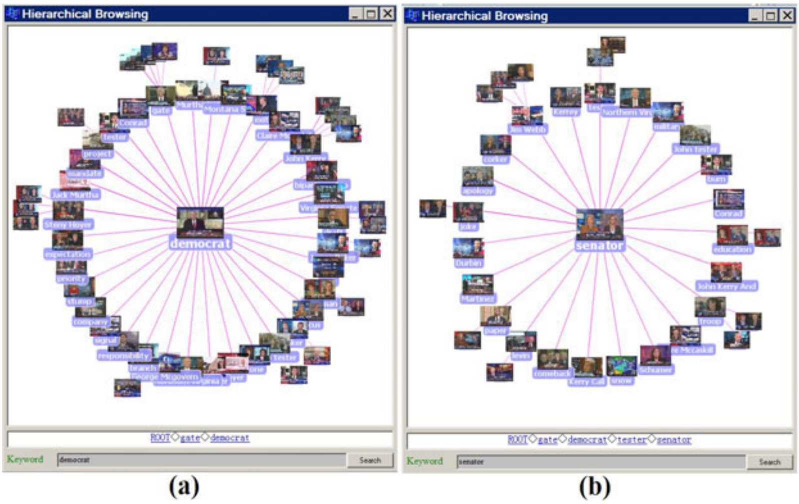


Fig. 7. The most relevant news topics for interestingness propagation

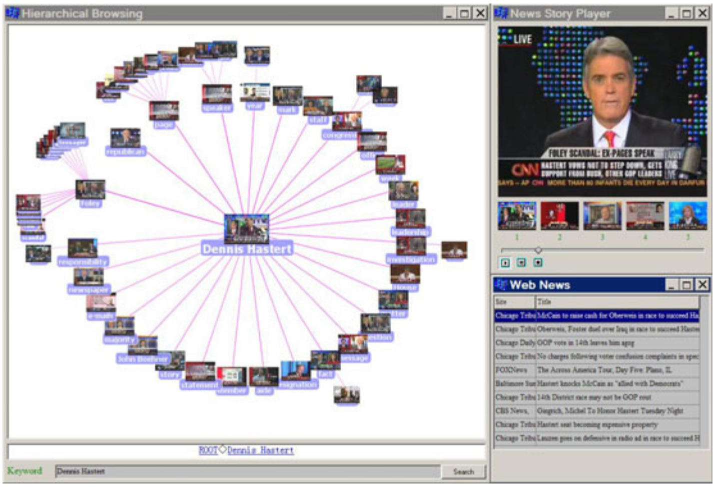
been accessed by the particular news seeker, and $\rho(C_j)$ is the interestingness score for the news topic C_j to be propagated. Thus the news topics, which have larger values of the personalized interestingness scores $\chi(\cdot)$ (strongly linked with some accessed news topics on the topic network), can be propagated adaptively.

By integrating the inter-topic correlations for automatic propagation of the preferences of news seeker, our proposed framework can precisely predict his/her hidden preferences (i.e., search intentions) from his/her current actions. Thus the user's profiles can be represented precisely by using the personalized topic network, where the interesting news topics can be highlighted according to their personalized interestingness scores as shown in Fig. 6 and Fig. 7. Such personalized topic network can further be used to recommend the news topics of interest for news seekers to make better future search decisions.

4 Personalized News Video Recommendation

Because the same news topic may be discussed many times in the same TV news program or be discussed simultaneously by multiple TV news programs, the amount of news videos under the same topic could be very large. Thus topic-based news search via keyword matching may return large amount of news videos which are relevant to the same news topic. To reduce the information overload, it is very important to develop new algorithms for ranking the news videos under the same news topic and recommending the most relevant news videos according to their importance and representativeness scores [18-19].

The news videos, which are relevant to the given news topic C_j, are ranked according to their importance and representiveness scores. For the given news topic



Keyword: Dennis Hastert

Site	Title
Chicago Tribu	McCain to raise cash for Oberweis in race to succeed Ha
Chicago Tribu	Oberweis, Foster duel over Iraq in race to succeed Haste
Chicago Daily	GOP vote in 14th leaves him agog
Chicago Tribu	No charges following voter confusion complaints in spec
FOXNews	The Across America Tour, Day Five: Plano, IL
Baltimore Sun	Hastert knocks McCain as "allied with Democrats"
Chicago Tribu	14th District race may not be GOP rout
CBS News,	Gingrich, Michel To Honor Hastert Tuesday Night
Chicago Tribu	Hastert seat becoming expensive property
Chicago Tribu	Lauzen goes on defensive in radio ad in race to succeed H

Fig. 8. Our system interface for supporting multi-modal news recommendation

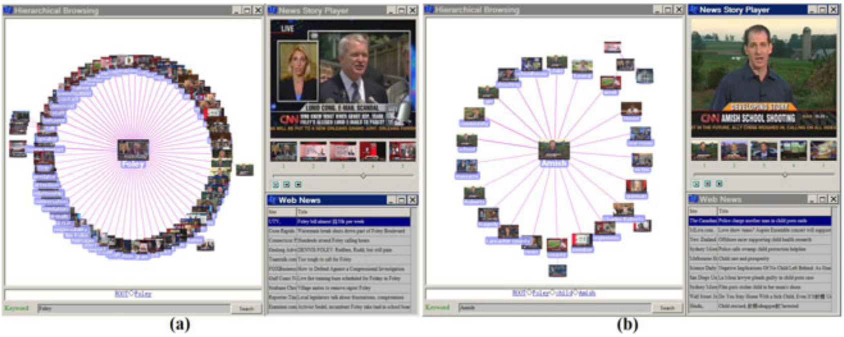
C_j, the importance and representativeness score $\varrho(x|C_j)$ for one particular news video x is defined as:

$$\varrho(x|C_j) = \alpha e^{-\Delta t} + (1-\alpha)\left\{\beta \frac{e^{v(x|C_j)} - e^{-v(x|C_j)}}{e^{v(x|C_j)} + e^{-v(x|C_j)}} + \gamma \frac{e^{r(x|C_j)} - e^{-r(x|C_j)}}{e^{r(x|C_j)} + e^{-r(x|C_j)}} \right. \\ \left. +\eta \frac{e^{q(x|C_j)} - e^{-q(x|C_j)}}{e^{q(x|C_j)} + e^{-q(x|C_j)}} \right\} \quad (8)$$

where $\beta + \lambda + \eta = 1$, Δt is the time difference between the time for the TV news programs to discuss and report the given news topic C_j and the time for the news seeker to submit their searches, $v(x|C_j)$ is the visiting times of the given news video x from all the news seekers, $r(x|C_j)$ is the rating score of the given news video x from all the news seekers, $q(x|C_j)$ is the quality of the given news video.

We separate the time factor from other factors for news video ranking because the time factor is more important than other factors for news video ranking (i.e., one topic can be treated as the news because it is new and tell people what is happening now or what is discussing now). The quality $q(x|C_j)$ is simply defined as the frame resolution and the length of the given news video x. If a news video has higher frame resolution and longer length (be discussed for longer time), it should be more important and representative for the given news topic.

After the search goals pf news seekers (i.e., which are represented by the accessed news topics) are captured interactively, our personalized news video recommendation system can: (a) recommend top 5 news videos according to their importance and representative scores; (b) recommend the news topics of interest on the topic network which are most relevant to the accessed news topic and suggest them as the future search directions according to the current preferences of news seekers, where the accessed news topic is set as the current focus (i.e., center of the topic network); (c) recommend the most relevant online text news which are relevant with the accessed news topic, so that news seekers can also read the most relevant online web news; (d) record the search history and preferences of news seekers for generating



(a) (b)

Fig. 9. Two examples for supporting multi-modal news recommendation

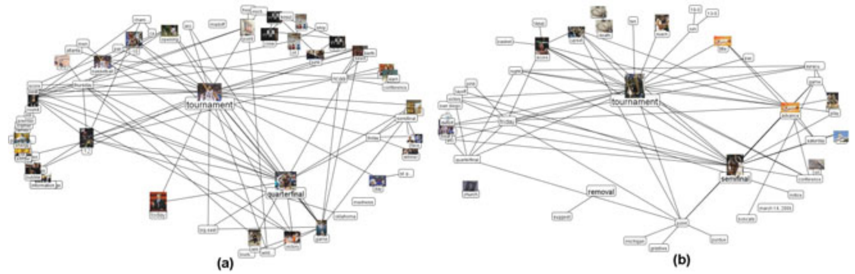
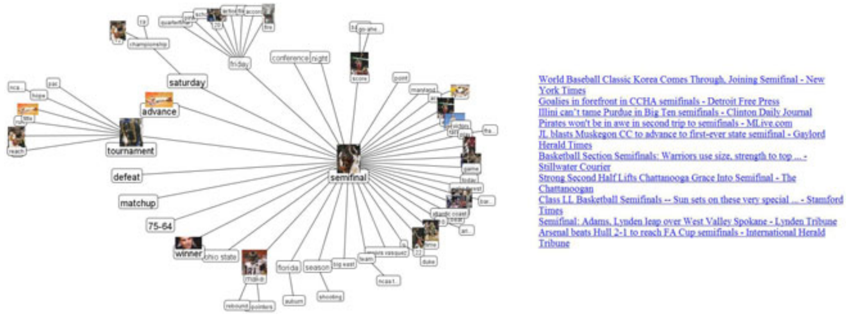


Fig. 10. Our system for supporting online news recommendation: (a) topic network for March 13; (b) topic network for March 14



World Baseball Classic Korea Comes Through, Joining Semifinal - New York Times
Goalies in forefront in CCHA semifinals - Detroit Free Press
Illini can't tame Purdue in Big Ten semifinals - Clinton Daily Journal
Pirates won't be in awe in second trip to semifinals - MLive.com
JL blasts Muskegon CC to advance to first-ever state semifinal - Gaylord Herald Times
Basketball Section Semifinals: Warriors use size, strength to top ... - Stillwater Courier
Strong Second Half Lifts Chattanooga Grace Into Semifinal - The Chattanoogan
Class LL Basketball Semifinals -- Sun sets on these very special ... - Stamford Times
Semifinal: Adams, Lynden leap over West Valley Spokane - Lynden Tribune
Arsenal beats Hull 2-1 to reach FA Cup semifinals - International Herald Tribune

Fig. 11. Our system for supporting online news recommendation: personalized topic network and the relevant online news sources

more reliable personalized topic network to make better recommendation in the future. Some experimental results are given in Fig. 8 and Fig. 9, one can conclude that our personalized news video recommendation system can effectively support multi-modal news recommendation from large-scale collections of news videos.

We have also extended our multi-modal news analysis tools to support personalized online news recommendation. First, the news topics of interest are extracted from large-scale online news collections and the inter-topic similarity contexts are determined for topic network generation as shown in Fig. 10, one can observe that such the topic network can represent the news topics of interest and their inter-topic similarity contexts effectively. By incorporating the inputs of news seekers for online news recommendation, the accessed news topic is set as the current focus and the most relevant news sources are recommended as shown in Fig. 11.

5 Algorithm Evaluation

We carry out our experimental studies by using large-scale news videos. The topic network which consists of 4000 most popular news topics is learned automatically from large-scale news videos. Our work on algorithm evaluation focus on: (1) evaluating the performance of our news topic detection algorithm and assessing the advantages for integrating multiple information sources for news topic detection; (2) evaluating the response time for supporting change of focus in our system, which is critical for supporting interactive navigation and exploration of large-scale topic network to enable user-adaptive topic recommendation; (3) evaluating the performance (efficiency and accuracy) of our system for allowing news seekers to look for some particular news videos of interest (i.e., personalized news video recommendation).

Automatic news topic detection plays an important role in our personalized news video recommendation system. However, automatic topic detection is still an open problem in natural language processing community. On the other hand, automatic video understanding via semantic classification is also an open issue in computer vision community. In this chapter, we have integrated multiple information sources (audio, video and closed captions) to exploit the cross-media advantages for achieving more reliable news topic detection.

Based on this observation, our algorithm evaluation for our automatic news topic detection algorithm focuses on comparing its performance difference by combining different information sources for news topic detection. We have compared three combination scenarios for news topic detection: (a) only the closed captions are used for news topic detection; (b) the closed captions and the audio channel are integrated and synchronized for news topic detection; (c) the closed captions, the audio channel and the video channel are seamlessly integrated and synchronized for news topic detection. As shown in Fig. 12, integrating multiple information sources for news topic detection can enhance the performance of our algorithm significantly.

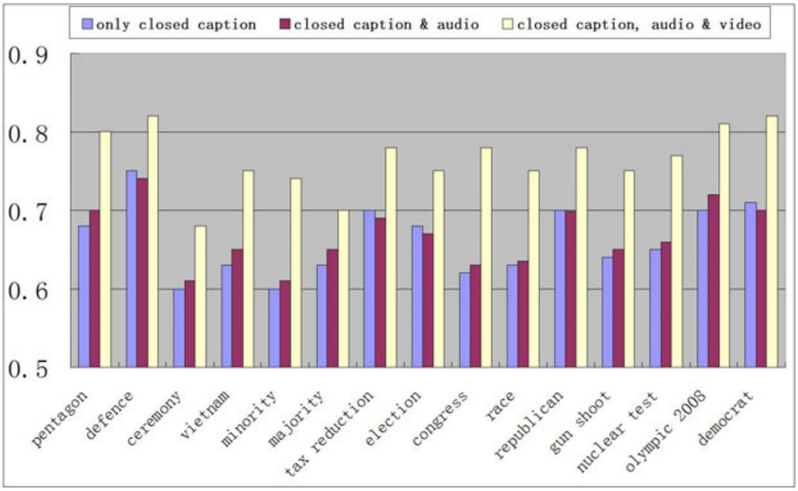


Fig. 12. The comparision results of our automatic news topic detection algorithm by integrating different sources

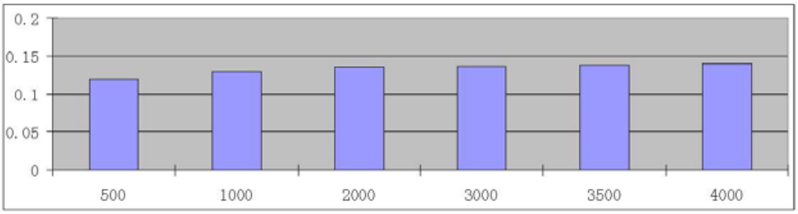


Fig. 13. The empirical relationship between the computational time T_1 (seconds) and the number of news topic nodes

One critical issue for evaluating our personalized news video recommendation system is the response time for supporting change of focus to enable interactive topic network navigation and exploration, which is critical for supporting user-adaptive topic recommendation. In our system, the change of focus is used for achieving interactive exploration and navigation of large-scale topic network. The *change of focus* is implemented by changing the Poincaré mapping of the news topic nodes from the hyperbolic plane to the display unit disk, and the positions of the news topic nodes in the hyerbolic plane need not to be altered during the focus manipulation. Thus the response time for supporting change of focus depends on two components: (a) The computational time T_1 for re-calculating the new Poincaré mapping of large-scale topic network from a hyperbolic plane to a 2D display unit disk, i.e., re-calculating the Poincaré position for each news topic node; (b) The visualization time T_2 for re-layouting and re-visualizing the new Poincaré mapping of large-scale topic network on the display disk unit. As shown in Fig. 13, one can find that the computational time T_1 is not sensitive to the number of news topics, and thus re-calculating the Poincaré mapping for large-scale topic network can almost be achieved in real time. We have also evaluated the empirical relationship between the visualization time T_2 and the number of news topic nodes. In our experiments, we have found that re-visualization of large-scale topic network is not sensitive to the number of news topics, and thus our system can support re-visualization of large-scale topic network in real time.

When the news topics of interest are recommended, our system can further allow news seekers to look for the most relevant news videos according to their importance and representative scores. For evaluating the effeciency and the accuracy of our personalized news video recommendation system, the *benchmark metric* includes *precision* P and *recall* R. The precision P is used to characterize the accuracy of our system for finding the particular news videos of interest, and the recall R is used to characterize the efficiency of our system for finding the particular news videos of interest. They are defined as:

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+TN} \tag{9}$$

where TP is the set of true positive news videos that are relevant to the need of news seeker and are recommended correctly, FP is the set of fause positive news

Table 1. The precision and recall for supporting personalized news video recommendation

news topics	**policy**	**pentagon**	**change**	**insult**
precision/recall	**95.6% /97.3%**	**98.5% /98.9%**	**100% /99.2%**	**92.8% /93.6%**
news topics	**implant**	**wedding**	**haggard**	**bob**
precision/recall	**90.2% /93.5%**	**96.3% /94.5%**	**96.5% /92.8%**	**90.3% /97.4%**
news topics	**gate**	**steny hoyer**	**democrat**	**urtha**
precision/recall	**95.9% /96.8%**	**96.5% /96.2%**	**96.3% /97.1%**	**93.6% /94.3%**
news topics	**majority**	**leader**	**confirmation**	**defence**
precision/recall	**99.2% /98.6%**	**93.8% /99.3%**	**94.5% /93.8%**	**100% /99.6%**
news topics	**secretary**	**veterm**	**ceremony**	**service**
precision/recall	**100% /98.8%**	**99.8% /99.2%**	**99.3% /96.6%**	**91.2% /93.2%**
news topics	**honor**	**vietnam**	**lesson**	**submit**
precision/recall	**91.2% /93.5%**	**98.8% /96.7%**	**90.3% /91.6%**	**91.2% /91.5%**
news topics	**minority**	**indonesia**	**president**	**trent lott**
precision/recall	**100% /99.6%**	**96.8% /97.7%**	**100% /96.8%**	**92.5% /92.3%**
news topics	**o.j. sinpson**	**trial**	**money**	**book**
precision/recall	**95.6% /99.4%**	**90.5% /90.3%**	**100% /90.6%**	**96.8% /93.6%**
news topics	**john kerry**	**military**	**race**	**mandate**
precision/recall	**100% /96.5%**	**100% /93.2%**	**100% /97.8%**	**92.6% /92.5%**
news topics	**election**	**leadship**	**school gun shoot**	**execution**
precision/recall	**100% /95.5%**	**92.8% /90.3%**	**100% /96.7%**	**90.6% /91.3%**
news topics	**responsibility**	**sex**	**message**	**congress**
precision/recall	**92.1% /91.5%**	**97.5% /98.2%**	**88.3% /87.6%**	**100% /96.3%**
news topics	**north korea**	**japan**	**china**	**white house**
precision/recall	**100% /99.3%**	**98.5% /95.6%**	**97.3% /95.2%**	**100% /94.8%**
news topics	**nuclear test**	**republican**	**amish**	**gun shoot**
precision/recall	**100% /97.6%**	**91.6% /92.8%**	**99.5% /91.6%**	**100% /99.8%**
news topics	**teacher**	**conduct**	**program**	**olmypic 2008**
precision/recall	**93.8% /94.5%**	**87.92% /88.3%**	**83.5% /90.2%**	**100% /99.3%**
news topics	**beijing**	**child**	**tax reduction**	**shooting**
precision/recall	**99.2% /97.3%**	**91.3% /91.5%**	**98.5% /96.9%**	**99.6% /98.4%**
news topics	**safety**	**investigation**	**ethic**	**committee**
precision/recall	**94.5% /94.8%**	**93.3% /96.5%**	**93.3% /95.6%**	**91.8% /95.2%**
news topics	**scandal**	**dennis hastert**	**preseident candidates**	**matter**
precision/recall	**96.6% /97.3%**	**95.3% /88.3%**	**98.5% /97.3%**	**85.2% /85.3%**

videos that are relevant to the need of news seeker and are not recommended, and TN is the set of true negative news videos that are relevant but are recommended incorrectly. Table 1 gives the precision and recall of our personalized news video recommendation system. From these experimental results, one can observe that our system can support personalized news video recommendation effectively, thus news seekers are allowed to search for some particular news videos of interest effectively.

6 Conclusions

In this chapter, we have developed an interactive framework to support personalized news video recommendation and allow news seekers to access large-scale news videos more effectively. To allow news seekers to obtain a good global overview of large-scale news videos at the topic level, topic network and hyperbolic visualization are seamlessly integrated to achieve user-adaptive topic recommendation. Thus news seekers can obtain the *news topics of interest* interactively, build up their mental search models easily and make better search decisions by selecting the visible news topics directly. Our system can also capture the search intentions of news seekers implicitly and further recommend the most relevant news videos according to their importance and representativeness scores. Our experiments on large-scale news videos have provided very positive results.

References

1. Marchionini, G.: Information seeking in electronic environments. Cambridge University Press, Cambridge (1997)
2. Fan, J., Keim, D., Gao, Y., Luo, H., Li, Z.: JustClick: Personalized Image Recommendation via Exploratory Search from Large-Scale Flickr Images. IEEE Trans. on Circuits and Systems for Video Technology 19(2), 273–288 (2009)
3. Yang, B., Mei, T., Hua, X.-S., Yang, L., Yang, S.-Q., Li, M.: Online video recommendation based on multimodal fusion and relevance feedback. In: ACM Conf. on Image and Video Retrieval (CIVR 2007), pp. 73–80 (2007)
4. Yang, H., Chaisorn, L., Zhao, Y., Neo, S.-Y., Chua, T.-S.: VideoQA: question answering on news video. In: ACM Multimedia, pp. 632–641 (2003)
5. Borthwick, A., Sterling, J., Agichtein, E., Grishman, R.: NYU: Description of the MENE named entity system as used in MUC-7. In: Proc. of the Seventh Message Understanding Conf, MUC-7 (1998)
6. McDonald, D., Chen, H.: Summary in context: Searching versus browsing. ACM Trans. Information Systems 24(1), 111–141 (2006)
7. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Intl. Conf. on New Methods in Language Processing (1994)
8. A.I. Inc. Lingpipe, http://www.aliasi.com/lingpipe/
9. Swan, R.C., Allan, J.: TimeMine: visualizing automatically constructed timelines. In: ACM SIGIR (2000)
10. Havre, S., Hetzler, B., Whitney, P., Nowell, L.: ThemeRiver: Visualizing thematic changes in large document collections. IEEE Trans. on Visualization and Computer Graphics 8(1), 9–20 (2002)
11. Luo, H., Fan, J., Yang, J., Ribarsky, W., Satoh, S.: Large-scale new video classification and hyperbolic visualization. In: IEEE Symposium on Visual Analytics Science and Technology (VAST 2007), pp. 107–114 (2007)
12. Luo, H., Fan, J., Yang, J., Ribarsky, W., Satoh, S.: Exploring large-scale video news via interactive visualization. In: IEEE Symposium on Visual Analytics Science and Technology (VAST 2006), pp. 75–82 (2006)
13. van Wijk, J.: Bridging the gaps. IEEE Computer Graphics and Applications 26(6), 6–9 (2006)

14. Walter, J.A., Ritter, H.: On interactive visualization of high-dimensional data using the hyperbolic plane. In: ACM SIGKDD (2002)
15. Lamping, J., Rao, R.: The hyperbolic browser: A focus+content technique for visualizing large hierarchies. Journal of Visual Languages and Computing 7, 33–55 (1996)
16. Furnas, G.W.: Generalized fisheye views. In: ACM CHI, pp. 16–23 (1986)
17. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: WWW (1998)
18. Wang, J., Chen, Z., Tao, L., Ma, W.-Y., Liu, W.: Ranking user's relevance to a topic through link analysis on web logs. In: WIDM, pp. 49–54 (2002)
19. Lai, W., Hua, X.-S., Ma, W.-Y.: Towards content-based relevance ranking for video search. In: ACM Multimedia, pp. 627–630 (2006)
20. Teevan, J., Dumais, S., Horvitz, E.: Personalized search via automated analysis of interests and activities. In: ACM SIGIR (2005)
21. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval 4(2), 133–151 (2001)
22. Mooney, R., Roy, L.: Content-based book recommending using learning for text categorization. In: ACM Conf. on Digital Libraries, pp. 195–204 (2000)
23. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Boston (1998)
24. Naphade, M., Smith, J.R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-scale concept ontology for multimedia. IEEE Multimedia (2006)
25. Fan, J., Gao, Y., Luo, H.: Integrating concept ontology and multi-task learning to achieve more effective classifier training for multi-level image annotation. IEEE Trans. on Image Processing 17(3) (2008)
26. Fan, J., Gao, Y., Luo, H., Jain, R.: Mining multi-level image semantics via hierarchical classification. IEEE Trans. on Multimedia 10(1), 167–187 (2008)
27. Fan, J., Luo, H., Gao, Y., Jain, R.: Incorporating concept ontology to boost hierarchical classifier training for automatic multi-level video annotation. IEEE Trans. on Multimedia 9(5), 939–957 (2007)
28. Fan, J., Yau, D.K.Y., Elmagarmid, A.K., Aref, W.G.: Automatic image segmentation by integrating color edge detection and seeded region growing. IEEE Trans. on Image Processing 10(10), 1454–1466 (2001)
29. Wactlar, H., Hauptmann, A., Gong, Y., Christel, M.: Lessons learned from the creation and deployment of a terabyte digital video library. IEEE Computer 32(2), 66–73 (1999)
30. Adams, W.H., Iyengar, G., Lin, C.-Y., Naphade, M.R., Neti, C., Nock, H.J., Smith, J.R.: Semantic indexing of multimedia content using visual, audio and text cues. EURASIP JASP 2, 170–185 (2003)
31. Naphade, M.R., Huang, T.S.: A probabilistic framework for semantic video indexing, filtering, and retrival. IEEE Trans. on Multimedia 3, 141–151 (2001)

Visual Quality Evaluation for Images and Videos

Songnan Li, Lawrence Chun-Man Mak, and King Ngi Ngan

The Chinese University of Hong Kong

1 Introduction

Information is exploding with technology progress. Compared with text and audio, image and video can represent information more vividly, which makes visual quality one of the most important aspects in determining user experience. A good visual quality evaluation method can assist in monitoring the quality of multimedia services and boosting user experience.

Visual quality evaluation plays its role in different stages of the visual information distribution chain, e.g., camera filter design for visual signal acquisition [51], quality monitoring during signal relaying [36], representation at the user site by display [39], printer [19], etc. In addition, the success of visual quality evaluation will provide guidance to a large number of image and video processing algorithms, e.g., compression, watermarking, image fusion, error protection, feature enhancement and detection, restoration, retrieval, graphic illumination, and so on. Due to its fundamental role, works on visual quality metrics can date back to half a century ago, and a vast number of objective quality metrics have been proposed over time.

Since pixel-based metrics, such as Mean Square Error (MSE), Signal-to-Noise Ratio (SNR) and Peak Signal-to-Noise Ratio (PSNR), are simple to calculate and easy to be incorporated into optimization process, they have been widely used in most visual-related products and services. However, it has been well acknowledged that these pixel-based difference measures do not correlate well with the Human Visual System's perception [18]. Distortions perceived by the human being are not always captured by MSE/SNR/PSNR, because these metrics operate on a pixel-by-pixel basis without considering the signal content, the viewing condition and the characteristics of the Human Visual System (HVS). These problems make the design of a better objective quality metric necessary, and progresses in recent vision research provide us with guidance to achieve this goal.

Different from objective quality metrics which work automatically without human intervention, subjective quality evaluation acquires quality judgment from the human observers in an off-line manner, and as a matter of course is considered to be the most accurate approach to measure visual quality. Although subjective quality evaluation is time-consuming and not feasible for on-line manipulation, its role in objective quality metric design is still irreplaceable: the perceptual visual quality derived from subjective evaluation can serve as a benchmark for the performance evaluation of different objective quality assessment algorithms, and can even direct the algorithm design. More and more subjectively-rated image and

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 497–544.
springerlink.com

video database become publicly available [59, 8, 20, 56, 10], which will speed up the advent of better objective quality metrics.

The goal of this chapter is to review both the objective and subjective visual quality evaluation methods, with Section 2 dedicated to the former and Sections 3 and 4 dedicated to the latter. In Section 2, objective visual quality metrics are reviewed on the basis of two distinct design approaches: HVS modeling approach and engineering based approach. Metrics for images and for videos are presented without clear partition due to their great similarities. Section 3 briefly describes the different aspects of a typical subjective evaluation procedure, while Section 4 presents an application of the subjective quality evaluation carried out recently by the authors: a performance comparison between two video coding standards — H.264 and AVS.

2 Objective Visual Quality Metrics

2.1 Classification

Objective visual quality metrics can be classified into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) metrics according to the availability of the reference information. For FR metrics, the original image or video sequence is fully available as the reference, and is considered to be of perfect quality. The distorted visual signal is compared against the reference and their similarity is measured so as to determine the quality of the distorted signal. FR metrics can be adopted in many image and video processing algorithms such as compression, watermarking, contrast enhancement, etc. However, in many practical applications, e.g., quality monitoring during transmission or at the user site, it is impossible to access the entire reference. RR and NR can fit in these situations. RR metrics only need partial reference signal. Features are extracted from the reference signal and transmitted in an ancillary channel for comparison against the corresponding features of the distorted signal extracted at the monitoring site. Compared with FR metrics, RR metrics are more flexible (weaker requirement on registration) and less expensive in terms of bandwidth requirement. NR metrics gauge quality without any reference information at all. They are free of registration requirement, and applicable to a wide range of applications. However, although human observers are good at NR quality evaluation, NR metric development turns out to be a difficult task, and only limited successes have been achieved so far.

From the design viewpoint, many classification methods have been proposed in literature: metrics designed by psychophysical approaches and by engineering approaches [89]; error sensitivity and structural sensitivity metrics [77]; bottom-up and top-down metrics [37], etc. Winkler et al. [90] proposed a comprehensive classification method which groups metrics into three categories: Data metrics, Picture metrics, Packet- and Bitstream-based metrics. Data metrics measure the fidelity of the signal without considering its content. The representatives are MSE/SNR/PSNR and their close families [15]. On the other hand, Picture metrics treat the visual data as the visual information that it contains. They may take the viewing conditions (viewing distance, ambient illumination, display properties, etc.), the HVS characteristics, and the signal content into consideration. Picture metrics can be further distinguished into two groups: metrics designed by vision modeling approaches (HVS-model-based metrics) and metrics designed by engineering approaches (engineering-based metrics). In this section, objective visual quality metrics will be separated into these two categories for a more detailed introduction. Packet- and Bitstream-based metrics [66, 30] are used in applications like internet streaming or IPTV. They measure the impact of network losses on visual signal quality. Different from traditional methods like bit error rate or packet loss rate, Packet- and Bitstream-based metrics distinguish the importance of the lost information to visual quality by checking the packet header and the encoded bitstream.

2.2 HVS-Model-Based Metrics

HVS-model-based metrics incorporate characteristics of the HVS which are obtained from psychophysical experiments of vision research. Although anatomy provides us with detailed physiological evidences about the front-end of the HVS (optics, retina, LGN etc.), a thorough understanding of the latter stages of the visual pathway (visual cortex, etc.) in charge of higher-level perception is still unachievable, which makes the construction of a complete physiological HVS model impossible. Consequently HVS models used by visual quality metrics are mostly based on psychophysical studies and only account for lower-level perception. Physiological and psychophysical factors typically incorporated into the HVS model include color perception, luminance adaptation, multi-channel decomposition, contrast sensitivity function (CSF), masking, pooling, etc., as shown in Fig. 1.

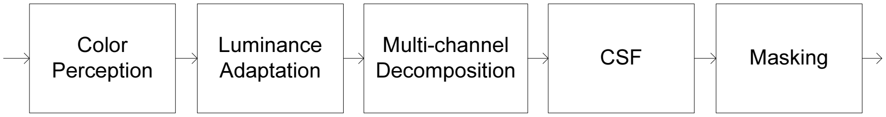


Fig. 1. A typical HVS model

Each of the perceptual factors listed above will be explained in detail in this section. But before that, it should be noted that visual quality metrics will also model different transformations of the visual signal, before it is perceived by the human eyes. For example, at the very beginning, visual signal usually is represented by pixel values. When displayed on a Cathode Ray Tube (CRT) or Liquid Crystal Display (LCD) monitor, these pixel values will be transformed into light intensities, which have a non-linear relationship with their corresponding pixel values. This non-linear relationship is determined by the gamma value of the display[1] and may be slightly different for the R, G, and B channels which will be introduced below.

[1] An excellent explanation on “gamma” can be found in [57].

2.2.1 Components of the HVS Model

- Color perception

R, G, and B stands for the three primary colors Red, Green, and Blue, which can be combined to create most of the visible colors. RGB color space is commonly employed by camera sensors and in computer graphics. There are physiological evidences that justify the use of this color space. When lights pass through the optics of the eye and arrive at the retina, they will be sampled and converted by two different types of photoreceptors: rods and cones. Rods and cones are responsible for vision at low light level and high light level, respectively. In addition, cones contribute in color perception. There are three types of cones: S cones, M cones, and L cones, which are sensitive to short, median, and long wavelengths, respectively, as shown in Fig. 2. They are often depicted as blue, green, and red receptors, although as a matter of fact they do not accurately correspond to these colors.

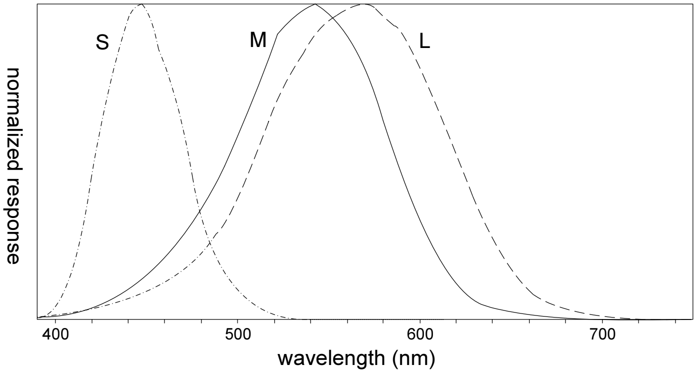


Fig. 2. Normalized response spectra of human cones, S, M, and L types, with wavelength given in nanometers [85]

Responses of the cones need to be further processed at a higher stage of the HVS for the purpose of "decorrelation". RGB channels are highly correlated with each other: by viewing the R, G, and B channels of a given image independently, you can find that each channel contains the entire image. Their correlations can also be seen from Fig. 2 where the three types of cones overlap in their sensitivities to the light wavelengths. Possibly for coding efficiency, HVS records the differences between the responses of the cones, rather than each type of cone's individual response. This is referred to as the *opponent processing theory* of color perception. According to this theory, there are three opponent channels: Black-White (B-W) channel, Red-Green (R-G) channel, and Blue-Yellow (B-Y) channel. Neural response to one color of an opponent channel is antagonistic to the other color in the same channel. This explains a number of perceptual phenomena, e.g., you can perceive a reddish yellow (orange) but you never perceive a reddish green. Physiological evidences support the existence of opponent channels: bipolar cells and ganglion cells of the retina may be involved in opponent color processing [88].

Besides the above mentioned RGB and B-W/R-G/B-Y, many other color spaces are developed for different purposes, e.g., CIELAB, HSL, YIQ, and so on. Most of these share the common characteristic that they treat visual information as a combination of luminance and chrominance components. Chrominance component is represented by two descriptors which usually have different physical meanings for different color spaces, such as a* (red-green balance) and b* (green-blue balance) for CIE L* a* b*, H (hue) and S (saturation) for HSL, and I (blue-green-orange balance) and Q (yellow-green-magenta balance) for YIQ. Luminance component on the other hand is more or less the same which is to simulate the B-W opponent channel of the HVS. Since B-W channel carries most of the visual information, many visual quality metrics only make use of luminance information for quality assessment. According to the performance comparison of different color spaces in visual quality metrics [86], there is only a slight performance loss due to abandoning the use of chrominance components, and on the other hand, the computational complexity can be reduced by a great amount.

- Luminance adaptation

It is well known that our perception is sensitive to luminance contrast rather than the luminance intensity. Given an image with a uniform background luminance *I* and a square at the center with a different luminance *l* + *dl*, if *dl* is the threshold value at which the central square can be distinguished from the background, then according to the Weber's law the ratio of *dl* divided by *l* is a constant for a wide range of luminance *l*. This implies that our sensitivity to luminance variation is dependent to the local mean luminance. In other words, the local mean luminance masks the luminance variation: the higher the local mean luminance, the stronger is the masking effect. That is the reason why the term "luminance masking" is preferred by some authors rather than "luminance adaptation". In practical implementations, the luminance of the original signal may serve as the masker to mask the luminance variation due to distortion.

If the ratio of *dl* to *l* is used to represent the differential change of the luminance contrast *dc,*

$$dc = \frac{K \times dl}{l}, \tag{1}$$

where *K* is a constant, then the following equation

$$c = K \times \log l + C, \tag{2}$$

can be obtained, which describes the relationship between the luminance contrast *c* and the luminance intensity *l*. Constants *K* and *C* can be determined experimentally. According to equation (2), the perceived luminance contrast is a non-linear function of the luminance intensity. For visual quality metrics, one approach to model luminance adaptation is by applying such a logarithmic function or other alternatives e.g., cube root, or square root function [13] before multi-channel decomposition.

Another way to simulate luminance adaptation is to convert the luminance intensity to the luminance contrast right after (rather than before) multi-channel decomposition, as shown in Fig. 3. Peli E. in [54] defines a local band limited contrast measure for complex images, which assigns a local contrast at every point of an image and at every frequency channel. Multi-channel decomposition creates a series of bandpass-filtered images and lowpass-filtered images. For bandpass-filtered image $a_k(x,y)$ and its corresponding lowpass-filtered image $l_{k-1}(x,y)$, the contrast at frequency subband k is expressed as a two-dimensional array $c_k(x,y)$:

$$c_k(x,y) = \frac{a_k(x,y)}{l_{k-1}(x,y)}. \tag{3}$$

This is one of the many attempts to define luminance contrast in complex images, and it has been adopted in visual quality metrics such as [39, 44] with some modifications.

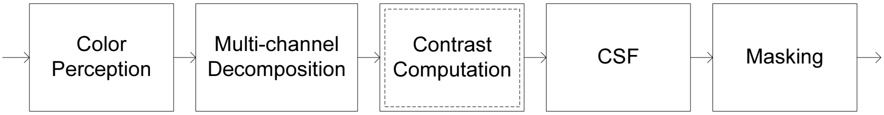


Fig. 3. HVS model using contrast computation

Other approaches to implement luminance adaptation can be found in Just Noticeable Difference (JND) modeling [35], which is closely related to visual quality assessment. In spatial domain JND, the visibility threshold of luminance variation can be obtained as a function of the background luminance, as shown in Fig. 4. In frequency domain JND, luminance adaptation effect is often implemented as a modification to the baseline threshold derived from the contrast sensitivity function, which will be introduced later.

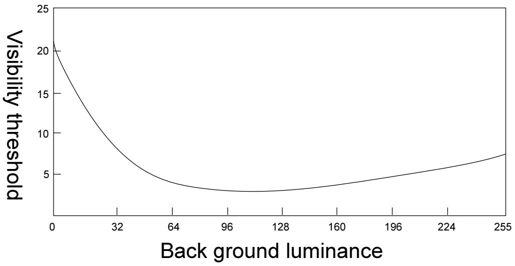


Fig. 4. Luminance adaption: visibility threshold versus background luminance [12]

- Multi-channel decomposition

Instead of employing just one channel as in the early works [41, 40], multi-channel decomposition has been widely used for HVS modeling nowadays. Multi-channel decomposition is justified by the discovery of the spatial frequency selectivity and orientation selectivity of the simple cells in the primary visual cortex. It can also be successfully used to explain empirical data from masking experiments.

Both temporal and spatial multi-channel decomposition mechanisms of the HVS have been investigated over time, with more efforts paid to the spatial one. For spatial multi-channel decomposition, most studies suggest that there exists several octave spacing radial frequency channels, each of which is further tuned by orientations with roughly 30 degree spacing [50]. Fig. 5 shows a typical decomposition scheme which is employed in [84] and is generated by cortex transform [79]. Many other decomposition algorithms serving this purpose exist, e.g., steerable pyramid transform [64], QMF (Quadrature Mirror Filters) transform [65], wavelet transform [31], DCT transform [80], etc. Some of these aim at accurately modeling the decomposition mechanism, while others are used due to their suitability for particular applications, e.g., compression [80]. A detailed comparison of these decomposition algorithms can be found in [85].

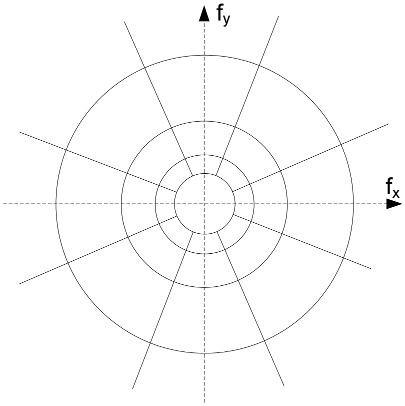


Fig. 5. Illustration of the partitioning of the spatial frequency plane by the steerable pyramid transform [31]

For temporal decomposition, it is generally believed that there exist two channels: one low-pass channel, namely sustained channel, and one band-pass channel, namely transient channel. Since most visual detailed information is carried in sustained channel, HVS models employed by some video quality metrics like those in [44, 94] only use a single low pass temporal filter to isolate the sustained channel, while the transient channel is disregarded. Temporal filters can be implemented as either Finite Impulse Response (FIR) filters [44] or Infinite Impulse Response (IIR) filters [32], either before [44] or after spatial decomposition [83].

- Contrast sensitivity function

Contrast sensitivity is the inverse of the contrast threshold – the minimum contrast value for an observer to detect a stimulus. These contrast thresholds are derived from psychophysical experiments using simple stimuli, like sine-wave gratings or Gabor patches. In these experiments, the stimulus is presented to an observer with its contrast increasing gradually. The contrast threshold is determined at the point where the observer can just detect the stimulus.

It has been proved by many psychophysical experiments that the HVS's contrast sensitivity depends on the characteristics of the visual stimulus: its spatial frequency, temporal frequency, color, and orientation, etc. Contrast sensitivity function (CSF) can be used to describe these dependences. Fig. 6 shows a typical CSF quantifying the dependency of contrast sensitivity on spatial frequency. The decreasing sensitivity for higher spatial frequency is a very important HVS property which has been widely applied in image and video compression: because the HVS is not sensitive to signals with higher spatial frequencies, larger quantization can be applied to them without introducing visible distortions. On the other hand, the decreasing sensitivity for lower frequencies is less crucial, and in many cases it has been neglected intentionally resulting in a low-pass version of the CSF [1]. CSF is more complex when the influences of other factors like temporal frequency or color are considered in conjunction with the spatial frequency [87].

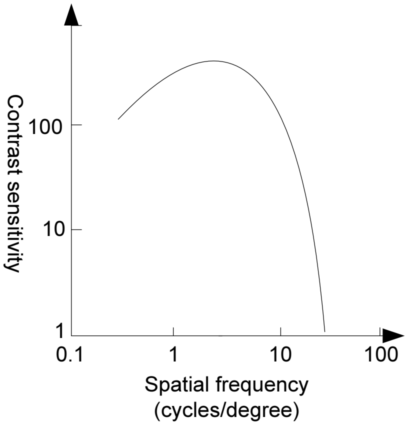


Fig. 6. A typical spatial CSF function

It should be noted that spatial frequency of a visual signal is a function of viewing distance. When the observer moves closer to the display or away from it, the spatial frequency of the visual signal will be changed. As a result, in order to make use of the CSF in the correct way, either viewing distance needs to be taken as a parameter for spatial frequency calculation, or the viewing distance should be fixed, e.g., 6 times image height for SDTV and 3 times image height for HDTV.

To incorporate it into the HVS model, CSF can be implemented either before or after the multi-channel decomposition. In the former case, CSF is implemented as linear filters with frequency response close to the CSF's. In the latter case, since visual signal has already been decomposed into different frequencies, CSF filtering can be approximated by multiplying each subband with a proper value. In JND models, CSF is often used to obtain the baseline contrast threshold, which will be further adjusted to account for luminance adaptation and the masking effect introduced below.

- Masking

Masking effect refers to the visibility threshold elevation of a target signal (the maskee) caused by the presence of a masker signal. It can be further divided into spatial masking and temporal masking.

In most spatial masking experiments, the target and masker stimuli are sinewaves or Gabor patches. The target stimulus is superposed onto the masker stimuli, and contrast threshold of the target stimulus are recorded, together with the masker information, including its contrast, spatial frequency, orientation, phase, etc. Many of these experiments verify that the threshold contrast of the target depends on the masking contrast, and also the other characteristics of the masker. Generally higher masking contrast and larger similarity between the masker and the target in their spatial frequencies, orientations, and phases will lead to higher masking effect, which is known as the contrast masking effect. Fig. 7 shows part of the contrast masking data from [33] describing the relationship between threshold contrast and the masking contrast. With the increase of the masking contrast, the threshold contrast reduces first and then increases, consistently for the three viewers. The threshold contrast reduction, referred to as *facilitation*, is often neglected in spatial masking models, resulting in a monotone increasing curve similar to Fig. 8.

One way to implement contrast masking is to make the original visual content act as the masker and the distortion as the target [13, 18]. In this case, usually contrast masking is assumed to occur only between stimuli located in the same channel (intra-channel masking) characterized by its unique combination of spatial frequency, orientation, phase, etc. The output of contrast masking function will

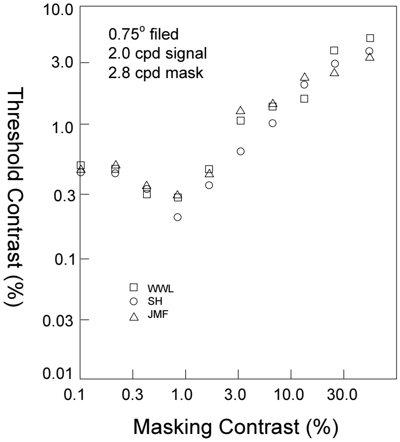


Fig. 7 Experimental data for contrast masking from [33]. WWL, SH, JMF represent three subjects.

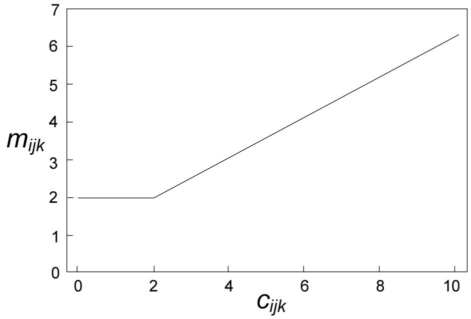


Fig. 8. Contrast masking function from [80], describing the masked threshold m_{ijk} as a function of DCT coefficient c_{ijk}.

be multiplied to the CSF baseline threshold to account for the contrast threshold elevation caused by contrast masking. In another type of contrast masking model, original visual content will no longer serve as the masker. Instead, the original and the distorted signals pass through the masking model separately. The outputs of the masking model simulate the response of cortical visual neurons to these visual contents, which will be compared directly in the next stage (pooling). The response of visual neuron can be modeled either by a saturating nonlinear transducer function [33] or by a contrast gain control process [65, 17, 81]. As an example of the latter case, Watson and Solomon integrate a variety of channel interactions into their model [81] (inter-channel masking), which is achieved by division of the excitatory signal from each neuron of one channel by an inhibitory signal that is a linear combination of responses of neurons within neighboring channels.

The sine-wave gratings or Gabor patches used to derive the above contrast masking models are oversimplifications with respect to natural images. Efforts have been made towards masking measurements with more realistic targets, e.g., quantization errors [47], and maskers, e.g., random noises, bandpass noises, and even natural images [82]. To emphasize their differences with the traditional contrast masking, different terms were used, such as noise masking, texture masking, or entropy masking, etc.

Compared with spatial masking, temporal masking has received less attention and is of less variety. In most of its implementations in video quality assessment, temporal masking strength is modeled as a function of temporal discontinuity in intensity: the higher the inter-frame difference, the stronger is the temporal masking effect. Particularly, the masking abilities of scene cut have been investigated in many experiments, with both of its "forward masking" and "backward masking" effects identified [2].

- Pooling

In vision system, pooling refers to the process of integrating information of different channels, which is believed to happen at the latter stages of the visual pathway. In visual quality assessment, pooling is used to term the error summation process which combines errors measured in different channels into a quality map or a single quality score. For image quality assessment, most approaches perform error summation across frequency and orientation channels first to produce a 2-D quality map, and then perform it across spaces to obtain a single score indicating the quality of the entire image. For video quality assessment, one more step is performed to combine quality scores for frames into a quality score for the video sequence.

Minkowski summation, as shown below, is the most popular approach to implement the error pooling process:

$$E = \left(\sum_i |e_i|^{\beta}\right)^{\frac{1}{\beta}}. \tag{4}$$

In the above equation, e_i represents error measured at channel/position/frame i; E is the integrated error; and β is the summation parameter which is assigned a value between 2 to 5 in most works. With a higher value of β, E will depend more on the larger e_is, which is consistent with the reality that visual quality is mostly determined by the stronger distortions.

In image quality metrics, higher-level characteristics of the vision system can be applied into quality metric by using spatial weighted pooling [74]:

$$E = \frac{\sum_i (w_i \times |e_i|)}{\sum_i w_i}, \tag{5}$$

where w_i is the weight given to the error e_i at spatial position i. It represents the significance of e_i to the visual quality of the image, and can be determined by cognitive behaviors of the vision system, such as visual attention [46]. In video quality metrics, temporal pooling can also integrate cognitive factors, such as the asymmetric behavior with respect to quality changes from bad to good and reverse [63].

2.2.2 Frameworks

Fig. 9 shows two different frameworks of HVS-model-based quality metrics. It should be noted that most HVS-model-based metrics are FR metrics, so their inputs include both the original and distorted visual signals.

In the first framework shown in Fig. 9 (a), the original and distorted signals pass through each of the HVS components separately, where the representations of the visual signals are changed sequentially simulating the processing of the HVS, until their differences are calculated and summed in the error pooling process. The summed error can be further converted to detection probability by using the probability summation rule as in [13], or converted to a quality score by using nonlinear regression as in [60].

In the second framework, JNDs need to be calculated in the domain where the original and the distorted signals are compared. JND is the short form for Just Noticeable Distortion which refers to the maximum distortion under the perceivable level. As shown in Fig. 9 (b), generally JND will be calculated as the product of

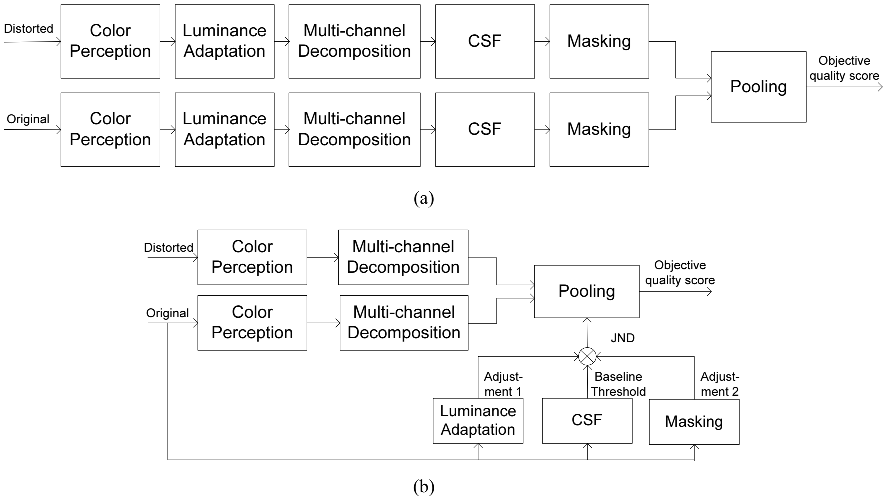


Fig. 9. Two frameworks of HVS-model-based quality metrics

the baseline contrast threshold obtained from the CSF and some adjustments obtained from luminance adaptation and various masking effects, such as contrast masking, temporal masking, and so on, which have been introduced in the last section. The errors of the distorted signal will be normalized (divided) by their corresponding JND values before they are combined in the error pooling process.

Frameworks different from the above mentioned exist, but often have slight differences. For example, the framework of DVQ [83] may be simplified as shown in Fig. 10, where the local contrast computation is added after multi-channel decomposition, and the luminance adaptation is removed compared with Fig. 9 (b).

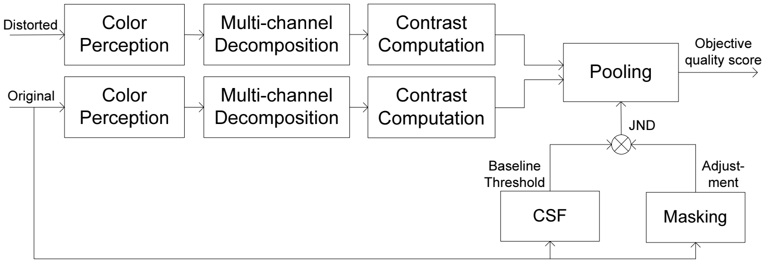


Fig. 10. DVQ's framework

2.2.3 Shortcomings

As stated above, the foundation of the HVS model mostly grounds on psychophysical experiments which use simple visual stimuli and target at contrast threshold evaluation. This leads to two major problems to visual quality metric that employs

HVS model as its kernel. Firstly, a natural image usually is a superposition of a large number of simple stimuli. Their interactions cannot be fully described by a model which is based on experimental data of only one or two simple stimuli. Secondly, there is no justification for the use of experimental data of contrast threshold evaluation in gauging visual quality, especially for images with supra-threshold distortions. For visual quality evaluation, it should be helpful to take higher-level behaviors of the vision system into consideration, but in most HVS models which target at contrast threshold prediction, only low-level perceptual factors are simulated. Besides the problems mentioned above, the high computational complexity is another disadvantage of the HVS-model-based quality metrics especially for video quality assessment.

2.3 Engineering-Based Metrics

To overcome these shortcomings brought by the vision model, recently many new visual quality metrics were designed by engineering approaches. Instead of founding on accurate experimental data from subjective viewing tests, these engineering-based quality metrics are based on (a) assumptions about, e.g., visual features that are closely related to visual quality; (b) prior knowledge about, e.g., the distortion properties or the statistics of the natural scenes. Since these features and prior knowledge are considered to be higher-level perceptual factors compared with lower-level ones used in the vision model, engineering-based quality metrics are also referred to as top-down quality metrics, and are considered to have the potential to better deal with supra-threshold distortions. In [27], International Telecommunication Union (ITU) recommends four video quality metrics after VQEG's FRTV Phase II tests [67], all of which belong to this category. This may serve as the evidence for the promising future of engineering-based quality metrics.

Unlike HVS-model-based metrics, most of which are FR, there are also many RR and NR engineering-based quality metrics. Since FR and RR metrics share great similarities in their processing routines, they will be reviewed together below, followed by the introduction of NR metrics.

2.3.1 FR and RR Metrics

Viewed conceptually, most engineering-based FR and RR quality metrics consist of three processing steps: (a) feature extraction; (b) feature comparison; and (c) quality determination, as shown in Fig. 11. The extracted features characterize the quality metric and determine its performance. These features may be either scalar ones or vectors, and their differences can be obtained in various ways, such as the absolute distance, the Euclidean distance, etc. In most quality metrics, the feature differences can quantify video distortions locally, by using one or several 2-D distortion maps. And these distortion maps will be combined together to generate a single quality score. Two methods are commonly used for the last step: in Fig. 12 (a), spatial/temporal pooling are performed first to generate several distortion factors each representing the intensity of a particular distortion type, and then these distortion factors will be combined at the end to generate a signal quality score for the entire image or video sequence; in Fig. 12 (b), distortions of different types are combined

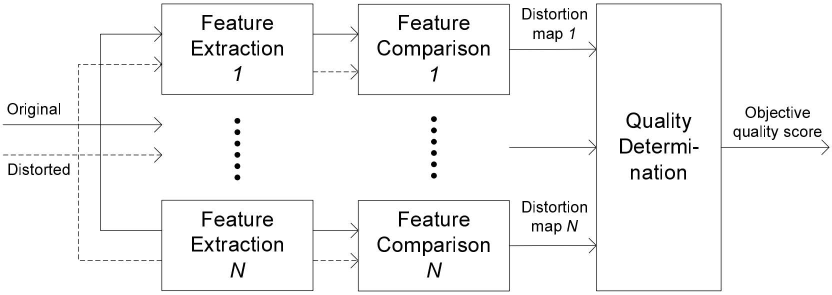


Fig. 11. A conceptual framework for FR and RR metrics

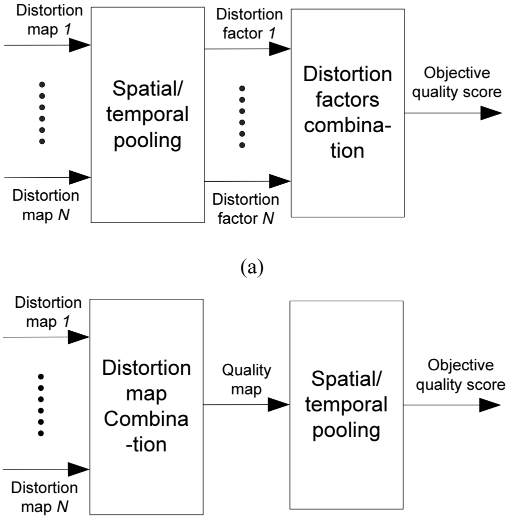


Fig. 12. Two methods for quality determination

together first to generate a quality map, and then spatial pooling is performed on this quality map to compute a single quality score.

In this section, four classic engineering-based image or video quality metrics will be briefly explained, with focus on their implementations of the three processing steps.

- Picture Quality Scale (PQS)

PQS [45] is a hybrid image quality metric employing both the HVS model and the engineering design approaches. Among the five distortion factors measured, three of them are obtained basically by using HVS models. Involved perceptual factors include luminance adaptation, CSF, and texture masking. The other two engineering-based distortion factors measure blockiness and error correlations.

To fit in the three processing steps introduced above, in PQS, non-linear mapped luminance values (to account for the luminance adaptation effect) are used as features, and feature comparison is implemented by direct subtraction. These feature differences are further processed by the CSF and by using prior knowledge about the locations of the distortions to produce two distortion maps measuring blockiness and local error correlations, respectively. In the last step, spatial pooling is performed separately on each of the two distortion maps, generating two engineering-based distortion factors. Together with the three HVS-model-based distortion factors, they are de-correlated by singular value decomposition and linearly combined to generate the PQS quality score.

Compared with the metrics that we will introduce below, the features and the comparison method used in PQS are very simple. In fact, it is not the features but the prior knowledge used, i.e., the locations of the distortions, that represents the idea of the engineering design approach.

- Video Quality Model (VQM)

VQM [55] is one of the best proponents of the VQEG FRTV Phase II tests [67]. For a video sequence, VQM generates seven distortion factors to measure the perceptual effects of a wide range of impairments, such as blurring, blockiness, jerky motion, noise and error blocks, etc. Viewed conceptually, VQM's distortion factors are all calculated in the same steps. Firstly, the video streams are divided into 3D Spatial-Temporal (S-T) sub-regions typically sized by 8 pixel × 8 lines × 0.2 second; then feature values will be extracted from each of these 3D S-T regions by using, e.g., statistics (mean, standard deviation, etc.) of the gradients obtained by a 13-coefficient spatial filter, and these feature values will be clipped to prevent them from measuring unperceivable distortions; Finally these feature values will be compared and their differences will be combined together for quality prediction.

Three feature comparison methods used by VQM are Euclidean distance, ratio comparison, and log comparison, as shown by equation (6), (7), (8), respectively, where f_o and f_{o2} are original feature values, and f_p and f_{p2} are the corresponding processed feature values. Euclidean distance is applied to 2D features (C_B-C_R vectors), and the other two are applied to scalar features (luminance values). The feature differences are integrated by spatial and temporal pooling first, generating seven distortion factors, which are then linearly combined at the last to yield the final VQM quality score.

$$p = \sqrt{(f_o - f_p)^2 + (f_{o2} - f_{p2})^2} \,. \tag{6}$$

$$p = (f_p - f_o) / f_o \,. \tag{7}$$

$$p = \log_{10}(\frac{f_p}{f_o}) \,. \tag{8}$$

- Structural Similarity Index (SSIM)

SSIM was first proposed in [77, 73] as a FR image quality metric. It has been extended to video quality metrics [78, 58] and applied to numerous vision-related applications.

The basic assumption of SSIM is that the HVS is highly adapted to extract structural information from the viewing field. SSIM divides the input images into overlapping image patches (e.g., 8×8 pixel blocks), and from each image patch three features were extracted. The first two scalar features are the mean μ and standard deviation σ of the luminance values of the image patch. The third feature can be regarded as a vector with its elements being luminance values normalized by σ. The extracted features from the reference image patch x and the distorted image patch y are compared by using the following equations[2]:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{{\mu_x}^2 + {\mu_y}^2 + C_1}, \tag{9}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{{\sigma_x}^2 + {\sigma_y}^2 + C_2}, \tag{10}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{11}$$

where $l(x,y)$, $c(x,y)$ and $s(x,y)$ are termed as the luminance similarity, the contrast similarity and the structural similarity, respectively, and constants C_1, C_2 and C_3 are used to avoid division by zero. Different from PQS and VQM, SSIM adopts the quality determination method shown in Fig. 12 (b): the three similarity factors were combined first before the spatial pooling was performed. This makes SSIM being able to produce a spatially varying quality map which indicates quality variations across the image.

- Visual Information Fidelity (VIF)

VIF [60] is a FR image quality metric grounding on the assumption that visual quality is related to the amount of information that the HVS can extract from an image. Briefly, VIF works in the wavelet domain and uses three models to model the original natural image, the distortions, and the HVS, respectively. As shown in Fig. 13, *C*, *D*, *E* and *F* are the modeling results for the original image, the distorted image, the perceived original image, and the perceived distorted image, respectively. Each of them is represented by a set of random fields, in such a way that the mutual information between any two of them is measurable. The mutual information between *C* and *E*, *I*(*C*,*E*), quantifies the information that the HVS can extract from the original image, whereas the mutual information between *C* and *F*,

[2] According to the explanations about the features used by SSIM, equation (11) actually involves both extraction and comparison of the third feature.

I(*C*,*F*), quantifies the information that can be extracted from the distorted image. The VIF quality score is given by equation (12). Viewed conceptually, the only feature used by VIF is the visual information. For implementation, the mutual visual information *I*(*C*,*E*) and *I*(*C*,*F*) are measured locally within each wavelet subband. Pooling over spaces and wavelet subbands is performed to obtain the numerator and the denominator of equation (12).

$$s_{VIF} = \frac{I(C,F)}{I(C,E)}. \tag{12}$$

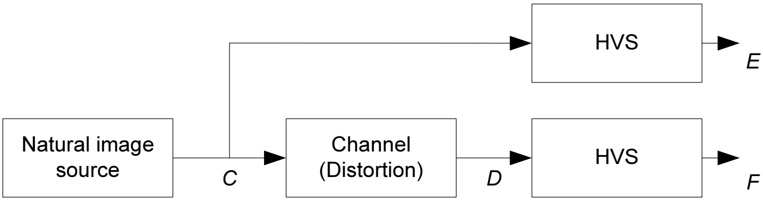


Fig. 13 Block-diagram of VIF from [60]

2.3.2 NR Metrics

As mentioned before, although human observers can easily assess quality without reference information, NR metric design is by no means an easy task. By contrary, NR quality assessment is so difficult that their applications are often limited to the cases where the prior knowledge about the distortion type is available. The distortion types that NR metrics often deal with include blocking, blurring, ringing, and jerky/jitter motion, etc., caused by signal acquisition, compression, or transmission. Another prior knowledge used by some NR metrics asserts that natural images belong to a small set in the space of all possible signals, and they can be described by statistical models fairly well. However, distortions may cause the modeling's inaccuracy, which in turn can be used as an indication of the distortion strength.

- Blocking

Blocking artifacts arise from block-based compression algorithms, like JPEG, MPEG1/2, H.26x, etc., running at low bit rates. Due to the fixed block structure commonly used, blockiness often appears as periodic horizontal and vertical edges whose positions are fixed on block boundaries. Most NR quality metrics detect and quantify blockiness in the spatial domain, by directly measuring differences of the boundary pixels. And in some NR metrics [91, 5, 95] these boundary differences are further adjusted to account for the luminance and texture masking effects. On the other hand, a handful of NR metrics detect blockiness in frequency domain. For example in [76], 1-D FFT is applied to the differences of adjacent rows or columns of the image. Periodic peaks caused by blockiness are identified in the resultant power spectrum and are used to assess blockiness. In [71], eight sub-images are constructed from the distorted image. Their similarities are measured in the Fourier transform domain to obtain two values describing inter-block similarity and intra-block similarity, respectively, with the former being closely related to the blockiness strength. The inter-block similarity is normalized by the intra-block similarity to yield the final blockiness measure.

- Blurring

Blurring artifacts can be caused by camera out-of-focus, fast camera motion, data compression, and so on. Unlike blockiness which can be easily localized, blurring is more image content dependent. Since blur affects edges most conspicuously, many spatial-domain blur metrics make use of this, by detecting step edges and then estimating the edge spread in the perpendicular direction [43, 49]. Also there are blur metrics that work in the frequency domain. For example in [42], the authors proposed a blur determination technique based on histograms of non-zero DCT coefficients of the JPEG or MPEG compressed signals. In [9], a blur metric was developed based on local frequency spectrum measurement, i.e., 2D kurtosis, around the edge regions.

- Ringing

Ringing is another common compression artifact caused by high frequency quantization. Analogous to the Gibbs phenomenon, ringing artifact manifests itself in the form of spurious oscillations, and appears most prominently in the smooth regions around the edges. Compared with blocking metrics or blurring metrics, ringing NR metrics are less investigated, and most existing ones [48, 11] follow a similar conceptual routine: identifying strong edges first and then detecting activities around them as the indication of the ringing artifact intensity. In an alternative approach [34], ringing strength is quantified by measuring the noise spectrum that is filtered out by anisotropic diffusion.

It should be noted that since ringing artifacts often coexist with other compression artifacts, e.g., blockiness or blur, and appear to be less annoying comparatively, quality metrics rarely aim at quantifying ringing artifact only. Instead most NR metrics that measure ringing artifacts will also consider other distortion types, and will balance their contributions to the final quality prediction.

- Jerky/jitter motion

In videos, besides the above mentioned spatial artifacts, temporal impairments like jerky or jitter motion may arise. Jerkiness is often used to describe the “regular” frame freezing followed by a discontinuous motion. Generally it is caused by consistent frame dropping on the encoder side (transcoder) serving as a bit rate or terminal capability adaptation strategy. On the other hand, jitter often describes “irregular” frame dropping due to packet loss during signal transmission. The influences of these motion fluidity impairments to visual quality have been investigated in [52, 38] by subjective viewing tests. Usually NR metrics quantifying motion impairments [53, 92] will consider the following factors: the frame dropping duration (the shorter the better for visual quality), the frame dropping density (with the same amount of frame loss, the more scattered the better), and the motion activity (the lower the better). Compared with spatial distortion NR metrics, temporal distortion

NR metrics can provide better quality prediction that is more consistent with the judgments of the human observers.

- Statistics of natural images

Different from texts, cartoons, computer graphics, x-ray images, CAT scans, etc., natural images and videos possess their unique statistical characteristics, which has led to the development of many NSS (Natural Scene Statistics) models to capture them. As argued in [61], since distortions may disturb the statistics of natural scenes, a deviation of a distorted signal from the expected natural statistics can be used to quantify the distortion intensity. NR metrics based on this philosophy are few but very enlightening. For example in [61], this philosophy was clearly stated for the first time, and a NR metric was developed for JPEG2000 coded images, in which a NSS model [7] was employed to capture the non-linear dependencies of wavelet coefficients across scales and orientations. In [75], the authors proposed a technique for coarse-to-fine phase prediction of wavelet coefficients. It was observed that the proposed phase prediction is highly effective in natural images and it can be used to measure blurring artifacts that will disrupt this phase coherence relationship.

3 Subjective Evaluation Standard

Objective quality metrics are developed to approximate human perceptions, but up to this moment, no single metric can completely represent human response in visual quality assessment. As a result, subjective evaluation is still the only mean to fully characterize the performance of different systems. A standard procedure is therefore needed for fair and reliable evaluations.

Several international standards are proposed for subjective video quality evaluation for different applications. The most commonly referenced standard is the ITU-R BT.500 defined by International Telecommunication Union (ITU) [26]. ITU-R BT.710 [25] is an extension of BT.500 dedicated for high-definition TV. ITU-T P.910 [29] is another standard which defines the standard procedure of digital video quality assessment with transmission rate below 1.5Mbit/s. These standards provide guidelines for various aspects of subjective evaluations, such as viewing condition, test sequence selection, assessment procedures, and statistical analysis of the results. The Video Quality Expert Group (VQEG) also proposed several subjective evaluation procedures to evaluate the performance of different objective quality metrics [68, 69, 70]. These proposed methods share many similarities to the BT.500 and P.910 standards.

A brief description of various aspects of the subjective evaluation standards will be discussed in this section. This includes the general viewing condition, observer selection, test sequence selection, test session structure, test procedure, and post-processing of scores.

3.1 Viewing Condition

The general viewing condition defines a viewing environment that is most suitable for visual quality assessment. It minimizes the environment's effect on the quality of the image or video under assessment. In BT.500, two environments (laboratory environment and home environment) are defined. Laboratory viewing environments is intended for system evaluation in critical conditions. Home viewing environment, on the other hand, is intended for quality evaluation at the consumer side of the TV chain. The viewing conditions are designed to represent a general home environment. Table 1 tabulates some parameters of viewing conditions used in different standards.

Table 1. General Viewing Conditions

	Condition	BT.710	BT.500 (lab env)	BT.500 (home env)	P.910
a	Ratio of viewing distance to picture height	3	-	Function of screen height	1-8H
b	Peak luminance on the screen (cd/m^2)	150-250	-	200	100-200
c	Ratio of luminance of inactive screen to peak luminance	≤0.02	≤0.02	≤0.02	≤0.05
d	Ratio of luminance on the screen when displaying only black level in a completely dark room, to that corresponding to peak white	Approx. 0.01	Approx. 0.01	-	≤0.1
e	Ratio of luminance of background behind picture monitor to peak luminance of picture	Approx. 0.15	Approx. 0.15	-	≤0.2
f	Illumination from other sources	low	low	200lux	≤20lux
g	Chromaticity of background	D_{65}	D_{65}	-	D_{65}
h	Arrangement of observers	Within ±30° horizontally from the center of the display. The vertical limit is under study	Within ±30° relative to the normal (only apply to CRT, other display are under study)	Within ±30° relative to the normal (only apply to CRT, other display are under study)	-
i	Display size	1.4m (55 in)	-	-	-
j	Display brightness and contrast	-	Setup via PLUGE [28, 24]	Setup via PLUGE [28, 24]	-

3.2 Candidate Observer Selection

To obtain reliable assessments from observers, certain requirements on the selection of observers must be met. Firstly, their eyesight should be either normal or has been corrected to normal by spectacles. Prior to the test session, observer must be screened for normal eyesight, which includes normal visual acuity and normal color vision. Normal visual acuity can be checked by the Snellen or Landolt chart. A person is said to have normal acuity when he/she can correctly recognize the symbols on the standard sized Snellen chart 20/20 line when standing 20 feet from the chart. At 20 feet, the symbols on the 20/20 line subtend five minutes of arc to the observers, and the thickness of the lines and spaces between lines subtends one minute of arc. Normal color vision can be checked by specially designed charts, for instance, the Ishihara charts. Numbers or patterns of different colors are printed on plates with colored background. A person with color deficiency will see numbers or patterns different from the person with normal color vision. The observer is classified as having normal vision if he/she has no more than a certain number of miss-identification of the patterns. In [29], it is required that observers cannot make more than 2 mistakes out of 12 test plates.

Another requirement to the observers is that they should be familiar with the language used in the test. The observer must be able to understand the instruction and provide valid response with semantic judgment terms expressed in that language.

In a formal subjective evaluation experiment, the observers should be non-experts, i.e., they should not be directly involved in video quality assessment as part of their works, and should not be experienced assessors. At least 15 observers should be used to provide statistically reliable results. However, in early phase of the development stage, an informal subjective evaluation with 4 to 8 expert observers can provide indicative results.

3.3 Test Sequence Selection

No single set of test material can satisfy all kinds of assessment problems. Particular types of test material should be chosen for particular assessment problems. For example, in a study of the overall performances of two video coding systems for digital TV broadcast, sequences with a broad range of contents and characteristics should be used. On the other hand, if the systems being assessed are targeted for video conference on mobile network, head-and-shoulder sequences should be chosen.

For general purpose video system assessment, a broad range of contents should be chosen. Video sequences from movies, sports, music videos, advertisements, animations, broadcasting news, home videos, documentaries, etc., can be included in the test set. Different characteristics of the test sequences should also be considered. The test set should have different levels of color, luminance, motion, spatial details, scene cuts, etc.

When impairment evaluation is performed, the sequences with different levels of impairment should be produced in order to generate tractable results for performance analysis.

3.3.1 Spatial and Temporal Information

Sequences with different levels of spatial and temporal information should be chosen for general purpose assessment. The P.910 standard defined a method to measure these kinds of information in a video sequence.

The spatial information (*SI*) is a simple measure of the spatial complexity in the sequence. Each luminance plane F_n in a video frame at time n is first filtered with the Sobel filter to obtain a filter output $Sobel(F_n)$. The standard deviation over all pixels in the frame, $std_{space}[Sobel(F_n)]$, is then computed. This operation is repeated for every frame in the sequence. The *SI* is defined as the maximum standard deviation of all frames, i.e.,

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\}. \tag{13}$$

The temporal information (*TI*) measures the change of intensity of the frames over time. Intensity change can be caused by motions of objects or background, change of lighting conditions, camera noises, etc. To compute *TI*, first we compute $M_n(i,j)$, the difference between pixel values of the luminance plane at the same location but in successive frames, i.e.,

$$M_n(i,j) = F_n(i,j)\text{-}\ F_{n\text{-}1}(i,j), \tag{14}$$

where $F_n\ (i,j)$ is the pixel value at i^{th} row and j^{th} column in the frame at time n. *TI* is then defined as

$$TI = max_{time}\{std_{space}[M_n(i,j)]\}. \tag{15}$$

If there are scene cuts in the sequence, two values of *TI* may be computed, one for sequence with scene cut, and one for sequence without scene cut. In addition, scenes with high *SI* are generally associated with relatively high *TI*, since motion in complex scenes usually results in large differences in intensity in successive frames.

3.4 Structure of Test Session

The maximum duration of a single test session is 30 minutes. A break of a few minutes should be given to the observer before the next test session starts. At the beginning of the first test session, about five stabilizing presentations should be introduced to stabilize assessor's opinions. Scores obtained from these presentations should not be taken into account. For subsequent sessions, about three stabilizing sequences should be presented at the beginning of each session. In addition, several training sequences should be introduced before the first session starts in order to familiarize the observers with the assessment procedure. The observers can ask questions regarding the assessment after the training sequences. The whole process is illustrated in Fig. 14.

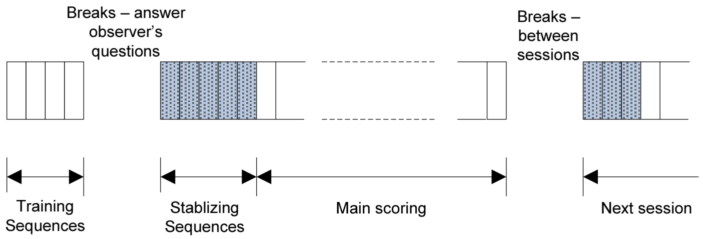


Fig. 14. Test session structure

A pseudo random order of the sequences should be used for each assessor. The order can be derived from Graeco-Latin squares or other means. This can reduce the effects of tiredness or adaptation on the grading process. Also, the same picture or sequence should not be used for consecutive presentations, even with different levels of distortions, to prevent ambiguity.

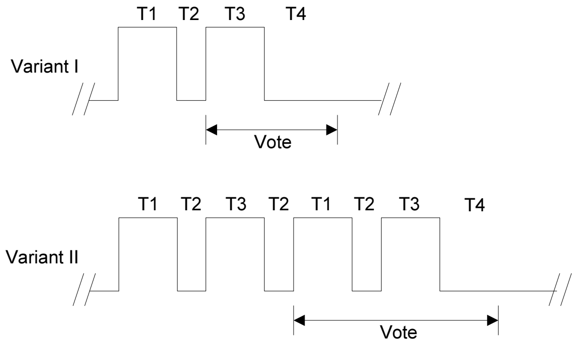
3.5 Assessment Procedure

There are numerous Full-Reference and No-Reference assessment procedures targeted for different problems. In the full-reference case, the two commonly used evaluation procedures are double-stimulus impairment scale (DSIS) and double-stimulus continuous quality-scale (DSCQS). DSIS is used for failure characterization, e.g., identifying the effect of certain impairment introduced in the video encoding or transmitting process. DSCQS is used when an overall system evaluation is need. No-Reference assessment can be performed using the single stimulus (SS) method. These three assessment procedures will be briefly described in this section. Readers should refer to [26] for more detailed descriptions of different assessment procedures.

3.5.1 Double-Stimulus Impairment Scale

DSIS method is a cyclic assessment procedure. An unimpaired reference image or sequence is presented to the observer first, followed by the same signal with impairments added by the system under test. The observer is asked to vote on the impaired one while keeping in mind the reference.

The structure of presentations is shown in Fig. 15. There are two variants of the structure. In variant I, the reference and the impaired picture or sequence are presented only once. In variant II, the reference and impaired material are presented



T1 = 10s	Reference sequence
T2 = 3s	Mid-grey
T3 = 10s	Test condition
T4 = 5-11s	Mid-grey

Fig. 15. Presentation structure of DSIS

twice. When the impairment is very small or when the presented material is a video sequence, variant II is preferred though it is more time consuming. The observers should be asked to look at the reference and impaired material for the whole duration of T1 and T3, respectively. A mid-gray screen (T2) is shown between T1 and T3 for a duration of about 3 seconds. If variant I is used, voting period starts immediately after the impaired material is presented. If variant II is used, voting period starts when the pair of reference and impaired material is shown at the second time. In either variant, a 5 to 11 seconds mid-gray (T4) is displayed before the next presentation.

The scores are recorded by using the five-grade impairment scale. The five grades and their distortion level descriptions are as follow:

5	imperceptible
4	perceptible, but not annoying
3	slightly annoying
2	annoying
1	very annoying

A form with clearly defined scale, numbered boxes or some other means to record the grading should be provided to the observers for score recording.

The impairments of the test material should have a broad range so that all five grades in the grading scale should be chosen by the majority of observers, and the impairments should be evenly distributed among the five grading levels.

3.5.2 Double-Stimulus Continuous Quality Scale

The double stimulus continuous quality scale (DSCQS) method is an effective evaluation method for overall system performance. The structure of the presentation is somewhat similar to that of DSIS, except that the pair of reference and impaired material is presented in random order. The observer does not have the knowledge of the display order, and he/she will give scores for both reference and impaired images or sequences.

There are two variants of the DSCQS presentation structure. In variant I, only one observer participates in a test session. The observer is free to switch between signal A and B until he/she is able to determine a quality score for each signal. This process may be performed two or three times for duration of up to 10 seconds. In variant II, at most three observers can assess the material simultaneously. If the material is a still picture, each picture will be displayed for 3 to 4 seconds with five repetitions. For video sequences, the duration of each sequence is about 10 seconds with two repetitions. This presentation structure is illustrated in Fig.16.

