Computer Analysis of Human Behavior

Albert Ali Salah • Theo Gevers
Editors

Computer Analysis of Human Behavior

Springer

Editors
Albert Ali Salah
Department of Computer Engineering
Boğaziçi University
Bebek, Istanbul 34342
Turkey
salah@boun.edu.tr

Theo Gevers
Informatics Institute
University of Amsterdam
Science Park 904
1098 XH Amsterdam
Netherlands
th.gevers@uva.nl

ISBN 978-0-85729-993-2 e-ISBN 978-0-85729-994-9
DOI 10.1007/978-0-85729-994-9
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2011939770

Cover design: VTeX UAB, Lithuania

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Overview and Goals

Human behavior is complex, but not random. Computer analysis of human behavior in its multiple scales and settings leads to a steady influx of new applications in diverse domains like human-computer interaction, affective computing, social signal processing, and ambient intelligence. We envision seamlessly integrated plug and play devices that can be used to endow a given environment with an awareness of the physical, functional, temporal and social organization of its internal domestic dynamics, as well as the personalities and social relationships of its inhabitants, providing a vast array of new services for ordinary people. We picture intuitive tools for social scientists, psychologists and doctors to observe and quantify human behavior. We desire realistic virtual agents and engaging robots that can analyze and properly respond to social context. We seek intelligent algorithms to process vast collections of multimedia information to retrieve relevant material in response to semantic queries.

The realization of all these tools and systems requires a fundamental grasp of the key issues, as well as knowledge and experience over the computational tools. The goal of this book is to provide a solid foundation toward achieving this.

Most significantly, the focus of the book is in advanced pattern recognition techniques to automatically interpret complex behavioral patterns generated when humans interact with machines or with others. This is a challenging problem where many issues are still open, including the joint modeling of behavioral cues taking place at different time scales, the inherent uncertainty of machine detectable evidences of human behavior, the mutual influence of people involved in interactions, the presence of long term dependencies in observations extracted from human behavior, and the important role of dynamics in human behavior understanding.

The contiguity of these problems with the field of pattern recognition is straightforward. The editors of the present volume (together with Alessandro Vinciarelli from the Univ. of Glasgow and Nicu Sebe from the Univ. of Trento) organized the First Workshop of Human Behavior Understanding as a satellite workshop to Int. Conference on Pattern Recognition (ICPR) in 2010, with wide attendance. Similarly,

the Human Communicative Behaviour Analysis Workshop is held as a satellite to IEEE's Computer Vision and Pattern Recognition (CVPR) Conference since 2008. The topics presented in the book are actively researched in the pattern recognition community.

Target Audience

The book is planned as a graduate textbook/guidebook on computer analysis of human behavior. Starting from the preliminaries, it covers major aspects concisely, introduces some of the most frequently used techniques and algorithms in detail, and provides examples of real applications. Each chapter is a stand-alone treatment of a relevant subject, discussing key issues, classic and recent approaches, as well as open questions and future research directions. Since the subject matter is very broad, we have restricted the number of chapters to ensure that the book can be covered in one semester and focused on providing both a good background and a comprehensive vision of the field. Each chapter is supplemented with educational material, including chapter summary, glossary, questions, and online lecture slides to help the instructor.

Organization of the Book

We have divided the book into four parts. The first part, called "*The Tools of the Trade*", is a selection of basic topics that the reader will repeatedly come across in human behavior analysis. These chapters are all written in an intuitive and didactic way, and pave the way for more advanced discussions.

The first of these four chapters introduces Bayesian methods for behavior analysis, because there are numerous uncertainties in measuring and assessing behavior, and one also needs to deal with idiosyncrasies and inconsistencies. In particular, Gaussian processes and Dirichlet processes are covered in this chapter.

Almost all human behavior is temporal, but the time scale may range from milliseconds (e.g. the movement of a facial muscle) to hours (e.g. sleep cycles) or months (e.g. habits). The temporal dimension is the sine qua non of human behavior, and subsequently, the second chapter introduces basic methods for temporal analysis, including Hidden Markov Models, Conditional Random Fields, and variants thereof. This chapter also gives a concise introduction to graphical models, their factorization and how to perform inference in graphical models. Subsequently, it complements the first chapter nicely.

The third chapter discusses how we can detect and track humans by computer vision methods to understand their actions, and is a prerequisite for most material presented in Parts II and III. The visual modality can provide a system with high-dimensional and dynamic data, creating some of the most formidable challenges in behavior analysis, and most analysis pipelines start with detection and tracking.

The fourth chapter is an introduction to computational visual attention, and explains how humans make sense of the immensely rich perceptual input (*the firehose of experience*, in the memorable expression of Ben Kuijpers), as well as how computer systems can mimic the process of attention to reduce their computation load, especially in terms of bottom-up (data driven) and top-down (semantically driven) approaches. The VOCUS attention system is described in detail, and applications are given from the field of mobile robotics.

Two major application areas of human behavior analysis are activity recognition, and analysis of social signals, including those that pertain to affect. The second part of the book is devoted to "*Analysis of Activities*", whereas the third part deals with social and affective behavior.

The first chapter in the second part (i.e. Chap. 5) is on gait and posture analysis, which is a relevant problem for clinical applications, surveillance, and ergonomics. In this chapter the reader will be introduced to a host of sensors (like gyroscopes and accelerometers) that can be used to measure useful physiological and movement related signals. Most of these sensors are easily integrated into smart phones, creating a huge potential for mobile phone applications (and games) that are based on human behavior analysis.

The second part continues with Chap. 6 on hand gesture analysis. Hand gestures can be used to define natural interfaces in human-computer interaction, but they are also rich sources of social and contextual cues during conversations. This chapter builds heavily on temporal analysis and tracking material of the first part.

Automatic analysis of complex human behavior is, as we mentioned earlier, one of the grand challenges of multimedia retrieval. Some behaviors are simple (e.g. walking) and can be detected by looking at simple cues. Some behaviors are complex (e.g. flirting) and require extensive knowledge and processing of context. Chapter 7 is on semantics of human behavior in image sequences, and discusses how environment influences the perceived activities, and how bottom-up and top-down approaches can be integrated for recognizing events.

With the third part of the book, "*Social and Affective Behaviors*", we move toward applications where pattern recognition and machine learning methods need to be complemented with psychological background knowledge and models. Social behaviors constitute a major research area, largely overlapping with the field of affective computing. This part opens up with a psychological treatise on social signals, written in a very accessible way for an audience with mainly computer science background. A wide range of signals like dominance, persuasion, shame, pride, and enthusiasm are introduced and discussed in Chap. 8.

Poggi and D'Errico define a *social signal* as a communicative or informative signal that, either directly or indirectly, conveys information about social actions, social interactions, social emotions, social attitudes and social relationships (2010). The audio modality is more frequently used for processing such signals compared to the visual modality. Chapter 9 is an extensive and technical discussion of voice and speech analysis for assessing human communicative behavior.

Chapter 10 is on analysis of affect from combined audiovisual cues. It discusses how affective signals can be measured and evaluated in a continuous manner, as

well as how to perform multimodal fusion. While the discussions of data annotation and experimental design in Chap. 10 pertain mainly to affective displays, identified issues and challenges are valid for almost the entire range of human behavior analysis.

The last chapter of the third part of the book, Chap. 11, discusses social interactions and group dynamics. Four case studies from a meeting scenario are presented, where the authors combine audiovisual cues to estimate the most and least dominant person, emerging leaders, and functional roles in the group, as well as the group dynamics as a whole.

The fourth part of the book is devoted to "*Selected Applications*" from three different research fields (ambient intelligence, biometrics, and gaming, respectively) that reflect the diversity and scope of behavioral cues and their usage.

Chapter 12 describes a vision of ambient assisted living, where a smart environment monitors the activities of its inhabitants for health care purposes. The aging population of developed countries call for technologies to allow elderly to remain longer in their home environments, giving them a higher quality of life, as well as reducing the costs of health care. Smart monitoring tools, provided that they deal with acceptance and privacy issues properly, are of great value.

Behavioral biometrics is the identification of a person via behavioral cues. Chapter 13 surveys this new field, and reveals that an astonishing number of behavioral cues, measured directly or indirectly, can be used to verify the identity of a person. This is a joint achievement of improvements in pattern recognition methods, as well as sensor technologies.

Finally, Chap. 14 deals with games, which are major economic drivers behind computer scientific research. Games do not only serve entertainment; there are games to exercise the body and mind, and sometimes a game is the means to a completely different end. Take for instance the robotics community, which uses robot soccer as a driver behind great advances in robot mechanics, coordination, planning, and a host of other challenges. This is an application area where both real and virtual human behavior can be analyzed for improving engagement and interaction, as well as for teaching computers and robots skills at human level.

Taken together, these chapters cover most of the field of human behavior analysis. It is possible to include many more tools and applications, as the field is positioned at a confluence of many different and mature research areas. Nonetheless, we hope that this collection will be a useful teaching tool for initiating newcomers, as well as a timely reference work that sums up recent research in this advancing area.

Amsterdam, The Netherlands

Albert Ali Salah
Theo Gevers

Contents

Part I Tools of the Trade

1 Bayesian Methods for the Analysis of Human Behaviour 3
Gwenn Englebienne

2 Introduction to Sequence Analysis for Human Behavior Understanding 21
Hugues Salamin and Alessandro Vinciarelli

3 Detecting and Tracking Action Content 41
Alper Yilmaz

4 Computational Visual Attention 69
Simone Frintrop

Part II Analysis of Activities

5 Methods and Technologies for Gait Analysis 105
Elif Surer and Alper Kose

6 Hand Gesture Analysis 125
Cem Keskin, Oya Aran, and Lale Akarun

7 Semantics of Human Behavior in Image Sequences 151
Nataliya Shapovalova, Carles Fernández, F. Xavier Roca, and Jordi Gonzàlez

Part III Social and Affective Behaviors

8 Social Signals: A Psychological Perspective 185
Isabella Poggi and Francesca D'Errico

9 Voice and Speech Analysis in Search of States and Traits 227
Björn Schuller

10 Continuous Analysis of Affect from Voice and Face 255
Hatice Gunes, Mihalis A. Nicolaou, and Maja Pantic

11 Analysis of Group Conversations: Modeling Social Verticality 293
Oya Aran and Daniel Gatica-Perez

Part IV Selected Applications

12 Activity Monitoring Systems in Health Care 325
Ben Kröse, Tim van Oosterhout, and Tim van Kasteren

13 Behavioral, Cognitive and Virtual Biometrics 347
Roman V. Yampolskiy

14 Human Behavior Analysis in Ambient Gaming and Playful Interaction 387
Ben A.M. Schouten, Rob Tieben, Antoine van de Ven, and David W. Schouten

Index 405

Contributors

Lale Akarun Computer Engineering Department, Boğaziçi University, Istanbul, Turkey, akarun@boun.edu.tr

Oya Aran Idiap Research Institute, Martigny, Switzerland, oya.aran@idiap.ch

Francesca D'Errico Roma Tre University, Rome, Italy, fderrico@uniroma3.it

Gwenn Englebienne University of Amsterdam, Science Park 904, Amsterdam, The Netherlands, G.Englebienne@uva.nl

Carles Fernández Departament de Ciències de la Computació and Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain, perno@cvc.uab.es

Simone Frintrop Institute of Computer Science III, Rheinische Friedrich-Wilhelms Universität Bonn, Römerstrasse 164, 53117 Bonn, Germany, frintrop@iai.uni-bonn.de

Daniel Gatica-Perez Idiap Research Institute, Martigny, Switzerland, gatica@idiap.ch

Jordi Gonzàlez Departament de Ciències de la Computació and Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain, poal@cvc.uab.es

Hatice Gunes Queen Mary University of London, London, UK, haticeg@ieee.org

Tim van Kasteren Boğaziçi University, Bebek, Istanbul, Turkey, tim0306@gmail.com

Cem Keskin Computer Engineering Department, Boğaziçi University, Istanbul, Turkey, keskinc@cmpe.boun.edu.tr

Alper Kose Department of Biomedical Sciences, University of Sassari, Sassari, Italy, akose@uniss.it

Ben Kröse University of Amsterdam, Amsterdam, The Netherlands, b.j.a.krose@uva.nl; Amsterdam University of Applied Science, Amsterdam, The Netherlands

Mihalis A. Nicolaou Imperial College, London, UK, mihalis@imperial.ac.uk

Tim van Oosterhout Amsterdam University of Applied Science, Amsterdam, The Netherlands, T.J.M.van.Oosterhout@hva.nl

Maja Pantic Imperial College, London, UK, m.pantic@imperial.ac.uk; University of Twente, Twente, The Netherlands

Isabella Poggi Roma Tre University, Rome, Italy

F. Xavier Roca Departament de Ciències de la Computació and Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain, xavir@cvc.uab.es

Hugues Salamin School of Computing Science, University of Glasgow, Glasgow, Scotland, hsalamin@dcs.gla.ac.uk

Ben A.M. Schouten Department of Industrial Design, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, b.a.m.schouten@tue.nl

David W. Schouten S. Nicolas Highschool, Prinses Irenestraat 21, 1077 WT Amsterdam, The Netherlands

Björn Schuller Institute for Human–Machine Communication, Technische Universität München, 80290 Munich, Germany, schuller@tum.de

Nataliya Shapovalova Departament de Ciències de la Computació and Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain, shapovalova@cvc.uab.es

Elif Surer Department of Biomedical Sciences, University of Sassari, Sassari, Italy, esurer@uniss.it

Rob Tieben Department of Industrial Design, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, r.tieben@tue.nl

Antoine van de Ven Fontys University of Applied Sciences, Postbus 347, 5600 AH Eindhoven, The Netherlands, antoine.vandeven@fontys.nl

Alessandro Vinciarelli School of Computing Science, University of Glasgow, Glasgow, Scotland, vincia@dcs.gla.ac.uk; Idiap Research Institute, Martigny, Switzerland

Roman V. Yampolskiy University of Louisville, Louisville, KY, USA, roman.yampolskiy@louisville.edu

Alper Yilmaz The Ohio State University, Columbus, OH 43035, USA, yilmaz.15@osu.edu

Acronyms

A-V	Arousal-valence
ADL	Activities of daily living
AF	Anatomical frame
ANN	Artificial neural network
ASR	Automatic speech recognition
BCI	Brain-computer interface
BLSTM-NN	Bi-directional long short-term memory neural network
BoW	Bag of words
CAST	Calibrated anatomical system technique
CRF	Conditional random field
DAG	Directed acylic graph
DCT	Discrete cosine transform
DLT	Direct linear transform
DoG	Difference of Gaussians
DP	Dirichlet process
DPMM	Dirichlet process mixture model
ECG	Electrocardiogram
EEG	Electroencephalography
EM	Expectation-Maximization algorithm
EMG	Electromyography
EOG	Electrooculogram
FEF	Frontal eye field
FFT	Fast Fourier transform
FIT	Feature Integration Theory
FMTL	Fuzzy metric temporal Horn logic
FOA	Focus of attention
FRCS	Functional role coding scheme
GF	Global frame
GP	Gaussian process
GSR	Galvanic skin response
HCI	Human–Computer interaction

HCRF	Hidden conditional random field
HEU	Human event understanding
HMM	Hidden Markov model
HNR	Harmonics-to-noise ratio
HoG	Histograms of oriented gradients
ICA	Independent components analysis
IOHMM	Input output hidden Markov model
IOR	Inhibition of return
IR	Infrared
IT	Infero-temporal cortex
KDE	Kernel density estimation
KF	Kalman filter
kNN	k-Nearest neighbor
LDA	Latent Dirichlet allocation
LDA	Linear discriminant analysis
LDCRF	Latent dynamic conditional random field
LGN	Lateral geniculate nucleus
LIP	Lateral intraparietal area
LLD	Low-level descriptors
LPC	Linear predictive coding
MAP	Maximum a posteriori
MARG	Magnetic, angular rate and gravity
MCMC	Markov chain Monte Carlo
MDF	Most discriminative features
MFCC	Mel-Frequency cepstrum coefficients
MLE	Maximum likelihood estimation
MMORPG	Massively multiplayer online role-playing games
MoG	Mixture of Gaussians
MRF	Markov random field
MSE	Mean squared error
MUD	Massively multi user dungeon
NMF	Non-negative matrix factorization
PAD	Pleasure-arousal-dominance
PCA	Principal components analysis
PDF	Probability density function
PF	Particle filter
PHOG	Pyramid of histograms of oriented gradients
PO	Parieto-occipital cortex
PP	Posterior parietal cortex
RCFL	Recursive coarse-to-fine localization
RF	Random forests
RFID	Radio frequency identification
RMSE	Root mean squared error
ROI	Region of interest
RPG	Role-playing video games

RT	Reaction time
SAL	Sensitive artificial listener
SAN	Social affiliation networks
SC	Superior colliculus
SGT	Situation graph tree
SIFT	Scale invariant feature transform
SLAM	Simultaneous localization and mapping
STIP	Space time interest points
SVM	Support vector machines
SVR	Support vector regression
TF	Technical frame
TOF	Time of flight
V1	Primary visual cortex
VDR	Visual Dominance Ratio
WoZ	Wizard-of-Oz
WTA	Winner-Take-All network

Part I
Tools of the Trade

Chapter 1
Bayesian Methods for the Analysis of Human Behaviour

Gwenn Englebienne

1.1 Bayesian Methods

Let us start by a brief recapitulation of the theory of Bayesian statistics, illustrating what the strengths of these methods are, and how their weaknesses have been overcome. At the core of Bayesian methods is the consistent application of the rules of probability in general, and Bayes' theorem in particular. Bayes' theorem,

$$p(B|A) = \frac{p(A|B)\, p(A)}{p(B)}, \tag{1.1}$$

states that the posterior probability of a random event B given event A, $p(B|A)$, is directly proportional to the *joint* probability of events A and B, $p(A, B) = p(A|B)p(B)$ and inversely proportional to the *marginal* probability of B. The marginal probability of a variable, in this case $p(B)$, is obtained by summing the joint probability $p(A, B)$ over all possible values of the other variables, A in this case, where a represents a particular value or instantiation of A:

$$p(B) = \sum_{a \in A} p(B, a). \tag{1.2}$$

If variables are continuous rather than discrete, the sum becomes an integral, but the general idea remains the same. This process is called *marginalisation*, and we describe what is happening in (1.2) as "marginalising out A", or computing the "marginal probability of B".