The observers are asked to assess the overall picture quality of each presentation by inserting a mark on a vertical scale. An example is shown in Fig. 17. The vertical scales are printed in pairs since both the reference and the impaired sequence must be assessed. The scales provide a continuous rating system for score from 0 to 100, which is different from the five-grade scale used in DSIS. They are divided into five equal lengths and associated descriptive terms are printed on the left of the first scale as general guidance to the observer. To avoid confusion, the

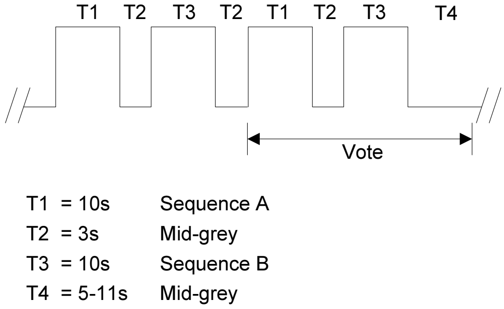


Fig. 16. Presentation structure of DSCQS

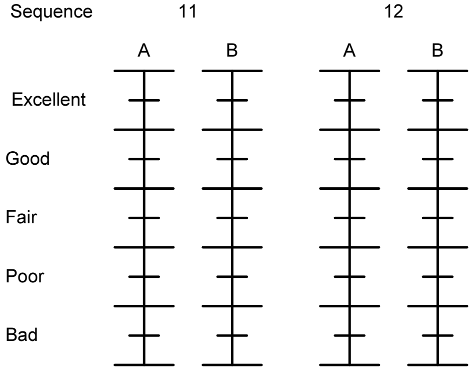


Fig. 17. Grading scale score sheet for DSCQS

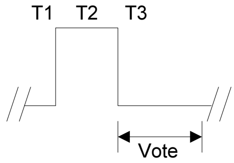
observer should use pen with color different from the printed scale. Electronic scoring tools can be used only if its display does not compromise the viewing conditions listed in Table 1.

3.5.3 Single Stimulus Methods

In single stimulus methods, the image or sequence is assessed without the reference source, and a presentation consists of three parts: a mid-gray adaptation field, a stimulus, i.e., the image or sequence being assessed, and a mid-gray post-exposure field. The durations of these parts are 3, 10, and 10 seconds, respectively. The voting of the stimulus can be performed during the display of the post-exposure field. The overall structure is illustrated in Fig. 18.

Depending on the applications, grading can be recorded by the 5-point impairment scale used in DSIS, an 11-grade numerical categorical scale described in ITU-R BT.1082 [23], or the continuous scale used in DSCQS.

Another variant of SS is that the whole set of test stimuli are presented three times, which means that a test session consists of three presentations. Each of



T1 = 3s Mid-gray adaptation
T2 = 10s Stimulus
T3 = 10s Mid-gray post-exposure

Fig. 18. Presentation structure of SS

them includes all the images or sequences to be tested only once. The first presentation serves as a stabilizing presentation. The scores obtained from this presentation will not be taken into account. The score of each image or sequence is the mean score obtained from the second and third presentation. The display orders of the images or sequences are randomized for all three presentations.

3.6 Post-Processing of Scores

After obtaining the scores from the observers, these data must be statistically summarized into meaningful form for analysis. In addition, observers should be screened and statistically unreasonable results should be discarded. A relationship between an objective measurement of the picture quality and the subjective score can also be deduced from the data obtained.

The following analysis is applicable to the results from the DSIS, DSCQS, and other methods which use numerical scales for grading. In the first case, the impairment is rated on a five-point or multi-point scale and the score range is from 1 to 5. In the second case, continuous rating scales are used and the results have integer values between 0 and 100.

3.6.1 Mean Scores and Confidence Interval Calculation

The common representation of the scores is the mean score and confidence interval. Let L be the number of presentations in the test, J be the number of test conditions applied to a picture or sequence, K be the number of test images or sequences, R be the number of repetitions of a test condition applied on a test picture or sequence. The mean score, $\bar{u}_{jkr}$, for each of the presentations is defined as

$$\bar{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \,, \tag{16}$$

where u_{ijkr} is the score of observer i for test condition j, sequence/image k, repetition r; and N is the number of observers. The overall mean scores, $\bar{u}_j$ and $\bar{u}_k$, could be calculated for each test condition and each test image or sequence in a similar manner.

The confidence intervals should be presented in addition to the mean scores to provide more information about the variability of the results. The confidence interval is derived from the standard deviation and sample size. [26] proposed to use the 95% confidence interval, which is defined as $\left\lfloor \bar{u}_{jkr} - \delta_{jkr}, \bar{u}_{jkr} + \delta_{jkr} \right\rfloor$ where:

$$\delta_{jkr} = 1.96 \frac{S_{jkr}}{\sqrt{N}}, \tag{17}$$

and the standard deviation for each presentation, S_{jkr}, is defined as:

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{\left(\bar{u}_{jkr} - u_{ijkr}\right)^2}{(N-1)}}\,. \tag{18}$$

The 95% confidence interval indicates that with a probability of 95%, the mean score will be within the interval if the experiment is repeated for a large number of times. As more samples are available, the confidence interval range gets smaller and the mean score becomes more reliable.

3.6.2 Screening of Observers

Sometimes scores obtained from certain observers may deviate from the distribution of the normal scores significantly. This kind of observers must be identified and their scores discarded from the test. The β_2 test is suggested in BT.500 to accomplish such task.

For each test presentation, we first compute the mean, $\bar{u}_{jkr}$, standard deviation, S_{jkr}, and kurtosis coefficient, β_{2jkr}, where β_{2jkr} is given by:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \text{ with } m_x = \frac{\sum_{i=1}^{N} \left(u_{ijkr} - \bar{u}_{jkr}\right)^x}{N}. \tag{19}$$

If $\beta_{2\mathrm{jkr}}$ is between 2 and 4, the distribution of the score can be assumed to be normal. Now for each observer, i, we need to find the number of score entries that lie outside of the score distribution of each test presentation. The valid range of distribution of each test presentation is defined as $\bar{u}_{jkr} \pm 2S_{jkr}$ if the distribution is normal. For non-normal distribution, the valid range is defined as $\bar{u}_{jkr} \pm \sqrt{20}S_{jkr}$. Let P_i and Q_i be the number of times that the score from observer i is above and below the valid range, respectively. P_i and Q_i can be computed by the following procedure:

```
for $j, k, r = 1, 1, 1$ to $J, K, R$
  if $2 \le \beta_{2\mathrm{jkr}} \le 4$, then:
    if $u_{ijkr} \ge \bar{u}_{jkr} + 2S_{jkr}$ then $P_i = P_i + 1$
    if $u_{ijkr} \le \bar{u}_{jkr} - 2S_{jkr}$ then $Q_i = Q_i + 1$
```

else:
 if $u_{ijkr} \geq \bar{u}_{jkr} + \sqrt{20} S_{jkr}$ then $P_i = P_i$+1
 if $u_{ijkr} \leq \bar{u}_{jkr} - \sqrt{20} S_{jkr}$ then $Q_i = Q_i$+1

where *J*, *K*, and *R* have the same meaning as in Section 3.6.1. After computing P_i and Q_i for observer *i*, if the following two conditions are met, then observer *i* will be rejected.

Condition 1: $\dfrac{P_i + Q_i}{J \cdot K \cdot R} > 0.05$

Condition 2: $\left| \dfrac{P_i - Q_i}{P_i + Q_i} \right| < 0.3$

The observer screening procedure should not be applied to the results of a given experiment more than once. In addition, it should be restricted to the experiment when there are relatively few observers (e.g., fewer than 20) participating the experiment and all of them are non-experts.

3.6.3 Relationship between the Mean Score and the Objective Measure of a Picture Distortion

When evaluating a relationship between the mean scores and a type of impairment at different levels, or between the mean scores and some objective measurements of distortion, it will be useful if the relationship can be represented by a simple continuous function with the mean score as the dependent variable.

In [26], the symmetric logistic function and a non-symmetric function are introduced to approximate the relationship. For both cases, the mean score *u* must first be normalized by taking a continuous variable *p* so that

$$p = \frac{(\bar{u} - u_{\min})}{(u_{\max} - u_{\min})}, \tag{20}$$

where u_{min} is the minimum score available on the *u*-scale for the worst quality; and u_{max} is the maximum score available on the *u*-scale for the best quality. The normalized mean score *p* can be estimated by a symmetric logistic function. Let $\hat{p}$ be the estimate of *p*. The function $\hat{p} = f(D)$, where *D* is the distortion parameter, can now be approximated by a judiciously chosen logistic function, as given by the general relation

$$\hat{p} = f(D) = \frac{1}{1 + e^{(D - D_M) \cdot G}}, \tag{21}$$

where D_M and *G* are constants and *G* may be positive or negative. To solve for D_M and *G*, we define

$$I = \frac{1}{p} - 1, \tag{22}$$

and its estimate

$$\hat{I} = \frac{1}{\hat{p}} - 1. \tag{23}$$

Combining (1.21) and (1.23), we obtained

$$\hat{I} = e^{(D - D_M) \cdot G}. \tag{24}$$

Let J be the natural log of I, and $\hat{J}$ be the natural log of $\hat{I}$, i.e.,

$$\hat{J} = \log_e I = (D - D_M) \cdot G. \tag{25}$$

A linear relationship between $\hat{J}$ and D is established. D_M and G can then be found by minimizing ε, the mean squared estimation error between $\hat{J}$ and J, which is defined as

$$\varepsilon = \frac{1}{N} \sum_{i=1}^{N} \left(J_i - \hat{J}_i \right)^2. \tag{26}$$

The simple least square method can be used to find the optimal D_M and G.

The symmetrical logistic function is particularly useful when the distortion parameter D can be measured in a related unit, e.g., the *S*/*N* (dB). If the distortion parameter was measured in a physical unit d, e.g., a time delay (ms), then the relationship between $\hat{p}$ and d can be defined as

$$\hat{p} = \frac{1}{1 + (d / d_M)^{1/G}}. \tag{27}$$

This is a non-symmetric approximation of the logistic function. Similar to the case of logistic function, we define $\hat{I}$ as

$$\hat{I} = \frac{1}{\hat{p}} - 1, \tag{28}$$

and the estimated $\hat{J}$ is

$$\hat{J} = \log(\hat{I}) = \frac{1}{G} \log\left(\frac{d}{d_M} \right). \tag{29}$$

The optimal values of d_M and G can be solved by minimizing the mean squared error ε defined in (26) using the Levenberg-Marquardt algorithm.

4 An Application of Quality Comparison: AVS versus H.264

4.1 Background

One important application of subjective quality evaluation is to compare the performance of two video encoding systems. Since human observers are the final judge of the quality of the encoded video, a subjective evaluation process is necessary to obtain a comprehensive understanding of the performance of the systems being tested.

In this section, we describe a comparison between two encoding systems, namely, the H.264/Advanced Video Coding (AVC) and the Audio and Video Coding Standard (AVS), based on objective distorion measurements and subjective evaluation. H.264/AVC is the most recent video coding standard jointly developed by the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) [21]. Various profiles are defined in the standard to suite different applications. For example, simple baseline profile provides the basic tools for video conferencing and mobile applications that run on low-cost embedded systems, while the main profile is used in general applications. More complex High, High 10, and High 4:2:2 profiles are introduced in the Fidelity Range Extension (FRext) of H.264 to further improve the coding efficiency for high-definition (HD) video and studio quality video encoding [62]. New coding tools, e.g., 8×8 block size transform, support of different chroma format, and precision higher than 8 bits, are utilized in these profiles.

AVS is a new compression standard developed by AVS Workgroup of China [3, 4, 22]. AVS Part 2 (AVS-P2) is designed for high-definition digital video broadcasting and high-density storage media. It is published as the national standard of China in February, 2006. Similar to the H.264/AVC, AVS is a hybrid DPCM-DCT coding system with compression tools like spatial and temporal prediction, integer transform, in-loop deblocking filter, entropy coding, etc [93]. The target applications of AVS include HD-DVD and satellite broadcast in China. AVS-P2 also introduced the X profile, which is designed for high quality video encoding. Several new tools are introduced: Macroblock-level adaptive frame/field coding (MBAFF), adaptive weighting quantization, adaptive scan order of transform coefficients, and arithmetic coding of most syntax elements.

In general, the structure of AVS and H.264 are very similar. The major difference is that many components in the AVS are less complex than the H.264 counterpart. For example, AVS utilizes only 4 different block sizes in motion estimation, while H.264 uses 7 block sizes. AVS has only 5 luminance block intra-prediction modes, compared to 13 modes in H.264. In addition, AVS utilizes a simpler in-loop deblocking filter, shorter tap filter for sub-pixel motion estimation, and other techniques to reduce the complexity of the encoder. The AVS encoder requires only about 30% of the computation load of H.264 for encoding, but it is able to achieve similar coding efficiency.

Objective comparisons between H.264-main profile and AVS-base profile were reported in several articles, e.g., in [93, 16, 72]. The results generally show that for smaller frame sized sequences, such as QCIF, and CIF, H.264 has a slight advantage over AVS. For SD and HD video, AVS and H.264 are similar in their rate distortion performance. However, no comparison has been made between the AVS-X profile and the H.264-High profile. In this section, the results of the objective and subjective performance comparisons between these two profiles will be presented. Section 4.2 describes the setup of the experiment. Objective and subjective evaluation results are presented in Section 4.3 and 4.4, respectively. In Section 4.5, we compare the accuracy of PSNR, and two other objective quality metrics that correlate better to human perception than PSNR, i.e., the Structural Similarity (SSIM) index and the Video Quality Metric (VQM), using the results obtained from subjective evaluation.

4.2 Test Setup

The performance comparison of AVS-X profile and H.264-high profile is divided into two parts: objective comparison and subjective comparison. In objective comparison, rate-distortion performance in terms of PSNR and bit rates is first used as the evaluation metric. Subjective evaluation of visual quality is also performed to compare their coding performances as perceived by human observers.

4.2.1 Sequence Information

Table 2 shows all the video sequences used in this comparison. Since the target of this comparison is for high-fidelity video, only HD video sequences were used in the test. For 1280×720 progressive (720p) sequences, the target bit rates used for the test were 4, 8, 10, and 15 Mbps, and the frame rate was 60Hz. Both 1920×1080 progressive (1080p) and 1920×1080 interlace (1080i) sequences were encoded at target bit rates of 6, 10, 15, and 20 Mbps, and at a frame rate of 25Hz.

Table 2. Test Sequences Used in Test

720p	1080p	1080i
City	*PedestrianArea*	*NewMobileCalendar*
Crew	*Riverbed*	*Parkrun*
Harbour	*Rushhour*	*Shields*
ShuttleStart	*Sunflower*	*StockholmPan*
SpinCalendar	*ToysCalendar*	*VintageCar*

4.2.2 Encoder Setting

The JM 14.0 reference H.264/AVC encoder and the rm6.2h reference AVS encoder are used to encode the test video sequences. The IBBPBBPBBP… GOP structure was used, with intra-frames inserted in every 0.5 sec, i.e., one intra-frame in every 30 frames for 720p, and in every 12 frames for 1080p and 1080i. Table 3 shows the general settings of the encoders. Note that due to memory limitation, only 2 reference frames were used when interlace videos are encoded with the H.264 encoder.

Table 3. Encoder Parameter Settings

Setting	H.264	AVS-P2
Encoder version	JM 14.0	rm6.2h
Profile	high	X
Number of reference frame	4 (2 for 1080i)	2
Block size	16×16, 16×8, 8×16, 8×8, 8×4, 4×8, 4×4	16×16, 16×8, 8×16, 8×8
Fast ME	Enabled	Enabled
ME search range	32	32
RD Optimization	Enabled	Enabled
Interlace mode	PAFF	PAFF
Loop filter	Enabled	Enabled
Adaptive scan	-	Enabled
Adaptive filter	-	Disabled

4.2.3 Subjective Test Setup

The subjective assessment was performed in a studio room with lighting condition satisfying the lab environment requirement of the ITU-R BT.500 standard, which is also briefly described in Section 3. The display monitor is a 65” Panasonic plasma display (TH-65PF9WK) and the viewing distance is 3 times the picture height. Background illumination has a D65 chromaticity.

Thirty-five non-expert observers participated in the subjective test, and about half of them were male. All of them did not work in video processing related jobs, and were not involved in any video quality assessment within the past four months. Their eyesight was either normal or had been corrected to be normal with spectacles.

Each observer compared 39 pairs of “reference” (H.264) and “processed” (AVS) sequences (13 pairs each for 720p, 1080p, and 1080i). The double-stimulus continuous quality scale (DSCQS) test method as described in Section 3.5.2 was used for this subjective test.

Note that the uncompressed sequences are not used as the references as in normal practice because we want to compare the visual quality of H.264 and AVS sequences directly. Direct comparison allows the observers to immediately identify the small differences in visual quality and record the scores.

4.3 Rate-Distortion Performance Comparisons Using PSNR, Bitrates

The average PSNR change (ΔPSNR) and bit rate change (ΔBitrate) computed by the method described in [6] are used to measure the objective performance of the two coding systems. The ΔPSNR is a measure of the difference in PSNR of the target and reference systems under the same bit rate range. Similarly, ΔBitrate measures the difference of bit rates under the same PSNR range. A sequence is encoded at four different bit rates by both the "reference" and the "target" systems, then ΔPSNR and ΔBitrate of the target system are computed from these rate-distortion points as illustrated in Fig. 19. In general, a ΔPSNR of 0.3dB is equivalent to a ΔBitrate of approximately 5%.

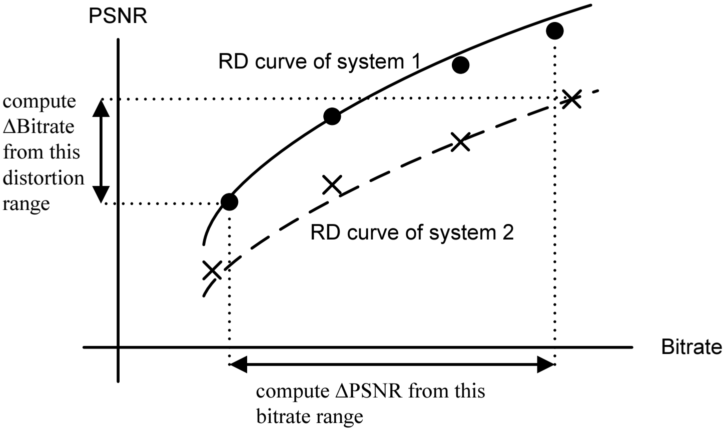

PSNR
RD curve of system 1
compute ΔBitrate from this distortion range
RD curve of system 2
Bitrate
compute ΔPSNR from this bitrate range

Fig. 19. Illustration of RD performance evaluation by ΔPSNR and ΔBitrate

The advantage of using this performance evaluation method is that it evaluates the performance of a system at multiple bit rates, so that a better understanding of the system under different bit rates can be acquired. In addition, both bit rate and the PSNR of encoded video sequences can be computed easily. Because of these advantages, this method is commonly used in encoding system performance comparison. In our experiment, sequences in H.264 format are defined as the reference. The ΔPSNR and ΔBitrate of all the test sequences encoded by AVS are shown in Table 4.

For the 720p sequences, the average ΔBitrate of all sequences is 0.71%, which implies that the overall coding efficiency of AVS and H.264 is very similar. However, the ΔBitrate fluctuates from -8.98% (*Crew*) to 10.72% (*SpinCalendar*). The coding efficiency of AVS depends quite heavily on the content of the sequence. For the 1080p sequences, the average ΔBitrate is -2.31%. AVS seems to have a slight advantage in encoding these sequences. The range of ΔBitrate varies only from -7.81% to 0.78%. The variation is less than that for 720p sequences. On the contrary, AVS has an average ΔBitrate of 5.48% for 1080i sequences, which means the coding efficiency is lower than that of H.264. In addition, a large variation in ΔBitrate is observed, with a range from -4.56% to 19.77%. The rate-distortion (RD) curves of several sequences are shown in Fig 20. The RD performance we obtained is similar to those reported in [16]. Although they were using older reference encoders, they also reported that AVS performs slightly worse on sequences like *City* and *SpinCalendar*, and slightly better on sequences like *Harbour* and *Crew*.

Table 4. ΔPSNR and ΔBit rate for all test sequences

Size	Sequence	ΔPSNR	ΔBitrate	*SI*
720p	*City*	-0.247	10.45%	77.41
	Crew	0.220	-8.98%	72.21
	Harbour	0.109	-3.60%	92.3
	ShuttleStart	0.092	-5.01%	35.21
	SpinCalendar	-0.205	10.72%	103.27
	Average	-0.006	0.71%	-
1080p	*PedestrianArea*	0.016	-1.10%	37.11
	Riverbed	0.402	-7.81%	39.46
	Rushhour	-0.040	-2.16%	26.74
	Sunflower	0.005	0.78%	39.39
	ToysCalendar	0.001	-1.26%	54.7
	Average	0.077	-2.31%	-
1080i	*NewMobileCalendar*	-0.398	19.77%	73.44
	Parkrun	-0.194	5.18%	123.01
	Shields	-0.188	10.08%	60.02
	StockholmPan	0.044	-3.08%	73.95
	VintageCar	0.090	-4.56%	58.3
	Average	-0.129	5.48%	-
	Overall Average	-0.019	1.29%	-

The RD performance shows that the content of the sequence has certain impact on the coding efficiency. It seems that H.264 performs better in sequences with lots of textures, such as *City*, *SpinCalendar*, *NewMobileCalendar*, and *Shields*. ΔBitrate of AVS are more than 10% in these sequences. The right-most column in Table 4 shows the spatial information (*SI*) of all sequences. *SI* is a simple measure of spatial texture introduced in section 3.3. Fig. 21 shows the relationship between ΔBitrate and *SI* for different sequences. Although there is no linear relationship between bitrate and *SI*, we can still see that positive ΔBitrate appears more frequently on sequences with high *SI*, i.e., highly textured sequences. In addition, the Pearson's correlation coefficient between ΔBitrate and SI is 0.402, suggesting that there is a positive relationship between *SI* and ΔBitrate. For other types of video, AVS is able to achieve a higher efficiency than H.264, e.g., *Crew* and *Riverbed*. The usage of more than 2 reference frames and more complex interpolation filter in H.264 does not seem to give a significant improvement in coding efficiency for these sequences. The simpler coding tools in AVS are sufficient to give similar efficiency compared with H.264.

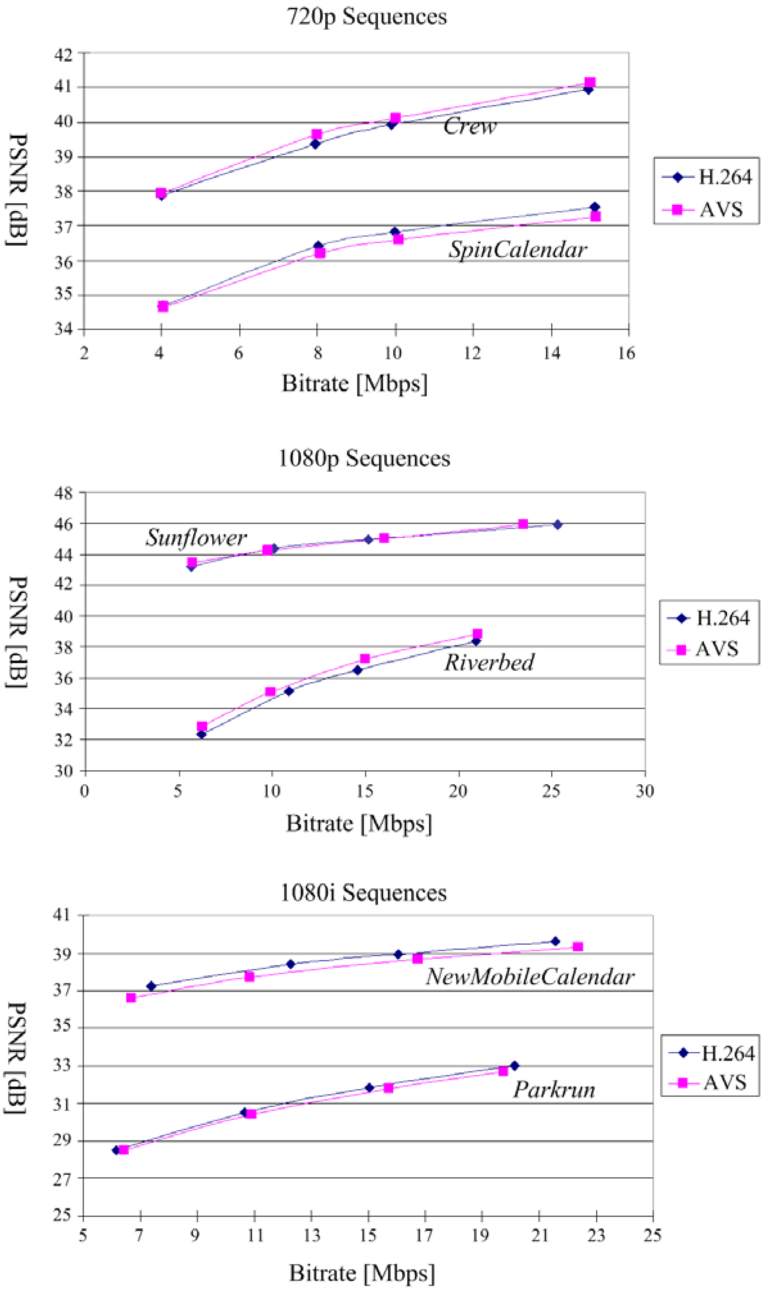


Fig. 20. RD Curves of some 720p, 1080p, and 1080i sequences

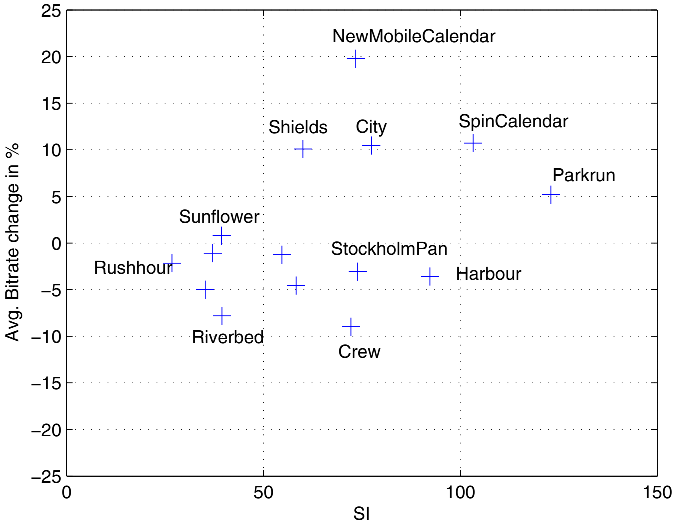


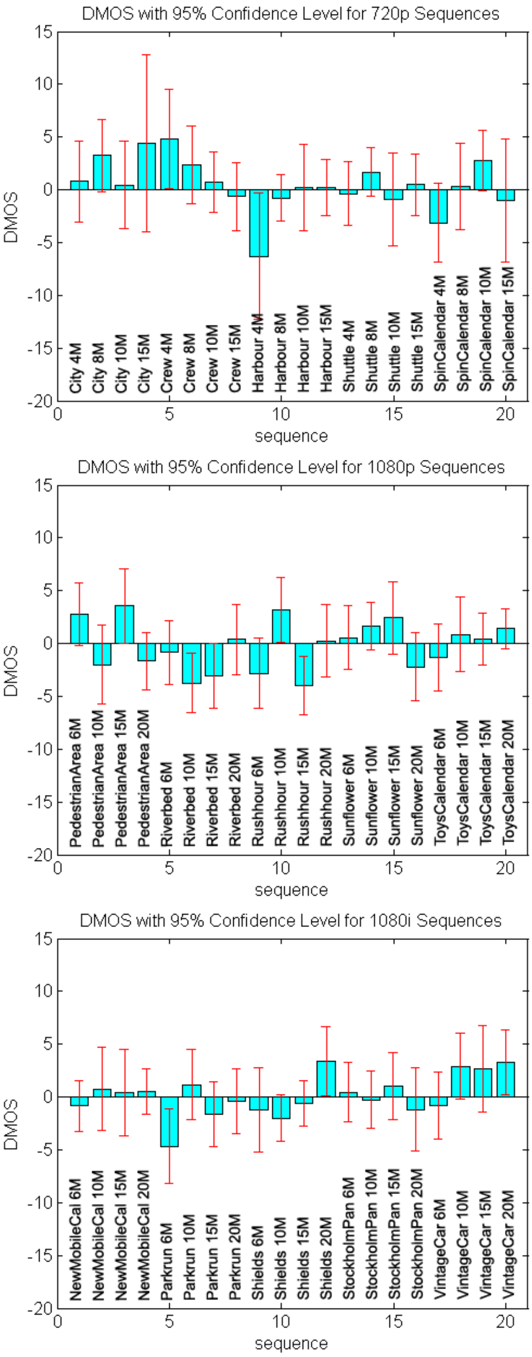
Fig. 21. ΔBitrate and SI relationship for different sequences

4.4 Subjective Evaluation Results

The scores given by the observers in the subjective test are used to evaluate the subjective quality of the sequences. The mean opinion score (MOS) of each sequence is first computed. Then the difference mean opinion score (DMOS), i.e., the difference in MOS between the AVS and the H.264 sequences, is used to compare the subjective quality of H.264 and AVS. A positive DMOS implies AVS has a better subjective quality than H.264. The DMOS for the all sequences, along with the 95% confidence level, are shown in Fig. 22.

The average DMOS of all sequences is 0.13 with standard deviation of 2.26. Note that the full range of the score is 100. This indicates that the overall visual quality of AVS and H.264 are very similar in many sequences. Fig. 23 shows the average DMOS of each sequence for the four bit rates used. The magnitudes of the average DMOS are all less than 3, which is also very small. We also computed the average DMOS of sequences with the same bit per pixel. The results are shown in Fig. 24. It is clear that on the average, AVS and H.264 have very similar performance in visual quality at different bit rates.

For 720p sequences, most of the sequences encoded by AVS have visual quality similar to those encoded by H.264. The DMOS scores range only from -5 to 5 except for one sequence: *Harbour* at 4 Mbps, which has a DMOS of -6.29. Although the PSNR of the AVS encoded sequence is only 0.17dB lower than the one encoded by H.264, the distortion is more obvious than other sequences. In fact, for *City* at 8 Mbps, the PSNR of AVS encoded sequence is 0.33 dB lower than H.264, but the DMOS is 3.24. As mentioned in Section 1.4.3, AVS performs worse in terms of RD performance for sequences containing highly-textured area.



DMOS with 95% Confidence Level for 720p Sequences

DMOS with 95% Confidence Level for 1080p Sequences

DMOS with 95% Confidence Level for 1080i Sequences

Fig. 22. DMOS for sequences in different frame sizes

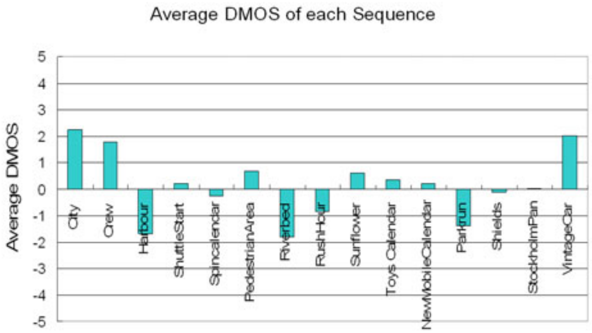


Fig. 23. Average DMOS of each sequence

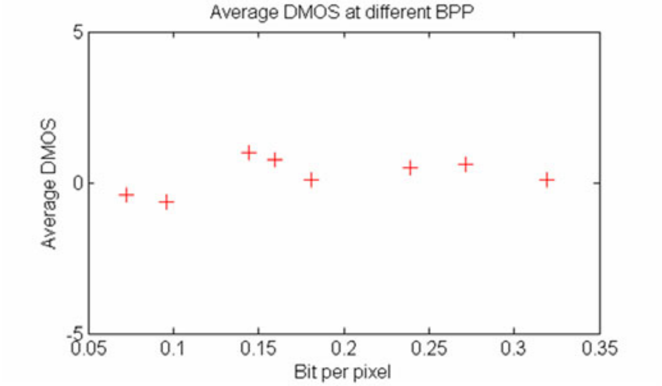


Fig. 24 Average DMOS of sequences with the same BPP

However, the subjective test results show that the visual quality is not affected by the reduction in coding efficiency. The DMOS for textured sequences, such as *City* and *SpinCalendar*, are all close to zero, even when their ΔBitrate are over 10%.

The DMOS obtained for 1080p and 1080i sequences have similar trend to that of 720p sequences. Although all sequences have non-zero DMOS, the magnitudes are all smaller than 5. No obvious difference in visual quality was observed from these sequences. Again, textured sequences such as *NewMobileCalendar* and *Shields* have DMOS close to zero. The ΔBitrate for *NewMobileCalendar* is close to 20% but the difference is visually unobservable. This phenomenon can be explained by the properties of the HVS. The spatial-temporal contrast sensitivity function of human eyes exhibits a non-separable band-pass characteristic with a low sensitivity at high spatial or temporal frequency [14]. Distortions in textured area, especially in moving regions, are less visible to human eyes than those in smooth and slow moving regions. As a result, even when ΔBitrate are over 10% for many sequences, the DMOS for them are all close to zero and the visual qualities are the same. The results clearly show that the commonly used RD performance based on PSNR and bit rates is not a good visual quality performance indicator for video coding systems.

4.5 The Use of PSNR, SSIM, and VQM in Quality Assessment

As discussed in Section 2, numerous objective quality metrics have been proposed to measure the visual quality of images or videos. Compared with PSNR, these metrics generally have a higher correlation with the subjective evaluation results. If an objective quality metric can accurately predict the perceived visual quality, then the difference of metric scores should be able to predict the visual difference of two sequences encoded by different systems. This can be illustrated in Fig. 25. Let D_{ref} and D_{tar} be the full-reference distortions computed for sequences encoded by the reference and target systems, respectively. The difference between D_{ref} and D_{tar}, denoted by D_{sys}, should also have a high correlation to the DMOS obtained from subjective evaluation. In this section, two popular metrics, SSIM and VQM, are tested to see if they can model human perception of distortion better than the conventional PSNR. Since SSIM is a metric designed for image, the average SSIM of all frames of the encoded sequence is used in our experiment.

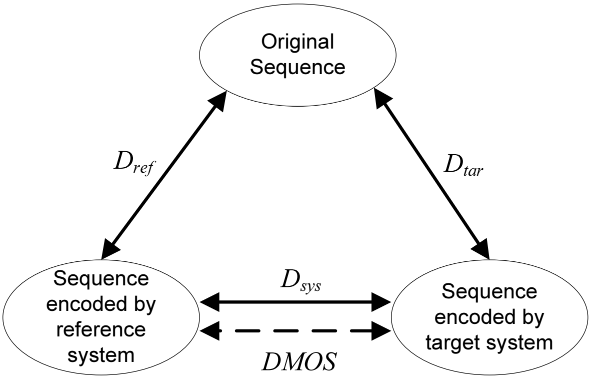

Original Sequence
D_{ref}
D_{tar}
Sequence encoded by reference system
D_{sys}
DMOS
Sequence encoded by target system

Fig. 25. Illustration of relationship of original sequence and sequence encoded by reference and target systems

The performance of an objective quality metric can be evaluated by its correlation to the MOS from subjective evaluation. Three measurements of correlation: Pearson's correlation coefficient (PCC), root mean square error (RMSE), and Spearman's rank order correlation coefficient (SROCC) are used for our evaluation. The PCC between two data sets, *X* and *Y,* is defined as

$$PCC(X,Y)=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}} \tag{30}$$

The RMSE between X and Y can be given as

$$RMSE(X,Y)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-Y_i\right)^2} \tag{31}$$

For the two sets of data X and Y, element X_i and Y_i are converted to rankings x_i and y_i, and SROCC is defined as the PCC of the ranks of X and Y.

A nonlinear mapping between the objective and subjective scores can be applied so that the objective metric can better predict the subjective quality. In both VQEG Phase-I and Phase-II testing and validation, nonlinear mapping is allowed, and the performance of the objective metric is computed after the mapping [60]. The mapping of an objective quality score x is mapped to $Q(x)$ by (32) and (33).

$$Q(x)=\beta_1\cdot\text{logistic}\left(\beta_2,\left(x-\beta_3\right)\right)+\beta_4\cdot x+\beta_5 \tag{32}$$

$$\text{logistic}(\tau,x)=\frac{1}{2}-\frac{1}{1+e^{\tau\cdot x}} \tag{33}$$

The parameters $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ can be found by minimizing the sum of squared difference between the mapped score $Q(x)$ and the corresponding MOS or DMOS. An illustration of the effect of this nonlinear mapping is shown in Fig 26.

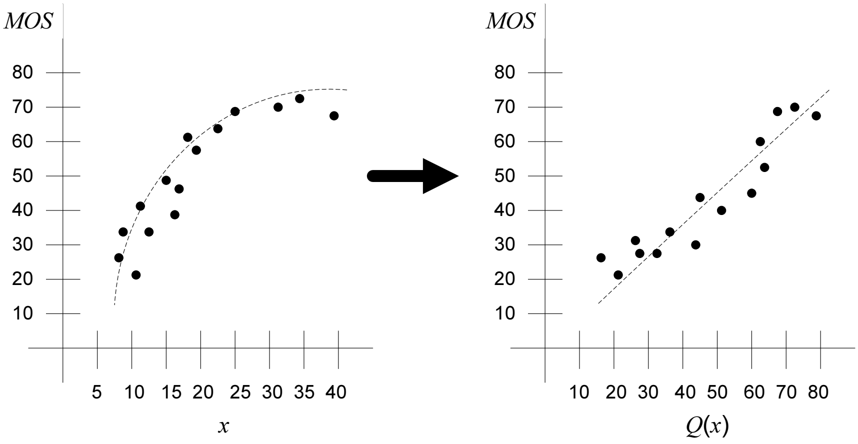


Fig. 26. Illustration of the effect of nonlinear mapping

The PSNR, SSIM, and VQM scores are first computed for all the H.264 and AVS encoded sequences. Then the score differences for the corresponding sequence pairs (that have the same content but encoded by different standards) denoted by DPSNR, DSSIM, and DVQM, are computed. Finally, the PCC, RMSE, and SROCC which measure the correlation between subjective quality scores DMOS and the nonlinearly mapped DPSNR, DSSIM and DVQM are computed. The results are shown in Table 5. VQM generates the highest PCC and SROCC, and the smallest RMSE, which indicates that VQM's correlation to the DMOS is the highest among the three quality metrics. SSIM outperforms PSNR in PCC and RMSE, but its SROCC is slightly lower than that of PSNR.

Table 5. PCC, RMSE, SROCC of three objective distortion metrics

	PCC	RMSE	SROCC
PSNR	0.1396	2.477	0.1718
SSIM	0.3421	2.350	0.1206
VQM	0.4428	2.243	0.1842

The nonlinearly mapped quality scores and the DMOS are plotted in Fig. 27 to Fig. 29. Even after the nonlinear mapping, we still cannot observe strong correlation between DMOS and the objective quality scores. The PCC and SROCC of all distortion metrics are relatively small, to be precise, below 0.5 and 0.2, respectively. This is quite different from the experimental results of other related works, e.g., [55, 77] where the correlation values are above 0.8 typically. The small correlation values are mainly due to the unperceivable differences between the H.264 and AVS encoded sequences. For many sequences, the DMOS is zero while the differences in the objective metrics, i.e., DPSNR, DSSIM, and DVQM, are non-zero. This is different from the experiments in [55, 77], where the visual quality differences between the reference and the distorted images or videos are much more obvious, and therefore the DMOS are usually non-zero. The large number of zero DMOS results in the bad performances of the tested objective quality metrics. Although VQM has comparably better performance, it still cannot accurately predict the DMOS which correspond to nearly unperceivable differences. Therefore for comparing two systems with similar performances, subjective evaluation is still a valuable tool.

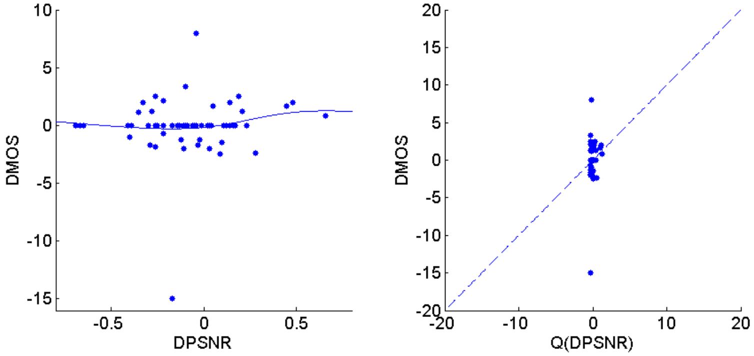


Fig. 27. DPSNR and nonlinear mapped DPSNR vs DMOS

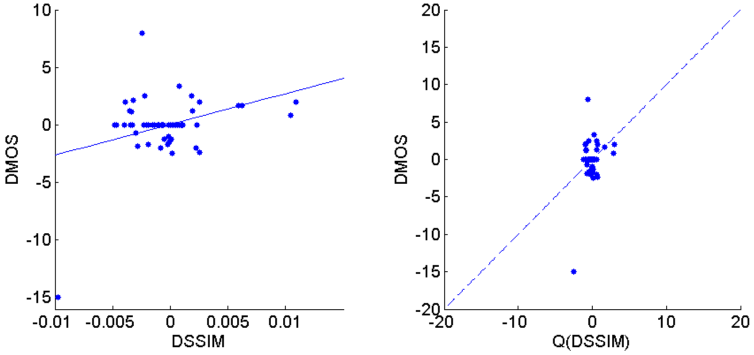


Fig. 28. DSSIM and nonlinear mapped DSSIM vs DMOS

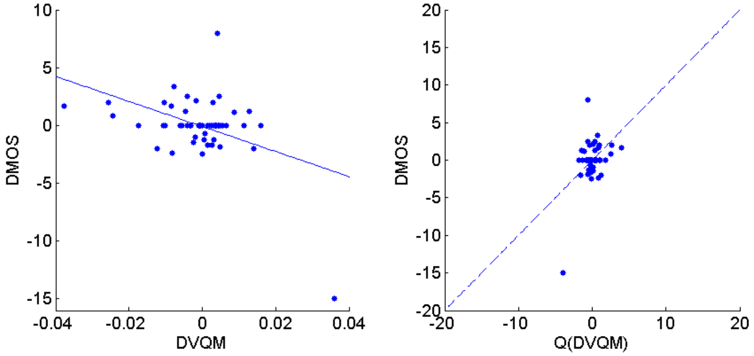


Fig. 29. DVQM and nonlinear mapped DVQM vs DMOS

5 Conclusions

Subjective evaluation is the most accurate method for measuring perceptual visual quality; however, it is time and money consuming, and is not applicable to most real-world applications. Therefore, objective visual quality metrics are developed, and many of them are discussed in this chapter. We categorized the objective visual quality metrics into two categories, i.e., the HVS-model-based metrics and engineering-based metrics, and introduced them separately. The HVS-model-based metrics account for various low-level characteristics of the HVS, such as luminance adaptation, CSF, contrast masking, etc., which are derived from physiological or psychophysical studies. These perceptual factors and also their implementations in visual quality metrics were discussed in detail. Two quality metric frameworks are presented to illustrate how these perceptual factors cooperate. Different from HVS-model-based quality metrics, engineering-based quality metrics are generally based on assumptions and prior knowledge, e.g., assumptions about the features which the HVS most likely correlate with visual quality, and the prior knowledge about the distortion properties. A conceptual framework was presented for FR and RR metrics, and four classic FR or RR visual quality metrics were summarized by fitting them into this framework. NR metrics were reviewed on the basis of the prior knowledge that they used, including the different distortion types and the statistics of the natural scenes.

An overview of standard subjective evaluation procedure is then presented. Specifically, we described the standard evaluation procedure ITU-R BT.500, which is one of the most commonly used procedures for subjective evaluation. The viewing environment, observer selection, test sequence selection, test procedure and score analysis, which are all important factors that affect the reliability and generality of the evaluation results, are discussed.

We also described an application of subjective video quality evaluation. Two recently developed video coding standards, H.264/AVC and AVS, are compared. Standard objective comparison method utilizing the rate-distortion performance shows that AVS has comparable performance to the H.264/AVC, except in some video sequences that have more complex textures. However, subjective evaluation shows that even on these sequences, the performance of the two systems are about the same. This demonstrates that the commonly used rate-distortion performance is not an accurate performance evaluation method. Two objective metrics, SSIM and VQM, are then utilized as the distortion measures and compared with PSNR. The results show that they are more correlated to the subjective evaluation results, but still cannot completely reflect the HVS perception.

With more knowledge on the psychophysical model of HVS, more accurate objective quality metrics can be developed in the future. However, modeling the extremely complex HVS is a challenging task. Until a thorough understanding of the human perception can be established, subjective evaluation will remain to be the most reliable method we can use for visual quality evaluation.

References

1. Ahumada, J.A.J., Peterson, H.A.: A Visual Detection Model for DCT Coefficient Quantization. In: The 9th AIAA Computing in Aerospace Conference, pp. 314–318 (1993)
2. Ahumada, J.A.J., Beard, B.L., Eriksson, R.: Spatio-temporal Discrimination Model Predicts Temporal Masking Functions. In: Proc. SPIE (1998), doi:10.1117/12.320103
3. AVS Video Expert Group, Draft of Advanced Audio Video Coding – Part 2: video, AVS_N1063 (2003)
4. AVS Video Expert Group, Information technology - Advanced coding of audio and video - Part 2: Video, GB/T 20090.2-2006 (2006)
5. Babu, R.V., Perkis, A.: An HVS-based no-reference Perceptual Quality Assessment Of Jpeg Coded Images Using Neural Networks. In: Proceedings of the International Conference on Image Processing, vol. 1, pp. 433–436 (2005)
6. Bjontegaard, G.: Calculation of average PSNR differences between RD curves, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, Doc. VCEG-M33 (2001)
7. Buccigrossi, R.W., Simoncelli, E.P.: Image Compression Via Joint Statistical Characterization in the Wavelet Domain. IEEE Transactions on Image Processing 8(12), 1688–1701 (1999)
8. Callet, P.L., Autrusseau, F.: Subjective Quality Assessment IRCCyN/IVC Database (2005), `http://www.irccyn.ec-nantes.fr/ivcdb/`
9. Caviedes, J., Oberti, F.: A New Sharpness Metric Based on Local Kurtosis, Edge and Energy Information. Signal Processing-Image Communication 19(2), 147–161 (2004)
10. Chandler, D.M., Hemami, S.S.: A57 Database, `http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html`
11. Cheng, H., Lubin, J.: Reference Free Objective Quality Metrics for Mpeg Coded Video. In: Human Vision and Electronic Imaging X, vol. 5666, pp. 160–167 (2005)
12. Chou, C.H., Li, Y.C.: A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion Profile. IEEE Transactions on Circuits and Systems for Video Technology 5(6), 467–476 (1995)
13. Daly, S.: The Visible Differences Predictor - an Algorithm for the Assessment of Image Fidelity. In: Human Vision, Visual Processing, and Digital Display III, vol. 1666, pp. 2–15 (1992)
14. Daly, S.: Engineering Observations from Spatiovelocity and Spatiotemporal Visual Models. In: Human Vision and Electronic Imaging III, vol. 3299, pp. 180–191 (1998)
15. Eskicioglu, A.M., Fisher, P.S.: Image Quality Measures and their Performance. IEEE Transactions on Communications 43(12), 2959–2965 (1995)
16. Fan, L., Ma, S.W., Wu, F.: Overview of AVS Video Standard. In: Proceedings of the IEEE International Conference on Multimedia and Expo., vol. 1, pp. 423–426 (2004)
17. Foley, J.M.: Human Luminance Pattern-Vision Mechanisms - Masking Experiments Require a New Model. Journal of the Optical Society of America a-Optics Image Science and Vision 11(6), 1710–1719 (1994)
18. Girod, B.: What's Wrong With Mean Squared Error? In: Watson, A.B. (ed.) Digital Images and Human Vision. The MIT Press, Cambridge (1993)
19. Grice, J., Allebach, J.P.: The Print Quality Toolkit: An Integrated Print Quality Assessment Tool. Journal of Imaging Science and Technology 43(2), 187–199 (1999)
20. Horita, Y., et al.: MICT Image Quality Evaluation Database, `http://mict.eng.u-toyama.ac.jp/mict/index2.html`

21. ISO/IEC 14496-10, Coding of Audio-visual Objects - Part 10: Advanced Video Coding. International Organization for Standardization, Geneva, Switzerland (2003)
22. ITU-T FG IPTV-ID-0082. Introductions for AVS-P2. 1st FG IPTV Meeting, ITU, Geneva, Switzerland (2006)
23. ITU-R Report BT.1082-1 Studies Toward The Unification of Picture Assessment Methodology. ITU, Geneva, Switzerland (1990)
24. ITU-R Recommendation BT.815-1 Specification of a Signal for Measurement of the Contrast Ratio Of Displays. ITU, Geneva, Switzerland (1994)
25. ITU-R Recommendation BT.710-4 Subjective Assessment Methods for Image Quality in High-Definition Television. ITU, Geneva, Switzerland (1998)
26. ITU-R Recommendation BT.500-11 Methodology for the Subjective Assessment of the Quality of Television Pictures. ITU, Geneva, Switzerland (2002)
27. ITU-R Recommendation BT.1683 Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference. ITU, Geneva, Switzerland (2004)
28. ITU-R Recommendation BT.814-2 Specifications and Alignment Procedures for Setting of Brightness and Contrast of Displays. ITU, Geneva, Switzerland (2007)
29. ITU-T Recommendation P.910 Subjective Video Quality Assessment Methods for Multimedia Applications. ITU, Geneva, Switzerland (2008)
30. Kanumuri, S., et al.: Modeling Packet-loss Visibility in MPEG-2 Video. IEEE Transactions on Multimedia 8(2), 341–355 (2006)
31. Lai, Y.K., Kuo, C.C.J.: A Haar Wavelet Approach to Compressed Image Quality Measurement. Journal of Visual Communication and Image Representation 11(1), 17–40 (2000)
32. Lambrecht, C.J.V.: Color Moving Pictures Quality Metric. In: Proceedings of International Conference on Image Processing, vol. I, pp. 885–888 (1996)
33. Legge, G.E., Foley, J.M.: Contrast Masking in Human-Vision. Journal of the Optical Society of America 70(12), 1458–1471 (1980)
34. Li, X.: Blind Image Quality Assessment. In: Proceedings of the International Conference on Image Processing, vol. 1, pp. 449–452 (2002)
35. Lin, W.S.: Computational Models for Just-Noticeable Difference. In: Wu, H.R. (ed.) Digital video image quality and perceptual coding. CRC Press, Boca Raton (2005)
36. Lin, W.S.: Gauging Image and Video Quality in Industrial Applications. In: Liu, Y. (ed.), SCI. Springer, Berlin (2008)
37. Lin, W.S., Li, D., Ping, X.: Visual Distortion Gauge Based on Discrimination of Noticeable Contrast Changes. IEEE Transactions on Circuits and Systems for Video Technology 15(7), 900–909 (2005)
38. Lu, Z.K., et al.: Perceptual Quality Evaluation on Periodic Frame-dropping Video. In: Proceedings of the IEEE International Conference on Image Processing, vol. 3, pp. 433–436 (2007)
39. Lubin, J.: The Use of Psychophysical Data and Models in the Analysis of Display System Performance. In: Watson, A.B. (ed.) Digital Images and Human Vision, The MIT Press, Cambridge (1993)
40. Lukas, F.X.J.: Picture Quality Prediction Based on a Visual Model. IEEE Transactions on Communications 30(7), 1679–1692 (1982)
41. Mannos, J.L., Sakrison, D.J.: The Effects of a Visual Fidelity Criterion on Encoding of Images. IEEE Transactions on Information Theory 20(4), 525–536 (1974)

42. Marichal, X.M., Ma, W.Y., Zhang, H.J.: Blur Determination in the Compressed Domain Using Dct Information. In: Proceedings of the International Conference on Image Processing, vol. 2, pp. 386–390 (1999)
43. Marziliano, P., et al.: Perceptual Blur and Ringing Metrics: Application to JPEG2000. Signal Processing-Image Communication 19(2), 163–172 (2004)
44. Masry, M., Hemami, S.S., Sermadevi, Y.: A Scalable Wavelet-based Video Distortion Metric and Applications. IEEE Transactions on Circuits and Systems for Video Technology 16(2), 260–273 (2006)
45. Miyahara, M., Kotani, K., Algazi, V.R.: Objective Picture Quality Scale (Pqs) for Image Coding. IEEE Transactions on Communications 46(9), 1215–1226 (1998)
46. Moorthy, A.K., Bovik, A.C.: Perceptually Significant Spatial Pooling Techniques for Image Quality Assessment. In: Proceedings of the SPIE Human Vision and Electronic Imaging XIV, vol. 7240, pp. 724012–724012-11 (2009)
47. Nadenau, M.J., Reichel, J., Kunt, M.: Performance Comparison of Masking Models Based on A New Psychovisual Test Method With Natural Scenery Stimuli. Signal Processing-Image Communication 17(10), 807–823 (2002)
48. Oguz, S.H., Hu, Y.H., Nguyen, T.Q.: Image Coding Ringing Artifact Reduction Using Morphological Post-Filtering. In: Proceedings of the IEEE Second Workshop on Multimedia Signal Processing, pp. 628–633 (1998)
49. Ong, E.P., et al.: A No-reference Quality Metric For Measuring Image Blur. In: Proceedings of the Seventh International Symposium on Signal Processing and Its Applications, vol. 1, pp. 469–472 (2003)
50. Pappas, T.N., Safranek, R.J.: Perceptual Criteria for Image Quality Evaluation. In: Bovik, A.C. (ed.) Handbook of Image and Video Processing. Academic Press, Orlando (2000)
51. Parmar, M., Reeves, S.J.: A Perceptually Based Design Methodology for Color Filter Arrays. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing III, pp. 473–476 (2004)
52. Pastrana-Vidal, R.R., et al.: Sporadic Frame Dropping Impact on Quality Perception. In: Human Vision and Electronic Imaging IX, vol. 5292, pp. 182–193 (2004)
53. Pastrana-Vidal, R.R., Gicquel, J.C.: Automative Quality Assessment of Video Fluidity Impairments Using a No-reference Metric. In: Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics (2006)
54. Peli, E.: Contrast in Complex Images. Journal of the Optical Society of America a-Optics Image Science and Vision 7(10), 2032–2040 (1990)
55. Pinson, M.H., Wolf, S.: A New Standardized Method for Objectively Measuring Video Quality. IEEE Transactions on Broadcasting 50(3), 312–322 (2004)
56. Ponomarenko, N., et al .: Tampere Image Database 2008 TID2008, version 1.0 (2008), `http://www.ponomarenko.info/tid2008.htm`
57. Poynton, C.: Gamma. In: Poynton, C. (ed.) A Technical Introduction to Digital Video. Wiley, New York (1996)
58. Seshadrinathan, K., Bovik, A.C.: A Structural Similarity Metric for Video Based on Motion Models. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 869–872 (2007)
59. Sheikh, H.R., et al.: LIVE Image Quality Assessment Database, Release 2 (2005), `http://live.ece.utexas.edu/research/quality`
60. Sheikh, H.R., Bovik, A.C.: Image Information and Visual Quality. IEEE Transactions on Image Processing 15(2), 430–444 (2006)

61. Sheikh, H.R., Bovik, A.C., Cormack, L.: No-reference Quality Assessment Using Natural Scene Statistics: JPEG2000. IEEE Transactions on Image Processing 14(11), 1918–1927 (2005)
62. Sullivan, G.J., Topiwala, P.N., Luthra, A.: The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to The Fidelity Range Extensions. In: Proceedings of the SPIE Applications of Digital Image Processing XXVII, vol. 5558, pp. 454–474 (2004)
63. Tan, K.T., Ghanbari, M., Pearson, D.E.: An Objective Measurement Tool for MPEG Video Quality. Signal Processing 70(3), 279–294 (1998)
64. Teo, P.C., Heeger, D.J.: Perceptual Image Distortion. In: Proceedings of IEEE International Conference on Image Processing, vol. 2, pp. 982–986 (1994)
65. Teo, P.C., Heeger, D.J.: Perceptual Image Distortion. In: Human Vision, Visual Processing, and Digital Display V, vol. 2179, pp. 127–141 (1994)
66. Verscheure, O., Frossard, P., Hamdi, M.: User-Oriented QoS Analysis in MPEG-2 Video Delivery. Real-Time Imaging 5(5), 305–314 (1999)
67. VQEG, Final report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment II. Video Quality Expert Group (2003), http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII/downloads/VQEGII_Final_Report.pdf (cited August 5, 2009)
68. VQEG, RRNR-TV Group Test Plan, Version 2.0. Video Quality Expert Group (2007), ftp://vqeg.its.bldrdoc.gov/Documents/Projects/rrnr-tv/RRNR-tv_draft_2.0_changes_accepted.doc (cited August 5, 2009)
69. VQEG, Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content, Draft Version 3.0.Video Quality Expert Group (2009), ftp://vqeg.its.bldrdoc.gov/Documents/Projects/hdtv/VQEG_HDTV_testplan_v3.doc (cited August 5, 2009)
70. VQEG, Hybrid Perceptual/Bitstream Group Test Plan, Version 1.3. Video Quality Expert Group (2009), ftp://vqeg.its.bldrdoc.gov/Documents/Projects/hybrid/VQEG_hybrid_testplan_v1_3_changes_highlighted.doc (cited August 5, 2009)
71. Vlachos, T.: Detection of Blocking Artifacts in Compressed Video. Electronics Letters 36(13), 1106–1108 (2000)
72. Wang, X.F., Zhao, D.B.: Performance Comparison of AVS and H. 264/AVC Video Coding Standards. Journal of Computer Science and Technology 21(3), 310–314 (2006)
73. Wang, Z., Bovik, A.C.: A Universal Image Quality Index. IEEE Signal Processing Letters 9(3), 81–84 (2002)
74. Wang, Z., Shang, X.L.: Spatial Pooling Strategies for Perceptual Image Quality Assessment. In: Proceedings of the International Conference on Image Processing, October7-10, vol. 1, pp. 2945–2948 (2006)
75. Wang, Z., Simoncelli, E.P.: Local Phase Coherence and the Perception of Blur. Advances in Neural Information Processing Systems 16, 1435–1442 (2004)
76. Wang, Z., Bovik, A.C., Evans, B.L.: Blind Measurement of Blocking Artifacts in Images. In: Proceedings of the International Conference on Image Processing, vol. 3, pp. 981–984 (2000)
77. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: from Error Visibility to Structural Similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)

78. Wang, Z., Lu, L., Bovik, A.C.: Video Quality Assessment Based on Structural Distortion Measurement. Signal Processing: Image Communication 19(2), 121–132 (2004)
79. Watson, A.B.: The Cortex Transform - Rapid Computation of Simulated Neural Images. In: Computer Vision Graphics and Image Processing, vol. 39(3), pp. 311–327 (1987)
80. Watson, A.B.: DCTune: A Technique for Visual Optimization of Dct Quantization Matrices for Individual Images. In: Proc. Soc. Information Display Dig. Tech. Papers XXIV, pp. 946–949 (1993)
81. Watson, A.B., Solomon, J.A.: Model of Visual Contrast Gain Control and Pattern Masking. Journal of the Optical Society of America a-Optics Image Science and Vision 14(9), 2379–2391 (1997)
82. Watson, A.B., Borthwick, R., Taylor, M.: Image Quality and Entropy Masking. In: Proc. SPIE (1997), doi:10.1117/12.274501
83. Watson, A.B., Hu, J., McGowan, J.F.: Digital Video Quality Metric Based on Human Vision. Journal of Electronic Imaging 10(1), 20–29 (2001)
84. Winkler, S.: A Perceptual Distortion Metric for Digital Color Video. In: Human Vision and Electronic Imaging IV, vol. 3644, pp. 175–184 (1999)
85. Winkler, S.: Issues in Vision Modeling for Perceptual Video Quality Assessment. Signal Processing 78(2), 231–252 (1999)
86. Winkler, S.: Metric Evaluation. In: Winkler, S. (ed.) Digital Video Quality: Vision Models and Metrics. Wiley, New York (2005)
87. Winkler, S.: Vision. In: Winkler, S. (ed.) Digital Video Quality: Vision Models and Metrics. Wiley, New York (2005)
88. Winkler, S.: Digital Video Quality: Vision Models and Metrics. Wiley, New York (2005)
89. Winkler, S.: Perceptual Video Quality Metrics - a Review. In: Wu, H.R. (ed.) Digital Video Image Quality and Perceptual Coding. CRC Press, Boca Raton (2005)
90. Winkler, S., Mohandas, P.: The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics. IEEE Transactions on Broadcasting 54(3), 660–668 (2008)
91. Wu, H.R., Yuen, M.: A Generalized Block-edge Impairment Metric for Video Coding. IEEE Signal Processing Letters 4(11), 317–320 (1997)
92. Yang, K.C., et al.: Perceptual Temporal Quality Metric for Compressed Video. IEEE Transactions on Multimedia 9(7), 1528–1535 (2007)
93. Yu, L., et al.: Overview of AVS-Video: Tools, Performance and Complexity. In: Proceedings of the SPIE Visual Communications and Image Processing, vol. 5960, pp. 679–690 (2005)
94. Yu, Z.H., et al.: Vision-model-based Impairment Metric to Evaluate Blocking Artifacts in Digital Video. Proceedings of the IEEE 90(1), 154–169 (2002)
95. Zhai, G.T., et al.: No-reference Noticeable Blockiness Estimation in Images. Signal Processing-Image Communication 23(6), 417–432 (2008)

Part III
Communications Related Processing

Scalable Video Coding and Its Applications

Naeem Ramzan and Ebroul Izquierdo

School of Electronic Engineering and Computer Science
Queen Mary University of London, Mile end road, London, United Kingdom

Abstract. Scalable video coding provides an efficient solution when video is delivered through heterogeneous networks to terminals with different computational and display capabilities. Scalable video bitstream can easily be adapted to required spatio-temporal resolution and quality, according to the transmission requirements. In this chapter, the Wavelet-based Scalable Video Coding (W-SVC) architecture is presented in detail. The W-SVC framework is based on wavelet based motion compensated approaches. The practical capabilities of the W-SVC are also demonstrated by using the error resilient transmission and surveillance applications. The experimental result shows that the W-SVC framework produces improved performance than existing method and provides full flexible architecture with respect to different application scenarios.

1 Introduction

Advances in video coding technology along with the rapid development in network infrastructure, storage capacity, and computing power are enabling an increasing number of video applications. In advanced communication systems, users may access and interact with multimedia content from different terminals and via different networks such as: video transmission and access over the Internet and handheld devices like mobile telephones and Personnel Digital Assistants (PDAs); multimedia broadcasting; and video services over wireless channels. In the scenario depicted in Fig.1, the video server requires video contents of different fidelities, such as high quality material for storage and future editing and lower bit-rate content for distribution. In traditional video communications over heterogeneous channels, the video is usually processed offline. Compression and storage are tailored to the targeted application according to the available bandwidth and potential end-user receiver or display characteristics. However, this process requires either transcoding of compressed content or storage of several different versions of the encoded video.

None of these alternatives represent an efficient solution. Furthermore, video delivery over error-prone heterogeneous channels meets additional challenges such as bit errors, packet loss and error propagation in both spatial and temporal domains. This has a significant impact on the decoded video quality after transmission and in some cases renders useless the received content. Consequently, concepts such as scalability, robustness and error resilience need to be re-assessed to allow for both efficiency and adaptability according to individual transmission bandwidth, user preferences and terminals.

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 547–559.
springerlink.com

Scalable Video Coding (SVC) promises to partially solve this problem by "encoding once and decoding many". SVC enables content organization in a hierarchical manner to allow decoding and interactivity at several granularity levels. That is, scalable coded bitstreams can efficiently adapt to the application requirements. The scenario shown in Fig. 1 can truncate the SVC encoded bitstream at different points and decode it. The truncated bitstream can be further truncated to some lower resolution, frame rate or quality. Thus, it is important to tackle the problems inherent to the diversity of bandwidth in heterogeneous networks and in order to provide an improved quality of services. Wavelet-based SVC provides a natural solution for error-prone transmissions with a truncateable bitstream.

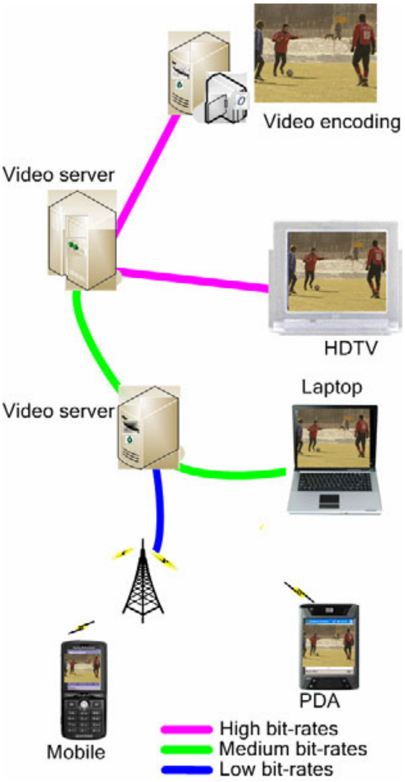


Fig. 1. Scalable video transmission: one video bitstream serves to different clients

The notion of extractor is the key in SVC. The extractor, as the name implies, extracts an adapted encoded video from the main encoded video. In this chapter, the wavelet-based SVC [1] framework is utilized and named as W-SVC. It is based on the wavelet transform performed in temporal and spatial domain. In this framework, the temporal and spatial scalability has been achieved by applying through Motion Compensated Temporal Filtering (MCTF) [2] in temporal domain and 2D Discrete Wavelet Transform (DWT) [3] in spatial domain respectively.

The MCTF results in motion information and wavelet coefficients that represent the texture of transformed frames. These wavelet coefficients are then bit-plane encoded [4, 5, 6] to achieve quality scalability.

In W-SVC, the video is divided into Group of Pictures (GOP). However, the GOP's of the video are interlinked, rendering dependency of following GOP on the proceeding GOP. In different applications, each GOP needs to be extracted with different scalability level according to requirement.

This flexible structure of SVC can be exploited in different application scenarios like Joint Source Channel Coding (JSCC) [7, 8] and event-based video coding in surveillance scenario [9].

The objective of JSCC is to jointly optimize the overall system performance subject to a constraint on the overall transmission bit-rate budget. As mentioned before, a more effective error resilient video transmission can be achieved if different channel coding rates are applied to different bit-stream layers, i.e., quality layers generated by the SVC encoding process. Furthermore, the parameters for Forward Error Correction (FEC) should be jointly optimized taking into account available and relevant source coding information.

In surveillance application, the surveillance videos are being processed using conventional video codecs which are designed to process videos regardless of the content of the video. However in many surveillance situations where the scene remains essentially static for seconds and even minutes in some cases. During these periods of time nothing interesting happens from the surveillance standpoint, and the video resembles a still picture for long periods of time with no other activity than random environmental motion. An alternative approach to reduce the bit-rate of the encoded video segments that are irrelevant from the surveillance standpoint are discussed in this chapter. This approach combines background subtraction and W-SVC. This produces a single scalable bit-stream that contains segments of video encoded at different qualities and / or spatio temporal resolutions. The irrelevant segments are encoded using low resolution / quality while the relevant segments are encoded at high resolution / quality. Additionally, the produced scalable bit-stream can easily be adapted for transmission purposes, without the need for computationally expensive transcoding.

This chapter is organized as follows: Section 2 explains the functionality of scalable coder and the scalability features used in existing standards. The architecture of W-SVC and used tools are described in Section 3. The performance evidence of W-SVC and its application in different scenario is demonstrated in Section 4. Section 5 concludes the chapter.

2 Scalability Functionality

Fundamental types of scalability are: temporal (frame rate); quality (SNR); and Spatial (resolution scalability). The main intend of SVC development was to meet the requirements for these basic types of scalability. Normally, the scalable bit-stream is organized progressively as the extraction/adaptation of the scalable bit-stream can be performed in a low complexity manner to achieve basic types of scalability. In temporal scalability mode, the scalable encoded bitstream can be extracted to one half of the frame rate, e.g from 50 fps (frames per second) to 25 fps. Similarly in spatial domain, the bitstream can be extracted to different resolution e.g. if the original sequence is of High Definition (HD) (1280x720 pixels) resolution, then one level lower resolution is HDS1 (640x360 pixels) and two levels lower resolution is HDS2 (320x180). An example of the original sequence and scaled sequences is shown in Fig. 2.

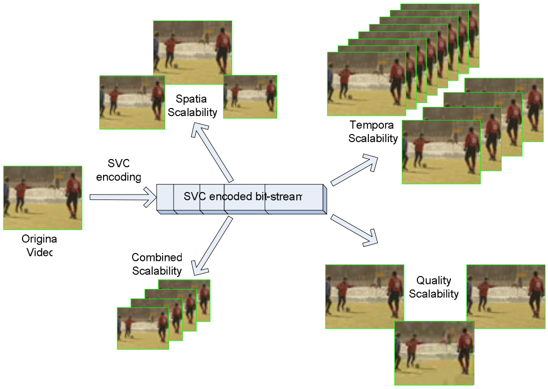


Fig. 2. Examples of basic scalabilities

Scalability features in existing standards: Some sort of scalability features are provided by different standards as shown in Table 1, however the full scalability features are provided by the emerging scalable extension of H.264/AVC or Wavelet-based SVC.

Table 1. Scalability features in existing video coding standards

Video codec	**Scalability features**
MPEG-1	No support
MPEG-2	Layered scalability (spatial, temporal, SNR)
MPEG-4	Layered and fine granular scalability (spatial, temporal, SNR)
H.264/A VC	Fully scalable extension
W-SVC	Fully scalable

Here, the detailed explanation of the wavelet-based SVC (W-SVC) is presented in next section.

3 Architectural Design of W-SVC

3.1 Main Modules of SVC

The W-SVC consists of three main modules:

Encoder: the input video is encoded by the W-SVC encoder, producing the bitstream of the maximum required quality which, if the application requires, can be up to quasi-lossless (resulting in imperceptible quality loss).

Extractor: the main aim of the W-SVC extractor is to truncate the scalable bitstream according to the scaling requirements and to generate the adapted bitstream and its description. The adapted bitstream is also scalable and can be fed back into the extractor for another stage of adaptation. This scenario corresponds to the situation of multiple-point adaptation where the adapted bitstream is sent to the next network node and is adapted by another extractor.

Decoder: W-SVC decoder is capable of decoding any adapted bitstream by W-SVC extractor or encoded by W-SVC encoder.

3.2 W-SVC Architecture

Most of the SVC frameworks consist of temporal decomposition using MCTF and spatial wavelet transform based on wavelets, producing a set of spatio-temporal sub-bands. In W-SVC high coding efficiency is achieved by using combinations of spatio-temporal transform techniques and 3D bit-plane coding. The Multi-Resolution (MR) structure resulting from MCTF and 2D sub-band decomposition enables temporal and spatial resolution scalabilities, respectively. Quality or SNR scalability is achieved by bit-plane coding.

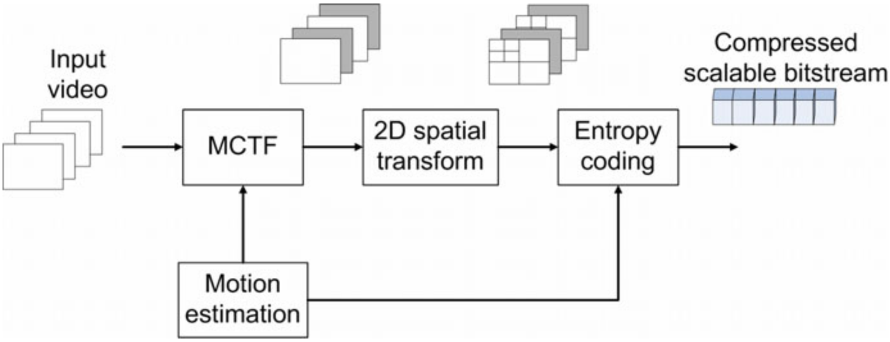


Fig. 3. t+2D W-SVC architecture

The order in which the spatial and temporal decompositions are performed is of crucial importance. Regarding the order in which these are applied, there are generally two basic types of non-redundant SVC architectures:

- t + 2D – temporal transform followed by the spatial transform,
- 2D + t – spatial transform followed by the temporal transform.

The t + 2D architecture provides higher compression performance but offers spatio-temporal mismatch [5] at lower resolution. Although the 2D+t architecture solves this problem at low resolution, it introduces shift variance in wavelet transforms that reduces the compression performance. Schemes that try to overcome limitations of these two architectures combine both approaches by performing several levels of down sampling, and are commonly known as 2D + t +2D architectures. On the other hand, Adami et al. [5] proposed an architecture that drops the requirement for the perfect reconstruction property in order to improve the spatial scalability performance. An example of the t+2D architecture is shown in Fig. 3. First, motion estimation is performed on the input frames and then temporal filtering is applied in the direction of motion vectors. Temporally decomposed frames are subjected to spatial decomposition based on 2D DWT. Since wavelet coefficients resulting from spatio-temporal transform are correlated, it is useful to apply some kind of compression scheme in combination with bit-plane coding of wavelet coefficients.

3.3 W-SVC Bitstream Organization

The input video is initially encoded with the maximum required quality. The compressed bitstream features a highly scalable yet simple structure. The smallest entity in the compressed bitstream is called an atom, which can be added or removed from the bitstream. The bitstream is divided into GOPs as shown in Fig. 4. Each GOP is composed of a GOP header, the atoms and an allocation table of all atoms. Each atom contains the atom header, motion vectors data (some atoms do not contain motion vector data) and texture data of a certain sub-band. Each atom can be represented in a 3D space with coordinates Q = quality, T = temporal resolution and S = spatial resolution. There exists a base atom in each domain that is called the 0-th atom and cannot be removed from the bitstream.

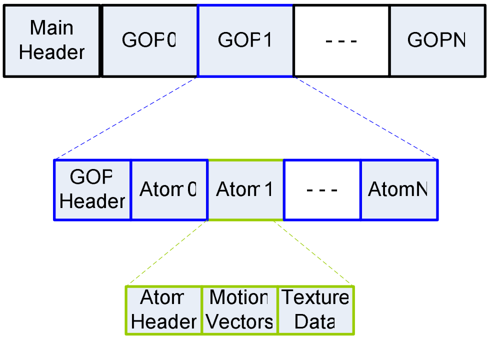


Fig. 4. Detailed description of used scalable bitstream

There are several progression orders for a multi-domain scalable bitstream, specifically for our case of quality, temporal and spatial scalability there are 3! = 6 orders. These are: QTS, QST, TQS, TSQ, SQT, STQ, where the convention is that the first letter refers to the domain which progresses most slowly, while the last refers to the one which progresses most quickly.

This flexible structure of scalable bitstream provides facilitation in different multimedia applications.

4 Applications

There are number of multimedia applications where SVC helps to reduce the complexity and provides natural solution of the problem.

4.1 Joint Source Channel Coding for Scalable Video

The progressive nature of scalable bitstream facilitates to apply the Unequal Error Protection (UEP) in the JSCC [10-12]. The JSCC consists of two main modules as shown in Fig. 5: scalable video encoding and channel encoding. At the sender side, the input video is coded using the W-SVC coder. The resulting bitstream is adapted according to channel capacities. The adaptation can also be driven by terminal or user requirements when this information is available. The adapted video stream is then passed to the channel encoding module where it is protected against channel errors. The channel coding module performs paketization, addition of CRC bits, and the channel error correction coding using a rate-distortion (R-D) optimization. After modulation, the video signal is transmitted over a lossy channel. At the receiver side, the inverse process is carried out. The main processing steps of the decoding are outlined in Fig. 5. Normally additive white Gaussian noise (AWGN) and Rayleigh fading channels are considered for research proposes.

Channel Coding: The main purpose of channel coding is to increase the reliability of data transmission. In channel coding, we normally add redundancy to the information data in order to provide error detection and correction capabilities at the receiver. Channel codes could be classified into two major categories: linear block codes; and convolutional codes. The encoder of the block code (n, k) divides the information data into blocks of k bits each and operates on them independently. It adds n-k redundant bits that are algebraically related to the k messages, thereby producing an overall encoded block of n bits called codeword with n>k, and R = k/n is called a code rate. The Reed-solomon and LDPC codes are the good example of block code. In contrast to block codes, convolutional codes have memory mr. An (n, k) convolutional encoder converts k information symbols into a codeword of length, which depend not only the k information symbols but also on the previous mr symbols. Nowadays, TCs [13] are one of the best practical channel codes because of their exceptional performance at low Signal to Noise Ratio (SNR). It is based on two convolutional encoders that are separated by an interleaver of length k. One reason for their better performance is that turbo codes

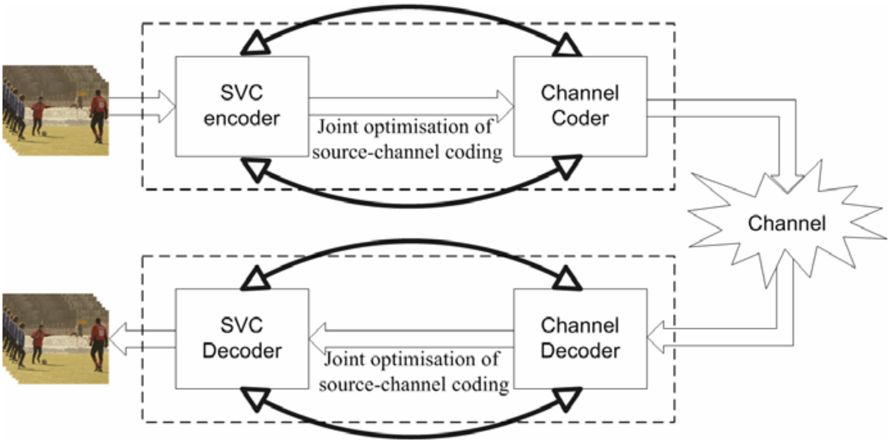

SVC encoder
Channel Coder
Joint optimisation of source-channel coding
Channel
SVC Decoder
Channel Decoder
Joint optimisation of source-channel coding

Fig. 5. Communication chain for video transmission

produce high weight codewords. Double binary TCs were introduced in the domain of TCs by Doulliard et al. [14]. These codes consist of two binary Recursive Systematic Convolutional (RSC) encoders of rate 2/3 and an interleaver of length k. Each binary RSC encoder encodes a pair of data bits and produces one redundancy bit, so the desired rate 1/2 is the natural rate of the double binary TCs.

Let us know formally state the problem, the R-D optimisation can be formulated as:

$$\min D_{s+c} \text{ subject to } R_{s+c} \leq R_{\max} \tag{1}$$

or

$$\max\left(PSNR\right)_{s+c} \text{ subject to } R_{s+c} \leq R_{\max} \tag{2}$$

for

$$R_{s+c} = R_s / R_c \text{,} \tag{3}$$

where D_{s+c} is the expected distortion at decoder, R_{s+c} is the overall system rate, R_s is the rate of the SVC bitstream, R_c is the channel coder rate and $R_{\max}$ is the given channel capacity. Here the index notation $s+c$ stands for combined source-channel information. The constrained optimization problem (1)-(3) can be solved by applying unconstrained Lagrangian optimization. Accordingly, JSCC aims at minimizing the following Lagrangian cost function J_{s+c} :

$$J_{s+c} = D_{s+c} + \lambda \cdot R_{s+c} \text{,} \tag{4}$$

Hence, the overall distortion can be explained by:

$$D_{s+c} = \sum_{i=0}^{Q} p_i \cdot D_{s,i-1} \ , \tag{5}$$

where p_i is the probability that the *i-th* quality layer is corrupted or lost, while the *j-th* layers are all correctly received for $j = 0,1,2,...,i-1$. Finally, p_i can be formulated as:

$$p_i = \left(\prod_{j=0}^{i-1} \left(1 - pl_j\right) \right) \cdot pl_i , \tag{6}$$

where pl_i is the probability of the *i-th* quality layer being corrupted or lost. pl_i can be regarded as the layer loss rate.

According to (6) the performance of the system depends on the layer loss rate, which in turn depends on the R_c. R_c depends upon how effectively the channel coding rate allocates the bit-rates to meet the channel capacity.

Now the problem converges to find the Optimal Protection Scheme (OPS) for channel encoding. There are number of methods [8, 11] are proposed to find the OPS.

An algorithm [8] is proposed to find the OPS efficiently. Initially, the optimal Equal Error Protection (EEP) is found and then protection is iteratively increased from the lowest quality layers and decreased from the highest quality layer. Hence the protection scheme converges to an OPS in this iterative process. In short, more protection is applied to the important part of the bitstream and a higher channel code rate is set for data with lower priority in the OPS. The lowest quality layer which contains the most important data (header, motion vectors and allocation tables) is protected with the lowest channel code rate and vice versa.

At the decoder side, if a packet is error-corrupted, the CRC fails after channel decoding. We then point out the corresponding atom in the SVC bitstream. If an atom (q_i, t_i, s_i) is corrupted after channel decoding or fails to qualify the CRC checks, then all the atoms which have higher index than *i* are removed by the error driven adaptation module. Finally, SVC decoding is performed to evaluate the overall performance of the system.

The performance of the JSCC framework has been extensively evaluated using the W-SVC codec. The JSCC proposed in [8], UEP optimal channel rate, packet size and interleaver for DBTC were estimated and used. The proposed technique is denoted as "ODBTC".

Two other advanced JSCC techniques were integrated into the same W-SVC codec for comparison. The first technique used serial concatenated convolutional codes of fixed packet size of 768 bytes and pseudo random interleaver [12]. It is denoted as "SCTC". Since product code was regarded as one of the most advanced in JSCC, the technique using product code proposed in [11] was used for the second comparison. This product code used Reed Solomon code as outer code and Turbo codes as inner code, so it is denoted by "RS+TC". A summary of PSNR results is shown in Fig. 6 and Fig. 7. These results show that the proposed UEP

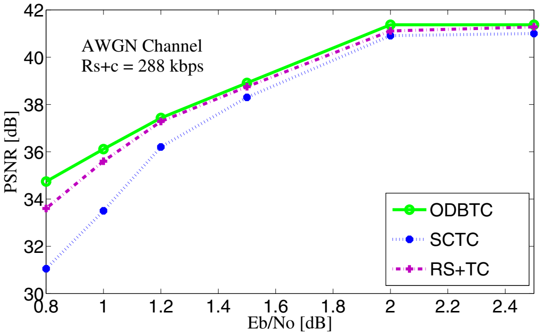


Fig. 6. Average PSNR for City QCIF sequence at 15 fps at different signal to noise ratio (Eb/No) for AWGN channel

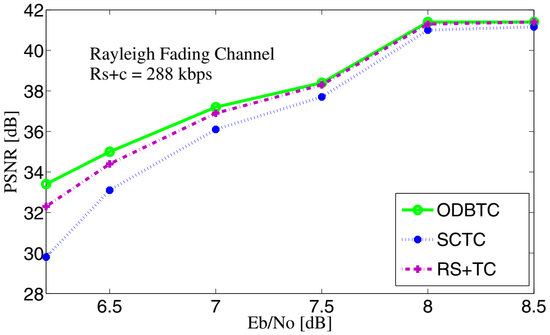


Fig. 7. Average PSNR for City QCIF sequence at 15 fps at different signal to noise ratio (Eb/No) for Rayleigh fading channel

ODBTC consistently outperforms SCTC and achieves PSNR gains at all signal-to-noise ratios (Eb/No) for both AWGN and Rayleigh fading channels.

4.2 Event-Based Scalable Coding of Surveillance Video

The basic principle behind the event-based scalable coding is to use different encoding settings for time segments representing different events in a surveillance video. For this purpose we classify temporal segments of the surveillance video into two types:

- temporal segments representing an essentially static scene (e.g. only random environmental motion is present – swaying trees, flags moving on the wind, etc.)
- temporal segments containing non-randomised motion activity (e.g. a vehicle is moving in a forbidden area).

To enable this classification, background subtraction and tracking module from [15] is used as Video Content Analysis (VCA). The output of this module defines parameters of compressed video. For actual encoding the W-SVC is employed.

VCA: Video background subtraction module based on Gaussian mixture model [15] is used as VCA. This module is able to deal robustly with light changes, bi-modal background like swaying trees and introduction or removal of objects from the scene. Value of each pixel is matched against weighted Gaussians of mixture. Pixels whose value is not within 2.5 standard deviations of the Gaussians representing background are declared as foreground.

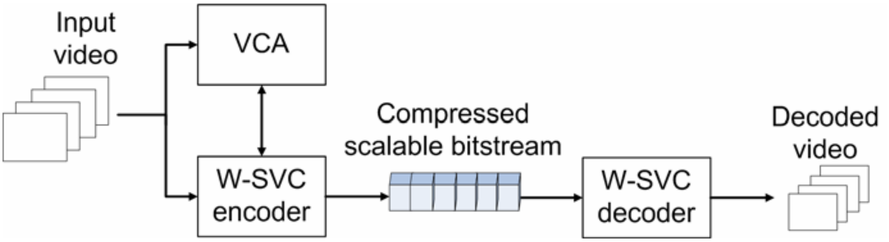


Fig. 8. Event-based scalable video encoding framework

At each time instance the W-SVC encoder communicates with the VCA module (background subtraction and tracking). When the input video is essentially static the output of the background subtraction does not contain foreground regions. This can be used to signal to the W-SVC encoder to adapt captured video at low spatio-temporal resolution and quality, as shown in Fig. 8. This allows, for instance, storing and/or transmitting the portions of the video containing long, boring, static scenes using low quality frame-rate and spatial resolution. On the other hand, when some activity in the captured video is detected, the VCA module notifies the W-SVC encoder to automatically switch its output to a desired much higher spatio-temporal resolution and quality video. Therefore, decoding and use of the video at different spatio-temporal resolutions and qualities corresponding to different events is achieved from a single bitstream, without multicasting or complex transcoding. Moreover, additional optional adaptation to lower bit-rate is also possible without re-encoding the video. This is, for instance, very useful in cases where video has to be delivered to a device with a low display capability. Using this approach, the bit-rate of video portions that are of low interest is kept low while the bit-rate of important parts is kept high. Since in many realistic applications it can be expected that large portions of the captured video have no events of interest, the proposed model leads to significant reduction of resources without jeopardizing the quality of any off-line event detection module that may be present at the decoder.

Subjective results of the event-based scalable encoding module are presented in Fig. 9. The top-left shows original frames in Fig 9a and Fig 9b. The top-right represents the output of the background subtraction module. The bottom row of Fig. 9a shows the reconstructed sequence whose essentially static segments (no event) were encoded at lower spatial resolution (bottom-left); at lower quality (bottom-right) and motion activities (event occurs) are encoded at higher spatial resolution (bottom-left); at higher quality (bottom-right) as shown in Fig. 9b.

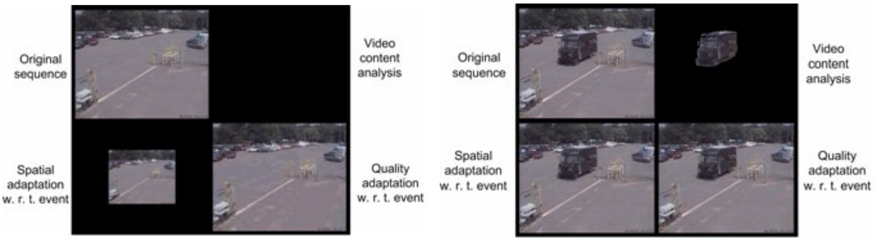

Original sequence
Video content analysis
Spatial adaptation w. r. t. event
Quality adaptation w. r. t. event
Original sequence
Video content analysis
Spatial adaptation w. r. t. event
Quality adaptation w. r. t. event

Fig. 9. Subjective result of event-based scalable encoding

a. there is no event in the video sequence. b. the event occurs in the video sequence.

5 Conclusions

This chapter has provided an overview of the different tools used in W-SVC. The architecture of W-SVC has presented in detail. Ffunctionality of scalable coder and the scalability features used in existing standards has explained in detail. The practical implementation of the flexible structure of W-SVC framework has demonstrated in surveillance application and error resilient transmission.

References

1. Mrak, M., Sprljan, N., Zgaljic, T., Ramzan, N., Wan, S., Izquierdo, E.: Performance evidence of software proposal for Wavelet Video Coding Exploration group. In: 76th MPEG Meeting ISO/IEC JTC1/SC29/WG11/ MPEG2006/M13146, Montreux, Switzerland (April 2006)
2. Ohm, J.-R.: Three-dimensional Subband Coding with Motion Compensation. IEEE Trans. Image Processing 3, 559–571 (1994)
3. Sweldens, W., Schroder, P.: Building your own wavelets at home. Wavelets in Computer Graphics, ACM SIGGRAPH Course notes, 15–87 (1996)
4. Zgaljic, T., Sprljan, N., Izquierdo, E.: Bitstream syntax description based adaptation of scalable video. In: Integration of Knowledge, Semantics and Digital Media Technology (EWIMT 2005), November 30, pp. 173–176 (2005)
5. Adami, N., Signoroni, A., Leonardi, R.: State-of-the-Art and Trends in Scalable Video Compression With Wavelet-Based Approaches. IEEE Transc. on Circuits and Systems for Video Technology 17(9) (September 2007)
6. Taubman, D.: High performance scalable image compression with EBCOT. IEEE Trans. Image Processing 9, 1158–1170 (2000)
7. Kondi, L.P., Ishtiaq, F., Katsaggelos, A.K.: Joint source-channel coding for motion-compensated DCT-based SNR scalable video. IEEE Trans. Image Process. 11(9), 1043–1052 (2002)

8. Ramzan, N., Wan, S., Izquierdo, E.: Joint Source-Channel Coding for Wavelet Based Scalable Video Transmission using an Adaptive Turbo Code. EURASIP Journal on Image and Video Processing, Article ID 47517, 12 pages (2007)
9. Zgaljic, T., Ramzan, N., Akram, M., Izquierdo, E., Caballero, R., Finn, A., Wang, H., Xiong, Z.: Surveillance Centric Coding. In: Proc. Of 5th International Conf. on Visual Information Engineering, VIE (July 2008)
10. Kim, J., Mersereau, R.M., Altunbasak, Y.: Error-resilient image and video transmission over the Internet using unequal error protection. IEEE Trans. Image Process. 12(2), 121–131 (2003)
11. Thomos, N., Boulgouris, N.V., Strintzis, M.G.: Wireless image transmission using turbo codes and optimal unequal error protection. IEEE Trans. Image Process. 14(11), 1890–1901 (2005)
12. Banister, B.A., Belzer, B., Fischer, T.R.: Robust video transmission over binary symmetric channels with packet erasures. In: Proc. Data Compression Conference, DCC 2002, pp. 162–171 (2002)
13. Berrou, C., Glavieux, A.: Near-optimum error-correction coding and decoding: Turbo codes. IEEE Trans. Commun. 44(10), 1261–1271 (1996)
14. Doulliard, C., Berrou, C.: Turbo codes with rate-m/(m+1) constituent convolutional codes. IEEE Trans. Commun. 53(10), 1630–1638 (2005)
15. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 747–757 (2000)

Auction Based Resource Allocation for Cooperative Wireless Video Transmission with Distributed Source Coding

Guan-Ming Su[1] and Zhu Han[2]

[1] Dolby Labs., Santa Clara, CA, USA
guanmingsu@ieee.org
[2] Electrical and Computer Engineering Department, University of Houston, Houston, TX, USA
zhan2@mail.uh.edu

With the recent rapid growth of communication, networking, and video compression technology, the real-time video streaming applications have evolved from traditional single-stream along simple transmitter-to-receiver path to complex multiple streams through advanced full-fledged cooperative networks. In this chapter, three major emerging advanced concepts are introduced: cooperative transmission, distributed source coding (DSC), and share auction based resource allocation. Cooperative transmission has been demonstrated as an effective transmission scheme to form virtual multiple-input and multiple-output (MIMO) system and provide diversity gains. Distributed source coding brings a new coding paradigm by letting the receiver jointly exploit the statistical dependencies among multiple streams sent from different sources without coding rate penalty. Share auction brings efficient way to allocate system resources in a distributed and collaborated manner to alleviate computation complexity. Based on these advanced concepts along with the advanced video processing ability for side information generation, a wireless multi-stream video transmission framework over full-fledged cooperative networks is presented.

1 Introduction

The concept of cooperative video communication [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14] has attracted significant attention lately. The basic idea is to efficiently take advantage of the broadcast nature of wireless networks, and to enable the relay nodes to play more intelligent and active roles in processing, transcoding, or re-adapting the received media information before transmitting to the next node. In addition, all the nodes in the network can serve at different roles simultaneously, as a transmitter, relay or receiver. The key benefit of such concept is to let nodes in a wireless network share information and transmit data cooperatively as a virtual antenna array to

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 561–583.
springerlink.com

improve the overall system performance. In [15], the authors bring together the unequal error protection (UEP) technique and the cooperative communications for layered video communication via the cross-layer design approach. The proposed layered-cooperation protects the base layer from channel errors through cooperation transmission to achieve higher error protection; while the enhancement layer is transmitted directly.

Distributed source coding (e.g. Slepian-Wolf coding for lossless compression and Wyner-Ziv coding for lossy compression) has been discussed in the literature for more than decades. Until recently, researchers start to incorporate the idea of distributed source coding into video transmission applications[16, 17]. Unlike the traditional source coding, where the whole source information is only observed and compressed by a single encoder, DSC investigates the problem that source information are observed by multiple locations and each observed signal is encoded individually in a distributed fashion. The decoder side will jointly reconstruct the original source according to all received compressed observations. By applying the DSC concept to the cooperative transmission scenario, the relay receives the broadcasted video stream from the source node and then performs video processing, such as transcoding, based on the received video packets to construct other observation of the video contents. The processed video stream at the relay node serves as additional side information at the destination to improve the decoded video quality. Combining the directly transmitted video stream and the side information from the relay, the destination can explore the source diversity to improve the reconstructed video quality. In this chapter, we present an integrated wireless video cooperative transmission framework that leveraging the benefits of both cooperative transmission and the idea of DSC.

When there are multiple users involved in the resource-limited full-fledged cooperative network, each user intents to compete the resources to maximize his/her own benefit [18, 19]. Note that the relay node can help improve video quality but its spectrum resource is also limited. The main issue is how to conduct resource allocation to utilize relay's resources. More specifically, each relay helps to connect a group of transmitters with a number of receivers. During the resource allocation process, the spectrum resources are first allocated for the transmitters which broadcast video packets to the relay and destination, and then for the relay nodes to transmit side information generated from the received packets to the destination. Thus, the resources used by the relays for each source are very critical for the overall network performance. In fact, we could adopt auction theory in our considered scenario. Auction theory is a subfield of the game theory attempting to mathematically capture behavior in strategic situations, in which an individual's success in making choices depends on the choices of others. The auction theory based solution has been successfully applied to the general cooperative data communications [5, 20] and gained attention for video communication applications via Vickrey-Clarke-Groves (VCG) auction [21, 22]. However, the computation complexity and communication overhead for VCG-based real-time video communication is very high. In this article, we propose a quasi-share auction based approach, which explores the concept of share auction into this new domain.

This article is organized as follows: In Section 2, the basics of cooperative transmission are studied, and the channel model and coding scheme are discussed. In Section 3, the cooperative video transmission protocol with DSC for one single transmitter-receiver pair is proposed and analyzed. In Section 4, the proposed resource allocation using quasi-share auction is demonstrated and analyzed for multiuser case. A performance upper bound is also presented to evaluate the proposed scheme. Simulations results are shown in Section 5 and conclusions are drawn in Section 6.

2 Background on Cooperative Communication Protocols

In this section, we use a single source-relay-destination case to review three traditional cooperative communication protocols; namely, direct transmission, amplify-and-forward, and decode-and-forward. Then we present the adopted channel model and forward error coding (FEC) code. The corresponding final coded bit error rate is discussed at the end of this section.

2.1 Cooperative Communication Protocols

We consider a single source-relay-destination cooperative communication environment as shown in Figure 1. In the considered environment, there are one source node s, one relay node r, and one destination node d. The cooperative transmission consists of two phases. In the first phase, source s broadcasts its information to both destination node d and relay node r. The received signals $Y_{s,d}$ and $Y_{s,r}$ at destination d and relay r can be expressed as

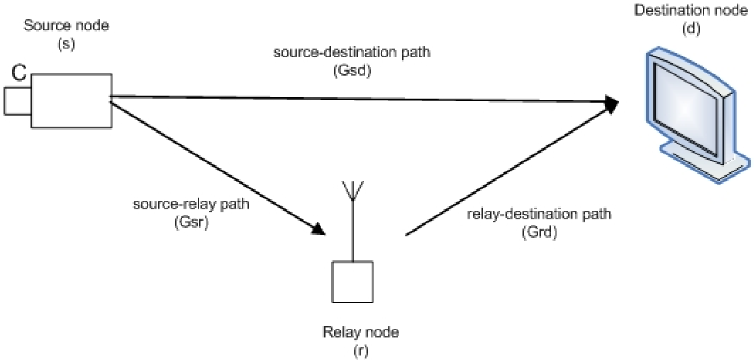


Fig. 1. Single source-relay-destination cooperative communication environment

$$Y_{s,d} = \sqrt{P_s G_{s,d}} X_{s,d} + n_d, \tag{1}$$

and

$$Y_{s,r} = \sqrt{P_s G_{s,r}} X_{s,d} + n_r, \tag{2}$$

respectively, where P_s represents the transmit power to the destination from the source, $X_{s,d}$ is the transmitted information symbol with unit energy at Phase one at the source, $G_{s,d}$ and $G_{s,r}$ are the channel gains from s to d and r respectively, and n_d and n_r are the additive white Gaussian noises (AWGN). Without loss of generality, we assume that the noise power is the same for all the links, denoted by σ^2. We also assume the channels are stable over each transmission time frame.

For *direct transmission (DT)*, without the relay node's help, the signal-to-noise ratio (SNR) that results from s to d can be expressed by

$$\Gamma^{DT} = \frac{P_s G_{s,d}}{\sigma^2}. \tag{3}$$

For the *amplify-and-forward (AF)* cooperation transmission, in Phase two, the relay amplifies $Y_{s,r}$ and forwards it to the destination with transmitted power P_r. The received signal at the destination is

$$Y_{r,d} = \sqrt{P_r G_{r,d}} X_{r,d} + n'_d, \tag{4}$$

where

$$X_{r,d} = \frac{Y_{s,r}}{|Y_{s,r}|} \tag{5}$$

is the energy-normalized transmitted signal from the source to the destination at Phase one, $G_{r,d}$ is the channel gain from the relay to the destination, and n'_d is the received noise at Phase two. Substituting (2) into (5), (4) can be rewritten as

$$Y_{r,d} = \frac{\sqrt{P_r G_{r,d}}(\sqrt{P_s G_{s,r}} X_{s,d} + n_r)}{\sqrt{P_s G_{s,r} + \sigma^2}} + n'_d. \tag{6}$$

Using (6), the relayed SNR at the destination for the source can be obtained by

$$\Gamma^{AF}_{s,r,d} = \frac{P_r P_s G_{r,d} G_{s,r}}{\sigma^2 (P_r G_{r,d} + P_s G_{s,r} + \sigma^2)}. \tag{7}$$

Therefore, by (3) and (7), we have the combined SNR at the output of maximal ratio combining (MRC) as

$$\Gamma^{AF} = \Gamma^{DT} + \Gamma^{AF}_{s,r,d}. \tag{8}$$

Notice that even though the SNR is improved, the bandwidth efficiency is reduced to half due to the half duplex of source transmission and relay transmission.

In the *decode-and-forward* (DF) cooperation transmission protocol, the relay decodes the source information transmitted in Phase one, re-encodes it, and retransmits the decoded information to the destination in Phase two. The destination combines the direct transmission information and re-encoded information together. We can express the SNR as

$$\Gamma^{DF} = \max_{0 \leq \rho \leq 1} \min\{(1-\rho^2)\frac{P_s G_{s,r}}{\sigma^2}, \frac{P_s G_{s,d}}{\sigma^2} + \frac{P_r G_{r,d}}{\sigma^2} + \frac{2\rho\sqrt{P_s G_{s,d} P_r G_{r,d}}}{\sigma^2}.\} \tag{9}$$

2.2 Channel Model and Forward Error Coding

In this article, we assume Rayleigh fading scenario. The bit error rate for a packet can be written as [23]

$$p_r = \frac{1}{2} - \frac{1}{2}\sqrt{\frac{\Gamma}{1+\Gamma}}, \tag{10}$$

where Γ is either Γ^{DT} in (3), Γ^{AF} in (7), or Γ^{DF} in (9), depending on the transmission protocol. If each packet has the length of L bits, the packet dropping rate is $1-(1-p_r)^L$.

Reed-Solomon (RS) code is an important subclass of the non-binary BCH error-correcting code in which the encoder operates on multiple bits rather than individual bits. An RS code is specified as RS(N, M). This means that the encoder takes M data symbols and adds parity symbols to make an N-symbol codeword. There are $N - M$ parity symbols. An RS decoder can correct up to t symbols that contain errors in a codeword, where $2t = N - M$. So by adapting t, we can have different level of channel protections. The coded bit error rate (BER) can be closely bounded by [23]

$$p_r^{RS} \leq \frac{1}{2}\left[1 - \sum_{i=0}^{t} \binom{N}{i} (p_r)^i (1-p_r)^{(N-i)}\right]. \tag{11}$$

Here we assume the BER is equal to 0.5 if the number of errors is greater than t. RS code can also be shortened to fit different coding length requirements.

3 Cooperative Wireless Video Transmission

In this section, we first present the proposed cooperative wireless video transmission framework for a single user scenario. Then a generic problem formulation for the proposed framework is shown in Section 3.2. With the background discussed in Section 2, we analyze the performance of different cooperative approaches adopted in the proposed framework in Section 3.3.

3.1 Framework of Cooperative Video Transmission

Compressed video exhibits many different characteristics from generic data, such as decoding dependency and delay constraint. For example, the visual

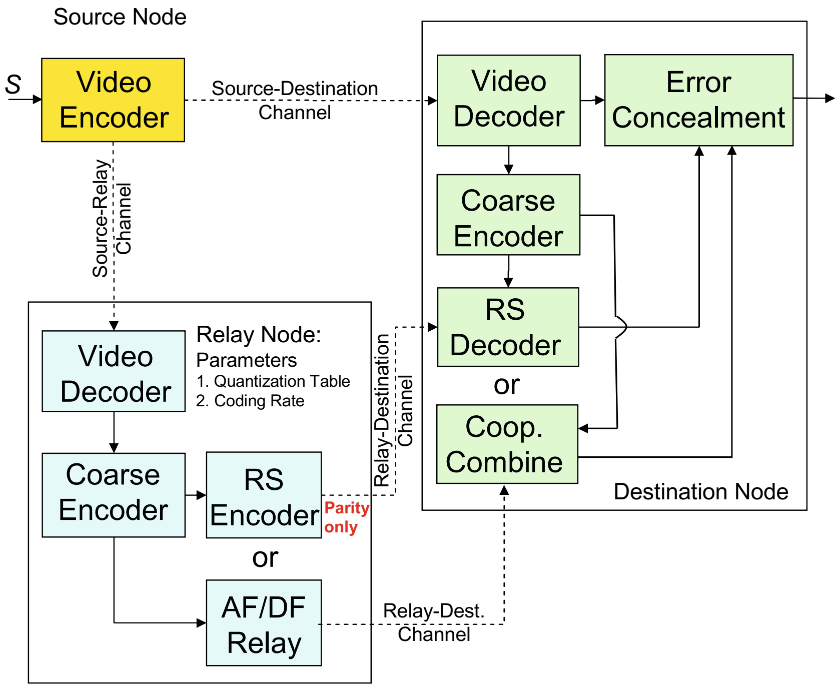


Fig. 2. Proposed Video Cooperation Transmission

quality depends on not only the transmission rate but also the video content, including the motion activity and complexity of texture. In addition, video frames can be further processed to adjust the quality or bit rate and still convey the main context of the delivered content. The proposed framework takes the aforementioned unique features of compressed video into consideration and integrates them into the proposed cooperative wireless video transmission framework.

We use Figure 2 to illustrate the proposed cooperative video transmission framework. We assume a control channel is available such that the destination knows the processing settings chosen by the relay node. The whole video capture-compression-transmission procedure is capsulated into two-stage pipeline for real-time streaming based on group of pictures (GOP). Denote the video refresh rate as F frames per second and number of frames in one GOP as G. Within the period to capture/compress one GOP video in the front-end video capture/compression component, the back-end transmission module delivers the compressed stream of previous GOP through the wireless channel.

There are two fundamental transmission phases for transmitting each GOP. In the first phase, the source broadcasts the video to the destination.

Because of the nature of broadcast, the source information is also delivered to the relay without any cost. To further take advantages of video unique characteristics, the relay node could own extra video processing tools, such as transcoder, and can convert the received image into a lower-resolution and/or lower-quality version. The processed version is then packed into packets and transmitted to the destination in the second phase. Note that although the source distortion induced along the source-relay-destination path is higher owing to transcoding, the end-to-end channel condition encountered may be better than the direct path along source to destination. The destination node will have two correlated video sources and have potential to further improve the video quality.

Based on the proposed framework, the relay can deploy different strategies:

1. For an embedded video stream whose main features consist of (1) any truncated segment of the stream can be decoded, and (2) the received quality is higher when more bits are received at the decoder, the relay can use AF or DF to relay the packets of first portion of the video stream. The destination combines the information from direct transmission and relay transmission to improve the quality of the received video. Note that by doing so, an unequal error protection scheme is constructed: first portion of stream is more important and a cooperative transmission protocol is adopted to provide better bit error rate for the first portion.
2. For both embedded and non-embedded video streams, the relay can deploy video processing, such as transcoding the received video to a coarse-quality video, and then encode the coarse-quality video using a systematic Reed-Solomon code. Only the parity bits are transmitted to the destination. The destination decodes the video transmitted in the first stage and transcodes it using the same coding parameters used by the relay node to construct the coarse-quality video. This coarse-quality video will be combined with the parity check bits sent from the relay to ensure the reconstructed coarse-quality video which will be utilized for error concealment. Notice that the relay might receive corrupted video packets. As a result, the relay might generate wrong parity bits and the performance at the destination can be impaired. To overcome this problem, a joint source-FEC-path optimization design can be deployed to protect the video stream in both transmission paths to maximize the final displayed video quality.

We can see that the proposed framework explores not only the inherited spatial diversity and multi-path diversity from cooperative transmission, but also the source diversity from the idea of DSC. Moreover, the proposed scheme is backward compatible, in the sense that the source-to-destination link is not modified. The current existing direct transmission scheme can coexist with the proposed scheme. This compatibility facilities the deployment of the proposed cooperative video transmission.

3.2 Generic Problem Formulation

In this article, we use 3D-SPIHT [24] as the video encoder, due to its advantage that SPIHT produces an embedded bitstream. Notice that if embedded video codecs are employed, the head segment of successful received packets serves as the coarse-quality version of the original video. Other video encoders can be implemented in a similar way.

Let us define D_{max} as the distortion without receiving any packets, $\eta = \frac{M}{N}$ as RS coding rate, $\Delta D_k(\eta)$ as the distortion reduction when receiving packet k after successfully receiving packet $1, 2, \ldots k-1$, and $p_k^{(X)}(\eta)$ as the probability that receiving all packets from packet 1 to k successfully using protocol X. The estimated distortion can be written as

$$E[D^{(X)}(\eta)] = D_{max} - \sum_{k=1}^{K} \Delta D_k(\eta) p_k^{(X)}(\eta), \tag{12}$$

where K is the maximal number of packets constrained by the bandwidth and tolerant delay. Notice that in order to decode the k^{th} packet, packet 1 to packet $k-1$ must be correctly decoded.

The problem is to minimize the expected received video distortion by turning the power and bandwidth usage at the relay node under the system bandwidth, transmission time delay, and overall power constraints. For the power constraint, we assume the overall power $P_s + P_r$ is bounded by P_0. For the delay constraint, we consider one GOP delay, i.e., G/F seconds, to cope with the intrinsic coding structure of 3D-SPIHT. For the bandwidth constraint, we suppose the source and relay share the same channel. For facilitate the discussion, we could further convert both the delay and bandwidth constraints to the total amount of packet constraint which limits the maximal number of packets, say K, to transmit if a fixed length of packet L is used. When the relay sends a total of $\bar{k} < K$ packets to the destination, the direct transmission has only $K - \bar{k}$ packets for transmission instead due to the total amount of packet constraint. By given a value of K, we are interested in how to assign the portion of packets for the direct transmission path and the relay transmission path. We further define a packet assignment parameter as

$$\theta = \frac{\bar{k}}{K}. \tag{13}$$

The whole cooperative video transmission problem can be formulated as

$$\min_{\theta, P_s, P_r, \eta} E[D] \tag{14}$$

$$\text{s.t.} \begin{cases} \text{packet assignment constraint: } 0 \leq \theta < 1, \\ \text{power constraint: } P_s + P_r \leq P_0, \\ \text{RS constraint: } 0 < \eta \leq 1. \end{cases}$$

The problem in (14) is a constrained optimization problem. The objective function $E[D]$ will be explained in the following subsection depending on the operation adopted in the relay node. Note that problem (14) is a non-linear optimization problem owing to the non-linear nature of video R-D function and coded BER function. Some standard nonlinear approaches such as interior-point-method [25] can be employed to solve the problem.

3.3 Performance Analysis

In this subsection, we study the details of different transmission protocols: direct transmission, relay transmission without combined decoding, relay transmission with combined decoding using AF/DF, and relay transmission with DSC. To overcome the strong decoding dependency exhibited in the video stream, all transmitted packets are protected by forward error coding. More specifically, the first three protocols are applied with RS(L, M_1), where L is the packet length and M_1 is the message length. We will also address the error protection scheme for the forth transmission protocol when we discuss the details.

Direct transmission

In the direct transmission, all the packet and power budget are allocated to the source-destination path. Thus $\theta = 0$ and $P_s = P_0$. The successful transmission probability for receiving all correct packet 1 to packet k can be written as

$$p_k^{(DT)}(\eta) = (1 - p_{s,d}(\eta))^k, \tag{15}$$

where $p_{s,d}(\eta)$ is the packet loss rate for sending a packet from the source node to destination node. $p_{s,d}(\eta)$ can be calculated from (3), (10), and (11). The distortion is

$$E[D^{(DT)}(\eta)] = D_{max} - \sum_{k=1}^{K} \varDelta D_k(\eta) p_k^{(DT)}(\eta). \tag{16}$$

Relay transmission without combined decoding

We use equal power for the source and relay in this scenario. Thus, $P_s = P_r = P_0/2$. Using this protocol, the relay simply forwards the received packets broadcast from the source and the destination node will not perform combined decoding (such as AF/DF) but have one extra copy from the relay besides the one from the direct path. A packet is lost if both the direct transmission and relay transmission fail. Thus,

$$p_k^{(RT)}(\eta) = \begin{cases} (1 - p_{s,d}(\eta)(1 - (1 - p_{s,r}(\eta))(1 - p_{r,d}(\eta)))^k, \ k \leq \bar{k} = \theta K; \\ p_{\bar{k}}^{RT}(\eta)(1 - p_{s,d}(\eta))^{k-\bar{k}}, \ K - \bar{k} \geq k > \bar{k}; \end{cases} \tag{17}$$

where $p_{s,r}(\eta)$ is the packet loss rate for sending a packet from source node to relay node and can be calculated from $\varGamma_{s,r} = \frac{P_s G_{s,r}}{\sigma^2}$ and (10); and $p_{r,d}(\eta)$ is the packet loss rate for sending a packet from the relay node to destination node and can be calculated from $\Gamma_{r,d} = \frac{P_r G_{r,d}}{\sigma^2}$ and (10).

In the aforementioned equation, the first case represents the situation where the relay retransmits the packets, while the second case represents the direct transmission only. The total number of transmitted packets from the source is reduced to $K - \bar{k}$, due to the relay transmission.

Then, the objective function becomes to minimize the expected distortion:

$$E[D^{(RT)}(\theta, \eta)] = D_{max} - \sum_{k=1}^{K-\bar{k}} \Delta D_k(\eta) p_k^{(RT)}(\eta). \tag{18}$$

Relay transmission with combined decoding using AF/DF

It can be proved that the power constraint and bandwidth constraint in (14) can be decoupled without loss of optimality. Due to the page limitation, we omit the proof. We assume the power is optimally allocated in this case. Under this protocol, the destination node will perform combined decoding if a packet is received from both direct path and relay forward path. Similarly to the previous case, we can write the expected distortion as

$$p_k^{(CD)}(\eta) = \begin{cases} (1 - p_{comb}(\eta))^k, & k \le \bar{k} = \theta K; \\ p_{\bar{k}}^{(CD)}(\eta)(1 - p_{s,d}(\eta))^{k-\bar{k}}, & K - \bar{k} \ge k > \bar{k}; \end{cases} \tag{19}$$

For AF, $p_{comb}(\eta)$ can be calculated by (8), (10), and (11). For DF, $p_{comb}(\eta)$ can be calculated by (9), (10), and (11).

The first case and second case have the similar physical meaning as (17). Similar to (18), we can also write

$$E[D^{(CD)}(\theta, \eta)] = D_{max} - \sum_{k=1}^{K-\bar{k}} \Delta D_k(\eta) p_k^{(CD)}(\eta). \tag{20}$$

Relay transmission with DSC

In the proposed DSC protocol, the packets received at the relay have length M_2 bits and are encoded as RS(M_2, M_1). The relay encodes the received packets with another layer of RS code with parameter RS(L, M_2), and sends the parity bits with length of $L - M_2$ only. The destination combines (1) M_2 bits from the direct source-to-destination transmission part, and (2) $L - M_2$ parity bits from the relay-to-destination transmission bits to improve the link quality. In this case, the packet assignment parameter becomes $\theta = \frac{L-M_2}{L}$. Denote $\varphi^m = \frac{M_1}{M_2}$ as the inner RS coding rate and $\varphi^n = \frac{M_2}{L}$ as the outer RS coding rate. Note that $\theta = 1 - \varphi^n$. Assuming the equal power allocation for the source and relay, the successful transmission probability for receiving all correct packet 1 to packet k can be written as

$$p_k^{(DSC)}(\varphi^m, \varphi^n) = (1 - p_{DSC}(\varphi^m, \varphi^n))^k, \tag{21}$$

where the packet error rate is the product of the successful packet transmission rate along source-to-relay path and the successful packet transmission rate after RS(L, M_2) decoding from the source to the destination, i.e.,

$$p_{DSC}(\varphi^m, \varphi^n) = 1 - (1 - p_{s,r}(\varphi^m))(1 - p_{s,r,d}^{RS}(\varphi^n))^L. \tag{22}$$

Define $t' = \frac{L-M_2}{2}$. We have the BER after the decoding of RS(L, M_2) code for both direct transmission and relay transmission as

$$p_{s,r,d}^{RS}(\varphi^n) \leq \frac{1}{2}\left[1 - \sum_{j=0}^{t'}\sum_{i=0}^{t'-j}\binom{M_2}{j}(p_{s,d})^j(1-p_{s,d})^{(M_2-j)}\right.$$
$$\left.\binom{L-M_2}{i}(p_{r,d})^i(1-p_{r,d})^{(L-M_2-i)}\right]. \tag{23}$$

We use the fact that the RS code can decode up to t' errors in either direct transmission part or the relay transmission part in (23). The expected distortion with DSC can be expressed as

$$E[D^{(DSC)}(\varphi^m, \varphi^n)] = D_{max} - \sum_{k=1}^{K} \Delta D_k(\varphi^m) p_k^{(DSC)}(\varphi^m, \varphi^n). \tag{24}$$

4 Quasi-share Auction Schemes for Multiple Sources

In the previous section, we study the single source-relay-destination case in which one relay tries to help one source-destination pair for the received video quality. In this section, based on the proposed cooperative video transmission scheme, we investigate multiuser case in which one relay tries to help a group of source-destination pairs to achieve the social optimum, i.e., the overall video quality. We first formulate the multiuser resource allocation problem in Section 4.1. Then, the quasi-share auction solution is proposed and analyzed in Section 4.2. To evaluate the optimality of the proposed solution, we discuss one existing approach in the literature to serve as one performance bound in Section 4.3.

4.1 Multiuser Resource Allocation for Relay Node

We consider the full-fledged cooperative network, in which each node can serve as transmitter, relay, or receiver. To make problem a bit simpler, we assume the nodes that play relay functions have been pre-determined (so the relay node determination problem is not in the scope of this article), so that each relay helps to connect a group of transmitters with a number of receivers

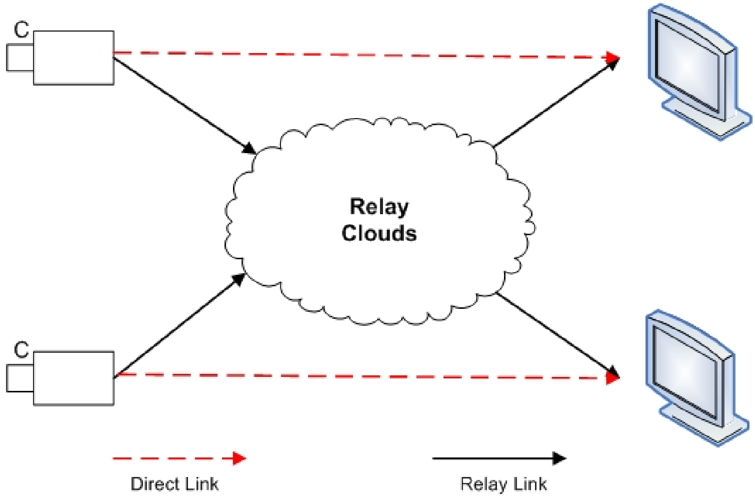


Fig. 3. Multiple User Resource Allocation in Relay

as shown in Figure 3. In this article, we suppose the cooperative transmission system has source node s_i, one relay node r and destination node d_i.

Denote set $\mathcal{I}$ as I source-destination pairs accessing one particular relay node in the network. To achieve real-time transmission, the overall allocated transmission time slots for all nodes to transmit I GOPs is set to the required time to playback one GOP. We reserve $t\%$ of time slots for relay node. The rest of time slots are allocated to each source-destination pair equally. Due to the distributed location of the nodes, each source-destination pair experiences different channel conditions in both direct and cooperative transmission path. Besides, since different sources transmit different video sequences and the contents are changing over time, the relay needs to dynamically adjust rate allocation to provide optimal video quality. The main issue is how to assign relay's time slots to each source-destination pair for delivering side information to achieve overall optimal video quality.

In this section, we take the relay transmission with DSC protocol as an example (which will be demonstrated in Section 5 to have best video quality among all protocols in most cases). Nevertheless, other protocols can be applied in a similar fashion. Define α_i as the fraction of relay's time slots assigned to source-destination pair i (including both source-relay and source-destination path). φ_i^m and φ_i^n are the inner and outer RS channel coding rate selected by the source i. We can formulate the considered network within each GOP time scale as

$$\min_{\alpha_i, \varphi_i^m, \varphi_i^n} \sum_{i \in \mathcal{I}} E[D_i(\alpha_i, \varphi_i^m, \varphi_i^n)] \tag{25}$$

$$\text{s.t.} \begin{cases} \sum_{i\in\mathcal{I}} \alpha_i \leq 1, \\ 0 < \varphi_i^m \leq 1, \forall i \in \mathcal{I}, \\ 0 < \varphi_i^n \leq 1, \forall i \in \mathcal{I}. \end{cases}$$

By a given α_i, the minimal achievable distortion for received video at destination i can be calculated as follows

$$ED_{s_i,r,d_i}(\alpha_i) = \min_{\varphi_i^m, \varphi_i^n} E[D_i(\alpha_i, \varphi_i^m, \varphi_i^n)]. \tag{26}$$

The problem in (26) can be solved locally in each source.

For the relay, the resource allocation problem is to optimize the overall distortion by dividing the relay's resources, which are the time slots. The problem can be formulated as

$$\min_{\alpha_i} \sum_{i\in\mathcal{I}} ED_{s_i,r,d_i}(\alpha_i) \tag{27}$$

$$\text{s.t.} \sum_{i\in\mathcal{I}} \alpha_i \leq 1.$$

In the next two subsections, we present the proposed solution to solve the problem in (27) and its corresponding performance bound.

4.2 Proposed Quasi-Share Auction

Theoretically, problem (27) can be solved in a centralized fashion by collecting all values of (26) from all users and performing optimization in one node with high computation ability. Note that when the bandwidth and delay increases, to gain the optimum of problem problem (27), the required information transmitting between source nodes to the managing node for function (26) becomes prohibitively high to consider every possible α_i. This leads us to investigate a distributed method to alleviate the information exchange.

In this subsection, we find a distributed solution to solve problem (27). Due to the distributed nature, different source-destination pairs try to optimize their own performances in a non-collaborate way. An auction is a decentralized market mechanism for allocating resources. The required information exchange between the user node and the moderator node will be limited to the bids sent from user and the bidding results. Note that although the auction can have several iterations with updated bid information; the overall consumed bandwidth is much less than the centralized scheme. Motivated by share auction, we propose a quasi-share auction that takes advantage of setting for cooperative video transmission. We will explain the differences between the shared auction and proposed quasi-share auction later. The rules of the quasi-share auctions are described below.

- *Information*: Public available information includes noise density σ^2 and maximal packet number K. The relay also announces a positive *reserve* *bid* (or reserve price in some literature) $\beta > 0$ and a *price* $\pi > 0$ to all sources. Each source i also knows the channel gains along direct and cooperative transmission path, namely, G_{s_i,d_i}, $G_{s_i,r}$, and G_{r,d_i}.

- *Bids*: Source i submits $b_i \geq 0$ to the relay.
- *Allocation*: Relay allocates proportions of time slot for source-destination pair i according to

$$\alpha_i = \frac{b_i}{\sum_{j \in \mathcal{I}} b_j + \beta}. \tag{28}$$

- *Payments*: In our case, source i pays the relay $C_i = \pi \alpha_i$.

A bidding profile is defined as the vector containing the sources' bids, $\mathbf{b} = (b_1, ..., b_I)$. The bidding profile of source i's opponents is defined as $b_{-i} = (b_1, ..., b_{i-1}, b_{i+1}, ..., b_I)$, so that $\mathbf{b} = (b_i; b_{-i})$. Each source i chooses bid b_i to maximize its payoff

$$S_i\left(b_i; b_{-i}, \pi\right) = \triangle E[D_{s_i,r,d_i}\left(\alpha_i\left(b_i; b_{-i}\right)\right)] - C_i\left(b_i; b_{-i}, \pi\right), \tag{29}$$

where

$$\triangle\, E[D_{s_i,r,d_i}\left(\alpha_i\left(b_i; b_{-i}\right)\right)] = E[D_{s_i,r,d_i}(0)] - E[D_{s_i,r,d_i}(\alpha_i\left(b_i; b_{-i}\right))]. \tag{30}$$

Since each source chooses its bid to maximize its payoff function in (29), from (28), the relay allocates more time slots to this user with higher bid to achieve better video quality. However, the cost C_i also increases. Consequently, if the other users do not change their bids, there is an optimal point to set the price.

Although video's rate-distortion (R-D) curve is often a *convex* decreasing function; however, (30) is generally not a *concave* increasing function owing to applying optimization over all possible channel coding rates for each α_i in (26). Notice that above payoff function for the quasi-share auction is similar to the soul of "Pricing Anarchy", in which the users pay the tax for their usage for the system resources.

Here, we omit the dependency on β. If the reserve bid $\beta = 0$, then the resource allocation in (28) only depends on the ratio of the bids. In other words, a bidding profile $k\mathbf{b}$ for any $k > 0$ leads to the same resource allocation, which is not desirable in practice. That is why we need a positive reserve bid. However, the value of β is not important as long as it is positive. For example, if we increase β to $k\beta$, then sources can just scale $\mathbf{b}$ to $k\mathbf{b}$, which results in the same resource allocation. For simplicity, we can simply choose $\beta = 1$ in the practice.

In (29), if the others' bids b_{-i} are fixed, source i can increase its time slot α_i in (28) by increasing b_i. As a result, the distortion is reduced and $\varDelta E[D_i]$ is improved. However, the payoff faction needs to pay the price for α_i. Depending one different price per unit π announced by the relay, there are three different scenarios:

1. If π is too small, the payoff function S_i in (29) is still an increasing function. As a result, the source tries to maximize its own benefit by setting price high. Consequently, $b_i \to \infty$.
2. If π is too large, the payoff function S_i is a decreasing function. As a result, the source would not participate in the bidding by setting $b_i = 0$.
3. If π is set to the right value, the payoff function S_i is a quasi-concave shape function, i.e., it increases first and then decreases within the feasible region. Consequently, there is an optimal b_i for the source to optimize its performance.

Based on the observation above, the quasi-share auction algorithm is shown as follows. The relay conducts line search for π from the situation in which $b_i = 0, \forall i$ to the situation in which $b_i = \infty, \forall i$. For each π, different sources set bids to optimize their own performances and report the expected distortion to the relay. By doing so, the computation is distributed to each source node. Among all π s', the relay selects the solution with the best overall system performance and announces the final α_i to each source i.

Compared with the share auction and the proposed quasi-share auction, the final results are the same if the bid update for share auction can be obtained and $\Delta E[D_i]$ is a concave increasing function. For data communication, the bids can be updated in a close form. However, due to the complexity to express the cooperative video end-to-end distortion, the close form update function cannot be obtained. As a result, we can only apply the quasi-share auction for the video cooperative transmission.

4.3 Performance Upper Bound

In this subsection, we investigate a performance upper bound similar to the VCG auction [26, 27, 28] proposed in the literature and compared with our proposed approach. In the performance upper bound, the relay asks all sources to reveal their evaluations of the relay's time slots, upon which the relay calculates the optimal resource allocation and allocates accordingly. A source pays the "performance loss" of other sources induced by its own participation of the auction. In the context of cooperative video transmissions, the performance upper bound can be described as follows:

- *Information*: Public available information includes noise density σ^2 and bandwidth W. Source s_i knows channel gain G_{s_i,d_i}. The relay knows channel gains G_{r,d_i} for all i, and can estimate the channel gains $G_{s_i,r}$ for all i when it receives bids from the sources.
- *Bids*: source s_i submits the function $Q_i\left(\alpha_i, G_{s_i,r}, G_{r,d_i}\right)$ to the relay, which represents the distortion decrease as a function of the relay parameter α_i and channel gains $G_{s_i,r}$ and G_{r,d_i}.
- *Allocation*: the relay determines the time slot allocation by solving the following problem (for notational simplicity we omit the dependence on $G_{s_i,r}$ and G_{r,d_i}),

$$\alpha^{*} = \arg\max_{\alpha} \sum_{j \in \mathcal{I}} Q_j\left(\alpha_j\right). \tag{31}$$

- *Payments*: For each source i, the relay solves the following problem

$$\alpha^{*/i} = \arg\max_{\alpha, \alpha_i = 0} \sum_{j} Q_j\left(\alpha_j\right), \tag{32}$$

i.e, the total distortion decreases without allocating resource to source i. The payment of source i is then

$$C_i = \sum_{j \neq i, j \in \mathcal{I}} Q_j\left(\alpha_j^{*/i}\right) - \sum_{j \neq i, j \in \mathcal{I}} Q_j\left(\alpha_j^{*}\right), \tag{33}$$

i.e., the performance loss of all other sources because of including source i in the allocation.

Source i in the performance upper bound obtains the *payoff* function as

$$Y_i = \triangle E[D_{s_i,r,d_i}\left(\alpha_i\right)] - C_i, \tag{34}$$

where

$$E[D_{s_i,r,d_i}\left(\alpha_i\right)] = E[D_{s_i,r,d_i}\left(0\right)] - E[D_{s_i,r,d_i}\left(\alpha_i\right)]. \tag{35}$$

Although a source can submit any function it wants, it has been shown that [26] that it is a (weakly) dominant strategy to bid truthfully, i.e., revealing the true function form of its distortion decrease

$$Q_i\left(\alpha_i\right) = \max\left\{E[D_{s_i,r,d_i}\left(0\right)] - E[D_{s_i,r,d_i}\left(\alpha_i\right)], 0\right\}. \tag{36}$$

As a result, the resource allocation of the performance upper bound as calculated in (31) achieves the *efficient* allocation [26].

Note that the sources do not need to know global network information, i.e., no need of knowing the channel gains related to other sources in the network. The auction can achieve the efficient allocation in one shot, by allowing the relay to gather a lot of information and perform heavy but local computation.

Although the performance upper bound has the desirable social optimal, it is usually computationally expensive for the relay to solve $I+1$ nonconvex optimization problems. To solve a nonconvex optimization, the common solution like interior point method needs a complexity of $O(I^2)$. As the result, the overall complexity for the performance upper bound is $O(I^3)$, while the proposed quasi-share auction has linear complexity. Furthermore, there is a significant communication overhead to submit $Q_i\left(\alpha_i\right)$ for each source i, which is proportional to the number of source nodes and reserved time slot for relay node. In the proposed scheme, the bids and the corresponding resource allocation are iteratively updated. This is similar to the distributed power control case, where the signal-to-interference-noise ratio and power update are iteratively obtained. As a result, the overall signalling can be reduced.

5 Simulation Results

In this section, we show the simulation results of the proposed framework. We first demonstrate the superior performance of the proposed cooperative video transmission in the single source-destination pair scenario. Then, we investigate the multiuser case for the proposed resource allocation using quasi share auction theory.

5.1 Single Source-Destination Pair Scenario

The simulation environment is set up as follows: The overall power $P_0 = 0.2$ Watt, the noise power is -100dbmw, and the propagation factor is 3. The source is located at the origin and the destination is located at $(1000m, 0m)$. The relay is moved from the $(100m, 400m)$ to $(900m, 400m)$. The packet length is $L = 255$. The tested video stream is *Foreman* in QCIF resolution (176x144) with refresh rate 30 frames per second.

In Figure 4, we show the peak-to-noise ratio (PSNR) as a function of the relay location for video *Foreman*. Here we normalize the relay location in x-axis over the distance from the source to the destination. From the figures, we can see that the direct transmission has the worst performance and provides unacceptable reconstructed video quality. The cooperative transmission without combined decoding at the receiver has the best performance when the relay is located at the middle of the source node and the destination node. For the AF protocol, the best performance is achieved when the relay

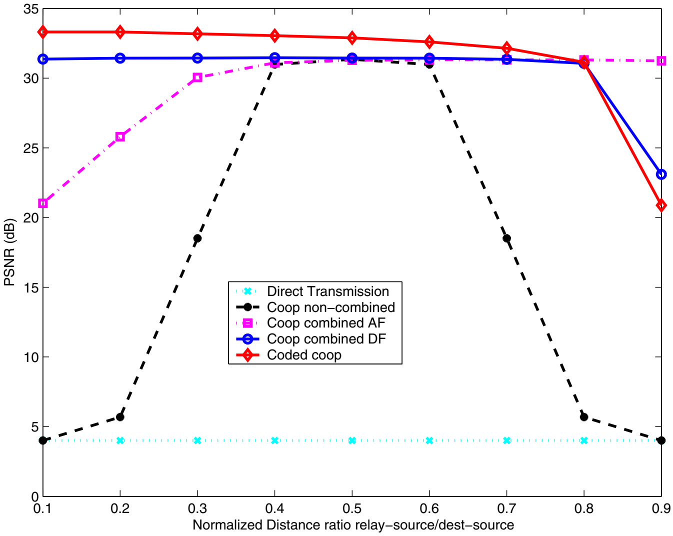


Fig. 4. PSNR vs. Relay Location (Video Foreman)

is relatively close to the destination; for the DF protocol, the optimal relay location is obtained toward the source node, and the DF protocol has better performance than the AF protocol when the relay node is close to the source node. These facts are very different from the generic-data domain cooperative transmission. Finally, the relay transmission with parity check bits (shown as coded coop) has superior performance (about 4dB gain) than the other protocols when the source node and relay node are close. However, when the relay node is far away from the source node, its performance degrades fast. This is because the performance is impaired by the source-relay link. On the whole, the proposed cooperative protocols can achieve better performances than direct transmission, and the characteristics of the performance improvement for video applications are very different from the ones from generic-data domain cooperative protocols.

Notice that the proposed cooperative framework will not always perform well in every relay location. The location of relay node needs to be close to the source-destination link. Otherwise, the cooperative transmission will not work, in the sense that the optimization in (14) degrades to traditional direct transmission with $\theta = 0$.

We are also interested in which protocol performs best under each particular scenario. For the AF/DF protocol, the received SNR can be significantly increased. This is especially true for low SNR case. However, the signal needs to be stored in the receiver for combining at the second state. This increases the implementation cost. For the relay transmission without combined decoding, the implementation cost is minor, but it has inferior performance when the SNR is low. The proposed scheme with parity check bits provides an improvement over the relay transmission without combined decoding in a cost effective manner. However, the relay needs to be close to the source to ensure a the good source-relay channel.

5.2 Multiple Source-Destination Pairs Scenario

For multiple user case, the simulations are setup as follows. The power for all source nodes and relay node is 0.1 Watt, the noise power is $5 * 10^{-10}$ Watt, and the propagation factor is 3. Source node 1 to node 3 are located at (-400m, 0m), (-300m, 50m), and (-200m, -20m), respectively. The corresponding destination node 1 to 3 are located at (200m, 0m), (400m, 100m), and (300m, 30m), respectively. The relay is located at the origin. We reserve 30% of bandwidth for relay to transmit the parity check bits. The packet length is L = 255. Again, we adopt 3D-SPIHT [24] codec as an example to compress video sequence in QCIF resolution (176x144) with refresh rate 30 frames per second. The GOP is set to 16 and each source node will transmit 10 GOPs to its corresponding destination node. To evaluate the performance under different video content and different level of motion activity in the video sequence, we compare three different sets of video sequences. The first set consists of low motion video sequences: *news*, *grandma*, and *akiyo*.

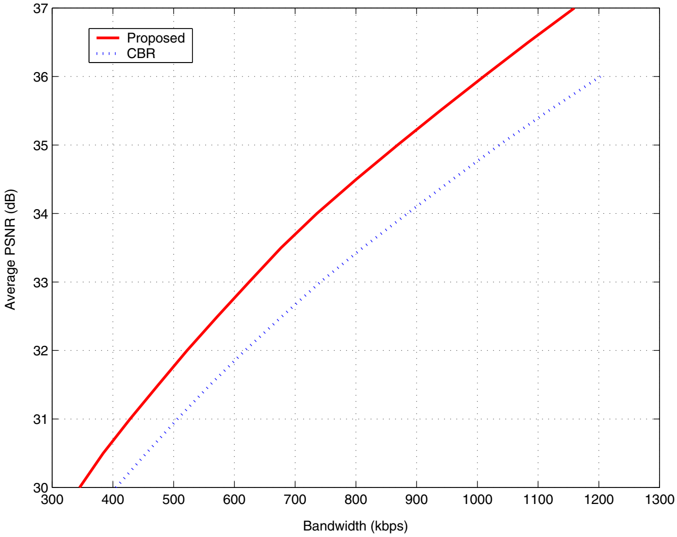


Fig. 5. Average PSNR vs. Bandwidth

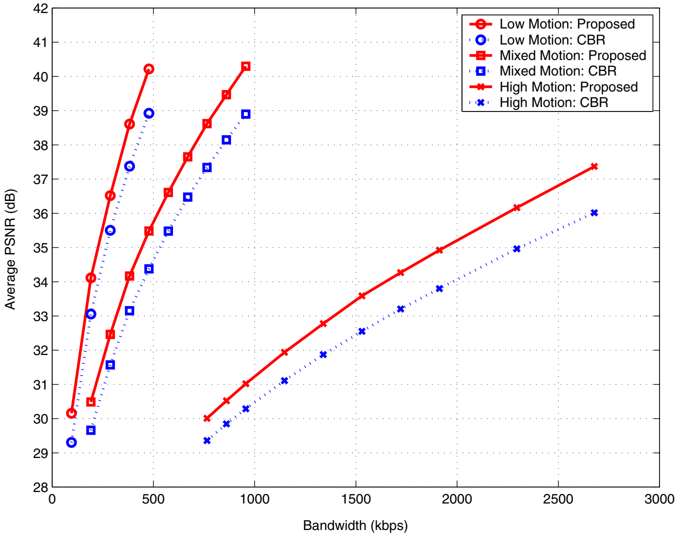


Fig. 6. PSNR vs. Bandwidth for Different Motion Activities

The second set contains *stefan*, *foreman*, and *coastguard*. The third set contains mixed level of motion video sequences, including *silent*, *foreman*, and *news*.

To demonstrate that the proposed framework can utilize the relay's bandwidth effectively to achieve better perceptual quality, we compare the constant bit rate (CBR) scheme which allocates equal amount of time slots for relay node to transmit parity check bits for each video source. In Figure 5, we show the average PSNR gain when we compare the proposed scheme and the CBR scheme for all three video sets. As we can see, the proposed scheme can have PSNR gain between 0.8dB and 1.3dB when the received video quality is between 30dB and 40dB, which is a noticeable quality improvement. The performance gain achieved by the proposed scheme is mainly contributed by jointly leveraging the diversity of different video source R-D characteristics and nodes' channel conditions; and dynamically allocating suitable amount of time slots to each video source. To further assess the impact of different level of motion activities, we show the PSNR performance for three different video sets in Figure 6. As revealed, the performance gain is consistent for all levels of motion activities owing to the dynamic resource allocation.

To evaluate how close the performance of the proposed scheme can approach to the optimal solution, we list the PSNR difference between the proposed scheme and the optimal solution in Table 1, 2, and 3. As shown in these three tables, the performance loss is only between 0.1dB and 0.3dB. Note that the computation complexity and the communication overhead to obtain the optimal solution are extremely high. The proposed distributed

Table 1. Performance gap: Low Motion

Bandwidth (kbps)	95.63	191.25	286.88	382.50	478.12
Optimal (dB)	30.36	34.29	36.83	38.82	40.44
Proposed (dB)	30.16	34.12	36.52	38.61	40.22
Gap (dB)	0.2	0.17	0.31	0.21	0.22

Table 2. Performance gap: High Motion

Bandwidth (kbps)	765	956.2	1530	1912.5	2677.5
Optimal (dB)	30.16	31.18	33.73	35.17	37.68
Proposed (dB)	30.01	31.02	33.59	34.92	37.37
Gap (dB)	0.15	0.16	0.14	0.15	0.31

Table 3. Performance gap: Mixed Motion

Bandwidth (kbps)	191.25	478.12	573.75	765.00	956.25
Optimal (dB)	30.65	34.34	36.89	38.89	40.54
Proposed (dB)	30.49	34.17	36.61	38.62	40.30
Gap (dB)	0.16	0.17	0.28	0.27	0.24

scheme can achieve similar video quality by requiring much lower computation and communication overhead.

6 Conclusions

In this article, we briefly introduce the concepts of cooperative communication, distributed source coding, and auction theory. We then present the proposed cooperative wireless video transmission protocols with distributed source coding using auction theory. The considered framework is formulated as an optimization problem to minimize the estimated distortion under the power and bandwidth constraints. We also evaluate four different cooperative schemes for the performance improvement over different scenarios. The proposed cooperative video transmission scheme has the best performance among all schemes, as long as the source and relay are closely located together. We further propose a quasi-auction based resource allocation for multi-user scenario. Compared to the performance upper bound which is complicated and unpractical, the proposed quasi-share auction can reduce the computation complexity, while the performance gap is only 0.1dB to 0.3dB. The future works include the extended study of multiple relays involved in the whole networks and comprehensive research on distributed protocols to handle the full-fledged cooperative networks where each node can serve as source, relay, and destination node simultaneously.

References

1. Sendonaris, A., Erkip, E., Aazhang, B.: User cooperation diversity, Part I: System description. IEEE Transactions on Communications 51, 1927–1938 (2003)
2. Cover, T.M., El Gamal, A.: Capacity theorems for the relay channel. IEEE Information Theory 25(5), 572–584 (1979)
3. Khojastepour, M.A., Sabharwal, A., Aazhang, B.: On the Capacity of 'Cheap' Relay Networks. In: Proc. 37th Annual Conference on Information Sciences and Systems, Baltimore, MD (March 2003)
4. Laneman, J.N., Tse, D., Wornell, G.W.: Cooperative diversity in wireless networks: Efficient protocols and outage behavior. IEEE Trans. on Information Theory 50(12), 3062–3080 (2004)
5. Huang, J., Han, Z., Chiang, M., Poor, H.V.: Auction-based Resource Allocation for Cooperative Communications. IEEE Journal on Selected Areas on Communications, Special Issue on Game Theory 26(7), 1226–1238 (2008)
6. Wang, B., Han, Z., Liu, K.J.R.: Distributed Relay Selection and Power Control for Multiuser Cooperative Communication Networks Using Buyer / Seller Game. In: Proceedings of Annual IEEE Conference on Computer Communications, INFOCOM, Anchorage, AK (May 2007)
7. Han, Z., Poor, H.V.: Coalition Game with Cooperative Transmission: A Cure for the Curse of Boundary Nodes in Selfish Packet-Forwarding Wireless Networks. In: Proceedings of 5th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt 2007), Limassol, Cyprus (April 2007)

8. Ng, T., Yu, W.: Joint optimization of relay strategies and resource allocations in cooperative cellular networks. IEEE Journal on Selected Areas in Communications 25(2), 328–339 (2007)
9. Bletsas, A., Lippman, A., Reed, D.P.: A simple distributed method for relay selection in cooperative diversity wireless networks, based on reciprocity and channel measurements. In: Proceedings of IEEE Vehicular Technology Conference Spring, Stockholm, Sweden (May 2005)
10. Savazzi, S., Spagnolini, U.: Energy Aware Power Allocation Strategies for Multihop-Cooperative Transmission Schemes. IEEE Journal on Selected Areas in Communications 25(2), 318–327 (2007)
11. Himsoon, T., Siriwongpairat, W., Han, Z., Liu, K.J.R.: Lifetime maximization framework by cooperative nodes and relay deployment in wireless networks. IEEE Journal on Selected Areas in Communications 25(2), 306–317 (2007)
12. Annavajjala, R., Cosman, P.C., Milstein, L.B.: Statistical channel knowledge-based optimum power allocation for relaying protocols in the high SNR regime. IEEE Journal on Selected Areas in Communications 25(2), 292–305 (2007)
13. Lin, B., Ho, P., Xie, L., Shen, X.: Optimal relay station placement in IEEE 802.16j networks. In: Proceedings of International Conference on Communications and Mobile Computing, Hawaii, USA (August 2007)
14. Su, W., Sadek, A.K., Liu, K.J.R.: Cooperative communication protocols in wireless networks: performance analysis and optimum power allocation. Wireless Personal Communications 44(2), 181–217 (2008)
15. Shutoy, H.Y., Gunduz, D., Erkip, E., Wang, Y.: Cooperative source and channel coding for wireless multimedia communications. IEEE Journal of Selected Topics in Signal Proc. 1(2), 295–307 (2007)
16. Girod, B., Aaron, A., Rane, S., Rebollo-Monedero, D.: Distributed video coding. In: Proceedings of the IEEE, Special Issue on Video Coding and Delivery, vol. 93(1), pp. 71–83 (January 2005)
17. Stankovic, V., Yang, Y., Xiong, Z.: Distributed source coding for multimedia multicast over heterogeneous networks. IEEE Journal on Selected Topics in Signal Processing 1(2), 220–230 (2007)
18. Han, Z., Liu, K.J.R.: Resource Allocation for Wireless Networks: Basics, Techniques, and Applications. Cambridge University Press, Cambridge (2008)
19. Su, G., Han, Z., Wu, M., Liu, K.J.R.: Multiuser Cross-Layer Resource Allocation for Video Transmission over Wireless Networks. IEEE Network Magazine, 21–27 (March/April 2006)
20. Huang, J., Berry, R., Honig, M.L.: Auction-based spectrum sharing. ACM/Kluwer Mobile Networks and Applications Journal (MONET) 11(3), 405–418 (2006)
21. Fattahi, A.R., Fu, F., van de Schaar, M., Paganini, F.: Mechanism-based resource allocation for multimedia transmission over spectrum agile wireless networks. IEEE Journal on Selected Areas in Communications 25(3), 601–612 (2007)
22. Fu, F., van der Schaar, M.: Noncollaborative resource management for wireless multimedia applications using mechanism design. IEEE Transactions on Multimedia 9(4), 851–868 (2007)
23. Proakis, J.: Digital Communications, 4th edn. Thomas Casson (2001)

24. Kim, B.-J., Xiong, Z., Pearlman, W.A.: Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT). IEEE Transactions on Circuits and Systems for Video Technology 10(8), 1374–1387 (2000)
25. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2006), http://www.stanford.edu/~boyd/cvxbook.html
26. Krishna, V.: Auction Theory. Academic Press, London (2002)
27. Binmore, K., Swierzbinski, J.E.: Treasury Auctions: Uniform or Discriminatory? Journal of Economic Design 5, 387–410 (2000)
28. Binmore, K., Swierzbinski, J.E.: A Little Behavioralism Can Go A Long Way. In: Binmore, K. (ed.) Does Game Theory Work? The Bargaining Challenge, ch. 8, pp. 257–276. MIT Press, Cambridge (2007)

Enterprise VoIP in Fixed Mobile Converged Networks

Kyungtae Kim and Chonggang Wang

NEC Laboratories America, 4 Independence Way, Suite 200, Princeton, NJ 08540
{kyungtae,cgwang}@nec-labs.com

Summary. Emerging markets are increasingly competitive under the heading of convergence and substitution, including device convergence, fixed-mobile convergence, fixed-mobile substitution, service convergence, VoIP (Voice over Internet Protocol) telephony substituting for circuit switched voice telephony, bundled offers of mobile, fixed, broadband and TV and, finally, truly unified communications and digital media services that are delivered irrespective of anytime, anywhere, and anyhow. With the increasing use of wireless networking, convergence technology combined with user mobility is expected to introduce a significant change in how and when people communicate, using a variety of new applications on ubiquitous wireless devices. This chapter aims to discuss VoIP over enterprise fixed mobile converged networks (FMCN) and points out some open issues. We first review VoIP basics including transport and signalling protocols, speech coding, and VoIP quality assessment model, followed by introducing enterprise FMCN architecture and service characteristics. Then we review the performance of VoIP over enterprise wireless networks and discuss voice over enterprise unified communications. Finally, several future directions including VoIP over femtocell, location/presence-enriched enterprise VoIP, and admission control in enterprise wireless network are briefly listed.

1 Voice Communication Networks

We have witnessed that voice services have been migrating from traditional circuit-based plain old telephone service (POTS) or public switched telephone networks (PSTN) telephony, and cellular telephony to recently emerging IP telephony or voice over IP (VoIP) for reasons including cost-efficiency and flexibility improvement, etc. POTS/PSTN and cellular networks[1] use circuit-switching technique, under which network resources are first allocated to establish a circuit from the sender to receiver before the start of communications. The allocated resources remain dedicated to the established circuit

[1] It's worth noting that future all-IP cellular networks use packet-switching technology to support voice services.

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 585–621.
springerlink.com

during the entire voice call: *circuit-switched network*. In packet-switched networks, each message is broken into and contained in smaller packets, each of which can take a different route to the destination where the packets are reassembled into the original message. Each packet consists of a portion of user data plus some control information in the form of header and/or tailer such as information for addressing and error correction: *packet switched network*. Voice calls over a traditional circuit-switched PSTN network are first digitized from analog voice signal, transmitted across thousands of miles, and finally converted back to analog once they get to the final destination (a home or office phone, for instance). During the transmission, several interconnected switches along the connected line remain open and occupied even while there is dead air and no conversation is taking place. The circuit is even open in both directions even if only one party is talking and the other is listening. This transmission method is very inefficient because it does not fully utilize the data transmission capacity of the dedicated line because of silence period in voice communications. Also it takes time to set up the connection. On the other hand, VoIP which relies on IP packet-switched data networks works differently. Rather than circuit switching, data packet switching used in VoIP sends and receives small chunks of voice data, called a packet, only when you need it - instead of in a constant stream. It also sends the data packets along whatever open Internet circuits that are available, which is much more efficient than using a dedicated line.

1.1 Diversified Communication Networks

Although POTS/PSTN telephony, cellular telephony, and IP telephony services have possessed their own territories and considered as a separate market respectively, people who subscribe and use those services with different voice terminal equipments should be able to communicate with each other even over different access networks and technologies such as POTS/PSTN networks, cellular networks, and IP networks. The same story applies for the same technology with the different service providers. As the wireless technologies advance shown in Figure 1, different systems in network design, architecture, standards, services, and terminals are developed and deployed especially in wireless world: from 1G to 4G including Time Division Multiple Access (TDMA), Code Division Multiple Access (CDMA), Orthogonal Frequency Division Multiple Access (OFDMA), Carrier Sensing Multiple Access (CSMA), satellite communication, sensor communication etc. In this situation, people have long sought out products and services that make their lives more convenient and simpler, give them access to new services while accessing multiple access technologies with a single device, and most importantly give them convenience as well as saving them money. Current network usage and device sales demonstrate the truth behind each consumer objective, like people's preference on hyper-converged mobile phones in Figure 2 that

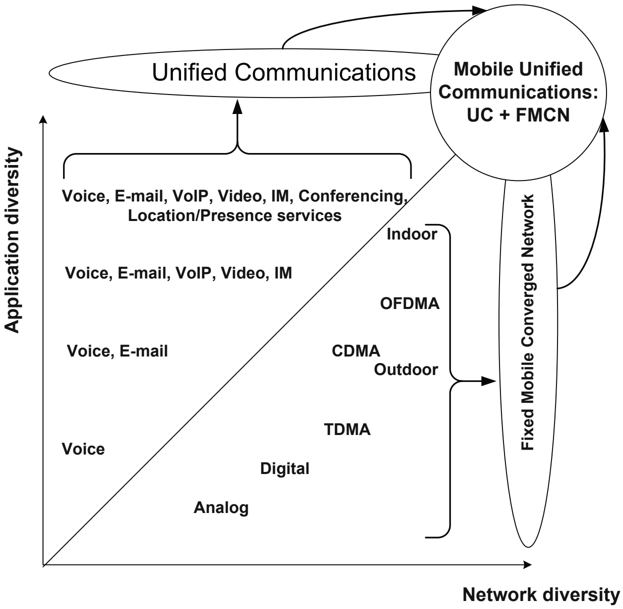

Unified Communications
Mobile Unified Communications: UC + FMCN
Application diversity
Voice, E-mail, VoIP, Video, IM, Conferencing, Location/Presence services
Voice, E-mail, VoIP, Video, IM
Voice, E-mail
Voice
Indoor
OFDMA
CDMA
Outdoor
TDMA
Digital
Analog
Fixed Mobile Converged Network
Network diversity

Fig. 1. Heterogenous networking technologies are merged into one platform: Fixed-Mobile Converged Network. Diversified communication tools/applications are integrated into one architecture: Unified Communications. Now these two architectures can be unified into Mobile Unified Communications in enterprise environment.

have a built-in camera, GPS, TV, music player, game player, projector, health inspector, wireless interfaces, and can manage your office documents. Device convergence is one of the most significant trends in current technology to substitute other computing and consumer electronics devices by integrating the functionalities of these devices into the phone platform. Especially these mobile phones are being equipped with various wireless interfaces from short-range communication technologies including infrared, Wi-Fi [1], UWB (Ultra-WideBand), Bluetooth, RFID (Radio-Frequency IDentification) [2] to long-range access networks, such as WiMAX (Worldwide Interoperability for Microwave Access) [3]. These technologies enable the phone to become a personal data gateway interacting with mobile networks in numerous ways. In this heterogeneous wireless access networking environment shown in Figure 3, convergence can be an attractive service to the customers who are suffered from selection of the best access network and poor quality of service due to his/her movement. Today, we view convergence as the migration of voice, data, and video services to a single consolidated network infrastructure which is based on both wired and wireless networks, of which network infrastructure is referred to as fixed-mobile converged networks (FMCN) [4].

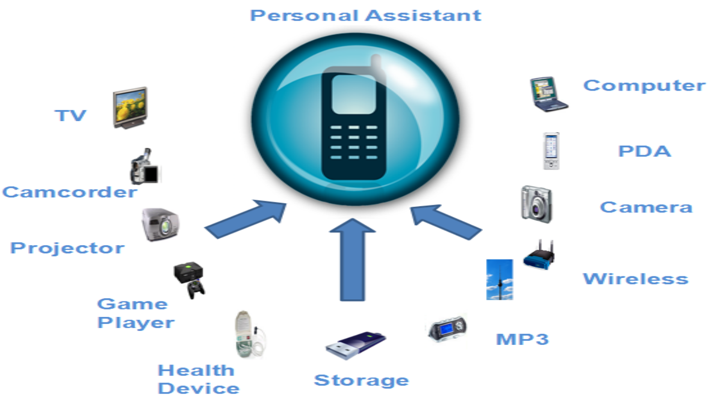


Fig. 2. Various multimedia functions converged into a single mobile device (Device Convergence), which is always with you and behaving as a personal assistant

1.2 Unification: Carrier-Centered vs. Enterprise-Centric

The aim of the FMCN is to provide the seamless transition of calls or communications in progress between outside cellular network and inside wired/wireless networks on a single-mode or multi-mode handset. The path of a call transits between outside networks and inside networks reaching to the cellular service provider over the subscribers' broadband network. To support communication continuity, there needs a mobility controller to handle the handover seamlessly over the path. This mobility controller can be located at and operated by either the service carrier or the enterprise. If it is in the service provider, it is called *carrier-centric convergence* and the seamless mobility can be supported through either unlicensed mobile access/generic access network (UMA/GAN) method or the voice call continuity (VCC) [5] technology of the IP multimedia subsystem (IMS) specification [6]. The UMA technology [7] encapsulates and transports the cellular packets over the broadband IP network, of which handset should implement the cellular protocol stack over the attached wireless networks. The carrier-based approach mainly aims at the consumer market because the subscriber doesn't need managing a call to control. The main requirements of the consumer are convenience and cheap. However the enterprise considers managing and controlling the communications as the most important features and does not allow its information to be processed and possessed by the service provider. Thus, the mobility controller should be located in the enterprise and enterprise IP-PBX can be an anchor point to support seamless continuity. Mobility aware IP-PBX extends the existing features and functionalities of the enterprise IP-PBX out to support mobile phones making them to behave like an extension over inside wireless network and this approach is called *enterprise-centric convergence.*

Fig. 3. Many customers may have no idea on which access technology is the best at the place. Convergence represents the coming together of all networks and management creating a unified network with call continuity with best access quality at anyplace, anytime, and anydevice.

UMA vs. VCC

The basic promise of convergence service is to enable subscribers using new dual-mode handsets or femtocell-aware single mode handset to automatically transition between outdoor mobile networks and indoor enterprise Wi-Fi/femtocell networks. Expanding coverage for better in-building usage is regarded as the most important reason for cellular operators to embrace these multi-mode handsets. One cannot expect a mobile subscriber to start a call in a Wi-Fi environment and hang up and re-initiate the call when they move outside into the cellular network from inside networks. Here it comes the need for seamless mobility and there are two approaches, mobile-centric UMA and fixed-centric VCC, to support service continuity [8].

UMA/GAN is a 3GPP standard defined by *the mobile community* to extend voice, data, and IMS services over IP access networks while having minimal impact on operators' core networks already in place. It allows mobile operators to leverage the cost and performance advantages of IP access technologies when delivering high-quality, low-cost mobile voice and data services in the location where subscribers are stationary such as in home, office or public hotspot. The underlying architecture is one in which the GSM/UMTS signalling and media streams are tunneled over an IP transport in attached wireless network which is non-cellular network. This is accomplished by the client establishing an IP link with a corresponding UMA point-of-presence network element called a UMA network controller (UNC) through encapsulating voice packet with cellular protocol. To the cellular network, such devices appear to be functionally identical to a standard cellular phone. Because of speaking cellular protocol between the UNC and the handset client, the functions supported at the handset are identical, regardless of whether they are in cellular or 802.11 coverage.

VCC is a 3GPP-defined specification defined by *the fixed community* that describes how an existing voice call continues even as a mobile phone moves between circuit-switched system (GSM/UMTS) and packet-switched radio domains including Wi-Fi, UMTS PS (packet-switched) network and vice versa with complete transparency and seamlessness from an end-user point of view. As with UMA, VCC is defined to take advantage of broadband and 802.11 networks in homes and businesses. It is the same as UMA that the call is anchored in the cellular network, but the major difference is that the VCC client is based on the SIP (Session Initiation Protocol) standard and not a cellular phone emulation model like UMA. *This approach has advantages in its compatibility with enterprise IP-PBX solutions and IMS networks, since most of these products use SIP protocol as a controlling protocol.* When the phone detects as available 802.11 signal or other wireless networks inside, it will use SIP to create the new session over the 802.11 broadband IP network. Thus, VCC differs from UMA in that VCC is a SIP-centric approach to the convergence.

1.3 Unifying Diverse Applications

Desktop and smart mobile phones are equipped with broad rage of applications to support communications: e-mail, voice mail, instant messaging (IM) and social networking, conferencing, video, and web which are shown in Figure 1. These applications are offering suitable capabilities to the right places and right time with their own communication platform like different methods and devices. However, it is challenging to find the best way at right place and time and even retrieve the messages from the different applications. Thus, people have sought a way where calls and messages would reach each user regardless of location or means of access technologies, allowing them to use the best communication method and to retrieve e-mail, instant messages, or voice mail from a common storage at any location. This is a promise of *Unified Communications* (UC).

1.4 Unifying Network and Application Diversity

Mobile unified communications (MUC) removes the barriers between fixed mobile network convergence and unified communications by unifying them into one architecture: voice, email, conferencing, video instant messaging, mobility, and location/presence services, allowing people to connect, communicate and collaborate from multiple locations with their available devices in Figure 1. By providing the tools to inside/outside employees to keep staying available and productive no matter where they are and no matter how they can be accessed, MUC is increasingly deployed into enterprise communication platform by integrating into enterprise communication controller, such as IP-PBX.