The different variables in the above equation can be vectors containing multiple variables, so that if we wanted to know the conditional probability of a subset B_1 of

G. Englebienne (✉)
University of Amsterdam, Science Park 904, Amsterdam, The Netherlands
e-mail: G.Englebienne@uva.nl

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_1,

$B = \{B_1, B_2\}$, $p(B_1|B_2)$, we would first marginalise out A from $p(A, B_1, B_2)$ and apply Bayes' theorem to the remainder of the variables.

$$p(B_1|B_2) = \frac{p(B_1, B_2)}{p(B_2)} \tag{1.3}$$

$$= \frac{\sum_{a\in A} p(A, B_1, B_2)}{\sum_{b\in B_1}\sum_{a\in A} p(A, B_1, B_2)}. \tag{1.4}$$

The key insight of Bayesian statistics is that the parameters of functions describing probability distributions are themselves random variables, which are not observed. They are, therefore, subject to the same rules of probability as any other random variables and should be treated in the same manner. For example, suppose that two variables B_1 and B_2 denote, respectively, what type of transport (car, bike, bus, train, ...) one uses and the current weather circumstances (precipitation, temperature, wind velocity, ...). Let us denote the set of parameters that parametrises this joint probability distribution by $\boldsymbol{\theta}$. The joint distribution can then be denoted, more explicitly, as $p(B_1, B_2|\boldsymbol{\theta})$,[1] and Bayesian statistics then say that if we wanted to predict the type of transport that a person will use given the current weather $p(B_1|B_2)$, we would first need to marginalise out the parameters to obtain

$$p(B_1, B_2) = \int p(B_1, B_2, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{1.5}$$

which can be factorised as

$$p(B_1, B_2) = \int p(B_1, B_2|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}, \tag{1.6}$$

where $p(\boldsymbol{\theta})$ captures how probable we believe a particular set of parameter values to be. We would then apply Bayes' theorem to the result, $p(B_1|B_2) = p(B_1, B_2)/p(B_2)$. We do not know the precise distribution of our data (because we never have infinite amounts of data to learn that distribution), and, by marginalising out the parameters in (1.6), we compute the *expectation* of the distribution instead.

If we have not observed any data, the distribution $p(\boldsymbol{\theta})$ is called the prior, as it reflects our belief, a priori, of what the parameter values should be. If we do observe a set of data points $\mathbf{X}$, we can compute the probability of seeing those data points for a particular value of $\boldsymbol{\theta}$, $p(\mathbf{X}|\boldsymbol{\theta})$. We can then use Bayes' theorem to compute

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{X})}, \tag{1.7}$$

from which we can compute the predictive probability $p(B_1|B_2)$. This is done by applying Bayes theorem to the joint probability given in (1.6), where we replace

[1] For the remainder of this chapter, we handle the convention that vectors are denoted by lowercase bold letters ($\mathbf{a}, \boldsymbol{\theta}, \ldots$), while matrices are denoted by uppercase bold letters ($\mathbf{A}, \boldsymbol{\Sigma}, \ldots$).

our prior belief about the parameters,[2] $p(\boldsymbol{\theta})$, with the posterior probability of the parameters after seeing the data $p(\boldsymbol{\theta}|\mathbf{X})$:

$$p(B_1, B_2|\mathbf{X}) = \int p(B_1, B_2|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta} \tag{1.8}$$

$$= \int p(B_1, B_2|\boldsymbol{\theta})\, \frac{p(\mathbf{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{X})}\, \mathrm{d}\boldsymbol{\theta} \tag{1.9}$$

In other words, the Bayesian approach forces us to think about our prior assumptions about the problem explicitly, and to encode these in our definition of a prior distribution over possible values of our parameters $p(\boldsymbol{\theta})$. If we know nothing about the problem, we can choose this prior to be uninformative. Whatever the prior we choose, however, the importance of the prior becomes negligible as enough data become available: as the number of data points in $\mathbf{X}$ increases, the importance of the likelihood term $p(\mathbf{X}|\boldsymbol{\theta})$ increases in (1.7) and the relative importance of the prior decreases. This is a satisfying result; when few data are available our model can be mostly directed by our prior assumptions (where our prior distribution itself encodes how confident we are in our assumptions), and large amounts of data will either confirm, correct or override our prior assumptions. Moreover, in any case, our uncertainty about our parameters is correctly reflected in the uncertainty about the prediction. In the limit for infinite amounts of training data, the posterior distribution over the model parameters, $p(\boldsymbol{\theta}|\mathbf{X})$, obtained by the Bayesian approach converges to a single impulse, located at the same value as obtained by traditional maximum a posteriori (MAP) methods and maximum likelihood (ML) methods.[3] Whenever we do not have infinite amounts of training data, however, Bayesian methods take into account all possible parameter values which could have produced those data.

The power of Bayesian methods stems from this simple fact: instead of trying to find the model that fits the training data best, possibly subject to prior regularising constraints that we want to enforce (as is the case for ML and MAP methods), we consider all possible models that could have produced our training data, and weigh those by the probability that they would indeed have produced those data. The resulting model is not prone to overfitting the training data, and even very complex models will only use as much effective expressive power as warranted by the available training data.

The difficulty, of course, is in performing the integration. For most functions of interest, there is no analytical solution to the integral. Numerical approximations are then required. Maximum likelihood and maximum a posteriori methods can be seen as one, very crude, approximation to the Bayesian paradigm, where the distribution over the parameters of the model is approximated by a Dirac impulse at the

[2] That is, the prior probability distribution that we had defined over the parameter values. In Bayesian statistics, probabilities are seen as a measure of our belief in the possible values of a variable.

[3] With infinite amounts of training data, the prior vanishes and both ML and MAP converge to the same solution.

mode of the posterior. In the limit of infinite amounts of training data, the maximum likelihood method converges to the same solution as fully Bayesian treatment of the parameters. However, when the amounts of training data are limited, less crude approximate integration will typically result in better models than maximisation of the likelihood, sometimes vastly so. There is, therefore, a lot of interest in developing techniques to perform good approximate integration. Yet, since maximum likelihood methods (whether with or without regularisation or priors over parameters) are computationally much cheaper, they are used in the vast majority of models used for behaviour modelling to date.

1.2 A Note on Bayesian Networks

Directed graphical models, also called Bayesian networks or belief networks are described in detail in Chap. 2. These networks provide an intuitive representation of the independence assumptions of a model between the different random variables of interest. Bayesian networks are very important in machine learning because they provide an easy and intuitive representation of some of the modelling assumptions and because, once the conditional distributions (or densities) of each variable given its parents has been specified, they provide us with automatic algorithms to perform inference with optimal efficiency. It is important to realise that the models themselves do not specify anything but the factorisation of the joint likelihood, and in particular the (conditional) probability distributions of the variables must be specified separately. These are almost always specified by parametric probability density function (PDF) whose parameters are learnt by maximum likelihood methods, rather than being marginalised out.

Despite their name, there is, therefore, nothing inherently Bayesian about Bayesian networks. The parameters of the distributions specified in the Bayesian networks could be considered nuisance variables, and be integrated out, but in practice they rarely are. Many applications of Bayesian networks simultaneously rely on sufficient amounts of training data, so that ML methods provide acceptable results, while the distributions they represent are sufficiently complex, so that the application of exact Bayesian inference is intractable.

Nevertheless, although ML techniques are typically used to perform inference in Bayesian networks, these rely on the structure of the network being fixed (or, at most, adapted to the length of a sequence). Yet the structure of Bayesian networks must not necessarily be specified a priori: it can be learnt from data, just as the parameters of the distributions can be learnt. Learning the structure of the network cannot, however, be done by purely ML methods since, for any realistic problem and finite amount of training data, the model that fits the data best must be the fully connected model. Moreover, the number of possible distinct networks is super-exponential in the number of variables, so that an exhaustive enumeration of all possible networks is impossible for even moderately-sized problems. This is a typical example of the overfitting inherent to ML methods, and Bayesian methods can provide an elegant solution to this particular problem.

In [3] for example, the network structure of Bayesian networks is learnt for the modelling of conversational gestures from data. Iconic gestures are identified automatically using multiple (non-Bayesian) methods, which leads to the creation of sparse networks to model the data. Sampling techniques are used to make the problem tractable, and it would be very interesting to compare the results of this work to fully Bayesian inference.

1.3 Approximating the Marginal Distribution

Because the exact computation of the marginal distributions is often impossible, there is a lot of interest in approximate integration methods. Various schemes exist to compute integration approximately. In the following section, we briefly introduce the most common schemes. These are general methods which are commonly applied; they are also of particular importance to the Dirichlet and Gaussian processes, described in the next section.

1.3.1 Sampling

Sampling is probably the simplest form of approximation. The idea of sampling is that the expectation of a function $f(x)$ under a probability distribution $p(x)$, $\mathbb{E}_{p(x)}[f(x)]$ can be approximated by point-wise evaluation of the function at points distributed according to $p(x)$:

$$\mathbb{E}_{p(x)}\big[f(x)\big] = \int p(x)\, f(x)\, \mathrm{d}x \tag{1.10}$$

$$\simeq \sum_{x \sim p(x)} f(x), \tag{1.11}$$

where x are independent samples taken from the distribution $p(x)$. Various schemes exist to generate samples with the desired distribution $p(x)$, each with performance characteristics that make them more or less well suited for a particular distribution.

Markov Chain Monte Carlo (MCMC) is a very commonly used technique where a simple proposal distribution (from which we can sample directly) is used repeatedly to generate a chain of samples with a more complex stationary distribution (from which we cannot sample directly). In order to obtain the desired distribution, the distribution of the generated samples is modified by discarding some of the samples, or by associating some variable weight to each sample, based on the difference between the target distribution and the stationary distribution of the proposal chain. The resulting samples are not independent, however. A large number of samples must, therefore, be generated, in order to obtain a handful of samples with the desired distribution that are suitably independent. Yet as the dimensionality of the

samples increases, other sampling methods, such as rejection sampling or importance sampling which generate independent samples, become extremely inefficient and the relative effectiveness of MCMC improves dramatically.

Sampling techniques are very useful for a number of reasons: they are comparatively simple to implement and the resulting approximation converges to the exact solution as the number of samples goes to infinity. More precise approximations are, therefore, simple to obtain by increasing the number of samples. Moreover, with some care, quite efficient and computationally effective sampling schemes can be devised.

The main problem of sampling methods is that it can be hard to evaluate whether the sampling scheme has converged to the desired distribution, or not. More effort can often be spent on ensuring the validity of the result than on the actual implementation itself. Sampling methods are nevertheless very mainstream and are used, for example, in particle filters [2] and Bayesian clustering methods [6].

1.3.2 Variational Approximations

The variational framework provides us with a technique to perform approximations analytically. The idea is that, when it is too hard to compute the function of interest, we can instead define some other function, which we can compute, and make it as close as possible to the function of interest. For probability distributions, we measure the difference between the intractable distribution p and the approximating distributions q in terms of the Kullback–Leibler divergence between the two:

$$KL(q \parallel p) = -\int q(x) \ln \frac{p(x)}{q(x)}. \tag{1.12}$$

For some functions and approximation functions of interest, this is possible without evaluating the original distribution p. If q is unrestricted, the optimal distribution is found for $q(x) = p(x)$, so that $KL(q \parallel p) = 0$, yet by restricting q to some tractable distribution, we obtain a workable approximation. Examples of typical restrictions are to restrict q to be a Gaussian distribution, or to factorise complex joint distributions into independent terms. In this scheme, the quality of the approximation is fixed beforehand by the restrictions imposed on $q(x)$, and care must be taken that the approximation be suitably close. In contrast, approximations using sampling methods can be made arbitrarily more accurate by sampling longer. The advantage of variational approximations is in the computational complexity, which can be made much lower than sampling.

1.4 Non-parametric Methods

One major development of the last decade in the area of probabilistic modelling, has been the arrival of effective sampling techniques for non-parametric methods such

as the Dirichlet process and the Gaussian process. These techniques are called non-parametric, because they do not use a fixed, parametric function to describe the distribution of the data. Instead, they consider an infinitely large number of parametrisations, and define a distribution over each possible parametrisation. Gaussian processes provide a distribution over continuous functions, while Dirichlet processes provide a distribution over discrete functions.

1.4.1 The Dirichlet Process

Dirichlet processes (DP [12]) are an extension of the Dirichlet distribution. The Dirichlet distribution is a continuous distribution over a set of non-negative numbers that sum up to one, and is defined as

$$\text{Dirichlet}(x_i \dots x_D; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{D} x_i^{(\alpha_i - 1)}, \tag{1.13}$$

where $\{x_1 \dots x_D\}$ are the random variables and $\boldsymbol{\alpha} = (\alpha_1 \dots \alpha_D)$ is the vector of parameters of the distribution. The Dirichlet distribution is therefore often used as a distribution over probabilities, where we are certain that one of D mutually exclusive values must be true. For example, the Dirichlet distribution is eminently suitable to parametrise the distribution over the priors of a generative clustering model: in this case, the prior probability that a data point belongs to a cluster sums up to one over all clusters. When clustering data, we do not know the precise values of these priors, however, and so we define a distribution over them. In this example, each cluster occupies its own partition of the data space: for any cluster, a new data point will have some probability of belonging to it, and the data point is certain to belong to exactly one of the clusters.[4]

Informally, the DP is an extension of the above example to the limit of infinitely many clusters. More formally, the DP is a distribution over functions, over a partitioning of the input space. As such, Dirichlet processes are often used for Bayesian non-parametric density estimation, as a prior over mixture components. Each mixture component is active in a subspace of the original input space, and one mixture component is active in every part of the space. The core result of the DP is that if we define a DP prior probability distribution over the partitioning of the space, the posterior distribution over the mixture components is also a DP.

Such DPs are defined by

$$\mathbf{x}_i | \boldsymbol{\theta}_i \sim F(\boldsymbol{\theta}_i), \tag{1.14}$$

$$\boldsymbol{\theta}_i | G \sim G, \tag{1.15}$$

[4] We may nevertheless not be certain which cluster a data point belongs to, so that neither the prior nor the posterior probability that a given data point belongs to a given cluster need to be zero or one.

$$G \sim \mathscr{DP}(G_0, \alpha). \tag{1.16}$$

The observations $\mathbf{x}_i$, which may be multivariate and continuous, are distributed according to a mixture of distributions, where each of the mixture components is of the (parametric) form $F(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denote the parameters of the distribution. We introduce a prior distribution G over the parameters $\boldsymbol{\theta}$, and use the Dirichlet process $\mathscr{DP}(G_0, \alpha)$ to specify a discrete prior over G. Here, G_0 is the base distribution and α is the concentration parameter. The DP is a stochastic process, producing an infinite stream of discrete samples G, which in turn specify the distribution over the parameters $\boldsymbol{\theta}$.

This model can be specified equivalently by introducing a nuisance variable c_i which indicates which mixture component (partition component of the space) the data point belongs to. If we take the limit of infinitely many mixture components, $K \to \infty$, and marginalise out c_i, we obtain an equivalent model. Such a clustering model can be defined as

$$x_i \mid c_i, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \sim F(x_i|\boldsymbol{\theta}_{c_i}), \tag{1.17}$$

$$\boldsymbol{\theta}_i \sim G_0, \tag{1.18}$$

$$c_i \sim \text{Multinomial}(\boldsymbol{\pi}), \tag{1.19}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right). \tag{1.20}$$

This model is depicted graphically in Fig. 1.1. Since we have infinitely many components, in practice we cannot represent the distribution of all the components. Yet we can never have more active mixture component than we have data points, and the number of data points is finite. We can therefore use approximation schemes to the base distribution. A great resource describing different sampling schemes and comparing their properties can be found in [14]. For variational approximations, see [5]. The key advantage of Dirichlet processes is that, although G_0 and each individual F can be continuous distributions, the GP distribution is discrete. As a consequence, there is a non-zero probability that two data points x_i and x_j belong to the same mixture component ($c_i = c_j$). This property of the DP allows us to have fewer mixture components than data points, and we can compute a posterior distribution over the number of mixture components. The strength of the Bayesian framework is highlighted in this model: we allow an infinite number of mixture components, hence making the model flexible and limiting the risks of underfitting. We also provide a prior over the partitioning of the space and over each component's distribution. We then marginalise out the partitioning and each particular component's distribution, thus avoiding the problem of overfitting.

The DP provides us with a beautiful framework for clustering, which can be extended very elegantly to hierarchies. That is, we can create clusters of subclusters of data points, where clusters may share subclusters. This framework of hierarchical DP [16], can also be extended to arbitrary tree structures [1]. In [13], hierarchical Dirichlet processes are used to automatically learn the number of states in a hidden

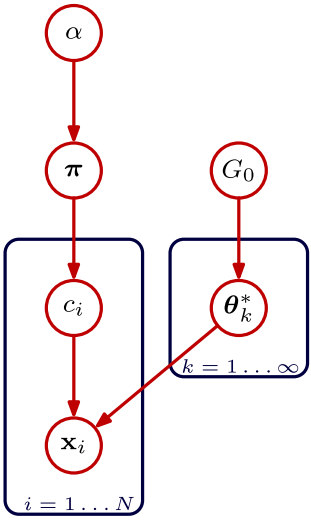


Fig. 1.1 Graphical model of a Dirichlet Process Mixture Model (DPMM)

Markov model (HMM, see Chap. 2 in this volume) and a linear-Gaussian state space model, and these are used to model the unconstrained behaviour of vehicles.