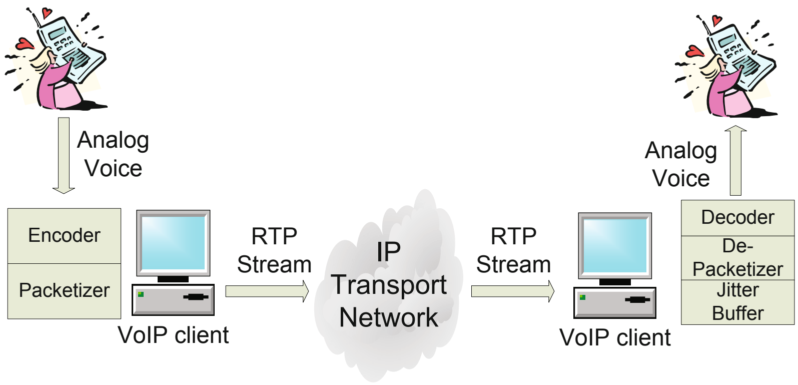


Fig. 4. VoIP System

There are some proposals about FMCN architecture and unified communications in literature, but what's the impact of FMCN on services especially VoIP and how VoIP could be supported better in such emerging FMCN are not well understood yet, especially for enterprise VoIP. This chapter aims to discuss enterprise VoIP in FMCN. Some background knowledge about VoIP is introduced in Section 2. Enterprise FMCN architecture is described in Section 3, where wireless networks play an important role in the full picture. VoIP over specific wireless networks and enterprise MUC (EMUC) are explained respectively in Sections 4 and 5. Finally, Section 6 concludes the whole chapter and lists several future directions.

2 Voice over IP (VoIP)

VoIP [9] is the transmission technology for voice communications over IP networks. It employs session control protocols to handle session establishment, termination, and negotiate session parameters, such as audio codecs which is responsible for converting analog voice signal to and from digitized voice bits. Voice packets are transmitted using Real-time Transport Protocol (RTP) and User Datagram Protocol (UDP). Figure 4 illustrates a typical VoIP systems. VoIP traffic is different from data traffic and has the following unique characteristics:

- **Low Bandwidth:** VoIP packet consumes very little bandwidth, for example 20 bytes voice payload for G.729a.
- **Large Overhead:** Compared to the voice payload, the header overhead for IP/UDP/RTP consumes the larger part of the bandwidth, i.e. 40 bytes.
- **Delay Sensitivity:** Non real-time data, even data streaming, can tolerate delays of a few seconds or more without impacting user experience. In contrast, VoIP traffic can only allow an end-to-end delay of 150 to 400 ms.

- **Jitter Sensitivity:** Jitter is defined as a statistical variance of the RTP data packet inter-arrival time and is one of the three most common VoIP problems, which are delay, jitter, and packet loss. If the jitter is so large that it causes packets to be received out of the range of the playout time and these packets are discarded, it causes the perceived voice quality to degrade.
- **Loss Sensitivity:** VoIP traffic is sensitive to frame loss rate due to no end-to-end retransmissions allowed and there is a threshold on delay vs. loss ratio to support the perceived voice quality. Although some codecs include the information to correct the error and can tolerate higher losses, common protocols such as G.711 or G.729 can only operate if the loss rate is within certain amount of threshold.
- **Uniform and Smooth Traffic:** VoIP produces the payload usually every fixed amount of time and needs regularity in the network delay and a low loss rate. This continuity should be kept in the receiver side while bursty data traffic interferes on-time transfer.

Playout/Jitter buffer: The network delivers voice packets in best way with variable delays. To be able to play the audio stream with reasonable quality, the receiving endpoint needs to make the variable delays into constant delays. This can be done by using a jitter or playout buffer. The jitter buffer places the voice packets to the queue to hold, not playing out as soon as it receives the voice packets. The delayed packets starts to play out after the buffer reach to the threshold, for example 60 or 100 ms. We can increase the size of jitter buffer to reduce buffer overflow or underflow, it simultaneously increases the end-to-end delay.

VoIP is susceptible to network conditions (delay, jitter and packet loss), which can degrade the voice quality to the point of being unacceptable to the average user. Managing voice quality over both IP-based wired and wireless networks has become a challenge, especially in a heterogeneous network environment. For a VoIP system the most significant influences on a user's perception of quality are a function of many factors that include the performance of codec system, errors and frame loss, echo cancelation, and delay. For example, a proper selection of the VoIP codecs allows the system to react to the random variations in capacity due to transmission rate changes, which are motivated by the wireless channel and the movement of users. There are several issues that need to be addressed in order to provide a toll-quality, PSTN equivalent end-to-end VoIP network. These include:

- **Quality of Service (QoS):** One of the key requirements for the widespread deployment of VoIP is the ability to offer a toll quality service equivalent to the existing PSTN. Perceived voice quality is very sensitive to three key performance criteria in a packet network: delay, jitter and packet loss. The followings are a set of minimal requirements for VoIP applications set by WiMAX Forum: a) Typical Data Rate: 4-384 Kbps; b) Delay: < 150 ms; c) Jitter: < 50 ms; d) Packet Loss: < 1.0 %.

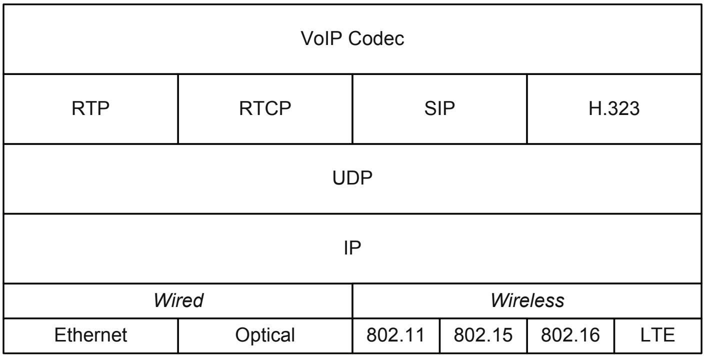


Fig. 5. VoIP Protocol stack over wired/wireless hybrid networks

- **Signalling Protocols:** Numerous different signaling protocols have been developed that are applicable to a VoIP solution, such as H.248, H.323, Media Gateway Control Protocol (MGCP), Session Initiation Protocol (SIP), etc.
- **Speech Coding Schemes:** Different voice codecs adapt to different network and application environments and lead to different voice quality.
- **Security:** PSTN has been very resistant to security attacks and have not suffered from significant problems since the introduction of SS7 out-of-band signaling. A VoIP network is much more susceptible to security attacks and must address the security issues including, denial of service, theft of service, and invasion of privacy.

2.1 VoIP Protocols

VoIP consists of a set of supporting protocols. As there are drivers to deploy VoIP in many different ways, it can be difficult to define a dominant VoIP signaling or encoding method. However, for the scope of this book chapter, we shall focus mainly on the signaling and media transport protocols used in the IETF, ITU-T, and 3GPP standards such as SIP and H.323 for signaling and RTP for media transportation.

Figure 5 shows the protocol suites to support VoIP. Two main types of traffic carrier upon IP are UDP and Transmission Control Protocol (TCP). In general, TCP is used when a reliable connection is required and UDP on simplicity. Due to the time-sensitive nature of voice traffic, UDP/IP is the logical choice to carry voice for media transport as well as signaling protocol. VoIP signaling protocols are used to set up and tear down calls, carry information required to locate users and negotiate capabilities. There are a few VoIP architectures, derived by various standard bodies and vendors, that are based on a few signaling protocol stacks, namely H.323 and SIP representatively. In the following section, the main characters of H.323 and SIP are explained.

RTP

RTP [10] was developed to enable the transport of real-time packets containing voice, video, or other information over IP. RTP is defined by IETF Proposed Standard RFC 3550. RTP does not provide any quality of service over IP network. RTP packets are handled the same as all other packets in an IP network. However, RTP allows for the detection of some of the impairments introduced by an IP network, such as packet loss and out of sequence packet arrival with sequence number, variable transport delay with time stamp.

H.323

H323 [11] is an ITU-T umbrella standard released in 1996, which consists of signaling and transport and coding protocols. H323 is a multimedia conferencing standard and mainly used in professional video conferencing systems, but also used for pure VoIP applications. Part of the design is to specifically tackle the interconnection with PSTN by means of a gateway.

SIP

SIP [11] is an application-layer *signalling* protocol that can be carried over UDP or TCP. As an application protocol, SIP operates independently of the network layer and requires only packet delivery service from the accessed network. The basic function of SIP is to locate the person, to ring the terminal, and to establish a connection which is already negotiated with the necessary parameters for media transport protocol. SIP has been designed with easy implementation, good scalability, and flexibility in mind. SIP endpoints are called user agents (UAs). Thus the user agent client (UAC) is the calling party, and the user agent server (UAS) is the called party. Except user agent, there are server components in SIP architecture. SIP servers are intermediary components that are located within the SIP-enabled network and assist user agents in session establishment and other functions. There are three types of SIP servers:

- **SIP Proxy Server:** it receives SIP request from a user agent or another proxy and forwards or proxies the request to another location.
- **SIP Redirect Server:** it receives a request from a user agent or proxy and returns a redirection response, indicating where the request should be redirected.
- **SIP Registrar Server:** it receives SIP registration requests and updates the user agent's information into a storage for location service.

SIP is not limited to establish a signaling path for voice communications and can in fact be used to set up video, text (for instant messaging), and other types of media sessions. The exact media components in the multimedia session are described using a Session Description Protocol (SDP) [12], which provides the two endpoints with information about media, such as the format of the media (G.711, Adaptive Multirate-AMR, etc.), the type of media (video, audio, etc.), and the transport protocol (RTP, etc.). Opposed to H.323 which is an umbrella standard, the purpose of SIP is just to make the communication possible. The communication itself must be achieved by other means and protocols/standards. SIP has been designed in conformance with the Internet model. It is an end-to-end-oriented signalling protocol which means that all the logic is stored in end-devices (except routing of SIP messages).

2.2 Speech Coding

VoIP works by taking analog audio signals and turns them into digital data which can then be transmitted over the Internet. The well-known principle of a digital modulation is to modulate an analogue signal with a digital sequence in order to transport this digital sequence over a given medium: fibre, radio link etc. This has great advantages with regard to classical analogue modulation: better resistance to noise, use of high-performance digital communication and coding algorithms etc. To digitize analog speech, an analog-digital converter (Coder) [13] samples the value of the analog signal repeatedly and encodes each result in a set of bits. Another identical codec (Decoder) at the far end of the communication converts the digital bitstream back into an analog signal. Most domestic PSTN networks operate with speech sampled at 8 kHz and an 8-bit nonlinear quantization according to ITU-T G.711. This encodes at 64 kbit/s and introduces little audible distortion for most types of signal. In a number of applications, however a much lower bit rate is desirable either because capacity is limited, i.e. wireless/mobile environment, or to maximize the amount of traffic or calls that can be supported under a limited bandwidth. Some popular codecs for VoIP are listed below[2].

G.711

G.711 is a high bit rate (64 kbit/s) ITU standard codec. It is the native language of the modern digital telephone network. The codec has two variants: A-law and U-law. It works best in local area networks where we have a lot of bandwidth available with the large size of the payload and low packet error rate. Its benefits include simple implementation which does not need much CPU power and a very good perceived audio quality. The downside is that it takes more bandwidth then other codecs, up to 84 Kbit/s including all UDP/IP overhead. G.711 is supported by most VoIP providers - the MOS value is 4.2.

[2] In addition, ITU recommendation G.718 is an embedded codec which provides variable rates at 8 kbit/s and 32 kbit/s for 8 and 16 kHz sampled speech and audio signals. ITU G.719 codec is a recently approved ITU standard for high-quality conversational fullband audio coding while introducing very low complexity and being hight practical.

G.729

G.729 is a popular codec that significantly reduces the bandwidth requirements for a IP voice signal but still provides good audio quality (MOS = 3.9) - from the standard payload size of 64 kbit/s of G.711 down to 8 kbit/s. There are various versions of G.729 (sometimes called G.729a, G.729b or G.729ab) that further reduce the voice payload size to 6.4 kbit/s or less. The codec algorithm encodes each frame to 10 bytes per 10 ms, so the resulting bitrate is 8 kbit/s for one direction. When used in VoIP, it usually sends 3-6 G.729 frames in each packet to reduce the overhead of packet headers. G.729 is the most widely used low bitrate codec and is universally supported by the major VoIP equipment manufacturers.

G.723.1

G.723.1 allows calls over 28.8 and 33 kbit/s modem links. It operates on audio frames of 30 milliseconds (i.e. 240 samples), its bitrate is 5.3 kbit/s with MOS=3.7.

GSM 06.10

GSM 06.10 is also known as GSM Full Rate and operates on audio frames 20 milliseconds long (i.e. 160 samples) and it compresses each frame to 33 bytes, so the resulting bitrate is 13 kbit/s with MOS=3.7.

Speex

Speex is an open source patent-free codec designed by the Xiph.org Foundation. It is designed to work with sampling rates of 8 kHz, 16 kHz, and 32 kHz and can compress audio signal to bitrates between 2 and 44 kbit/s. For use in VoIP telephony, the most usual choice is the 8 kHz (narrow band) variant.

iLBC (internet Low Bit Rate Codec)

iLBC is a free codec developed by Global IP Solutions. The codec is defined in RF C3951. With iLBC, you can choose to use either 20 ms or 30 ms frames and the resulting bitrate is 15.2 kbit/s and 13.33 kbit/s respectively. Much like Speex and GSM 06.10, you will find iLBC in many open source VoIP applications.

2.3 VoIP Quality Assessment: E-Model [14]

A VoIP system shown in Figure 4 consists of an encoder-decoder pair and an IP transport network. The choice of vocoder is important because it has to fit the particularities of the transport network (loss and delay). One of the

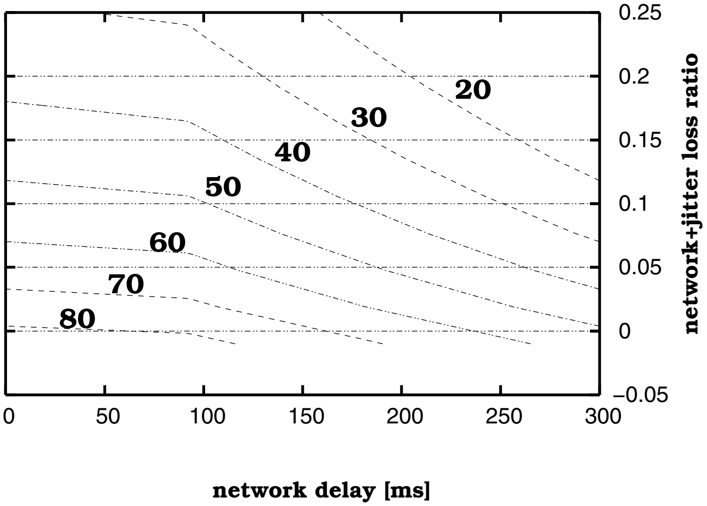


Fig. 6. *R-score* for 60ms jitter buffer [15]

popular voice encoders is G.729, which uses 10ms or 20ms frames. It is used by some available 802.11 VoIP phones. VoIP application with G.729 codec sends 50 packets per second, of 20 bytes each. Although a 30% utilization increase is generally expected when accounting for periods of silence when no packets are sent, we do not consider silence periods in this book chapter. To measure the quality of a call, we use a metric called E-Model, which takes into account mouth to ear delay, loss rate, and the type of the encoder. Quality is defined by the *R-score* (a scalar measure that ranges from 0 (poor) to 100 (excellent)), which for medium quality should provide a value above 70:

$$\begin{aligned} R = {} & 94.2 - 0.024d \\ & - 0.11(d - 177.3)H(d - 177.3) \\ & - 11 - 40\log(1 + 10e) \end{aligned}$$

where:

- $d = 25 + d_{jitter\_buffer} + d_{network}$ is the total ear to mouth delay comprising 25 ms vocoder delay, delay in the de-jitter buffer, and network delay
- $e = e_{network} + (1 - e_{network})e_{jitter}$ is the total loss including network and jitter losses
- $H(x) = 1\, if\, x > 0;\, 0\, otherwise$ is the Heaviside function
- the parameters used are specific to the G.729a encoder with uniformly distributed loss

The constants consider the delay introduced by the encoder for its lookahead buffer, and the delay introduced by the jitter buffer, which has two contradictory effects: it increases end to end delay, therefore degrading the quality, but it also reduces the delay variance - jitter, which has an overall better effect. The *R-score* is finally computed only from the loss and the delay in the network, which can be measured directly. Loss in the jitter buffer is considered as the fraction of packets which do not meet their deadlines. In order to compute loss probabilities and average delay in the network, all packets from all flows in a tested network are considered together.

Figure 6 shows the values of the *R-score* with respect to network delay and total loss for 60 ms jitter buffer and 25 ms vocoder delay. The interpretation of the iso-*R-score* curves is that for example to obtain an *R-score* of 70, the network has to deliver all packets in less than 160 ms, or deliver 98% in less than 104 ms [16]. From the figure we can see that the quality is sensitive to even a couple of percents of loss, whereas the delay tolerates differences in tens of milliseconds. For example, in wireless network, loss has a high variance, as it depends on the quality of the channels and the interface cards, and on the interference from external or internal sources. In the FMCN scenario, end to end loss is difficult to control and needs to be maintained under 2%. However, using the retry mechanism of lower layer within the delay budget, this loss can be reduced at the cost of increasing delay.

3 Enterprise FMCN

Modern automated offices of SOHO (Small Office, Home Office) or enterprise may contain desk phones, fax machines, PCs surely with Internet connections (DSL, Cable, or Fiber), PDAs (Personal Device Assistants), smartphones, laptops with broadband wireless access (WiMAX or Long Term Evolution (LTE) [17]), and wireless cellular telephones (and in the drawer, more than one cellular phones for international trip for business): *multiple devices and heterogeneous access networks.* Here people hardly consider the desk phones as the analog telephones, but VoIP telephones or even video telephone. Currently, the users or subscribers depend on their intelligence to choose the right devices after checking out what access technologies are available at the place. There is no other choice for the users: *all intelligence to choose the right access network is on users or subscribers.* Then, given the proliferation of advanced technologies and devices, there is no surprise that people are now looking for ways to unify and converge these devices, wireless/wired services, communication technologies and access networks in enterprise environment. The user brings just one device which handles all sophisticated process to locate the best access network at the right place and establish the communication at the best quality. This convenience helps the business process their work easily while eliminating several devices to bring as well as providing always being connected to the access network at the right time: *the access network collaborates with the device to provide the best network, while hiding all details from the user.* This is why Enterprise FMCN is getting attention and VoIP over FMCN is becoming a viable solution for mainstream consumers - both businesses and individuals. For subscriber, this service provides a single wireless

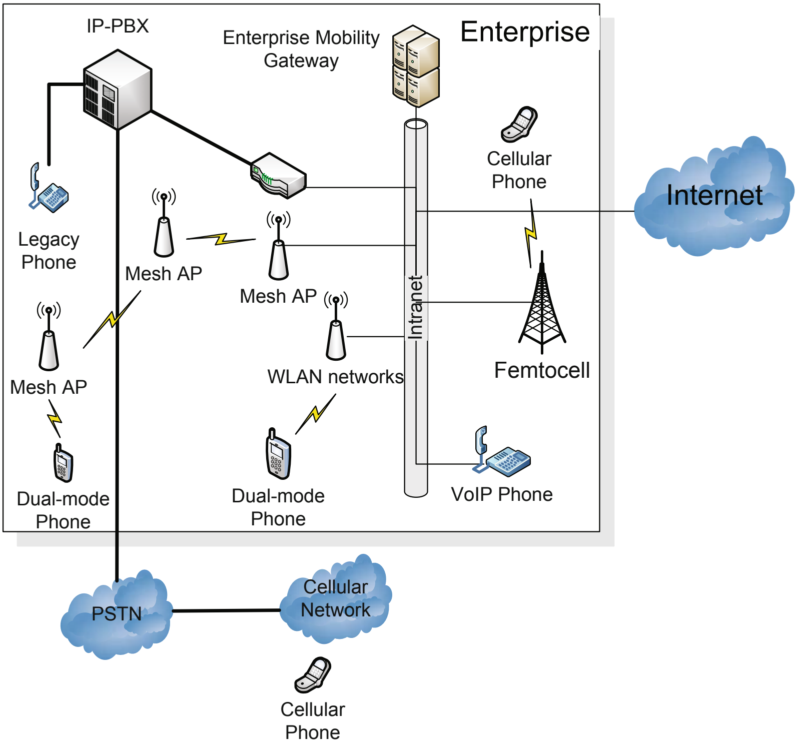


Fig. 7. Enterprise Fixed-Mobile Networks: people communicate over fixed network - Ethernet as well as over mobile networks - Wi-Fi, wireless mesh, and femtocell. All communications are managed by Enterprise IP-PBX, which is the heart and brain of an enterprise communication network.

phone and a personal number while increasing mobile convenience. In the office, people uses Wi-Fi or small size base station, Femtocell [18], to connect to a broadband service, such as Internet and hands over from/to outside cellular networks. Here are the benefits. First, easy access to communication and Internet services everywhere: *convenience*, second, quick, seamless sharing of work, information and experiences within enterprise: *higher performance for employees and better management and control with additional features*, then a single bill from one trusted service provider that offers the whole range of preferred communication and Internet services in both fixed and mobile environment: *saving money.*

3.1 Convergence of Fixed and Mobile Services/Networks

FMCN aims to provide seamless connectivity between fixed and wireless telecommunications networks. Here, "Fixed" and "Mobile" are service viewpoint, while wireline and wireless are technology viewpoint. The technological convergence of wireline and wireless offers a way to connect a mobile phone to a fixed line infrastructure so that operators can provide services to their users irrespective of their location, access technology and terminal: *network-level convergence.* Fixed refers to communication that requires the user to maintain its physical location relatively constant. Mobile communications assume it is possible for a mobile user to maintain seamless connectivity to a network or a communication peer, or continue using communication services while changing location. With the converged services across fixed and mobile environments, a mobile user would stay connected and oblivious to changing conditions as if these events were not happening and that user were not moving at all across fixed and mobile networks while accessing the services: *service-level convergence.*

For fixed operator, FMCN provides acquisition of new subscribers by new services, mainly with existing infrastructure or partly shared infrastructure. For example, cable operators can acquire new voice subscribers and offer new broadband services by exploiting unlicensed wireless access (Wi-Fi) and existing fixed broadband lines into homes and businesses. Thus, fixed operator can sell mobility in fixed network *with dual-mode handset.* Mobile operator reaches home or enterprise in mobile network. FMCN allows mobile operator to replace the wireline device without causing wireless network congestion with mobile handset. The handset uses the wireline broadband for in-home or in-building calls *with small-size femtocell station.*

FMCN first offers the fixed and mobile operators to retain their customers and increase revenues with easy-to-use services accessible via any network, any device and any place, secondly reduce investment and operation costs, taking advantage of common networks and central maintenance. Handsets have become multifunction devices built-in cameras, with location with GPS, MP3 players, voice recognition, and high-quality video displays, while functioning as a personal secretary. Nowadays, the mobile device and wireless network have become central to the converged lifestyle communications of today's people. Undoubtedly, FMCN has become catching the latest all terms to refer to the concept of merging the cellular network to a non cellular network, such as Wi-Fi and the next big step in the evolution in telecommunication networks.

FMCN enables subscribers to use resources from the mobile and fixed network transparently to the users as well as allow to move active voice calls seamlessly between fixed and mobile networks. FMCN has only recently begun to gain attraction as a viable service offering, driven by the interaction of three industry trends:

- **Mobile Phone Everywhere:** According to U.N. figures published at 2009, more people are using cell phones and other mobile devices for calls and high-speed data service than that for fixed network and this trends will continue to increase over time. Mobile phones offer more than just mobility; they serve the role of a general-purpose communications device.

- **Wireless Everywhere:** Wi-Fi and femtocell have become a common approach for connecting mobile phones, PCs and other devices to broadband communications resources, both in homes and enterprises.
- **VoIP Everywhere:** With the tremendous growth of the Internet and private data networks, service providers realized that it is more cost-effective to carry all of their traffic, including voice, over packet-based IP networks than carrying them over circuit switches. The trend toward VoIP started in the core voice infrastructure but is now spreading to the edge and becoming a popular choice for residential and enterprise telephone service. Eventually, even mobile phones are expected to become VoIP endpoints.

Given these trends, many service providers are looking forward to a convergence of fixed and mobile wireless communication, i.e. FMCN. The benefits they hope to attain include:

- **Wider Coverage:** Wireless access to a fixed network, via Wi-Fi, femtocell, or other technology, can provide connectivity when the handset is out of the range of a conventional macro cellular signal, i.e. inside the building or underground.
- **Better Voice Quality:** A number of factors, such as being inside a building, can degrade cellular signals and compromise voice quality. Wi-Fi and femtocell can deliver a cleaner signal for a superior subscriber experience.
- **Optimized Cost and Control:** Many enterprises already support wireless access to the corporate backbone in their plants and offices. Such companies can reduce their communication costs and gain greater control over usage if their employees send voice traffic over the internal network when at work or on traveling. Moreover, with VoIP, enterprises can exploit low-cost public IP networks, including the Internet, to further reduce their phone bills.
- **Conservation of Scarce Resources:** The radio spectrum available for mobile networks is limited. Steering calls to/from a stationary subscriber over a fixed network frees up precious bandwidth for other subscribers and services on the move.

These factors, as well as the rapid industry-wide acceptance of all IP-based networking system, have sparked interest in FMCN. The internationally recognized all IP-based communication standard (IP Multimedia Subsystem, IMS) was originally specified by the 3GPP, but has since been extended to all fixed and mobile networks and has been accepted by many other standards bodies including 3GPP2, ETSI and CableLabs. Indeed, IMS is increasingly being viewed as the core of all next-generation network architecture.

3.2 Service Characteristics in FMCN

While characteristics of convergence services vary among operators and vendors, a viable FMCN service must have three major features:

- **One Unique Number with single voicemail:** The subscriber must be reachable with a single phone number regardless of the network the handset is connected to at the time. Subscriber benefits from the convenience of single-number and single-voicemail service coupled with the freedom to control call handling depending on context (work, home, mobile).
- **Service Continuity:** An active call must continue without interruption as a subscriber moves between fixed and mobile networks.
- **Service Consistence:** Most or all services should present a common user experience across fixed and mobile networks.

4 Voice over Enterprise Wireless Networks

4.1 Challenges of VoIP over Wireless Networks

Although VoIP has become very popular and successful in wireline systems, it is still in its infancy in the wireless system, and many technical challenges remain. Under the environment which requires good perceived voice quality and guaranteed QoS like enterprise, it still has many challenging issues to solve.

- **Loss, Delay, and Jitter Control:** In wireline systems, channels are typically clean and end-to-end transmission can be almost error-free in case of no congestion, requiring no retransmissions. However, a wireless channel could be lossy, resulting in bit errors and corrupted/lost packets. Packets may have to be retransmitted multiple times to ensure successful reception at the lower layer, i.e. link layer, and the number of retransmissions depends on the wireless link condition. This could introduce significant delay and large delay variations which cause the perceived voice quality to be deteriorated. Further, unlike the circuit channels, which has a dedicated fixed bandwidth for continuous transmission, packet transmissions are typically bursty and share a common channel that allows multiplexing for efficient channel utilization. This operation also results in load-dependent delay and jitter.
- **Spectral Efficiency:** In a wireline VoIP system, bandwidth is abundant, and it is often used to trade-off a shorter delay. In fact, more bandwidth-guaranteed circuit-switched transmissions have been abandoned in favor of the flexibility of packet-switched transmissions, even though packet transmissions incur extra overheads. In wireless systems, however, the spectrum resource is generally regarded as the most expensive resource in the network, and high-spectral efficiency is vitally important for service providers or network operators. Therefore, mobile VoIP systems must be

designed such that they can control delay and jitter without sacrificing spectral efficiency. Packet transmission overheads must also be kept to a minimum over the air interface.

- **Mobility Management:** In many packet-based transmission systems, mobility management has been designed mainly for data applications. When mobile users move among cell sites, the handover procedure follows either the break-before-make or the make-before-break principle. The make-before-break provides better performance while requiring more resources. The break-before-make method leads to a larger transmission gap when the mobile is being handed over from one cell site to another. While a transmission gap is often acceptable in data applications, it can be unacceptable for real-time applications like voice. To support VoIP applications, the handoff design must be optimized so that the transmission gap during the handoff is minimized and does not impact perceived voice quality [19] [20].

4.2 VoIP over IEEE 802.11-Based Wi-Fi Networks

Wi-Fi/WLAN network has been one of the most successful technologies in the recent years. It is a fact that it has been deployed everywhere: companies, educational institutions, airports, cafeterias, homes, and especially enterprises. Also, it is frequently reported that the government of a city is to commence a project to offer public Internet access, i.e. hotspots, by using Wi-Fi technology. Two kinds of WLAN architectures or operation modes exist: Infrastructure mode, which is the most common mode and uses a centralized coordination station, usually called an Access Point (AP), for the scheduling of transmissions. All traffic goes via the AP. As another operating mode, an ad hoc mode network works without this centralized element and therefore it needs a routing protocol and special coordination method to provide reliable end-to-end communications between users, which is used for wireless multihop network. In this book chapter, we use the term Wi-Fi and WLAN interchangeably.

VoIP Capacity

As a base protocol in 802.11 networks, 802.11b is the first of the Wi-Fi standards to become popular and is being used for the testbed in this chapter. Intuitively an 802.11b 11 Mbit/s channel should be able to support up to 85 G.711 VoIP calls: Half of the bandwidth: 5.5 Mbit/s / 1 unidirectional G.711: 64 kbit/s $\simeq$ 85 flows. However, in reality, only 6 calls are supported and this means some overhead consumes the large part of bandwidth in Wi-Fi network. The reason is that the 802.11 MAC protocol overhead (protocol procedure overhead + MAC header overhead) consumes large portion of available bandwidth [21].

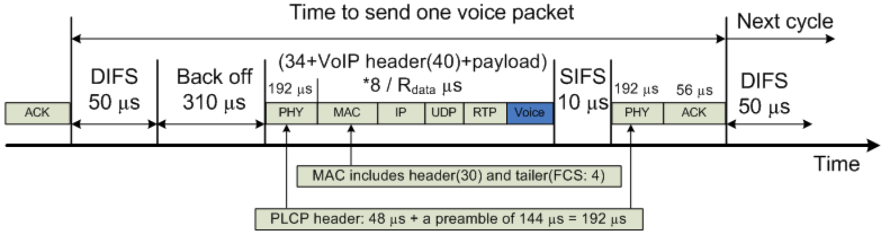

Time to send one voice packet
Next cycle
ACK
DIFS
50 μs
Back off
310 μs
(34+VoIP header(40)+payload)
*8 / R_data μs
192 μs
PHY
MAC
IP
UDP
RTP
Voice
SIFS
10 μs
192 μs
56 μs
PHY
ACK
DIFS
50 μs
Time
MAC includes header(30) and tailer(FCS: 4)
PLCP header: 48 μs + a preamble of 144 μs = 192 μs

Fig. 8. Basic Operation of 802.11 DCF

VoIP Overhead

The 802.11 MAC Layer defines two different access methods, the mandatory contention-based Distributed Coordination Function (DCF) and the optional polling-based Point Coordination Function (PCF). At present, DCF is the dominant MAC mechanism implemented by IEEE 802.11-compliant products. Furthermore, PCF is basically considered unsuitable for wider deployment, because of the lack of centralized control and the requirement of global time synchronization. Therefore, in this book chapter, we consider on DCF access method only, of which basic access mechanism is a Carrier Sense Multiple Access with Collision Avoidance mechanism (CSMA/CA). While operating in distributed way, CSMA/CA protocol intrinsically has comparative overhead to send a packet. The current IEEE 802.11b standard can support date rates up to 11 Mbps. A VoIP flows, when G.729a codec is used, requires 8 kbit/s. Ideally, the number of simultaneous VoIP flows that can be supported by an 802.11b WLAN is around (11Mbit/s)/(8Kbit/s) = 1375, which corresponds to about 687 calls, each with two VoIP flows. However, it turns out that the current WLAN can support only a few VoIP sessions. For example the maximum number of VoIP calls with G.729a in 11 Mbit/s is around 12, a far apart from the estimate. This result is mainly due to the added packet-header overheads as the short VoIP packets traverse the various layers of the standard protocol stack, as well as in the inefficiency of the 802.11 MAC protocol.

As most vocoders use samples of 10-100 ms, a node is expected to get a large volume of small packet traffic. However, 802.11 networks incur a high overhead to transfer one packet, therefore small sizes of packets reduce the network utilization. The problem with small payloads is that most of the time spent by the 802.11 MAC is for sending headers and acknowledgments, waiting for separation DIFS and SIFS, and contending for the medium. With the basic understanding of Wi-Fi, we calculate 802.11b protocol overhead needed for transferring one VoIP packet. Let us assume a VoIP packet consists of 40 byte IP/UDP/RTP headers (20+8+12) and a payload 20 bytes per 20 ms when using G.729a codec. The IEEE 802.11/802.11b standard defines SIFS to be 10 usec. A slot time is 20 usec and the value of DIFS is defined to be the value of SIFS plus two slot times which is 50 usec. The size of an acknowledgment frame is 14 bytes which take about 10 usec to transmit at 11 Mbit/s. However, each transmitted frame also needs some physical layer overhead (PLCP header of 48 usec and a preamble of 144 usec) which is about 192 usec. Thus, the total time to transmit an acknowledgment is 203 usec. The IEEE 802.11b standard defines CWmin to be 31. Therefore, in the scenario of a single node constantly transmitting, the average random back-off time is 15.5 slots which equals 310 usec. For the actual data frame we have an overhead of 34 bytes for the 802.11 MAC header, 20 bytes of IP header, 8 bytes of UDP header and 12 bytes of RTP header totaling 74 bytes which take about 54 usec to transmit at 11 Mbit/s. Together with the 192 usec physical layer overhead this amounts to 246 usec. Summing up these values, the time needed per VoIP frame as illustrated in Figure 8 can be calculated as 50 + 310 + 192 + 53 + (data/11Mbps) + 10 + 192 + 11 = 818 + payload/11Mbps usec. If G.729a codec is used for the payload (20bytes per packet = 20 x 8 / 11,000,000 = 14.5 usec) and 50 packets in one second, the maximum VoIP capacity in one hop is 12 calls (1 second / 832.5 usec x 50 x 2 = 12.01 12 calls). Capacity of Wi-Fi system is dependent on many factors, and one of factors is the header overhead associated with small packet sizes. The per frame overhead in the IEEE 802.11 WLAN standard significantly limits capacity on the network. At 2Mbit/s, a similar computation leads to 8 calls. When sending x byte voice samples, the overhead incurred is given by:

- RTP/UDP/IP 12+8+20=40 bytes
- MAC header + ACK = 38 bytes
- MAC/PHY procedure overhead = $754\mu s$
  - DIFS($50\mu s$), SIFS($10\mu s$)
  - preamble + PLCP ($192\mu s$) for data and ACK
  - contention (approx $310\mu s$)

The throughput in Mbps is given by the relation

$$T(x) = \frac{8x}{754 + (78 + x)\frac{8}{B}}$$

where x is the payload size in bytes, and B is the raw bandwidth of the channel (1,2,5.5, or 11). When using 20 byte voice payload in a 2 Mbps network, the capacity of the network is only 10% of the maximum possible.

4.3 VoIP over IEEE 802.11-Based Multihop Networks

Providing VoIP users with true mobile phone services having the freedom of roaming requires wide area wireless coverage, and IEEE 802.11-based multi-hop wireless mesh networks have been considered a practical solution for the enterprise. The benefits of mesh network compared to wired LAN connecting Wi-Fi access points are: i) ease of deployment and expansion; ii) better coverage; iii) resilience to node failure; iv) reduced cost of maintenance. Such a

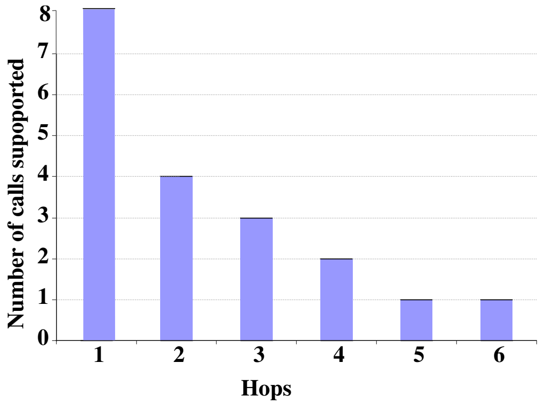


Fig. 9. In a linear topology, capacity degrades with the number of hops [15]

mesh network has the potential of creating an enterprise-scale or community-scale wireless backbone supporting multiple users while driving these users from using fixed phones to wireless VoIP phones shown in Figure 7. Each mesh nodes have more than two interfaces, one for the VoIP clients operating as a Wi-Fi access point, others to connect the mesh nodes behaving as an intermediate nodes. As a backbone network, wireless mesh network transfers the voice packets from APs to the IP-PBX which inter-connects with PSTN or operates with the Internet directly through SIP/RTP protocol suites.

However, supporting delay sensitive realtime applications such as VoIP over wireless mesh networks is challenging. The problems to support voice service over Wi-Fi which are pointed out at the previous section become even more severe when supporting VoIP over multihop mesh networks. In a multihop wireless network operating on a single channel, the UDP throughput decreases with number of hops for properly spaced nodes and is shown to be between 1/4 and 1/7 that of single hop capacity. This phenomenon of self interference is produced by different packets of the same flow competing for medium access at different nodes. When all nodes are within interference range, the UDP throughput in a linear topology can degrade to $\frac{1}{n}$, where n is the number of hops.

As shown in Figure 9, our experiment on a real mesh testbed with G.729 encoded VoIP calls indicates that the number of supported medium quality calls decreases with the number of hops for a simple linear topology. In a mesh network with 2Mbps link speed, the number of supported calls reduces from 8 calls in single hop to one call after 5 hops. This significant reduction in the number of supported calls can be attributed to following factors: a) decrease in the UDP throughput because of self interference; b) packet loss over multiple hops and c) high protocol overhead for small VoIP packets.

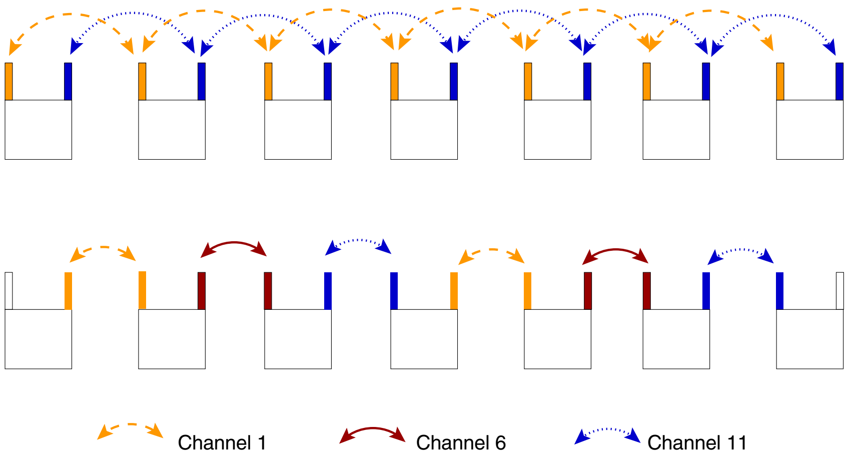


Fig. 10. Each node with two 802.11b interface. case A: two non overlapping channels for forward and reverse direction; case B: three channels used with reduced self interference [15].

Voice services over wireless mesh network faces a number of technical problems: a) providing QoS sensitive VoIP traffic b) packet loss due to channel interference and c) high overhead of the protocol stack - 802.11/IP/UDP/RTP for each VoIP packet with small payload. The above problems become even more severe when supporting VoIP over multihop mesh networks. In a multihop wireless network, the UDP throughput decreases with number of hops for properly spaced nodes and is shown to be between 1/4 and 1/7 that of single hop capacity. However, when nodes are interfering, the UDP throughput along a multihop string can degrade up to $\frac{1}{n}$, where n is the number of hops. Our experiment on a real mesh testbed with G.729a encoded VoIP calls indicates: a) the number of supported medium quality calls decreases with the number of hops for a simple linear topology; b) in a mesh network with 2Mbps link speed, the number of supported calls reduces from 8 calls in single hop to one call after 4 hops. This significant reduction in the number of supported calls can be attributed to following factors: a) decrease in the UDP throughput because of self interference; b) packet loss over multiple hops and c) high protocol overhead for small VoIP packets. But, VoIP capacity over multi-hop wireless networks can be improved using the following techniques.

Multiple Radio Interfaces

Referring back to Figure 9, we see that the main problem in a multihop network is performance degradation with increasing number of hops. A simple idea for improvement would be to just increase the number of interfaces in

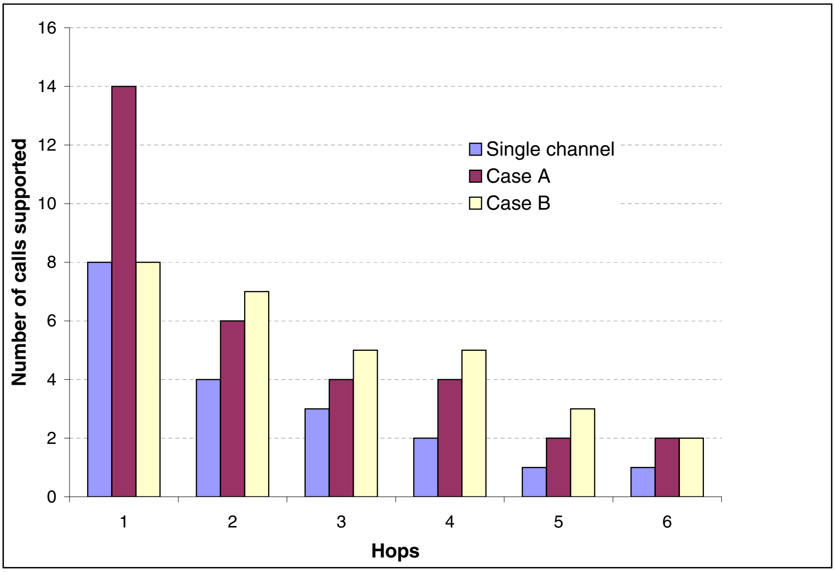


Fig. 11. Channel diversity: use of multiple cards with independent channels

each node. A naive use of multiple interfaces in a string would be to use one interface on a channel for the forward traffic and a second interface on a second channel for the reverse traffic, which should provide double capacity [15].

We verified this in our testbed on a string of six hops. However, for each of these flows, the same behavior as in Figure 9 is created by interference with neighbors which have cards on the same channel. An alternate method is to use more independent channels as shown in Figure 10. However, using 802.11b, only three channels are available to avoid interference, which limits the achievable improvement. Operating with only two backhaul interfaces and only three independent channels offered by 802.11b, we evaluated the following situations. Case A: two independent channels for forward and reverse traffic: (1,6)-(1,6)-(1,6)-(1,6)-(1,6)-(1,6)-(1,6). Case B: reduced self interference channel allocation: (1)-(1,6)-(6,11)-(11,1)-(1,6)-(6,11)-(11). The two solutions produce notable improvements, especially for longer paths (Figure 11). The lack of improvement for shorter paths is explained by a shortcoming of our testbed node, which only supports a limited number of interrupts per second. Using a better architecture, roughly a doubling of performance is expected with the addition of a second card, at least for the solution A. Solution B has even greater potential of improvement when more independent channels are available. If more channels were available, like in 802.11a, interference may be completely eliminated in a string, because a channel can be reused after 11 hops, which in most cases will be out of the interference range.

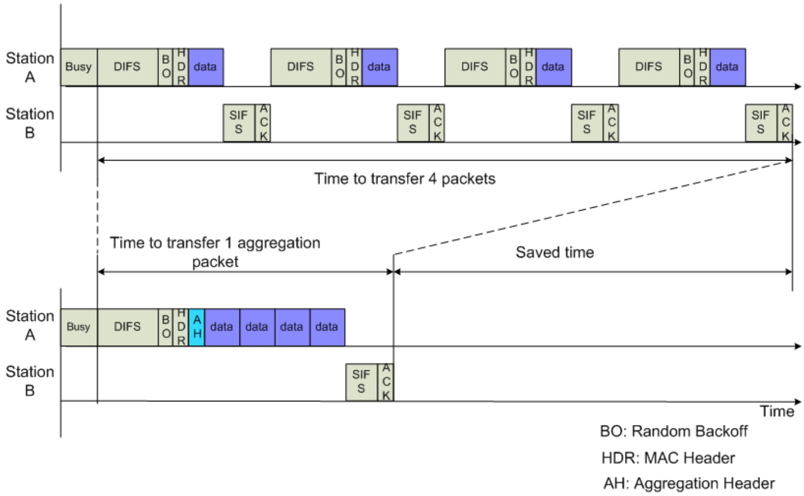


Fig. 12. Aggregating multiple packets [22]

Packet Aggregation

One way to improve throughput of the network, for example a network for VoIP applications, with small packets is to use packet aggregation as shown in Figure 12.

Aggregation of small packets has been researched and several algorithms, such as end-to-end packet aggregation, hop-by-hop packet aggregation, and node-to-end packet aggregation were proposed for general Internet and 802.11 networks as well [22].

These algorithms can provide good network utilization with the creation of larger aggregation packets while increasing the number of VoIP calls a lot, however it also requires the computational complexity and hardware resources, such as CPU power, battery usage etc, because every packet is aggregated at one node and deaggregated at the next intermediate node until the packets arrive at the destination. Also, this technique must define how long aggregation should be delayed at every node, which is hard to obtain in wireless mesh network. Also header compression with aggregation over the wireless mesh network can support much higher throughput [23].

5 Mobile Unified Communications in Enterprise

5.1 What Is Unified Communications

Inside and outside the office, a communications system must connect employees, mobile workers, remote workers, departmental workgroups and branch locations to make the enterprise a seamless one. Thereafter, business communications must unify calls, e-mails, chats, messaging, applications and even management. Thus, a new complete enterprise communications solution is required to allow communication to be managed more intelligently: *heterogeneous communication tools into all-in-one tool.*

With these requirements, Unified Communications (UC) becomes a promising technology architecture where communication tools are integrated so that both businesses and individuals can manage all their communications in one entity instead of separately. The integration of voice, video, data communication, multiple devices, and the services on a shared IP-based infrastructure must offer organizations significant gains in business productivity by removing latency in communications between customers and service providers, between team members, and with partners and consultants.

5.2 Benefits of Unified Communications

Unified Communications intends to help businesses to streamline information delivery and ensure ease of use. Through business-optimized Unified Communications, human latency in business processes are minimized or eliminated, resulting in better, faster interaction and service-delivery for the customer, and cost savings for the business. UC also allows for easier, more direct collaboration between co-workers and suppliers and clients, even if they are not physically on the same site. This allows for possible reductions in business travel, especially with multi-party video communications, reducing an organization's budget and time for traveling. Given the sophistication of UC technology, its uses are myriad for businesses.

5.3 Mobile Unified Communications

As more people own multiple devices, ranging from laptop computers to mobile phones to mobile e-mail devices, they spend more time managing their communications across different phone numbers, voice mailboxes, and e-mail accounts, limiting their ability to accomplish work efficiently. A few years ago, the demand for mobility might have applied only to a few employees such as highly mobile workers who needed access to resources wherever they were. Today, the demand for mobile and wireless technologies in business is pervasive. In enterprise, as landline voice traffic continues its migration to wireless/mobile phones - at any given time, 35% of all workers are only available by mobile - enterprise communications infrastructure should be provisioned to provide the same services to the mobile devices.

Business sectors across the globe, from retail businesses to warehouses to field service technicians, have embraced mobile phones, smartphones, personal digital assistants (PDAs), wireless-equipped laptop computers, and other devices for their convenience, portability, and efficiency. While mobility applications are rapidly deployed into the business area, UC is still in its early-adoption stage. Mobility will play an important role in leveraging the value of all UC applications. Thus, when making a business case for UC, applications with ubiquitous devices that facilitate location-awareness with mobility, such as softphones, smartphones, fixed-wireless dual-mobility devices and mobility clients, definitely contribute the most significant benefits to achieve.

Now, Enterprise Mobile Unified Communication (EMUC) is more than Fixed and Mobile Convergence, which only focuses on enabling voice call continuity between wired and wireless devices. EMUC anchors and integrates services within the enterprise and extends the power of enterprise-based unified communication into the mobile world with the additional functionality such as location/presence enriched services.

With the extension of location/presence enriched services, it enables users to know where their colleagues are physically located (say, their car or home office). They also have the ability to see which mode of communication the recipient prefers to use at any given time (perhaps their cell phone, or email, or instant messaging). A user could seamlessly set up a real-time collaboration on a document they are producing with co-workers, or in a retail setting, a worker might do a price-check on a product using a hand-held device and need to consult with a co-worker based on a customer inquiry.

5.4 IP-PBX Extension to Mobile Unified Communications

As the heart and brains of an enterprise communications network, the IP-PBX (Internet Protocol - Private Branch Exchange) can be the vital link that interfaces businesses and their customers. An IP-PBX system is a business telephone system designed to deliver voice, video and data over a communication network, such as LAN, and inter-operate with the normal PSTN network, while providing many additional functionalities, such as the ability to conference, call transfer, Interactive Voice Response (IVR), etc. In case of PSTN network and Internet together, if you want to initiate a call from a circuit-switched network, i.e. PSTN, to a peer in Internet or vice versa, a media gateway should come in the middle of the call path. The media gateway operates to connect different types of networks, one of its main furcation is to convert between different transmission and coding techniques. In enterprise, IP-PBX performs the conversion between PSTN voice to a media streaming protocol, such as RTP, as well as a signalling protocol used in the VoIP system. While mainly controlling VoIP systems in enterprise, IP-PBX performs advanced functions including IVR, audio/video conferencing, click-to-call, and call logging/tracking, presence services etc. Also it includes mobility support while merging the fixed communication system and mobile world into all-in-one system.

Asterisk [24] is a complete open source software PBX system created by Digium. It runs on Linux operating system and transfers voice and data among the peers, with or without requiring additional hardware for VoIP. In case of allowing a voice call with PSTN, it should be equipped with asterisk-compatible hardware which bridges conventional telephone networks to VoIP telephone networks. It supports PBX switching, codec translation, voicemail, conferencing bridging, IVR, and video telephony. Asterisk can be functioning in enterprise as:

- **As an IP-PBX:** while switching calls, managing call routes, enabling additional features, and connecting callers over analog or digital connections. Its intelligence can be extended to understand an enterprise wireless networks with location over the building inside or outside and call capacity over the accessed wireless network on the user.
- **As a Gateway:** while bridging two parties having different codecs and the communication systems, such as between PSTN and Internet callers or Internet calls having two different codecs. The wireless intrinsic capacity limitation lets the gateway to decide the optimum codec to choose at the point over the network on the call connection or even while talking on the phone: *Wireless-aware IPPBX*.

5.5 IP-PBX as a B2BUA

A SIP proxy server located at the enterprise is an intermediary entity that mainly plays the role of routing. A SIP Proxy server may not be allowed to alter SIP message and changes message headers or body even in case of being required by enterprise management system. Additionally, a SIP proxy server may not initiate or disconnect a call between both SIP user agents in mid-call. Thus, this basic SIP proxy server may not satisfy the requirement of the enterprise which wants to control and manage all conversation. A SIP Back-To-Back User Agent (B2BUA) behaves both a standard SIP entity running as a SIP server and a SIP user agent simultaneously, however manipulates the communication path and call parameters, which is operating under enterprise's control. The B2BUA enables enterprise to manage and track a call from beginning to end, integrate and offer new additional features. It resides between both end points of a call and separates the communication session into two call legs (one for one end user and B2BUA and the other for B2BUA and the other user end point) and bridges all SIP signalling or audio/video media streams between both ends of the call, while hiding the call leg in the enterprise side from the outside caller. A B2BUA provides the following features:

- Call control and management: call tracking and logging, admission control etc.
- Protocol/capacity conversion: codec conversion, protocol conversion, wireless bandwidth management etc.
- Protecting the internal network: hiding the details of the enterprise network architecture, etc.

Since B2BUA gives the enterprise with a flexible tool to control and manage VoIP system with IP-PBX system, SIP B2BUA server is a natural choice for the enterprise communication platform.

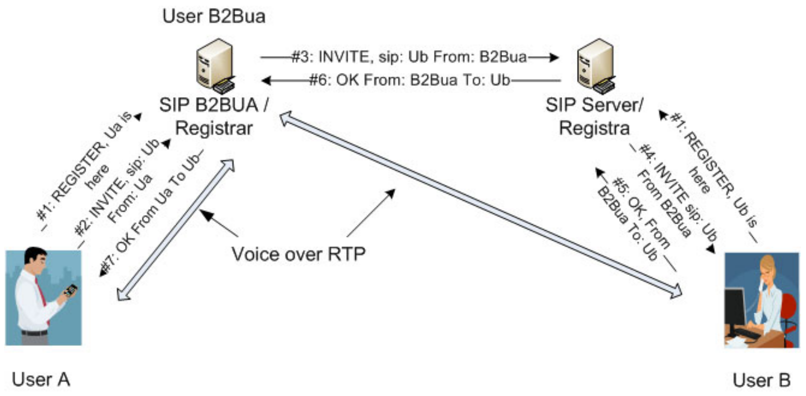


Fig. 13. Call Setup using SIP

5.6 Call Setup Using Enterprise IP-PBX

Let's assume User A and User B are on different networks or companies and that the two networks are connected by a router. Each network has a SIP server on the local LAN in the scenario shown in Figure 13. Let's assume the SIP server managing User A as a SIP B2BUA, which divides the communication session into two call legs. In the originating call leg the B2BUA acts as a user agent server (UAS) for User A and processes the request, generates the new requests (INVITE, BYE etc.) as a user agent client (UAC) to the destination end User B, handling the signaling back-to-back between end points.

- #1: When U_a and U_b turn on their terminals, the user agent software on each terminals registers them with their local SIP registrar server (REGISTER), such as their IP addresses.
- #2: U_a initiates the call, and the user agent on his terminal transmits an invitation message (INVITE) to B2BUA server, which acts as a user agent server. This invitation contains the session description information with SDP body, such as media capability (audio/video), the codec information etc.
- #3: B2BUA initiates the call (INVITE) on the behalf of U_a behaving a user agent client, B2Bua, while forwarding the invitation to every SIP server it knows how to reach to U_b.
- #4, #5, #6, #7: SIP server which U_b registered her location forwards the invitation message to U_b. U_b answers the call (200 OK), which returns acknowledgement with her media capability in SDP over the same path the invitation traveled.
- Now they have exchanged the media capability with SDP message, so they have the IP address and port information to contact the other party

and now can transmit RTP media directly each other. Contrary to normal SIP server's operation, B2BUA is in the middle of the session, providing media conversion and translation if needed.

5.7 Wireless-Aware IP-PBX

By integrating the management of desktop and Wi-Fi phones at the enterprise and cellular communication that is currently managed by a cellular provider into the enterprise IP-PBX, business communications can be managed more efficiently. Employees and workgroup users are more efficient when they are able to manage incoming and outgoing calls, chats and e-mails using one interface. They also become more collaborative and responsive when the same interface gives them real-time presence management controls, corporate and workgroup directories and conferencing. Whether a mobile employee is using a laptop or smartphone at the airport or a home PC, Enterprise UC enables the employee to connect to the corporate system to handle calls, chats and e-mails, and access the customer and business data he/she needs while providing the same interface: *one interface for all applications*.

One-number Follow-me and Find-me keeps mobile users connected to customers and colleagues, while remote speech-enabled messaging lets them easily access and conveniently manage e-mails, voice mails, and corporate directories and status setting. The service upgrade for the unified applications can easily be controlled by a single entity, the enterprise IP-PBX. With this architecture, a handoff of a call from the cellular network to the private network within an enterprise or vice versa can be controlled by enterprise mobile IP-PBX.

This vision comes true with Mobile UC-enabled IP-PBX. The mobile UC enhanced IP-PBX system architecture to support and manage business communications allowing users to make and receive calls using both the enterprise business cellular number and the desktop phone is shown in Figure 14 and 15. This architecture allows the IT department within an enterprise to control all calls from/to the employees with the single/dual mode mobile handsets in the enterprise including handover between the Wi-Fi and the 3G network. In Figure 14, the dual-mode handset is on an active call established with a peer using WLAN. When it roams to the edge of WLAN coverage, WLAN infrastructure instructs enterprise IP-PBX and mobility controller to initiate handover. IP-PBX initiates a new call leg to the cellular network and bridges the new leg to include the active call and releases the WLAN connection. A new call path is thus established and the end user manually or automatically answers the call.

Figure 15 shows roaming from 3G to WLAN while a call is active over 3G. At first, the dual-mode mobile device requests a call establishment to the enterprise IP-PBX by transmitting a phone number of the peer. The Enterprise IP-PBX establishes the new call leg to the PSTN network and bridges the new leg to the dual-mode device waiting for the call to be established. When the user moves into the Wi-Fi coverage of the enterprise, the WLAN

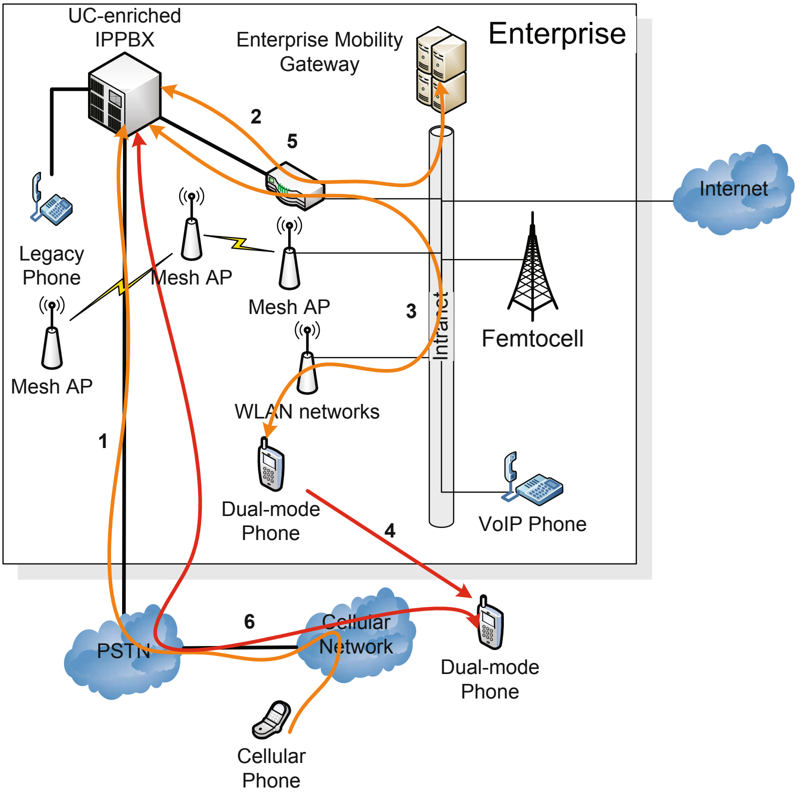


Fig. 14. Roaming from 802.11 to 3G networks

infrastructure monitors and instructs IP-PBX and mobility controller to initiate handover. IP-PBX initiates a new call leg to the Wi-Fi network and bridges the new leg to activate the call over WLAN and release the cellular connection.

5.8 Inside Wi-Fi (FMC) vs. Inside Cellular (FMS)

In FMCN, you can walk into your office talking on a cellular phone, sit down, and continue the conversation on the same call, while using enterprise wired/wireless networks inside instead of outside cellular service. With the development and deployment of new wireless technologies, several alternatives can be supported to enable seamless call continuity.

Fixed-Mobile Substitution

If the wireless connection remains cellular - *Fixed-Mobile Substitution*: this is a big saving for the user (no need fixed line), however a big losing for the

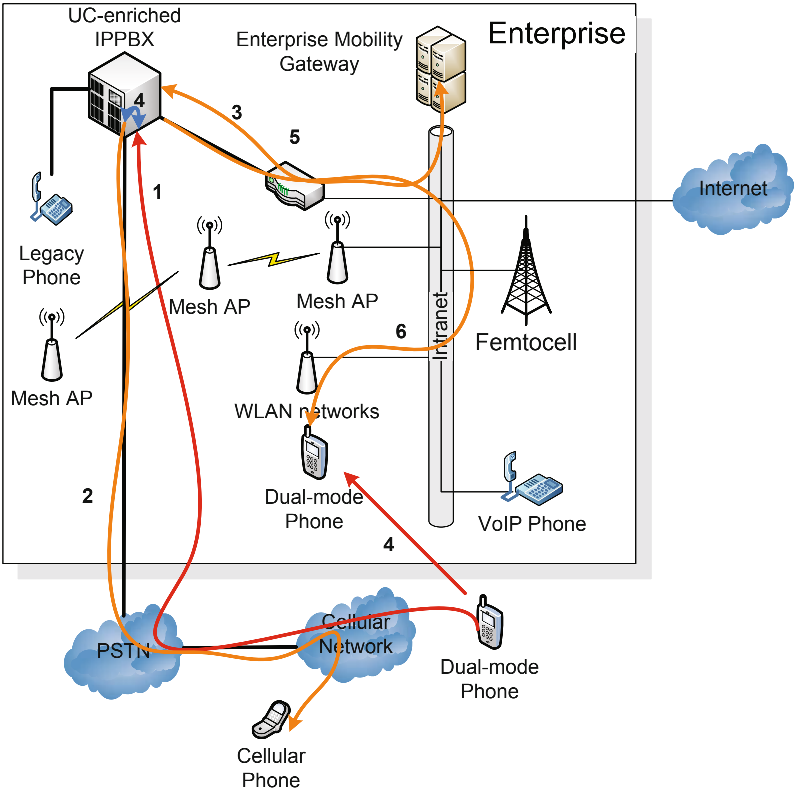


Fig. 15. Roaming from 3G to 802.11 networks

fixed operator. A femtocell allows service providers to extend service coverage indoors, especially where access would be limited or unavailable.

Fixed-Mobile Convergence

If it seamlessly transitions to Wi-Fi or Bluetooth, and stops using cellular minutes and transferring the call over the wireline - *Fixed-Mobile Convergence*: this is a cost-saving for the user (no need for cellular service) while utilizing the existing fixed broadband service, however can be a primary factor of the subscriber loss for the mobile operator.

Enterprise Viewpoint

Enterprises have been waiting the arrival of dual-mode (Wi-Fi and cellular) or single mode (femtocell-aware) handsets, *convergence-aware handset* that can potentially replace the traditional PBX phone. These convergence-aware handsets can make voice calls over Wi-Fi or femtocell, while retaining the functions and benefits of a traditional cellular phone. With these convergence-aware handsets, workers only need one device with a single number.

What enterprise wants from FMCN is to have the ability for an IP-PBX to treat a cellular phone as an extension, and the ability for a cellular phone to behave like a IP-PBX extension phone. Extension to cellular phone means that the system seamlessly bridges office phone services to mobile devices, enabling the user of one phone number and one voice mailbox. The most important things to enterprise is whether the convergence services can be integrated into the enterprise calling system while acquiring the control and management of the communications. To enable this service, there are some requirements to provide:

- **Session Continuity:** is moving a call in progress from outside call to inside call or vice versa, as much the same way as you might transfer a call from one extension to another. Dual-model handset or a single-mode cellular phone can completely hide the session handover. Depending on the deployed architecture, you might control the process with the help of IP-PBX or give all control to cellular operator. To provide additional services in enterprise FMCN, IP-PBX should can play an important role more than moving the calls from outside to inside or the other way around.
- **Mobile IP-PBX:** treats the cellular phone as the PSTN extension and allows the employee to invoke IP-PBX features.
- **Mobility Controller:** session continuity requires a component in the network that routes and reroutes the call over either the enterprise network or cellular network as needed while keeping track of the call. This component can collaborate with IP-PBX to support seamless session continuity.

6 Conclusion and Future Directions

This chapter discussed VoIP in enterprise fixed mobile converged networks. First, VoIP basics including VoIP related transport and signal protocols, speech coding schemes, and VoIP quality assessment model were introduced. Then enterprise fixed mobile converged networks were explained with focuses on the system architecture and service characteristics. Challenges and performance of VoIP over enterprise fixed mobile converged networks were reviewed. Mobile Unified Communications with voice supporting over enterprise fixed mobile converged networks were also detailed by utilizing the principles of enterprise IP-PBX.

Several future directions are listed below.

- **VoIP over Femtocell:** Femtocells are small cellular base stations intended to extend service coverage and offload the mobile macro network to home, small office, and enterprise environments. When someone is in the home or enterprise covered by femtocell, people would be able to make phone calls directly through their own femtocell instead of the overlaying marcocell. It results in a significantly improved signal quality - *close base station* and substantial cost savings - *without using macro cellular network.* Femtocells are self-installing, self-optimizing, self-healing, and plug-and-play devices deployed by users similar to Wi-Fi access points and use IP broadband connections for backhaul to cellular networks. One of the main challenges of the femtocell is to mitigate the interference between marcocell and femtocell, or even among femtocells, allocating intelligently the spectrum to the femtocell under the control of the service provider. In a conventional network the radio resource is centrally managed; however this will not be the case for the femtocells which are much more autonomous. Hence, on the deployment, the femtocell should be equipped with the certain requirements. For example, femtocell will need to have not only zero-touch configuration, remote management, software upgrades and remote debugging, but also the intelligent monitoring capabilities for location determination, automatic topology discovery, and neighborhood watch. Another challenging issue consists of sharing the radio resource between femtocells. This problem is likely to happen in an environment where femtocells are densely deployed such as in big residential building containing large number of residential or enterprise building consisting of large number of femtocell base station to cover the whole area.

Currently femtocell is based mainly on 2G or 3G technology, however in near future, 4G-based femtocell offering higher throughput is expected to deploy in enterprise. Even though 4G networks such as LTE are expected to provide much higher speed than 2G and 3G and pave the way for data applications, voice services will still be the major application in the sense of contributing primary revenue in an perspective of telecom operators. However since 4G networks are all-IP designed, there is no circuit-switching any more. The question for the operators is how to guarantee circuit-like carrier-level VoIP quality over LTE networks. Some approaches like “Circuit Switched (CS) Fallback”, “Voice over LTE via Generic Access (VoLGA)” and “IP Multimedia Subsystem” are possible solutions; however, they either need to rely on existing circuit voice networks or require long-term evolution.

- **Location-Aware Enterprise VoIP:** Enterprise mobility becomes an emerging but essential reality. An essential prerequisite to mobile Unified Communications is the ability to gather information about the current location or position of users and their mobile devices. As the efforts of enabling mobility in enterprise, considerable interest has recently emerged in indoor location based services (LBS). LBS requires reliable information about user position and her environment; such information is available from an indoor location system. Indoor positioning requires the deployment of specialized equipment or product that integrates with enterprise communication networks. A desirable indoor location system should be characterized by high accuracy, short training phase, cost-effectiveness,

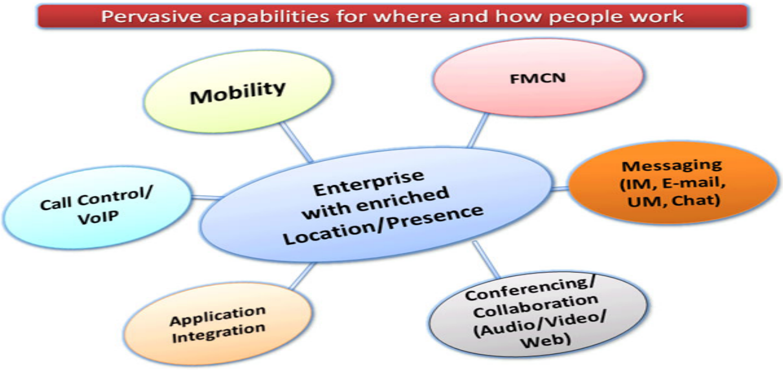

Pervasive capabilities for where and how people work
Mobility
FMCN
Call Control/ VoIP
Enterprise with enriched Location/Presence
Messaging (IM, E-mail, UM, Chat)
Application Integration
Conferencing/ Collaboration (Audio/Video/ Web)

Fig. 16. Enterprise Mobile Unified Communications are the integration of numerous applications including collaboration, conferencing, unified messaging, contact center, "mobility" and location-aware presence. Enterprise UC with wireless mobility has the pervasive capabilities for where and how people work.

and robustness in the face of unexpected situation. Combining information about the availability and the preferable means of quick communication using text, audio and video at one's location provides an important advantage for users in addition to traditional communication media such as the phone or e-mail. There are two different approaches on pursuing indoor positioning and localization research exploiting 1) a single RF technology using the Wi-Fi enabled phone, 2) the heterogeneity of ubiquitous computing infrastructures, sensor assisted RF-based indoor positioning system. The system can be designed to utilize mobile phones carried by people in their everyday lives as mobile sensors to track mobile events. The sensor-enabled mobile phones are best suited to deliver sensing services, such as tracking in enterprise building, than more traditional solutions like tag based positioning systems, which are limited in scale, deployment, and cost. Now, far more than a single technology or platform, mobile UC equipped with location and presence features is rather an architecture approach to seamless collaboration and communication across all media and places - desktop, business and office applications, fixed and mobile voice, video as well as messaging, desk-sharing and conferencing of all kinds shown in Figure 16. Two important components in the figure which enable and empower mobile UC is mobility extension for business support and location-aware enriched presence. This architecture makes it possible to select the most cost-effective service according to the location-aware presence information (anywhere, anytime and anymedia).

- **Admission Control:** The traditional telephones control the user access via call admission control. However, most of the current IP telephony system has no admission control and can only offer best effort service with the assumption of sufficient bandwidth in wired network. However it can allow a new traffic to keep entering the network even beyond the network capacity limitation; consequently causing both the existing and the new flows to degrade their call quality due to packet delay and loss. Normally because the enterprise network is deployed with the large capacity of its network, the call quality degradation due to no admission control happens rarely. However, VoIP over enterprise wired/wireless converged network, which has limited bandwidth, should be provided to prevent perceived call quality degradation, thus admission control mechanism should be at place.

References

1. Muller, N.J. (ed.): Wi-Fi for the Enterprise: Maximizing 802.11 For Business. McGraw-Hill Professional, New York (2008)
2. Miles, S.B., Sarma, S.E., Williams, J.R.: RFID Technology and Applications. Cambridge University Press, Cambridge (2008)
3. Wang, F., Ghosh, A., Sankaran, C., Fleming, P.J., Hsieh, F., Benes, S.J.: Mobile wimax systems: performance and evolution. IEEE Communications Magazine 46(10), 41–49 (2008)
4. Casati, A., Shneyderman, A., Faynberg, I.: Fixed Mobile Convergence: Voice Over Wi-Fi, IMS, UMA/GAN, Femtocells, and Other Enablers. O'Reilly Media, Inc., Sebastopol (2009)
5. Voice Call Continuity between the Circuit-Switched Domain and IP Multimedia Core Network Subsystems. stage 3, version 7.4.0. 3GPP specification (May 2008)
6. Qadeer, M.A., Khan, A.H., Ansari, J.A., Waheed, S.: IMS Network Architecture. In: International Conference on Future Computer and Communication, ICFCC 2009, pp. 329–333 (April 2009)
7. Generic Access Network (GAN). stage 2, version 8.1.0. 3GPP specification (May 2008)
8. Watson, R.: Fixed/Mobile Convergence and Beyond: Unbounded Mobile Communications. Newnes (November 2008)
9. Burgy, L., Consel, C., Latry, F., Réveillère, L., Palix, N.: Telephony over IP: Experience and Challenges. ERCIM News 63, 53–54 (2005)
10. Schulzrinne, H., Casner, S., Frederick, R., Van Jacobson.: RTP: A Transport Protocol for Real-Time Applications (1996)
11. Glasmann, J., Kellerer, W., Muller, H.: Service development and deployment in H.323 and SIP. In: Proceedings of the Sixth IEEE Symposium on Computers and Communications, pp. 378–385 (2001)
12. Handley, M., Jacobson, V.: SDP: Session Description Protocol (1998)
13. Cisco technical note. Voice Over IP - Per Call Bandwidth Consumption (2006)
14. International Telecommunication Union. ITU-T Recommendation G.107: The E-model, a computation model for use in transmission planning. Technical report (August 2008)
15. Ganguly, S., Navda, V., Kim, K., Kashyap, A., Niculescu, D., Izmailov, R., Hong, S., Das, S.R.: Performance Optimizations for Deploying VoIP Services in Mesh Networks. IEEE Journal on Selected Areas in Communications 24(11), 2147–2158 (2006)

16. Cole, R.G., Rosenbluth, J.H.: Voice over IP performance monitoring. SIGCOMM Comput. Commun. Rev. 31(2), 9–24 (2001)
17. Larmo, A., Lindstrom, M., Meyer, M., Pelletier, G., Torsner, J., Wiemann, H.: The LTE link-layer design - LTE part II: 3GPP release 8. Communications Magazine 47(4), 52–59 (2009)
18. Chandrasekhar, V., Andrews, J., Gatherer, A.: Femtocell networks: a survey. Communications Magazine 46(9), 59–67 (2008)
19. Akyildiz, I.F., Xie, J., Mohanty, S.: A Survey of Mobility Management in Next-Generation All-IP-based Wireless Systems. IEEE Wireless Communications, see also IEEE Personal Communications 11(4), 16–28 (2004)
20. Saha, D., Mukherjee, A., Misra, I.S., Chakraborty, M., Subhash, N.: Mobility support in IP: a survey of related protocols. IEEE Network 18, 34–40 (2004)
21. Anjum, F., Elaoud, M., Famolari, D., Ghosh, A., Vaidyanathan, R., Dutta, A., Agrawal, P., Kodama, T., Katsube, Y.: Voice performance in WLAN networks - an experimental study. In: Global Telecommunications Conference GLOBECOM 2003, vol. 6, pp. 3504–3508. IEEE, Los Alamitos (2003)
22. Kim, K., Ganguly, S., Izmailov, R., Hong, S.: On Packet Aggregation Mechanisms for Improving VoIP Quality in Mesh Networks. In: IEEE 63rd Vehicular Technology Conference, VTC 2006-Spring, vol. 2, pp. 891–895 (May 2006)
23. Kim, K., Hong, S.: VoMESH: voice over wireless MESH networks. In: Wireless Communications and Networking Conference, WCNC 2006, vol. 1, pp. 193–198. IEEE, Los Alamitos (2006)
24. Madsen, L.: Asterisk: The Future of Telephony. McGraw-Hill/Osborne Media (January 2008)

Speech Quality Assessment

Philipos C. Loizou

University of Texas-Dallas,
Department of Electrical Engineering, Richardson, TX, USA

Abstract. This chapter provides an overview of the various methods and techniques used for assessment of speech quality. A summary is given of some of the most commonly used listening tests designed to obtain reliable ratings of the quality of processed speech from human listeners. Considerations for conducting successful subjective listening tests are given along with cautions that need to be exercised. While the listening tests are considered the gold standard in terms of assessment of speech quality, they can be costly and time consuming. For that reason, much research effort has been placed on devising objective measures that correlate highly with subjective rating scores. An overview of some of the most commonly used objective measures is provided along with a discussion on how well they correlate with subjective listening tests.

The rapid increase in usage of speech processing algorithms in multi-media and telecommunications applications raises the need for speech quality evaluation. Accurate and reliable assessment of speech quality is thus becoming vital for the satisfaction of the end-user or customer of the deployed speech processing systems (e.g., cell phone, speech synthesis system, etc.).

Assessment of speech quality can be done using subjective listening tests or using objective quality measures. Subjective evaluation involves comparisons of original and processed speech signals by a group of listeners who are asked to rate the quality of speech along a pre-determined scale. Objective evaluation involves a mathematical comparison of the original and processed speech signals. Objective measures quantify quality by measuring the numerical "distance" between the original and processed signals. Clearly, for the objective measure to be valid, it needs to correlate well with subjective listening tests, and for that reason, much research has been focused on developing objective measures that modeled various aspects of the auditory system. This Chapter provides an overview of the various subjective and objective measures proposed in the literature [1] [2, Ch. 10] for assessing the quality of processed speech.

Quality is only one of many attributes of the speech signal. Intelligibility is a different attribute and the two are not equivalent. For that reason, different assessment methods are used to evaluate quality and intelligibility of processed speech. Quality is highly subjective in nature and it is difficult to evaluate reliably. This is partly because individual listeners have different internal standards of what constitutes "good" or "poor" quality, resulting in large variability in rating scores

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 623–654.
springerlink.com

among listeners. Quality measures assess "how" a speaker produces an utterance, and includes attributes such as "natural", "raspy", "hoarse", "scratchy", and so on. Quality is known to possess many dimensions, encompassing many attributes of the processed signal such as "naturalness", "clarity", "pleasantness", "brightness", etc. For practical purposes we typically restrict ourselves to only a few dimensions of speech quality depending on the application. Intelligibility measures assess "what" the speaker said, i.e., the meaning or the content of the spoken words. In brief, speech quality and speech intelligibility are not synonymous terms, hence different methods need to be used to assess the quality and intelligibility of processed speech.

The present Chapter focuses on assessment of speech quality, as affected by distortions introduced by speech codecs, background noise, noise-suppression algorithms and packet loss in telecommunication systems.

1 Factors Influencing Speech Quality

There is a host of factors that can influence speech quality. These factors depend largely on the application at hand and can affect to some degree listening and talking difficulty. In telecommunication applications, for instance, degradation factors that can cause a decrease in speech quality and subsequently increase listening difficulty include distortions due to speech codecs, packet loss, speech clipping and listener echo [3]. The distortions alone introduced by speech codecs vary widely depending on the coding rate [1, Ch. 4]. The distortions introduced, for instance, by waveform coders (e.g., ADPCM) operating at high bit rates (e.g., 16 kbps) differ from those introduced by linear-predictive based coders (e.g., CELP) operating at relatively lower bit rates (4-8 kbps).

The distortions introduced by hearing aids include peak and center clipping, Automatic Gain Control (AGC), and output limiting. The AGC circuit itself introduces non-linear distortions dictated primarily by the values of attack and release time constants. Finally, the distortions introduced by the majority of speech-enhancement algorithms depend on the background noise and the suppression function used (note that some enhancement algorithms can not be expressed in terms of a suppression function). The choice of the suppression function can affect both the background noise and speech signal itself, leading to background and speech distortions. The suppression function of spectral-subtractive type of algorithms, for instance, is known to introduce "musical noise" distortion [4].

In summary, there are many factors influencing speech quality and the source of those factors depends on the application. Hence, caution needs to be exercised when choosing subjective or objective measures to evaluate speech quality.

2 Subjective Listening Tests

Several methods for evaluating speech quality have been proposed in the literature [1]. These methods can be broadly classified into two categories: those that are based on relative preference tasks and those that are based on assigning a numerical value on the quality of the speech stimuli, i.e., based on quality ratings. In the relative preference tests, listeners are presented with a pair of speech stimuli consisting of the test stimuli and the reference stimuli. The reference stimuli are typically constructed by degrading the original speech signal in a systematic fashion, either by filtering or by adding noise. Listeners are asked to select the stimuli they prefer the most. In the rating tests, listeners are presented with the test speech stimuli and asked to rate the quality of the stimuli on a numerical scale, typically a 5-point scale with one indicating poor quality and a five indicating excellent quality. No reference stimuli are needed in the rating tests. As we will see next, these tests have their strengths and weaknesses, and in practice, the best test might depend on the application at hand. In the following sections, we describe in more detail the relative preference and quality rating tests which can be used to assess the quality of degraded speech.

2.1 Relative Preference Methods

Perhaps the simplest form of paired comparison test is the forced-choice paired comparison test. In this test, listeners are presented with pairs of signals produced by systems A and B, and asked to indicate which of the two signals they prefer. The same signal is processed by both systems A and B. Results are reported in terms of percent of time system A is preferred over system B. Such a method is typically used when interested in evaluating the preference of system A over other systems. The main drawback of this simple method is that it is not easy to compare the performance of system A with the performance of other systems obtained in other labs.

While the above AB preference test tells us whether system A is preferred over system B, it does not tell us by how much. That is, the magnitude of the difference in preference is not quantified. The comparison category rating (CCR) test is designed to quantify the magnitude of the preference difference on a 4-point scale with the rating of 0 indicating no difference, 1 indicating small difference, 2 indicating a large difference and 3 indicating a very large difference. Table 1 shows the category ratings [5,6]. This scale is also referred to as the comparison mean opinion score (CMOS). Positive and negative numbers are used to account for both directions of preference.

2.2 Absolute Category Rating Methods

Preference tests typically answer the question: “How well does an average listener like a particular test signal over another signal or over a reference signal which can be easily reproduced?” Listeners must choose between two sequentially presented signals, but do not need to indicate the magnitude of their preference (except in the CCR test, Table 1) or the reason(s) for their decision. In some applications, however, knowing the reason why a particular signal is preferred over another is more important that the preference score itself. Another shortcoming of the preference methods is that the reference signals do not always allow for a wide range of distortions as they only capture a limited scope of speech distortions that could be encountered. This could potentially result in most of the test signals being preferred (or disliked) over the reference signals, thereby introducing a bias in the quality evaluation. Lastly, most preference tests produce a *relative* measure of quality (e.g., relative to a reference signal) rather than an absolute measure. As such, it is difficult to compare preference scores obtained in different labs without having access to the same reference signals. The above shortcomings of the preference tests can be addressed by the use of absolute judgment quality tests in which judgments of overall quality are solicited from the listeners without the need for reference comparisons. These tests are described next.

Table 1. Comparison category ratings used in the comparison mean opinion score (CMOS) test

Rating	**Quality of second stimulus compared to the first is:**
3	Much better
2	Better
1	Slightly better
0	About the same
-1	Slightly worse
-2	Worse
-3	Much worse

2.2.1 Mean Opinion Scores (MOS)

The most widely used direct method of subjective quality evaluation is the category judgment method in which listeners rate the quality of the test signal using a five-point numerical scale (see Table 2), with 5 indicating “excellent” quality and 1 indicating “unsatisfactory” or “bad” quality. This method is one of the methods recommended by the IEEE Subcommittee on Subjective Methods [7] as well as by ITU [6,8]. The measured quality of the test signal is obtained by averaging the scores obtained from all listeners. This average score is commonly referred to as the Mean Opinion Score (MOS).

The MOS test is administered in two phases: training and evaluation. In the training phase, listeners hear a set of reference signals that exemplify the high (excellent), the low (bad) and the middle judgment categories. This phase, also known as “anchoring phase”, is very important as it is needed to equalize the subjective range of quality ratings of all listeners. That is, the training phase should in principle equalize the “goodness” scales of all listeners to ensure, to the extent possible, that what is perceived “good” by one listener is perceived “good” by the other listeners. A standard set of reference signals need to be used and described when reporting the MOS scores [9]. In the evaluation phase, subjects listen to the test signal and rate the quality of the signal in terms of the five quality categories (1-5) shown in **Table** Table 2.

Table 2. MOS rating scale

Rating	Speech quality	Level of distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

Detailed guidelines and recommendations for administering the MOS test can be found in the ITU-R BS.562-3 standard [6] and include:

1. **Selection of listening crew**: Different number of listeners is recommended depending on whether the listeners had extensive experience in assessing sound quality. Minimum number of non-expert listeners should be 20 and minimum number of expert listeners should be 10. The listeners need to be native speakers of the language of the speech materials tested, and should not have any hearing impairments.
2. **Test procedure and duration**: Speech material (original and degraded) should be presented in random order to subjects, and the test session should not last more than 20 minutes without interruption. This step is necessary to reduce listening fatigue.
3. **Choice of reproduction device**: Headphones are recommended over loudspeakers, since headphone reproduction is independent of the geometric and acoustic properties of the test room. If loudspeakers are used, the dimensions and reverberation time of the room need to be reported.

Further guidelines pertaining the choice of speech input levels, noise and reference conditions, etc. for proper evaluation of the quality of narrow- and wide-band speech codecs can be found in the ITU standard [5] as well as in [10].

Reference signals can be used to better facilitate comparisons between MOS tests conducted at different times, different laboratories and different languages [11]. MOS scores can be obtained, for instance, using different Modulated Noise Reference Unit (MNRU) reference signals[1] for various values of Q (S/N) ranging from 5 to 35 [5,11]. A plot of MOS scores as a function of Q can be constructed to transform the raw MOS scores to an equivalent Q value. The Q equivalent values can then be used to compare performance among systems in different labs.

The MOS test is based on a five-category rating of the speech quality (Table 2). The quality scale is in a way quantized into five discrete steps, one for each category. Listeners are therefore forced to describe the complex impressions of speech

[1] The MNRU reference signals are generated by adding to the input signal random noise with amplitude proportional to the instantaneous signal amplitude as follows: $r(n) = x(n)\left[1 + 10^{-Q/20} d(n)\right]$ where *x(n)* is the input speech signal, *d(n)* is the random noise and Q is the desired SNR.

quality in terms of the five categories. It is implicitly assumed that these five steps (categories) are uniformly spaced, i.e., that they equidistant from each other. This assumption, however, might not be true, in general. For these reasons, some have suggested modifying the above test to ask the listeners to evaluate the test signals in terms of real numbers from 0 to 10, where zero indicates "bad" quality and 10 indicates "excellent" quality [12]. In this test, no quantization of the quality scale is done since the listeners are allowed to use fractions between integers, if they so desire.

A variant of the MOS test that addresses to some degree the low resolution issue stated above, is the degradation mean opinion score (DMOS) test [13]. In this test, the listeners are presented with both the unprocessed signal (which is used as a reference) and the processed signal. Listeners are asked to rate the perceived degradation of the processed signal relative to the unprocessed signal on a 5-point scale (Table 3). This test is suitable for situations in which the signal degradations or impairments are small.

Table 3. Degradation rating scales

Rating	Degradation
1	Very annoying
2	Annoying
3	Slightly annoying
4	Audible but not annoying
5	Inaudible

2.2.2 Diagnostic Acceptability Measure

The absolute category judgment method (e.g., MOS test) is based on ratings of the *overall* quality of the test speech signal. These ratings, however, do not convey any information about the listeners' bases for judgment of quality. Two different listeners, for instance, may base their ratings on different attributes of the signal, and still give identical overall quality rating. Similarly, a listener might give the same rating for two signals produced by two different algorithms, but base his judgments on different attributes of each signal. In brief, the MOS score alone does not tell us which attribute of the signal affected the rating. The MOS test is therefore considered to be a single-dimensional approach to quality evaluation, and as such it can not be used as a diagnostic tool to improve the quality of speech enhancement or speech coding algorithms.

A multi-dimensional approach to quality evaluation was proposed by Voiers [14] based on the Diagnostic Acceptability Measure (DAM). The DAM test evaluates the speech quality on three different scales classified as *parametric, metametric* and *isometric* [1,15]. These three scales yield a total of 16 measurements on speech quality covering several attributes of the signal and background. The metametric and isometric scales represent the conventional category judgment approach where speech is rated relative to "intelligibility", "pleasantness" and "acceptability". The parametric scale provides fine-grained measurements of the signal and background distortions. Listeners are asked to rate the signal distortion on six different dimensions and the background distortion on four dimensions. Listeners are asked for instance to rate on a scale of 0 to 100 how muffled or how nasal the signal sounds ignoring any other signal or background distortions present. Listeners are also asked to rate separately on a scale of 0 to 100 the amount of hissing, buzzing, chirping or rumbling present in the background. The composite acceptability measure summarizes all the information gathered from all the scales into a single number, and is computed as a weighted average of the individual scales.

The parametric portion of the DAM test relies on the listeners' ability to *detect*, perhaps more reliably, specific distortions present in the signal or in the background rather than providing preference judgments of these distortions. It therefore relies on the assumption that people tend to agree better on *what they hear* rather than on *how well they like it* [15]. To borrow an example from daily life, it is easier to get people to agree on the color of a car than how much they like it. As argued in [15], the parametric approach tends to give more accurate – more reliable – scores of speech quality as it avoids the individual listener's "taste" or preference for specific attributes of the signal from entering the subjective quality evaluation.

Compared to the MOS test, the DAM test is time consuming and requires carefully trained listeners. Prior to each listening session, listeners are asked to rate two "anchor" and four "probe" signals. The "anchors" consist of examples of high and low quality speech and give the listeners a frame of reference. The "probes" are used to detect any coincidental errors which may affect the results in a particular session. In addition to the presentation of "anchors" and "probes", listeners are selected on the basis that they give consistent ratings over time and have a moderately high correlation to the listening crew's historical average rating [1]. The selected listeners are calibrated prior to the testing session so as to determine their own subjective origin or reference relative to the historical average listener's ratings.

2.2.3 The ITU-T P.835 Standard for Evaluating Noise-Suppression Algorithms

The above subjective listening tests (DAM and MOS) were designed primarily for the evaluation of speech coders. The speech coders, however, are evaluated mainly in quiet and generally introduce different types of distortion than those encountered in noise suppression algorithms. Speech enhancement algorithms typically degrade the speech signal component while suppressing the background noise, particularly in low SNR conditions. That is, while the background noise may be suppressed, and in some cases rendered inaudible, the speech signal may get degraded in the process. This situation complicates the subjective evaluation of speech enhancement algorithms since it is not clear as to whether listeners base their overall quality judgments on the signal distortion component, noise distortion component or both. This uncertainty regarding the different weight individual listeners place on the signal and noise distortion components introduces additional error variance in the subjects' ratings of overall quality resulting and consequently decreases the reliability of the ratings. These concerns were addressed by the

ITU-T standard (P. 835) [16] that was designed to lead the listeners to integrate the effects of both signal and background distortion in making their ratings of overall quality.

The methodology proposed in [16] reduces the listener's uncertainty by requiring him/her to successively attend to and rate the waveform on: the *speech signal* alone, the *background noise* alone, and the *overall effect* of speech and noise on quality. More precisely, the ITU-T P.835 method instructs the listener to successively attend to and rate the enhanced speech signal on:

1. the speech signal alone using a five-point scale of signal distortion (SIG) – see Table 4.
2. the background noise alone using a five-point scale of background intrusiveness (BAK) – see Table 5,
3. the overall (OVL) effect using the scale of the Mean Opinion Score - [1=bad, 2=poor, 3=fair, 4=good, 5=excellent].

Table 4. Scale of signal distortion (SIG)

Rating	Description
5	Very natural, no degradation
4	Fairly natural, little degradation
3	Somewhat natural, somewhat degraded
2	Fairly unnatural, fairly degraded
1	Very unnatural, very degraded

Table 5. Scale of background intrusiveness (BAK)

Rating	Description
5	Not noticeable
4	Somewhat noticeable
3	Noticeable but not intrusive
2	Fairly conspicuous, somewhat intrusive
1	Very conspicuous, very intrusive

Each trial contains a three-sentence sample of speech laid out in the format shown in Figure 1. Each sample of speech is followed by a silent period during which the listener rates the signal according to the SIG, BAK or OVL scales. In the example shown in the figure, each sample of speech is approximately four seconds in duration and includes: one second of preceding background noise alone, two seconds of noisy speech (roughly the duration of a single sentence), and one second of background noise alone. Each sample of speech is followed by an appropriate silent interval for rating. For the first two samples, listeners rate either the signal **or** the background depending on the rating scale order specified for that trial. For the signal distortion rating, for instance, subjects are instructed to attend *only* to the speech signal and rate the speech on the five-category distortion scale shown in Table 4. For the background distortion rating, subjects are instructed to attend *only* to the background and rate the background on the five-category intrusiveness scale shown in Table 5. Finally, for the third sample in each trial, subjects are instructed to listen to the noisy speech signal and rate it on the five-category overall quality scale used in MOS tests (Table 2). To control for the effects of rating scale order, the order of the rating scales needs to be balanced. That is, the scale order should be "Signal, Background, Overall Effect" for half of the trials, and "Background, Signal, Overall Effect" for the other half. The ITU-T P.835 standard was used in [17] to evaluate and compare the performance of 13 different speech enhancement algorithms.

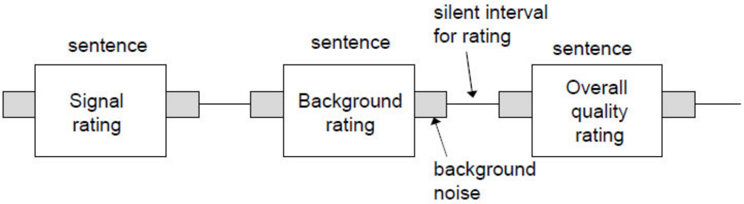


Fig. 1. Stimulus presentation format for the listening tests conducted according to the ITU-T P.835 standard

2.3 Considerations in Subjective Listening Tests

2.3.1 Evaluating the Reliability of Quality Judgments: Recommended Practice

In the above subjective tests, listeners rate the quality of the processed speech on a 5-point discrete scale (MOS test) or on a 0-100 continuous scale (DAM test). For the ratings to be meaningful, however, listeners must use the scales consistently. A given listener must rate a specific speech sample the same way every time he or she hears it. That is, we would like the *intra-rater reliability* of quality judgments to be high. Listeners need, in other words, to be self-consistent in their assessment of quality. Various statistics have been used to evaluate intra-rater reliability [18,19]. The two most common statistics are the *Pearson's correlation coefficient* between the first and second ratings, and the test-retest *percent agreement*.

Additionally, all listeners must rate a given speech sample in a similar way. We would thus like the *inter-rater reliability* of quality judgments to be high. A number of *inter-rater reliability* measures have been used [18] and include among others the Cronbach's alpha [20], Kendall's coefficient of Concordance [21] and the intraclass correlation coefficient [22,23].

The measurements of *intra-* and *inter-rater* reliability are critically important as they indirectly indicate the confidence we place on the listeners' (i.e., the raters) quality judgments. High values of *inter-rater* reliability, for instance, would suggest that another sample of listeners would produce the same mean rating score for the same speech material. In other words, high inter-rater reliability implies high reproducibility of results. In contrast, a low value of *inter-rater* reliability would suggest that the listeners were not consistent in their quality judgments.

The efficacy of reliability measures has been studied extensively in behavioral sciences (see reviews in [19,24]) as well as in voice research where pathological voices are rated by clinicians in terms of breathiness or roughness [18,25,26]. More detailed description about the *intra-* and *inter-rater* reliability measures can be found in [2, Chap. 10].

2.3.2 Using Statistical Tests to Assess Significant Differences: Required Practice

After conducting subjective quality tests and collecting the ratings from all subjects, we often want to compare the performance of various algorithms. At the very least, we are interested in knowing whether a specific algorithm improves the speech quality over the baseline condition (i.e., un-processed speech). Consider for instance the MOS ratings scores obtained by 10 listeners in Table 6 when presented with speech processed by different algorithms. The mean MOS score for speech processed by algorithm A was 3.24, and the mean rating score for speech processed by algorithm B was 3.76. For this example, can we safely say with confidence that algorithm B improved the subjective speech quality relative to algorithm A? The answer is no, as it depends largely on the inter-rater reliability of quality judgments or grossly on the variance of the rating scores. Consider the Example 2 in Table 6 contrasting the rating scores of speech processed by say two different algorithms, C and D. The mean rating scores are identical to those obtained by algorithms A and B, however, the variance of the rating scores is high, suggesting that the inter-rater reliability in Example 2 was low (i.e., subjects were not consistent with each other when making quality judgments). In brief, we can not reach a conclusion, based solely on the mean rating scores, as to which algorithm performs better without first performing the appropriate statistical test.

Table 6. Example MOS ratings of 10 listeners for speech processed by algorithms A-D

	Example 1		Example 2	
Subjects	Alg. A	Alg. B	Alg. C	Alg. D
1	3.10	3.60	1.80	1.80
2	3.20	3.70	2.60	1.50
3	3.50	4.00	3.50	4.00
4	3.30	3.80	4.50	4.90
5	3.40	3.90	2.50	3.70
6	3.20	3.70	3.50	3.90
7	3.50	4.00	4.10	4.50
8	3.10	3.60	4.60	5.00
9	3.00	3.50	2.10	4.60
10	3.10	3.80	3.20	3.70
Mean	3.24	3.76	3.24	3.76
Variance	0.03	0.03	0.96	1.46

Statistical techniques [27, ch. 4] can be used to draw inferences about the means of two populations, which in our case correspond to the ratings of processed and un-processed speech or more generally to ratings obtained using two different algorithms. The t-statistic can often be used to test two hypothesis, the null hypothesis that the means are equal, and the alternate hypothesis that the means are different. The computed value of *t* will determine if we will accept or reject the null hypotheses. If the value of *t* is found to be greater than a critical value (found in statistics tables), then we reject the null hypothesis and therefore conclude that the means of the two populations are different. For the example in Table 6, if *t* is found to be larger than the critical value, we conclude that there is a *statistically significant* difference in quality and that algorithm B produced better speech quality than algorithm A. If the value of *t* is found to be smaller than the critical value, then we accept the null hypothesis and conclude that the means of the two populations do not differ, i.e., performance (quality) of algorithm A is as good as performance of algorithm B. For the Example 1 in Table 6, t-tests revealed that the rating scores of algorithm B are significantly higher than the ratings of algorithm A, i.e., algorithm B performed better than algorithm A. For the Example 2 in Table 6, however, t-tests revealed non-significant differences between the ratings of algorithms C and D. In other words, algorithm D did not improve speech quality relative to algorithm C. As the examples in Table 6 illustrate, we can not draw conclusions as to which algorithm improves quality based solely on the mean rating scores (the mean scores were identical in examples 1 and 2).

The above t-test applies only when we want to compare the means of two populations. It is tempting to run pair-wise comparisons of the population means using multiple t-tests to answer the above questions. However, the probability of falsely rejecting *at least one* of the hypotheses increases as the number of *t* tests increases. That is, although we may set the probability of Type I error at the $\alpha = 0.05$ level for each individual test, the probability of falsely rejecting *at least one* of those tests might be much larger than 0.05. For the above reason, multiple pairwise comparisons are recommended with Bonferroni correction. The Bonferroni test is based on Student's t statistic and adjusts the observed significance level based on the fact that multiple comparisons are made. This is simply done by multiplying the observed significance level by the number of comparisons made. Alternate statistical tests, including the analysis of variance, are described in [2, Ch. 10].

For the relative preference listening tests, one-sided t-tests need to be run to assess whether algorithm A is preferred over algorithm B beyond the chance level, which is 50%.

In summary, no reliable conclusions can be drawn based solely on the mean rating scores collected from subjective listening tests. The appropriate statistical test needs to be run to truly assess whether a particular algorithm improved (or not) speech quality.

3 Objective Quality Measures

Subjective listening tests provide perhaps the most reliable method for assessment of speech quality. These tests, however, can be time consuming requiring in most cases access to trained listeners. For these reasons, several researchers have investigated the possibility of devising objective, rather than subjective, measures of speech quality [1, ch. 2]. Ideally, the objective measure should be able to assess the quality of the processed speech without needing access to the original speech signal. The objective measure should incorporate knowledge from different levels of processing including low-level processing (e.g., psychoacoustics) and higher level processing such as prosodics, semantics, linguistics and pragmatics. The ideal measure should predict with high accuracy the results obtained from subjective listening tests with normal-hearing listeners. In addition, it should take into account inherent differences between languages (e.g., Western languages vs. tonal languages) [28].

Much progress has been done in developing such an objective measure [1]. In fact, one such measure has been standardized [29]. Current objective measures are limited in that most require access to the original speech signal, and some can only model the low-level processing (e.g., masking effects) of the auditory system. Yet, despite these limitations some of these objective measures have been found to correlate well with subjective listening tests (e.g., MOS scores). A different class of measures, known as non-intrusive measures, does not require access to the original signal. Figure 2 shows how the conventional (also referred to as intrusive) measures and the non-intrusive measures are computed. This Chapter will focus primarily on the intrusive measures, as those measures have been studied the most. A brief introduction and literature review on non-intrusive measures will also be given.

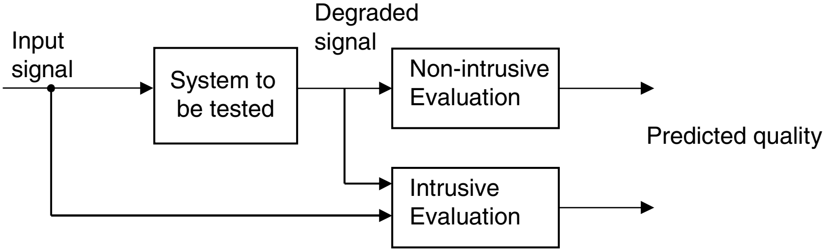


Fig. 2. Computation of intrusive and non-intrusive objective measures

Most objective measures of speech quality are implemented by first segmenting the speech signal into 10-30 ms frames, and then computing a distortion measure between the original and processed signals. A single, global measure of speech distortion is computed by averaging the distortion measures of each speech frame. More sophisticated objective measures [30,31] deviate from the above short-time frame-processing framework and also involve a time-delay estimation block for aligning the two signals prior to the distortion measure computation. As we will see shortly, the distortion measure computation can be done either in the time domain (e.g., signal-to-noise ratio measures) or in the frequency domain (e.g., LPC spectral distance measures). For the frequency-domain measures, it is assumed that any distortions or differences detected in the magnitude spectra are correlated with speech quality. Note that the distortion measures are not distance measures in the strict sense, as they do not obey all properties of a distance metric. For one, these measures are not necessarily symmetric and some (e.g., log spectral distance measure) yield negative values. Psychoacoustic experiments [32] suggest that the distance measures should not be symmetric [33].

A large number of objective measures has been evaluated, particularly for speech coding [1] and speech enhancement [34] applications. Reviews of objective measures can be found in [35-38]. Next, we focus on a subset of those measures.

3.1 Time and Frequency Signal-to-Noise Ratio Measures

The segmental signal-to-noise ratio can be evaluated either in the time or frequency domain. The time-domain measure is perhaps one of the simplest objective measures used to evaluate speech enhancement or speech coding algorithms. For this measure to be meaningful it is important that the original and processed signals be aligned in time and that any phase errors present be corrected. The segmental signal-to-noise (SNRseg) is defined as:

$$\text{SNRseg}=\frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum\limits_{n=Nm}^{Nm+N-1}x^2(n)}{\sum\limits_{n=Nm}^{Nm+N-1}(x(n)-\hat{x}(n))^2} \tag{1}$$

where $x(n)$ is the original (clean) signal, $\hat{x}(n)$ is the enhanced signal, N is the frame length (typically chosen to be 15-20 msecs), and M is the number of frames in the signal.

One potential problem with the estimation of SNRseg is that the signal energy during intervals of silence in the speech signal (which are abundant in conversational speech) will be very small resulting in large negative SNRseg values, which will bias the overall measure. One way to remedy this is to exclude the silent frames from the sum in Eq. (1) by comparing short-time energy measurements against a threshold or by flooring the SNRseg values to a small value. In [39], the SNRseg values were limited in the range of [-10 dB, 35 dB] thereby avoiding the need for a speech/silence detector.

The segmental SNR can be extended in the frequency domain to produce the frequency-weighted segmental SNR (fwSNRseg) [40]:

$$\text{fwSNRseg}=\frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum\limits_{j=1}^{K}B_j\log_{10}\left[\dfrac{F^2(m,j)}{(F(m,j)-\hat{F}(m,j))^2}\right]}{\sum\limits_{j=1}^{K}B_j} \tag{2}$$

where B_j is the weight placed on the jth frequency band, K is the number of bands, M is the total number of frames in the signal, $F(m, j)$ is the filter-bank amplitude (excitation spectrum) of the clean signal in the jth frequency band at the mth frame, and $\hat{F}(m, j)$ is the filter-bank amplitude of the enhanced signal in the same band. The main advantage in using the frequency-based segmental SNR over the time-domain SNRseg (Eq. (1)) is the added flexibility to place different weights for different frequency bands of the spectrum. There is also the flexibility in choosing perceptually-motivated frequency spacing such as critical-band spacing.

Various forms of weighting functions B_j were suggested in [1,40]. One possibility is to choose the weights B_j based on articulation index studies [41]. Such an approach was suggested in [1] with the summation in Eq. (2) taken over 16 articulation bands spanning the telephone bandwidth (300-3400 Hz).

3.2 Spectral Distance Measures Based on LPC

Several objective measures were proposed based on the dissimilarity between all-pole models of the clean and enhanced speech signals [1]. These measures assume that over short-time intervals speech can be represented by a pth order all-pole model of the form:

$$x(n) = \sum_{i=1}^{p} a_x(i)x(n-i) + G_x u(n) \tag{3}$$

where $a_x(i)$ are the coefficients of the all-pole filter (determined using linear prediction techniques), G_x is the filter gain and $u(n)$ is a unit variance white noise excitation. Perhaps two of the most common all-pole based measures used to evaluate speech-enhancement algorithms are the log likelihood ratio and Itakura-Saito measures. Cepstral distance measures derived from the LPC coefficients were also used.

The log-likelihood ratio (LLR) measure is defined as:

$$d_{LLR}(\mathbf{a}_x, \bar{\mathbf{a}}_{\hat{x}}) = \log \frac{\bar{\mathbf{a}}_{\hat{x}}^T \mathbf{R}_x \bar{\mathbf{a}}_{\hat{x}}}{\mathbf{a}_x^T \mathbf{R}_x \mathbf{a}_x} \tag{4}$$

where $\mathbf{a}_x^T$ are the LPC coefficients of the clean signal, $\bar{\mathbf{a}}_{\hat{x}}^T$ are the coefficients of the enhanced signal, and $\mathbf{R}_x$ is the $(p+1)\times(p+1)$ autocorrelation matrix (Toeplitz) of the clean signal. This measure penalizes differences in formant peak locations.

The Itakura-Saito (IS) measure is defined as follows:

$$d_{IS}(\mathbf{a}_x, \bar{\mathbf{a}}_{\hat{x}}) = \frac{G_x}{\bar{G}_{\hat{x}}} \frac{\bar{\mathbf{a}}_{\hat{x}}^T \mathbf{R}_x \bar{\mathbf{a}}_{\hat{x}}}{\mathbf{a}_x^T \mathbf{R}_x \mathbf{a}_x} + \log\left(\frac{\bar{G}_{\hat{x}}}{G_x}\right) - 1 \tag{5}$$

where G_x and $\overline{G}_{\hat{x}}$ are the all-pole gains of the clean and enhanced signals respectively. Note that unlike the LLR measure, the IS measure penalizes differences in all-pole gains, i.e., differences in overall spectral levels of the clean and enhanced signals. This can be considered as a drawback of the IS measure, since psychoacoustic studies [42] have shown that differences in spectral level have minimal effect on quality.

A gain-normalized spectral distortion (SD) measure is often used to assess the quality of coded speech spectra. The SD measure evaluates the similarity of the LPC spectra of the clean and processed signals [3,33].

The LPC coefficients can also be used to derive a distance measure based on cepstrum coefficients. This distance provides an estimate of the log spectral distance between two spectra. The cepstrum coefficients can be obtained recursively from the LPC coefficients $\{a_j\}$ using the following expression [43, p. 442]:

$$c(m) = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c(k) a_{m-k} \qquad 1 \le m \le p \tag{6}$$

where p is the order of the LPC analysis (Eq. (3)). A measure based on cepstrum coefficients can be computed as follows [44]:

$$d_{cep}(\mathbf{c}_x, \overline{\mathbf{c}}_{\hat{x}}) = \frac{10}{\log_e 10} \sqrt{2 \sum_{k=1}^{p} \left[c_x(k) - c_{\hat{x}}(k) \right]^2} \tag{7}$$

where $c_x(k)$ and $c_{\hat{x}}(k)$ are the cepstrum coefficients of the clean and enhanced signals respectively.

3.3 Perceptually-Motivated Measures

The above objective measures are attractive in that they are simple to implement and easy to evaluate. However, their ability to predict subjective quality is limited as they do not closely emulate the signal processing involved at the auditory periphery. For one, the normal-hearing frequency selectivity as well as the perceived loudness were not explicitly modeled or incorporated in the measures. Much research [42,45-50] has been done to develop objective measures based on models of human auditory speech perception, and in this section we describe some of these perceptually-motivated measures.

3.3.1 Bark Distortion Measures

Much progress has been made on modeling several stages of the auditory processing, based on existing knowledge from psychoacoustics about how human listeners process tones and bands of noise [51, ch. 3]. Specifically, these new objective measures take into account the fact that:

1. The ear's frequency resolution is not uniform, i.e., the frequency analysis of acoustic signals is not based on a linear frequency scale. This can be modeled by pre-processing the signal through a bank of bandpass filters with center frequencies and bandwidths increasing with frequency. These filters have come be known in the psychoacoustics literature as critical-band filters and the corresponding frequency spacing as critical-band spacing.
2. Loudness is related to signal intensity in a nonlinear fashion. This takes into account the fact that the perceived loudness varies with frequency [52,53].

One such measure that takes the above into account is the Bark distortion measure (BSD). The BSD measure for frame *k* is based on the difference between the loudness spectra and is computed as follows:

$$BSD(k) = \sum_{b=1}^{N_b} \left[S_k(b) - \overline{S}_k(b) \right]^2 \tag{8}$$

where $S_k(b)$ and $\overline{S}_k(b)$ are the loudness spectra of the clean and enhanced signals respectively and N_b is the number of critical bands. The mean BSD measure is finally computed by averaging the frame BSD measures across the sentence. Experiments in [46] indicated that the BSD measure yields large values for the low-energy (unvoiced) segments of speech. This problem can be avoided by excluding the low-energy segments of speech from the BSD computation using a voiced/unvoiced detector. Improvements to the BSD measure were reported in [47,54,55] leading to the modified BSD measure (MBSD). Experiments in [46,47] indicated that both BSD and MBSD measures yielded a high correlation ($\rho > 0.9$) with MOS scores. Further improvements to the MBSD measure were proposed in [54,56].

3.3.2 Perceptual Evaluation of Speech Quality (PESQ) Measure

Most of the above objective measures have been found to be suitable for assessing only a limited range of distortions which do not include distortions commonly encountered when speech goes through telecommunication networks. Packet loss, for instance, signal delays and codec distortions would cause most objective measures to produce inaccurate predictions of speech quality. A number of objective measures were proposed in the 1990s focusing on this type of distortions as well as filtering effects and variable signal delays [31,57,58].

A competition was held in 2000 by the ITU-T study group 12 to select a new objective measure capable of performing reliably across a wide range of codec and network conditions. The perceptual evaluation of speech quality (PESQ) measure, described in [30], was selected as the ITU-T recommendation P.862 [29] replacing the old P.861 recommendation [59]. The latter recommendation proposed a quality assessment algorithm called perceptual speech quality measure (PSQM). The scope of PSQM is limited to assessing distortions introduced by higher-bit speech codecs operating over error-free channels.

The structure of the PESQ measure is shown in Figure 3. The original (clean) and degraded signals are first level equalized to a standard listening level, and filtered by a filter with response similar to a standard telephone handset. The signals are aligned in time to correct for time delays, and then processed through an auditory transform, similar to that of BSD, to obtain the loudness spectra. The absolute difference between the degraded and original loudness spectra is used as a measure of audible error in the next stage of PESQ computation. Note that unlike most objective measures (e.g., the BSD measure) which treat positive and negative loudness differences the same (by squaring the difference), the PESQ measure treats these differences differently. This is because positive and negative loudness differences affect the perceived quality differently. A positive difference would indicate that a component, such as noise, has been added to the spectrum, while a negative difference would indicate that a spectral component has been omitted or heavily attenuated. Compared to additive components, the omitted components are not as easily perceived due to masking effects, leading to a less objectionable form of distortion. Consequently, different weights are applied to positive and negative differences. The differences, termed the disturbances, between the loudness spectra is computed and averaged over time and frequency to produce the prediction of subjective MOS score. The final PESQ score is computed as a linear combination of the average disturbance value d_{sym} and the average asymmetrical disturbance value d_{asym} as follows:

$$PESQ = a_0 + a_1 \cdot d_{sym} + a_2 \cdot d_{asym} \tag{9}$$

where $a_0 = 4.5$, $a_1 = -0.1$ and $a_2 = -0.0309$. The range of the PESQ score is –0.5 to 4.5, although for most cases the output range will be a MOS-like score, i.e., a score between 1.0 and 4.5. High correlations ($\rho > 0.92$) with subjective listening tests were reported in [30] using the above PESQ measure for a large number of testing conditions taken from mobile, fixed and voice over IP (VoIP) applications. The PESQ can be used reliably to predict the subjective speech quality of codecs (waveform and CELP-type coders) in situations where there are transmission channel errors, packet loss or varying delays in the signal. It should be noted that the PESQ measure does not provide a comprehensive evaluation of telephone transmission quality, as it only reflects the effects of one-way speech or noise distortion perceived by the end-user. Effects such as loudness loss, sidetone and talker echo are not reflected in the PESQ scores. More details regarding the PESQ computation can be found in [2, Ch. 10].

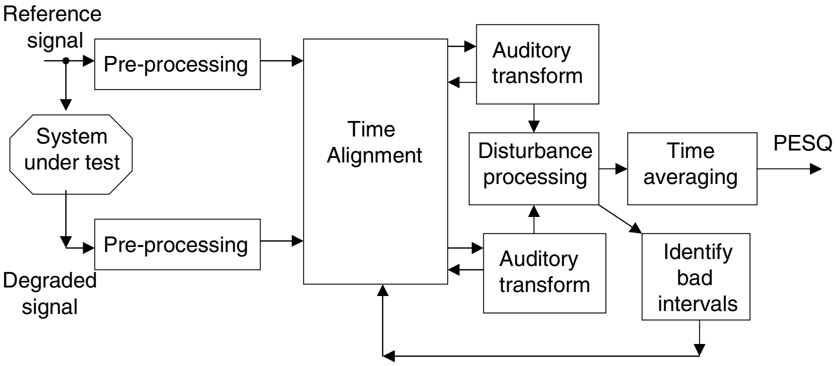


Fig. 3. Block diagram of the PESQ measure computation

3.4 Composite Measures

In addition to the above measures, one can form the so called *composite measures* [1, Ch. 9] by combining multiple objective measures. The rational behind the use of composite measures is that different objective measures capture different characteristics of the distorted signal, and therefore combining them in a linear or nonlinear fashion can potentially yield significant gains in correlations. Regression analysis can be used to compute the optimum combination of objective measures for maximum correlation. One possibility is to use the following linear regression model:

$$\begin{aligned} y_i &= f(\mathbf{x}) + \varepsilon_i \\ &= \alpha_0 + \sum_{j=1}^{P} \alpha_j x_{ij} + \varepsilon_i \end{aligned} \tag{10}$$

where $f(\mathbf{x})$ is the mapping function presumed to be linear, P is the number of objective measures involved, $\{y_i\}_{i=1}^{N}$ are the dependent variables corresponding to the subjective ratings of N samples of degraded speech, x_{ij} is the independent (predictor) variable corresponding to the jth objective measure computed for the ith observation (degraded sample or condition), and ε_i is a random error associated with each observation. The regression coefficients α_i can be estimated to provide the best fit with the data using a least-squares approach [1, p. 184]. The P objective measures considered in (10) may include, among other measures, the LPC-based measures (e.g., IS, LLR), segmental SNR measures (e.g., SNRseg) or the PESQ measure. The selection of objective measures to include in the composite measure is not straightforward and in some cases it is based solely on experimental evidence (trial and error) and intuition. Ideally, we would like to include objective measures that capture complementary information about the underlying distortions present in the degraded signal.

A linear function $f(\mathbf{x})$ was assumed in (10) for mapping P objective measures to the observed subjective ratings, $\{y_i\}_{i=1}^{N}$. Such a model is accurate only when the true form of the underlying function is linear. If it is not, then the modeling error will likely be large and the fit will be poor. Non-parametric models which make no assumptions about the form of the mapping function can alternatively be used. More specifically, models based on multivariate adaptive regression splines (MARS) have been found to yield better performance for arbitrary data sets [60]. Unlike linear and polynomial regression analysis, the MARS modeling technique is data driven and derives the functional form from the data. The basic idea of the MARS modeling technique is to recursively partition the domain into smaller sub-regions and use spline functions to locally fit the data in each region. The number of splines used in each sub-region is automatically determined from the data. The MARS model has the following form:

$$y_i = \alpha_0 + \sum_{j=1}^{M} \alpha_j B_j(\mathbf{x}) + \varepsilon_i \tag{11}$$

where $B_j(\mathbf{x})$ are the basis functions and M is the number of basis functions which are automatically determined from the data (note that M could be larger than the number of objective measures). The MARS technique has been successfully applied to speech quality evaluation in [34,61]. Radial basis functions were used in [49,50] for $B_j(\mathbf{x})$. Good correlations were obtained in [50] in terms of predicting the quality of noise-suppressed speech.

While the composite measures always improve the correlation, caution needs to be exercised in as far using these measures with test speech materials and distortions other than the ones that have been validated. The reason for this is that the composite measures need to be cross-validated with conditions not included in the training stage, hence they will perform the best when tested with the same speech materials containing processed speech with similar distortions.

3.5 Non-intrusive Objective Quality Measures

The above objective measures for evaluating speech quality are "intrusive" in nature as they require access to the input (clean) signal. These measures predict speech quality by estimating the "distortion" between the input (clean) and output (processed) signals and then mapping the estimated "distortion" value to a quality metric. In some applications, however, the input (clean) signal is not readily available and therefore the above objective measures are not practical or useful. In VoIP applications, for instance, where we are interested in monitoring continuously the performance of telecommunication networks (in terms of speech quality), we only have access to the output signal. In such cases, a non-intrusive objective measure of speech quality would be highly desirable for continuous monitoring of quality of speech delivered to a customer or to a particular point in the network. Based on such quality assessment, network traffic can be routed, for instance, through less congested parts of the network and therefore improve the quality of service.

A fundamentally different approach is required to analyze a processed signal when the clean (reference) input signal is not available, and several *non-intrusive* measures have been proposed in the literature [61-67]. Some methods are based on comparing the output signal to an artificial reference signal derived from an appropriate codebook [65,66]. Other methods use vocal-tract models to identify distortions [63]. This latter method [63] first extracts a set of vocal-tract shape parameters (e.g., area functions, cavity size) from the signal, and then evaluates these parameters for physical production violations, i.e., whether the parameters could have been generated by the human speech-production system. Distortions are identified when the vocal-tract parameters yield implausible shape and cavity sizes. A variant of the vocal-tract method was adopted as the ITU-T P.563 [68] standard for non-intrusive evaluation of speech quality. More information on non-intrusive methods can be found in [62].

3.6 Evaluation of Objective Quality Measures

So far we have not yet discussed what makes a certain objective measure better than other. Some objective measures are "optimized" for a particular type of distortion and may not be meaningful for another type of distortion. The task of evaluating the validity of objective measures over a wide range of distortions is immense [1]. A suggested process to follow is to create a large database of speech distorted in various ways and evaluate the objective measure for each file in the database and for each type of distortion [1, ch 1]. At the same time, the distorted database needs to be evaluated by human listeners using one of the subjective listening tests (e.g., MOS test) described above. Statistical analysis needs to be used to assess the correlation between subjective scores and the values of the objective measures. For the objective measure to be valid and useful, it needs to correlate well with subjective listening tests. A discussion is given next on how to assess the predictive power of objective measures followed by a presentation of some of the measures that have been found to correlate well with listening tests.

3.6.1 Figures of Merit

The correlation between subjective listening scores and objective measures can be obtained using the Pearson's correlation coefficient which is computed as follows:

$$\rho = \frac{\sum_d (S_d - \bar{S}_d)(O_d - \bar{O}_d)}{[\sum_d (S_d - \bar{S}_d)^2]^{1/2}[\sum_d (O_d - \bar{O}_d)^2]^{1/2}} \qquad (12)$$

where S_d are the subjective quality ratings, O_d are the values of the objective measure, and $\bar{S}_d$ and $\bar{O}_d$ are the mean values of S_d and O_d respectively. This correlation coefficient ρ can be used to predict the subjective results based on the values of the objectives measures as follows:

$$P_k = \bar{P} + \rho \frac{\sigma_P}{\sigma_O}\left(O_k - \bar{O}\right) \tag{13}$$

where O_k denotes the value of the objective measure obtained for the *k*th speech file in the database, P_k denotes the predicted subjective listening score, σ_P and σ_O denote the standard deviations of the subjective and objective scores respectively, $\bar{P}$ and $\bar{O}$ denote the mean values of the subjective and objective scores respectively. Note that Eq. (13) is based on first-order linear regression analysis assuming a single objective measurement. Higher order polynomial regression analysis could also be used if the objective measure is composed of multiple measurements [1, ch. 4.5].

A second figure-of-merit is an estimate of the standard deviation of the prediction error obtained by using the objective measures to predict the subjective listening scores. This figure-of-merit is computed as:

$$\sigma_e = \sigma_P \sqrt{1-\rho^2} \tag{14}$$

where σ_e is the *standard error of the estimate.* The standard error of the estimate of the subjective scores provides a measure of variability of the subjective scores about the regression line, averaged over all objective scores. For good predictability of the subjective scores, we would like the objective measure to yield a small value of σ_e. Both figures of merit, i.e., correlation coefficient and standard error of the estimate σ_e, need to be reported when evaluating objective measures. In some cases, histograms of the absolute residual errors, computed as the difference between the predicted and actual scores, can provide valuable information similar to that provided by σ_e. Such histograms can provide a good view of how frequently errors of different magnitudes occur.

An alternative figure-of-merit to σ_e is the root-mean-square error (RMSE) between the per condition averaged objective measure and subjective ratings computed over all conditions:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}\left(\bar{S}_i - \bar{O}_i\right)^2}{M}}$$

where $\overline{S}_i$ indicates the averaged subjective score in *i*th condition, $\overline{O}_i$ indicates the averaged objective score in *i*th condition and *M* is the total number of conditions.

The above analysis assumes that the objective and subjective scores are *linearly* related (see Eq. (13)). This is not always the case, however, and in practice, it is not easy to uncover the best-fitting function or the true relationship between the objective and subjective measurements. Scatter plots of the rating scores vs. objective scores can provide valuable insights in terms of unveiling the relationship between the objective and subjective measurements. Some found a better fit with a quadratic relationship [44,46] while others found a good fit with a logistic function [69]. Kitawaki *et al* .[44], for instance, derived a quadratic expression for predicting MOS scores from cepstral distance measures for Japanese speech. Nonparametric regression techniques, such as the MARS technique [60] can alternatively be used to uncover the mapping function between (multiple) objective measures and subjective ratings (see Section 3.4).

3.6.2 Correlations of Objective Measures with Subjective Listening Tests

Objective measures need to be validated with ratings obtained in subjective listening tests with human listeners. The choice of objective measures needs to be made carefully depending on the application, language and type of distortions present in the processed speech.

For distortions introduced by speech coders, for instance, the objective measures investigated in [1] are appropriate. High correlations (ρ >0.9) were obtained primarily with composite and frequency-variant measures. The LPC-based measures performed modestly well ($\rho < 0.62$). The SNRseg measure performed well, but only for distortions introduced by waveform speech coders (e.g., ADPCM). This suggests that the SNRseg measure is *only* appropriate for evaluating speech processed via waveform coders. For distortions, such as clipping, introduced by hearing aids the coherence-based measures reported in [70,71] are appropriate.

For distortions introduced by speech-enhancement algorithms, the objective measures discussed and evaluated in [34] are appropriate. These measures were evaluated using the publicly available noisy speech corpus (NOIZEUS[2]), which was used in a comprehensive subjective quality evaluation [72] of 13 different speech enhancement algorithms encompassing four different classes of algorithms: spectral subtractive, subspace, statistical-model based and Wiener-filtering type algorithms. The enhanced speech files were sent to Dynastat, Inc (Austin, TX) for subjective evaluation using the standardized methodology for evaluating noise suppression algorithms based on ITU-T P.835 [16]. The use of ITU-T P.835 methodology yielded three rating scores for each algorithm: an overall quality rating, a signal distortion rating and a background distortion rating. A summary of the resulting correlations is given in Table 7 for a subset of the objective measures tested.

[2] Available at: `http://www.utdallas.edu/~loizou/speech/noizeus/`

Table 7. Estimated correlation coefficients ($|\rho|$) of objective measures with overall quality, signal distortion and background noise distortion [34]

Objective measure	Overall quality	Signal distortion	Background distortion
SegSNR	0.36	0.22	0.56
Weighted spectral slope (WSS) [87]	0.64	0.59	0.62
PESQ	0.89	0.81	0.76
Log-likelihood ratio (LLR)	0.85	0.88	0.51
Itakura-Saito distance (IS)	0.60	0.73	0.09
Cepstrum distance (CEP)	0.79	0.84	0.41
fwSNRseg	0.85	0.87	0.59
Modified PESQ	0.92	0.89	0.76

In addition to several conventional objective measures (most of which were described in this Section), modifications to the PESQ measure were also considered in [34]. As it was not expected that the PESQ measure would correlate highly with all three rating scores (speech distortion, noise distortion and overall quality), the PESQ measure was optimized for each of the three rating scales by choosing a different set of parameters (a_0, a_1, a_2) in Eq. (9) for each rating scale. Multiple linear regression analysis was used to determine the values of the parameters a_0, a_1, a_2. Of the seven basic objective measures tested, the PESQ measure yielded the highest correlation ($\rho = 0.89$) on overall quality, followed by the fwSNRseg and LLR measures ($\rho = 0.85$). Even higher correlation with overall quality was obtained with the modified PESQ measure ($\rho = 0.92$). The majority of the basic objective measures predicted equally well signal distortion and overall quality, but not background distortion. This was not surprising given that most measures take into account both speech-active and speech-absent segments in their computation. Measures that would place more emphasis on the speech-absent segments would be more appropriate and likely more successful in predicting noise distortion. The SNRseg measure, which is widely used for evaluating the performance of speech enhancement algorithms, yielded a very poor correlation coefficient ($\rho = 0.31$) with overall quality. This outcome suggests that the SNRseg measure is unsuitable for evaluating the performance of enhancement algorithms.

In summary, the PESQ measure has proved to be the most reliable measure for assessing speech quality. Consistently high correlations were noted for speech processed by speech codecs and telephone networks [30] as well as for noisy speech processed by speech-enhancement algorithms [34]. High correlations were also obtained with the PESQ measure in Mandarin Chinese speech processed through various speech codecs [73]. Although not designed to predict speech intelligibility, the PESQ measure has also yielded a modestly high correlation ($\rho = 0.79$) with intelligibility scores [74], at least when tested with English speech. Modifications of the PESQ measure for Mandarin Chinese were reported in [28]. High correlation with speech intelligibility was also obtained with the fwSNRseg measure (Eq. (2)).

4 Challenges and Future Directions in Objective Quality Evaluation

Presently, there is no single objective measure that correlates well with subjective listening evaluations for a wide range of speech distortions. Most measures have been validated for a specific type of distortion and for a specific language. Some measures correlate well with distortions introduced by speech coders while others (e.g., PESQ measure) correlate well with distortions introduced by telecommunication networks and speech-enhancement algorithms. While the PESQ measure has been shown to be a robust objective measure, it is computationally demanding and requires access to the whole utterance. In some applications, this might not be acceptable. Ideally, the objective measure should predict the quality of speech independent of the type of distortions introduced by the system whether be a network, a speech coder or a speech enhancement algorithm. This is extremely challenging and would require a deeper understanding of the human perceptual processes involved in quality assessment.

For one, little is known as to how we should best integrate or somehow combine the frame computed distance measures to a single global distortion value. The simplest approach used in most objective measures is to compute the arithmetic mean of the distortions computed in each frame, i.e.,

$$D = \frac{1}{M} \sum_{k=0}^{M-1} d(\mathbf{x}_k, \overline{\mathbf{x}}_k) \tag{15}$$

where M is the total number of frames, D denotes the global (aggregate) distortion, and $d(\mathbf{x}_k, \overline{\mathbf{x}}_k)$ denotes the distance between the clean and processed signals in the kth frame. This distance measure could take, for instance, the form of either (4), (5), or (8). The averaging in Eq. (15) implicitly assumes that all frames (voiced, unvoiced and silence) should be weighted equally, but this is not necessarily consistent with quality judgments. For one, the above averaging does not take into account temporal (forward or backward) masking effects.

Alternatively, we can consider using a time-weighted averaging approach to estimate the global distortion, i.e.,

$$D_W = \frac{\sum_{k=0}^{M-1} w(k) d(\mathbf{x}_k, \overline{\mathbf{x}}_k)}{\sum_{k=0}^{M-1} w(k)} \tag{16}$$

where $w(k)$ represents the weighting applied to the kth frame. Computing the frame weights, $w(k)$, however, is not straightforward and no optimal methods (at least in the perceptual sense) exist to do that.

Accurate computation of $w(k)$ would require a deeper understanding of the factors influencing quality judgments at least at two conceptual levels: the suprasegmental (spanning syllables or sentences) and the segmental (spanning a single phoneme) levels. At the suprasegmental level we need to know how humans integrate information across time, considering at the very least temporal (non-simultaneous) masking effects such as forward and backward masking. Forward masking is an auditory phenomenon which occurs when large energy stimuli (maskers) precede in time, and suppress (i.e., mask) later arriving and lower energy stimuli from detection. In the context of speech enhancement, this means that the distortion introduced by the noise-reduction algorithm may be detectable beyond the time window in which the signal and distortion are simultaneously present. Masking may also occur before the masker onset and the corresponding effect is called backward masking [75,ch. 4]. Back-ward masking effects are relatively short (less than 20ms), but forward-masking effects can last longer than 100 msecs [75, ch. 4.4] and its effects are more dominant. Attempts to model forward masking effects were reported in [1, p. 265,45,55].

At the segmental (phoneme) level, we need to know which spectral characteristics (e.g., formants, spectral tilt, etc) of the signal affect quality judgments the most. These characteristics might also be language dependent [76], and the objective measure needs to take that into account (e.g., [28]). We know much about the effect of spectral manipulations on perceived vowel quality but comparatively little on consonant quality [42,77]. Klatt [42] demonstrated that of all spectral manipulations (e.g., low-pass filtering, notch filtering, spectral tilt) applied to vowels, the formant frequency changes had the largest effect on quality judgments. His findings, however, were only applicable to vowels and not necessarily to stop consonants or any other sound class. For one, Klatt concluded that spectral tilt is unimportant in vowel perception [42], but that is not the case however in stop-consonant perception. We know from the speech perception literature that spectral tilt is a major cue to stop place of articulation [78, ch. 6,79]. Some [79] explored the idea of constructing a spectral template that could be associated with each place of stop articulation, and used those templates to classify stops. In brief, the stop consonants, and possibly the other consonants, need to be treated differently than vowels, since different cues are used to perceive consonants.

There has been a limited number of proposals in the literature on how to estimate the weights $w(k)$ in (16) or how to best combine the local distortions to a single global distortion value [1,1, ch. 7,45,69,80,81]. In [80,82], the weights

$w(k)$ were set proportional to the frame energy (raised to a power) thereby placing more emphasis on voiced segments. This approach, however, did not yield any significant benefits as far as obtaining a better correlation with subjective listening tests [1, p. 221,82]. A more successful approach was taken in [83] for assessing distortions introduced by hearing aids. Individual frames were classified into three regions relative to the overall RMS level of the utterance, and the objective measure was computed separately for each region. The high-level region consisted of segments at or above the overall RMS level of the whole utterance. The mid-level region consisted of segments ranging from the overall RMS level to 10 dB below, and the low-level region consisted of segments ranging from RMS-10 dB to RMS-30 dB. A similar approach was also proposed in [84].

Rather than focusing on finding suitable weights for Eq. (16), some have proposed alternative methods to combine the local distortions into a single global distortion value. In [80], a classifier was used to divide the speech frames into four distinct phonemic categories: vocalic, nasal, fricative and silence. A separate distortion measure was used for each phonemic class and the global distortion was constructed by linearly combining the distortions of the four classes. A similar approach was also proposed in [81] based on statistical pattern-recognition principles. The underlying assumption in these segmentation-based methods is that the distortion in various classes of sounds is perceived differently, and therefore a different weight ought to be placed to each class. It is not yet clear what those weights should be, and further research based on psychoacoustic experiments is needed to determine that.

A different approach for combining local distortions was proposed in [69] based on the assumption that the overall perceived distortion consists of two components. The first component takes the average distortion into account by treating all segments (frames) and all frequencies equally. The second component takes into account the distribution of the distortion over time and frequency. That is, it takes into consideration the possibility that the distortion might not be uniformly distributed across time/frequency but concentrated into a local time or frequency region. The latter distortion is computed using an information-theoretic measure borrowed from the video coding literature [85]. This measure, which is based on entropy, quantifies roughly the amount of information contained in each time-frequency cell and assigns the appropriate weight accordingly. The measuring normalizing blocks (MNB) algorithm [31] utilizes a simple perceptual transform, and a hierarchical structure of integration of distance measurements over a range of time and frequency intervals.

In most objective quality measures, the distortion is computed as the difference between the auditory spectra of the clean and processed signals or as the difference of their all-pole spectra (e.g., LPC) representations. This difference is commonly squared to ensure positivity of the distance measure. Squaring this difference, however, assumes that the positive and negative differences contribute equally to the perceived quality. But as mentioned earlier, that is not the case. A positive difference might sometimes be perceived more harshly and therefore be more objectionable than a negative difference. This is because the omitted components (produced by a negative difference) might sometimes be masked and therefore become inaudible. Objective measures should therefore treat positive and negative distortions differently. Yet, only a few objective measures take into account this asymmetrical effect of auditory spectra differences on quality judgments [30,31,86].

To summarize, further research is needed to address the following issues and questions for better objective quality evaluation:

1. At the suprasegmental level, we need a perceptually meaningful way to compute the weights $w(k)$ in (16), modeling at the very least temporal (forward) masking effects.
2. At the segmental (phoneme) level, we need to treat consonants differently than vowels since perceptually we use different cues to identify consonants and vowels. Certain spectral characteristics of the consonants and vowels need to be emphasized or deemphasized in the distortion calculation, and these characteristics will likely be different.
3. A different weight needs to be placed on positive and negative differences of the auditory spectral representation of the clean and processed signals.

To address the above issues, it will require a better understanding of the factors influencing human listeners in making quality judgments. For that, perception experiments similar to those reported in [42,45,77] need to be conducted.

5 Summary

This Chapter presented an overview of the various techniques and procedures that have been used to evaluate the quality of processed speech. A number of subjective listening tests were described for evaluating speech quality. These tests included relative preference methods and absolute category rating methods (e.g., MOS, DAM). The ITU-T P.835 standard established for evaluating quality of speech processed by noise-reduction algorithms was also described. Lastly, a description of common objective quality measures was provided. This included segmental SNR measures, spectral distance measures based on LPC (e.g., Itakura-Saito measure) and perceptually motivated measures (e.g., bark distortion measure, PESQ measure). The segmental SNR measure, which is often used to assess speech quality, was not found to correlate well with subjective rating scores obtained by human listeners, and should not be used. The PESQ measure has been proven to be the most reliable objective measure for assessment of speech quality [30,34], and to some degree, speech intelligibility [74].

References

[1] Quackenbush, S., Barnwell, T., Clements, M.: Objective measures of speech quality. Prentice Hall, Englewood Cliffs (1988)

[2] Loizou, P.: Speech Enhancement: Theory and Practice. CRC Press LLC, Boca Raton (2007)

[3] Grancharov, V., Kleijn, W.: Speech Quality Assessment. In: Benesty, J., Sondhi, M., Huang, Y. (eds.) Handbook of Speech Processing, pp. 83–99. Springer, Heidelberg (2008)

[4] Berouti, M., Schwartz, M., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, pp. 208–211 (1979)

[5] ITU-T, Subjective performance assessment of telephone band and wide-band digital codecs, ITU-T Recommendation p. 830 (1996)

[6] International Telecommunication Union - Radiocommunication Sector, Recommendation BS. 562-3, Subjective assessment of sound quality (1990)

[7] IEEE Subcommittee, IEEE Recommended Practice for Speech Quality Measurements. IEEE Trans. Audio and Electroacoustics AU-17(3), 225–246 (1969)

[8] International Telecommunication Union - Telecommunication Sector, Recommendation, Subjective performance assessment of telephone band and wideband digital codecs p. 830 (1998)

[9] IEEE Recommended Practice for Speech Quality Measurements. IEEE Trans. Audio and Electroacoustics AU-17(3),225–246 (1969)

[10] Coleman, A., Gleiss, N., Usai, P.: A subjective testing methodology for evaluating medium rate codecs for digital mobile radio applications. Speech Communication 7(2), 151–166 (1988)

[11] Goodman, D., Nash, R.: Subjective quality of the same speech transmission conditions in seven different countries. IEEE Trans. Communications COm-30(4), 642–654 (1982)

[12] Rothauser, E., Urbanek, G., Pachl, W.: A comparison of preference measurement methods. J. Acoust. Soc. Am. 49(4), 1297–1308 (1970)

[13] ITU-T, Methods for subjective determination of transmission quality, ITU-T Recommendation p. 800 (1996)

[14] Voiers, W.D.: Diagnostic Acceptability Measure for speech communication systems. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, pp. 204–207 (1977)

[15] Voiers, W.D., Sharpley, A., Panzer, I.: Evaluating the effects of noise on voice communication systems. In: Davis, G. (ed.) Noise Reduction in Speech Applications, pp. 125–152. CRC Press, Boca Raton (2002)

[16] ITU-T, Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, ITU-T Recommendation p. 835 (2003)

[17] Hu, Y., Loizou, P.: Subjective comparison of speech enhancement al-gorithms. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, vol. I, pp. 153–156 (2006)

[18] Kreiman, J., Kempster, G., Erman, A., Berke, G.: Perceptual evaluation of voice quality: Review, tutorial and a framework for future research. J. Speech Hear. Res. 36(2), 21–40 (1993)

[19] Suen, H.: Agreement, reliability, accuracy and vailidity: Toward a clarification. Behavioral Assessment 10, 343–366 (1988)

[20] Cronbach, L.: Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334 (1951)

[21] Kendall, M.: Rank correlation methods. Hafner Publishing Co., New York (1955)

[22] Shrout, P., Fleiss, J.: Intraclass correlations: Uses in assessing rater re-liability. Psychological Bulletin 86(2), 420–428 (1979)

[23] McGraw, K., Wong, S.: Forming inferences about some intraclass correlation coefficients. Psychological Methods 1(1),30–46 (1996)

[24] Tinsley, H., Weiss, D.: Interrater reliability and agreement of subjective judgments. J. Counseling Psychology 22(4), 358–376 (1975)

[25] Gerratt, B., Kreiman, J., Antonanzas-Barroso, N., Berke, G.: Compar-ing internal and external standards in voice quality judgments. J. Speech Hear. Res. 36, 14–20 (1993)

[26] Kreiman, J., Gerratt, B.: Validity of raing scale measures of voice quality. J. Acoust. Soc. Am. 104(3), 1598–1608 (1998)

[27] Ott, L.: An introduction to statistical methods and data analysis, 3rd edn. PWS-Kent Publishing Company, Boston (1988)

[28] Chong, F., McLoughlin, I., Pawlikoski, K.: A Methodology for Improving PESQ accuracy for Chinese Speech. In: TENCON Conference, pp. 1–6 (2005)

[29] ITU, Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation p. 862 (2000)

[30] Rix, A., Beerends, J., Hollier, M., Hekstra, A.: Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs. In: Proc. IEEE Int. Conf. Acoust, Speech, Signal Processing, vol. 2, pp. 749–752 (2001)

[31] Voran, S.: Objective estimation of perceived speech quality - Part I: Development of the measuring normalizing block technique. IEEE Transactions on Speech and Audio Processing 7(4), 371–382 (1999)

[32] Flanagan, J.: A difference limen for vowel formant frequency. J. Acoust. Soc. Am. 27, 613–617 (1955)

[33] Viswanathan, R., Makhoul, J., Russell, W.: Towards perceptually consistent measures of spectral distance. In: Proc. IEEE Int. Conf. Acoust, Speech, Signal Processing, vol. 1, pp. 485–488 (1976)

[34] Hu, Y., Loizou, P.: Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang Processing 16(1), 229–238 (2008)

[35] Dimolitsas, S.: Objective speech distortion measures and their relevance to speech quality assessments. In: IEE Proc. - Vision, Image and Signal Processing, vol. 136(5), pp. 317–324 (1989)

[36] Kubichek, R., Atkinson, D., Webster, A.: Advances in objective voice quality assessment. In: Proc. Global Telecommunications Conference, vol. 3, pp. 1765–1770 (1991)

[37] Kitawaki, N.: Quality assessment of coded speech. In: Furui, S., Sondhi, M. (eds.) Advances in Speech Signal Processing, pp. 357–385. Marcel Dekker, New York (1991)

[38] Barnwell, T.: Objective measures for speech quality testing. J. Acoust. Soc. Am. 66(6), 1658–1663 (1979)

[39] Hansen, J., Pellom, B.: An effective quality evaluation protocol for speech enhancement algorithms. In: Proc. Inter. Conf. on Spoken Language Processing, vol. 7, pp. 2819–2822 (1998)

[40] Tribolet, J., Noll, P., McDermott, B., Crochiere, R.E.: A study of complexity and quality of speech waveform coders. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, pp. 586–590 (1978)

[41] Kryter, K.: Methods for calculation and use of the articulation index. J. Acoust. Soc. Am. 34(11), 1689–1697 (1962)

[42] Klatt, D.: Prediction of perceived phonetic distance from critical band spectra. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, vol. 7, pp. 1278–1281 (1982)

[43] Rabiner, L., Schafer, R.: Digital processing of speech signals. Prentice Hall, Englewood Cliffs (1978)

[44] Kitawaki, N., Nagabuchi, H., Itoh, K.: Objective quality evaluation for low bit-rate speech coding systems. IEEE J. Select. Areas in Comm. 6(2), 262–273 (1988)

[45] Karjalainen, M.: A new auditory model for the evaluation of sound quality of audio system. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, vol. 10, pp. 608–611 (1985)

[46] Wang, S., Sekey, A., Gersho, A.: An objective measure for predicting subjective quality of speech coders. IEEE J. on Select. Areas in Comm. 10(5), 819–829 (1992)

[47] Yang, W., Benbouchta, M., Yantorno, R.: Performance of the modified Bark spectral distortion as an objective speech quality measure. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, pp. 541–544 (1998)

[48] Karjalainen, M.: Sound quality measurements of audio systems based on models of auditory perception. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol. 9, pp. 132–135 (1984)

[49] Chen, G., Parsa, V.: Loudness pattern-based speech quality evaluation using Bayesian modelling and Markov chain Monte Carlo methods. J. Acoust., Soc. Am. 121(2), 77–83 (2007)

[50] Pourmand, N., Suelzle, D., Parsa, V., Hu, Y., Loizou, P.: On the use of Bayesian modeling for predicting noise reduction performance. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 3873–3876 (2009)

[51] Moore, B.: An introduction to the psychology of hearing, 5th edn. Academic Press, London (2003)

[52] Fletcher, H., Munson, W.: Loudness, its definition, measurement and calculation. J. Acoust. Soc. Am. 5, 82–108 (1933)

[53] Robinson, D., Dadson, R.: A re-determination of the equal-loudness relations for pure tones. Brit. J. Appl. Phys. 7, 166–181 (1956)

[54] Yang, W.: Enhanced modified Bark spectral distortion (EMBSD): An objective speech quality measure based on audible distortion and cognition model. Ph.D., Temple University (1999)

[55] Novorita, B.: Incorporation of temporal masking effects into bark spectral distortion measure. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, vol. 2, pp. 665–668 (1999)

[56] Yang, W., Yantorno, R.: Improvement of MBSD by scaling noise masking threshold and correlation analysis with MOS difference instead of MOS. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol. 2, pp. 673–676 (1999)

[57] Rix, A., Hollier, M.: The perceptual analysis measurement for robust end-to-end speech quality assessment. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol. 3, pp. 1515–1518 (2000)

[58] Bartels, R., Stewart, G.: Solution of the matrix equation AX+XB=C. Comm. of ACM 15(9), 820–826 (1972)

[59] Beerends, J., Stemerdink, J.: A perceptual speech-quality measure based on a psychoacoustic sound representation. J. Audio Eng. Soc. 42(3), 115–123 (1994)

[60] Friedman, J.: Multivariate adaptive regression splines. Annals Statistics 19(1), 1–67 (1991)

[61] Falk, T.H., Chan, W.: Single-Ended Speech Quality Measurement Using Machine Learning Methods. IEEE Trans. Audio Speech Lang. Processing 14(6), 1935–1947 (2006)

[62] Rix, A.: Perceptual speech quality assessment - A review. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol. 3, pp. 1056–1059 (2004)

[63] Gray, P., Hollier, M., Massara, R.: Non-intrusive speech quality as-sessment using vocal-tract models. In: IEE Proc. - Vision, Image and Signal Processing, vol. 147(6), pp. 493–501 (2000)

[64] Chen, G., Parsa, V.: Nonintrusive speech quality evaluation using an adaptive neurofuzzy inference system. IEEE Signal Processing Letters 12(5), 403–406 (2005)

[65] Jin, C., Kubichek, R.: Vector quantization techniques for output-based objective speech quality. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 491–494 (1996)

[66] Picovici, D., Madhi, A.: Output-based objective speech quality measure using self-organizing map. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 476–479 (2003)

[67] Kim, D., Tarraf, A.: Perceptual model for nonintrusive speech quality assessment. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol. 3, pp. 1060–1063 (2004)

[68] ITU, Single ended method for objective speech quality assessment in narrow-band telephony applications. ITU-T Recommendation p. 563 (2004)

[69] Hollier, M., Hawksford, M., Guard, D.: Error activity and error en-tropy as a measure of psychoacoustic significance in the perceptual domain. In: IEE Proc. - Vision, Image and Signal Processing, vol. 141(3), pp. 203–208 (1994)

[70] Arehart, K., Kates, J., Anderson, M., Harvey, L.: Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 122, 1150–1164 (2007)

[71] Kates, J.: On using coherence to measure distortion in hearing aids. J. Acoust. Soc. Am. 91, 2236–2244 (1992)

[72] Hu, Y., Loizou, P.: Subjective comparison and evaluation of speech enhancement algorithms. Speech Communication 49, 588–601 (2007)

[73] Holub, J., Jianjun, L.: Intrusive Speech Transmission Quality Measurement in Chinese Environment. In: Intern. Conf. on Information, Communications and Signal Processing, pp. 1–3 (2007)

[74] Ma, J., Hu, Y., Loizou, P.: Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 125(5), 3387–3405 (2009)

[75] Zwicker, E., Fastl, H.: Pschoacoustics: Facts and Models, 2nd edn. Springer, Heidelberg (1999)

[76] Kang, J.: Comparison of speech intelligibility between English and Chinese. J. Acoust. Soc. Am. 103(2), 1213–1216 (1998)

[77] Bladon, R., Lindblom, B.: Modeling the judgment of vowel quality differences. J. Acoust. Soc. Am. 69(5), 1414–1422 (1981)

[78] Kent, R., Read, C.: The Acoustic Analysis of Speech. Singular Publishing Group, San Diego (1992)

[79] Stevens, K., Blumstein, S.: Invariant cues for the place of articulation in stop consonants. J. Acoust. Soc. Am. 64, 1358–1368 (1978)

[80] Breitkopf, P., Barnwell, T.: Segmental preclassification for improved objective speech quality measures. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 1101–1104 (1981)

[81] Kubichek, R., Quincy, E., Kiser, K.: Speech quality asessment using expert pattern recognition techniques. In: IEEE Pacific Rim Conf. on Comm. Computers, Sign. Proc., pp. 208–211 (1989)

[82] Barnwell, T.: A comparison of parametrically different objective speech quality measures using correlation analysis with subjective listening tests. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 710–713 (1980)

[83] Kates, J., Arehart, K.: Coherence and the speech intelligibility index. J. Acoust. Soc. Am. 117, 2224–2237 (2005)

[84] Mattila, V.: Objective measures for the characterization of the basic functioning of noise suppression algorithms. In: Proc. of online workshop on Measurement Speech and Audio Quality in Networks (2003)

[85] Mester, R., Franke, U.: Spectral entropy-activity classification in adaptive transform coding. IEEE J. Sel. Areas Comm. 10(5), 913–917 (1992)

[86] Voran, S.: Objective estimation of perceived speech quality - Parti I: Development of the measuring normalizing block technique. IEEE Transactions on Speech and Audio Processing 7(4), 371–382 (1999)

[87] Klatt, D.H.: Prediction of perceived phonetic distance from critical-band spectra:A first step. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 1278–1281 (1982)

Document-Aware Graph Models for Query-Oriented Multi-document Summarization

Furu Wei[1,2], Wenjie Li[1], and Yanxiang He[3]

[1] Department of Computing,
The Hong Kong Polytechnic University, Hong Kong,
{csfwei,cswjli}@comp.polyu.edu.hk
[2] IBM China Research Laboratory, Beijing, China
weifuru@cn.ibm.com
[3] Department of Computer Science and Technology,
Wuhan University, Wuhan, China,
yxhe@whu.edu.cn

Abstract. Sentence ranking is the issue of most concern in document summarization. In recent years, graph-based summarization models and sentence ranking algorithms have drawn considerable attention from the extractive summarization community due to their capability of recursively calculating sentence significance from the entire text graph that links sentences together rather than relying on single sentence alone. However, when dealing with multi-document summarization, existing sentence ranking algorithms often assemble a set of documents into one large file. The document dimension is ignored. In this work, we develop two alternative models to integrate the document dimension into existing sentence ranking algorithms. They are the one-layer (i.e. sentence layer) document-sensitive model and the two-layer (i.e. document and sentence layers) mutual reinforcement model. While the former implicitly incorporates the document's influence in sentence ranking, the latter explicitly formulates the mutual reinforcement among sentence and document during ranking. The effectiveness of the proposed models and algorithms are examined on the DUC query-oriented multi-document summarization data sets.

Keywords: Query-oriented multi-document summarization, document-sensitive sentence ranking, mutual-reinforcement sentence ranking.

1 Introduction

The explosion of the WWW has brought with it a vast board of information. It has become virtually impossible for anyone to read and understand large numbers of individual documents that are abundantly available. Automatic document summarization [17] [7] provides an effective means to manage such an exponentially increased collection of information and to support information seeking and condensing goals. The main evaluation forum providing benchmarks for researchers who work on document summarization to exchange their ideas and

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 655–678.
springerlink.com

experiences is the Document Understanding Conferences (DUC [3] [7]). The goals of the DUC are to enable researchers to participate in large-scale experiments upon the standard benchmark and to increase the availability of appropriate evaluation techniques. Over the past years, the DUC evaluations have gradually evolved from single-document summarization to multi-document summarization and from generic summarization to query-oriented summarization [20].

Up to the present, the dominant approaches in document summarization regardless of the nature and the goals of the tasks have still been built upon the sentence extraction framework. Under this framework, sentence ranking is the issue of most concern. It computes sentence significance based on certain criteria and ranks the sentences according to significance. Most previous work in the literature addresses the ranking issue by examining the features of each individual sentence, such as its content, its grammatical structure and etc. Recently, the relations among the sentences have been emphasized in the graph-based models that represent a document or a set of documents as a text graph. The text graph is normally constructed by taking a sentence as a node, and the similarity between the two sentences as edges. The significance of a node in the graph is then estimated by the graph-based ranking algorithms that normally take into account the global information recursively computed from the entire graph rather than merely relying on the local information within a single sentence. So far, the most popular graph-based ranking algorithms applied in document summarization are Google's PageRank [2] and its variations. LexRank [5] developed for generic summarization is one example of the PageRank-like algorithms. LexRank has also been extended to its topic-sensitive version [18] to accommodate the new challenge of query-oriented summarization, which is initiated by the DUC in 2005.

In general, existing PageRank-like algorithms can well model the phenomena that a sentence is important if it is linked by other important sentences. Or to say, they are capable of modeling the reinforcement among the sentences in a text graph. However, when dealing with multi-document summarization, these algorithms often assemble a set of documents into one large file and ignore the intrinsic difference between single document summarization and multi-document summarization. Or to say, the information carried by the document dimension is totally ignored in sentence ranking. We argue that since text is always organized and structured in a certain way to deliver information, sentence and document are not independent. A document carries global and contextual information necessary for understanding the sentences in that document. Consequently, the document dimension will (and should) influence sentence ranking. How to effectively integrate document dimension in sentence ranking is our major concern in this work. We explore two alternative models to integrate the document dimension into existing sentence ranking algorithms. The first one is a document-sensitive graph model which is developed to emphasize the difference among documents and the influence of global document set information on local sentence ranking. In this model, a document is implicit in a one-layer graph that links sentences together. The second one is a mutual reinforcement model, which is a two-layer graph with two layers corresponding to the documents and the sentences, respectively. In this model, a document is explicit in the graph. Sentence ranking and document ranking are simultaneous. Later we will show that sentence ranking indeed benefits from document ranking. We evaluate the effectiveness of the proposed models and algorithms in the context of DUC query-oriented multi-document summarization task, which aims to produce a short and concise summary for a set of relevant documents according to a given query that describes a user's information need. Significant results are achieved.

The remainder of this paper is organized as follows. Section 2 reviews existing graph-based ranking approaches applied in document summarization. Next, Section 3 and Section 4 introduce document-sensitive graph model and document-sentence mutual reinforcement graph model, respectively, as well as their applications in query-oriented multi-document summarization. Then, Section 5 presents evaluations and discussions. Finally, Section 6 concludes the paper.

2 Related Work

Sentence ranking is the issue of most concern under the framework of extractive summarization. Traditional feature-based approaches evaluated sentence significance and ranked the sentences relying on the features that were elaborately designed to characterize the different aspects of the sentences. A variety of statistical and linguistic features, such as term frequency (distribution), sentence dependency structure, sentence position, query relevance and etc., have been extensively investigated in the past due to their easy implementation and the ability to achieve promising ROUGE (i.e. DUC automatic) evaluation results. Among them, centroid introduced by [22] and signature term introduced by [11] are most remarkable. As for query-oriented summarization, the query term feature has been proved to be extremely useful. Besides, certain kinds of clustering techniques were also involved to expand the topic title and query keywords. As a matter of fact, feature-based approaches have been most widely used in the top five participating systems in DUC 2005-2007. Please refer to the online DUC reports for more details [4].

The features were often linearly combined and the weights of them were either experimentally tuned or automatically derived by applying certain learning-based mechanism [19] [28]. Learning-based approaches were popular in recent DUC competitions, such as the discriminative training model used to learn weights for a variety of sentence level features, the Support Vector Regression (SVR) model used for automatic feature weight selection and the log-linear model by maximizing metrics of sentence goodness, etc. [4]. Learning-based systems have achieved encouraging results in DUC 2007.

Newly emerged graph-based approaches like LexRank [5] and TextRank [14] [15] modeled a document or a set of documents as a weighed text graph. Different from feature-based approaches, graph-based approaches took into account the global information and recursively calculated sentence significance from the entire text graph rather than only relying on unconnected individual sentences. These approaches were actually inspired by PageRank [2], which has been successfully used for ranking web pages in the Web graph. The effectiveness of the PageRank-like approaches came from the advantage of making use of the link structure information. It further promoted the use of topic-sensitive PageRank [6], an extension of PageRank, in query-oriented summarization [18] [24] [13].

While the PageRank-like approaches normally considered the similarity or the association of the sentences, Zha [30], in contrast, proposed a mutual reinforcement principle that is capable of extracting significant sentences and key phrases at the same time. In his work, a weighted bipartite document graph was built by linking together the sentences in a document and the terms appearing in those sentences. Zha argued that a term should have a high salience score if it appears in many sentences with high salience scores while a sentence should have a high salience score if it contains many terms with high salience scores. This mutual reinforcement principle was reduced to a solution for the singular vectors of the transition matrix of the bipartite graph. In fact, as early in 1998, the similar idea has been used in HITS algorithm [16] to identify hub and authority web pages in a small subset of the web graph. Zha's work was later advanced by Wan et al [25] who additionally calculated the links among the sentences and the links among the terms. Zha's and Wan's works are the ones most relevant to our studies presented in this paper. But they all concentrated on single-document generic summarization. Later, we [27] integrated the notion of mutual reinforcement into PageRank-like algorithms. They introduce a unified mutual reinforcement chain, where reinforcement among terms, sentences and documents are considered simultaneously.

The use of the PageRank family was also very popular in event-based summarization approaches [9] [23] [29] [10]. In contrast to conventional sentence-based approaches, event-based approaches took event terms, such as verbs and action nouns and their associated named entities as graph nodes, and connected nodes according to their co-occurrence information or semantic dependency relations. They were able to provide finer text representation and thus could be in favor of sentence compression which was targeted to include more informative contents in a fixed-length summary. Nevertheless, these advantages largely lied on appropriately defining and selecting event terms.

3 Document-Sensitive Graph-Based Model

3.1 Introduction

It is worth noting that the sentence edges can be naturally differentiated into the edges linking the inter-document sentences (i.e. the sentences within the same documents) or the edges linking the intra-document sentences (i.e. the sentences in the different documents) by considering document in a conventional text graph that links sentences together. Such a distinction is meaningful. When evaluating sentence significance for multi-document summarization, as already observed, the sentences from many other documents can contain more useful and globally informative information than the other sentences within the same document, and therefore the recommendations of them are supposed to be more important. This is determined by the nature of multi-document summarization which requires the information included in the summary to be globally important on the whole document set. We further argue that the recommendations of the sentences from the documents on very similar topics are more reliable than the recommendations of the sentences from the documents telling of quite different stories. In addition, the document/sentence inclusion relation also allows for the impact from documents on sentences to be integrated in the ranking algorithms by imposing the document global information to the local sentence evaluation. Unfortunately, the above-mentioned document-level effects are neglected by almost all the previous graph-based ranking algorithms in the summarization literature. These observations motivate us to study how to make better use of the information provided in the whole text graph for the task of query-oriented multi-document summarization.

In our document-sensitive graph model, a set of document *D* is represented as a text similarity graph $G=\left(V,C,E^V,E^C,\alpha,\beta,\phi,\varphi\right)$, where *V* and *C* represent the sentence vertex set and the document vertex set, respectively. $E^V \subseteq V\times V$ and $E^C \subseteq C\times C$ are the sentence edge set and the document edge set. $\alpha: V \to \Re_+^*$, $\beta: C \to \Re_+^*$ are two functions defined to label the sentence vertices and document vertices, while $\phi: E^V \to \Re_+^*$ and $\varphi: E^C \to \Re_+^*$ are two functions defined to label sentence edges and the document edges. When we add the document information into the conventional graph model, the document-level relations that have been ignored in the past become visible and their influence on the sentence evaluation can then be used to enhance the existing PageRank-like algorithms.