1.4.2 The Gaussian Process

Where the Dirichlet process used a Dirichlet distribution to specify a discrete probability distribution over continuous functions, the Gaussian process (GP) provides a continuous distribution over continuous functions. The concept is slightly arcane, but is very elegant and worth understanding. The "bible" of Gaussian processes is Rasmussen and William's [15], which provides a brilliant overview of the history, derivations, and practical considerations of Gaussian processes. Gaussian processes have a long history, and have been well-known in areas such as meteorology and geostatistics (where GP prediction is known as "kriging") for 40 years. A good overview of the use of Gaussian processes for spatial data can be found in [9].

Gaussian processes have recently been used successfully to extract human pose [10], a problem that is known to be strongly multimodal and hence hard to model [11], and to create complex models of human motion [17]. Even more recently, Gaussian processes have been used successfully for activity recognition in groups of people, where interactions between the participants add to the complexity [7]. The flexibility of GPs comes from the wide range of kernel functions that can be used, which makes it possible to model very complex signals, with multiple levels of superposition. As early as 1969, GPs were used to model the firing patterns of human muscle [8].

The idea of Gaussian processes is to consider a function $f(\mathbf{x})$ as an infinite vector, $\mathbf{f}$, of function values: one for every possible value of $\mathbf{x}$. In this representation, we can define a distribution over the function as a distribution over this vector, rather

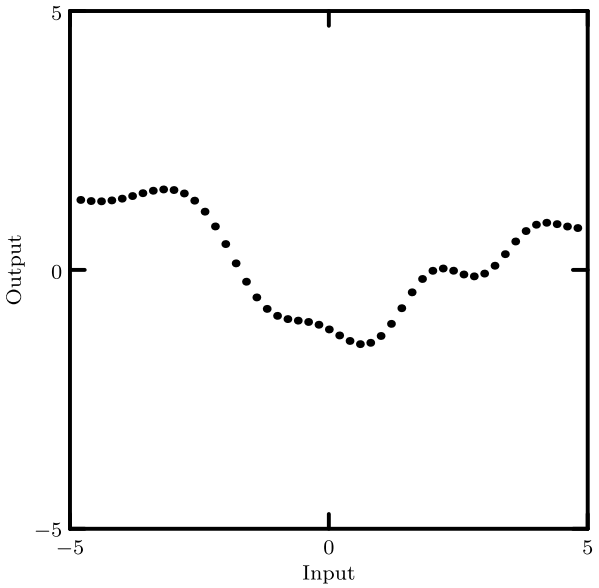


Fig. 1.2 Illustration of a single sample drawn from a Gaussian process. Here, we created a 49-dimensional Gaussian distribution, with a zero vector for the mean, and a 49×49 matrix $\boldsymbol{\Sigma}$, where the elements of the covariance matrix were computed using some function $k(x, x')$—more on this function later. Each dimension of the 49-dimensional Gaussian therefore corresponds to a value x, through the function k. In this case, our inputs x are sampled to be uniformly spaced between -5 and 5. We drew a sample from this Gaussian distribution, and plotted the obtained values (i.e. 49 output points) as a function of the corresponding x

than as a distribution over the parameters of some parametrisation of the function.[5] Now distributions over infinite-dimensional vectors may seem impossible to deal with, but it turns out that, if we assume that this distribution is Gaussian, it is actually possible to manipulate such an object in practice. Moreover, the resulting distribution is closely related to a distribution over parametric functions.

The Gaussian process defines the conditional joint distribution of all elements in our infinite-dimensional vector $\mathbf{f}$ as being Gaussian, with some mean and covariance.[6] The question, then, is of course how we could represent such a distribution. The answer relies on two key observations.

1. We only ever need to evaluate a function for a finite number of function values. Fig. 1.2 illustrates this idea by showing a single sample from some Gaussian process. In this case, a 49-dimensional sample of $f(\mathbf{x})$ is drawn from the process, for a given set of 49 one-dimensional inputs. We could have drawn a much

[5] For the purpose of this explanation, we will consider functions of a single, scalar variable, but the concept easily extends to multiple dimensions.

[6] To be more formally exact, it is defines the distribution over any finite subset of the variables in that vector to be Gaussian.

higher-dimensional sample, or a sample that spans a much larger or smaller input range: the point is that we never need to compute *all* the (infinitely many) numeric values of the function.
2. One of the properties of the multivariate Gaussian distribution is that the marginal distribution of a subset of its dimensions is again a Gaussian distribution, with mean and covariance equal to the relevant elements of the original mean vector and covariance matrix. Therefore, the distribution returned by the Gaussian process over the points where we do evaluate the function, is the same whether or not we take all other points (where we do not evaluate the function) into account. All we need to know, is how our finite set of points covary.

We cannot define a mean vector and covariance matrix over all the points in our function vector, as an unfortunate side effect of its infinite size, but we can provide a functional description of its mean vector and covariance matrix conditionally on the input values $\mathbf{x}$ at hand. These functions lie at the heart of the Gaussian process, and define its properties. The capacity to ignore the points where the function is not evaluated is what makes the Gaussian process tractable and, indeed, very efficient for small data sets. The combination of computational tractability and a formally consistent Bayesian treatment of the model makes Gaussian processes very appealing.

1.4.2.1 Gaussian Process as Bayesian Linear Regression

Consider standard linear regression with Gaussian noise:

$$y = f(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right), \tag{1.21}$$

$$= \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{w} + \varepsilon. \tag{1.22}$$

The target output y is a noisy version of a linear function of the parameters $\mathbf{w}$ and the feature representation $\boldsymbol{\phi}(\mathbf{x})$ of the input data point $\mathbf{x}$. We assume that the output noise is Gaussian, with zero mean and σ^2 variance.

We are given a set of training data points $\mathbf{X}$, which we represent in a design matrix, $\boldsymbol{\Phi}$, and a set of targets $\mathbf{y}$. From this, we want to learn to predict the output $f(\mathbf{x}_*)$ for a given input $\mathbf{x}_*$. We specify the design matrix as

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^\top \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\top \end{bmatrix} \tag{1.23}$$

so that each row contains the features of one data point. In the Bayesian framework, we need to introduce a prior distribution over the parameters $\mathbf{w}$, and choose a zero-mean Gaussian prior, with covariance matrix $\boldsymbol{\Sigma}$: $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Since both the likelihood $p(\mathbf{y}|\boldsymbol{\Phi}, \mathbf{w})$ and the prior $p(\mathbf{w})$ are Gaussian, the posterior $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\Phi})$

is also Gaussian. Moreover, for a given input value $\mathbf{x}_*$, the predictive distribution over the function values $\mathbf{f}(\mathbf{x}_*)$ is given by

$$p\big(f(\mathbf{x}_*)|\mathbf{x}_*, \boldsymbol{\Phi}, \mathbf{y}\big) = \int p\big(f(\mathbf{x}_*)|\mathbf{x}_*, \mathbf{w}\big) p(\mathbf{w}|\boldsymbol{\Phi}, \mathbf{y})\, \mathrm{d}\mathbf{w}. \tag{1.24}$$

It is a standard result that the convolution of two Gaussians is again Gaussian, so that this predictive distribution over function values is Gaussian. The resulting distribution is given by

$$p\big(f(\mathbf{x}_*)|\mathbf{x}_*, \boldsymbol{\Phi}, \mathbf{y}\big) = \mathcal{N}\left(\frac{1}{\sigma^2}\boldsymbol{\phi}(\mathbf{x}_*)^\top \mathbf{A}^{-1}\boldsymbol{\Phi}\mathbf{y}, \boldsymbol{\phi}(\mathbf{x}_*)^\top \mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}_*)\right), \tag{1.25}$$

where $\mathbf{A} = \frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{\Sigma}^{-1}$. Moreover, this can be rewritten as

$$\begin{aligned} &p\big(f(\mathbf{x}_*)|\mathbf{x}_*, \boldsymbol{\Phi}, \mathbf{y}\big) \\ &\quad= \mathcal{N}\big(\boldsymbol{\phi}_*^\top \boldsymbol{\Sigma}\boldsymbol{\phi}\big(\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top + \sigma^2 I\big)^{-1}\mathbf{y}, \boldsymbol{\phi}_*^\top\boldsymbol{\Sigma}\boldsymbol{\phi}_* \\ &\qquad- \boldsymbol{\phi}_*^\top\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top\big(\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top + \sigma^2 I\big)^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\phi}_*\big), \end{aligned} \tag{1.26}$$

where we used $\boldsymbol{\phi}_*$ as shorthand for $\boldsymbol{\phi}(\mathbf{x}_*)$. This last form may look daunting at first sight, but it is advantageous when the number of features is larger than the number of data points and, since $\boldsymbol{\Sigma}$ is positive-definite, we can rewrite the multiplications of the form $\boldsymbol{\phi}(\mathbf{x})^\top\boldsymbol{\Sigma}\boldsymbol{\phi}(\mathbf{x}')$ as $\boldsymbol{\psi}(\mathbf{x})\cdot\boldsymbol{\psi}(\mathbf{x}')$ for some vector of feature functions $\boldsymbol{\psi}(\mathbf{x})$. Notice that $\boldsymbol{\phi}(\mathbf{x})$ only occurs in multiplications of that form in (1.26), so that we can use the *kernel trick* and fully specify our predictive distribution with a kernel function $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})\cdot\boldsymbol{\psi}(\mathbf{x}')$. For every set of feature functions, we can compute the corresponding kernel function. Moreover, for every kernel function there exists a (possibly infinite) expansion in feature functions. This infinite expansion is not a problem in a Bayesian framework, because the (implicit) integration over the parameter values prevent the model from overfitting on the training data.

The Gaussian process is fully specified by its mean function, $\mu(\mathbf{x})$, and its covariance or kernel function $k(\mathbf{x}, \mathbf{x}')$. The functions $f(\mathbf{x})$ are distributed as

$$f(\mathbf{x}) \sim \mathcal{GP}\big(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big) \tag{1.27}$$

where, by definition, $\mu(\mathbf{x}) = \mathbb{E}[f(x)]$ and $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))]$. In practice, the mean function is often taken to be zero, for simplicity as much as for symmetry in the function space. This is not a restriction of the Gaussian process itself, however, and sometimes a non-zero-mean function is indeed specified. Notice that a zero mean does not make the mean of a particular function equal to zero; rather, for every point in the input space the expectation over the value of all functions at that point, is zero. If we are given a set of inputs $\mathbf{X}$, a set of targets $\mathbf{y}$ and a set of test data points $\mathbf{X}_*$, we can directly specify the covariance matrix of our joint distribution over the function values for the training data points and the test

data points as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma^2 I & \mathbf{K}(\mathbf{X},\mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*,\mathbf{X}) & \mathbf{K}(\mathbf{X}_*,\mathbf{X}_*) \end{bmatrix}\right), \tag{1.28}$$

where each element (i, j) of the matrix $\mathbf{K}(\mathbf{X},\mathbf{X}') = k(\mathbf{x}_i, \mathbf{x}'_j)$ for $\mathbf{x}_i$ being data point i in the set $\mathbf{X}$. Using the standard result for the Gaussian conditional distribution, we can compute

$$p(\mathbf{f}_*|\mathbf{X},\mathbf{y},\mathbf{X}_*) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}), \quad \text{where} \tag{1.29}$$

$$\boldsymbol{\mu} = \mathbf{K}(\mathbf{X}_*,\mathbf{X})\big(\mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma^2 I\big)^{-1}\mathbf{y} \quad \text{and} \tag{1.30}$$

$$\boldsymbol{\Sigma} = \mathbf{K}(\mathbf{X}_*,\mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*,\mathbf{X})\big(\mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma^2 I\big)^{-1}\mathbf{K}(\mathbf{X},\mathbf{X}_*). \tag{1.31}$$

Notice how the mean of the predictive distribution is non-zero, and depends on the training data and targets. The covariance consists of a term representing the prior covariance of the test data points, which is diminished by a term that depends on the training and test data points, and on the noise of the targets; it does not depend on the value of the training targets. Also notice how the matrix inversion, which dominates the computational complexity of (1.30) and (1.31), does not depend on the test data: it needs only be computed once. Training a Gaussian process, therefore, consists of: (1) selecting the correct properties of the function of interest (by selecting the right kernel function and optimising its parameters), and (2) computing the said matrix inverse.

1.4.2.2 Kernel Functions

The kernel function fully specifies the (zero-mean) Gaussian process prior. It captures our prior beliefs about the type of function we are looking at; most importantly, we need to specify beforehand how "smooth" we believe the underlying function to be. Figure 1.3 illustrates how different kernel functions result in different styles of functions: the plots on the left hand side show samples from the prior, while the corresponding plots on the right hand side show samples from a GP with the same covariance function, conditional on the three depicted training data points. The kernel functions depicted here are stationary: they depend solely on the *difference* between the two input vectors. They are, therefore, invariant to translations. Many other kernel functions are possible, including non-stationary ones: the only requirement of a function for it to be a valid kernel function is that it should be positive semidefinite. Informally, this means that it must be a function that leads to a positive semidefinite covariance matrix. The easiest way to derive new kernel functions is to modify known kernel functions using operations which are known to preserve the positive-semidefiniteness of the function [4].

Kernel functions often have parameters. These parameters are not affected by the Gaussian process, and are part of its specification. One recurrent parameter is

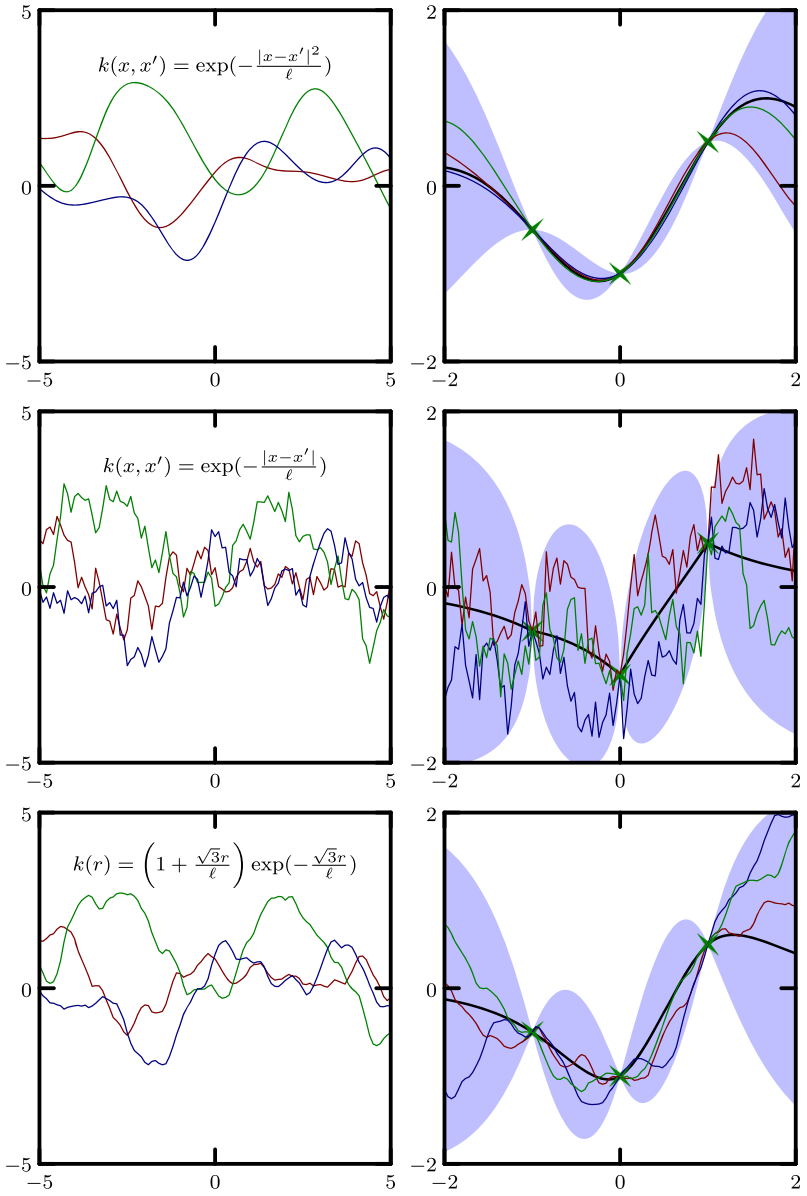


Fig. 1.3 The kernel function specifies the properties of the modelled function. The plots on the *left hand side* show three different samples from the Gaussian process prior (before we observe any data). The corresponding plots on the *right hand side* show samples of the posterior distribution (*thin lines*), given the training data (*green crosses*), as well as the mean function (*thick black line*) and two standard deviations around the mean (*shaded area*). In all plots the length scale ℓ was set to one, and r is defined for notational convenience, as $r = |\mathbf{x} - \mathbf{x}'|$

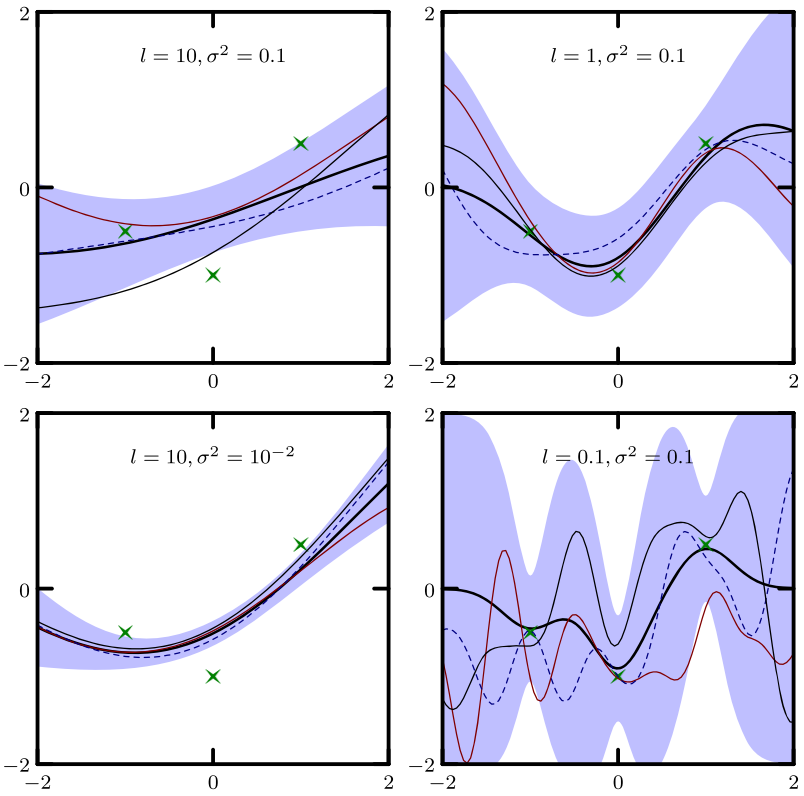


Fig. 1.4 Illustration of Gaussian process regression, with one-dimensional input and one-dimensional output, and three training data points. Changing the length scale of the kernel function affects the overall "smoothness" of the sampled functions, while the variance of the error on the training data points affects how close the functions are to the training data

the length scale ℓ, which basically sets how the distance between data points affect the way the corresponding function outputs covary. The length scale, therefore, affects how smoothly the function varies. Figure 1.4 shows samples from four different Gaussian processes, all with squared exponential kernel functions and different length scales, illustrating how this parameter affects the distribution over the functions. A short length scale increases the variance in areas away from the training data, and consequently also increases the effect of the prior in those areas.