In the previous work [5], the document set *D*={ $d_1, d_2, ...d_N$ } (*N* is the total number of the documents to be summarized) is represented as a simple weighted undirected text graph G by taking sentences in D as vertices and adding a edge to connect the two vertices if the two sentences concerned are similar enough. There is only one kind of objects (i.e. sentences) and one kind of object relations (i.e. sentence similarity relations) in this graph. Sentences from the same or the different documents are treated equally. When the concept of document is emphasized, one more kind of objects (i.e. documents) is added into the graph. One can then easily obtain the following three important but previously ignored information: (1) the inclusion relation between sentences and the document they belong to; (2) the similarity relation among documents; and (3) what's more, the sentence-sentence similarity relations are divided into two categories, i.e. the one within the document (called intra-document relation) and the one cross over two documents (called inter-document relation). As for Query-oriented multi-document summarization, an additional kind of object (i.e. queries) is involved. Alike, the document-query relevance and the sentence-query relevance can be formulated separately so that the impact of the document-query relevance on the sentence-query relevance can be taken into account.

3.2 Existing PageRank-Like Algorithms in Document Summarization

PageRank [2] has been adapted to rank the undirected graphs in the community of document summarization. For instance, Erkan and Radev [5] proposed the

LexRank algorithm for generic summarization. Let R denote the ranking vector of N sentences in the similarity graph G, M denote the normalized affinity matrix of G and $\vec{p}$ denote the preference probability vector where each element is positive and the sum of all the elements equals to 1. The PageRank ranking scheme is defined as:

$$R = d \cdot M \bullet R + (1-d) \cdot \vec{p} \tag{1}$$

where d is the damping factor between 0 and 1. There are many variations of the PageRank algorithm that follow the same ranking scheme presented in Equation (1). The difference among those algorithms lies in their different use of M and $\vec{p}$.

PageRank can be determined by the stationary solution of the Markov Chain with P as the transition matrix, or found by the eigenvector problem.

$$P \bullet R = \lambda \cdot R \text{, and } P = d \cdot M + (1-d) \cdot \vec{p} \bullet 1^T \tag{2}$$

where **1** denotes a $N \times 1$ vector of all 1's. It is obvious that P is both columns stochastic and irreducible. Meanwhile, P is primitive because all the elements in P are positive. As a result, based on Perron's Theorem [8], the dominant eigenvector of P is unique with 1 as the eigenvalue. Moreover, the power iteration method applied to P in Equation (2) will converge to its dominant eigenvector. We can use the power method to compute the PageRank vector.

Following the spirit of topic-sensitive PageRank introduced in Haveliwala [6], Otterbacher et al. [18] proposed the query-sensitive version of LexRank (i.e. Q-LexRank), which is then followed by Wan et al. [25]. Let $\vec{p}_q$ denote a $N \times 1$ vector such that $\vec{p}_q(i) = \alpha(s_i)$, we summarize the existing PageRank-like algorithms in document summarization as follows. The difference between the LexRank and its query-sensitive version (we call it Q-LexRank) is that they used different preference vector. The algorithms used in [25] were analogous to Q-LexRank and LexRank. Wan et al. used the inter-document links and the intra-document links to construct two separated graph independently, and then combined the calculated PageRank values by a simple linear combination function.

3.3 Graph-Based Document-Sensitive Ranking Algorithm (DsR)

The idea of the document-sensitive ranking algorithm is inspired by the work of [31], where a weighted inter-cluster edge ranking (WICER) was proposed for the clustered graphs. The major contributions of their work are to weight the edges according to whether they are the inter-cluster or the intra-cluster edges and to weight the vertices based on the number of clusters they connect. WICER computes the graph with its internal relations and structures but does not concern how the external factors cause the change of graph computing. Also, there is no mathematical analysis on WICER and the algorithm is not guaranteed to converge.

We borrow the spirit of it and develop a new ranking algorithm in our summarization model. We also prove the convergence of the solution.

We emphasize the document dimension in the PageRank-like algorithm in the following two ways. One is on the sentence affinity matrix and the other is on the preference vector.

To reflect the impact of the document dimension on the sentence affinity matrix, different sentence edges are differentiated corresponding to the fact that the recommendations exist in two familiar communities are more credible. So the sentence edges are additionally weighted by the similarity between the two documents they connect. Let *N* denotes the number of the documents involved, then we compare two sentence affinity matrices,

$$M_o = \begin{bmatrix} M_{11} & \cdots & \cdots & \cdots & M_{1N} \\ M_{21} & M_{22} & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ M_{N1} & \cdots & \cdots & \cdots & M_{NN} \end{bmatrix}, \text{ and } \quad M = \begin{bmatrix} w_{11}\cdot M_{11} & \cdots & \cdots & \cdots & w_{1N}\cdot M_{1N} \\ w_{21}\cdot M_{21} & w_{22}\cdot M_{22} & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{N1}\cdot M_{N1} & \cdots & \cdots & \cdots & w_{NN}\cdot M_{NN} \end{bmatrix}$$

The block matrix M_{ii} denotes the affinity matrix of the sentences in document *i*, M_{ij} ($i \neq j$) denotes the cross-document (*i* and *j*) affinity matrix, and so on. Notice that M_o corresponds to the original sentence affinity matrix (i.e. the sentence similarity matrix) used in LexRank and Q-LexRank. The key to encode the document dimension into the affinity matrix is to emphasize the document influence on the sentence edges that connect different documents, as illustrated in *M*. The weight matrix $W_M = \begin{bmatrix} w_{11} & \cdots & \cdots & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{N1} & \cdots & \cdots & \cdots & w_{NN} \end{bmatrix}$ is used to discriminate the cross-document sentence edges. Typically, the diagonal elements in *W* are set to 1, which denotes the relative weight of the intra-document sentence edges. On the contrary, the non-diagonal elements are determined by the relations between the two corresponding documents. In our work, *W* is defined as $W(i,j) = 1 + \varphi(d(s_i), d(s_j))$, where $d(s_i)$ denotes the document that contains the sentence s_i.

To reflect the impact of the document dimension on the preference vector $\vec{p}$, we believe that a sentence from the document with higher significance should be ranked higher. This can be explained as that a recommendation from a reputable person should be more important. Accordingly, the centroid-based weight of the document in generic summarization, or the relevance of the document to the query in query-oriented summarization, is taken as the weight on the preference vector $\vec{p}$. See the two preference vectors,

$$\overrightarrow{p_o} = \begin{bmatrix} \overrightarrow{p_1} & \overrightarrow{p_2} & \cdots & \overrightarrow{p_N} \end{bmatrix}^T \text{, and } \vec{p} = \left(\begin{bmatrix} \overrightarrow{p_1} & \overrightarrow{p_2} & \cdots & \overrightarrow{p_N} \end{bmatrix} \bullet \begin{bmatrix} w_1 & 0 & \cdots & \cdots & 0 \\ 0 & w_2 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & w_N \end{bmatrix} \right)^T$$

where $\overrightarrow{p_i}$ denotes the sub-preference vector of the sentences from the document *i*.

The weight matrix $W_P = \begin{bmatrix} w_1 & 0 & \cdots & \cdots & 0 \\ 0 & w_2 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & w_N \end{bmatrix}$ is designed to differentiate the bias to sentences from different documents. This weight matrix is determined with respect to the documents. In this work, it is defined as $W_p(i) = 1 + \beta(d(s_i))$.

To guarantee the solution of our newly designed algorithm, we should first make $\vec{p}$ a preference probability vector.

Lemma 1. $\vec{p}$ is a preference probability vector, if *Wp* is positive and the diagonal elements in *Wp* sum to 1.

Proof: Since $\overrightarrow{p_i}$ is a probability vector, we have $\left|\overrightarrow{p_i}\right| = 1$. Then,

$$\begin{aligned} |\vec{p}| &= |w_1 \cdot \vec{p}_1 + w_1 \cdot \vec{p}_2 + \cdots + w_N \cdot \vec{p}_N| = w_1 \cdot |\vec{p}_1| + w_1 \cdot |\vec{p}_2| + \cdots + w_N \cdot |\vec{p}_N| \\ &= \sum_i W_P(i) = 1 \end{aligned}.$$

Second, we should make the matrix *M* column stochastic and irreducible. To make *M* column stochastic, we force each of the block matrices (i.e. M_{ij}) column stochastic. They are normalized by columns such that any column in these matrices sums to 1. There may be zero columns in these matrices. In these cases, we replace the zero columns, as in PageRank, with the preference vector $\vec{p}$.

Lemma 2. *M* is column stochastic, provided that the weight matrix *W* is column stochastic.

Proof: Let M^{1k}, M^{2k}, …, M^{Nk}($k \in [1, N]$) denote the block matrices of the k-th column *M*, then

$$\sum_i M_{ij} = w_{k1} \cdot \sum_i X_{ij}^{1k} + w_{k2} \cdot \sum_i X_{ij}^{2k} + \cdots + w_{kN} \cdot \sum_i X_{ij}^{Nk} = w_{k1} + w_{k2} + \cdots + w_{kN} = 1 .$$

To make *M* irreducible, we make the block matrices in *M* irreducible by adding additional links between any two sentences, which is also adopted in PageRank. Then we have,

Lemma 3. *M* is irreducible.

Proof: Since the graphs corresponding to the diagonal block matrices in *M* are strongly connected (i.e. they are irreducible) and the edges connecting the graphs are bidirectional, the graph corresponding to *M* is obviously strongly connected. Thus *M* must be also irreducible.

Finally, we obtain $P = d \cdot M + (1-d) \cdot \vec{p} \bullet 1^T$. Obviously, *P* is stochastic, irreducible and primitive. As a result, we can compute the unique dominant vector (with 1 as the eigenvalue) of *P*. It is well known that the power iteration method applied to *P* converges to *R*.

Until now, the document dimension has been integrated into the existing PageRank-like algorithms with a solid mathematical foundation. Let DsR denote the algorithm designed for query-oriented summarization, we summarize the labeling functions involved in the following table.

$\alpha(s_i)$	$rel(s_i \mid q) = \dfrac{sim(s_i, q)}{\sum_{s_k \in D} sim(s_k, q)}$
$\phi(s_i, s_j) = sim\_norm(s_i, s_j)$	$\dfrac{sim(s_i, s_j)}{\sum_{s_k \in D \cap k \neq i} sim(s_i, s_k)}$
$\beta(d_i)$	$rel(d_i \mid q) = \dfrac{sim(d_i, q)}{\sum_{d_k \in D} sim(d_k, q)}$
$\varphi(d_i, d_j) = sim\_norm(d_i, d_j)$	$\dfrac{sim(d_i, d_j)}{\sum_{d_k \in D \cap k \neq i} sim(d_i, d_k)}$

4 Mutual Reinforcement Graph-Based Model

4.1 Introduction

In many text processing applications, such as information retrieval, question answering and document summarization, the text people often manipulate and evaluate is of two different granularities. They are document and sentence. While document ranking is indispensable to information retrieval, sentence ranking is one of the most fundamental issues in document summarization. Comparatively speaking, sentence ranking is more challenging than document ranking since a sentence carries much less information than a document for measuring the similarity of text. However, the sentence does not stand alone in the text without the context.

It is an unarguable fact that the text is always organized and structured in a certain way so that the core information would be easily identified. The assumption that document and sentence are independent of each other in delivering meanings is untenable. Therefore, even when the sentence ranking result is the only concern in summarization, the mutual constraints and the influences between document and sentence could not be ignored. In this section, we propose a new sentence ranking algorithm based on the Mutual Reinforcement (MR) of Document (D) and Sentence (S). We define the reinforcement between document and sentence as the external reinforcement.

In addition to the external reinforcement, the proposed sentence ranking algorithm also supports the calculation of the internal reinforcement within a set of documents or a set of sentences, i.e. the document-level reinforcement and the sentence-level reinforcement. The existing PageRank-like algorithms employed in summarization can be viewed as sentence-level reinforcement instances. In the past, the importance of sentence relations have been stressed in graph-based summarization models and their contribution to the performance improvement has been recognized [5]. We put them forward to the relations at both the document level and the sentence level and move towards a more unified reinforcement model. To sum up, the external and the internal reinforcement together form a complete two-level document and sentence mutual reinforcement (D-S MR or MR for short) framework.

The mutual reinforcement framework is developed with an attempt to capture the following intuitions: 1. A document is important if (1) it correlates to important sentences; (2) it associates to other important documents; 2. A sentence is important if (1) it correlates to importance documents; (2) it associates to other important sentences. Then, the ranking of documents and sentences can be iteratively derived from the D-S MR. Let R_D and R_S denote the ranking scores of the document set D and the sentence set S, respectively, the iterative ranking can be formulated as follows:

$$\begin{cases} R_D^{(k+1)} = \alpha_1 \cdot D_D \cdot R_D^{(k)} + \beta_1 \cdot D_S \cdot R_S^{(k)} \\ R_S^{(k+1)} = \beta_2 \cdot S_D \cdot R_D^{(k)} + \alpha_2 \cdot S_S \cdot R_S^{(k)} \end{cases} \tag{4}$$

where D_D denotes the D-D affinity matrix, D_S denotes the D-S affinity matrix, and so on. The calculation of the four affinity matrices in Equation (1) will be detailed later in Section 4. $W = \begin{bmatrix} \alpha_1 & \beta_1 \\ \beta_2 & \alpha_2 \end{bmatrix}$ is the weight matrix used to balance the relative weights of document and sentence in D-S MR. The coefficients in Equation (1) corresponds to a block matrix.

$$M = \begin{bmatrix} \alpha_1 D_D & \beta_1 D_S \\ \beta_2 S_D & \alpha_2 S_S \end{bmatrix} \tag{4'}$$

Let $R = \begin{bmatrix} R_D \\ R_S \end{bmatrix}$, then R can be computed as the dominant eigenvector of M, i.e.

$$M \cdot R = \lambda \cdot R \tag{5}$$

Given that the corresponding graph of M is not bipartite, we must force M stochastic, irreducible and primitive[1] in order to guarantee a unique solution of R. On this account, the necessary matrix transformation explained below must be performed. We will prove to readers that the new transformed M is stochastic, irreducible, and more strictly, primitive for certain.

A sufficient condition for a stochastic M is to make the four affinity block matrices in M column stochastic. For the sake of simplicity, we let X be either of the two diagonal block matrices (i.e. D_D and S_S) and Y be either of the remaining two block matrices (i.e. S_D and D_S).

We first delete the rows and the columns that do not contain any non-zero element in X. This manipulation is analogous to the strategy used in PageRank to cope with the dangling pages in the Web graph that do not have outgoing links. Since X is symmetric, if the out-degree of a document or a sentence node is zero, its in-degree must be zero as well. Such a node is actually an isolated node in a text graph. Therefore, the ranking results will not be influenced when the isolated nodes are removed. On the other hand, it is noted that there are no zero columns in Y. Let us take S_D for example. The affinity of the sentence s and the document d is at least greater than zero if d contains s. Now, we are ready to normalize both X and Y by columns to their column stochastic versions $\overline{X}$ and $\overline{Y}$. We replace X and Y with $\overline{X}$ and $\overline{Y}$ in M, and denote the new matrix as $\overline{M}$.

Next, we manage to make $\overline{M}$ irreducible. Let $\overline{X}$ denote either of the two new diagonal block matrices in $\overline{M}$. Similar to the treatment used in PageRank calculation, we make the graph corresponding to $\overline{X}$ strongly connected by adding (artificial) links for every pair of nodes with a probability vector $\vec{p}$. After such an adjustment, the revised $\overline{X}$ becomes

$$\overline{\overline{X}} = d \cdot \overline{X} + (1-d)E \text{ and } E = \vec{p} \times [1]_{1\times k} \tag{6}$$

where $0 < d < 1$, d is usually set to 0.85 according to PageRank. k is the order of $\overline{X}$. The probability vector $\vec{p}$ can be defined in many different ways. A typical definition is to assume a uniform distribution over all elements, i.e. $\vec{p} = [1/k]_{k\times 1}$. By

[1] A matrix is irreducible if its graph shows that every node is reachable from every other node. A non-negative, irreducible matrix is primitive if it has one eigenvalue on its spectral circle. An irreducible Markov chain with a primitive transition matrix is called an aperiodic chain. Please refer to [8] for more details.

doing so, $\overline{\overline{X}}$ becomes both stochastic and irreducible. We finally replace $\overline{X}$ with $\overline{\overline{X}}$ in $\overline{M}$, and let $\overline{\overline{M}}$ denote the latest matrix.

After the above-mentioned transformations on the matrices, we now can prove that the final $\overline{\overline{M}}$ is column stochastic, irreducible and primitive. For the sake of simplicity, we re-write $\overline{\overline{M}}$ as $\overline{P} = \left[\begin{array}{c|c} \alpha_1 \cdot P_{11} & \beta_1 \cdot P_{12} \\ \hline \beta_2 \cdot P_{21} & \alpha_2 \cdot P_{22} \end{array}\right]$ and let $P = \left[\begin{array}{c|c} P_{11} & P_{12} \\ \hline P_{21} & P_{22} \end{array}\right]$ and $W = \begin{bmatrix} \alpha_1 & \beta_1 \\ \beta_2 & \alpha_2 \end{bmatrix}$. From the previous analysis, we have,

(1) $P_{11}(m \times m) > 0$, $P_{22}(m \times m) > 0$, $P_{12}(m \times n) \geq 0$, $P_{21}(n \times m) \geq 0$;
(2) P_{11}, P_{12}, P_{21} and P_{22} are column stochastic;
(3) $\forall i \in [1,n]$ and $\exists j \in [1,m]$ such that $P_{12}(j,i) > 0$[2];
(4) P_{12} and P_{21} satisfy $P_{12}(i,j) > 0 \Leftrightarrow P_{21}(j,i) > 0$ and $P_{12}(i,j) = 0 \Leftrightarrow P_{21}(j,i) = 0$; and
(5) It is easy to ensure $W > 0$ and make W column stochastic.

Lemma 1. $\overline{P}$ is also column stochastic if the weight matrix W is column stochastic.

Proof: Let A and B denote the two block matrices in any column of $\overline{P}$ under concern, α and β the corresponding weight coefficient with respect to A and B, then $\sum_i \overline{P}_{ij} = \alpha \sum_i A_{ij} + \beta \sum_i B_{ij} = \alpha + \beta = 1$. □

Lemma 2. $\overline{P}$ is irreducible.

Proof: Since the two graphs corresponding to the two diagonal block matrices in $\overline{P}$ are strongly connected (i.e. they are irreducible) and the links connecting the two graphs are bidirectional, obviously the graph corresponding to $\overline{P}$ is also strongly connected. Thus, $\overline{P}$ must be irreducible. □

Now the matrix $\overline{P}$ is both stochastic and irreducible. More strictly, we have

Lemma 3. $\overline{P}$ is primitive.

Proof: Considering $\overline{P}^2 = \left[\begin{array}{c|c} \alpha_1 \cdot P_{11} & \beta_1 \cdot P_{12} \\ \hline \beta_2 \cdot P_{21} & \alpha_2 \cdot P_{22} \end{array}\right] \bullet \left[\begin{array}{c|c} \alpha_1 \cdot P_{11} & \beta_1 \cdot P_{12} \\ \hline \beta_2 \cdot P_{21} & \alpha_2 \cdot P_{22} \end{array}\right]$

$$= \left[\begin{array}{c|c} \alpha_1^2 \cdot P_{11}^2 + \beta_1 \beta_2 \cdot P_{12} \bullet P_{21} & \alpha_1 \beta_1 \cdot P_{11} \bullet P_{12} + \alpha_2 \beta_1 \cdot P_{12} \bullet P_{22} \\ \hline \alpha_1 \alpha_2 \cdot P_{21} \bullet P_{11} + \beta_1 \beta_2 \cdot P_{22} \bullet P_{21} & \alpha_2 \beta_1 \cdot P_{21} \bullet P_{12} + \alpha_2^2 \cdot P_{22}^2 \end{array}\right]$$

[2] For each sentence, there exists at least one document that contain that sentence such that the element in the affinity matrix is a positive value because the affinity between them is positive, and vice versa.

we have

(1) $\alpha_1^2 \cdot P_{11}^2 + \beta_1\beta_2 \cdot P_{11} \bullet P_{21} > 0$ $\quad (P_{11}^2 > 0)$

(2) $\alpha_2\beta_1 \cdot P_{21} \bullet P_{12} + \alpha_2^2 \cdot P_{22}^2 > 0$ $\quad (P_{22}^2 > 0)$

(3) $\alpha_1\beta_1 \cdot P_{11} \bullet P_{12} + \alpha_2\beta_1 \cdot P_{12} \bullet P_{22} > 0$ $\quad (P_{11} \bullet P_{12} > 0)$

(4) $\alpha_1\alpha_2 \cdot P_{21} \bullet P_{11} + \beta_1\beta_2 \cdot P_{22} \bullet P_{21} > 0$ $\quad (P_{22} \bullet P_{21} > 0)$

It is easy to deduce that $\overline{P}^2 > 0$ and $\overline{P}$ is primitive. □

The above proof can be understood from the perspective of graph. Let G_1 denote the graph corresponding to the matrix P_{11}, G_2 denote the graph corresponding to the matrix P_{22}, and G denote the graph corresponding to the matrix $\overline{P}$. P_{12} and P_{21} can be viewed as the links connecting the nodes between G_1 and G_2. Notice that any two nodes in G_1 or G_2 are connected and there is at least one link from the nodes in G_1 to G_2, and vice versa. The nodes in G have been divided into two sets, i.e. G_1 and G_2. There is no question that any two nodes in the same set (i.e. within G_1 or within G_2) are able to reach each other in exact two steps. We then consider the case that a node n_{1i} in G_1 links to a node n_{2j} in G_2. There exists at least one node, say n_{1k} for example, in G_1 such that there is a direct path from n_{1k}. to n_{2j}. Therefore, n_{1i} is also able to reach n_{2j} in exact two steps given that n_{1i} and n_{1k} are connected in G_1. This conclusion also holds for the case that a node in G_2 links to a node in G_1 because the paths are reversible in G. In conclusion, any two nodes in G are able to reach each other in exact two steps, which means the matrix $\overline{P}^2 > 0$.

As a result, we can compute the unique dominant eigenvector (with 1 as the eigenvalue) of $\overline{\overline{M}}$. It is well-known that the power method applied to $\overline{\overline{M}}$ will converge to R.

$$\overline{\overline{M}} \cdot R = \lambda \cdot R. \tag{7}$$

Eventually, we can develop an iterative algorithm to solve Equation (1).

4.2 Weight Matrix Design

A critical issue in implementing Equation (1) is to design the appropriate weight matrix W. Essentially a positive column stochastic matrix is expected. We design a symmetric weight matrix, i.e. $W = \begin{bmatrix} \alpha & \beta \\ \beta & \alpha \end{bmatrix}$. Although W is necessary to be column stochastic and positive in our previous analysis, we use the weight matrix before it is normalized to be column stochastic for the ease of illustration and explanation. Generally speaking, α (set to 1 as reference) indicates the weight of the internal reinforcement and β ($\leq \alpha$) the external reinforcement. The motivation of this design is straightforward. It is reasonable to assume that the internal reinforcement is more important than the external reinforcement. But it seems unnecessary to further distinguish the weights of different kinds of internal reinforcement (i.e. documents-to-document and sentence-to-sentence reinforcement) or external reinforcement (i.e. documents-to-sentence and sentence-to-document reinforcement).

By designing such a symmetric weight matrix, we come up with the following interesting and important significance conservation property of the ranking solution in Algorithm 1.

Proposition 1. The significance is re-distributed within the scope of document set or sentence set in each iteration. However, the sum of the scores in each set remains the same during the ranking iterations. In other words, sentences (or documents) compete with one another for a higher significance, but they will not jump across the set boundary to grab the significance from documents (or sentences). Formally, let $R^{(0)}=[R_1^{(0)} \quad R_2^{(0)}]$, we have $\left|R_1^{(n)}\right|=\left|R_2^{(n)}\right|=\gamma$ if $\left|R_1^{(0)}\right|=\left|R_2^{(0)}\right|=\gamma$. γ can be any arbitrary positive value. Later in the experiments, we set it to 1/2 so as to ensure $\left|R^{(0)}\right|=1$.

Proof. We complete the proof by mathematical induction.

(1) Given that, $\left|R_1^{(0)}\right|=\gamma$ and $\left|R_2^{(0)}\right|=\gamma$;

(2) Assume $\left|R_1^{(k)}\right|=\gamma$ and $\left|R_1^{(k)}\right|=\gamma$ when $n=k$, then,

$\left|R_1^{(k+1)}\right|=\left|\alpha \cdot P_{11} \cdot R_1^{(k)}+\beta \cdot P_{12} \cdot R_2^{(k)}\right|=(\alpha+\beta)\cdot\gamma=\gamma$, and

$\left|R_2^{(k+1)}\right|=\left|\beta \cdot P_{21} \cdot R_1^{(k)}+\alpha \cdot P_{22} \cdot R_2^{(k)}\right|=(\alpha+\beta)\cdot\gamma=\gamma$.

It means that the conservation of total significance in a document set or a sentence set also holds at $n=k+1$ if it holds at $n=k$. Therefore, Proposition 1 is true. □

This proposition is meaningful in the context. Given the initial individual significance of a sentence (or a document) and the accumulated total significance of all the sentences (or the documents), the ranking Algorithm 1 can be viewed as iteratively re-distributing the total significance among the sentences (or the documents) by the mutual reinforcement of document and sentence (including both external and internal) according to the link structure (i.e. the affinity graph) of them. We believe that the documents influence the ranking of sentences, and vice versa. In other words, the external reinforcement from the documents provides useful hints to guide the sentence internal rank competition, and the other way around. However, this does not mean that the total significance of the sentences (or the documents) would change during ranking iterations. The re-distribution of the total significance should not cross over the set boundary of the sentences (or the documents). In short, document and sentence are interactive during ranking iterations but they still have certain independence. It is meaningless for the text with different granularities to compete with each other for a higher significance. Significance of a document and a sentence are not comparable.

Another advantage of using the symmetric weight matrix is that we only need to tune and fix one parameter when we design the weight matrix. In this context, we only need to determine the proportion of the internal-reinforcement and the external-reinforcement weight. We will discuss the parameter issues later in Section 5.2.3.

4.3 Query-Sensitive D-S MR (Qs-MR)

In the previously introduced D-S MR framework, the reinforcement of document and sentence is query-unaware. That is only the content of the text is concerned. However, for the tasks like query-oriented summarization, how the reinforcement is biased to an external context (such as a user's query) is often of great interest.

A general way to incorporate the query information into the general D-S MR framework is to impose the influence of a user's query on each text unit (document or sentence) such that it works in the internal reinforcement. This somewhat can be viewed as a topic-sensitive PageRank [6] at each level of text granularity. The key to make ranking biased towards the query rests with the definition of the query-sensitive probability vector $\vec{p}$. A simple yet effective solution is to define $\vec{p}$ as

$$\vec{p}_i = \begin{cases} rel(t_i \mid q) & \text{if } rel(t_i \mid q) \neq 0 \\ \theta & \text{otherwise} \end{cases} \tag{8}$$

where t_i can be either a document or a sentence, $rel(t_i \mid q)$ denotes the relevance of t_i to q and can be calculated by cosine similarity, which is widely used in information retrieval [1]. θ is an extremely small real number to avoid zero elements in $\vec{p}$. $\vec{p}$ is further normalized to 1 in order for it to be a probability vector.

Existing query-oriented summarization approaches basically follow the same processes: (1) first calculate the significance of the sentences with reference to the given query from different perspectives with/without using some sorts of sentence relations; (2) then rank the sentences according to certain criteria and measures; (3) finally extract the top-ranked but non-redundant sentences from the original documents to produce a summary. Under this extractive framework, undoubtedly the two critical processes involved are sentence ranking and sentence selection. We summarize the sentence ranking algorithm in Algorithm 2 and present the sentence selection strategy in Section 4.5.

```
Algorithm 2: RankSentence(D, S, q)
Input: The document set D, the sentence set S, and the query q.
Output: The ranking vectors of $R_D$ and $R_S$.
1: Construct the affinity matrices $D_D$, $D_S$, $S_D$ and $S_S$;
2: Transform the four block matrices as mentioned in Section 4.1;
3: Design the symmetric weight matrix W;
3: Choose (randomly) the initial non-negative vectors $R_D^{(0)}$ and $R_S^{(0)}$, such that $|R_D^{(0)}|$=1/2 and $|R_S^{(0)}|$=1/2;
4: Return Rank($D_D$, $D_S$, $S_D$, $S_S$, W, $R^{(0)}$).
```

4.4 Sentence Selection by Removing Redundancy

In multi-document summarization, the number of the documents to be summarized can be very large. This makes information redundancy problem appear to be more serious in multi-document summarization than in single-document summarization. Redundancy removal becomes an inevitable process. Since our focus in this study is the design of effective (sentence) ranking algorithm, we apply the following straightforward yet effective sentence selection principle. We incrementally add into the summary the highest ranked sentence of concern if it doesn't significantly repeat the information already included in the summary until the word limitation of the summary is reached[3].

5 Experiments

5.1 Experiment Set-Up

We conduct the experiments on the DUC 2005 and DUC 2006 data sets. Table 1 shows the basic statistics of the data sets. Each set of documents is accompanied with a query description representing a user's information need. The query usually consists of one or more interrogative and/or narrative sentences. Here is a query example from the DUC 2005 document set "d331f".

```
<topic>
<num> d331f </num>
<title> World Bank criticism and response </title>
<narrative>
Who has criticized the World Bank and what criticisms have they made of
World Bank policies, activities or personnel. What has the Bank done to
respond to the criticisms?
</narrative>
<granularity> specific </granularity>
</topic>
```

According to the task definitions, system-generated summaries are strictly limited to 250 words in length.

Table 1. Basic Statistics of the DUC Data Sets

	Total Number of Document Sets	Average Number of Documents per Set	Average Number of Sentences per Set
DUC 2005	50	31.86	1002.54
DUC 2006	50	25	815.22

[3] A sentence is discarded if the cosine similarity of it to any sentence already selected into the summary is greater than 0.9.

As for the evaluation metric, it is difficult to come up with a universally accepted method to measure the quality of machine-generated summaries. In fact, summary evaluation methods themselves are still an ongoing research in the summarization community. Many literatures have addressed different methods for automatic evaluations other than human judges. Among them, ROUGE [12] is supposed to produce the most reliable scores in correspondence with human evaluations. More important, it offers the advantage of being readily applied to compare the performance of different approaches on the same data set. Given the fact that judgments by humans are time-consuming and labor-intensive and ROUGE has been officially adopted by the DUC for automatic evaluations since 2005, like the other researchers, we also use it as the evaluation criteria in this paper.

Documents and queries are pre-processed by segmenting sentences and splitting words. Stop-words are then removed[4] and the remaining words are stemmed with Porter Stemmer [21]. In all the following experiments, both text units (i.e. documents or sentences) and queries are represented as the vectors of terms. Notice that the term weights are normally measured in summarization models by the TF*IDF scheme as in conventional vector space models (VSM). However, we argue that it would be more reasonable to use the sentence-level inverse sentence frequency (ISF) instead of the document-level IDF when dealing with a sentence-level text processing application. This has been verified in our early study [26]. We define $isf_w = \log(N/sf_w)$ where N is the total number of the sentences in the document set, and sf_w is the number of the sentences where the word w appears. Then, the weight of w is computed as $tf_w \cdot isf_w$. θ in Equation (6) is assigned to 20% of the minimum value of the relevance of the documents (or the sentences) to the query in a document set.

5.2 Evaluation on One-Layer Graph-based Model

In the proposed document sensitive graph model, two new components are introduced. They are the relevance of documents (denoted by A in Tables 2), the different weigh treatment for the edges combining the different documents (denoted by B). The aim of this first set of experiments is to examine the individual or combined contributions of these two new components (i.e. A and B). Table 2 below shows the results of the average recall scores of ROUGE-1, ROUGE-2 and ROUGE-SU4, along with the 95% confidential intervals within the square brackets on the DUC 2005 data set. Let Q-LexRank denotes the query sensitive LexRank [18].

[4] A list of 199 words is used to filter stop words

Table 2. Model selection on DUC 2005 data set

	ROUGE-1	ROUGE-2	ROUGE-SU4
Q-LexRank	0.3702 [0.3672,0.3772]	0.0725 [0.0704,0.0766]	0.1306 [0.1274,0.1341]
A	0.3736 [0.3686,0.3784]	0.0756 [0.0726,0.0788]	0.1308 [0.1278,0.1337]
B	0.3751 [0.3695,0.3804]	0.0745 [0.0712,0.0777]	0.1308 [0.1277,0.1339]
A $\cup$ B / DsR	0.3785 [0.3731,0.3840]	0.0771 [0.0734,0.0808]	0.1337 [0.1303,0.1373]

As shown in the Table 2, Q-LexRank can already achieve considerable results on the DUC 2005 evaluation. However, it is encouraging to see that there are still improvements when the two new components are added. The best results are achieved when both of them are considered. DsR improves ROUGE performance over Q-LexRank noticeably. It is 2.24% increase in ROUGE-1, 6.34% increase in ROUGE-2 and 2.37% increase in ROUGE-SU4. These results demonstrate the effectiveness of our extensions to the document-sensitive graph model and the corresponding ranking algorithm.

The aim of the second set of experiments is to examine our proposed model and ranking algorithm on the different summarization tasks. We use the same configuration as the one in the above section. Tables 3 below show the results for the DUC 2006 data sets.

Table 3. Model evaluation on DUC 2006 data set

	ROUGE-1	ROUGE-2	ROUGE-SU4
Q-LexRank	0.3899 [0.3833,0.3964]	0.0856 [0.0813,0.0899]	0.1394 [0.1353,0.1438]
DsR	0.3955 [0.3897,0.4012]	0.0899 [0.0857,0.0943]	0.1427 [0.1391,0.1464]

Table 4. Summary of improvements by DSR on DUC data set

	ROUGE-1	ROUGE-2	ROUGE-SU4
DUC 2005	+2.24%	**+6.34%**	+2.37%
DUC 2006	+1.44%	**+5.02%**	+2.37%

5.3 Evaluation on Two-Layer Graph-based Model

As for the two-layer graph based model, the aim of the following experiments is to examine and fix the involved parameters on the DUC 2005 data set. There are three parameters in our algorithm. According to [32], we set the damping factor to 0.75 used in the internal reinforcement. Meanwhile, to avoid to link-by-chance phenomena, we only insert edges to the text graph when the similarity between the two nodes (i.e. sentences) is greater than 0.03. We focus on examining the weight matrix parameters α and β for balancing the internal and external reinforcement.

The aim of the following set of experiments is to examine the weight matrix. For the simplicity of illustration, we use the weight matrix before it is normalized to be stochastic for presentation in this section. In our implementation, the corresponding normalized version is utilized. Recall that the weight matrix W we design is symmetric, where parameters α and β reflect the relative importance of the internal reinforcement and the external reinforcement. We let α =1 and then tune the values of β. In these experiments, the damping factor d is set to 0.75 and similarity threshold is set to 0.03. Table 5 shows the results of the average recall scores of ROUGE-1, ROUGE-2 and ROUGE-SU4 along with their 95% confidence intervals included within square brackets.

We can see from Table 5 that the ranking algorithm can produce stable and promising results in the range of 0.4-0.7 for β. We also test the cases that the external reinforcement is considered more important than the internal

Table 5. Experiments on Weight Matrix

β	ROUGE-1	ROUGE-2	ROUGE-SU4
0.1	0.3805 [0.3744, 0.3863]	0.0777 [0.0739, 0.0815]	0.1344 [0.1309, 0.1381]
0.2	0.3831 [0.3772,0.3891]	0.0786 [0.0748,0.0825]	0.1355 [0.1319,0.1391]
0.3	0.3835 [0.3772, 0.3899]	0.0797 [0.0757, 0.0840]	0.1361 [0.1322, 0.1400]
0.4	0.3840 [0.3774, 0.3902]	0.0803 [0.0762, 0.0846]	0.1368 [0.1327, 0.1411]
0.5	**0.3861** **[0.3797, 0.3924]**	**0.0814** **[0.0774, 0.0857]**	**0.1384** **[0.1344, 0.1426]**
0.6	0.3868 [0.3806, 0.3932]	0.0806 [0.0767, 0.0848]	0.1384 [0.1346, 0.1424]
0.7	0.3860 [0.3797, 0.3925]	0.0800 [0.0761, 0.0841]	0.1378 [0.1339, 0.1417]
0.8	0.3855 [0.3793, 0.3918]	0.0797 [0.0758, 0.0836]	0.1376 [0.1339, 0.1416]
0.9	0.3851 [0.3788, 0.3914]	0.0792 [0.0753, 0.0832]	0.1373 [0.1335, 0.1413]
1.0	0.3859 [0.3796, 0.3923]	0.0786 [0.0747, 0.0826]	0.1372 [0.1335, 0.1412]
2.0	0.3859 [0.3797, 0.3921]	0.0793 [0.0752, 0.0835]	0.1370 [0.1332, 0.1411]
3.0	0.3817 [0.3756, 0.3877]	0.0772 [0.0735, 0.0807]	0.1338 [0.1302, 0.1373]
4.0	0.3796 [0.3736, 0.3858]	0.0764 [0.0728, 0.0798]	0.1329 [0.1294, 0.1363]
5.0	0.3787 [0.3727, 0.3846]	0.0758 [0.0724, 0.0793]	0.1324 [0.1290, 0.1359]

reinforcement (i.e. $\beta > 1$). Also from the following Table 5, the trend decline of the ROUGE results is observed when β gets bigger and bigger. It suggests $\beta < 1$ is a better choice than $\beta > 1$. This observation supports the common sense that the internal reinforcement should be more important than the external reinforcement in our D-S MR framework.

We are also interested to know the difference between the symmetric and the asymmetric versions of the weight matrix W. Now let $W = \begin{bmatrix} \alpha & \mu \\ \beta & \alpha \end{bmatrix}$ denote the weight matrix before normalization as before. We set α =1 and $\beta = 0.5$ according to results from the previous experiments (see Table 5) and then re-run the algorithm by setting the range of μ to 0.1 and 1.0 and the step size to 0. 1. We fix β (i.e. the weight of external reinforcement from sentence to document) in these experiments because the focus here is to rank the sentences for query-oriented multi-document summarization. The following Table 8 shows the ROUGE results.

Table 6. Experiments on Weight Matrix

μ	ROUGE-1	ROUGE-2	ROUGE-SU4
0.1	0.3805 [0.3745, 0.3864]	0.0777 [0.0738, 0.0816]	0.1346 [0.1310, 0.1383]
0.2	0.3831 [0.3772, 0.3892]	0.0786 [0.0749, 0.0825]	0.1355 [0.1319, 0.1391]
0.3	0.3835 [0.3772, 0.3899]	0.0798 [0.0758, 0.0840]	0.1361 [0.1322, 0.1401]
0.4	0.3840 [0.3774, 0.3902]	0.0803 [0.0762, 0.0846]	0.1368 [0.1327, 0.1411]
0.5	**0.3861** **[0.3797, 0.3924]**	**0.0814** **[0.0774, 0.0857]**	**0.1384** **[0.1344, 0.1426]**
0.6	0.3860 [0.3796, 0.3922]	0.0806 [0.0767, 0.0847]	0.1381 [0.1342, 0.1425]
0.7	0.3859 [0.3795, 0.3925]	0.0800 [0.0760, 0.0841]	0.1378 [0.1340, 0.1418]
0.8	0.3856 [0.3795, 0.3920]	0.0795 [0.0756, 0.0836]	0.1376 [0.1338, 0.1415]
0.9	0.3854 [0.3791, 0.3916]	0.0791 [0.0752, 0.0831]	0.1375 [0.1337, 0.1414]
1.0	0.3859 [0.3797, 0.3922]	0.0787 [0.0748, 0.0826]	0.1375 [0.1337, 0.1415]

As shown, the best performance is achieved at μ =0.5 when W is a symmetric matrix. The experiments here demonstrate the effectiveness of the symmetric weight matrix from the empirical perspective, while the mathematical analysis in Section 4.2 provides important properties of the symmetric weight matrix from the theoretical perspective.

In this section, we examine the effectiveness of the proposed Qs-MR based ranking algorithm for the task of query-oriented multi-document summarization. For comparison purpose, we also implement another two widely-used and well-performed ranking strategies. One is to rank the sentences according to their relevance to the query (denoted by QR). The other one is the PageRank deduced iterative ranking algorithm introduced in [18] (denoted by Q-LexRank). In the following experiments, we use the parameter setting obtained from the previous experiments, i.e. 0.75 for the damping factor d, 0.03 for the similarity threshold and the normalized version of $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ for the weight matrix. Table 7 shows the ROUGE evaluation results on DUC 2005 and DUC 2006 data sets, respectively.

Table 7. Comparison of Ranking Strategies

	ROUGE-1	ROUGE-2	ROUGE-SU4
	Experiments on DUC 2005 data set		
Qs-MR	0.3861 [0.3797, 0.3924]	0.0814 [0.0774, 0.0857]	0.1384 [0.1344, 0.1426]
Q-LexRank	0.3702 [0.3672,0.3772]	0.0725 [0.0704,0.0766]	0.1306 [0.1274,0.1341]
QR	0.3579 [0.3540, 0.3654]	0.0664 [0.0630, 0.0697]	0.1229 [0.1196, 0.1261]
	Experiments on DUC 2006 data set		
Qs-MR	0.4012 [0.3954, 0.4069]	0.0914 [0.0873, 0.0956]	0.1444 [0.1409, 0.1479]
Q-LexRank	0.3899 [0.3833,0.3964]	0.0856 [0.0813,0.0899]	0.1394 [0.1353,0.1438]
QR	0.3805 [0.3751, 0.3860]	0.0781 [0.0743, 0.0817]	0.1326 [0.1292, 0.1359]

As seen from Table 7, our proposed algorithm outperforms the QR algorithm significantly. Meanwhile, it can also outperform the traditional graph-based ranking algorithm (i.e. Q-LexRank). We summarize the improvements as follows in Table 8.

Table 8. Summary of improvements by DSR on DUC data set

	ROUGE-1	ROUGE-2	ROUGE-SU4
	Improvements over QR		
DUC 2005	**+7.88%**	**+22.59%**	**+12.61%**
DUC 2006	**+5.44%**	**+17.03%**	**+8.90%**
	Improvements over Q-LexRank		
DUC 2005	+4.29%	**+12.28%**	+5.97%
DUC 2006	+2.90%	**+6.78%**	+3.59%

5.4 Comparison with DUC Systems

We then compare our results with the DUC participating systems. To provide a global picture, we present the following representative ROUGE results of (1) the worst-scoring human summary (denoted by H), which reflects the margin between the machine-generated summaries and the human summaries; (2) the top five and worst participating systems according to their ROUGE-2 scores (e.g. S15, S17 etc.); and (3) the NIST baseline which simply selects the first sentences as summaries from the documents until the summary length is achieved. We can then easily locate the positions of our system developed based on Qs-MR among them. Notice that the ROUGE-1 scores are not officially released by the DUC.

Table 9. Comparison with DUC Participating Systems in the DUC 2005

	ROUGE-1	ROUGE-2	ROUGE-SU4
H	-	0.0897	0.1510
Qs-MR	**-**	**0.0814**	**0.1384**
DsR	**-**	**0.0771**	**0.1337**
S15	-	0.0725	0.1316
S17	-	0.0717	0.1297
S10	-	0.0698	0.1253
S8		0.0696	0.1279
S4		0.0686	0.1277
NIST Baseline	-	0.0403	0.0872

Table 10. Comparison with DUC Participating Systems in the DUC 2006

	ROUGE-1	ROUGE-2	ROUGE-SU4
H	-	0.1036	0.1683
S24	-	0.0956	0.1553
Qs-MR	**-**	**0.0914**	**0.1444**
S15	-	0.0910	0.1473
DsR	**-**	**0.0899**	**0.1427**
S12	-	0.0898	0.1476
S8		0.0895	0.1460
S23		0.0879	0.1449
NIST Baseline	-	0.0495	0.0979

As shown in Table 9 and 10, we can conclude that both Qs-MR and DsR outperform or are comparable to the top participating systems in both DUC 2005 and 2006 evaluations.

6 Conclusion

In this paper, we propose two alternative models to integrate the document dimension into existing sentence ranking algorithms, namely, the one-layer (i.e. sentence layer) document-sensitive model and the two-layer (i.e. document and sentence layers) mutual reinforcement model. While the former implicitly incorporates the document's influence in sentence ranking, the latter explicitly formulates the mutual reinforcement among sentence and document during ranking. When evaluated on the DUC 2005 and 2006 query-oriented multi-document summarization data sets, promising results are achieved.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. The ACM Press, New York (1999)
2. Brin, S., Page, L.: The Anatomy of a Large-scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems 30(1-7), 107–117 (1998)
3. DUC, `http://duc.nist.gov/`
4. DUC Reports, `http://www-nlpir.nist.gov/projects/duc/pubs.html`
5. Erkan, G., Radev, D.R.: LexRank: Graph-based Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22, 457–479 (2004)
6. Haveliwala, T.H.: Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering 15(4), 784–796 (2003)
7. Jones, K.S.: Automatic Summarising: The State of the art. Information Processing and Management 43, 1449–1481 (2007)
8. Langville, A.N., Meyer, C.D.: Deeper Inside PageRank. Journal of Internet Mathematics 1(3), 335–380 (2004)
9. Leskovec, J., Grobelnik, M., Milic-Frayling, N.: Learning Sub-structures of Document Semantic Graphs for Document Summarization. In: Proceedings of Link KDD Workshop, pp. 133–138 (2004)
10. Li, W.J., Wu, M.L., Lu, Q., Xu, W., Yuan, C.F.: Extractive Summarization using Intra- and Inter-Event Relevance. In: Proceedings of ACL/COLING, pp. 369–376 (2006)
11. Lin, C.Y., Hovy, E.: The Automated Acquisition of Topic Signature for Text Summarization. In: Proceedings of 18th COLING, pp. 495–501 (2000)
12. Lin, C.Y., Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: Proceedings of HLT-NAACL, pp. 71–78 (2003)
13. Lin, Z.H., Chua, T.S., Kan, M.Y., Lee, W.S., Qiu, L., Ye, S.R.: NUS at DUC 2007: Using Evolutionary Models for Text. In: Proceedings of Document Understanding Conference (2007)
14. Mihalcea, R.: Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In: Proceedings of ACL 2004, Article No. 20 (2004)
15. Mihalcea, R.: Language Independent Extractive Summarization. In: Proceedings of ACL 2005, pp. 49–52 (2005)
16. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. In: Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677 (1999)

17. Mani, I., Maybury, M.T. (eds.): Advances in Automatic Summarization. The MIT Press, Cambridge (1999)
18. Otterbacher, J., Erkan, G., Radev, D.R.: Using Random Walks for Question-focused Sentence Retrieval. In: Proceedings of HLT/EMNLP, pp. 915–922 (2005)
19. Ouyang, Y., Li, S.Y., Li, W.J.: Developing Learning Strategies for Topic-Based Summarization. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management, pp. 79–86 (2007)
20. Over, P., Dang, H., Harman, D.: DUC in Context. Information Processing and Management 43(6), 1506–1520 (2007)
21. Porter Stemmer, `http://www.tartarus.org/~martin/PorterStemmer`
22. Radev, D.R., Jing, H.Y., Stys, M., Tam, D.: Centroid-based Summarization of Multiple Documents. Information Processing and Management 40, 919–938 (2004)
23. Vanderwende, L., Banko, M., Menezes, A.: Event-Centric Summary Generation. In: Working Notes of DUC 2004 (2004)
24. Wan, X.J., Yang, J.W., Xiao, J.G.: Using Cross-document Random Walks for Topic-focused Multi-document Summarization. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 1012–1018 (2006)
25. Wan, X.J., Yang, J.W., Xiao, J.G.: Towards Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. In: Proceedings of ACL, pp. 552–559 (2007)
26. Wei, F.R., Li, W.J., Lu, Q., He, Y.X.: A Cluster-Sensitive Graph Model for Query-Oriented Multi-document Summarization. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 446–453. Springer, Heidelberg (2008)
27. Wei, F.R., Li, W.J., Lu, Q., He, Y.X.: Query-Sensitive Mutual Reinforcement Chain with Its Application in Query-Oriented Multi-Document Summarization. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 283–290 (2008)
28. Wong, K.F., Wu, M.L., Li, W.J.: Extractive Summarization Using Supervised and Semi-Supervised Learning. In: Proceedings of the 22nd International Conference on Computational Linguistics, pp. 985–992 (2008)
29. Yoshioka, M., Haraguchi, M.: Multiple News Articles Summarization based on Event Reference Information. In: Working Notes of NTCIR-4 (2004)
30. Zha, H.Y.: Generic Summarization and Key Phrase Extraction using Mutual Reinforcement Principle and Sentence Clustering. In: Proceedings of the 25th ACM SIGIR, pp. 113–120 (2002)
31. Padmanabhan, D., Desikan, P., Srivastava, J., Riaz, K.: WICER: A Weighted Inter-Cluster Edge Ranking for Clustered Graphs. In: Proceedings of 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 522–528 (2005)
32. Wei, F.R., Li, W.J., Lu, Q., He, Y.X.: Applying Two-Level Mutual Reinforcement Ranking in Query-Oriented Multi-document Summarization. Journal of the American Society for Information Science and Technology (2009) (in press)

Elements of Visual Concept Analysis

Lei Wu[1] and Xian-Sheng Hua[2]

[1] MOE-MS Key Lab of MCC,
University of Science and Technology of China
leiwu@live.com

[2] Microsoft Research Asia
xshua@microsoft.com

Visual concept analysis and measurements consist of low level visual analysis (image representation), image distance measurements (inter-image representation), semantic level concept modeling (concept representation) and concept distance measurements (inter-concept representation), which are four aspects of the fundamental visual concept analysis techniques. In the low level visual analysis, we discuss the visual feature, visual words, and image representations, based on which, we further discuss the image distance measurement. Beyond the low level analysis is the semantic level analysis, where we focus on the concept modeling and concept distance measurements. The methods for semantic level concept modeling can be roughly divided into generative model and discriminative models. In order to facilitate the following discussion on concept distance measurements, we mainly emphasize the generative models, such as bag-of-words model, 2D hidden markov model, visual language model. These models have been applied to the large scale real world Web image annotation and tagging tasks, and all of them represent the concepts in the form of distributions, so that they can be directly applied to the state of the art concept distance measurements, i.e. Flickr distance. These models and measurements are useful in numerous applications, i.e. image clustering, similarity search, object retrieval, annotation, tagging recommendation, indexing, etc. Some related applications are given to illustrate the usage of these measurements in real world problems.

1 Preliminary

In recent years, the success of information retrieval techniques in text analysis has aroused much interest in applying them to image retrieval tasks. One of the most successful models in text mining is the "bag of words" model (BoW), which greatly facilitates the document representation and makes the large scale indexing practical in text domain.

Much effort has been made to apply this model in visual domain, however there are several difficulties. The first challenge is the image representation.

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 679–717.
springerlink.com

Different from a text document, an image does not consist of the semantic units, i.e. the "words", but of low level pixels. To imitate the expression of text document, a concept named "visual word" is proposed as the semantic unit in visual domain.

There are quite a few approaches to generate the visual words. Although they may adopt different techniques to detect and represent the visual words, the general processes are similar. An image is firstly represented by sets of local appearance features or shape descriptors [37], [40], [3], [57], which are extracted either densely, at random, or sparsely according to some local salience criteria from images [42]. Then these features and descriptors are quantized or clustered into a collection of compact vectors, called "visual words". Each feature corresponds to one visual word, and all the visual words form a vocabulary named "codebook". With each visual word drawn from a fixed codebook, an image can then be viewed as a "bag-of-visual-words" representation. For simplicity, in this book, we also use "BoW" as abbreviation for the "bag-of-visual-word" model.

The BoW representation allows the leverage of text data mining techniques in visual domain including two typical topic detection techniques, i.e. the probabilistic Latent Semantic Analysis (pLSA) of Hoffmann [23], and the Latent Dirichlet Allocation (LDA) of Blei et al [8]. Recently, both methods have been exploited for visual categorization [10], [15], [17], [50], [53]. However, a desirable property of these topic detection techniques is that it is possible to achieve object localization or segmentation by investigating the topic posteriors of the visual words as shown in [10]. This requires that each semantic unit should correspond to certain semantic meaning as ordinary words do in text domain. Although much effort has been made on generating informative visual words, these visual semantic units still can not correspond to semantic concepts in nature language. This casts a big question on whether the direct application of these topic detection techniques on the meaningless visual words makes any sense, also known as the "semantic gap" problem.

Each visual word may not contain constant semantic meaning, while a group of related visual words may be more meaningful. This belief leads to the research on the correlation between visual words in an image. Although visual words do not contain specific meaning comparing with the ordinary words, there is one advantage for visual words, that is the informative spatial correlation. Comparing with words in text document, which have order in one dimension, visual words have spatial correlation with their neighbors in more dimensions. These "bag-of-visual-words" models [53], however, assume that the local appearance descriptors are independent with each other conditioned on the object class or the latent topic. Although this assumption greatly simplifies the computations, the ignorance of the spatial co-occurrence of local features may reduce the performance of the algorithms. Objects with different appearance but similar statistics of visual words tend to be confused. This motivates the research on exploring the complex spatial correlation between visual words into a uniform statistical language model.

Some related work has considered the co-occurrence of local descriptors, such as co-occurrence of pairs of visual words[31],"doublets" [53], correlograms of pixels [51], co-occurrence of multiple (more than two) neighboring descriptors [1], visual phrases [65], two-dimensional multi-resolution hidden Markov models (2-D MHMM) [33], Markov random fields (MRF) [34], conditional random fields (CRF) [29], and recently the efficient visual language model (VLM), which systematically incorporated the neighboring correlation of visual features into the BoW model. To handle the object scaling and views, scale invariant visual language model[59] and latent topic visual language model[60] are proposed.

Based on these efficient models, the relations between concepts can be well captured, which is known as the concept distance. Further we discuss several concept distance measurements, such as WordNet, Google distance, Flickr distance, and their applications to multimedia.

2 Low Level Visual Analysis

Low level visual analysis aims to calculate the visual property of certain regions in an image by the pixel level operations. It is the basic of the visual concept analysis, which aims to generate the statistical models from these region based visual properties.

Image consists of pixels, while each single pixel does not provide much information about the content of the image. A group of related pixels form a region in an image. The property of the regions provides some useful information about the image. Measuring the property of these regions is the so called low level visual analysis. This section discusses the low level image analysis, including visual features, visual words, and image representation.

2.1 Visual Features

Visual feature is defined as the global or local operations applied to an image to generate certain quantitative measurements, which are helpful for solving the computational tasks. According to the types of operations, the visual feature can be categorized into global feature and local feature. Global feature measures the property of the whole image, and local feature measures the property of local regions in the image.

Some commonly used global features include mean gray, gray histogram, image moment, texture histogram, etc. Each global feature can also be deemed as an operator L on the image I, denoted as $L(I)$. For example, let L_1 be the mean operator, and L_2 denotes the gray histogram, the mean gray and gray histogram features of given image I are represented by $L_1(I)$ and $L_2(I)$ respectively. These operators can nest, i.e. mean of gray histogram of image I is represented by $(L_1 * L_2)(I)$. Lots of types of global features can be derived by nesting different operators.

If these operators are applied to the regions of the image, the local feature can be generated. For example, let $r_i, i = 1, \cdots, n$ be a series of regions in image I. The operator L on each of the regions $L(r_i)$ will be the local features. However, the main difference of local feature from global feature does not lie on the descriptor, but on the localization and segmentation of these local regions, which are called feature detection. The process specializes the informative regions, such as lines, edges, angle, and movements etc. Some most commonly used feature detection methods include Moravec corner detection [45], Harris and Stephens corner detection [22], multi-scale Harris operator [6], level curve curvature approach[9], Laplacian of Gaussian (LoG) [35], difference of Gaussians approach (DoG) [43], determinant of the Hessian (DoH) feature detection[7], hybrid Laplacian and determinant of the Hessian operator (Hessian-Laplace)[43], Wang and Brady corner detection[21], SUSAN corner detector [54], Maximally stable extremum regions (MSER)[41], Affine-adapted interest point operator[43] etc. The well-known scale-invariant feature transform (SIFT) feature [38] is based on the DoG detection, with additional noise depression process and multi-scale keypoint descriptor.

In the following, we divide the visual features in two categories, global features and local features.

Global features

Mean Gray

The mean gray feature calculates the average gray level of the image or region. Given an image I, let I_{xy} be the pixel in the x-th column and y-th row in the image. The mean gray feature is defined as follows.

$$L_{MG}(I) = \frac{1}{mn} \sum_{x,y} I_{xy}$$

where $m \times n$ is the size of the image.

Image Moment

The image moment calculates the average pixel intensity by a particular weighting.

$$L_{M_{ij}}(I) = \sum_x \sum_y x^i y^j I_{xy}$$

where $i, j = 0, 1, \cdots$ and the moment sequence is uniquely determined by the image I.

Texture Histogram

A commonly used texture histogram is the 8-bin histogram. It divides the 2D space into 8 phases as shown in Fig. 1, denoted as $B_1, \cdots, , B_8$. The angle of each direction phase is set to be 45^o. Then it calculates the direction of texture d at each pixel by Eq. (1)

$$d_{xy} = \arctan \frac{dy_{xy}}{dx_{xy}}; x = 1, \cdots, m; y = 1, \cdots, n \qquad (1)$$

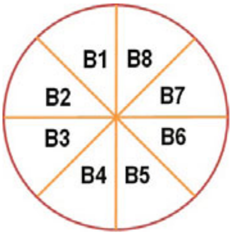


Fig. 1. 8-bin texture histogram

$$dx_{xy} = I_{xy} - I_{x+1,y} \tag{2}$$

$$dy_{xy} = I_{xy} - I_{x,y+1} \tag{3}$$

If the textural direction lies inside any of the 8 phases, the pixel is put into the corresponding bin, i.e. if $d_{xy} \in B_k$, then $\delta(I_{xy}, B_k) = 1$.

$$\delta(I_{xy}, B_k) = \begin{cases} 1, \; d_{xy} \in B_k; \\ 0, \text{ otherwise.} \end{cases}$$

Finally, the texture histogram $H = [h_1, \cdots, h_8]$ of the image is formed by calculating the number of pixels in each of the bins.

$$h_k = \sum_{xy} \delta(I_{xy}, B_k)$$

The above texture histogram does not consider the gradient magnitude of the texture. An improvement is to weight the texture histogram by gradient magnitude $\mathbf{m}_{xy}$ of each pixel (Eq. (4)). The magnitude is calculated in Eq.(5).

$$h_k = \sum_{I_{xy} \in B_k} m_{xy} \tag{4}$$

$$m_{xy} = \sqrt{(dx_{xy})^2 + (dy_{xy})^2} \tag{5}$$

Rotation Invariant Texture Histogram[62]

Comparing with mean gray and gray histogram, texture histogram contains more structural information. Since most objects and informative regions are lying on edges or corners, this feature is more discriminative for describing informative regions. However, it is sensitive to rotation. Suppose an object is rotated a little in the image, the texture histogram may be altered dramatically. For this reason the rotation invariant texture histogram is used.

In order to make the texture histogram resistant to rotation variance, the average textural direction D of an image is firstly calculated.

$$D = \arctan \frac{\sum_{xy} dy_{xy}}{\sum_{xy} dx_{xy}} \tag{6}$$

$$dx_{xy} = I_{xy} - I_{x+1,y} \tag{7}$$

$$dy_{xy} = I_{xy} - I_{x,y+1} \tag{8}$$

Then the patch is rotated to make the average textural direction vertical pointing to the top. Starting from the average direction, eight direction phases are defined as $[B_1, B_2, \cdots, B_8]$. The angle of each direction phase is set to be 45^o. Then the texture at each pixel is calculated and quantized into each bin by the same means as discussed previously.

Local Features

Scale-Invariant Feature Transform (SIFT) [38]

The Scale-Invariant Feature Transform (SIFT) is one of the widely used local features. It is believed to be invariant to both image scaling and rotation. There are several steps to generate the SIFT feature.

First step is keypoint detection. The SIFT feature adopts the Difference of Gaussian (DoG) method to help detect the keypoints. DoG actually calculates the difference of the Gaussian-blurred images L_G at scales $k_i\sigma$ and $k_j\sigma$. The difference between the Gaussian-blurred images is defined as the DoG image, denoted L_{DoG}.

$$L_{DoG}(x, y, \sigma) = L_G(x, y, k_i\sigma) - L_G(x, y, k_j\sigma)$$

$$L_G(x, y, k\sigma) = G(x, y, k\sigma) \times I(x, y)$$

where $G(x, y, k\sigma)$ is the Gaussian blur at scale $k\sigma$. Based on the DoG image, scale-space extrema detection [35] is used to locate the keypoints. This algorithm defines a local region of nearest 26 neighbors in a discrete scale-space volume, and finds the points that are local extrema with respect to both space and scale.

Second step is noisy point filtering. The noisy points are categorized into two folds. One fold contains points that lie on some low contrast regions. The other fold contains points which are poorly located but have high edge responses. To remove the first type of noise, the method filters the scale-space extremas by contrast, and points in high contrast regions are preserved. To handle the second type of noise, it filters the extremas by principal curvature. The points located along edges with large curvature is preserved.

Third step is the orientation assignment. In order to achieve invariance to rotation, each keypoint is assigned an orientation relative to the neighboring region. Under each scale $k\sigma$, the gradient magnitude m_{xy} and orientation $D(x, y)$ with 36-bins are calculated at each keypoint. Then within a neighboring region around each keypoint, a histogram of 36-bin orientations is formed and weighted by gradient magnitude. The highest peak is assigned to the keypoint as its orientation.

Final step is the feature descriptor. Each keypoint corresponds to a feature descriptor, which contains 4×4 array of 8-bin histogram around each keypoint. So the descriptor is of 128 dimensions.

2.2 From Visual Feature to Visual Word

Previously we have introduced several types of well-known visual features. These visual features capture the low level property of the image, however, some of the visual features are in high dimensional space, which are hard to store and calculate. High dimensional features often confront the sparseness problem and the noisy problem. In this section we discuss the mapping from high dimensional visual feature to the lower dimensional form which is called *visual word*. By the dimension reduction and coding technology, visual words are easy to store and efficient for indexing and calculation.

Generally there are several methods to map the visual feature into visual words, such as the principle component analysis, clustering, hash coding, etc. We will discuss each of the methods in the following.

Mapping by PCA

Principle component analysis (PCA) can be used to map the high dimensional feature into low dimensional visual words. The basic assumption of using PCA for the mapping is that high dimensions are correlated to each other. The intuitive explanation is there are much redundant in the feature dimensions, and removing certain dimensions will not loss much information. PCA is able to transform a number of possibly correlated dimensions into a smaller number of uncorrelated dimensions, which are called *principal dimensions*.

Given a feature matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ of size $m \times n$, the target is to find a mapping $\mathbf{W}$

$$\mathbf{X} = \mathbf{W} \mathit{\Sigma} \mathbf{V}^\top$$

where the diagonal entries of $\mathit{\Sigma}$ are known as the *singular values* of $\mathbf{X}$, and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m]$ is an $m \times m$ unitary matrix. The solve of the problem is equivalent to finding the singular value decomposition of the data matrix $\mathbf{X}$.

From another perspective, the mapping by PCA method can be interpreted by the following iterative optimization process. Firstly, the method will try to find a principle dimension $\mathbf{w}_1$ to maximizing the variances between dimensions.

$$\mathbf{w}_1 = \arg \max_{\|\mathbf{w}\|=1} var \left\{ \mathbf{w}^\top \mathbf{X} \right\}$$

Then the method will try to find the second principle dimension w_2 to maximizing the variance between data subtracting the previous principle dimensions.

$$\mathbf{w}_2 = \arg \max_{\|\mathbf{w}\|=1} var \left\{ \mathbf{w}^\top \hat{\mathbf{X}}_1 \right\}$$

$$\hat{\mathbf{X}}_1 = \mathbf{X} - \mathbf{w}_1^\top \mathbf{X}$$

So the k-th principle dimension $\mathbf{w}_k$ can be calculated as follows.

$$\mathbf{w}_k = \arg \max_{\|\mathbf{w}\|=1} var \left\{ \mathbf{w}^\top \hat{\mathbf{X}}_{k-1} \right\}$$

$$\hat{\mathbf{X}}_{k-1} = \mathbf{X} - \sum_{i=1}^{k-1} \mathbf{w}_i^\top \hat{\mathbf{X}}_{i-1}$$

Given the visual feature $\mathbf{x}_i$, the final visual word $w_L(\mathbf{x}_i) = \mathbf{W}_L^\top \mathbf{x}_i$, where L is the dimension of the visual word space, and $\mathbf{W}_L = [\mathbf{w}_1, \cdots, \mathbf{w}_L]$.

Mapping by Clustering

Another method for the feature-word mapping is by clustering. The assumption of using the clustering method is that the features in a cluster has similar meaning. In other words, the nearby features in the space are redundant and can be represented by only one feature at the center. However, in some cases, the assumption does not hold, and the mapping by clustering may lead to semantic loss, which means the visual word may not be as discriminative as the visual features. Besides, the number of visual words is hard to choose. So clustering the visual features in the high dimensional feature space is a simple but far from an ideal way to generate the visual words.

K-means Clustering

K-means is a commonly used method for clustering the visual features into visual words, i.e. the well-known bag-of-words model adopts this approach. It aims to partition n observations (features) into k clusters (visual words as cluster centers $(w_1, \cdots, w_k)$ where each observation belongs to the cluster with the nearest visual words. Given a set of visual features $(x_1, \cdots, x_n)$, where each feature is represented as a d dimensional vector. The algorithm aims to partition the set into k clusters $\mathbf{S} = (S_1, \cdots, S_k)$ so that the within-cluster sum of square is minimized.

$$\mathbf{S}^* = \arg \min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{j=1}^{n} \gamma_{ji} \|x_j - w_i\|^2$$

where γ_{jk} indicates whether the j-th feature belongs to the k-th cluster.

An EM like iterative process is used to solve the clustering. Firstly some initial values for w_i is randomly chosen. In the E (expectation) process, we fix the visual words w_i, and minimize the object function with respect to γ_{ji}, which is calculated by Eq. (1).

$$\gamma_{jk} = \begin{cases} 1,\ w_k = \arg\min_{w_i} \|x_j - w_i\|^2; \\ 0,\ \text{otherwise.} \end{cases} \tag{9}$$

In the M (maximization) process, we fix parameter γ_{ji}, and minimize the object function with respect to w_i, which obtains the updating equation 2.

$$w_i = \frac{\sum_j \gamma_{ji} x_j}{\sum_j \gamma_{ji}} \tag{10}$$

The E step and M step will perform iteratively until convergence.

Gaussian Mixture Models

Another commonly used method for clustering the visual features is the Gaussian mixture model, the discussion of which will provide deeper insight into the clustering based mapping methods. Rather than the single Gaussian density assumption, the Gaussian mixture model assumes that the visual features are distributed in multiple Gaussian densities, which is written as

$$p(x_i) = \sum_{k=1}^{K} \phi_k \mathcal{N}(x_i | w_k, \sigma_k).$$

where $p(x_i)$ is the density function of visual feature x_i. Each visual word w_k is the mean of a Gaussian distribution $\mathcal{N}(x_i|w_k, \sigma_k)$, and σ_k is the variance of the k-th Gaussian. ϕ_k is the mixture coefficient which determines the contribution of each Gaussian distribution to the visual feature density. And the mixture coefficient ϕ_k should meet the following conditions:

$$0 \leq \phi_k \leq 1$$

and

$$\sum_{k=1}^{K} \phi_k = 1$$

If we consider each Gaussian is a state z_k of the visual feature, $p(z_k) = \phi_k$ represents the prior that the state z_k appears, and $p(x_i|z_k)\mathcal{N}(x_i|w_k, \sigma_k)$ denotes the posterior of the visual feature given that k-th visual word. The probability of assigning the visual feature x_i to the visual word w_k is

$$p(z_k|x_i) = \frac{\phi_k \mathcal{N}(x_i|w_k, \sigma_k)}{\sum_j \phi_j \mathcal{N}(x_i|w_j, \sigma_j)}.$$

The Gaussian mixture model can also be formulated as a maximization problem.

$$w^* = \arg\max_{w,\sigma,\phi} p(\mathbf{X}|\phi, w, \sigma) = \arg\max_{w,\sigma,\phi} \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} \phi_k \mathcal{N}(x_i|w_k, \sigma_k) \right\}$$

This problem can also be solved by the EM algorithm. We obtain,

$$w_k = \frac{\sum_{i=1}^{N} p(z_k|x_i) x_i}{\sum_{i=1}^{N} p(z_k|x_i)} \tag{11}$$

$$\sigma_k = \frac{\sum_{i=1}^{N} p(z_k|x_i)(x_i - w_k)(x_i - w_k)^\top}{\sum_{i=1}^{N} p(z_k|x_i)}$$

$$\phi_k = \frac{\sum_{i=1}^{N} p(z_k|x_i)}{N}$$

Comparing the result of k-means (Eq.(2)) and the result of Gaussian mixture model (Eq.(3)), we find that the k-means clustering is in fact the "hard" version of the Gaussian mixture model. In k-means, the weight γ_{jk} is a binary indicator, while in Gaussian mixture model, the weight $p(z_k|x_i)$ is a continuous probability.

Besides these unsupervised clustering methods, there are also some semi-supervised clustering methods that may be helpful in generating the visual words. However, as the label information for the visual features is hard to obtain, currently most models adopt the unsupervised clustering methods. Further study on adopting semi-supervised clustering to handle this problem would be an interesting research topic.

Mapping by Hash Coding

Hash coding method will transform the high dimensional visual feature into a brief datum or a simple integer to help index the features into an array. The mapping is generally performed by a well-defined hash function, which takes a visual feature as input, and outputs an integer naming *hash code* or *visual word*. Hash functions are widely used in indexing and data compressing. This method is very efficient and suitable for large scale dataset, however, the hash function may map different visual features into one visual word. In other words, the hash coding method will lose information during the mapping procedure. In the following, we will discuss two of the well-known hash coding methods.

Binary code

The binary code represents each dimension of the visual feature into a binary bit (0 or 1) based on a threshold. It is usually used in indexing the text documents. Here we show its usage in representing the visual features.

Given a set of visual feature $\{f_i\}_{i=1}^n$, where f_i is a d-dimensional vector $f_i = [f_i(1), \cdots, f_i(d)]$. The hashing function is defined in the follows.

$$h_i(k) = \begin{cases} 1, \ f_i(k) > mean_j(f_j(k)); \\ 0, \text{ otherwise.} \end{cases}$$

Here we set the threshold as the mean value of this dimension. If the feature intensity is above the average level, this dimension is coded 1; otherwise coded 0. Then put all the d bits together to form a integer $\mathbf{h} = [h_i(1) \cdots h_i(d)]$. In this way the high dimensional visual feature can be coded into a short integer for indexing.

Locality-Sensitive Hashing (LSH)

Locality-Sensitive Hashing (LSH)[20] aims to map the visual features into some buckets, so that the similar visual features are in the same buckets with high probability. The LSH method defines a hash function family $\mathcal{H}$ and a similarity function ϕ. For any two features f_i, f_j in the feature space $\mathcal{F}$, the hash function $h_k \in \mathcal{H}$ is chosen according to certain distribution P, which satisfies the property that

$$P[h_k(f_i) = h_k(f_j)] = \phi(f_i, f_j)$$

In some cases, we define a threshold on the distance between features to generate a simple representation of the LSH method. Suppose we define a threshold R, a family $\mathcal{H}$ is called $(R, cR, p1, p2)$-*sensitive* if for any features $f_i, f_j \in \mathcal{F}$

$$P_{\mathcal{H}}[h_k(f_i) = h_k(f_j)] \begin{cases} \geq p_1, \ \|f_i - f_j\| < R; \\ \leq p_2, \ \|f_i - f_j\| > cR. \end{cases}$$

where $p1$ and $p2$ are two probabilities, $p1 > p2$ to ensure the LSH family useful.

One of the easiest ways to construct an LSH family is by bit sampling [25]. Given a visual feature f_i, the hash function $h(f_i) = f_i(k)$, where $f_i(k)$ is the k-th dimension of the visual feature. The random hash function h actually maps the feature to one of its dimensions. This LSH family has the following parameters.

$$p_1 = 1 - R/d; \ p_2 = 1 - cR/d$$

There are also many other approaches to construct the LSH family which are beyond the discussion of this book. Please refer to [20] for further reading.

Mapping by Multiple Methods

Multiple mapping methods can be combined to generate the visual words. For example in [56], the authors combines PCA with binary hashing methods to map a visual feature into a visual word.

2.3 Image Representation

Image representation is to depict an image by its intrinsic characteristics. It is used to index and search images, discern an image after modification, or differentiate an image from different ones. There are several types of image representations, such as pixel level representation, global feature representation, and local feature representation.

The easiest approach to represent an image is by recording the RGB information of its pixels. Given an image I of size $n \times m$, the pixel level representation is an $n \times m \times 3$ dimensional vector, recording the RGB of each pixel in the image. This kind of representation is simple but have several drawbacks. Firstly, it is a high dimensional vector. If an image of size 640×480, the representation consists of 921,600 dimensions. The high dimensional representation is hard to store as well as calculation. Secondly, this representation is sensitive to noise. All the backgrounds pixels, which may not be relevant to the topic of the image, are recorded in the vector. Thirdly, it is sensitive to illumination changes, scaling, rotation, etc. Regulation of the color, light and contrast will alter the vector dramatically.

To reduce the dimensionality of the image representation, global feature representation is proposed. The global feature represents an image by certain statistical measurements, which can both depress the noise and also reduce the dimensionality. Some of the common global features are discussed in Section 2. The dimension of the representation is determined by the dimension of the global features. This kind of global representation is compact and efficient, but they ignore much information of the image. This rough representation is not capable to describe the objects, which only exists in some local regions inside the image.

To improve the discrimination of the representation so that the objects can be detected in the images, local representation is proposed. Different from the global representation, the local representation focuses on not only describing the statistical characteristics of the image but also locating the meaningful objects inside the image. So the main challenging problem with local representation is the region localization and segmentation. There are generally four types of methods to localize the regions.

Uniform patch

One easy way to divide an image into regions is by uniform patches. Given an image of size $n \times m$, the local patch size is defined as $k \times l$, then the image can be divided into $\frac{n}{k} \times \frac{m}{l}$ equal-sized non-overlapping patches. The advantage of this scheme is its usage of all the information in the image. The disadvantage is its sensitivity to object motion, rotation and scaling. This approach is tested effective and reported performing not worse than the complex regions of interest (ROI) method [16] in object recognition.

Random windows

In some application, such as the near duplicate image detection, the image may be cropped in one dimension or scaled. Although it is modified a little, this modified image should be considered near-duplicate with the original image. However, as this modification will completely change the patches if we use the fixed sized patch, these near-duplicates may be taken as different images. To avoid this from happening, random windows scheme is used. The random windows scheme will generate a series of subregions with random location and random size. When the number of random windows is large, this seemly chaos representation can be robust to such modifications as scaling, cropping, motion, etc. However, the problem with the random windows scheme is that there should be a large number of windows. It is generally impractical for large scale dataset; otherwise, the random windows may bring great noise, i.e. background and meaningless regions, to represent the image.

Segmentation

To avoid bringing irrelevant regions into the image representation, segmentation is necessary. Image segmentation will segment the image into regions by the boundaries and edges. The image segmentation provides possibility to focus only on the meaningful regions, and somewhat prevents the semantic integrated regions from broken. However, it is a time consuming process, and current image segmentation does not provide reliable results.

Regions of Interest

The region of interest scheme will find the informative regions which contain interesting features, such as edges, corners, blobs, Ridges, etc. ROI is widely used in local feature detection. With complex region detectors, such as Laplacian of Gaussian (LoG) [35], difference of Gaussians approach (DoG) [43], determinant of the Hessian (DoH) feature detection[7], hybrid Laplacian and determinant of the Hessian operator (Hessian-Laplace)[43], the ROI representation can focus on the object regions, but it may also loss some context information. This representation is especially suitable for BoW model, but not suitable for context model, since the spacial relation between the regions is hard to retrieve.

3 Image Distance Measurement

In the previous section, we discussed low level visual analysis, which extract information from pixels within an image. Now, we can proceed to calculate the relationship between images by measuring the image distance.

Given an image I, which is represented as a d-dimensional vector, the image distance measures the difference between this image vectors. According to the choice of metric space, the image distance measurement can be divided into static distance and dynamic distance. Static distance measurement measures the image distance in a fixed metric space. For example, Euclidian distance measures the image distance in a unique euclidian space, Mahalanobis distance project the feature by a Mahalanobis matrix and then measure the distance in the fixed space. Dynamic distance measurement measures the image distance in different metric spaces. The method will adaptively choose one of the subspaces to define the distance metric. In some cases, the algorithm may also measure the image distances in multiple subspaces iteratively.

3.1 Static Distance Measurement

One of the simplest way to calculate the image distance is to calculate the L2 distance between image vectors, which is also called Euclidian distance.

Let v_i and v_j be the vectors for the i, j-th images, the Euclidian distance between the two images is:

$$D_{L2} = \sqrt{\|v_i - v_j\|^2}$$

Sometimes, the Euclidian distance does not truly reflect the semantic distance between two images. The distance should be measured in a warped space so that the semantically similar images are close to each other and irrelevant images are alienated. To meet this demand, Mahalanobis distance is proposed. The basic idea of Mahalanobis distance is to take into account the correlation of the data set and is scale-invariant, i.e. independent on the scale of the measurement.

$$D_M(v_i, v_j) = \sqrt{(v_i - v_j)^\top M(v_i - v_j)}$$

where M is the covariance matrix of the dataset.

Due to the semantic gap, the covariance matrix of the low level features in the Mahalanobis distance does not reflect the relation between semantic objects. This kind of Mahalanobis distance only measures the distance of the low level features and the sematic relations are not well measured. To provide more meaningful distance measure for semantic images, researchers work hard on learning a Mahalanobis distance with side information to bridge the semantic gap. The main idea is to incorporate the label information into the distance measurement. This label information indicates which pair of images contain similar objects and are considered as positive constraint, and which pair of images are irrelevant and are considered as negative constraints. Then the method try to learn such a covariance matrix M, so that the distance between images with positive constraints are minimized and the distance between images with negative constraints are maximized.

This work is also called distance metric learning (DML). Some of the successful DML methods include PDGM[63], NCA[26], RCA[5], DCA[24], LMNN[58], ITML[13], DistBoost[55], etc. Please refer to the respective papers for further readings.

3.2 Dynamic Distance Measurement

The previously discussed image distance measurements only measure the distances in fixed warped space, i.e. given the dataset and constraints, the covariant matrix M is fixed for any pair of images. In certain cases we need to measure distance between two images in multiple spaces or adaptively choose a space for distance measurement with respect to the property of the test images. This leads to the research on the dynamic image similarity measurement.

One of the well known method for dynamic distance measurement is the query oriented subspace shifting algorithm (QOSS)[62]. The basic assumption of the QOSS algorithm is that if two images are near-duplicate to each other they should be similar in multiple subspaces; if two images are not near-duplicate they are only similar in certain subspaces. So measuring the distance in multiple subspaces is more robust than distance measurement in only one subspace. The challenging problem with dynamic distance measurement is how to choose the subspaces. Apparently, we can not try every subspace, since the number of subspaces may be infinite.

QOSS firstly measures the distance in the subspace where data are mostly separated. Then it removes the irrelevant data by a threshold on the distance and keep the related data in the loop. In the next iteration, it measures the distance in the subspace where the related data are mostly separated. This process goes on until convergence. It uses the maximum distance between two images in all iterations to measure their distance. This dynamic distance measurement can detach the irrelevant images from each other, and keep low distance between highly relevant images.

This scheme is presented in details as follows.

Step 1: Calculate the closeness threshold ϵ in the subspace κ by the same means in rough filtering;

Step 2: Select the query surrounding samples and update the Q_s;

$$Q_s = \{I_j | \|PQ - PI_j\|_\kappa < \epsilon\} \tag{12}$$

where P is the projection matrix.

Step 3: Update the projection matrix P based on the query surrounding collection;

$$P_i \leftarrow eigenvector(cov(Q_s), i) \tag{13}$$

$$P = [P_0, P_1, \cdots, P_d] \tag{14}$$

where $eigenvector(cov(Q_s), i)$ is the i^{th} sorted eigenvector of the covariance matrix for query surrounding collection, and d is the dimension of the low dimensional space.

Step 4: Repeat Step 1 and 3, until the query surrounding collection Q_s does not change. So far, we believe all the non-duplicates surrounding the query image are filtered, and the algorithm finishes.

4 Semantic Level Concept Modeling

In the previous sections, we have discussed low level visual analysis and image distance measurement. Low level visual analysis only deal with the representation of an image, and image distance measures the relation between images. However, in many tasks, we not only need image representation and their pairwise relationship, but also have to deal with the representation of a group of semantically related images. Here “semantically related images” denotes that these images contains same objects or related to the same topic. We call the representation of these semantically related images as “concept modeling”, which is quite useful in object detection, classification, annotation, tagging, etc. Different from the low level analysis which analyze the property within one image, the concept modeling analyzes the statistical property of visual words over a collection of images. In the following, we will start with the concept definition, and discuss several state of the art generative modeling methods.

Here we only focus on three of the generative modeling methods, because they are related to the following section on concept distance measurement. There are also many other modeling methods, including correlative model [49], transfer model [30], Max-Margin Hough Transform [39], Multiple Instance model [4], discriminative model [14], kernel methods [64], Hierarchical Representations [47] among others. These models are out of the scope of this section, interested readers may refer to the references for further reading.

4.1 Concept Definition

According to contemporary philosophy, a concept is a cognitive unit of meaning, which is divided into objects and abstract objects. Generally, an object is an entity that could be perceived by human sensors, like sense of sight or tactile organ. Abstract objects are abstract ideas or mental symbols also named as the "unit of knowledge". Here the concept only refers to the objects, such as concrete objects, events, scenes, or motions. Since these concepts are frequently appeared and can be recorded by photos. Only the recordable concepts can be modeled and measured by our technology.

Each concept is related to some of the other concepts. For example, when talking about the concept "airplane", people always think about the "airport", or when we think of a "dog" and often can imagine the "head" and the "legs". This conceptual relationship enbodies the distance between concepts in the semantic space, where semantically related concepts are closer to each other, i.e. "airplane" and "airport" is closer than "airplane" and "dog".

Concept is a semantic element in human cognition, and the distance between concepts is also a measurement in human cognition. It is difficult to calculate it directly. One of the possible means to measure this distance is to simulate the human cognition. The cognition is formed based on multiple information sources, such as from reading text documents, photos, videos, and verbal communications. The basic idea of measuring the conceptual distance is by data mining from a large pool of these multimedia knowledges.

4.2 Bag of Words Model (BoW)

A concept can be represented in many forms. Here we refer to the representation of a concept from its visual characteristics. Specifically, we focus on two of well-known models, the bag of words model (BoW) and the visual language model (VLM).

The bag of words model is a simple assumption firstly used on nature language processing and information retrieval. In this model, each document is considered as a “bag”. The words in the document are considered independent to each other, and thus the order of words is ignored. The model assumes that a document (text) can be well represented by a collection of words, ignoring their orders and grammar. This simple assumption makes the model easy to be adopted into both the Naïve Bayes framework and the Hierarchical Bayes framework. Since the order of words is ignored, this assumption leads to a simple conditional independence Bayes model.

Naïve Bayes framework

A general classification task can be performed by maximizing the likelihood function as follows.

$$p(C|I) = \frac{p(I|C)p(C)}{P(I)} \propto \prod_i p(w_i|C)p(C)$$

where w_i is the i-th word in the document I, and C is the concept (category). The model calculates the likelihood for each of the concepts and choose the concept that maximizes the likelihood function as the optimal concept. This generative model is very efficient in dealing with the large scale text data.

Hierarchical Bayes framework

To better analysis and detect multiple unknown objects in one image, i.e. a scene containing people, building, cars, dogs, the hierarchical Bayes model can be used. Here we introduce two commonly used hierarchical Bayes models, probabilistic latent semantic analysis (pLSA) and latent dirichlet analysis (LDA) model.

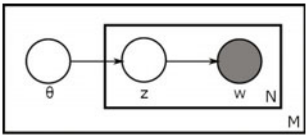


Fig. 2. The graphical model for pLSA

The pLSA model[23] was first introduced in 1999 by Jan Puzicha and Thomas Hofmann. pLSA is one of the well known topic models, which is commonly used in the information retrieval and object recognition. It assumes there are fixed number of latent topics in each document. Both the distribution of words and the distribution of documents are conditioned on these latent topics, as shown in Figure 2. The model takes each word in the document as a sample from a mixture model, whose parameters are multinomial variables, which is also deemed as the representation of the “topics”. Given the observation of co-occurrence of words and documents, the pLSA models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions:

$$p(w, I) = \sum_z p(w|z)p(I|z)p(z) = p(I) \sum_z p(w|z)p(z|I)$$

where z represents the latent topic. The first formulation is the symmetric formulation, which assumes given the latent topic z, the words and documents are independent, and word w and document I are both conditioned on the latent topic z. The second formulation is the asymmetric formulation, where the topic z is dependent on the document, and the words are generated according to $p(w|z)$. Thus each word generates from one topic, and different words in a document can generate from different topics from a topic collection of fixed size. In this way, each document is represented as a list of mixing proportions of a fixed set of latent topics.

This form has reduced the description of a document, however there are several drawbacks with pLSA model. Firstly, there is no generative probabilistic model for the mixing proportions, i.e. it is unknown how to assign mixing proportions or topic distributions to document outside the training set. Secondly, the number of parameters grows linearly with the size of the corpus as well as the number of documents, which leads to a serious problem of overfitting.

To calculate the parameters $p(w|z)$ and $p(z|I)$ for this model, an EM algorithm is used, which iteratively update each parameter to maximizing the likelihood.

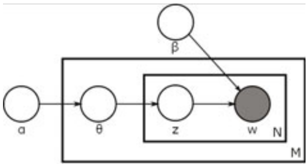


Fig. 3. The graphical model for LDA

LDA model is a three layer hierarchical Bayes model, which is proposed to handle the overfitting problem by adding a Dirichlet prior on the per-document topic distribution. In LDA, each document is also assumed a mixture of multiple latent topics, which is similar to the pLSA model except that LDA further assumes a Dirichlet prior on the topic distribution, the graphical model of which is shown in Figure 3, where α is the parameter of the uniform Dirichlet prior on the per-document topic distributions, and β is the parameter of the uniform Dirichlet prior on the per-topic word distribution. θ_i is the topic distribution for document i. z_{ij} is the topic for the j-th word in document i. The only observations in the model are the words w_{ij}.

BoW model for object recognition

Later, Fei-Fei et al. applied the BoW model and related hierarchical Bayes models to object recognition, where the latent topics are considered as the semantic objects in the image, the words correspond to the visual words in the documents, and the documents are the images. This model achieves success on the Caltech dataset, but it is also reported that pLSA has severe overfitting problems [8]. The number of parameters grows linearly with the number of documents. In addition, although pLSA is a generative model of the documents in the collection it is estimated on, it is not a generative model of new documents.

4.3 2D Hidden Markov Model

In the hidden markov model (HMM) (Fig. 4), There are states and observations. The states can transfer between each other with certain probability but these states are invisible (hidden), and only the observations are visible. Each state has a probability distribution over the possible observations. Thus given a sequence of observations, the possible sequence of states can be estimated. Jia et al. [32] firstly adopted the 2D HMM model for image segmentation and region classification. With some modification, this approach can also be used to represent a concept.

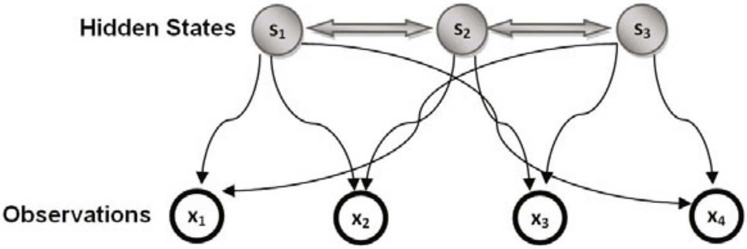


Fig. 4. The illustration of Hidden Markov Model

To model a concept, it firstly divides the image into non-overlapping blocks, each of which is further represented by the Wavelet feature x_{ij}, where i, j indicate the vertical and horizontal location of the block. It defines $x_{i'j'}$ is previous of x_{ij}, if $i' \leq i$, $j' < j$, which is also denoted $(i', j') < (i, j)$. This model takes these feature vectors as observations, and assume that there are M possible states for each block, denoted as s_{ij}.

There are three assumptions to simplify the calculation of the 2D HMM model.

Firstly, it assumes that the state for each block is conditioned on the states and features of the immediate previous blocks.

$$P(s_{ij}|I) = P(s_{ij}|s_{i',j'}, u_{i',j'}, (i', j') < (i, j)) = a_{mnl}, \tag{15}$$

$$m = s_{i-1,j}, n = s_{i,j-1}, l = s_{ij} \tag{16}$$

where I is an given image, and P represents a probability of an event. a_{mnl} is the transitional probability from immediate previous state $s_{i-1,j}$ and $s_{i,j-1}$

to current state s_{ij}. The class for each block c_{ij} is uniquely determined once the states are known. In other words, the state of each block indicates the concept it belongs to.

Secondly, it assumes that given the state of a block s_{ij}, the feature vector x_{ij} is independent to those of other blocks.

$$P(x_{ij}, (i,j) \in \mathcal{N} | s_{ij}, (i,j) \in \mathcal{N}) = \prod_{(i,j)\in\mathcal{N}} P(x_{ij}|s_{ij})$$

where $\mathcal{N} = \{(i,j), 0 \leq i < n_x, 0 \leq j < n_y\}$ denotes the collection of all blocks in an image.

Thirdly, it assumes that once the state of a patch is known s_{ij}, the feature vector follows the Gaussian distribution.

$$P(x_{ij}|s_{ij}) = \frac{1}{\sqrt{(2\pi)^n \| \sum_{s_{ij}} \|}} e^{-\frac{1}{2}(x_{ij}-u_{s_{ij}})^\top \sum_{s_{ij}}^{-1}(x_{ij}-u_{s_{ij}})}$$

where $\sum_{s_{ij}}$ is the covariance matrix and $u_{s_{ij}}$ is the mean vector.

According to these assumptions, the calculation of $P(s_{ij}, x_{ij}, (i,j) \in \mathcal{N})$ can be calculated as:

$$P(s_{ij}x_{ij}, (i,j) \in \mathcal{N}) = P(s_{ij}, (i,j) \in \mathcal{N}) \prod_{(i,j)\in\mathcal{N}} P(x_{ij}|s_{ij}) \tag{17}$$

The probability of a state sequence of the image can be estimated by:

$$P(s_{ij}, (i,j) \in \ ) = P(T_0)P(T_1|T_0)P(T_2|T_1) \cdots P(T_{n_x+n_y-2}|T_{n_x+n_y-3})$$

where T_i denotes the sequence of states for blocks lying on i-th diagonal as shown in Fig. 5.

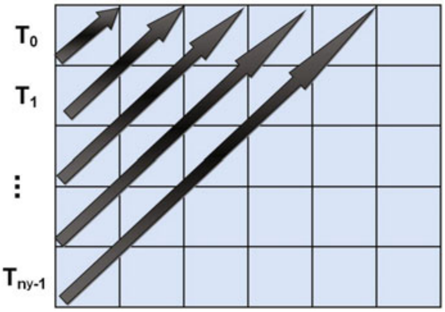


Fig. 5. Sequence of blocks lying on diagonals.

Finally, concepts can be represented as a joint distribution of states (concepts) and their local features in the image in Eq.(17).

4.4 Visual Language Model (VLM)

The visual language model [61] adopts the uniformly distributed equal-sized patches to represent an image (Fig. 6). This representation will facilitate calculate of conditional probability in VLM. The model assumes the patches are generated from top to bottom, and from left to right. Each patch is related to its previous patches. In this assumption, the spatial dependence between image patches are modeled in the form of patches' conditional probability.

The training process will model the following conditional probability

$$p(w_{ij}|w_{00}w_{01}\cdots w_{mn}) = p(w_{ij}|w_{00}w_{01}\cdots w_{i,j-1}) \tag{18}$$

The calculation of this conditional probability is still not efficient enough. Inspired by 2D HMM [46] used in face recognition, the model assumes that each patch depends only on its immediate vertical and horizontal neighbors. Although there may be some statistical dependency on visual words with larger interval, the description of this dependency will make the model too complex to implement. This model ignores this kind of dependency, just as the language model does for text classification.

According to how much dependency information is considered in the model, there are three kinds of visual language models, i.e. unigram, bigram and trigram. In unigram model, the visual words in an image are regarded independent with each other. In bigram model, the dependency between two neighboring words is considered. And in trigram model a word is assumed to depend on both the word on the left and the word above it.

These three models are expressed in Eq. (19)-(21) respectively.

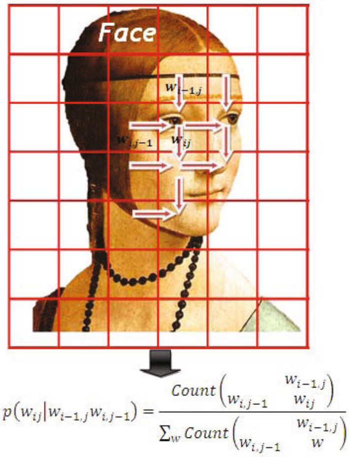

Face
$w_{i-1,j}$
$w_{i,j-1}$
w_{ij}
$p(w_{ij}|w_{i-1,j}w_{i,j-1}) = \frac{Count\begin{pmatrix} & w_{i-1,j} \\ w_{i,j-1} & w_{ij} \end{pmatrix}}{\sum_w Count\begin{pmatrix} & w_{i-1,j} \\ w_{i,j-1} & w \end{pmatrix}}$

Fig. 6. Process of trigram language model training

$$p(w_{ij}|w_{00}w_{01}\cdots w_{mn}) = p(w_{ij}) \tag{19}$$

$$p(w_{ij}|w_{00}w_{01}\cdots w_{mn}) = p(w_{ij}|w_{i-1,j}) \tag{20}$$

$$p(w_{ij}|w_{00}w_{01}\cdots w_{mn}) = p(w_{ij}|w_{i-1,j}w_{i,j-1}) \tag{21}$$

Where w_{ij} represents the visual word at Row i, Column j in the word matrix. In the following, we will discuss in details about the training process for the three models.

Unigram Model

For each concept, a unigram model characterizes the distribution of individual visual words under the concept. In the training process, we calculate $p(w_i|C_k)$ using

$$p(w_i|C_k) = \frac{F_N(w_i|C_k)}{\sum_{w\in V} F_N(w|C_k)}, k = 1, 2, \ldots, K \tag{22}$$

To avoid zero probability which would cause the classifier to fail, we assign a small prior probability to each unseen word in the concept. Accordingly, the amount of this prior probability should be discounted from the probabilities of the words appearing in the concept, so that the sum of probabilities is 1. The smoothed words distribution is represented by Eq. (23).

$$p(w_i|C_k) = \begin{cases} \frac{F_N(w_i|C_k)\times(1-\frac{1}{N})}{\sum_{w\in V} F_N(w|C_k)}, & F_N(w_i|C_k) > 0; \\ \frac{1}{NR}, & otherwise. \end{cases} \tag{23}$$

where N is the total number of words in the training set and R is the number of words that do not appear in class C_k. This probabilistic model tells how likely a word appears in an image belonging to that concept.

Bigram Model

Unlike the unigram model, a bigram model assumes that each visual word is conditionally dependent on its left neighbor. So the training process is to learn the conditional probability by Eq. (24).

$$p(w_{ij}|w_{i,j-1}, C_k) = \frac{F_N(w_{i,j-1}, w_{ij}|C_k)}{F_N(w_{i,j-1}|C_k)} \tag{24}$$

$w_{i,j-1}$ is the left neighbor of $w_{i,j}$ in the visual words matrix. However, the bigrams are usually quite sparse in the probability space. The maximum likelihood estimation is usually biased higher for observed samples and biased lower for unobserved samples. Thus smoothing technique is needed to provide better estimation of the infrequent or unseen bigrams. Instead of just assigning a small constant prior probability, we adopt more accurate smoothing method [12], which combines back-off and discounting [27].

$$p(w_{ij}|w_{i,j-1},C_k) = \begin{cases} \beta(w_{i,j-1}) \times p(w_{ij}|C_k), & F_N(w_{i,j-1}w_{ij}|C_k) = 0; \\ \hat{p}(w_{ij}|w_{i,j-1},C_k), & \textit{otherwise.} \end{cases} \tag{25}$$

$$\beta(w_{i,j-1}) = \frac{1 - \sum_{F_N(w_{i,j-1}w)>0} \hat{p}(w|w_{i,j-1},C_k)}{1 - \sum_{F_N(w_{i,j-1}w)>0} p(w|C_k)} \tag{26}$$

$$\hat{p}(w_{ij}|w_{i,j-1},C_k) = d_r \times \frac{F_N(w_{i,j-1}w_{ij}|C_k)}{F_N(w_{i,j-1}|C_k)} \tag{27}$$

Back-off method is represented in Eq. (25) (26), and discounting is represented in Eq. (27). If a bigram does not appear in the concept, back-off method is applied to calculate the bigram model from the unigram model. $\beta(w_{i,j-1})$ is called back-off factor. If bigram $w_{i,j-1}w_{ij} \in V$ appears in concept C_k, the discounting method is used to depress the estimation of its conditional probability. d_r is called the discounting coefficient.

$$d_r = \frac{r-b}{r} \tag{28}$$

$$b = \frac{n_1}{n_1 + 2n_2} \tag{29}$$

r is the number of times a bigram appears; and n_i is the number of visual words that appear i times in the concept.

Trigram Model

The above two modeling processes are almost the same as the statistical language models used in text classification, However, the trigram model for visual language is different from that for natural language processing. In text analysis, trigram is a sequence of three words $< w_{i-2}, w_{i-1}, w_i >$, while in visual document, which is a two dimensional matrix, we assume each word is conditionally dependent on its left neighbor and the word above it. So these three words form a trigram $< w_{i-1,j}, w_{i,j-1}, w_{ij} >$. The training process of a trigram model is illustrated in the following equation.

$$p(w_{ij}|w_{i-1,j},w_{i,j-1},C_k) = \frac{F_N(w_{i-1,j}w_{i,j-1}w_{ij}|C_k)}{F_N(w_{i-1,j}w_{i,j-1}|C_k)} \tag{30}$$

For the same reason with bigram model, discounting and back-off methods are also introduced for trigram model.

$$p(w_{ij}|w_{i-1,j}w_{i,j-1},C_k) = \begin{cases} \beta(w_{i-1,j}w_{i,j-1})p(w_{ij}|w_{i,j-1},C_k), & F_N(w_{ij}^3|C_k) = 0; \\ \hat{p}(w_{ij}|w_{i-1,j}w_{i,j-1},C_k), & \textit{otherwise.} \end{cases} \tag{31}$$

$$\beta(w_{i-1,j}w_{i,j-1}) = \frac{1 - \sum_{F_N(w_{i-1,j}w_{i,j-1}w)>0} \hat{p}(w|w_{i-1,j}w_{i,j-1}, C_k)}{1 - \sum_{F_N(w_{i-1,j}w_{i,j-1}w)>0} \hat{p}(w|w_{i,j-1}, C_k)} \tag{32}$$

$$\hat{p}(w_{ij}|w_{i-1,j}w_{i,j-1}, C_k) = d_r \times \frac{F_N(w_{i-1,j}w_{i,j-1}w_{ij}|C_k)}{F_N(w_{i-1,j}w_{i,j-1}|C_k)} \tag{33}$$

w_{ij}^3 represents the trigram $< w_{i-1,j}w_{i,j-1}w_{ij} >$. The spatial correlation between visual words is captured by the conditional probabilities of trigrams. In summary, the training procedure is as follows:

a. Divide each training image into patches;

b. Generate a hash code for each patch to form a visual document;

c. Build visual language models for each concept by calculating the conditional distribution of unigram, bigram and trigram.

The process of building trigram visual language model is illustrated in Fig. 6. It is worth noting that not all visual words are useful for classification. Therefore, we introduce a feature selection process during visual language training. Words are selected according to their term frequency (TF) and inverse document frequency (IDF).

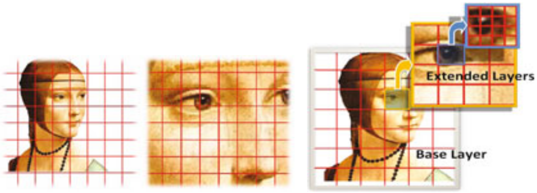
$$tf_i^k = \frac{n_i}{\sum_j n_j} \tag{34}$$

$$idf_i = \log \frac{|V|}{|d : w_i \in d|} \tag{35}$$

where tf_i^k measures the frequency a word w_i appears in images belonging to concept C_k and idf_i reflects the frequency of word w_i in the dataset. The $tf - idf$ weight $tf_i^k \times idf_i$ evaluates how important the word w_i is to concept C_k. For each concept, we select the words with $tf - idf$ weight bigger than a threshold. And the words from different concepts are combined. This approach can depress the influence of random background and reduce the size of vocabulary.

4.5 Scale Invarient Visual Language Model (m-VLM)

The VLM has efficiently captured the spatial dependence information, however it has some drawbacks. One of the biggest problems of the previous visual language model is the object scaling problem. The same object or scene with different scales may generate completely different visual matrixes. To make the model more resistant to scale variation of objects, multi-layer extension is introduced to the visual language model, denoted as scale invarient visual language model (m-VLM)[59]. Instead of extracting image patches of a single size, the model extract different sizes of patches from an image, which could



(a) The object scaling problem. (b) Image representation.

Fig. 7. Scaling problem and Multi-layer image representation

catch object characters in different scales. The visual language model built on these patches models the spatial co-occurrence relationship between them.

The basic idea of scale invariant modeling method is to train the language model based on various scales of patches; so that given any image, the words conditional distribution of the object region can be best matched. For a multi-layer patch representation, the patches on the same layer are of the same size, while those on a higher layer are twice of the size. For example, we use 8×8 for the first layer, 16×16 for the next, and so on. The first layer is called the base layer. Other layers are called extended layers as shown in Fig. 7. All these patches are transformed into the visual words in the same way as the monolayer language model training process.

For multi-layer visual language modeling method, three assumptions are made corresponding to multi-layer unigram model (m-unigram), multi-layer bigram model (m-bigram) and multi-layer trigram model (m-trigram) respectively.

Assumption 1. In m-unigram model, visual words on different layers are independent with each other.

Assumption 2. In m-bigram model, each visual word only depends on its left neighbor in the same layer.

Assumption 3. In the m-trigram model, each visual word depends on its left neighbor and the word above it in the same layer.

The training processes of these three models are formulated as the following three equations correspondingly. For Multi-layer unigram model,

$$p(w_1|C_k) = \begin{cases} \frac{\sum_{l=0}^{m-1} F_N(w_1|L_l,C_k)\times(1-\frac{1}{R})}{\sum_{w\in V}\sum_{l=0}^{m-1} F_N(w|L_l,C_k)}, & F_N(w_1|C_k) > 0; \\ \frac{1}{NR}, & otherwise. \end{cases} \tag{36}$$

For Multi-layer bigram model,

$$p(w_2|w_1, C_k) = \begin{cases} \beta(w_1) \times p(w_1|C_k), & F_N(w_1 w_2|C_k) > 0; \\ \hat{p}(w_2|w_1, C_k), & otherwise. \end{cases} \tag{37}$$

$$\hat{p}(w_2|w_1, C_k) = d_r \times \frac{\sum_{l=0}^{m-1} F_N(w_1 w_2|L_l, C_k)}{\sum_{l=0}^{m-1} F_N(w_1|L_l, C_k)} \tag{38}$$

For Multi-layer trigram model,

$$p(w_3|w_1w_2, C_k) = \begin{cases} \beta(w_1w_2)p(w_2|w_1, C_k), & F_N(w_1w_2w_3|C_k) > 0; \\ \hat{p}(w_3|w_1w_2, C_k), & otherwise. \end{cases} \tag{39}$$

$$\hat{p}(w_3|w_1w_2, C_k) = d_r \times \frac{\sum_{l=0}^{m-1} F_N(w_1w_2w_3|L_l, C_k)}{\sum_{l=0}^{m-1} F_N(w_1w_2|L_l, C_k)} \tag{40}$$

w_1, w_2, w_3 represent any three words in the vocabulary. L_l is the l^{th} layer. $\beta(\cdot)$ and d_r are the same definition as in monolayer VLM. m is the number of layers. For each document D in concept C, we divide it into m layers, and count n-grams on all layers. The parameter m is determined experimentally.

The multi-layer visual language modeling method has brought many favorable properties to VLM. Since the patches are of various sizes, the object scaling is no longer a problem with VLM. Moreover, m-VLM does not increase the computational time in classification phase. Since all additional computational cost are introduced in the training process, the classification process takes the same time as the monolayer VLM. During training, the models are built under various scales, however, we only need to extract mono-scale patches from a test image. So the classification process is exactly the same as the monolayer VLM classifier.

5 Conceptual Distance Measurement

In the previous section, we discussed the semantic level concept modeling methods. These modeling methods only can handle the representation of each single concept, however, in certain cases the relationship between concepts are also quite useful. In this section, we will further discuss a general concept relationship measurement, named conceptual distance measurement.

Here conceptual distance should measure four kinds of common conceptual correlations, synonym, similarity, meronymy and concurrence. Synonym denotes the same object with different names, like football and soccer. Similarity denotes that two concepts are visually similar, i.e. horse and donkey. Meronymy represents one concept is part of another, i.e. wheel and car. Concurrence denotes situations when two concepts appear simultaneously in daily life.

As far as we know, there are currently four kinds of conceptual distance measurements. Conceptual distance generation based on human labor, i.e. WordNet distance; conceptual distance based on Web textual information, i.e. Google distance; conceptual distance based on Web tag information, i.e. tag concurrence distance; and concept distance based on visual information, i.e. Flickr distance.

5.1 WordNet Distance

The first kind of conceptual distance is generated by human experts, such as the WordNet Distance [44], which was developed by the Cognitive Science

Laboratory of Princeton University. It is widely used to exploit semantic relationship of common concepts. Since WordNet is defined by human experts, the semantic distance based on WordNet is very close to human perception. However, it can only support a relatively limited number of concepts (around 150,000) comparing to the overall concepts on the Web. Also it is very expensive to extend the corpus in WordNet. For example, there are 10^9 different online tags on the photo sharing website like Flickr, and this number is still increasing every day. Since it is expensive to update the conceptual relations in database manually, WordNet can hardly catch up with the proliferation of concepts on the web.

5.2 Google Distance

The second kind conceptual distance is generated by Web textual information, such as Normalized Google Distance (NGD)[11]. NGD was proposed by Cilibrasi and Vitányi to measure the conceptual distance by the Google page counts when querying both concepts to the Google search engine. It actually reflects how often two concepts will co-occur in same web-pages. This distance costs little human effort and covers almost any concept on the Web.

$$NGD(x,y) = \frac{max(\log f(x), \log f(y)) - \log f(x,y)}{\log N - min(\log f(x), \log f(y))} \tag{41}$$

where $NGD(x,y)$ represents the Normalized Google distance between concepts x and y. $f(x)$, $f(y)$, and $f(x,y)$ denotes the number of pages containing x, y, both x and y. N is the total number of web pages indexed by Google. However, NGD assumes the conceptual relationship only depends on the co-occurrence of these concepts among online textual documents. This assumption is simple, and cannot cover the cases of meronymy and concurrence, as shown in Table 1.

5.3 Tag Concurrence Distance

The Tag Concurrence Distance (TCD)[36] directly applies the idea of Google distance to the tags. It treats the tag list of each image as a document, and calculates the TCD distance the same way as Google distance. This is a very intuitive way to measure the conceptual distance based on tags. The inverse of the mutual information between two tags can be further calculated as a distance of two different tags according to their association with the images. This distance reflects the frequency of two tags occurring in the same image. However, there is the sparseness problem that many correlations are missing due to the incompleteness of tags. In Flickr, there are less than 10 tags per image. It is far less than the number of words in a typical web document. Therefore, many semantic relations may not be covered by this Tag Concurrent Distance.

5.4 Flickr Distance

The fourth kind distance measurement is generated from the visual correlation, such as the proposed Flickr Distance (FD) [60], which generates the latent topic visual language model (LTVLM) to represent each tag by analyzing the visual words correlation to its related images.

To simulate the concurrence of concepts in human cognition, the calculation of conceptual correlation should be performed in daily life environment. To achieve this, FD try to mine the statistical semantic relations between concepts from a large pool of the daily life photos. To obtain a less biased estimation, the image pool should be very large and the source of the images should be independent. Luckily, the on-line photo sharing website Flickr meets both conditions. There are more than 10^9 photos on Flickr, and these photos are uploaded by independent users. Besides, each photo is manually tagged by the users, which provides well connections between the photos and the semantic concepts (tags), which makes Flickr images an ideal dataset for learning the visual conceptual relations. That is the reason for the name "Flickr distance".

The basic idea of Flickr distance is to calculate the concept distance by the divergence between the LTVLMs of the concepts. Each concept corresponds to a visual language model, which consists of the trigram conditional distributions under different latent topics. Kullback-Leibler divergence (KL divergence) is a common measurement of the difference between two probability distributions. However, as it does not meet the constraint of triangular inequality, it is in fact not a strict metric. Based on KL divergence, a more strict metric Jensen-Shannon (JS) divergence is defined. This divergence is symmetric and its square root is demonstrated a strict metric. The visual correlation is defined as the inverse of average square root of the JS divergence between the latent topic VLMs.

Let $P_{z_i^{C_1}}$ and $P_{z_j^{C_2}}$ be the trigram distributions under latent topic $z_i^{C_1}$ and $z_j^{C_2}$ respectively. $z_i^{C_1}$ represents the i^{th} latent topic of concept C_1. The K-L divergence between them is defined to be

$$D_{KL}(P_{z_i^{C_1}}||P_{z_j^{C_2}}) = \sum_l P_{z_i^{C_1}}(l) \log \frac{P_{z_i^{C_1}}(l)}{P_{z_j^{C_2}}(l)} \tag{42}$$

where $P_{z_i^{C_1}}(l)$,$P_{z_j^{C_2}}(l)$ correspond to the probability density of the l^{th} trigram in these two distributions respectively. In the view of information theory, the KL divergency is in fact a measurement of the mutual entropy between the two visual language models.

$$\begin{aligned} &D_{KL}(P_{z_i^{C_1}}||P_{z_j^{C_2}}) \\ &= -\sum_l P_{z_i^{C_1}}(l) \log P_{z_j^{C_2}}(l) + \sum_l P_{z_i^{C_1}}(l) \log P_{z_i^{C_1}}(l) \\ &= H(P_{z_i^{C_1}}, P_{z_j^{C_2}}) - H(P_{z_i^{C_1}}) \end{aligned} \tag{43}$$

where $H(P_{z_i^{C_1}}, P_{z_j^{C_2}})$ is the cross entropy of the two distributions, and $H(P_{z_i^{C_1}})$ is the entropy of $P_{z_i^{C_1}}$. According to the Gibbs' inequality, $D_{KL}(P_{z_i^{C_1}}||P_{z_j^{C_2}}) \geq 0$. It is zero if and only if $P_{z_i^{C_1}}$ equals $P_{z_j^{C_2}}$.

JS divergence is defined based on KL divergence to measure the distance metric between these visual language models (Eq. (44)).

$$D_{JS}(P_{z_i^{C1}}||P_{z_j^{C2}}) = \frac{1}{2}D_{KL}(P_{z_i^{C_1}}||M) + \frac{1}{2}D_{KL}(P_{z_j^{C_2}}||M) \tag{44}$$

$$M = \frac{1}{2}(P_{z_i^{C_1}} + P_{z_j^{C_2}}) \tag{45}$$

where M is the average of $P_{z_i^{C_1}}$ and $P_{z_j^{C_2}}$. It is demonstrated that the square root of the Jensen-Shannon divergence is a metric. Thus the Flickr distance between two concepts C_1 and C_2 is calculated as the average square root of the JS divergence between the latent topic VLM of concept C_1 and that of concept C_2.

$$D_{Flickr}(C_1, C_2) = \sqrt{\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{1}{K^2}D_{JS}(P_{z_i^{C_1}}||P_{z_j^{C_2}})} \tag{46}$$

More generally, the conditional distribution of topics within each concept can be used to weight the distance. This topic distribution is generated by the LTVLM.

$$D_{Flickr}(C_1, C_2) = \sqrt{\sum_{i=1}^{K}\sum_{j=1}^{K}P(z_i^{C_1}|C_1)P(z_j^{C_2}|C_2)D_{JS}(P_{z_i^{C_1}}||P_{z_j^{C_2}})} \tag{47}$$

Thus the visual correlation (VC) between two concepts C_1 and C_2 is inverse to the Flickr distance as Eq.(48), where δ is a very small constant.

$$VC(C_1, C_2) = \frac{1}{\delta + \sqrt{D_{Flickr}(C_1, C_2)}} \tag{48}$$

6 Applications

Previously, we have discussed the low level visual analysis (image representation), image distance measurement (inter-image representation), semantic level concept modeling (concept representation), and conceptual distance measurement (inter-concept representation). In this section, we will discuss the utility of these models and measurements in real world applications.

Table 1. Illustration of NGD, TCD and FD. The first column shows the conceptual relationship. The second column listed some conceptual pairs. The third to the fifth columns are the conceptual distance under different measurements. Improper scores are marked in bold.

Relationship	Conceptual pair	NGD	TCD	FD
None	Airplane–dog	0.2562	0.8532	0.5151
Synonym	Football–soccer	0.1905	0.1739	0.0315
Similarity	Horse–donkey	0.2147	0.4513	0.4231
Concurrence	Airplane–airport	**0.3094**	0.1833	0.0576
Meronymy	Car–wheel	**0.3146**	**0.9617**	0.0708

6.1 Near Duplicate Image Detection

Visual distance measurement can be widely used in multiple applications. One of its most common usages is the near duplicate image detection. In this subsection, we focus on applying both the static image distance measurement and the dynamic image distance measurement to the near-duplicate detection task. We take Bin's method [56] as an example to illustrate the static image distance measurement, and take QOSS [62] as an example to illustrate the usage of dynamic image distance measurement.

With the advances of web technology, the diffusion of web images has increased exponentially. This greatly aggravates the problem of image copyright infringement. Although watermark schemes have been proposed [19] to protect the copyrighted images and trademarks, this kind of protection will become inefficacy when the content of the copyrighted image is slightly modified and then republished. To detect these slightly modified images, which is also called near-duplicates, the content-based image replica recognition scheme is proposed [48]. Given a copyrighted image as a query, the task is to find all the accessible duplicates and near-duplicates on the web.

The main issues with the near-duplicates detection focus on two aspects, efficient image features and similarity measurement. Considering the efficiency, most features used in the large-scale near-duplicates detection task are simple, such as mean gray, color histogram, texture histogram etc. To measure the similarity, many distance functions are proposed, i.e. Minkowski-like metrics, Histogram Cosine distance, Fuzzy logic etc. However, these methods frequently overlook the near-duplicate images. Later, some advanced methods are proposed, such as [28, 66]. Although these methods are reasonable, they are not efficient enough for large-scale near-duplicates detection.

Static distance for near duplicate detection

Recently, Bin et al. [56] proposed a large-scale duplicates detection algorithm. This method divides the image into patches, and uses the mean gray of each patch as the feature. The hash code is generated from the most distinguishing feature dimensions picked by principle component analysis (PCA) to facilitate fast similarity comparison. Hamming distance is adopted for similarity measurement. This algorithm is reported efficient and still capable to maintain high precision. Yet, as the distinguishing features picked by PCA only characterize the whole dataset, the specific property of the query image is not well utilized.

Dynamic distance for near duplicate detection

Considering both the efficiency and effectiveness, the whole approach consists of two phases, offline indexing and online detection. In the offline indexing phase, the main objective is to provide efficient index of the whole dataset. To achieve this, we transform each image in the database into a low dimensional feature vector, which can be further represented as a compact hash code. The PCA projection matrix for feature dimension reduction is generated in advance from a static sufficiently large image collection. In the online detection phase, we aim at improving the effectiveness of the method without too much cost of efficiency. For this reason, firstly a rough filtering is performed based on the fast hash code matching to remove the major proportion of non-duplicates. Then on the relatively small remaining set, the proposed iterative subspace shifting algorithm is used to refine the roughly filter.

Suppose the query image is $q \in R^n$. An image I_j is believed relevant to the query image if and only if

$$\|H(Pq) - H(PI_j)\|_{\|} < \epsilon$$

where P is the static projection matrix. $H(\bullet)$ is the hash coding function. κ represents the corresponding subspace, and ϵ is the threshold to determine whether the image is close to the query or not in the subspace. The set of samples which are close to the query are called *query surrounding samples.* All the query surrounding images form the *query surrounding collection* Q_s.

In order to determine the loose threshold ϵ for rough filtering, several random transformations are generated from each query image and represented in hash code in the same subspace projected with the static PCA matrix. The largest Hamming distance between the query and its transformations is set as the threshold.

$$\epsilon = \max_l \|Pq_j - Pq_j^{(l)}\|_\kappa \tag{49}$$

where $q_j^{(l)}$ is the l^{th} random transformation of the query image q_j.

Since the hash code matching has provided a much smaller query surrounding collection, we can use an iterative scheme to detect the near-duplicates from this collection. For each iteration, PCA eigenspace of the query surrounding samples is selected as the optimal subspace for measuring the similarity among the query surrounding samples. This subspace keeps as much of the variance of the collection as possible. The remote samples will then be excluded from the query surrounding collection. As the collection is updated, the eigenspace will of course shift. So in the next iteration, the similarity measurement will be performed in another eigenspace. It is more probably that the near-duplicates would remain close to the query after the subspace has shifted, while non-duplicated images which may form a cluster in a previous subspace will scatter in the subsequent spaces.

Comparison between static and dynamic distances

In order to facilitate the comparison, we compare QOSS with the following methods. G-HC represents Bin's hash coding approach [56] with gray feature. T-HC denotes Bin's approach using the texture histogram feature. QOSS is the query oriented subspace shifting algorithm with texture histogram feature. The results are given in Table 2.

Table 2. Near-duplicates detection performance

Methods	G-HC	T-HC	QOSS
Precision	96.57	96.82	96.85
Recall	69.97	73.15	90.34

Table 2 shows that comparing with static distance metric (Bin's method), the QOSS method has greatly improved the recall while keeps similar precision. For Bin's method, the similarity measure is done in a single subspace. In order to keep relatively high precision, the near-duplicate criterion should be strict. Even some near-duplicates may not follow. For the proposed method, the similarity is measured on multiple subspaces iteratively, and in each subspace the criterion may not necessarily be strict to maintain high precision.

6.2 Conceptual Clustering

In this application scenario, we apply the Flickr distance based visual correlation on conceptual clustering. Conceptual clustering is widely used for topic detection and summarization. Since there are lots of tags and descriptions associated with web images, we are able to use conceptual clustering to detect the main topic of these images. However, the focus of the topic summarization in image may not be the same with that for the text. For example, the image is more likely to focus on the main object or scene, while in the text document it focuses more on the story or point of view of the author. Thus an applicable conceptual distance measurement for textual domain, i.e. Google distance, may not perform as well as the specific distance measurement for visual domain. Here we compare the conceptual clustering results of NGD, TCD, and FD.

Three groups of concepts are selected: Space related terms (4 concepts), Ball games (10 concepts), and Animals (9 concepts). We choose these concepts because all users agree that these concepts are grouped without ambiguous in the user study. In total there are 23 concepts in the experiment.

The task is to group these concepts into clusters automatically. One of the key issues with conceptual clustering is the conceptual distance measurement. In most cases, WordNet is used to measure the conceptual distances. However, due to the limitation of WordNet lexicon, a large portion of concepts, i.e. famous movies, brand, game, sport star, singer, etc. are inextricable. In this experiment, we build two different networks between these concepts with Google distance and Flickr distance separately. Based on these two conceptual networks, spectral clustering is adopted to generate the conceptual clusters. We adopt spectral clustering rather than the commonly used K-means algorithm, because it is hard to calculate the cluster centers of these concepts in K-means algorithm, while the spectral clustering only use the relationship between the samples. The results of the conceptual clusters are shown in Table 3.

Table 3. Result of conceptual clustering. The bold font denotes the miss-clustered concepts.

Clustering by NGD			Clustering by TCD			Clustering by FD		
bears	**bowling**	baseball	moon	baseball	basketball	moon	bears	baseball
horses	dolphin	basketball	space	**donkey**	**bears**	saturn	dolphin	basketball
moon	donkey	football	Venus	softball	bowling	space	donkey	football
space	**saturn**	golf	**whale**	**wolf**	**dolphin**	venus	**golf**	**snake**
-	sharks	soccer	-	-	football	-	horses	soccer
-	snake	tennis	-	-	golf	-	sharks	bowling
-	**softball**	volleyball	-	-	**horses**	-	spiders	softball
-	spiders	-	-	-	**Saturn**	-	**tennis**	volleyball
-	turtle	-	-	-	**sharks**	-	turtle	-
-	**venus**	-	-	-	soccer	-	whale	-
-	whale	-	-	-	**spiders**	-	wolf	-
-	wolf	-	-	-	tennis	-	-	-
-	-	-	-	-	**turtle**	-	-	-
-	-	-	-	-	volleyball	-	-	-

Table 3 shows that the Flickr distance based spectral clustering can effectively generate the conceptual clusters. After the conceptual clustering, we know the three categories of images are about space, animals, and ball games. The errors are marked in **bold** in Table 3. 6 out of the 23 total concepts are mistakenly clustered by NGD; 9 errors for TCD; and 3 errors in the result of FD. Comparing with the clustering results based on NGD and TCD, the results by FD is more promising.

6.3 Social Tagging Recommendation

Recently, social tagging is one of the most promising approaches for web image annotation and indexing. Social tagging requires all the users in social network label the web resources with their own keywords and share with others. This labeling operation is named "tagging". Different from ontology based annotation; there is no pre-defined ontology or taxonomy in social tagging. Thus this task is more convenient for ordinary users.

Although social tagging is easy to perform, the user created tags contain much noise, such as ambiguous tags, misspelling tags, improper tag, due to the lack of effective recommendation. Tag recommendation will provide the related or more proper tags for the users to choose in tagging an image. The quality of tag recommendation is quite important [2] to final quality of tagging, since one reason users do not tag properly is because they cannot think of any proper tags [52].

One of the critical problems in social tagging recommendation is the correlation measurement between the tags. Current recommendation is only based on the tag co-occurrence, while visual correlation is ignored in the recommendation process. In this experiment, we would like to demonstrate the usefulness of Flickr distance in the tag recommendation task.

Given the image and some of its initial user created tags, we would like to recommend a list of related tags which may be also applicable to the image. We denote the set of initial tags as $\mathcal{OT}$, and the set of remaining tags as $\mathcal{UT}$. The relevance of the tags is represented in two domains. The average tag co-occurrence in Flickr dataset is deemed as text domain correlation, and the Flickr distance between tags is deemed as a measurement of visual domain correlation. Then for these domains, we generate several ranking features $\{f_l\}_{l=1}^{3n}$ (n is the number of initial tags). The first n ranking features are based on NGD, and the following n by TCD, and the last n features by FD.

$$f_l(t_i, t_l) = W_{NGD}^s(t_i, t_l),\ t_l \in \mathcal{OT},\ t_i \in \mathcal{UT},\ l = 1, \cdots, n \tag{50}$$

$$f_{n+l}(t_i, t_l) = W_{TCD}^s(t_i, t_l),\ t_l \in \mathcal{OT},\ t_i \in \mathcal{UT},\ l = 1, \cdots, n \tag{51}$$

$$f_{2n+l}(t_i, t_l) = W_{FD}^s(t_i, t_l),\ t_l \in \mathcal{OT},\ t_i \in \mathcal{UT},\ l = 1, \cdots, n \tag{52}$$

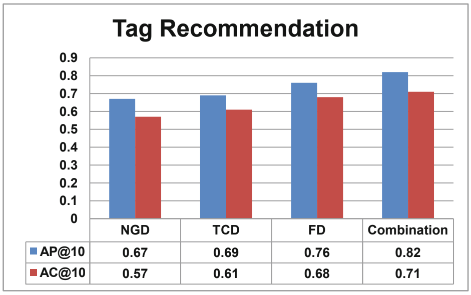
where W_{NGD}^s, W_{TCD}^s, W_{FD}^s are the weights of t_i, t_l in the corresponding conceptual network. These ranking features of multi-domains are firstly used separately and then combined in the Rankboost framework [18], which considers only the order of instances and not scores, to generate the most related keywords for social tagging. Performance under different correlation features and their combinations are shown in Fig.8.

We randomly select 10,000 images and associated 5,000 tags from Flickr as the test data, and the rest images and associated tags are used as the training data. To train the tag co-occurrence model, we count the co-occurrence frequency between every pair of tags from the collection, and then normalize them to $[0, 1]$. In the visual domain, we generate the ranking features based on Flickr distance. Then we combine these weak rankers form different domains using the Rankboost algorithm.

Each recommendation approach will generate an ordered list of relevant tags for each image. Then a group of volunteers are required to evaluate these recommended tags. If a tag is relevant to the image, it will be marked true; otherwise false. The average precision of the top 10 recommendations and the coverage over all correct recommendations are adopted to measure the performance of each recommendation method. The coverage is defined as the proportion of correct tags (including all correct tags by both methods and initial tags) that are recommended by the specific method. We adopt the coverage rather than the recall, because the recall is inapplicable for the recommendation task.

$$Coverage(m_i) = \frac{\#correct\,tags\,by\,method\,m_i}{\#correct\,tags\,in\,total} \tag{53}$$

Figure 8 compares the performance of the three methods. In Figure 8, "AP@10" is the average precision at top 10 recommendations, and "AC@10" is the average coverage at top 10 recommendations. We find FD outperforms NGD by 13.46% in precision and 19.24% in coverage, and outperforms TCD by 9.83 in precision and 11.43 in coverage. These results demonstrate the effectiveness of the Flickr distance in tag recommendation task.



	NGD	TCD	FD	Combination
■ AP@10	0.67	0.69	0.76	0.82
■ AC@10	0.57	0.61	0.68	0.71

Fig. 8. Performance of the tag recommendation. The Flickr distance based tag recommendation outperforms Google distance based recommendation by 16% and outperforms tag concurrence based recommendation by 14%.

7 Summary

In this chapter, we focus on four aspects of visual concept analysis, low level visual analysis (image representation), image distance measurement (inter-image representation), semantic level concept modeling (concept representation), and conceptual distance measurement (inter-concept representation).

These four aspects also forms two layer of analysis. The first layer is the visual layer, which focuses on image representation and image distance measurements. The second layer is the concept layer, which focuses on the concept representation and concept distance measurement. In the first layer, we discussed the visual feature, visual words, visual similarity measurements. In the second layer, we concentrate on the semantic level concept modeling methods, such as BoW model, 2D HMM model, and VLM. Based on these models, we further investigate several concept distance measurements, including WordNet distance, Google distance, tag concurrence distance, and Flickr distance. Various applications related to multimedia research have shown the usage of these analysis methods, models and distance measurements.

References

1. Agarwal, A., Triggs, B.: Hyperfeatures – multilevel local coding for visual recognition. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 30–43. Springer, Heidelberg (2006)
2. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: CHI 2007 (2007)
3. Opelt, A., Pinz, A., Zisserman, A.: A boundary-fragment-model for object detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 575–588. Springer, Heidelberg (2006)
4. Babenko, B., Yang, M.-H., Belongie, S.: Visual Tracking with Online Multiple Instance Learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009 (2009)
5. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning a mahalanobis metric from equivalence constraints. Journal of Machine Learning Research 6, 937–965 (2005)
6. Baumberg, A.: Reliable feature matching across widely separated views (2000)
7. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
8. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3(5), 993–1022 (2003)
9. Bretzner, L., Lindeberg, T.: Feature tracking with automatic selection of spatial scales. Comput. Vis. Image Underst. 71(3), 385–392 (1998)
10. Cao, L., Fei-Fei, L.: Spatially coherent latent topic model for concurrent object segmentation and classification. In: Proc. of ICCV 2007 (2007)
11. Cilibrasi, R., Vitanyi, P.M.B.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19, 370 (2007)
12. Clarkson, P., Rosenfeld, R.: Statistical language modeling using the CMU–cambridge toolkit, pp. 2707–2710 (1997)
13. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proc. of ICML 2007, Corvalis, Oregon, pp. 209–216 (2007)
14. Duan, L., Tsang, I.W., Xu, D., Maybank, S.J.: Domain Transfer SVM for Video Concept Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009 (2009)

15. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: Proc. of CVPR 2004, vol. 12, p. 178 (2004)
16. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2005)
17. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: Proc. of ICCV 2005, vol. 2, pp. 1816–1823 (2005)
18. Freund, Y., Iyer, R., Schapire, R.E., Singer, Y., Dietterich, G.: An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 170–178 (2003)
19. Garcia-molina, H., Ketchpel, S.P., Shivakumar, N.: Safeguarding and charging for information on the internet. In: Proc. of ICDE 1998 (1998)
20. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. pp. 518–529 (1999)
21. Han, W., Brady, M.: Real-time corner detection algorithm for motion estimation. Image and Vision Computing, pp. 695–703 (November 1995)
22. Harris, C., Stephens, M.: A combined corner and edge detection. In: Proceedings of The Fourth Alvey Vision Conference, pp. 147–151 (1988)
23. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval, Berkeley, California, August 1999, pp. 50–57 (1999)
24. Hoi, S.C.H., Liu, W., Lyu, M.R., Ma, W.-Y.: Learning distance metrics with contextual constraints for image retrieval. In: Proc. of CVPR 2006 (2006)
25. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: STOC 1998: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, pp. 604–613. ACM Press, New York (1998)
26. Goldberger, G.H.J., Roweis, S., Salakhutdinov, R.: Neighbourhood components analysis. In: Advances in Neural Information Processing Systems (2005)
27. Katz, S.: Estimation of probabilities from sparse data for the language model component of a speech recognizer
28. Ke, Y., Sukthankar, R., Huston, L., Ke, Y., Sukthankar, R.: Efficient near-duplicate detection and sub-image retrieval. In: Proc. of ACM Multimedia 2004, pp. 869–876 (2004)
29. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
30. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009 (2009)
31. Lazebnik, S., Schmid, C., Ponce, J.: A maximum entropy framework for part-based texture and object recognition. In: Proc. of ICCV 2005, pp. 832–838 (2005)
32. Li, J., Najmi, A., Gray, R.M.: Image classification by a two dimensional hidden markov model. IEEE Trans. Signal Processing 48, 517–533 (1998)
33. Li, J., Wang, J.: Automatic linguistic indexing of pictures by a statistical modeling approach
34. Li, S.Z.: Markov Random Field Modeling in Image Analysis. Springer, New York (2001)

35. Lindeberg, T.: Scale-Space Theory in Computer Vision. Kluwer Academic Publishers, Norwell (1994)
36. Liu, D., Hua, X.-S., Yang, L., Wang, M., Zhang, H.-J.: Tag ranking. In: Proc. of World Wide Web 2009, WWW 2009 (2009)
37. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. of ICCV 1999, pp. 1150–1157 (1999)
38. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
39. Maji, S., Malik, J.: Object Detection using a Max-Margin Hough Transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009 (2009)
40. Marée, R., Geurts, P., Piater, J., Wehenkel, L.: Random subwindows for robust image classification. In: Proc. of CVPR 2005, vol. 1, pp. 34–40 (2005)
41. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: British Machine Vision Conference, vol. 1, pp. 384–393 (2002)
42. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 128–142. Springer, Heidelberg (2002)
43. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. Int. J. Comput. Vision 60(1), 63–86 (2004)
44. Miller, G.A., et al.: Wordnet, a lexical database for the english language. Cognition Science Lab. Princeton University, Princeton (1995)
45. Moravec, H.P.: Obstacle avoidance and navigation in the real world by a seeing robot rover. PhD thesis, Stanford, CA, USA (1980)
46. Otluman, H., Aboulnasr, T.: Low complexity 2-d hidden markov model for face recognition. In: Proc. of IEEE Conference on International Symposium on Computer Architecture (2000)
47. Paul Schnitzspan, S.R. B.S., Fritz, M.:
48. Qamra, A., Meng, Y.: Enhanced perceptual distance functions and indexing for image replica recognition. IEEE Trans. Pattern Anal. Mach. Intell. 27(3), 379–391 (2005); Senior Member-Chang, Edward Y
49. Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Mei, T., Zhang, H.-J.: Correlative multi-label video annotation. In: Proc. of ACM Multimedia 2007 (2007)
50. Russell, B.C., Freeman, W.T., Efros, A.A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006), pp. 1605–1614 (2006)
51. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: Proc. of CVPR 2006, pp. 2033–2040 (2006)
52. Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J.: Tagging, communities, vocabulary, evolution. In: CSCW 2006 (2006)
53. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: Proc. of ICCV 2005, pp. 370–377 (2005)
54. Smith, S.M., Brady, J.M.: Susan - a new approach to low level image processing. International Journal of Computer Vision 23, 45–78 (1995)

55. Tomboy, T.H., Bar-hillel, A., Weinshall, D.: Boosting margin based distance functions for clustering. In: Proc. of ICML 2004, pp. 393–400 (2004)
56. Wang, B., Li, Z., Li, M., Ma, W.-Y.: Large-scale duplicate detection for web image search. In: Proc. of ICME 2006 (2006)
57. Wang, J.Z., Li, J., Wiederhold, G.: SIMPLIcity: Semantics-sensitive integrated matching for picture LIbraries. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(9), 947–963 (2001)
58. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems, pp. 1473–1480 (2006)
59. Wu, L., Hu, Y., Li, M., Yu, N., Hua, X.-S.: Scale-invariant visual language modeling for object categorization. IEEE Transactions on Multimedia 11(2), 286–294 (2009)
60. Wu, L., Hua, X.-S., Yu, N., Ma, W.-Y., Li, S.: Flickr distance. In: MM 2008: Proceeding of the 16th ACM International Conference on Multimedia, pp. 31–40. ACM, New York (2008)
61. Wu, L., Li, M., Li, Z., Ma, W.-Y., Yu, N.: Visual language modeling for image classification. In: Proc. of 9th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR 2007 (2007)
62. Wu, L., Liu, J., Li, M., Yu, N.: Query oriented subspace shifting for near-duplicate image detection. In: Proc. of ICME 2008 (2008)
63. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems (2002)
64. Xu, D., Chang, S.-F.: Video event recognition using kernel methods with multilevel temporal alignment. IEEE Trans. Pattern Anal. Mach. Intell. 30(11), 1985–1997 (2008)
65. Yuan, Y.Y.M., Wu, J.: Discovery of collocation patterns: from visual words to visual phrases. In: Proc. of CVPR 2007 (2007)
66. Zhang, D.-Q., Chang, S.-F.: Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In: Proc. of ACM Multimedia 2004, pp. 877–884. ACM, New York (2004)

Error Propagation and Distortion Modeling in Loss-Affected Predictive Video Coding

Jacob Chakareski

Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
jakov.cakareski@epfl.ch

Summary. The highly complex prediction dependency structure employed in current video coding algorithms makes the resulting compressed bitstreams highly susceptible to data loss or corruption. Caused by transmission over imperfect communication channels or faulty storage devices these errors propagate then into other segments of the video bitstream thereby causing wild variations in quality of the reconstructed content. This chapter reviews the state-of-the-art in modeling the above error propagation phenomenon in predictive video coding and the resulting increase in video distortion across the affected media presentation. We focus in greater detail on the most important recent advances in packet-based distortion estimation techniques and examine some of the most interesting related discoveries. We show that video distortion is not only affected by the amount of data lost but also by the spatio-temporal distribution of the affected data segments. Furthermore, we illustrate cases where contrary to common belief subsequent packet loss actually leads to a reduction in video distortion and where surprisingly increased burstiness of the loss process again contributes to a smaller drop in video quality.

Keywords: error propagation, distortion modeling, video coding, lossy transmission, packet loss, burst packet loss errors, Markov models.

1 Introduction

Video compression has enabled a plethora of novel multimedia applications. From entertainment (DVD, IPTV, VoD), to video monitoring and surveillance for commercial and scientific applications, to video conferencing and telecommuting, they have all benefited man and society in numerous ways. In fact, many aspects of our personal and professional lives have been profoundly affected by the proliferation of digital content made possible by video coding.

The continuous demand for ever more efficient video compression has been matched by a consistent increase in complexity and computational capability of the related algorithms introduced to deliver it [1–7]. However, as the number of redundant bits removed from the multimedia data has been steadily

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 719–758.
springerlink.com

growing so has the amount of interdependency between the segments comprising the thereby created compressed content. This can be attributed in major part to data prediction techniques that every subsequent video coding standard has been increasingly taking advantage of.

In particular, as early as 1929 [8] it has been recognized that substantial compression gains can be achieved if a video picture[1] is differentially encoded[2] relative to its nearest temporal predecessor. The modern video coding community has generalized this concept by introducing the notion of a macroblock, a square region of a video frame, that is differentially encoded with respect to its most similar[3] companion macroblock in the previous video frame. This approach has been known under the joint name of motion estimation and motion compensation in video compression parlance [9, 10] referring to the fact that it accounts for the motion of an image segment between two successive video frames. Further compression efficiency can be achieved by simultaneously employing multiple reference macroblocks in previous and subsequent pictures for the macroblock being currently encoded in the present picture. The term multi-hypothesis motion compensation [11] has been coined for this generalization of the technique. Finally, a macroblock can also be differentially encoded relative to similar macroblocks within the same picture, a technique denoted intra-prediction and introduced in the latest video compression standard H.264 AVC [6].

Though the techniques highlighted above provide significant performance advances by taking advantage of the spatio-temporal correlation of the video content at the same time they make the compressed bitstream vulnerable to data loss or corruption[4]. In particular, the complex set of data dependencies that these techniques employ impose a strict decoding order that must be followed when reconstructing the content. In other words, the bitstream corresponding to a macroblock cannot be decoded unless all of its reference macroblocks in the data prediction technique have been decoded previously. Therefore, an error introduced in a decoded macroblock[5] can propagate to all other macroblocks that will be subsequently decoded and that have used this macroblock as a reference at compression. This error propagation phenomenon can strongly degrade the video quality of the thereby reconstructed content causing erratic fluctuations of its value over long periods of time.

Consider for example the illustration in Figure 1 showing a generic dependency graph between the data units (packets) comprising the compressed

[1] A video frame in video coding terminology.

[2] Simply described, rather than compressing the signals associated with the two pictures independently, we encode instead the first signal and its difference relative to the second one.

[3] According to some similarity metric.

[4] Caused by transmission over imperfect channels or faulty storage devices.

[5] Caused either by the loss of the original macroblock and its subsequent concealment by another already decoded macroblock or by erroneously reconstructing the original macroblock due to corrupted content.

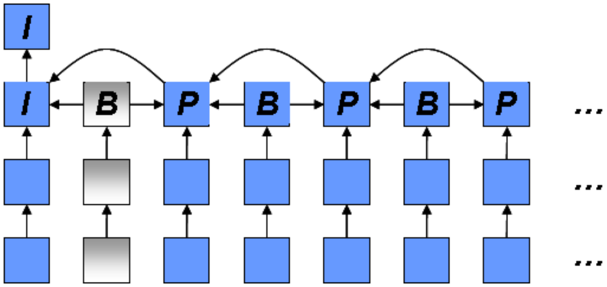


Fig. 1. Dependency graph for data units of a media presentation

representation of a video content. These units can correspond to the individual compressed macroblocks in the bitstream, to groups of compressed macroblocks (in compression terms denoted as slices), or to the compressed data of the individual video frames. Each node in the graph represents a data unit, and an arrow from data unit l to data unit l' in the graph signifies that for decoding data unit l, data unit l' must be decoded first. Then, for instance, an error affecting the first B-packet, denoted in shaded gray in Figure 1, will propagate to all its descendants in the graph, also denoted in shaded gray in the figure.

Now, computing the additional distortion affecting a predictively encoded content in the case of packet loss is a challenging task. First, the extensive use of differential encoding, as described earlier, makes tracking the propagation of error very difficult. In addition, there are other non-linearities involved in the process of compressing and decoding the content, such as quantization and filtering, that augment the complexity of the task even further. Finally, error concealment, the process of replacing data missing at reconstruction with other already available data, also needs to be carefully taken into consideration as it substantially impacts the distortion assessment.

Still, having an accurate estimate of packet-loss induced video distortion is very important as we increasingly send multimedia content over our data networks. In particular, it is estimated that online video accounts for 33% of the network traffic today. Furthermore, the networked video portion is expected to skyrocket to 90% by year 2012 [12]. Therefore, it is crucial for network operators and content providers to be able to assess the video quality of the content that they deliver over their networks without any involvement on the consumer side. That is because a solution involving measurements on the customer's premises is impractical from several aspects: (i) The original content needs to be present and compared to what the client is actually receiving, (ii) Some customers may not be willing for such an arrangement, seeing it as interference with their privacy, and (iii) The scalability of this approach can be an issue as well. Hence, content and service providers would strongly prefer to have a method of determining video quality of the content they serve either on the sending end or within the network, as a function of network quality parameters, such as packet loss. An accurate video quality estimation technique would help them determine the appropriate course of action with respect to their network management operations in order to provide a consistent content quality to their customers. Similarly, accurate distortion models are at the root of every efficient technique for video communication, e.g., [13–17]. Having such models improves the error resilience of such schemes while simultaneously allowing them to compute effective resource allocation decisions.

Early works on video distortion modeling considered that in the case of transmission over packet lossy networks the effects of the individual losses on the resulting degradation of the video content are additive [18, 19]. In particular, each packet loss during transmission contributes independently to the overall video distortion affecting the reconstructed stream. Though simple and computationally easily tractable, such models do not take into account the complex interaction between the error processes associated with the individual lost packets. Due to the many non-linearities involved in creating the compressed content, as described earlier, these loss processes exhibit a very complex set of interdependencies that are impossible to capture with an additive model.

Consider for example the propagation of additional distortion into subsequent frames by the loss of video frame k of an encoding created by predictively encoding every frame only with reference to its nearest temporal predecessor. As illustrated in Figure 2a, the increase in distortion associated with frame k is the largest since this is the content that will be replaced (concealed) by frame $k-1$ by the decoder at reconstruction. The decoder then proceeds by decoding the subsequent frames that have been correctly received. However, due to the predictive encoding of these frames, with reference to the lost frame k, they cannot be decoded error-free as the decoder is using a different reference $(k-1)$ now. This propagation of error due to incorrect reference frame manifests itself as additional distortion affecting frames $k+1, k+2, \ldots$, as illustrated in Figure 2a. The reasons why the additional distortion decays as we move further away from the lost frame k will be explained later on and are irrelevant for the discussion here.

Now, let the total increase in distortion affecting the video stream associated with the loss of frame k be denoted as $D(k)$. Consider next the case of losing two video frames k_1 and k_2. The resulting increase in video distortion per video frame of the reconstructed content is illustrated in Figure 2b. Here, the loss of frame k_1 is concealed by frame k_1-1 by the decoder who then proceeds reconstructing the content. The decoder then encounters another lost frame, $k_2 > k_1$, and replaces this missing content with the previously reconstructed frame k_2-1. Note that the content corresponding to the original frame k_2-1 is not the same now, as the frame indexed with k_2-1 at the decoder has been reconstructed using an incorrect reference in the past, i.e., frame k_1 has been replaced with frame k_1-1. Hence, in the case of

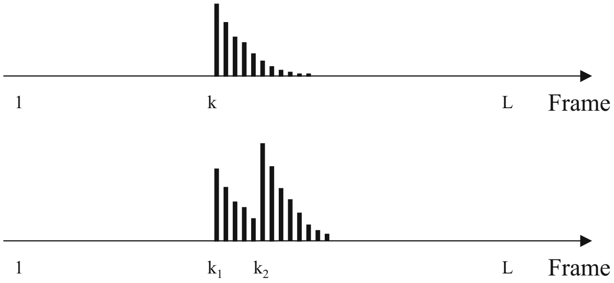


Fig. 2. (a) Loss of single frame k induces distortion in later frames. $D(k)$ is the total distortion summed over all affected frames. (b) $D(k_1, k_2)$ is total distortion summed over all frames caused by losing frames k_1 and k_2.

frames $k_2, k_2 + 1, \ldots$ we have the influence of two error events at frames k_1 and k_2 combining together, as illustrated in Figure 2b. Therefore, the overall increase in distortion $D(k_1, k_2)$ associated with the loss of the pair k_1, k_2 cannot be written as the sum of the contributions of the individual losses, i.e., $D(k_1) + D(k_2)$.

To the best of our knowledge, the first work that ventured to explore beyond superposition or linear video distortion models of the type described earlier is [20]. In particular, the authors recognized that the packet loss pattern and the loss burst length are important for an accurate estimation of video distortion in communication over lossy packet networks. The authors develop a distortion model describing the cross-correlation between different error events as a function of a few system parameters. Subsequently, the work in [21] proposed an alternative characterization of the video distortion as a function of the underlying packet loss process. Specifically, the authors develop a distortion model inspired by the concept of Markov Chains from statistics that captures the memory phenomenon of the distortion-loss process associated with lossy video transmission. Finally, [22] is the most recent relevant contribution to the field of non-linear video distortion modeling. The authors in this work employ a trellis model in order to account not only for the effect of error propagation but also for the fact that different packet loss events will typically have different likelihoods of occurrence. The expected video distortion can then be accurately synthesized by traversing the trellis in a left-to-right manner generating along the way the different prospective error patterns and their associated probabilities.

The works cited above provided some very interesting discoveries and insights that were not apparent before through the additive models. Therefore, in the subsequent three sections of this chapter we will respectively examine each one of them in greater detail. In particular, we will investigate their main characteristics of operation examining along the way their performance advantages and respective weaknesses. Implementation aspects and complexity will also be taken into consideration in the analysis. Then, in Section 5 we will provide a comprehensive and in greater depth discussion of related work, covering both linear and non-linear distortion models. Finally, a set of concluding remarks and a summary of all exciting challenges in distortion modeling that still remain open are included at the end in Section 6. We believe that the three important works covered in major part here will pave the way for subsequent new investigations that could help us explore further the complex interaction between video compression and packet loss.

2 Does Burst-Length Matter?

As mentioned earlier, the work in [20, 23] was the first to recognize the importance of characterizing the correlation between the error processes associated with individual packet losses. In order to describe the modeling framework proposed by the authors, we need to introduce some preliminaries and notation first.

The authors consider the case where each video frame is predictively encoded (P-frame) relative to its nearest temporal predecessor save for the very first frame in the content that is independently encoded (I-frame). For increased error-resilience it is assumed that a number of macroblocks is intra-encoded in every consecutive P-frame. Let the corresponding intra-refresh period be denoted as N[6]. It is assumed further that every $P - frame$ is mapped to a single transmission packet. Therefore, a loss of a packet corresponds to the loss of an entire frame.

For convenience, the video signal associated with a frame is considered as a 1-D vector of size $M = M_1 \times M_2$ pixels, where M_1 and M_2 are the number of rows and columns, respectively, of the pixel image matrix associated with a video frame. Now, let $f[k], \hat{f}[k], g[k]$ denote respectively the original video signal of frame k, its error free reconstruction, and its reconstructed value at the decoder in the case of loss concealment. The initial error frame caused by the loss of frame k is defined as

$$e[k] = g[k] - \hat{f}[k] \, .$$

Assuming that $e[k]$ is a zero-mean random process its variance is then equal to the Mean Square Error (MSE) associated with the loss of frame k, i.e.,

$$\sigma^2[k] = e^{\mathrm{T}}[k] \cdot e[k] \, / \, M \ .$$

The authors measure the quantities described above for every video frame by simulating the corresponding error events at the encoder employing the

[6] This is the number of subsequent frames after which the video signal will be completely intra-refreshed. For example, for QCIF video if subsequent rows of macroblocks are respectively intra coded in every consecutive frame, then after nine frames the whole image associated with a video frame will be intra-refreshed.

same error concealment strategy that the decoder would employ to replace a missing frame. In this particular case, copy previous frame concealment is considered, i.e., $g[k]-\hat{f}[k-1]$. For the rest of the exposition in this section, let the initial error frame and the corresponding MSE associated with the single loss of frame k be denoted as $e_S[k]$ and $\sigma_S^2[k]$, respectively. We will employ $e[k]$ and $\sigma^2[k]$ for the more general case of multiple losses. Similarly, let D denote the total additional distortion affecting the video content in the case of a general packet loss pattern, while D_S will represent the corresponding quantity for the case of a single frame loss.

Now, in order to calculate $D_S[k]$ we need a model that describes how the error power due to the loss of frame k propagates into subsequent frames. To this end, inspired by the approach in [18] the authors characterize the propagated error power at frame $k+l$ as

$$\sigma^2[k+l] = \begin{cases} \sigma_S^2[k] \cdot r^l \cdot (1-l/N) & : \quad 0 \leq l \leq N, \\ 0 & : \quad \text{otherwise}, \end{cases} \tag{1}$$

where N is the intra update period (in number of frames) introduced earlier and $r < 1$ is the attenuation factor that accounts for the effect of spatial filtering employed in the prediction loop of a video encoder. Here, it is assumed that the error is completely removed by intra update after N frames. The reduction in error power due to spatial filtering is actually dependent on the strength of the loop filter and the actual error signal. However, for simplicity the authors assume r to be constant for a given burst length. Then, the total distortion $D_S[k]$ can be computed as

$$D_S[k] = \sum_{i=k}^{\infty} \sigma^2[i] = \sum_{i=0}^{N-1} \sigma_S^2[k] \cdot r^i \cdot (1-i/N) = \alpha \cdot \sigma_S^2[k] \tag{2}$$

where $\alpha = D_S[k]/\sigma_S^2[k]$ captures the amount of error power propagated through the stream due to the single loss at k.

The main idea of the authors is to employ the single loss values described thus far, as measured and computed at the encoder, to synthesize the more general error frame and distortion quantities in the case of multiple losses. We will describe the approach that they take to this end next.

2.1 Burst Losses of Length Two

Let two consecutive losses of frames k and $k-1$ be experienced. The corresponding single loss error frames are given by

$$\begin{aligned} e_S[k-1] &= g[k-1] - \hat{f}[k-1] = \hat{f}[k-2] - \hat{f}[k-2]\,, \\ e_S[k] &= g[k] - \hat{f}[k] = \hat{f}[k-1] - \hat{f}[k]\,. \end{aligned}$$

Therefore, a burst loss of length two affecting frames $k-1$ and k contributes to a residual error frame k given by

$$\begin{aligned} e[k] &= g[k] - \hat{f}[k] = \hat{f}[k-2] - \hat{f}[k] \\ &= e_S[k-1] + e_S[k]\,. \end{aligned}$$

Furthermore, the corresponding increase[7] in MSE associated with frame k in this case can be computed as

$$\sigma^2[k] = \sigma_S^2[k-1] + \sigma_S^2[k] + 2\rho_{k-1,k} \cdot \sigma_S[k-1] \cdot \sigma_S[k]\,, \tag{3}$$

where

$$\rho_{k-1,k} = \frac{{e_S}^{\mathrm{T}}[k-1] \cdot e_S[k] \,/\, M}{\sigma_S[k-1] \cdot \sigma_S[k]}$$

represents the correlation coefficient between error frames $k-1$ and k. As evident from (3), the loss-inflicted error affecting frame k is not any longer a sum of the individual loss contributions, as assumed in previous additive models. The third term in (3) captures the influence of the cross-correlation between the two loss events on the resulting distortion affecting frame k.

Finally, following an analogy with $D_S[k]$ the total distortion associated with the two losses at $k-1$ and k can be computed as

$$\begin{aligned} D[k-1,k] &= \sum_{i=k-1}^{\infty} \sigma^2[i] = \sigma_S^2[k-1] + \alpha \cdot \sigma^2[k] \\ &= \sigma_S^2[k-1] + D_S[k-1] + D_S[k] + 2\rho_{k-1,k} \cdot \sqrt{D_S[k-1] \cdot D_S[k]}\,, \end{aligned}$$

where again (1) was employed to model the propagation of error into subsequent frames that have been successfully received. However, in this case we replaced $\sigma_S^2[k]$ with $\sigma^2[k]$ in (1). Similarly to the case of $\sigma^2[k]$ in (3), the expression for $D[k-1,k]$ above also contains terms that have not been accounted for in previous additive models. In particular, these earlier techniques compute $D[k-1,k]$ solely as the addition of the two terms $D_S[k-1]$ and $D_S[k]$. Therefore, the important cross-correlation influence of the two error events at frames k and $k-1$ is missed.

2.2 Burst Losses of Length B

The authors then extend the approach from Section 2.1 to the general case of burst length $B > 2$. In particular, let B consecutive frames from $k-B+1$ to k be lost. Following an analogy with (3), an expression for the MSE at frame k is proposed as follows

[7] To emphasize the fact that quantization error that is already present in the compressed content is not part of the analysis.

$$\sigma^2[k] = \sum_{i=k-B+1}^{k} \sigma_S^2[i] + 2 \sum_{i=k-B+1}^{k} \sum_{j=i+1}^{k} \rho_{i,j} \cdot \sigma_S[i] \cdot \sigma_S[j] \,. \tag{4}$$

Furthermore, the total distortion is computed using

$$D[k-B+1,\ldots,k] = \sum_{i=k-B+1}^{k-1} \sigma^2[i] + \alpha(B) \cdot \sigma^2[k] \,. \tag{5}$$

Note that in (1) it was assumed that r is constant across the video frames. Therefore, the multiplicative factor α that is computed through r was also considered to be constant. In reality, r can exhibit certain variation over the individual frames and even more importantly it is strongly affected by the power spectrum density (PSD) of the error signal. The work in [18] modeled r as the fraction of error power not removed by the prediction loop filtering. As the number of loss events increases the PSD of the error signal shifts its spectrum toward lower frequencies that are not removed by the filtering. Hence, r is a function of the burst length. To account for this, the authors corrected α in (5) by modeling it as a linear function of B. According to their empirical measurements, this simple model provided a satisfactory performance.

2.3 Two Losses Separated by a Short Lag

Finally, the author derive an expression describing the overall distortion for the case of two non-consecutive packet losses separated by a certain lag l. It is important to have such a characterization in order to study the video distortion caused by an arbitrary loss pattern where the the individual loss events are not necessarily consecutive. It is only necessary to study the case when $l \leq N$ as otherwise the two losses can be considered independent and therefore the total distortion will be additive.

Now, let two separate losses at frames $k-l$ and k occur, where $2 \leq l \leq N$. Then, using the exposition presented thus far the total distortion can be derived and written as

$$D[k-l,k] = \beta(N,l,r) D_S[k-l] + \frac{\sigma^2[k]}{\sigma_S^2[k]} D_S[k] \tag{6}$$

where the expression for $\beta(N,l,r)$ can be obtained in a similar fashion as α was derived in (2). Again, compared to an additive model, where the contributions of the two loss events $D_S[k-l]$ and $D_S[k]$ would be simply summed up (equally), here we can see that their overall contribution is rather represented as a weighted sum of the two individual terms.

2.4 Empirical Evaluation

Armed with the modeling machinery derived thus far, we study now its prediction accuracy by applying packet loss patterns to actual video sequences.

The content employed in the experiments comprises four standard test sequences in QCIF format, *Foreman*, *Mother-Daughter*, *Salesman*, and *Claire*. For each sequence, 280 frames are encoded using JM 2.0 of the H.264 standard [6] at 30 fps and with a constant quantization level for an average luminance PSNR of about 36 dB. The first frame of a sequence is intra-coded, followed by P-frames. Every 4 frames a slice is intra updated to improve error-resilience by reducing error propagation (as recommended in JM 2.0), corresponding to an intra-frame update period of N = 4 × 9 = 36 frames.

In addition to the full-blown model proposed by the authors, and denoted henceforth *local estimation* (LE) they also examined the performance of a sub-sampled version of it, denoted henceforth *global estimation* (GE). In particular, to collect the parameters for the LE model the authors run in total $L \times N$ decodings at the encoder simulating various single or double loss events at every decoding. From these measurements, the parameters $\sigma_S[k]$, $D[k,l]$[8], and α are obtained, for $k = 0, 1, \ldots, L-1$ and $l = 1, 2, \ldots, N$, where L is the length of the sequence in frames. In addition, for the case of burst loss $B > 2$ additional $L \times 2$ decodings are run in order to fit the linear model for $\alpha(B)$. The two parameters of the model are then stored for use in the subsequent experiments. Finally, for the GE model, sub-sampled versions of the above measurement procedures are run in order to reduce the complexity of the related simulations and the amount of data that needs to be stored as parameters. Specifically, for the index of the first loss event only $k = 10, 20, 30, \ldots$ is employed and the stored parameters represent instead average values over all possible choices for k.

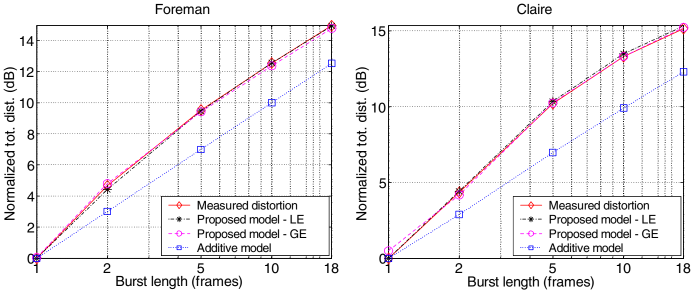


Fig. 3. Measured versus estimated total distortion as a function of burst loss length, normalized by total distortion for a single loss. (left) Foreman, (right) Claire.

In Figure 3, we examine the prediction performance of the proposed GE and LE models against the actual measured distortion and the conventional additive model as a function of the burst length (in frames). It can be seen

[8] Precisely, it is in fact the MSE $\sigma^2[k]$ for the second loss in (3) and (6) that is measured and stored.

that in the case of both, *Foreman* and *Claire*, the two proposed models provide accurate estimates of the overall distortion afflicting the video stream as the burst length increases. Contrarily, we can see from Figure 3 that the additive model provides a distortion estimate that is consistently below the actual value. By not taking into account the cross-correlation between the individual loss events, as explained earlier, this model omits to account for a significant portion of the error process arising in such situations.

Next, in Figure 4 we examine the performance of the GE and LE models in the case of two losses separate by a lag as a function of the lag size (in frames). It can be seen that again in the case of both, *Mother-Daughter* and *Salesman*, the proposed models outperform the additive model by providing a more accurate estimate of the actual distortion especially for shorter lag lengths. This is expected as for longer temporal separation between the two loss events their influence becomes less correlated, i.e., they become more independent, as argued earlier.