The parameters of the kernel function, also called hyperparameters, are fixed for a particular GP, but can, of course, themselves be learnt from data. In the Bayesian framework, the way to do this is to place a prior distribution over the hyperparameters, and to integrate out the hyperparameters. This integral typically cannot be done analytically, and approximations are then required. When the posterior distribution over the parameters is strongly peaked, one acceptable approximation to the posterior is the Dirac impulse: the integral then becomes the maximum likelihood function. Since the posterior distribution over the hyperparameters is more often

strongly peaked than the posterior over the parameters, the values of the hyperparameters are often found by maximising the marginal likelihood of the training data with respect to the hyperparameters. Such an optimisation is called type II Maximum Likelihood and, although this procedure re-introduces a risk of overfitting, this risk is far lower than with maximum likelihood optimisation of the parameters.

1.4.2.3 Classification

Gaussian processes lend themselves very naturally for regression, but can also be used very effectively for classification. Instead of having an unrestricted function output as in the case of regression, the output of a two-class classifier is constrained to lie in the range $[0 \ldots 1]$, so that it can be interpreted as the probability of one of the classes given the inputs. This is typically done using a "squashing function", such as the logistic sigmoidal:

$$\sigma(a) = \frac{1}{1 + e^{-a}} \tag{1.32}$$

For multi-class problems, this can be extended very naturally to the soft-max function.

Training a GP for classification is more involved than for regression, however. Because the output is pushed through a non-linear function, the posterior distribution is not Gaussian, and cannot be derived in closed form anymore. Approximation schemes are therefore required, and both variational approximation schemes, as well as sampling methods, are routinely used for this.

1.5 Human Behaviour Modelling

Bayesian methods in general, and non-parametric models in particular, present a very promising avenue of research in human activity modelling. As we attempt to recognise and model ever more subtle behaviours, advanced and flexible models become necessary. The systems with which our machines observe the world become ever more sophisticated and ever cheaper, simultaneously leading to an explosion of the amounts of collected data, and of the dimensionality of the data points. Yet manually labelling training data remains just as tedious, expensive and error-prone.

Bayesian inference provides us with extremely powerful techniques with which to address these problems: they do not force us to keep our models artificially simple, yet they can learn from little training data without overfitting. In recent years, we have seen increasing use of Gaussian processes and Dirichlet processes in the field, and it is to be expected that this trend will continue in the foreseeable future.

1.6 Summary

In this chapter, we have introduced the general Bayesian modelling framework, and have exposed its modelling strengths and computational weaknesses. We have exposed the important development of non-parametric methods, and have described two classes of Bayesian non-parametric models, which are commonly used in behavioural modelling: Dirichlet mixture models and Gaussian processes.

1.7 Questions

1. What are Bayesian methods?
2. When using Gaussian processes, the parameters of the kernel functions are adapted to the training data. Why, then, are Gaussian processes called non-parametric methods?
3. Naive Bayes is a probabilistic model that uses Bayes' rule to classify data points. Is it a Bayesian model, though?

1.8 Glossary

- *Bayesian model*: A probabilistic model in which Bayes' rule is applied consistently. Parameters of the model are considered as nuisance variables, and are marginalised out.
- *DP*: Dirichlet Process. A stochastic process over functions in a partitioning of the observation space. Such a process provides a distribution over (discrete) distributions in the space
- *GP*: Gaussian Process. A stochastic process over continuous variables. This provides a distribution over functions of the input space.
- *HMM*: Hidden Markov Model. A stochastic model of sequential data, where each observation is assumed to depend on a corresponding discrete, latent state, and the sequence of latent states forms a Markov chain.

References

1. Adams, R., Ghahramani, Z., Jordan, M.: Tree-structured stick breaking processes for hierarchical modeling. In: Neural Information Processing Systems. MIT Press, Vancouver (2009)
2. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., Sci, D., Organ, T., Adelaide, S.A.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. **50**(2), 174–188 (2002). See also IEEE Transactions on Acoustics, Speech, and Signal Processing
3. Bergmann, K., Kopp, S.: Modeling the production of coverbal iconic gestures by learning bayesian decision networks. Appl. Artif. Intell. **24**(6), 530–551 (2010)
4. Bishop, C.M.: Pattern Recognition and Machine Learning, 1st edn. Springer, Berlin (2007)

5. Blei, D.M., Jordan, M.I.: Variational inference for Dirichlet process mixtures. Bayesian Anal. **1**(1), 121–144 (2006)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
7. Cheng, Z., Qin, L., Huang, Q., Jiang, S., Tian, Q.: Group activity recognition by Gaussian process estimation. In: Proceedings of the International Conference on Pattern Recognition, pp. 3228–3231 (2010)
8. Clamann, H.P.: Statistical analysis of motor unit firing patterns in a human skeletal muscle. Biophys. J. **9**(10), 1233–1251 (1969)
9. Cressie, N.A.C.: Statistics for Spatial Data. Wiley Series in Probability and Statistics. Wiley-Interscience, New York (1993). Rev sub edition
10. Ek, C.H., Torr, P.H., Lawrence, N.D.: Gaussian process latent variable models for human pose estimation. In: Proceedings of the 2007 Conference on Machine Learning for Multimodal Interfaces (2007)
11. Ek, C.H., Rihan, J., Torr, P.H., Lawrence, N.D.: Ambiguity modeling in latent spaces. In: Machine Learning for Multimodal Interaction. Lecture Notes in Computer Science, vol. 5237, pp. 62–73. Springer, Berlin (2008)
12. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**(2), 209–230 (1973)
13. Fox, E.B., Sudderth, E.B., Willsky, A.S.: Hierarchical Dirichlet processes for tracking maneuvering targets. In: 10th Int. Conf. on Information Fusion, pp. 1–8 (2007)
14. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**(2), 249–265 (2000)
15. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
16. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc. **101**(476), 1566–1581 (2006)
17. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 283–298 (2008)

Chapter 2
Introduction to Sequence Analysis for Human Behavior Understanding

Hugues Salamin and Alessandro Vinciarelli

2.1 Introduction

Human sciences recognize sequence analysis as a key aspect of any serious attempt of understanding human behavior [1]. While recognizing that nonsequential analysis can provide important insights, the literature still observes that taking into account sequential aspects "*provide*[*s*] *an additional level of information about whatever behavior we are observing, a level that is not accessible to nonsequential analyses*." [2]. The emphasis on sequential aspects is even higher when it comes to domains related to social interactions like, e.g., Conversation Analysis: "[...] *it is through the knowledge of the place of an action in a sequence that one reaches an understanding of what the action was* (*or turned out to be*)." [4]. Furthermore, social interactions are typically defined as "*sequences of social actions*" in the cognitive psychology literature [21].

In parallel, and independently of human sciences, sequence analysis is an important topic in machine learning and pattern recognition [5, 7]. Probabilistic sequential models, i.e. probability distributions defined over sequences of discrete or continuous stochastic variables, have been shown to be effective in a wide range of problems involving sequential information like, e.g., speech and handwriting recognition [6], bioinformatics [3] and, more recently, Social Signal Processing and social behavior understanding [27].

H. Salamin (✉) · A. Vinciarelli
School of Computing Science, University of Glasgow, Glasgow, Scotland
e-mail: hsalamin@dcs.gla.ac.uk

A. Vinciarelli
e-mail: vincia@dcs.gla.ac.uk

A. Vinciarelli
Idiap Research Institute, Martigny, Switzerland

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_2,

Given a sequence $\mathbf{X} = (x_1, \ldots, x_N)$, where x_t is generally a D-dimensional vector with continuous components, the sequence analysis problem (in machine learning) takes typically two forms: The first is called *classification* and it consists in assigning $\mathbf{X}$ a class c belonging to a predefined set $C = \{c_1, \ldots, c_K\}$. The second is called *labeling* and it corresponds to mapping $\mathbf{X}$ into a sequence $\mathbf{Z} = (z_1, \ldots, z_N)$ of the same length as $\mathbf{X}$, where each z_t belongs to a discrete set $S = \{s_1, \ldots, s_T\}$. An example of classification in Human Behavior Understanding is the recognition of gestures, where a sequence of hand positions is mapped into a particular gesture (e.g., hand waving) [29]. An example of labeling is role recognition in conversations, where a sequence of turns is mapped into a sequence of roles assigned to the speaker of each turn [24].

In both cases, the problem can be thought of as finding the value $\mathbf{Y}^*$ satisfying the equation

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \mathrm{P}(\mathbf{X}, \mathbf{Y}), \tag{2.1}$$

where $\mathbf{Y}^*$ can be one of the classes belonging to C, or a sequence $\mathbf{Z}$ of the same length as $\mathbf{X}$. In this respect, the main problem is to find a model $\mathrm{P}(\mathbf{X}, \mathbf{Y})$ suitable for the problem at hand, i.e. an actual expression of the probability to be used in the equation above. This chapter adopts the unifying framework of *graphical models* [14] to introduce two of the most common probabilistic sequential models used to estimate $\mathrm{P}(\mathbf{X}, \mathbf{Y})$, namely Bayesian Networks (in particular Markov Models and Hidden Markov Models [10, 23]) and Conditional Random Fields [15, 26].

The chapter focuses in particular on two major aspects of the sequence analysis problem: On one hand, the role that conditional independence assumptions have in making the problem tractable, and, on the other hand, the relationship between independence assumptions and the particular factorization that the models mentioned above show. The text provides some details of inference and training as well, including pointers to the relevant literature.

The rest of the chapter is organized as follows: Sect. 2.2 describes the graphical models framework, Sects. 2.3 and 2.4 introduce Bayesian Networks and Conditional Random Fields, respectively, Sect. 2.5 proposes training and inference methods and Sect. 2.6 draws some conclusions.

2.2 Graphical Models

The main problem in estimating $\mathrm{P}(\mathbf{X}, \mathbf{Y})$ is that the state spaces of the random variables $\mathbf{X}$ and $\mathbf{Y}$ increase exponentially with the length of $\mathbf{X}$. The resulting challenge is to find a suitable trade-off between two conflicting needs: to use a compact and tractable representation of $\mathrm{P}(\mathbf{X}, \mathbf{Y})$ on one side and to take into account (possibly long-term) time dependencies on the other side. Probability theory offers two main means to tackle the above, the first is to *factorize* the probability distribution, i.e. to express it as a product of factors that involve only part of the random variables in $\mathbf{X}$ and $\mathbf{Y}$ (e.g., only a subsequence of $\mathbf{X}$). In this way, the global problem is broken into

small, possibly simpler, problems. The second is to make *independence assumptions* about the random variables, i.e. to make hypotheses about which variables actually influence one another in the problem.

As an example of how factorization and independence assumptions can be effective, consider the simple case where $\mathbf{Y}$ is a sequence of binary variables. By applying the chain rule, it is possible to write the following:

$$\mathrm{P}(Y_1, \ldots, Y_N) = \mathrm{P}(Y_1) \prod_{i=2}^{N} \mathrm{P}(Y_i | Y_1, \ldots, Y_{i-1}). \tag{2.2}$$

As the number of possible sequences is 2^N, a probability distribution expressed as a table of experimental frequencies (the percentage of times each sequence is observed) requires $2^N - 1$ parameters.

In this respect, the factorization helps to concentrate on a subset of the variables at a time and maybe to better understand the problem (if there is a good way of selecting the order of the variables), but still it does not help in making the representation more compact, the number of the parameters is the same as before the factorization. In order to decrease the number of parameters, it is necessary to make independence assumptions like, e.g., the following (known as *Markov property*):

$$\mathrm{P}(Y_i | Y_1, \ldots, Y_{i-1}) = \mathrm{P}(Y_i | Y_{i-1}). \tag{2.3}$$

The above transforms (2.2) into

$$\mathrm{P}(Y_1, \ldots, Y_N) = \mathrm{P}(Y_1) \prod_{i=2}^{N} \mathrm{P}(Y_i | Y_{i-1}), \tag{2.4}$$

where the number of parameters is only $2(N-1)+1$, much less than the original $2^N - 1$. The number of parameters can be reduced to just three if we consider that $\mathrm{P}(Y_i | Y_{i-1})$ is independent of i, thus it does not change depending on the particular point of the sequence. The combination of factorization and independence assumptions has thus made it possible to reduce the number of parameters and model long sequences with a compact and tractable representation.

Probabilistic graphical models offer a theoretic framework where factorization and independence assumptions are equivalent. Distributions $\mathrm{P}(\mathbf{X}, \mathbf{Y})$ are represented with graphs where the nodes correspond to the random variables and the missing edges account for the independence assumptions. More in particular, the graph acts as a filter that out of all possible $\mathrm{P}(\mathbf{X}, \mathbf{Y})$ selects only the set *DF* of those that *factorize over the graph* (see below what this means depending on the type of graph). In parallel the graph acts as a filter that selects the set *DI* of those distributions $\mathrm{P}(\mathbf{X}, \mathbf{Y})$ that respect the independence assumptions encoded by the graph (see below how to identify such independence assumptions). The main advantage of graphical models is that $DF = DI$, i.e. factorization and independence assumptions are equivalent (see [5] for an extensive description of this point). Furthermore, inference and training techniques developed for a certain type of graph can be extended to all of the

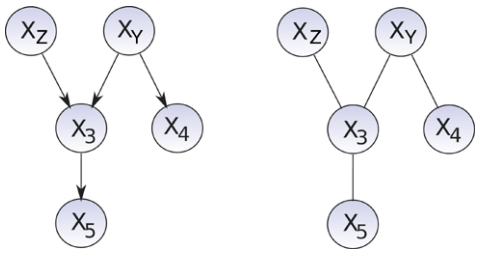


Fig. 2.1 Probabilistic graphical models: each node corresponds to a random variable and the graph represents the joint probability distribution over all of the variables. The edges can be directed (*left graph*) or undirected (*right graph*)

distributions encompassed by the same type of graph (see [13] for an extensive account of training techniques in graphical models).

The rest of this section introduces notions and terminology that will be used throughout the rest of this chapter.

2.2.1 Graph Theory

The basic data structure used in the chapter is the graph.

Definition 2.1 A *graph* is a data structure composed of a set of nodes and a set of edges. Two nodes can be connected by a directed or undirected edge.

We will denote by $G = (\mathbf{N}, \mathbf{E})$ a graph, where $\mathbf{N}$ is the set of nodes and $\mathbf{E}$ is the set of the edges. We write $n_i \rightarrow n_j$ when two nodes are connected by a directed edge and n_i—n_j when they are connected by an undirected one. If there is an edge between n_i and n_j, we say that these are *connected* and we write that $n_i \rightleftharpoons n_j$. An element of $\mathbf{E}$ is denoted with (i, j) meaning that nodes n_i and n_j are connected.

Definition 2.2 If $n \rightleftharpoons m$, then m is said to be a *neighbor* of n (and vice versa). The set of all neighbors of n is called the *neighborhood* and it is denoted by $\mathrm{Nb}(n)$. The set of the *parents* of a node n contains all nodes m such that $m \rightarrow n$. This set is denoted by $\mathrm{Pa}(n)$. Similarly, the set of the *children* of a node n contains all nodes m such that $n \rightarrow m$. This set is denoted by $\mathrm{Ch}(n)$.

Definition 2.3 A *path* is a list of nodes $(p_1, \ldots, p_n)$ such that $p_i \rightarrow p_{i+1}$ or p_i—p_{i+1} holds for all i. A *trail* is a list of nodes $(p_1, \ldots, p_n)$ such that $p_i \rightleftharpoons p_{i+1}$ holds for all i.

The difference between a trail and a path is that a trail can contain $p_i \leftarrow p_{i+1}$ edges. In other words, in a trail it is possible to follow a directed edge in the wrong direction. In undirected graphs, there is no difference between paths and trails.

Definition 2.4 A *cycle* is a path $(p_1, \ldots, p_n)$ such that $p_1 = p_n$. A graph is *acyclic* if there are no cycles in it.

2.2.2 Conditional Independence

Consider two random variables X and Y that can take values in $\text{Val}(X)$ and $\text{Val}(Y)$, respectively.

Definition 2.5 Two random variables X and Y are *independent*, if and only if $\text{P}(Y|X) = \text{P}(X)\ \forall x \in \text{Val}(X), \forall y \in \text{Val}(Y)$. When X and Y are independent, we write that $\text{P} \models (X \perp Y)$.

The definition can be easily extended to sets of variables $\mathbf{X}$ and $\mathbf{Y}$:

Definition 2.6 Two sets of random variables $\mathbf{X}$ and $\mathbf{Y}$ are *independent*, if and only if $\text{P}(Y|X) = \text{P}(X)\ \forall X \in \text{Val}(\mathbf{X}), \forall Y \in \text{Val}(\mathbf{Y})$. When $\mathbf{X}$ and $\mathbf{Y}$ are independent, we write that $\text{P} \models (\mathbf{X} \perp \mathbf{Y})$.

Definition 2.7 Let $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be sets of random variables. We say that $\mathbf{X}$ is *conditionally independent* of $\mathbf{Y}$ given $\mathbf{Z}$ if and only if:

$$\text{P}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \text{P}(\mathbf{X}|\mathbf{Z})\text{P}(\mathbf{Y}|\mathbf{Z})$$

We write that $\text{P} \models (\mathbf{X} \perp \mathbf{Y}|\mathbf{Z})$.

The rest of the chapter shows how the notion of conditional independence is more useful, in practice, than the simple independence. For example, the Markov property (see above) can be seen as a conditional independence assumption where the future X_{t+1} is conditionally independent of the past $(X_1, \ldots, X_{t-1})$ given the present X_t. Such an assumption might not be true in reality (X_t is likely to be dependent on $X_1, \ldots, X_{t-1}$), but it introduces a simplification that makes the simple model of (2.4) tractable.

2.3 Bayesian Networks

Bayesian Networks [11, 12, 20] are probabilistic graphical models encompassed by Directed Acyclic Graphs (DAGs), i.e. those graphs where the edges are directed and no cycles are allowed. The rest of the section shows how a probability distribution factorizes over a DAG and how the structure of the edges encodes conditional independence assumptions. As factorization and independence assumptions are equivalent for graphical models, it is possible to say that all of the distributions that factorize over a DAG respect the conditional independence assumptions that the DAG encodes. Inference and training approaches will not be presented for directed models because each directed graph can be transformed into an equivalent undirected one and related inference and training approaches can be applied. The interested reader can refer to [9, 13] for extensive surveys of these aspects.

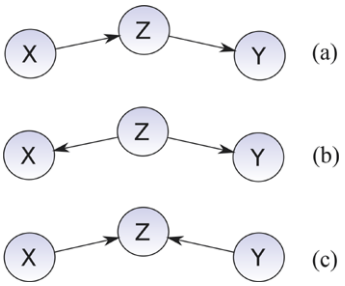


Fig. 2.2 The picture shows the three ways it is possible to pass through a node (Z in this case) along a trail going from X to Y: head-to-tail, tail-to-tail and head-to-head

2.3.1 Factorization

Definition 2.8 Let $\mathbf{X} = (X_i, \ldots, X_N)$ be a set of random variables and G be a DAG whose node set is $\mathbf{X}$. The probability distribution P over $\mathbf{X}$ is said to *factorize* over G if

$$P(\mathbf{X}) = \prod_{i=1}^{n} \mathrm{P}\big(X_i | \mathrm{Pa}(X_i)\big). \tag{2.5}$$

A pair (G, P) where P factorizes over G is called *Bayesian Network*.

2.3.2 The d-Separation Criterion

A DAG allows one to read conditional independence assumptions through the concept of *d-separation* for directed graphs.

Definition 2.9 Let (G, P) be a Bayesian Network and $X_1 \rightleftharpoons \ldots \rightleftharpoons X_N$ a path in G. Let $\mathbf{Z}$ be a subset of variables. The path is blocked by $\mathbf{Z}$ if there is a node W such that either:

- W has converging arrows along the path ($\rightarrow W \leftarrow$) and neither W nor its descendants are in $\mathbf{Z}$
- W does not have converging arrows ($\rightarrow W \rightarrow$ or $\leftarrow W \rightarrow$), and $W \in \mathbf{Z}$

Definition 2.10 The set $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ if every undirected path between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ is blocked by $\mathbf{Z}$

The definition is more clear if we consider the three structures depicted in Fig. 2.2. In the case of Fig. 2.2(a), $\mathbf{Z}$, d-separates X and Y and we can write the following:

$$\mathrm{P}(X, Y, Z) = \mathrm{P}(X)\mathrm{P}(Z|X)\mathrm{P}(Y|Z) = \mathrm{P}(Z)\mathrm{P}(X|Z)\mathrm{P}(Y|Z). \tag{2.6}$$

As $\mathrm{P}(X, Y, Z) = \mathrm{P}(X, Y|Z)\mathrm{P}(Z)$, the above means that $\mathrm{P} \models (X \perp Y|Z)$. The case of Fig. 2.2(b) leads to the same result (the demonstration is left to the reader), while

the structure of Fig. 2.2(c) has a different outcome:

$$\mathrm{P}(X, Y|Z) = \mathrm{P}(X|Z)\mathrm{P}(Y|Z)\mathrm{P}(Z). \tag{2.7}$$

In this case, Z does not d-separate X and Y and it is not true that $\mathrm{P} \models (X \perp Y|Z)$, even if $\mathrm{P} \models (X \perp Y)$. This phenomenon is called *explaining away* and it is the reason of the condition about the nodes with converging arrows in the definition of d-separation. In more general terms, the equivalence between d-separation and conditional independence is stated as follows:

Theorem 2.1 *Let* (G, P) *be a Bayesian Network. Then if* $\mathbf{Z}$ *d-separates* $\mathbf{X}$ *and* $\mathbf{Y}$, $\mathrm{P} \models (\mathbf{X} \perp \mathbf{Y}|\mathbf{Z})$ *holds.*

Thus, the conditional independence assumptions underlying a Bayesian Network can be obtained by simply applying the d-separation criterion to the corresponding directed graph.

2.3.3 Hidden Markov Models

The example presented in Sect. 2.2, known as *Markov Model* (see Fig. 2.3), can be thought of as a Bayesian Network where $\mathrm{Pa}(Y_t) = \{Y_{t-1}\}$:

$$\mathrm{P}(Y_1, \ldots, Y_N) = \mathrm{P}(Y_1) \prod_{i=2}^{N} \mathrm{P}(Y_i|Y_{i-1}) = \prod_{i=1}^{N} \mathrm{P}\big(Y_i|\mathrm{Pa}(Y_i)\big). \tag{2.8}$$

The DAG corresponding to this distribution is a linear chain of random variables.

An important related model is the Hidden Markov Model (HMM) [10, 23], where the variables can be split into two sets, the states $\mathbf{Y}$ and the observations $\mathbf{X}$:

$$\mathrm{P}(\mathbf{X}, \mathbf{Y}) = \mathrm{P}(Y_1)\mathrm{P}(X_1|Y_1) \prod_{t=2}^{N} \mathrm{P}(Y_t|Y_{t-1})\mathrm{P}(X_t|Y_t), \tag{2.9}$$

where the terms $\mathrm{P}(Y_t|Y_{t-1})$ are called *transition probabilities*, the terms $\mathrm{P}(X_t|Y_t)$ are called *emission probability functions*, and the term $\mathrm{P}(Y_1)$ is called *initial state probability*. The underlying assumptions are the Markov Property for the states and, for what concerns the observations, the conditional independence of one observation with respect to all of the others given the state at the same time.

HMMs have been used extensively for both classification and labeling problems. In the first case, one class is assigned to the whole sequence $\mathbf{X}$. For C classes, different sequences of states $\mathbf{Y^i}$ are used to estimate the probability $\mathrm{P}(\mathbf{X}, \mathbf{Y^i})$ and the one leading to the highest value is retained as the winning one:

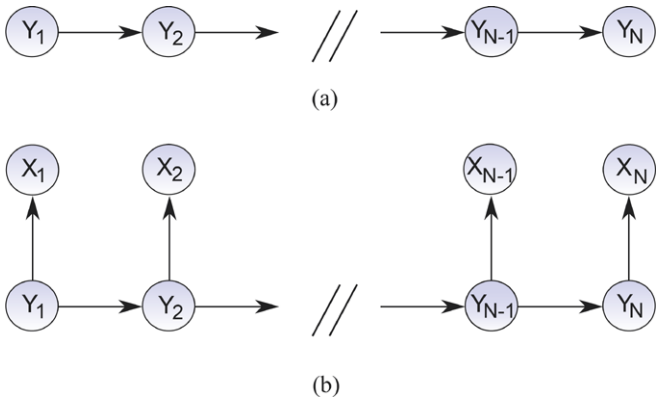


Fig. 2.3 The figure depicts the Bayesian Networks representing a Markov Model (**a**) and a Hidden Markov Model (**b**)

$$k = \arg \max_{i \in [1,C]} = \mathrm{P}\left(\mathbf{X}, \mathbf{Y^i}\right), \tag{2.10}$$

where k is assigned to $\mathbf{X}$ as class. In the labeling case, the sequence of states $\hat{\mathbf{Y}}$ that satisfies

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y} \in \mathscr{Y}} = \mathrm{P}(\mathbf{X}, \mathbf{Y}), \tag{2.11}$$

is used to label the observations of $\mathbf{X}$ ($\mathscr{Y}$ is the set of the state sequences of the same length as $\mathbf{X}$). Each element X_t is labeled with the value y_t of variable $\hat{Y}_t$ in $\hat{\mathbf{Y}}$.

HMM have been widely used for speaker diarization (i.e. the task of segmenting an audio recording in speaker turn). In this scenario, the HMM is used as an unsupervised clustering algorithm. The hidden states $\mathbf{Y}$ of the model correspond to the speakers and the observations are features extracted from the audio spectrum (usually Mel-frequency cepstral coefficients [17]). For a description of a state of the art system using this approach see [8].

HMM suffers from two main limitations. The first is that the observations are assumed to be independent given the states. In the case of human behavior analysis, this assumption does not generally hold. The model presented in the next section, the Conditional Random Field, can address this problem.

The second limitation is that the Markov property makes it difficult to model the duration of the hidden states, i.e. the number of consecutive observations labeled with the same state. The reason is that the probability of transition to a state y_t depends only on y_{t-1}. The Hidden Semi-Markov Model [25] was developed to address this limitation. A complete description of this model is beyond the scope of this chapter, but the key idea is to have the transition probabilities to y_t that depend not only on y_{t-1}, but also on the number of consecutive observations that have been labeled with y_{t-1}.

2.4 Conditional Random Fields

Conditional Random Fields [14, 15, 26] differ from Bayesian Networks mainly in two aspects: The first is that they are encompassed by undirected graphical models, the second is that they are *discriminative*, i.e. they model $P(\mathbf{Y}|\mathbf{X})$ and not $P(\mathbf{X}, \mathbf{Y})$. The former aspect influences the factorization, as well as the way the graph encodes conditional independence assumptions. The latter aspect brings the important advantage that no assumptions about $\mathbf{X}$ need to be made (see below for more details).

2.4.1 Factorization and Conditional Independence

Definition 2.11 Let $G = (\mathbf{N}, \mathbf{E})$ be a graph such that the random variables in $\mathbf{Y}$ correspond to the nodes of G and let P be a joint probability distribution defined over $\mathbf{Y}$. A pair (G, P) is a Markov Random Field if:

$$\mathrm{P}\big(Y|\mathbf{Y} \setminus \{Y\}\big) = \mathrm{P}\big(Y|\mathrm{Nb}(Y)\big) \quad \forall Y \in \mathbf{Y}. \tag{2.12}$$

The factorization of P is given by the following theorem:

Theorem 2.2 *Let (G, P) be a Markov Random Field, then there exists a set of functions $\{\varphi_c | c$ is a clique of $G\}$ such that*

$$\mathrm{P}(\mathbf{Y}) = \frac{1}{Z} \prod_c \varphi_c(\mathbf{Y}|_c), \tag{2.13}$$

where $\mathbf{Y}|_c$ is the subset of $\mathbf{Y}$ that includes only variables associated to the nodes in c, and Z is a normalization constant:

$$Z = \sum_{\mathbf{y}} \prod_c \varphi_c(\mathbf{y}|_c), \tag{2.14}$$

where $\mathbf{y}$ iterates over all possible assignments on $\mathbf{Y}$.

The functions φ_c are often called potentials. They need to be positive functions but they do not necessarily need to be probabilities, i.e. they are not bound to range between 0 and 1. The conditional independence assumptions underlying the factorization above can be inferred by considering the definition of the Markov Network. Each variable is conditionally independent of all of the others given those who correspond to the nodes in its neighborhood: $\mathrm{P} \models (Y \perp \mathbf{Y} \setminus \{Y, \mathrm{Nb}(Y)\} | \mathrm{Nb}(Y))$.

Conditional Random Fields (see Fig. 2.4) are based on Markov Networks and are defined as follows:

Definition 2.12 Let $G = (\mathbf{N}, \mathbf{E})$ be a graph such that the random variables in $\mathbf{Y}$ correspond to the nodes of G. The pair $(\mathbf{X}, \mathbf{Y})$ is a *Conditional Random Field* (CRF)

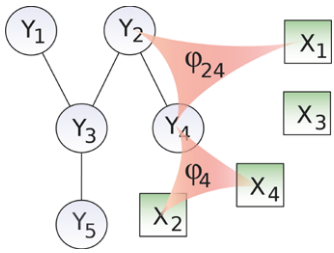


Fig. 2.4 Conditional Random Fields. The potentials are defined over cliques and have as argument the variables corresponding to the nodes of the clique and an arbitrary subset of the observation sequence X

if the random variables in $\mathbf{Y}$ obey the Markov property with respect to the graph G when conditioned on X:

$$\mathrm{P}(Y|\mathbf{X}, \mathbf{Y} \setminus Y) = \mathrm{P}\big(Y|\mathbf{X}, \mathrm{Nb}(Y)\big) \tag{2.15}$$

the variables in $\mathbf{X}$ are called *observations* and those in $\mathbf{Y}$ *labels*.

The definition above does not require any assumption about $\mathbf{X}$ and this is an important advantage. In both labeling and classification problems, $\mathbf{X}$ is a constant and the value of $\mathrm{P}(\mathbf{X}, \mathbf{Y})$ must be maximized with respect to $\mathbf{Y}$:

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \mathrm{P}(\mathbf{Y}|\mathbf{X})\mathrm{P}(\mathbf{X}) = \arg\max_{\mathbf{Y}} \mathrm{P}(\mathbf{Y}|\mathbf{X}). \tag{2.16}$$

Thus, modeling $\mathbf{X}$ explicitly (like it happens, e.g., in Hidden Markov Models) is not really necessary. The model does not require conditional independence assumptions for the observations that might make the models too restrictive for the data and affect the performance negatively. In this respect, modeling $\mathrm{P}(\mathbf{Y}|\mathbf{X})$ makes the model more fit to the actual needs of labeling and classification (see equation above) and limits the need of conditional independence assumptions to the only $\mathbf{Y}$.

The factorization of Conditional Random Fields is as follows:

Theorem 2.3 *Let* (G, P) *be a Markov Network*; *then there exists a set of functions* $\{\varphi_c | c \text{ is a clique of } G\}$ *such that*

$$\mathrm{P}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_c \varphi_c(\mathbf{y}|_c, \mathbf{x}). \tag{2.17}$$

Z *is a normalization constant called the partition function*:

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_c \varphi_c(\mathbf{y}|_c, \mathbf{x}), \tag{2.18}$$

where $\mathbf{y}$ *iterates over all possible assignments on* $\mathbf{Y}$.

The problem left open so far is the definition of the potentials. As this chapter focuses on sequence analysis, the rest of this section will consider the particular case of *Linear Chain Conditional Random Fields*, one of the models most commonly applied for the sequence labeling problem.

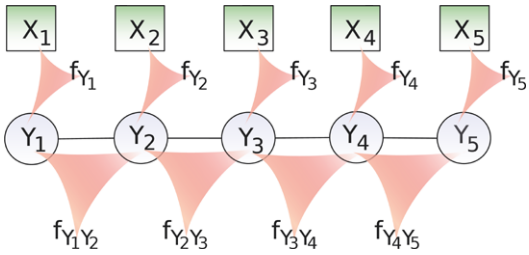


Fig. 2.5 Linear Chain Conditional Random Fields. The cliques in a chain are pair of adjacent labels or individual labels. The potentials are functions of (i) adjacent nodes or (ii) a node and the corresponding observation

2.4.2 Linear Chain Conditional Random Fields

In linear chain CRFs, the cliques are pairs of nodes corresponding to adjacent elements in the sequence of the labels or individual nodes (see Fig. 2.5):

Definition 2.13 A graph is a *chain* if and only if $E = \{(y_i, y_{i+1}), 1 \leq i < |Y|\}$.

Here E is the set of the edges and (y_i, y_{i+1}) represents the edge between the nodes corresponding to elements Y_i and Y_{i+1} in $\mathbf{Y}$.

The following assumptions must be made about the potentials to make the model tractable:

1. The potential over $\{y_t, y_{t+1}\}$ depends only on y_t and y_{t+1}.
2. The potential over $\{y_t\}$ depends only on y_t and x_t.
3. The potentials are the same for all t.
4. The potentials are never zero.