In summary, it can be seen that accounting for the interaction between individual loss events can provide significant benefits in terms of distortion prediction accuracy. However, the authors did not examine how their models would perform for arbitrary loss patterns comprising prospectively multiple burst loss and individual loss events separated with different lag values. Moreover, the performance results presented in [20, 23] and included here in Figure 3 describe only the average performance of the models. In particular, the reported distortion estimate associated with a model is obtained as the average predicted distortion across different loss realizations of the same burst length normalized with the corresponding average distortion for a single loss event. Therefore, it is difficult to assess the variations in prediction accuracy and the distribution of prediction error for a given model and burst length. Finally, the findings presented here imply that overall distortion always increases with burst length. However, as shown later on in Section 4 this is in fact not always true and is content dependent.

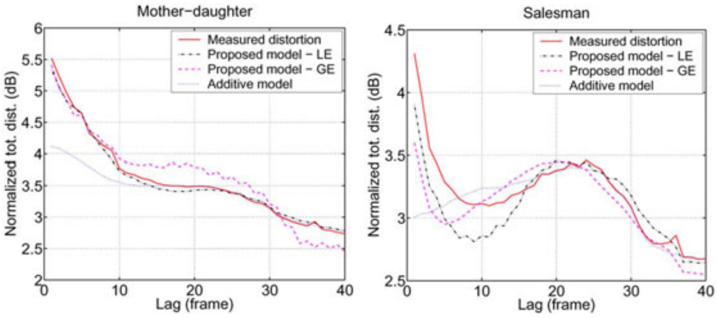


Fig. 4. Measured versus estimated total distortion for two losses separated by a lag. (left) MotherDaughter, (right) Salesman.

3 Distortion Chains

The approach presented here formalizes the concept of distortion memory in lossy video communication. It provides a modeling framework that generalizes to higher-order terms the notion of distortion cross-correlation between packet loss events. In particular, the key idea of the Distortion Chains framework [21, 24] is to model the additional distortion associated with the loss of a present packet as dependent on the k nearest previous losses where k denotes the order of the chain. When describing this model subsequently, we will take advantage of the notation already introduced earlier in Section 2.

Let L be the length of a video sequence in frames and let $\boldsymbol{k} = (k_1, k_2, \ldots, k_N)$ denote a loss pattern of length N, i.e., N frames are lost during transmission where $k_i < k_j$, for $i < j$, are the indices of the lost frames. Then, the total distortion, denoted by $D(\boldsymbol{k})$, due to the loss pattern is the sum of the MSEs over all the frames affected by the loss pattern $\boldsymbol{k}$, i.e.,

$$D(\boldsymbol{k}) = \sum_{l=1}^{L} \sigma^2[l] = \sum_{l=k_1}^{L} \sigma^2[l] \,. \tag{7}$$

For example, $D(k)$ and $D(k_1, k_2)$ would correspond to the total distortion values associated with the loss of frame k and frames k_1, k_2, respectively, mentioned in the context of Figure 2. Then, $D(k_{N+1}|\boldsymbol{k})$ is defined as the additional increase in distortion due to losing frame $k_{N+1} > k_N$ given that frames $k_1, \ldots, k_N$ are already lost, i.e.,

$$D(k_{N+1}|\boldsymbol{k}) = D(k_1, \ldots, k_{N+1}) - D(k_1, \ldots, k_N). \tag{8}$$

A Distortion Chain model of order N (denoted DC^N henceforth) comprises the distortion values $D(\boldsymbol{k})$ for every loss pattern $\boldsymbol{k}$ of length N satisfying $k_i < k_j$, for $i < j$, and the conditional distortions $D(k_{N+1}|\boldsymbol{k})$ for every loss pattern $(\boldsymbol{k}, k_{N+1})$ of length $N + 1$ satisfying $k_N < k_{N+1}$. Using DC^N an estimate, $\widetilde{D}(\boldsymbol{k})$, of the total distortion associated with an arbitrary packet loss pattern $\boldsymbol{k} = (k_1, \ldots, k_P)$ with P losses, where $N < P \leq L$, can be computed as

$$\begin{aligned} \widetilde{D}(\boldsymbol{k}) = {} & D(k_1, \ldots, k_N) + \\ & \sum_{i=N}^{P-1} D(k_{i+1}|k_{i-N+1}, \ldots, k_i) \,. \end{aligned} \tag{9}$$

3.1 Complexity and Implementation

As in the case of the framework from Section 2, the distortion quantities comprising a Distortion Chain can be generated at the encoder by simulating the corresponding loss events, decoding the video sequence, and then computing the resulting distortions. Though this may be impractical for large N, the authors have established that good prediction accuracy can still be obtained even with Distortion Chains of small order. Moreover, when packet losses are spaced far apart (farther than the intra refresh period of the encoded video sequence), they become decoupled since their effects are independent, as discussed earlier. This reduces the complexity of the algorithm associated with generating the model, as explained next.

For instance, for the Distortion Chain DC^1 one needs to store the distortion values $D(k)$ associated with losing frame $k = 1, \ldots, L$, illustrated in Figure 2(a). In addition, one also needs to store the quantities $D(j|k)$ from (8), which represent the additional increase in distortion when frame j is lost, given that frame k is already lost, for $1 \leq k < j \leq L$. Note that storing $D(j|k)$ is equivalent to storing $D(k, j)$ as apparent from (8), where $D(k, j)$ is the total distortion associated with losing frames k and j, illustrated in Figure 2(b). Now, if $D(k, j)$ was stored for every possible pair (k, j), then the total storage cost would be quadratic in L since there are L isolated losses contributing to $D(k)$ and $L(L-1)/2$ distinct $D(k_1, k_2)$ values. However, since the distortion coupling between dropped packets decreases as the distance between the packets increases, one practical simplification is to recognize that $D(k, j) = D(k) + D(j)$ for $|j - k| > M + 1$, where M depends on the compression. For example, for a video encoding with a Group Of Pictures (GOP) structure, M is at most the number of packets in the GOP. On the other hand, if all frames are predictively encoded, M corresponds then to the intra refresh period of the encoding. This simplification reduces the required storage and computation for DC^1 to being linear in L, precisely $(L - M - 1)(M + 1) + M(M - 1)/2$.

3.2 Performance Assessment

Here, we investigate the distortion prediction performance of the Distortion Chains framework by simulating different packet loss patterns on actual compressed video content. We compare the measured total distortion for each pattern with that predicted by Distortion Chain models of different order $N = 0, 1, 2$. The video sequences used in the experiments are coded using JM 2.1 of the JVT/H.264 video compression standard [6], using coding tools of the Main profile. Two standard test sequences in QCIF format are used, Foreman and Carphone. Each has at least 300 frames at 30 fps, and is coded with a constant quantization level at an average luminance (Y) PSNR of about 36 dB. The first frame of each sequence is intra-coded, followed by all P-frames. Every 4 frames a slice is intra updated to improve error-resilience by reducing error propagation (as recommended in JM 2.1), corresponding to an intra-frame update period of $M = 4 \times 9 = 36$ frames.

First, in Figure 5 we show the distortion values $D(k)$, for $k = 1, \ldots, L$, comprising the Distortion Chain DC^0 for the video sequences Foreman and Carphone. Notice the huge variability in total distortion that results from losing different frames in the sequence. In Figure 5 only the first 300 frames,

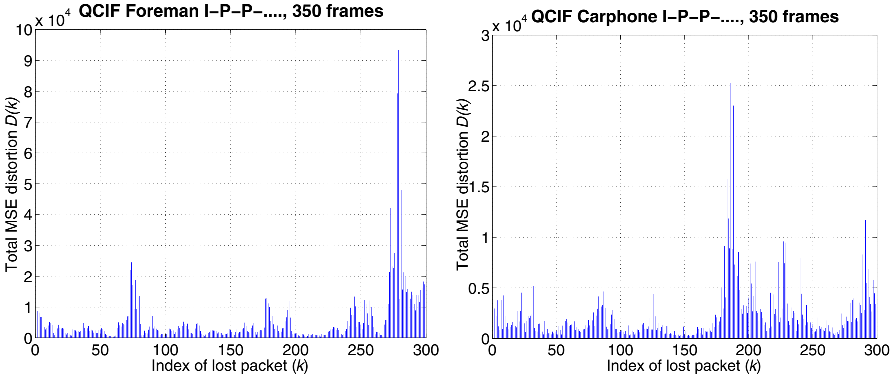


Fig. 5. Example of DC^0 for Foreman and Carphone video sequences. Each sample point in the graphs identifies the total distortion $D(k)$ associated with the loss of a single frame k.

out of a 350 frame video sequence, are considered as possible candidates to be lost in order to properly account for the error propagation effect.

This variability is quantified in Table 1 where we see that there exists significant variation in the total distortion produced by the loss of different P-frames.

Table 1. Mean of the total MSE distortion $D(k)$, and mean-normalized versions respectively of the minimum, median, 95-percentile, and maximum values of $D(k)$ for DC^0 for different sequences.

Sequence	Mean	Min	Median	95%	Max
Foreman	5615.88	0.04	0.49	3.18	16.64
Mother & Daughter	247.77	0.06	0.61	3.71	6.92
Carphone	2254.80	0.10	0.60	3.33	11.19
Salesman	284.38	0.06	0.61	3.35	5.86

Then, in Figure 6 we examine the conditional distortion values employed by the Distortion Chain model DC^1. Notice that there are some values of $D(j|k)$ which are negative. This is an interesting phenomenon that has not been reported earlier in works on distortion modelling, save for the study on burst losses [20] examined in Section 2. The authors of this earlier work reported that negative correlation was identified to sometimes exist in neighboring lost frames. Having negative conditional distortions leads to the surprising result that sometimes it is better to drop two frames instead of dropping only one frame. For example, sometimes it is better to drop the two frames k and j together, instead of only dropping the single frame k, since the total distortion for dropping both frames k and j is less than that for dropping

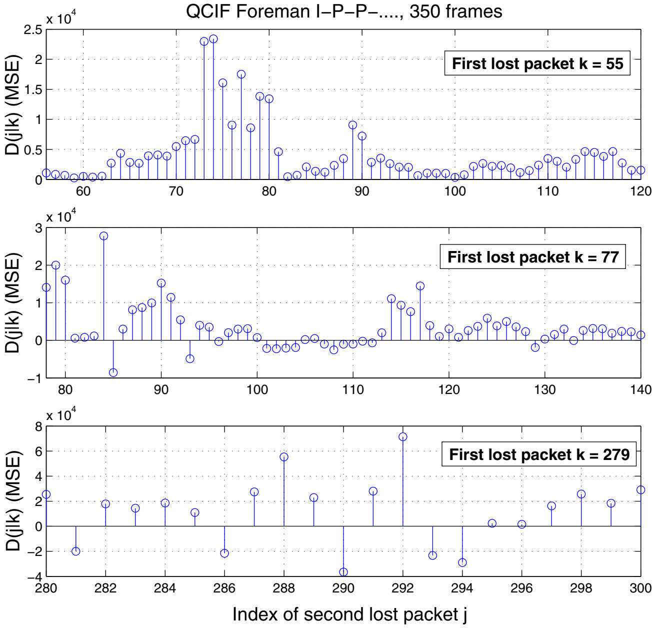


Fig. 6. Example of DC^1 information for the Foreman video sequence. Each sample point in the graph is $D(j|k)$ which corresponds to the increase in total distortion due to the loss of frame j given that frame $k < j$ is already lost. Notice that there are a number of $D(j|k)$ which are negative, for example $D(290|279)$. In these cases, instead of only dropping one frame (e.g. 279), it is better to drop two frames (e.g. 279 and 290) since that will produce a smaller total distortion.

frame k only. Having this knowledge can be very useful for adaptive video streaming.

This section proceeds by examining the prediction accuracy of the Distortion Chain models DC^0, DC^1, and DC^2. In the investigation, we also explore the performance of a liner model as considered in early works on distortion modeling [18, 19]. In particular, these works model the total distortion afflicting the video sequence as being proportional to the number of lost packets that occur, as described earlier. Then, with this *stationary linear* model, the expected total distortion (D_{Linear}) is computed as

$$D_{Linear} = \#Losses \cdot \frac{1}{L} \sum_{l=1}^{L} D(l), \tag{10}$$

where $D(l)$ is the total distortion that is associated with the loss of packet l (assuming that all other packets are correctly received), L is the total number of packets in the video sequence, and $(1/L)\sum_{l=1}^{L} D(l)$ is the average single packet loss total distortion. Finally, given that the total number of losses is linearly (hence the name of the model) related to the average packet loss rate (PLR), i.e., $\#Losses = PLR \cdot L$ (for $PLR \leq 1$) we can write (10) as

$$D_{Linear} = PLR \cdot \sum_{l=1}^{L} D(l). \tag{11}$$

In the first set of experiments, conducted using the Foreman sequence, the prediction performance is examined across the range of packet loss rates (PLR) 3 - 10%. Note that for lower PLRs the distortion chain framework can often perfectly predict the distortion since it can typically exactly account for the lost packets at the low PLRs. For each packet loss rate a corresponding set of 50,000 random packet loss patterns is generated. For each loss pattern $\boldsymbol{k} = (k_1, k_2, \ldots)$ the video is decoded and the resulting total MSE distortion $D(\boldsymbol{k})$ of the luminance component of the video is recorded. At the same time, we generate predictions of $D(\boldsymbol{k})$ using respectively the Linear model (as defined in Equation (11) above) and the proposed Distortion Chains DC^0, DC^1, and DC^2. The predicted distortion values are denoted $\widetilde{D}(\boldsymbol{k})$, as introduced earlier . Finally, we compute the PSNR of these quantities using $10 \log_{10} \frac{255^2}{D/N_F}$, where D is either $D(\boldsymbol{k})$ or $\widetilde{D}(\boldsymbol{k})$ and N_F is the number of frames in the video sequence.

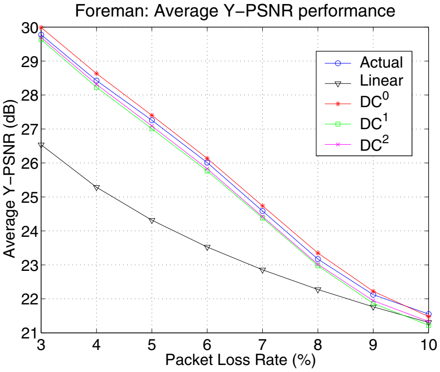


Fig. 7. PSNR of the actual and predicted total MSE distortions for *Foreman*.

In Figure 7, we show these PSNR values, averaged over all 50,000 loss patterns that correspond to a particular loss rate, as a function of the PLR. There are a few observations that follow from Figure 7. First, all of the Distortion Chains provide better predictions of the expected distortion than the Linear model. Second, on average DC^1 and DC^2 underestimate the Y-PSNR as computed above, while DC^0 overestimates it, i.e., on average DC^1 and DC^2 overestimate the actual distortion, while DC^0 underestimates it.

Note that the performance difference between Linear and the Distortion Chain models is larger for low packet loss rates and it gradually decreases as the packet loss rate increases. Specifically, at $PLR = 3\%$ the Distortion Chain models provide a performance gain of roughly 3.5 dB, while at $PLR = 10\%$ the gain is practically negligible. In essence, this is due to the large variability in total distortion produced as a function of the specific packet that is lost (see Figure 5). For example, let us assume that we lose only *one* packet in the sequence. Then, based on the specific lost packet l, for some l the total distortion will be much larger than the average single packet loss total distortion, $(1/L)\sum_{l=1}^{L} D(l)$, while for other l the total distortion will be much less than the average. Hence, a Distortion Chain allows us to explicitly capture this variability as a function of l, while the Linear model does not provide that. On the other hand, as the number of losses increases (assuming for simplicity that the loss effects are independent) the resulting total distortion will approach $\#Losses \cdot (1/L)\sum_{l=1}^{L} D(l)$, since more averaging (over the lost packets) occurs and therefore the penalty that the Linear model pays decreases.

Next, we define $\Delta D(\boldsymbol{k}) = \frac{|D(\boldsymbol{k}) - \widetilde{D}(\boldsymbol{k})|}{D(\boldsymbol{k})}$ to be the relative error of a predicted distortion $\widetilde{D}(\boldsymbol{k})$ for a packet loss pattern $\boldsymbol{k}$. In essence, the relative error informs us how big the prediction error of $\widetilde{D}(\boldsymbol{k})$ is relative to the actual value $D(\boldsymbol{k})$ for a given loss pattern $\boldsymbol{k}$. We next examine the distribution of the relative error $\Delta D(\boldsymbol{k})$ over the 50,000 packet loss patterns $\boldsymbol{k}$ that correspond to a given PLR. Figure 8 shows the Cumulative Density Functions (CDFs) of the relative errors for all four distortion models considered here, for both PLR = 3% and 8%. The first observation is that all of the distortion chain models perform significantly better than the linear model. In addition, for PLR = 3% we see that DC^0, DC^1, and DC^2 provide estimates that are within a 10% error bound 40%, 75%, and 95% of the time, respectively, while the linear model achieves this less than 10% of the time. Similarly, DC^0, DC^1, DC^2 provide estimates that are within a 20% error bound 74%, 93%, and 99% of the time, respectively, while the linear model does that only 5% of the time. Figure 8 also shows that the distortion chain models provide improved accuracy as compared to the linear model at 8% PLR, though the improvement is lower due to the reduced accuracy as a result of the higher packet loss rate.

In conclusion, the Distortion Chains framework provides improved prediction accuracy relative to prior linear models. The framework extends the concept of distortion-loss correlation to higher order terms, when compared to the burst loss model from Section 2 that only considers the first order terms. In addition, the present study raised the awareness of the interesting phenomenon of negative distortion values observed in certain specific cases of packet loss. Still, no formal analysis of this phenomenon has been carried out within the modeling framework of Distortion Chains. Moreover, the two distortion studies presented thus far do not consider the influence of

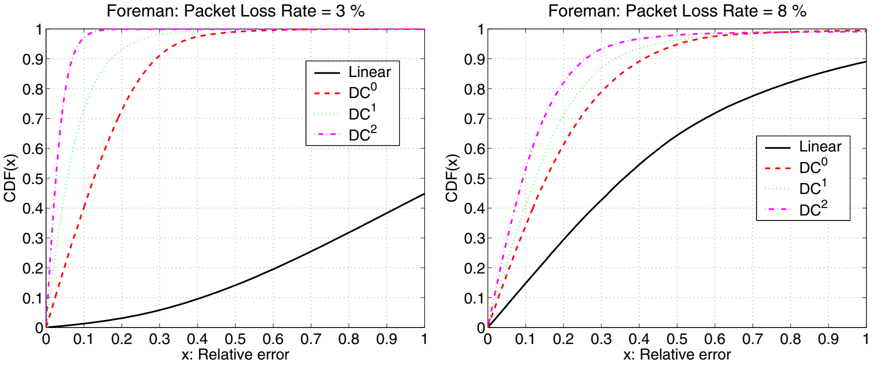


Fig. 8. CDF of $\Delta D(\boldsymbol{k})$ for PLR = 3 % (left) and PLR = 8 % (right).

correlated packet loss probabilities prospectively experienced during transmission. Having such a state-based channel model would certainly alter the contributions of certain packet loss patterns relative to others to the overall expected distortion experienced by the transmitted content. The work presented in the next section overcomes these shortcomings by developing a framework that incorporates models for both missing aspects described above.

4 Distortion Trellis

The work in [22] addresses the problem of distortion modeling for video transmission over burst-loss channels characterized by a finite state Markov chain. Based on a detailed analysis of the error propagation and the bursty losses, a Distortion Trellis model is proposed, enabling estimating at both frame level and sequence level the expected mean-square error (MSE) distortion caused by Markov-model bursty packet losses. The model takes into account the temporal dependencies induced by both the motion compensated coding scheme and the Markov-model channel losses. The model is applicable to most block-based motion compensated encoders, and most Markov-model lossy channels as long as the loss pattern probabilities for that channel are computable. In addition, based on the study of the decaying behavior of the error propagation, a sliding window algorithm is developed to perform the MSE estimation with low complexity. In order to describe the proposed modeling framework in greater detail, some preliminaries need to be covered first.

4.1 Preliminaries

General assumptions

It is assumed that the encoder compresses a raw video sequence into groups of pictures (GOPs) each comprising an I-frame followed by subsequent

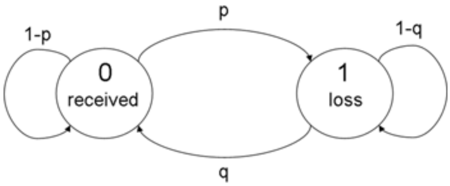


Fig. 9. Gilbert channel model

P-frames. In a P-frame, macroblock (MB) intra-refreshing can be used for either coding efficiency or error resilience. All MBs in a frame are grouped into one slice and each slice is coded into one network packet. At the decoder, a certain temporal error concealment strategy is applied whenever a frame is lost. Motivated by the presence of temporal memory and correlation in packet losses in wired/wireless Internet the authors employ a Gilbert model [25] to describe the long-term network packet loss. In this model, the channel switches between an error state and an error-free state. When the channel is in the error state, the transmitted packet is always lost while in the error-free state the packet is always correctly received. Let State 0 and State 1 respectively denote the error-free and the error states. As shown in Figure 9, the parameter p is the transition probability from State 0 to State 1, and q denotes the probability of the opposite transition. Normally $p+q<1$. If $p+q=1$, the Gilbert model reduces to a Bernoulli model. From the definition, the stationary probability for State 0 and 1, denoted by π_0 and π_1, can be computed as $\pi_0 = q/(p+q)$ and $\pi_1 = p/(p+q)$, respectively. Then, the mean packet loss ratio PLR equals π_1, and the average burst length ABL is given by $1/q$.

Problem formulation

Let x_n^i and y_n^i denote the reconstructed pixel values for frame n and pixel i at the encoder and at the decoder, respectively. Then, the average MSE distortion for frame n for channel realization c can be calculated as

$$d_n^c = E_i\left\{\left(x_n^i - y_n^i\right)^2\right\} = \frac{1}{XY}\sum_{i=1}^{XY}\left(x_n^i - y_n^i\right)^2, \tag{12}$$

where $E_i\{\cdot\}$ denotes the computation of the average MSE over all pixels in frame n, X and Y respectively denote the frame width and height in pixels. Finally, the expected distortion of frame n can be defined as

$$d_n = E_c\left\{d_n^c\right\} = E_c\left\{E_i\left\{\left(x_n^i - y_n^i\right)^2\right\}\right\}, \tag{13}$$

where $E_c\{\cdot\}$ denotes the expectation taken over all possible channel realizations. Note that the definition of d_n is generic and hence applies to most existing coding technologies and channel realizations.

When calculating d_n for a Bernoulli channel[9], an important problem is to model the error propagation due to decoding dependencies between temporally adjacent frames. In the case of a Gilbert channel, the packet losses also exhibit temporal dependencies. Hence, when calculating d_n for a Gilbert channel, the decoding dependencies and the loss dependencies should both be considered (jointly). Therefore, it is more complex to model d_n in the latter case.

4.2 Framework of the Distortion Trellis model

In motion compensated video coding, decoding error in a previous frame may propagate into the current frame. In such a case, the distortion of the current frame is affected not only by the transmission state ("Lost" and "Received") of the current frame, but also by the transmission states of all previous frames in the same GOP. In other words, it is affected by the loss patterns of all transmitted frames in the same GOP (including the current frame). For a frame sequence of length n, the total number of all possible loss patterns is 2^n. Thus, theoretically, after decoding the n-th frame in a GOP, the total number of all possible distortion values of the n-th frame at the decoder is also 2^n.

In a Bernoulli channel, a packet is either lost with a probability PLR or received with a probability $1-PLR$, independently of other loss events. Thus, when calculating d_n, we do not need to calculate all 2^n possible distortions. Instead, most existing models define another two distortions, d_n^L and d_n^R. The former is the expected distortion given that frame n is lost, while the latter denotes the expected distortion for the case when frame n is received. Often, d_n^L and d_n^R are calculated in a recursive approach to account for the error propagation. In such a case, we only need to calculate two distortion values for each frame. Finally, d_n is calculated as

$$d_n = E_c\left\{d_n^c\right\} = PLR \cdot d_n^L + (1 - PLR) \cdot d_n^R, \tag{14}$$

In a Gilbert channel, packet losses are no more i.i.d. but exhibit dependencies over time. Observed from the sender the loss process of all P-frames in a GOP is a two-state Markov process[10]. In such a case, when calculating d_n, we need to consider all 2^n cases for frame n, which adumbrates a rather elaborate calculation process.

Now, consider the impairments for a transmitted packet (frame) sequence of length n as an n-bit binary random variable $K_n = \{B_j\}_{j=1}^n$. The random variable B_j is over the binary alphabet $\{0,\ 1\}$. $B_j = 1$ indicates that the j-th frame is lost. Then, the total number of all possible values of K_n is 2^n. Define moreover an ordered set $\mathbf{I}_n = \{k_n^r\}$, $r = 1, \ldots, 2^n$, where k_n^r is an n-bit

[9] That is in the case of independent packet losses as assumed in the two previous studies on distortion modeling presented in this chapter.

[10] A correct reception of the I-frame is assumed to be guaranteed.

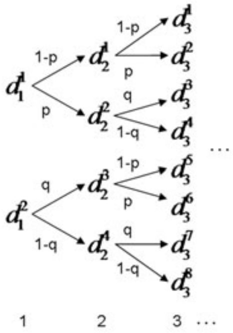


Fig. 10. Statistical dependencies $\{d_n^r, n = 1, 2, \ldots\}$

binary number and $k_n^1 = \overbrace{0\ldots0}^{nbits}$, $k_n^r = 1 + k_n^{r-1}$, $r = 2, \ldots, 2^n$. Furthermore, we assume that the r-th value of K^n is k_n^r, the r-th element in $\mathbf{I}_n$. Note that in the present analysis this is an important assumption, based on which we can recursively derive $\mathbf{I}_n$ from $\mathbf{I}_{n-1}$ in a simple way. Hereafter, we refer to k_n^r as the r-th loss pattern of a frame sequence of length n.

Let $P(k_n^r)$ denote the probability that loss pattern k_n^r occurs, i.e. $P(k_n^r) = \mathrm{P_r}(K_n = k_n^r)$. Note that different loss patterns lead to different distortion values. Let d_n^r be the decoder distortion of the n-th frame in a frame sequence of length n under loss pattern k_n^r. Then, d_n^r can be defined as

$$d_n^r = E_i\left\{\left(x_n^i - y_{n,r}^i\right)^2\right\}, \tag{15}$$

where $y_{n,r}^i$ denotes the decoder reconstructed value of pixel i in the n-th frame for an n-length frame sequence under loss pattern k_n^r. Thus, from the definition of d_n^r, we can obtain an important probability relation as follows: Pr(at the decoder the distortion of frame n is d_n^r)= $P(k_n^r)$.

In essence, the definitions of k_n^r and d_n^r lay a foundation for the proposed model. First, they establish the relation between various loss patterns and their corresponding decoding distortions. Second, they enable us to recursively analyze the loss dependencies and decoding dependencies. Figure 10 illustrates the statistical dependencies between the elements of the set $\{d_n^r, n = 1, 2, \ldots\}$. Since the packet loss dependencies of the channel loss process and the distortion/decoding dependencies of the video frames can both be depicted by a trellis graph, we refer to the proposed distortion estimation method as the Distortion Trellis model.

Then, we can calculate the expected distortion of frame n by taking an expectation over all possible decoder distortion values for frame n,

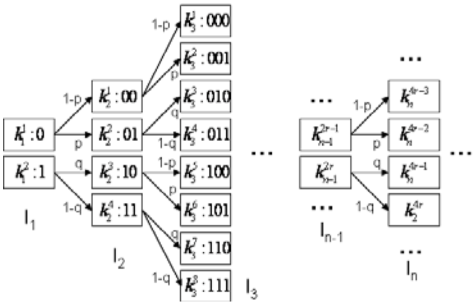


Fig. 11. Computation of the loss pattern probabilities $\{P(k_n^r), n = 1, 2, \ldots\}$

$$\begin{aligned} d_n = E_c\{d_n^c\} &= \sum_{r=1}^{2^n} d_n^r \cdot \Pr\left(\text{distortion of frame } n \text{ is } d_n^r\right) \\ &= \sum_{r=1}^{2^n} d_n^r \cdot P\left(k_n^r\right), n = 1, 2, \ldots \end{aligned} \tag{16}$$

The formula in (16) is the general form of the proposed Distortion Trellis model. From (16), it is clear that the computation of d_n necessitates knowledge of both d_n^r and $P\left(k_n^r\right), r = 1, \ldots, 2^n$. We must emphasize that (16) is applicable to most channel models and hence is general. For different channel models, the only difference in using the Distortion Trellis model is the computation of $P\left(k_n^r\right)$, because generally the same loss pattern occurs with different probabilities in different channels. On the other hand, d_n^r is uncorrelated with a specific channel model but only depends on the video sequence. That is why the Distortion Trellis model can be easily extended any arbitrary finite state Markov loss model. For a Gilbert channel, $P\left(k_n^r\right)$ can be derived recursively as follows. From the definition of $\{k_n^r\}$, it is clear that given $P\left(k_{n-1}^t\right), t = 1, \ldots, 2^{n-1}$, the loss pattern probabilities can be written as

$$\begin{cases} P\left(k_n^{4r-3}\right) = (1-p) \cdot P\left(k_{n-1}^{2r-1}\right) \\ P\left(k_n^{4r-2}\right) = p \cdot P\left(k_{n-1}^{2r-1}\right) \\ P\left(k_n^{4r-1}\right) = q \cdot P\left(k_{n-1}^{2r}\right) \\ \quad P\left(k_n^{4r}\right) = (1-q) \cdot P\left(k_{n-1}^{2r}\right), r = 1, \ldots, 2^{n-2}. \end{cases} \tag{17}$$

The computation of the loss pattern probabilities is illustrated in Fig. 11, which reveals the loss dependencies in the case of a Gilbert channel. The remaining task is how to calculate d_n^r.

Recursive computation of d_n^r

From the definition of k_n^r, it can be observed that for loss pattern k_n^{2r-1}, the n-th packet is received, while for loss pattern k_n^{2r}, the n-th packet is lost, where $r = 1, , 2^{n-1}$. Thus, given that the loss pattern of the previous $n-1$ frames is k_{n-1}^r, d_n^{2r-1} is the frame-average distortion if the n-th frame is received while d_n^{2r} denotes the same quantity for the case when the n-th frame is lost. Next, we separately consider computing d_n^{2r-1} and d_n^{2r}.

(i) Computation of d_n^{2r}

If a frame is lost, all MBs in this frame are recovered using some temporal error concealment strategy, regardless whether they are coded in inter or intra mode. Let $f_l(i)$ denote the index of the l-th pixel in frame $n-1$ that is used to estimate pixel i in frame n. Then the final concealed value of $y_{n,2r}^i$ can be expressed as $\Phi_l\left(y_{n-1,r}^{f_l(i)}\right)$, where Φ_l represents the pixel operation on $y_{n-1,r}^{f_l(i)}$ for all l used in obtaining the final concealed value of $y_{n,2r}^i$. For example, in video coders using sub-pixel motion estimation, Φ_l denotes the interpolation operation. For another example, in video coders using deblocking filters, Φ_l denotes the deblocking operation. For previous frame copy concealment, $\Phi_l\left(y_{n-1,r}^{f_l(i)}\right) = y_{n-1,r}^i$. Φ_l could also denote weighted prediction and so on. It is a reasonable assumption that Φ_l is a linear pixel filtering operation and can be considered the same for different frames. Then, d_n^{2r}can be derived as follows:

$$\begin{aligned} d_n^{2r} &= E_i\left\{\left(x_n^i - \Phi_l\left(y_{n-1,r}^{f_l(i)}\right)\right)^2\right\} \\ &= E_i\left\{\left(x_n^i - \Phi_l\left(x_{n-1}^{f_l(i)}\right) + \Phi_l\left(x_{n-1}^{f_l(i)}\right) - \Phi_l\left(y_{n-1,r}^{f_l(i)}\right)\right)^2\right\} \\ &= E_i\left\{\left(x_n^i - \Phi_l\left(x_{n-1}^{f_l(i)}\right)\right)^2\right\} + E_i\left\{\left(\Phi_l\left(x_{n-1}^{f_l(i)}\right) - \Phi_l\left(y_{n-1,r}^{f_l(i)}\right)\right)^2\right\} \\ &= ECD_n + E_i\left\{\left(\Phi_l\left(x_{n-1}^{f_l(i)} - y_{n-1,r}^{f_l(i)}\right)\right)^2\right\}, r = 1, \ldots, 2^{n-1}, \end{aligned} \tag{18}$$

where $ECD_n = E_i\left\{\left(x_n^i - \Phi_l\left(x_{n-1}^{f_l(i)}\right)\right)^2\right\}$. Note that ECD_n is the average error concealment distortion of frame n. Given a specific coding scheme and error concealment strategy, Φ_l and $f_l(i)$ are determined and then ECD_n is determined. It is worth noticing that ECD_n is the new added distortion if frame n is lost. $E_i\left\{\left(\Phi_l\left(x_{n-1}^{f_l(i)} - y_{n-1,r}^{f_l(i)}\right)\right)^2\right\}$ is the temporal propagation distortion from frame $n-1$. Note that the third identity in (18) is based on the assumption that the concealment error $x_n^i - \Phi_l\left(x_{n-1}^{f_l(i)}\right)$ and the propagation error $\Phi_l\left(x_{n-1}^{f_l(i)}\right) - \Phi_l\left(y_{n-1,r}^{f_l(i)}\right)$ are uncorrelated [26, 27]. The fourth identity is based on the assumption that Φ_l is linear which is quite reasonable in most cases.

Furthermore, as explained earlier in this chapter error propagation into subsequent frames is typically attenuated by the adoption of some coding schemes such as de-blocking filtering and sub-pixel motion estimation [28], whose effect can be regarded as a spatial filter or more precisely as an error attenuator. Therefore, the propagated distortion in a present frame can be considered as the filtered output of the distortion in its temporal predecessor. Following this reasoning, the term $E_i\left\{\left(\Phi_l\left(x_{n-1}^{f_l(i)} - y_{n-1,r}^{f_l(i)}\right)\right)^2\right\}$ in (18) can be approximated as

$$\begin{aligned} &E_i\left\{\left(\Phi_l\left(x_{n-1}^{f_l(i)} - y_{n-1,r}^{f_l(i)}\right)\right)^2\right\} \\ &= u \cdot E_i\left\{\left(x_{n-1}^{i} - y_{n-1,r}^{i}\right)^2\right\} = u \cdot d_{n-1}^{r}. \end{aligned} \tag{19}$$

Then (18) can be rewritten as

$$d_n^{2r} = ECD_n + u \cdot d_{n-1}^{r}, r = 1, \ldots, 2^{n-1}, \tag{20}$$

where u is the error attenuation factor for a lost frame.

Therefore, d_n^{2r} can be estimated as a sum of two separate parts. One part is the average concealment distortion ECD_n, which can be directly calculated at the encoder just after encoding frame n. The second term in (20) denotes the temporal distortion propagation and indicates the relation between d_n^{2r} and d_{n-1}^{r}. In particular, this term reveals the numerical relationship between the distortions of frame $n-1$ and n when frame n is lost. For a practical application, the parameter u has to be estimated for the specific video coder and content that are employed.

(ii) Computation of d_n^{2r-1}

We now turn to computing d_n^{2r-1}. As is well known, a received frame may still contain distortion due to error propagation from an impaired previous frame. In such a case, the coding modes should be considered because the distortions in received inter-coded MBs and intra-coded MBs are different. We first consider the case when all MBs are coded in inter mode and then we will extend our result to the more general case of having mixed MB coding modes in a frame. Let $g_l(i)$ denote the index of the l-th pixel in frame $n-1$ that is used to estimate pixel i in frame n. Note that $g_l(i)$ may differ from $f_l(i)$. Then, at the encoder, the predicted value of x_n^i can be expressed as $\Psi_l\left(x_{n-1}^{g_l(i)}\right)$, where Ψ_l represents the pixel operation on all $x_{n-1}^{g_l(i)}$ used for obtaining the predicted value of x_n^i, such as when performing interpolation or deblocking filtering. We also assume that Ψ_l is linear and has the same form for different frames. Similarly, at the decoder, the predicted value of $y^i_{n,2r-1}$ is $\Psi_l\left(y^{g_l(i)}_{n-1,r}\right)$. Then, d^{2r-1}_n can be derived as follows:

$$\begin{aligned} d^{2r-1}_n &= E_i\left\{\left(\Psi_l\left(x^{g_l(i)}_{n-1}\right) - \Psi_l\left(y^{g_l(i)}_{n-1,r}\right)\right)^2\right\} \\ &= E_i\left\{\left(\Psi_l\left(x^{g_l(i)}_{n-1} - y^{g_l(i)}_{n-1,r}\right)\right)^2\right\}, r = 1, \ldots, 2^{n-1}. \end{aligned} \tag{21}$$

As in the case of d^{2r}_n, the operator Ψ_l can be regarded as a spatial filter that will attenuate the error propagation. Hence, we similarly employ $v_0 \cdot d^r_{n-1}$ to approximate $E_i\left\{\left(\Psi_l\left(x^{g_l(i)}_{n-1} - y^{g_l(i)}_{n-1,r}\right)\right)^2\right\}$ and therefore we can rewrite (21) as

$$d^{2r-1}_n = v_0 \cdot d^r_{n-1}, r = 1, \ldots, 2^{n-1}, \tag{22}$$

where v_0 is the error attenuation factor for a received frame, in which all MBs are coded in inter mode.

The development of (22) assumes that all MBs in a P-frame are coded in an inter mode. However, a P-frame often contains intra-coded MBs, which will effectively attenuate the error propagation, as explained earlier in this chapter. Therefore, to take this into account we introduce a new constant λ and rewrite (22) as

$$d^{2r-1}_n = v \cdot d^r_{n-1},\ r = 1, \ldots, 2^{n-1}, \tag{23}$$

where $v = \lambda \cdot v_0$.

In essence, (23) describes the numerical relationship between the distortions of frame $n-1$ and n when frame n is received. As in the case of u from (20), the parameter v needs to be estimated as well.

Recursive computation of d_n and further analysis

Based on (20) and (23), we can recursively obtain the distortion d^r_n, for $r = 1, \ldots, 2^n$. The loss pattern probability $P(k^r_n)$ can be recursively calculated with (17). Then, using (16), the expected distortion d_n for Gilbert channel packet losses can be estimated as

$$\begin{aligned} d_n &= \sum_{t=1}^{2^n} d^t_n \cdot P\left(k^t_n\right) \\ &= \sum_{r=1}^{2^{n-2}} \left[P\left(k^{4r-3}_n\right) d^{4r-3}_n + P\left(k^{4r-2}_n\right) d^{4r-2}_n \right. \\ &\quad \left. + P\left(k^{4r-1}_n\right) d^{4r-1}_n + P\left(k^{4r}_n\right) d^{4r}_n\right], \end{aligned} \tag{24}$$

where

$$\begin{cases} P\left(k_n^{4r-3}\right) = (1-p)\,P\left(k_{n-1}^{2r-1}\right), & d_n^{4r-3} = v \cdot d_{n-1}^{2r-1} \\ P\left(k_n^{4r-2}\right) = p \cdot P\left(k_{n-1}^{2r-1}\right), & d_n^{4r-2} = ECD_n + u \cdot d_{n-1}^{2r-1} \\ P\left(k_n^{4r-1}\right) = q \cdot P\left(k_{n-1}^{2r}\right), & d_n^{4r-1} = v \cdot d_{n-1}^{2r} \\ P\left(k_n^{4r}\right) = (1-q)\,P\left(k_{n-1}^{2r}\right), & d_n^{4r} = ECD_n + u \cdot d_{n-1}^{2r} \\ r = 1, \ldots, 2^{n-2}. \end{cases} \tag{25}$$

It can be seen that d_n depends on u, v, ECD_n, p, and q. The former three parameters depend on the video sequence. The parameter pair p and q is used to describe the Gilbert channel and is equivalent to another parameter pair, PLR and ABL, which are more commonly used. Then, for video transmission over a Gilbert channel, given the average packet loss ratio PLR, the average burst length ABL, the initial probability distribution $P(k_1^1)$ and $P(k_1^2)$, and the initial distortion distribution d_1^1 and d_1^2, the expected distortion of each frame in a GOP can be estimated via a frame recursion approach using (24) and (25).

Next, the cumulative expected distortion over the entire GOP D_N can be defined as $D_N = \sum_{n=1}^{N} d_n$. Note that the sequence level expected distortion D_N can be used as an objective metric to assess the average video quality. Using the proposed method, D_N can be directly derived as explained next. We define $D(k_n^r)$ as the total distortion of a sequence from the first P-frame to the n-th P-frame, for a given loss pattern k_n^r. Hence, D_N can be estimated by taking an expectation over all possible loss patterns as follows:

$$D_N = \sum_{r=1}^{2^N} D\left(k_N^r\right) \cdot P\left(K_N^r\right). \tag{26}$$

With the help of the distortion model in (20) and (23), $D(k_n^r)$ can be calculated as follows,

$$\begin{cases} D\left(k_n^{2r}\right) = D\left(k_{n-1}^{r}\right) + d_n^{2r}, \\ D\left(k_n^{2r-1}\right) = D\left(k_{n-1}^{r}\right) + d_n^{2r-1}, \quad \text{for } r = 1, 2, \ldots, 2^{N-1}. \end{cases} \tag{27}$$

where $P(k_N^r)$ in (26) and d_n^{2r}, d_n^{2r-1} in (27) can be calculated using (25). Then, using (27), the total distortion for an arbitrary loss pattern can be calculated. The formula in (26) provides a way to estimate and analyze the impact of the bursty loss behavior on the average video quality.

Using the Distortion Trellis model, one can estimate the expected distortion d_n caused by Markov-model bursty losses, at the encoder/sender. However, the model often fails to compute d_n within acceptable time. In particular, when calculating d_n, one needs to compute the terms d_n^r and $P(k_n^r)$ associated with $r = 1, \ldots, 2^n$. Consequently, the complexity for calculating d_n is $O(2^n)$ while that for calculating D_N is $O(N2^n)$. Thus, it is desirable to develop a low-complexity algorithm for distortion estimation which is described next.

4.3 Sliding Window Algorithm

The Distortion Trellis model considers that all previous frames in the same GOP could affect the distortion of the current frame due to error propagation. However, the propagation of error typically decays in magnitude over the subsequent frames due to the effects of intra refreshing and spatial filtering. Therefore, it is a reasonable assumption that the distortion of frame n is independent of the transmission state of frame m, when $|m-n| > W$, where W is an integer constant, as argued and verified throughout this chapter. Based on this assumption, the authors propose a sliding window (SW) algorithm to calculate d_n for $n > W$ with low complexity. For $n \leq W$, we employ the same approach described previously to calculate d_n.

In the spirit described above, for every frame n, for $n > W$, the SW algorithm associates a corresponding sequence segment, or a window W, comprising frames $n-W+1$ to n, which loss patterns are exclusively considered for calculating the corresponding distortion d_n. Specifically, we assume that the first frame in the segment $n-W+1$ is either received with probability $P(k_1^1)$ or lost with probability $P(k_1^2)$, independently of any frame prior to it, and the corresponding distortion is d_1^1 and d_1^2, respectively. The loss process of the frames within W is also considered to be a two-state Markov process, or a Gilbert process. In such a case, there are in total 2^W loss patterns that should be considered for each frame $n > W$. This means that when calculating d_n using (16), we only need to calculate 2^W corresponding decoder distortion values rather than 2^n. The window slides ahead one frame at a time, and the expected distortion d_n for all $n > W$ can be obtained in this manner. It can be seen that, instead of considering the loss process of all P-frames in a GOP as Markovian, the SW algorithm limits the Markov loss process within each window W and ignores the frames outside the window. The overall SW algorithm is summarized in Algorithm 1 below.

Algorithm 1: SW for calculating d_n for $n = 1, \ldots, N$

```
Input: $PLR, ABL, u, v, W, N, \{ECD_n, n = 1, \ldots, N\}$;
Output: the expected distortion $d_n$ for $n = 1, \ldots, N$;
Procedure:
Initialization: $d_1^1 = 0, d_1^2 = ECD_1, P(k_1^1) = 1 - PLR,$
  $P(k_1^2) = PLR, p = PLR/(ABL(1 - PLR)), q = 1/ABL$;
for $n = 1$ to $N$ do
  if $n \leq W$ then
    Compute $d_n$ using (24) and (25), note that when
    $n = W$, $P(k_W^j), j = 1, \ldots, 2^W$ are obtained;
  else
    Reset the initial distortion values: $d_1^1 = 0$,
    $d_1^2 = ECD_{n-W+1}$;
    for $i = 2$ to $W$ do
```

```
            for $j = 1$ to $2^{i-1}$ do
                $d_i^{2j-1} = vd_{i-1}^j, d_i^{2j} = ECD_{n-W+i} + ud_{i-1}^j$;
            end for
        end for
        After the two for loops, $d_W^j, j = 1, \ldots, 2^W$
        are obtained, then the output is
        $d_n = \sum_{j=1}^{2^W} P\left(k_W^j\right) d_W^j$;
    end if
end for
```

The window length W is an important parameter of the SW algorithm. Generally, a big W leads to more accurate prediction but increases the algorithm's complexity. An appropriate W implies that the tradeoff between the estimation accuracy and the computation complexity is achieved. From (20) and (23), we can see that parameters u and v determine the distortion fading speed over consecutive frames, which can be considered when selecting the appropriate W. Generally, a small to middle u and v indicate quick fading, in which case a relatively small W may be acceptable.

The SW algorithm provides an efficient way to calculate d_n. To obtain d_n for $n > W$ based on the SW algorithm, only the quantities d_n^r and $P(k_n^r)$ associated with $r = 1, \ldots, 2^W$ need to be computed. Together with (26) and (27), the GOP level expected distortion can also be calculated. In [22] the authors establish that a window size $W \leq 16$ is sufficient to achieve acceptable prediction accuracy for most examined cases. Hence, the computation cost is reduced significantly compared to the original Distortion Trellis model. For example, when the SW algorithm with $W = 15$ is used to calculate $d_n, n = 1, \ldots, N$ for a GOP with $N = 36$, the number of iteration cycles that need to be run reduces from $\sum_{n=1}^{36} 2^n$ to $\sum_{n=1}^{15} 2^n + (36-15) \cdot 2^{15}$. This translates to a more than 90% reduction in computational complexity.

4.4 Performance Evaluation

Here, we conduct a series of comprehensive simulation experiments in order to asses the prediction performance of the Distortion Trellis framework as a function of various systems parameters and for different video contents. The performance of the low-complexity alternative is also carefully examined.

Simulation Setup

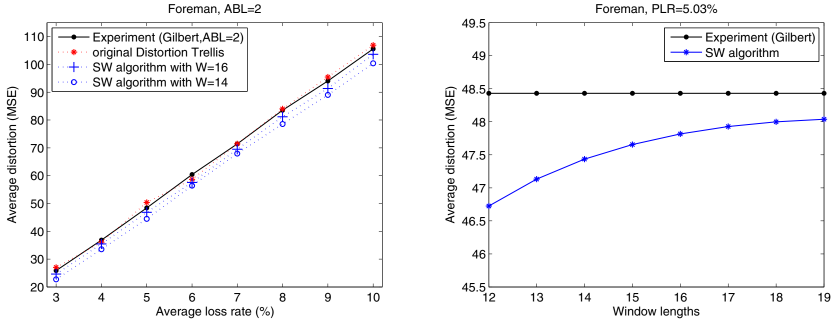
The H.264 reference software encoder JM12.2 [29] with the Baseline profile is used to encode the test sequences used in these experiments. Four QCIF sequences are used, the low motion sequence “News”, the moderate motion sequence “Foreman”, and the high motion sequences “Stefan” and “Football”. The former three are encoded at 15 fps while the sequence “Football” is coded at 30 fps. All sequences are encoded with a constant QP=28. The first frame is encoded as an I-frame, while the remaining frames are coded as P-frames with a forced intra refresh rate of 9/99 (every nine frames a row of MBs is intra refreshed in a round-robin fashion). Intra prediction and 1/4-pel motion estimation are enabled. We use one slice per frame and one frame per packet. At the decoder, the simple frame-copy scheme is used for concealment, so that the factor ECD_n in (20) can be easily pre-measured using $ECD_n = \frac{1}{XY}\sum_{i=1}^{XY}\left(x_n^i - x_{n-1}^i\right)$, where X and Y once again respectively denote the width and the height of frame n in pixels. The concealment frame is displayed instead of the missing frame, and is also stored in the reference frame buffer for decoding subsequent frames.

In the experiments, a range of values for the average packet loss rate is examined from 3% to 10%, and similarly the average burst length is varied in the range 1, 1.5, 2,. . . , 5. Each pair of PLR and ABL values is translated into the corresponding p and q values for the Gilbert channel. Then, with each pair of p and q we simulate a Gilbert packet loss process and generate 50,000 to 90,000 loss traces with random loss patterns. For each loss trace, we decode the video and calculate the MSE distortion between each transmitted and received P-frame. The expected distortion for each frame is then obtained by averaging the distortion of that frame over all traces. The GOP size for each sequence used is 390 for “Foreman”, 240 for “Football” and 200 for both “News” and “Stefan”. Finally, to estimate the model parameters u and v, in (20) and (23), respectively, we use a least square fitting method applied on the compressed video data.

Simulation results and analysis

In the first set of experiments, the measured average MSE distortion, the estimate using the original Distortion Trellis model, and the estimate based on the SW algorithm are all compared. Figure 12a plots the average expected distortion for PLR values from 3% to 10% at $ABL = 2$. Due to the high complexity of the original Distortion Trellis model, which is used as the performance benchmark here, we test the model over short sequence segments in this simulation. Particularly, we encode 20-frame segment starting at different positions in the original sequence. For each tested PLR and ABL pair, we generate 50,000 loss traces for each segment. The average expected distortion is then obtained by averaging over all segments and loss traces.

It can be seen that the original Distortion Trellis model provides better prediction of the expected distortion than the SW algorithm along the whole tested PLR range. Nonetheless, although the SW algorithm is less accurate, it still matches the measured expected distortion quite well. As expected, we see that the estimation curve using the SW algorithm is always under the experimental curve and that its performance improves as the window length increases. These plots indicate that the SW algorithm could be used as a



(a) (b)

Fig. 12. (a) Average distortion comparison; (b) average expected distortion over all P-frames of "Foreman" versus window length

very good approximation of the original model, especially at larger window lengths.

To examine the influence of the window length on the accuracy of the SW algorithm, Figure 12b plots the average expected distortion over all P-frames of "Foreman" versus the window lengths from 12 to 19 at $PLR = 5\%$ and $ABL = 2$, based on the same simulation data set as used in Figure 12a. We clearly see that the SW algorithm generally underestimates the expected distortion. This is because when calculating the expected distortion of the current frame, the distortion from frames outside the sliding window is ignored by the SW algorithm. In particular, we observe that smaller window lengths lead to larger estimation error values, because a smaller window ignores more distortion components from the past. We also observe that with the increase of the window length W from 12 to 16, the performance of the SW algorithm increases gradually, as discussed in Section 4.3. However, increasing the window length further does not bring as much performance gain. We believe that this is because the fading behavior of the impulse channel distortion often follows an exponential decay curve [30], or at least follows a similar degrading trend. That is why the SW algorithm performance does not increase linearly with the window length. In the experiments, the SW algorithm with $W \leq 16$ provides a fairly good performance for most of the examined cases. In addition, for some fast decaying an sequences even smaller W also provides acceptable results. Therefore, only the SW algorithm will be employed for estimating the expected distortion in the rest of the experiments considered in this section.

Next, the average expected distortion over all P-frames (i.e., the quantity D_N/N) versus PLR from 3% to 10%, for both ABL=2 and 5 is plotted in Figure 13. The tested sequences include "Foreman" and "Football". For "Foreman" the first 390 frames are coded while for "Football" the first 240

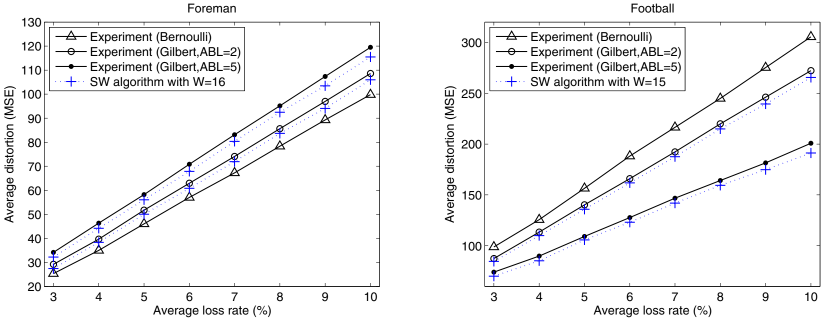
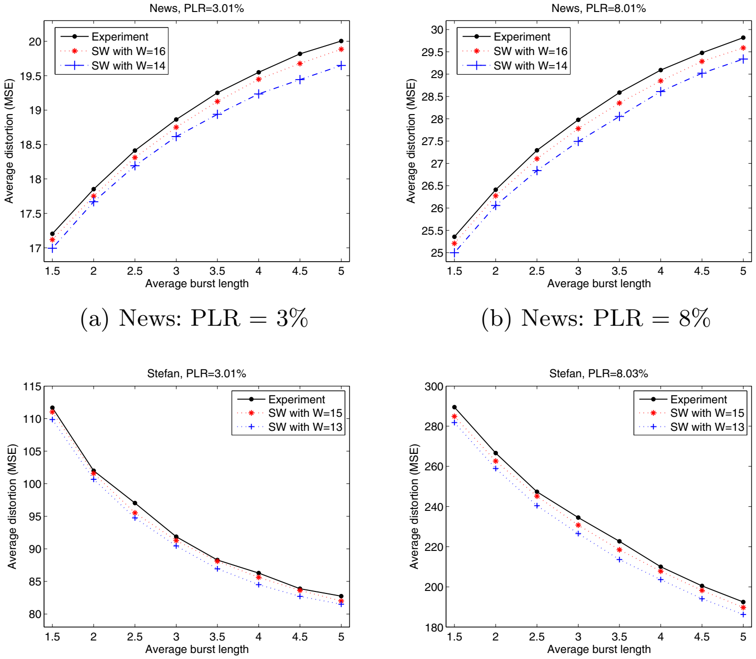


Fig. 13. Average distortion versus PLR: (left) Foreman and (right) Football. Corresponding distortions for a Bernoulli channel are also shown.

frames are coded. For each tested PLR and ABL pair, we simulate a Gilbert loss process and generate 90,000 random loss patterns. For each loss pattern, the distortion model in (27) is used to predict the decoder distortion. The window length used in the SW algorithm is 16 for "Foreman" and 15 for "Football". The average expected distortion for the case of a Bernoulli channel at the same loss rate is also plotted in the same figure for comparison, where 1000 loss traces are generated at each loss rate for this channel model.

We can see that the SW algorithm accurately estimates the average expected distortion over most of the range of the average loss rate. At high loss rate, the SW is less accurate, but still well matches the actual distortion curve. The good match between the theoretical data and the measured data tells us that the proposed model can be used to estimate and analyze the impact of bursty losses on the average video quality. Moreover, we see that though the window length used for "Football" is smaller than that for "Foreman", the accuracy of the SW algorithm is similar. We will discuss this later. From both figures, we also see that at the same ABL, the average distortion increases linearly with the PLR.

Interestingly, we observe from Figure 13 that for the same average loss rate the expected distortion for the Gilbert channel can be smaller or larger than that for the Bernoulli channel depending on the video content. Specifically, for the "Foreman" sequence the former is larger while for the "Football" sequence the opposite is true. Even more interestingly, we observe that increasing the average burst length does not always contribute to a larger expected distortion, for a given average loss rate. For example, in the case of "Foreman" a larger ABL leads to a larger expected distortion, for the same PLR, as seen from Figure 13 (left). However, the opposite holds true in the case of "Football", as seen from Figure 13 (right). To the best of our knowledge, the aforementioned experimental results has been reported for the first time in the present work. Notice that the latter result in fact contrasts the earlier finding reported in the context of the first distortion model reviewed



(a) News: PLR = 3%

(b) News: PLR = 8%

(c) Stefan: PLR = 3%

(d) Stefan: PLR = 8%

Fig. 14. Average expected distortion versus ABL for PLR = 3% or 8%

in Section 2 that video distortion always increases as a function of the burst length. Still, the results presented in Figure 13 prove again that burst length does matter, i.e., it does affect video quality, as also corroborated by this earlier model [20, 23].

To study further the impact of the average burst length on the average video quality, in Fig. 14 we show the average expected distortion over all P-frames (D_N/N) versus burst length of 1, 1.5,. . . ,5 for a given packet loss rate (3% or 8%). The tested sequences comprise “News” and “Stefan”[11]. For each sequence the first 200 frames are coded. The examined loss rates include 3% and 8%. For each tested PLR and ABL pair, 60,000 loss traces are generated. It can be seen from Fig. 14 that the estimated average distortion matches the measured data well along the whole range of tested burst length values and for the two packet loss rates. In addition, we can clearly observe that the average expected distortion D_N/N does NOT always increase as the average burst length increases for a given average packet loss rate. Specifically, in the case of “Stefan” increasing the average burst length reduces D_N/N, while for “News” the opposite is true. These results confirm

[11] Due to space constraints, the corresponding results for “Foreman” and “Football” are not included. Still, they only confirm the findings reported here.

the earlier findings from Fig. 13 that for the same average loss rate, a larger average burst length does not always lead to a larger distortion in the case of a Gilbert channel.

As mentioned earlier, the preceding observations seem different from those reported in [20, 23], where it was found that "longer burst length always causes larger MSE distortion". We believe that this is due to a difference in the experimental setup. In particular, the work in [20, 23] aims to test if burst length matters, and therefore it measures the total distortion at the decoder versus varying burst lengths, implying that the average loss rates that are used are proportional to each burst length. While in the present experiments the authors considered how the average burst length affects the expected distortion if the average loss rate remains constant. In particular, from the Gilbert channel model shown in Fig. 9, we can see that increasing ABL while keeping PLR constant implies at the same time reducing the parameters q and p proportionally (recall that $ABL = 1/q$ and $PLR = p/(p+q)$).

In the following, we employ distortion component analysis to study further this last set of results. In particular, from (26), it is clear that D_N is the sum of 2^N components, i.e., $D_N = \sum_{r=1}^{2^N} D\left(k_N^r\right) \cdot P\left(h_N^r\right)$. Define $\mathcal{D}_r = P\left(k_N^r\right) \cdot D\left(k_N^r\right)$ as the r-th component. We observe that several components are much larger than almost all the other components. For the D_N versus ABL curve, these large components will determine its trend. Many other components are relatively small and thus make less contribution to D_N, though they may still affect the shape of the D_N versus ABL curve.

We first try to explain why D_N/N is an increasing function of ABL for "Foreman" and "News". Specifically, we employ multiple 10-frame segments taken at different positions in the "Foreman" sequence to calculate an "average D_{10}" and plot all components of the average D_{10} versus ABL in Fig. 15. We discover that some relatively large components monotonically increase with ABL, which makes D_{10} an increasing function of ABL. Although many other components decrease with ABL, they are relatively small and thus cannot influence the overall trend of D_N as a function of ABL. Note that the component $\mathcal{D}_{2^{10}}$ (the red curve in Fig. 15 increases quite quickly and becomes much larger than all the others starting from $ABL = 3.5$. The quantity $D_{10} - \mathcal{D}_{2^{10}}$ is also plotted with a dashed line. We can see that without $\mathcal{D}_{2^{10}}$, the total distortion becomes a decreasing function of ABL for $ABL \geq 4$. This implies that the single component $\mathcal{D}_{2^N}$ contributes the most to make D_N increase with ABL at high average burst lengths. Additionally, note that $\mathcal{D}_{1+2^9}$ (plotted in yellow) is the main decreasing component. The same analysis applies to the "News" sequence. These results are not included here to conserve space.

Furthermore, it is an interesting observation from Figs. 14(a-b) that the rate of increase of the average distortion D_N/N gradually decreases as ABL increases. We believe that this is because there are still many components of D_N decreasing with ABL, as shown in Fig. 15. Though these components are too small to make the trend of D_N versus ABL curve change from upward

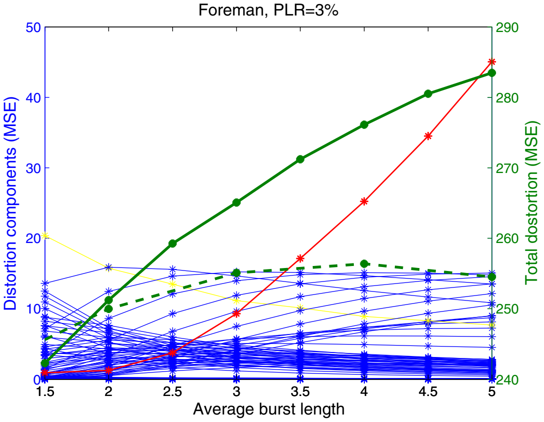


Fig. 15. Distortion components $\mathcal{D}_r, r = 1, \ldots, 2^{10}$ of D_{10} versus ABL. The red curve is $\mathcal{D}_{2^{10}}$. The green solid curve is D_{10}. The green dash curve is $D_{10} - \mathcal{D}_{2^{10}}$. The yellow curve is $\mathcal{D}_{1+2^9}$.

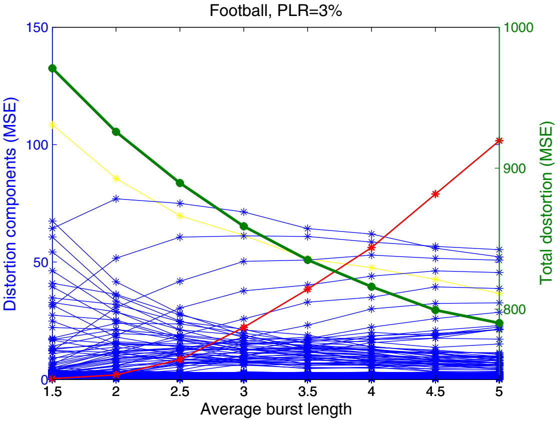


Fig. 16. Distortion components $\mathcal{D}_r, r = 1, \ldots, 2^{10}$ of D_{10} versus ABL. The red curve is $\mathcal{D}_{2^{10}}$. The green solid curve is D_{10}. The yellow curve is $\mathcal{D}_{1+2^9}$.

to downward, they still slow down the increasing rate of D_N as a function of ABL. In other words, they gradually decrease the slope of the D_N versus ABL curve.

Next, we apply the same analysis to explain why D_N/N is a decreasing function of ABL for "Stefan" and "Football". Using the same approach for creating Fig. 15, we compute the 10-frame average D_{10} for the "Football" sequence and plot in Fig. 16 all components of D_{10} for ABL from 1.5 to 5 at $PLR = 3\%$. Compared to the corresponding graphs from Fig. 15 it can be seen that now many distortion components in Fig. 16 have a similar shape, however exhibit different relative magnitudes. For example, the component $\mathcal{D}_{2^{10}}$ (the red curve in Fig. 16) is not that large now while $\mathcal{D}_{1+2^9}$ becomes a large and important component. Finally, from Figs. 14(c-d)[12] we can see that the rate of decrease of the average distortion D_N/N reduces as ABL increases. This is due to the fact that many small but increasing components slow down the decreasing rate of D_N as a function of ABL, as shown in Fig. 16. The results for "Stefan" can be explained in a similar fashion and are not included here for space considerations.

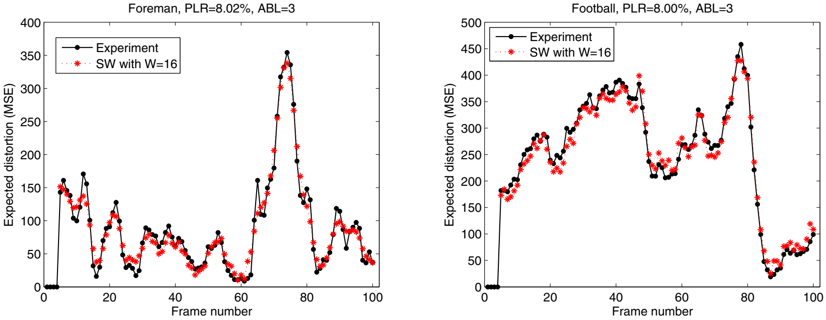


Fig. 17. Expected distortion versus frame number for ABL = 3 and PLR = 8% in the case of Foreman (left) and Football (right).

Finally, Fig. 17 shows the expected distortion versus frame number for "Foreman" and "Football" at $PLR = 8\%$ and $ABL = 3$. For each sequence the first 100 frames are coded. We can see that the estimated distortion using the SW algorithm with $W = 16$ fits the measured distortion values well. This tells us that the proposed model predicts well the frame level expected distortion and therefore can be employed to design efficient frame-based error resilient techniques for video transmission over burst loss channels.

In summary, it can be seen that the distortion trellis framework provides an accurate description of the loss-induced MSE affecting a transmitted video content, both at the frame-level as well as overall on a sequence level. Furthermore, we saw that through this framework we can analytically relate some very interesting discoveries observed in the simulation experiments such as increasing or decreasing video distortion as a function of burst loss length. The authors in [22] provided further experimental evidence supporting the findings included here. In addition, they extended their framework to the more general case of a finite-state Markov model and demonstrated its improved accuracy for distortion prediction in actual Internet experiments. These last two sets of results are not included here due to space constraints.

[12] The distortion vs. ABL curves are analogous for "Football" and "Stefan".

5 Related Work

Depending on the content level (pixel, macro-block, frame/slice/packet, or GOP) at which a model performs its computations, all related work on error propagation and distortion modeling in lossy video communication can be broadly classified into four categories. From the pixel-based approaches, we cite here the recursive optimal per-pixel estimation (ROPE) method introduced in [31] and its extensions [32–35]. The model proposes to compute the frame-level and sequence-level distortion values based on recursive per-pixel estimates of the expected reconstruction MSE affecting the video content in the event of packet loss. Similarly, [36, 37] employ the same recursive principle to compute and track the average distortion at each macro-block of a video frame. However, no consideration is given to the specific packet loss pattern nor to the statistical dependency of the individual losses that the transmitted video content experiences in any of the works above[13].

As mentioned throughout the chapter, most of the packet-level distortion estimation techniques proposed to date consider a linear model. That is the overall distortion affecting the reconstructed sequences is proportional to the number of packet losses experienced during transmission, i.e., the average packet loss rate. For instance, [18, 38] model the influence of intra-refresh rate and spatial filtering on the error propagation associated with a single packet loss. Then, linearity and superposition are assumed to synthesize the distortion associated with multiple losses. Similarly, the work in [19] employs such a liner superposition model in the context of wireless video streaming. In the spirit of [31], the works in [26, 27] propose frame-based recursive techniques for distortion estimation. However, again only the average packet loss rate is considered in the analysis disregarding therefore the influence of specific packet loss patterns and the statistical correlation of the channel induced packet loss. Finally, the studies in [39, 40] design methods for computing the GOP-level transmission-induced distortion in mobile video streaming.

Works where distortion models have been employed for other applications include [41] that employs a linear model for in-network monitoring of video quality of compressed streams subject to packet loss. In particular, content specific information and the effect of error propagation are directly assessed from the compressed stream. This information is then mapped to a specific video quality level using a linear distortion model that only takes into account the overall loss rate, as explained earlier. Similarly, the work in [42] develops a novel perceptually-driven video quality metric that again takes into account the effect of multiple losses through a linear additive relationship. Still, the effect of inter-packet loss separation is taken into consideration by augmenting the metric with a heuristic multiplying factor denoted cluster degree. Though no formal analysis is provided, from their experiments,

[13] It should be mentioned that [34] is the only work reviewed here that considers correlated packet loses. However, the employed recursive computation still prevents differentiating the impact of various loss patterns.

the authors observe that closely grouped losses have a stronger influence on perceptual quality, which confirms the analytical and simulation findings of the studies covered in depth in this chapter. Finally, distortion models that circumvent the effect of error-propagation by assuming frame-freeze concealment at the decoder[14] have been considered, e.g., in [43–45] in the context of rate-distortion optimized video streaming.

6 Conclusions and Open Challenges

The distortion models studied in this chapter have substantially advanced the state-of-the-art in characterizing the impact of data loss on reconstructed video quality. Some of their most important discoveries include accounting for the effect of burst loss patterns, demonstrating the existence of negative distortion in the event of further packet loss, and characterizing the impact of the statistical dependencies of the communication channel. By capturing many subtle effects related to video compression, packet loss processes, and their interplay these advanced models have been able to provide a more accurate prediction performance and a deeper understanding of many complexities that arise when predictively encoded content is transmitted over an unreliable channel. Still, there are a few important challenges that remain to be tackled.

For instance, in high data rate video a single frame may be broken down into multiple packets before transmission. This reduces the impact of burst loss dramatically as recognized, e.g., in [41, 46] since the multiple packets lost in a row may still belong to the same video frame. Therefore, from the perspective of the compressed content such losses will be analogous to the loss of a single frame. Similarly, video content is frequently transmitted protectively encoded, i.e., in concert with forward-error correction (FEC) data in order to alleviate the effect of channel-induced packet loss. Therefore, an interesting direction of further research in distortion modeling would be to extend the present models to capture the impact of FEC. As in the case of high date-rate video, adding such protective layers of data would reduce the level of end-to-end burstiness in terms of packet loss that the video decoder observes on the receiving end. Yet another relevant issue that would need to be carefully addressed is the fact that different video codecs employ at present different error concealment techniques. Specifically, in the event of loss of some slices of a frame MPEG-2 [2] conceals the whole frame with its reconstructed predecessor. In contrast, H.264 [6] still employs the correctly received slices to reconstruct the present picture where the missing data is interpolated from the these slices and suitably selected content from the previous frame. Clearly, these two different concealment strategies will result

[14] The first loss-affected frame is concealed with its temporally nearest predecessor that is already decoded. This content remains displayed, i.e., frozen (hence the name) until the successful decoding of a subsequent intra-coded frame.

in two different distortion-loss dependencies that will need to be individually recognized. Finally, the advent of multi-view imaging opens up another prospective avenue of interesting research on distortion modeling. In such a setting, it is crucial to account for the additional inter-view dependencies that arise if an accurate assessment of the reconstruction quality in the event of data loss is desired.

References

1. Telecom. Standardization Sector of ITU, Video coding for low bitrate communication. ITU-T Recommendation H.261 (1990)
2. ISO/IEC, Information technology — generic coding of moving pictures and associated audio information: Video (MPEG-2) International Standard 13818-2:2000 (1996)
3. Telecom. Standardization Sector of ITU, Video coding for low bitrate communication ITU-T Recommendation H.263 (March 1996)
4. Telecom. Video coding for low bitrate communication, ITU-TRecommendation H.263 Version 2 (February 1998)
5. ISO/IEC, Information technology — coding of audio-visual objects part 2: Visual (MPEG-4), JTC1/SC29/WG11 N4350, International Standard 14496-2 (July 2001)
6. Telecom. Standardization Sector of ITU, Video coding for low bitrate communication, Draft ITU-T Recommendation H.264 (March 2003)
7. ITU-T and ISO/IEC JTC 1, Advanced video coding for generic audiovisual services, amendment 3: Scalable video coding, Draft ITU-T Recommendation H.264 - ISO/IEC 14496-10(AVC) (April 2005)
8. Kell, R.D.: Improvements relating to electric picture transmission systems. UK Patent Specification No. 341,811
9. Mitchell, J.L., Pennebaker, W.B., Fogg, C.F., LeGall, D.J.: MPEG Video Compression Standard. Chapman and Hall, Boca Raton (1997)
10. Haskell, B.G., Puri, A.: Digital Video: An Introduction to MPEG-2. Chapman & Hall, New York (1997)
11. Wiegand, T., Girod, B.: Multi Frame Motion-Compensated Prediction for Video Transmission. Springer, Heidelberg (2001)
12. Approaching the zettabyte era. Cisco Visual Networking Index. Cisco Inc. (June 2008)
13. Albanese, A., Blömer, J., Edmonds, J., Luby, M., Sudan, M.: Priority encoding transmission. IEEE Trans. Information Theory 42, 1737–1744 (1996)
14. Budagavi, M., Gibson, J.D.: Multiframe video coding for improved performance over wireless channels. IEEE Trans. Image Processing 10(2), 252–265 (2001)
15. Chakareski, J., Chou, P., Aazhang, B.: Computing rate-distortion optimized policies for streaming media to wireless clients. In: Proc. Data Compression Conference, pp. 53–62. IEEE Computer Society, Snowbird (2002)
16. Chou, P.A., Miao, Z.: Rate-distortion optimized streaming of packetized media. IEEE Trans. Multimedia 8(2), 390–404 (2006)

17. Apostolopoulos, J.: Reliable video communication over lossy packet networks using multiple state encoding and path diversity. In: Proc. Conf. on Visual Communications and Image Processing, vol. 4310, pp. 392–409. SPIE, San Jose (2001)
18. Stuhlmüller, K., Färber, N., Link, M., Girod, B.: Analysis of video transmission over lossy channels. IEEE J. Selected Areas in Communications 18(6), 1012–1032 (2000)
19. Kim, I.-M., Kim, H.-M.: A new resource allocation scheme based on a PSNR criterion for wireless video transmission to stationary receivers over gaussian channels. IEEE Trans. Wireless Communications 1(3), 393–401 (2002)
20. Liang, Y., Apostolopoulos, J., Girod, B.: Analysis of packet loss for compressed video: Does burst-length matter. In: Proc. Int'l Conf. Acoustics, Speech, and Signal Processing, vol. 5, pp. 684–687. IEEE, Hong Kong (2003)
21. Chakareski, J., Apostolopoulos, J., Tan, W.-T., Wee, S., Girod, B.: Distortion chains for predicting the video distortion for general packet loss patterns. In: Proc. Int'l Conf. Acoustics, Speech, and Signal Processing, vol. 5, pp. 1001–1004. IEEE, Montreal (2004)
22. Li, Z., Chakareski, J., Niu, X., Xiao, G., Zhang, Y., Gu, W.: Modeling and analysis of distortion caused by Markov-model burst packet losses in video transmission. IEEE Trans. Circuits and Systems for Video Technology (August 2008) (to appear)
23. Liang, Y., Apostolopoulos, J., Girod, B.: Analysis of packet loss for compressed video: Does burst-length matter. IEEE Trans. Circuits and Systems for Video Technology 18(7), 861–874 (2008)
24. Chakareski, J., Apostolopoulos, J., Wee, S., Tan, W.-T., Girod, B.: Rate-distortion hint tracks for adaptive video streaming. IEEE Trans. Circuits and Systems for Video Technology 15(10), 1257–1269 (2005); special issue on Analysis and Understanding for Video Adaptation
25. Gilbert, E.N.: Capacity of a burst-noise channel. Bell Syst. Tech. Journal 39, 1253–1266 (1960)
26. He, Z.H., Cai, J.F., Chen, C.W.: Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding. IEEE Trans. Circuits and Systems for Video Technology 12(6), 511–523 (2002)
27. Wang, Y., Wu, Z., Boyce, J.M.: Modeling of transmission-loss-induced distortion in decoded video. IEEE Trans. Circuits and Systems for Video Technology 16(6), 716–732 (2006)
28. Girod, B., Färber, N.: Feedback-based error control for mobile video transmission. Proceedings of the IEEE 87(10), 1707–1723 (1999)
29. H.264/AVC Reference Software JM12.2, http://iphome.hhi.de/suehring/tml/download/old_jm/jm12.2.zip
30. He, Z.H., Xiong, H.K.: Transmission distortion analysis for real-time video encoding and streaming over wireless networks. IEEE Trans. Circuits and Systems for Video Technology 16(9), 1051–1062 (2006)
31. Zhang, R., Regunathan, S.L., Rose, K.: Video coding with optimal inter/intra-mode switching for packet loss resilience. IEEE J. Selected Areas in Communications 18(6), 966–976 (2000)
32. Reibman, A.: Optimizing multiple description video coders in a packet loss environment. In: Proc. Int'l Packet Video Workshop, Pittsburgh, USA (April 2002)

33. Yang, H., Rose, K.: Recursive end-to-end distortion estimation with model-based cross-correlation approximation. In: Proc. Int'l Conf. Image Processing, vol. 3, pp. 469–472. IEEE, Barcelona (2003)
34. Heng, B.A., Apostolopoulos, J.G., Lim, J.S.: End-to-end rate-distortion optimized mode selection for multiple description video coding. In: Proc. Int'l Conf. Acoustics, Speech, and Signal Processing, vol. 5, pp. 905–908. IEEE, Philadelphia (2005)
35. Yang, H., Rose, K.: Advances in recursive per-pixel end-to-end distortion estimation for robust video coding in H.264/AVC. IEEE Trans. Circuits and Systems for Video Technology 17(7), 845–856 (2007)
36. Côté, G., Shirani, S., Kossentini, F.: Optimal mode selection and synchronization for robust video communications over error-prone networks. IEEE J. Selected Areas in Communications 18(6), 952–965 (2000)
37. Ekmekci, S., Sikora, T.: Recursive decoder distortion estimation based on AR(1) source modeling for video. In: Proc. Int'l Conf. Image Processing, vol. 1, pp. 187–190. IEEE, Singapore (2004)
38. Färber, N., Stuhlmüller, K., Girod, B.: Analysis of error propagation in hybrid video coding with application to error resilience. In: Proc. Int'l Conf. Image Processing, vol. 2, pp. 550–554. IEEE, Kobe (1999)
39. Zhang, C., Yang, H., Yu, S., Yang, X.: Gop-level transmission distortion modeling for mobile streaming video. Signal Processing: Image Communication 23(2), 116–126 (2008)
40. Ivrlač, M.T., Choi, L.U., Steinbach, E., Nossek, J.A.: Models and analysis of streaming video transmission over wireless fading channels. Signal Processing: Image Communication (June 2009)
41. Reibman, A., Vaishampayan, V.: Quality monitoring for compressed video subjected to packet loss. In: Proc. Int'l Conf. Multimedia and Exhibition, vol. 1, pp. 17–20. IEEE, Baltimore (2003)
42. Liu, T., Wang, Y., Boyce, J.M., Yang, H., Wu, Z.: A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts. IEEE J. Selected Areas in Communications 3(2), 280–293 (2009)
43. Chakareski, J., Girod, B.: Rate-distortion optimized packet scheduling and routing for media streaming with path diversity. In: Proc. Data Compression Conference, pp. 203–212. IEEE Computer Society, UT (2003)
44. Cheung, G., Tan, W.-T.: Directed acyclic graph based source modeling for data unit selection of streaming media over qos networks. In: Proc. Int'l Conf. Multimedia and Exhibition, vol. 2, pp. 81–84. IEEE, Lausanne (2002)
45. Tu, W., Chakareski, J., Steinbach, E.: Rate-distortion optimized frame dropping for multi-user streaming and conversational video. Hindawi Journal on Advances in Multimedia (2) (January 2008); special issue on Collaboration and Optimization for Multimedia Communications
46. Tao, S., Apostolopoulos, J., Guérin, R.: Real-time monitoring of video quality in IP networks. IEEE/ACM Trans. Networking 16(5), 1052–1065 (2008)

Index

Author Index