These first three assumptions mean that the marginal distribution for y_t is fully determined by y_{t-1}, y_{t+1} and x_t. The fourth assumption means that every sequence of labels $\mathbf{Y}$ has a probability strictly greater than zero. This last assumption is important in practice, because it allows the product of potentials to be replaced by the exponential of a sum as [14]

$$\mathrm{P}(Y|X) = \frac{\exp(\sum_{t=1}^{N} f_1(y_t, \mathbf{x}_t) + \sum_{t=1}^{N-1} f_2(y_t, y_{t+1}))}{Z(X)},$$

$$Z(X) = \sum_{Y \in \mathcal{Y}^N} \exp\left(\sum_{t=1}^{N} f_1(y_t, \mathbf{x}_t) + \sum_{t=1}^{N-1} f_2(y_t, y_{t+1})\right),$$

where f_1 and f_2 represent potentials having as argument only one label y_t or a pair of adjacent labels $\{y_t, y_{t+1}\}$. Thus, the potentials have been represented as a linear combination of simpler terms called *feature functions*.

In general, the feature functions used for f_1 are

$$f_{y,t}(y_t, \mathbf{x}) = \begin{cases} x_t & \text{if } y_t = y, \\ 0 & \text{otherwise,} \end{cases} \tag{2.19}$$

where x_t is the observation at time t. This family of feature functions can capture linear relations between a label and an observation x_t. For f_2, the feature functions are typically

$$f_{y,y'}(y_t, y_{t+1}) = \begin{cases} 1 & \text{if } y_t = y \text{ and } y_{t+1} = y', \\ 0 & \text{otherwise.} \end{cases} \tag{2.20}$$

In summary, Linear Chain CRFs estimate $p(\mathbf{Y}|\mathbf{X})$ as

$$p(\mathbf{Y}|\mathbf{X}, \alpha) = \frac{1}{Z(\mathbf{X})} \exp\left(\begin{array}{l} \displaystyle\sum_{t=1}^{N} \sum_{y \in \mathscr{Y}} \alpha_y f_{y,t}(y_t, x_t) \\ \quad + \displaystyle\sum_{t=1}^{N-1} \sum_{(y,y') \in \mathscr{Y}^2} \alpha_{y,y'} f_{y,y'}(y_t, y_{t+1}) \end{array} \right). \tag{2.21}$$

The weights α_y of the feature functions of form $f_{y,t}(\mathbf{X}, \mathbf{Y})$ account for how much the value of a given observation is related to a particular label. The weights $\alpha_{y,y'}$ of the feature functions of form $f_{y,y'}(\mathbf{X}, \mathbf{Y})$ account for how frequent it is to find label y followed by label y'.

Linear Chain CRF have been used with success in role recognition [24], where the goal is to map each turn into a role. In this case, the labels correspond to a sequence of roles. The observations are feature vectors accounting for prosody and turn taking patterns associated to each turn.

CRFs have several extensions aimed at addressing the weaknesses of the basic model, in particular the impossibility of labeling sequences as a whole and of modeling latent dynamics. Two effective extensions are obtained by introducing latent variables in the model. The first of these extensions is the hidden Conditional Random Field (HCRF) [22] and it aims at labeling a sequence as a whole. The HCRFs are based on linear chain CRFs, where the chain of labels $\mathbf{Y}$ is latent and a new variable C is added (see Fig. 2.6). The new variable C represents the class of the observations and is connected to every label. All of the potentials are modified to depend on the class C (see Fig. 2.6).

The second extension aims at modeling latent dynamics like, for example, a single gesture (e.g., hand waving) that can have several states (hand moving left and hand moving right) associated with a single label. CRFs cannot model these states and the dynamics associated with them. The Latent Discriminative Conditional Random Fields (LDCRF) [18] were introduced to overcome this drawback. LDCRF introduce a linear chain of latent variables between the observations and the labels (see Fig. 2.7). The labels are disconnected and thus assumed to be conditionally independent given the hidden states. Also, the labels are not directly connected to the observations.

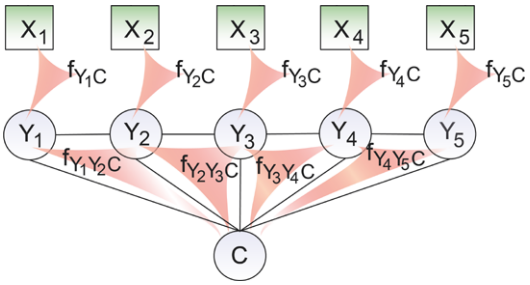


Fig. 2.6 Hidden Conditional Random Fields. The class is represented by the C. The variables Y_i are not observed. The potentials are functions of (i) adjacent nodes and the class ($f_{Y_i Y_{i+1} C}$) or (ii) a node, the corresponding observation, and the class ($f_{Y_i C}$). The potentials $f_{Y_i C}$ are not drawn connected to C to keep the figure readable

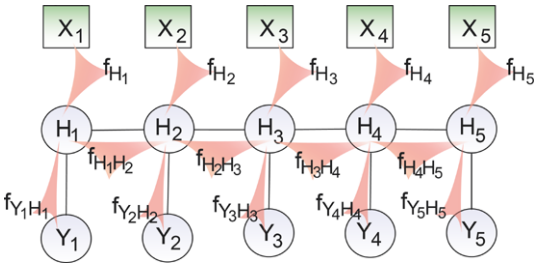


Fig. 2.7 Latent Dynamic Conditional Random Fields. The variables H_i are not observed and capture the latent dynamic. The potentials are functions of (i) adjacent hidden states, (ii) a hidden state and the corresponding label, or (iii) a hidden state and the corresponding observation

2.5 Training and Inference

The models presented so far cannot be used without appropriate *training* and *inference* techniques. The training consists in finding the parameters of a model (e.g., the transition probabilities in a Hidden Markov Model or the α coefficients in a Conditional Random Field) that *better fit* the data of a training set, i.e. a collection of pairs $\mathscr{T} = \{(\mathbf{X^i}, \mathbf{Y^i})\}$ $(i = 1, \ldots, |\mathscr{T}|)$ where each observation is accompanied by a label supposed to be true. By "better fit" we mean the optimization of some criterion like, e.g., the maximization of the likelihood or the maximization of the entropy (see below for more details).

The inference consists in finding the value of $\mathbf{Y}$ that better fits an observation sequence $\mathbf{X}$, whether this means to find the individual value of each Y_j that better matches each $\mathbf{X}$:

$$\mathrm{P}(Y_j = y|\mathbf{X}) = \sum_{\mathbf{Y} \in \{\mathbf{Y}, Y_j = y\}} \mathrm{P}(\mathbf{Y}|\mathbf{X}) \tag{2.22}$$

or finding the sequence $\mathbf{Y}^*$ that globally better matches $\mathbf{X}$:

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \mathrm{P}(\mathbf{Y}|\mathbf{X}). \tag{2.23}$$

The number of possible sequences increases exponentially with $\mathbf{Y}$, thus training and inference cannot be performed by simply estimating $\mathrm{P}(\mathbf{Y}|\mathbf{X})$ for every possible $\mathbf{Y}$. The next two sections introduce some of the key techniques necessary to address both tasks with a reasonable computational load.

2.5.1 Message Passing

One of the main issues in both training and inference is to estimate the probability $\mathrm{P}(Y_j = y)$ that a given label Y_j takes the value y. The *Message Passing* algorithm allows one to perform such a task in an efficient way by exploiting the local structure of the graph around the node corresponding to Y_j (see [30] for an extensive survey of the subject). In particular, the key idea is that the marginal distribution of a node Y_j can be determined if the value of the variables corresponding to its neighboring nodes are known. In practice, those values are unknown, but it is possible to estimate the *belief* that measures the relative probability of the different values. For this reason, the message passing algorithm is sometimes referred to as *belief propagation*.

This section will focus in particular on the message passing algorithm for Pairwise Markov Networks, namely Markov Networks where the cliques include no more than two nodes. While being an important constraint, still it includes cases of major practical importance such as chains, trees and grids (the Linear Chain Conditional Random Fields fall in this class).

The beliefs are defined by

$$b_j(y_j) = \varphi_j(y_j) \prod_{k \in \mathrm{Nb}(Y_j)} m_{kj}(y_j), \tag{2.24}$$

where $\varphi_j(y_j)$ is the potential for node Y_j, m_{kj} is the message from node Y_k to node Y_j (see below for the definition of the messages). Formally, a belief is a function that maps each possible value of Y_j into a real number.

A message is another function that maps the value of one node into a real number and it represents the influence that the sending node has on the receiving one:

$$m_{kj}(y_j) = \sum_{y_k} \left(\varphi_k(y_k) \varphi_{jk}(y_j, y_k) \prod_{n \in \mathrm{Nb}(Y_k) \setminus \{Y_j\}} m_{nk}(y_k) \right) \tag{2.25}$$

where φ_{jk} is the potential of the clique including Y_j and Y_k (this equation motivates the name *sum-product* algorithm that it is used sometimes for this algorithm).

The belief propagation requires the variables to be ordered and this might create problems when a graph contain cycles. When cycles are absent (which is the case

for the models considered in this chapter), the following procedures allow one to find a suitable ordering:

1. Choose a root node
2. Compute messages starting at the leaf, moving to the root
3. Compute messages starting at the root, going to the leafs

It is important to note that the value of the message is independent of the order in which the messages are passed.

At the end of the procedure, each node is associated with a belief that can be used to compute the marginal probabilities as shown by the following:

Theorem 2.4 *Let G be a pairwise random field on $\mathbf{Y}$ and b_j the beliefs computed using the message passing algorithm, then the following holds*:

$$\mathrm{P}(Y_j = y_j) = \frac{b_j(y_j)}{\sum_{y_i} b_j(y_i)}. \tag{2.26}$$

In the case of Conditional Random Fields, the observations in $\mathbf{X}$ have to be taken into account. The message and the beliefs are now dependent on $\mathbf{X}$:

$$b_j(y_j, \mathbf{X}) = \varphi_j(y_j, \mathbf{X}) \prod_{Y_k \in \mathrm{Nb}(Y_j)} m_{kj}(y_j, \mathbf{X}), \tag{2.27}$$

$$m_{kj}(y_j, \mathbf{X}) = \sum_{y_k, \mathbf{X}} \left(\varphi_k(y_k, \mathbf{X}) \varphi_{jk}(y_j, y_k, \mathbf{X}) \prod_{Y_n \in \mathrm{Nb}(Y_k) \setminus \{Y_j\}} m_{nk}(y_k, \mathbf{X}) \right). \tag{2.28}$$

As $\mathbf{X}$ is a constant and as it is known a priori, it is possible to apply exactly the same equations as those used for the Markov Networks.

2.5.1.1 Inference

There are two possible inference scenarios (see beginning of this section): The first consists in finding, for each label, the assignment that maximizes the marginal probability. The second consists in finding the assignment that maximizes the joint probability distribution over the entire label sequence $\mathbf{Y}$.

The first case is a straightforward application of the message passing algorithm. For a given label Y_j, it is sufficient to use the beliefs to find the particular value y^* that maximizes the following probability:

$$y^* = \arg\max_y \mathrm{P}(Y_j = y) = \arg\max_y b_j(y). \tag{2.29}$$

It can be demonstrated that this particular way of assigning the values to the labels minimizes the misclassification rate.

In the second case, the expression of the messages in (2.25) must be modified as follows:

$$m_{kj}(y_j) = \max_{y_k}\left(\varphi_k(y_k)\varphi_{jk}(y_j, y_k)\prod_{n\in \mathrm{Nb}(Y_k)\setminus\{Y_j\}} m_{nk}(y_k)\right), \tag{2.30}$$

where the initial sum has been changed to a maximization. This ensures that the message received by the node corresponding to label Y_j brings information about the sequence $(Y_1, \ldots, Y_{j-1})$ with the highest possible probability rather than about the sum of the probabilities over all possible sequences.

It is again possible to assign to each Y_j, the value y_j^* that maximize the beliefs obtained using the modified messages:

$$y_j^* = \arg\max_y b_j(y). \tag{2.31}$$

It can be shown that the resulting assignment $Y^* = \{y_1^*, \ldots, y_n^*\}$ is the sequence with the maximum probability:

$$Y^* = \arg\max_Y \mathrm{P}(Y). \tag{2.32}$$

2.5.1.2 Training

The last important aspect of probabilistic sequential models is the training. The topic is way too extensive to be covered in detail and the section will focus in particular on Markov Networks as this can be a good starting point toward training Conditional Random Fields. If the assumption is made that the potentials are strictly greater than zero, then Markov Networks can be factorized as

$$\mathrm{P}(\mathbf{Y}|\alpha) = \frac{1}{Z}\exp\left(\sum_c\sum_{i=1}^{n_c}\alpha_c^i f_c^i(\mathbf{Y}|_c)\right), \tag{2.33}$$

$$Z = \sum_{\mathbf{Y}}\exp\left(\sum_c\sum_{i=1}^{n_c}\alpha_c^i f_c^i(\mathbf{Y}|_c)\right), \tag{2.34}$$

where the $f_c^i(Y|_c)$ are feature functions defined over a clique c. The same expression as the same as the one used for Conditional Random Fields, but without the observations $\mathbf{X}$.

Training such a model it means to find the values of the coefficients α that optimize some criteria over a training set. This section considers in particular the maximization of the likelihood:

$$\alpha^* = \arg\max_\alpha \sum_j \log \mathrm{P}\left(\mathbf{Y^j}|\alpha\right), \tag{2.35}$$

where the $\mathbf{Y^j}$ are the sequences of the training set.

The main problem is that solving the above equation leads to an expression for the α coefficients which is not in closed form, thus it is necessary to apply gradient ascent techniques. On the other hand, these are effective because of the following:

Theorem 2.5 *The log-likelihood function is concave with respect to the weights.*

In practice, the limited memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm [16] works well and this has two main motivations: The first is that the algorithm approximates the second derivative and thus converges faster, the second is that it has a low memory usage and it works well on large scale problems. One of the main steps of the LFBGS is the estimation of the derivative of the log-likelihood with respect to α.

$$\frac{\partial}{\partial \alpha_i^c} \sum_j \log \mathrm{P}\left(\mathbf{Y^j}\right) = \frac{\partial}{\partial \alpha_i^c} \sum_j \log\left(\frac{1}{Z} \exp\left(\sum_c \sum_{i=1}^{n_c} \alpha_c^i f_c^i\left(\mathbf{Y^j}|_c\right)\right)\right) \tag{2.36}$$

$$= \frac{\partial}{\partial \alpha_i^c} \sum_j \left(\sum_c \sum_{i=1}^{n_c} \alpha_c^i f_c^i\left(\mathbf{Y^j}|_c\right)\right) - \frac{\partial}{\partial \alpha_i^c} \sum_j \log Z \tag{2.37}$$

$$= \sum_j \left(f_c^i\left(\mathbf{Y^j}|_c\right) - E\left[f_c^i\right]\right). \tag{2.38}$$

The equation above shows that the optimal solution is the one where the theoretical expected value of the feature functions is equal to their empirical expected value. This corresponds to the application of the Maximum Entropy Principle and it further explains the close relationship between Conditional Random Fields and Maximum Entropy Principle introduced in this section.

2.6 Summary

This chapter has introduced the problem of sequence analysis in machine learning. The problem has been formulated in terms of two major issues, namely classification (assigning a label to an entire sequence of observations) and labeling (assigning a label to each observation in a sequence). The chapter has introduced some of the most important statistical models for sequence analysis, Hidden Markov Models and Conditional Random Fields. The unifying framework of Probabilistic Graphical Models has been used in both cases and the emphasis has been on factorization and conditional independence assumptions. Some details of training and inference issues have been provided for Conditional Random Fields, and more generally, for undirected graphical models.

The models introduced in this chapter are not aimed in particular at human behavior understanding, but they have been used successfully in the domain (see [27]

for an extensive survey of the domain). Sequences arise naturally in many behavior analysis problems, especially in the case of social interactions where two or more individuals react to one another and produce sequences of social actions [21].

While trying to provide an extensive description of the sequence analysis problem in machine learning, this chapter cannot be considered exhaustive. However, the chapter, and the references therein, can be considered a good starting point toward a deeper understanding of the problem. In particular, graphical models have been the subject of several tutorials (see, e.g., [19] and Chap. 8 of [5]) and dedicated monographs [14]. The same applies to Hidden Markov Models (see, e.g., [23] for a tutorial and [10] for a monograph) and Conditional Random Fields (see, e.g., [28] for a tutorial and [14] for a monograph).

Last, but not least, so far Human Sciences and Computing Science (in particular machine learning) have looked at the sequence analysis problem in an independent way. As the cross-pollination between the two domains improves, it is likely that models more explicitly aimed at the human behavior understanding problem will emerge.

2.7 Questions

Question 2.1 What is the *rationale* behind (2.1)?

Question 2.2 Consider the graph represented in Fig. 2.2(c). Let X, Y and Z be binary random variables. Let the probability of the Bayesian Network be defined by the following conditional probabilities:

X	$P(X)$
0	0.6
1	0.4

Y	$P(Y)$
0	0.5
1	0.5

X	Y	$P(Z=0\|X,Y)$	$P(Z=1\|X,Y)$
0	0	0.8	0.2
0	1	0.6	0.4
1	0	0.5	0.5
1	1	0.6	0.4

Without using Theorem 2.1, prove the following:

1. $P \models (X \perp Y)$.
2. $P \not\models (X \perp Y | Z)$.

Question 2.3 Consider the Markov Model (MM) and the Hidden Markov Model (HMM) presented in Fig. 2.3. Find a smallest possible set that:

1. d-separates Y_1 from Y_N in the case of MM.
2. d-separates Y_1 from Y_N in the case of HMM.

Prove that there is no subset of the observations $\mathbf{X}$ that d-separates Y_1 from Y_N in the case of HMMs.

Question 2.4 What is the conditional independence assumption made by the Linear Chain Conditional Random Fields?

Question 2.5 Let (G, P) be a Markov Random Field, where G is the undirected graph in Fig. 2.1. By applying (2.25) and (2.24) give the expressions for:

1. $m_{45}(y_5)$.
2. $b_5(y_5)$.
3. $\sum_{y_5} b_5(y_5)$.

Mark that the product in the third case can be rearranged to yield Z as this is a special case of Theorem 2.4.

Question 2.6 Prove Theorem 2.5: The log-likelihood function is concave with respect to the weights. This proof requires some background in analysis and should use materials not presented in this chapter. A proof is given in [14, Chap. 20.3].

2.8 Glossary

- *Probabilistic Sequential Model*: Probability distribution defined over sequences of continuous or discrete random variables.
- *Sequence*: Ordered set of continuous or discrete random variables, typically corresponding to measurements collected at regular steps in time or space.
- *Probabilistic Graphical Model*: Joint probability distribution defined over a set of random variables corresponding to the nodes of a (directed or undirected) graph.
- *Graph*: Data structure composed of a set of nodes and a set of edges, where two nodes can be connected by a directed or undirected edge.
- *Conditional Independence*: Let $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be sets of random variables. We say that $\mathbf{X}$ is *conditionally independent* of $\mathbf{Y}$ given $\mathbf{Z}$ if and only if:

$$\mathrm{P}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \mathrm{P}(\mathbf{X}|\mathbf{Z})\mathrm{P}(\mathbf{Y}|\mathbf{Z}).$$

References

1. Abbott, A.: Sequence analysis: New methods for old ideas. Annu. Rev. Sociol. **21**, 93–113 (1995)
2. Bakeman, R., Gottman, J.M.: Observing Interaction: An Introduction to Sequential Analysis. Cambridge University Press, Cambridge (1986)
3. Baldi, P., Brunak, S.: Bioinformatics: the machine learning approach. MIT Press, Cambridge (2001)
4. Bilmes, J.: The concept of preference in conversation analysis. Lang. Soc. **17**(2), 161–181 (1988)
5. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
6. Camastra, F., Vinciarelli, A.: Machine Learning for Audio, Image and Video Analysis: Theory and Applications. Springer, Berlin (2008)

7. Dietterich, T.: Machine learning for sequential data: A review. In: Caelli, T., Amin, A., Duin, R., de Ridder, D., Kamel, M. (eds.) Structural, Syntactic, and Statistical Pattern Recognition. Lecture Notes in Computer Science, vol. 2396, pp. 227–246. Springer, Berlin (2002)
8. Friedland, G., Vinyals, O., Huang, Y., Muller, C.: Prosodic and other long-term features for speaker diarization. IEEE Trans. Audio Speech Lang. Process. **17**(5), 985–993 (2009)
9. Heckerman, D.: A tutorial on learning with bayesian networks. In: Holmes, D., Jain, L. (eds.) Innovations in Bayesian Networks, pp. 33–82. Springer, Berlin (2008)
10. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge (1997)
11. Jensen, F.V.: An Introduction to Bayesian Networks. UCL Press, London (1996)
12. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs. Springer, Berlin (2007)
13. Jordan, M.I.: Learning in Graphical Models. Kluwer Academic, Dordrecht (1998)
14. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge (2009)
15. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning, pp. 282–289 (2001)
16. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. **45**, 503–528 (1989)
17. Mermelstein, P.: Distance measures for speech recognition, psychological and instrumental. Pattern Recognition and Artificial Intelligence **116** (1976)
18. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press, New York (2007)
19. Murphy, K.: An introduction to graphical models. Technical Report, University of British Columbia (2001)
20. Pearl, J.: Bayesian networks: A model of self-activated memory for evidential reasoning. Computer Science Department, University of California (1985)
21. Poggi, I., D'Errico, F.: Cognitive modelling of human social signals. In: Proceedings of the 2nd International Workshop on Social Signal Processing, pp. 21–26 (2010)
22. Quattoni, A., Wang, S., Morency, L.P., Collins, M., Darrell, T.: Hidden conditional random fields. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1848–1852 (2007)
23. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)
24. Salamin, H., Vinciarelli, A., Truong, K., Mohammadi, G.: Automatic role recognition based on conversational and prosodic behaviour. In: Proceedings of the ACM International Conference on Multimedia, pp. 847–850. ACM, New York (2010)
25. Sansom, J., Thomson, P.: Fitting hidden semi-Markov models to breakpoint rainfall data. J. Appl. Probab. **38**, 142–157 (2001)
26. Sutton, C., McCallum, A.: An introduction to conditional random fields for relational learning. In: Getoor, L., Taskar, B. (eds.) Introduction to Statistical Relational Learning. MIT Press, Cambridge (2007)
27. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009)
28. Wallach, H.M.: Conditional random fields: an introduction. Technical Report MS-CIS-04-21, Department of Computer and Information Science, University of Pennsylvania (2004)
29. Wu, Y., Huang, T.: Vision-based gesture recognition: A review. In: Braffort, A., Gherbi, R., Gibet, S., Teil, D., Richardson, J. (eds.) Gesture-Based Communication in Human-Computer Interaction. Lecture Notes in Computer Science, vol. 1739, pp. 103–115. Springer, Berlin (1999)
30. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Understanding belief propagation and its generalizations. In: Lakemeyer, G., Nebel, B. (eds.) Exploring Artificial Intelligence in the New Millennium, pp. 239–270. Morgan Kaufman, San Mateo (2003)

Chapter 3
Detecting and Tracking Action Content

Alper Yilmaz

3.1 Introduction

Object detection and tracking are very active areas of research in the field of computer vision with significant number of papers being published in major conferences and journals every year. The goal of this chapter is to introduce the reader to main trends in the field rather than providing a list of approaches, to give an insight to underlying ideas, as well as to show their limitations in the hopes of creating interest for conducting research to overcome these shortcomings.

A video sequence containing activities performed by an actor also includes regions that belong to the *background*. At a low feature level, the background does not carry any information related to activities and is generally ignored during analysis, by removing redundancy or by finding spatio-temporal regions containing significant motion content. Either one of these approaches provides us with the ability to extract useful information, which may be directly utilized for representing the activity content or may be tracked for gaining insight into underlying motion. The type of features extracted and the tracking approach used to locate them depend on the choice of appearance and shape representations.

In order to facilitate the discussion on finding and tracking the action content, we will first introduce shape representations commonly employed in context of action analysis. Following this discussion, we will first introduce the appearance descriptors that model observed shape representations. These are followed by detection and tracking of the introduced representations modeled by the descriptors. We will conclude by discussing some open topics.

A. Yilmaz (✉)
The Ohio State University, Columbus, OH 43035, USA
e-mail: yilmaz.15@osu.edu

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_3,

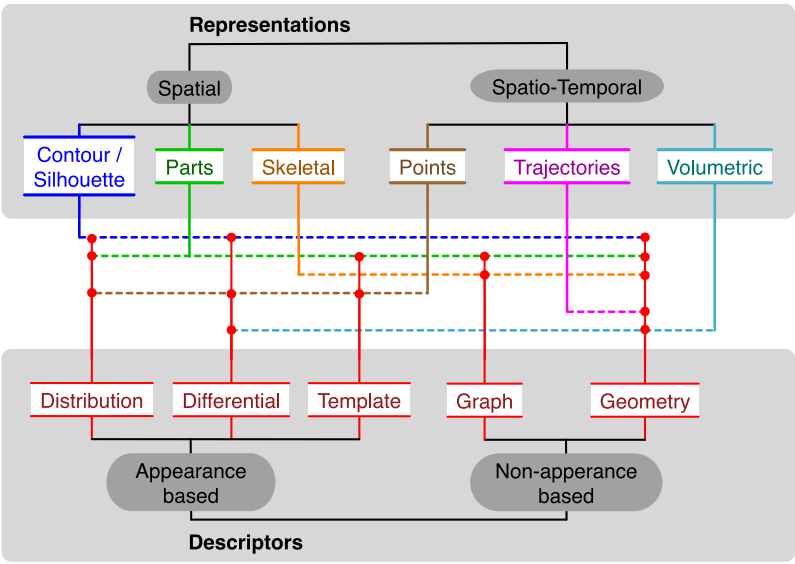


Fig. 3.1 Taxonomy of representations used in action recognition and the list of descriptors traditionally used for each representation. The *colored lines* denote the representation type and each line connecting to it provides the types of descriptors researcher have used for this representation. Counting the connections in the form of a *red bullet* • shows the popularity of the descriptors. For instance geometry based descriptors are adopted by five different representations including contour, parts, skeletal, trajectory and volumetric, but they have not been used for point-based representations

3.2 Representations

There is a strong relation between the activities performed by humans and the representation chosen to define the activity content. Broadly speaking, the types of representations for actions either represent the performers' shape or the motion in the video. As illustrated in Fig. 3.1, depending on the approach used to define the action content, representations are classified into *spatial only* and *spatio-temporal* categories. Either of these classes can be used to represent the action content using a single image or multiples of images.

3.2.1 Spatial Representations

Methods that rely on using a single image, arguably, hypothesize existence of a unique posture that characterizes the action content. The posture of the performer is manifested in spatial shape models, representing the performer's:

- Bounding contour or the silhouette.
- Arrangement of the body parts.
- Skeleton.

Fig. 3.2 Spatial shape (posture) representations. From *left* to *right*: performer template, bounding contour, silhouette, skeletal, parts representations

These representations are illustrated in Fig. 3.2.

Silhouette is usually in the form of a mask with value 1 denoting an object pixel and 0 denoting a non-object pixel [33]. It is also used to define the object contour, which is the bounding curve defining its outline. Contour representation requires a special data structure that explicitly or implicitly defines its shape [61]. A traditional explicit contour structure used since the late 1970s is composed of a set of points (tuples), (x_i, y_i) on the contour along with a set of spline equations, which are real functions fit to the control points generated from the tuples [24]. A well studied choice (from among many others) is the natural cubic spline defined using $n-1$ piecewise cubic polynomials, $S_i(x)$ between n tuples [46]:

$$y = S(x) = \begin{cases} S_0(x) & x \in [x_0, x_1], \\ S_1(x) & x \in [x_1, x_2], \\ \vdots & \\ S_{n-1}(x) & x \in [x_{n-1}, x_n], \end{cases}$$

where the functions are defined as

$$\begin{aligned} S_i(x) = &\frac{z_{i+1}(x - x_i)^3}{6h_i} + \left(\frac{y_{i+1}}{h_i} - \frac{h_i}{6} z_{i+1}\right)(x - x_i) \\ &+ \frac{z_i(x_{i+1} - x)^3}{6h_i} + \left(\frac{y_i}{h_i} - \frac{h_i}{6} z_i\right)(x_{i+1} - x). \end{aligned} \tag{3.1}$$

In (3.1), $h_i = x_{i+1} - x_i$ and without loss of generality, z_i can be estimated for six tuples by solving the following system of equations:

$$\begin{bmatrix} 2(h_0 + h_1) & h_1 & 0 & 0 & 0 \\ h_1 & 2(h_1 + h_2) & h_2 & 0 & 0 \\ 0 & h_2 & 2(h_2 + h_3) & h_3 & 0 \\ 0 & 0 & h_3 & 2(h_3 + h_4) & h_4 \\ 0 & 0 & 0 & h_4 & 2(h_4 + h_5) \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix}$$

$$= 6 \begin{bmatrix} \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\ \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\ \frac{y_4 - y_3}{h_3} - \frac{y_3 - y_2}{h_2} \\ \frac{y_5 - y_4}{h_4} - \frac{y_4 - y_3}{h_3} \\ \frac{y_6 - y_5}{h_5} - \frac{y_5 - y_4}{h_4} \end{bmatrix}.$$

The piecewise analytical contour in this form provides the ability to estimate differential descriptors with ease.

The analytical form, however, does not allow for changes in the contour topology (merging and splitting), which are necessary during contour evolution, as will be discussed later in the chapter. Alternatively, contour in an image can be implicitly defined using the level set formalism [49]. In this formalism, the position of a contour, Γ, is embedded as the zero level set in a two dimensional function $\phi[\mathbf{x}]$ satisfying:

$$\Gamma = \{\mathbf{x} | \phi[\mathbf{x}, t] = 0\}, \tag{3.2}$$

$$\texttt{inside}\ \Gamma = \{\mathbf{x} | \phi[\mathbf{x}, t] > 0\}, \tag{3.3}$$

$$\texttt{outside}\ \Gamma = \{\mathbf{x} | \phi[\mathbf{x}, t] < 0\}, \tag{3.4}$$

where $\mathbf{x} = (x, y)$. In this equation, t denotes iteration number during evolution and will be discussed in more detail later in the chapter. The value at a grid point (x, y) is commonly set to its distance from the closest contour location and is computed by applying a distance transform. For more detail on distance transform, we refer the reader to a comparative survey by Fabbri et al. [13]. Alternatively, researchers have replaced the distance transform with an indicator function to reduce the computational complexity of the representation [34]. In recent years, the level set representation has attracted much attention due to

1. Its ability to adapt to topology changes
2. Direct computation of differential contour features, and
3. Extendibility to higher dimensions with no change in formulation

For example, the contour curvature, which is a commonly used differential contour feature, is computed by

$$\kappa[\mathbf{x}, t] = \nabla \cdot \left(\frac{\nabla \phi[\mathbf{x}, t]}{|\nabla \phi[\mathbf{x}, t]|} \right), \tag{3.5}$$

where ∇ is derivative operator. Depending on the granularity of choice, parts of a performer's body can be grouped into head, torso, arms and legs, which can be defined as geometric primitives, such as rectangles, cylinders and ellipses [3]. The join of the body parts follow a special configuration—head is above torso, arms are connected to shoulders, etc.—which reduces the degrees of freedom for types of motion that can be performed using kinematic motion models. An important issue that needs explicit handling is the occlusion problem that arises when one part is behind the

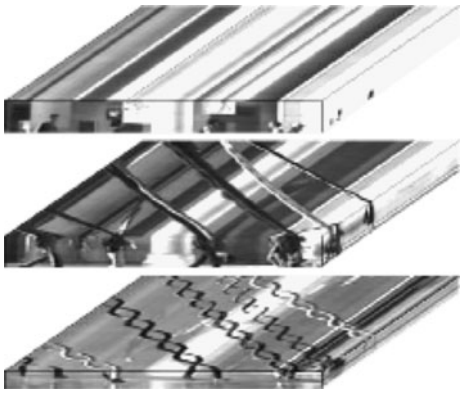


Fig. 3.3 Space-time cube in (x, y, t) domain and its (yt) slices at two different y values revealing different motion content used for analysis of walking action (copyright [1994] IEEE, [39])

others making it invisible. Occluded parts constitute missing observations and can be dealt with by applying heuristics or by enforcing learned part-arrangements. In addition, the degree of articulation increases the complexity of the models.

Another common spatial representation is the set of skeletal models which are commonly used to animate characters and humans in graphics. Skeleton, in our context, is an articulated structure with a set of line/curve segments and joints connecting them [2]. Similar to the contour, object silhouette is required to find the skeleton. A common algorithm used to estimate the skeleton representation is the medial axis transform, which takes the object silhouette and iteratively computes the set of points lying on its topological skeleton such that each point has more than one closest points to the bounding contour. Alternatively, the medial axis is defined as the loci of centers of bi-tangent circles that fit within the object silhouette.

3.2.2 Spatio-Temporal Representations

While the spatial representations are used in analysis of motion content, there are specific representations that are defined in the (x, y, t) space and inherently provide the motion information [63]. These representations use a special data structure called the space-time cube, which is generated by stacking video frames (see Fig. 3.3 for illustration). The space-time cube can be considered a 3D image, where cutting it at any x value results in an image taking (y, t) as its parameters or at any y value gives a (x, t) domain image. Specifically, the (x, t) domain image has been used as early as three decades ago for analysis of cyclic motion. An important observation about the space-time cube is that aside from providing temporal information, it also carries a unique view geometric information when we have two cameras observing a common scene. The geometry is out of the scope of this chapter and we suggest the interested reader to look at additional sources [20].

The spatio-temporal representations can be extracted by local analysis or by looking at the space-time cube globally. Local representations are composed of a set of

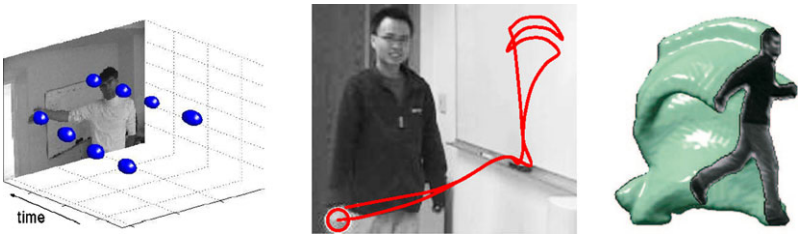


Fig. 3.4 Spatio-temporal representations. *Left* to *right*: spatio-temporal points generated from high frequency hand motion (copyright [26]), trajectory of hand during erasing board, volumetric representation generated from running action (copyright [2005] IEEE, [7])

points $\mathscr{N} = \{(x_i, y_i, t_i)|0 \leq i \leq n\}$ that present characteristic motion and appearance content [26]. In the literature, while there have been a small number of exceptions, the point set $\mathscr{N}$ is commonly treated as a bag of features without temporal order [29]. The bag of features representation, despite the temporal ambiguity, has seen increased adoption over the last few years by many researchers.

Alternative to using temporally sparse point sets, one can consider progression of the point set in time by extracting their trajectories [44]. This representation is referred to as the *trajectory representation* and has been used in many domains within computer vision research including, but not limited to, 3D recovery, video segmentation and action recognition. Trajectory representation is constituted of a temporally ordered point series $\mathscr{T}_i = \{\mathbf{x}_i^0, \mathbf{x}_i^1, \ldots, \mathbf{x}_i^m\}$, which represent the position of the ith point starting from its initial observation at $\mathbf{x}_i^0$ until it disappears from the scene at mth frame [44]. Due to its redundancy, such that it represents the same point multiple times, it can be considered as a superset of the point representation. An intuitive example of a trajectory can be exemplified by the motion of wrist joint during execution of an action. The resulting trajectory draws a curve in the (x, y, t) space, providing instantaneous as well as holistic motion content (see middle image in Fig. 3.4).

The last of the spatio-temporal representations is the *volumetric representation*. In its simplest form, the space-time cube can be considered as a volume providing a holistic representation of the scene. Aside from extracting other representations from it, researchers have used the complete space-time volume to match action under the constraint that the volume only contains motion that relates to the action. This limitation triggered another volumetric representation which considered the subvolume from the space-time cube containing the action content or in other words the performer. The subvolume constitutes a generalized cylinder composed of the performer's posture, such as Γ in (3.2), for each frame from beginning until the completion of the action [7, 54, 60]. The generalized cylinder can be represented by the set of points defining the volume surface $\mathscr{V} = \{[\mathbf{x}, t]|\mathbf{x} = \Gamma^t(s)\}$, or the set of points defining pixels inside it $\mathscr{V} = \{[\mathbf{x}, t]|\mathbf{x} < \Gamma^t(s)\}$ where s denotes the contour arc length (see rightmost image in Fig. 3.4).

3.3 Descriptors

The selection of descriptors is closely related to the choice of representations discussed in the previous section, and it plays a critical role in detection, tracking and action recognition. Generally speaking, a desirable property of a descriptor is its uniqueness, so that they can be easily distinguished from others depending on the task at hand. For instance, template as a descriptor is commonly exploited for the tracking and detection problems, while they have not been extensively used for the action recognition problem. Instead, researchers extract differential features or distributions from the template and use it to represent the action content.

In this section, we base our discussion on the relations between the descriptors and the representations given in Fig. 3.1. As can be noticed from the figure, each descriptor is used for different representations, and they are grouped into two classes based on the use of appearance. This classification will motivate the organization of the following discussion.

3.3.1 Non-appearance Based Descriptors

The descriptors falling under this category do not require additional processing. They are generally readily available once the representation is extracted from the image. The *Euclidean vector* (or vector in short) is a natural example of this category. Euclidean vector is a *geometric descriptor* defined by a starting point, direction and magnitude (or length). In computer vision, a prominent example for a vector is the optical flow vector $\mathbf{u} = (u, v)$ that defines the spatial motion of a point $\mathbf{x}_i^t$:

$$u = x_i^{t+1} - x_i^t, \qquad v = y_i^{t+1} - y_i^t. \tag{3.6}$$

This property makes it a unique descriptor for trajectory based representation, such that $\mathscr{T}_i = \{\mathbf{u}_i^0, \mathbf{u}_i^1, \ldots, \mathbf{u}_i^m\}$.

Geometric descriptors are used to provide information about the shape and size of representations defined in the Euclidean or non-Euclidean coordinates. The latter is generally referred to as the differential geometry.[1] In addition to the Euclidean vector representation of trajectories, another typical geometric descriptor is the relative angles and positions computed between line segments such as the ones in the skeletal representation and representation by parts. The spatial geometric arrangement defined by relative angles and positions can also be extended to the temporal domain by computing the change in the relative angle from one frame to the next. These changes provide insight to how the performer is moving during the course of an action.

The *differential geometry*, on the other hand, studies the shape, size and distances on curved surfaces. Let us consider a performer executing an action. The motion of

[1] In this chapter, differential geometric descriptors are treated different from differential descriptors, as will be discussed under appearance based descriptors.

the performer's body or joints generally lie on a nonlinear manifold. For example, the motion of a wrist joint during hand waving generates a curved trajectory on a one-dimensional manifold defined by the arc-length parameter. Without loss of generality, differential geometric descriptors can be generated in the tangent space by computing the local curvatures. For the trajectory representation, the curvature is computed by

$$\kappa = \frac{x'y'' - y'x''}{(x'^2 + y'^2)^{3/2}}, \tag{3.7}$$

where x' and y' denote first order derivatives and x'' and y'' denote second order derivatives. For the volumetric representations, however, since the degree of freedom is higher, first we define a shape operator S:

$$S = \frac{1}{EG - F^2}\begin{bmatrix}(GL - FM) & (GM - FN)\\ (EM - FL) & (EN - FM)\end{bmatrix}, \tag{3.8}$$

for chosen orthogonal directions (s_p, t_p), unit surface normal $\mathbf{n}$, and

$$E = \mathbf{x}_s \cdot \mathbf{x}_s, \tag{3.9}$$

$$F = \mathbf{x}_s \cdot \mathbf{x}_t, \tag{3.10}$$

$$G = \mathbf{x}_t \cdot \mathbf{x}_t, \tag{3.11}$$

$$L = \mathbf{x}_{ss} \cdot \mathbf{n}, \tag{3.12}$$

$$M = \mathbf{x}_{st} \cdot \mathbf{n}, \tag{3.13}$$

$$N = \mathbf{x}_{tt} \cdot \mathbf{n}, \tag{3.14}$$

where the single and double subscript, respectively, denote the first and second order derivatives. The two algebraic invariants (determinant and trace) of the shape operator define the Gaussian curvature, K, and the mean curvature, H, of the surface:

$$K = \det(S) = \frac{LN - M^2}{EG - F^2}, \tag{3.15}$$

$$H = \frac{1}{2}\text{trace}(S) = \frac{EN + GL + 2FM}{2(EG - F^2)}. \tag{3.16}$$

Alternative to the geometric descriptors, *graphs* provide descriptors that define connectedness of the features for a chosen representation. A graph $G(V, E)$ is an ordered pair composed of vertices V and edges E. The edges $E_i = (V_j, V_k)$ are two element subsets of V and define connectedness between the two elements. The data structure used to define the connectivity between the vertices is the *affinity matrix* which is a symmetric matrix for undirected graphs. The affinity matrix contains 0 and 1 values denoting existence of an edge between vertex pairs. Graph descriptors are explicitly used in the skeletal representations and the representation by parts,

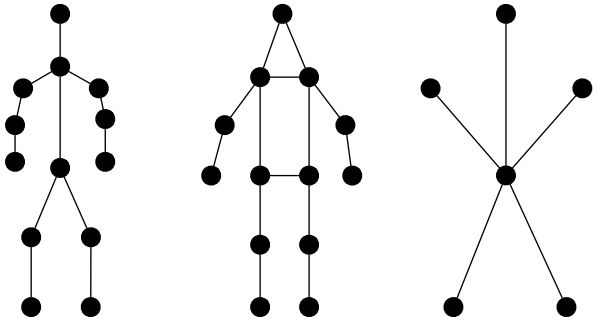


Fig. 3.5 Graph representations used in the order of popularity. From *left* to *right*: tree (most frequently used graph type), cyclic graph, and star graph

where the vertices represent the joints (connections between the parts) or intersections between the curve segments (sub-skeletons). In the case when the representation considers the shoulder as a single joint, the graph representation simplifies to a tree, which does not contain any cycles. Trees as descriptors have been commonly used in detection and tracking of human performers. Other graph types used by researchers are illustrated in Fig. 3.5.

3.3.2 Appearance Based Descriptors

In recent years the field has seen a shift toward appearance based descriptors which model the color observations spatially and/or temporally. These descriptors, broadly speaking, can be deterministic or stochastic. The deterministic descriptors represent the appearance by a feature vector, which is considered as a point in a high dimensional space. In contrast, stochastic descriptors model the appearance by taking domain (x, y) and range (R, G, B) as random variables to generate probability distributions. While an argument on either one of these methods' superiority may be too far-fetched, it is not a mistake to say that deterministic descriptors are more accepted in action analysis. Despite that, stochastic descriptors are widely used for the tracking problem and will be discussed later in the chapter.

Among the three descriptors shown in Fig. 3.1, the most intuitive and commonly adopted descriptor is the *template*. Templates can be 2D (spatial) or 3D (spatio-temporal) depending on their use, and commonly have a shape in the form of a geometric primitive, such as a rectangle, square, ellipse, circle or their 3D versions. A unique property of a template is that it is an ordered list of appearance observations inside the region bounded by the limits of the shape, for example a 7×7 color template or $20 \times 20 \times 20$ derivative template. This property naturally provides the template descriptor with the capability to carry both spatial and appearance information. The spatial order of appearance, however, comes at the cost of being sensitive to changes in the camera viewpoint. Thus, they are more suitable for problems

where the viewing angle of the camera and the performer's action remain constant or change very slowly.

Distribution based descriptors estimate the probability distribution from the observations within a spatial or spatio-temporal region defined by a template, silhouette, or a volume. The observations considered can be raw color values, derivative information, or texture measures. Color distributions are generally estimated non-parametrically by a histogram, or parametrically by mixture models. The histogram, which is a common choice, can be generated by first defining the number of bins (quantization levels) and counting the number of observations that fall into respective bins. While a histogram can be generated using raw color or intensity values, they may need to be processed, such as mean color adjustment, prior to estimating the histogram to remove the illumination and shadowing effects [10]. This adjustment can be achieved by subtracting the mean color computed in the neighborhood of the region of interest: $\mu(R,G,B) = \frac{1}{C}\sum_x \sum_y \sum_t I(x,y,t)$ where C is the volume of the region.

Alternative to using the raw color values, it is also customary to use the gradient for generating a *distribution based descriptor*. Two closely related approaches adopted by many in the field are the scale-invariant feature transform (SIFT) descriptors [31] and the histogram of oriented gradients (HOG) [10, 16]. Both of these approaches compute the gradient of intensity and generate a histogram of gradient orientations weighted by the gradient magnitude. Shared steps between the two approaches can be listed as follows:

Input: Image and regions of interest
Output: Histograms
foreach *Region* **do**
  foreach $(x, y) \in$ *region* **do**
    compute the gradient: $\nabla I(x,y) = (I_x, I_y) =$
    $(I(x-1,y) - I(x+1,y), I(x,y-1) - I(x,y+1))$;
    compute gradient direction: $\alpha(\nabla I(x,y) = \arctan(I_x, I_y)$;
    compute gradient magnitude:
    $|\nabla I(x,y)| = (I_x * I_x + I_y * I_y)^{1/2}$;
    if $|\nabla I(x,y)| \geq \tau$ **then**
      Increment histogram bin for $\alpha(\nabla I(x,y))$;
    end
  end
  smooth histogram;
end

Aside from the common steps outlined above, SIFT approach computes the major orientation from the resulting histogram and subtracts it from computed orientations to achieve rotation invariance. HOG, on the other hand does not perform

orientation normalization. The size of the region and the distance of the performer from the camera play a significant role in achieving discriminative descriptors. Generally speaking, the closer the performer is to the camera, the more discriminative the descriptor will be.

The descriptors above characterize the appearance content of the chosen representation. Another possibility is to generate the histogram defining the shape of the performer. *Shape histograms* model the spatial relation between the pixels lying on the contour. The spatial relation between a reference point (x_r, y_r) on the contour with respect to other points (x_i, y_i) can be modeled by generating a histogram S_r. A collection of such histograms generated by taking all contour points individually as a reference (or a randomly selected subset of points) provides a distribution based descriptor, which can be used to match postures. The spatial relation between two points can be measured by computing the angle and magnitude of the vector joining them which can later be used to generate a 2D histogram taking angle and magnitude as its parameters. Alternatively, shape context uses a concentric circular template centered on a reference contour point, which provide the bins of the histogram in the polar coordinates [4].

3.4 Finding Action Content

The action content manifests itself in the performer's posture and motion; hence analysis of this content requires its extraction on a frame by frame basis or using the space-time volume. This section is dedicated to the extraction of this content and is closely coupled with the representations of the performer and action content discussed in Sect. 3.2. Considering the vast amount of published articles in this topic, the treatment provided here should by no means be considered a complete survey; rather it is only a subset that relates to analysis of actions.

3.4.1 Point Detection

Point detection methods have long been used in computer vision research starting as early as late 1970s [38]. They are, traditionally, used to find interest points in images which have an expressive texture inside templates centered on them. By and large, point detection is still an open area of research due to the problems related to their extraction and association, especially, due to their sensitivity to changes in the camera viewpoint. Here, we will introduce commonly employed spatial and temporal point detection methods starting with the Harris corner detector [19].

The *Harris detector* is one of the earliest and most commonly used point detection methods due to its low computational complexity and ease of implementation. The Harris corner detector, like many others, defines texturedness by conjecturing

that the change in the color content of pixels in the locality of a candidate interest point should be high:

$$E(x, y) = \sum_u \sum_v \big(I(x+u, y+v) - I(x, y)\big)^2. \tag{3.17}$$

The Taylor series approximation of this equation around (x, y) results in

$$E(u, v) = [u\ v] \underbrace{\begin{bmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{bmatrix}}_{\mathtt{M}} \begin{bmatrix} u \\ v \end{bmatrix}. \tag{3.18}$$

This equation contains the commonly termed structure tensor $\mathtt{M}$, which is a second moment computed from the template around the candidate. This matrix defines an ellipse with minor and major axes denoted by its eigenvectors and their extent by respective eigenvalues. The eigenvalues, λ_i of $\mathtt{M}$ are computed from its characteristic equation: $\lambda^2 + \det(\mathtt{M}) - \lambda \cdot \mathrm{trace}(\mathtt{M}) = 0$, which suggests that using determinant and trace of $\mathtt{M}$ should suffice in marking interest points as stated in [19]; hence a traditional texturedness measure is $R(x, y) = \min(\lambda_1, \lambda_2)$ approximated by $R(x, y) \approx \det(\mathtt{M}) - k \cdot \mathrm{trace}(\mathtt{M})^2$ for constant k. The texturedness measure is computed for all pixels and it is subjected to nonmaximal suppression which removes weak interest point candidates and eliminates multiple candidates in small neighborhoods. Harris detector, when applied in scale space, such as by convolving the image with a set of different scaled Gaussian filters, provides feature points at multiple scales. The interest points coexisting at different scales can be combined to provide scale-invariant interest points. Considering that the shape tensor is invariant to rotations, the Harris detector becomes invariant to similarity transform.

The spatial point detection scheme outlined for Harris detector is later extended to spatio-temporal coordinates by introducing the time as an additional dimension to the formulation [26]. This addition resulted in

$$E(u, v, w) = [u\ v\ w] \underbrace{\begin{bmatrix} \sum I_x^2 & \sum I_x I_y & \sum I_x I_t \\ \sum I_x I_y & \sum I_y^2 & \sum I_y I_t \\ \sum I_x I_t & \sum I_y I_t & \sum I_t^2 \end{bmatrix}}_{\tilde{\mathtt{M}}} \begin{bmatrix} u \\ v \\ w \end{bmatrix}, \tag{3.19}$$

where $\tilde{\mathtt{M}}$ defines a sphere and its smallest eigenvalue defines the strength of texturedness as well as *motionness* within the space-time cube for each point: $\tilde{R}(x, y, t) = \mathtt{min}(\lambda_1, \lambda_2, \lambda_3)$. Application of the nonmaximal suppression in the spatio-temporal coordinates results in *space-time interest points* (STIP). STIP features have recently seen increased interest in representing the action content and have been successfully applied to action recognition problem.

Limitations of the Harris detector include: its inability to locate interest points at subpixel level,[2] and the number of interest points it detects. Both of these limitations

[2]Subpixel refers to spatial locations that are not integer.

can be eliminated at the cost of increased computation by using the *scale-invariant feature transform* (SIFT) [31]. Scale invariant feature transform is composed of two major steps for detecting interest points (or the so called keypoints):

1. Scale space peak selection.
2. Key point localization.

Similar to Harris detector applied to multiple scales, SIFT first generates a scale space of minimum four different scales σ_i at different image resolutions, called "octaves". Consecutive scales in each octave are then used to generate difference-of-Gaussian (DoG) images. The difference images provide the domain to select candidate interest points which correspond to the minima and maxima within a $3 \times 3 \times 3$ cubic template in (x, y, σ) coordinates. These candidates are at pixel coordinates and their location can be updated by estimating the locally maximum fit of the DoG in the vicinity of candidate points. The local maxima are computed by expanding the DoG at candidate location $\mathbf{x} = (x, y, \sigma)^\top$ into a Taylor series:

$$\mathrm{DoG}(\mathbf{x}) = \mathrm{DoG}(\mathbf{x}) + \left(\frac{\partial \mathrm{DoG}(\mathbf{x})}{\partial \mathbf{x}}\right)^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top \frac{\partial^2 \mathrm{DoG}(\mathbf{x})}{\partial \mathbf{x}^2}\mathbf{x}. \tag{3.20}$$

The local maximum of the expansion, hence the new position of the candidate, $\tilde{\mathbf{x}}$, can be estimated as

$$\tilde{\mathbf{x}} = -\left(\frac{\partial^2 \mathrm{DoG}(\mathbf{x})}{\partial \mathbf{x}^2}\right)^{-1} \frac{\partial \mathrm{DoG}(\mathbf{x})}{\partial \mathbf{x}}. \tag{3.21}$$

The new extremum value at the new location $\mathrm{DoG}(\tilde{\mathbf{x}})$ is then thresholded to remove insignificant candidates. As a final step interest point candidates along the edges are pruned by fitting a quadratic surface to their localities and by computing the principal curvatures, which are the eigenvalues of the Hessian matrix

$$\mathrm{H} = \begin{bmatrix} \mathrm{DoG}_{xx} & \mathrm{DoG}_{xy} \\ \mathrm{DoG}_{xy} & \mathrm{DoG}_{yy} \end{bmatrix}. \tag{3.22}$$

The eigenvalues are then subjected to texturedness measure by evaluating $R(\tilde{\mathbf{x}}) = \frac{\lambda_1+\lambda_2}{\lambda_1\lambda_2}$. The final two steps omitted in this discussion relate to generating descriptors for detected interest points. SIFT detector generates more number of interest points compared to other interest point detectors. This is due to the fact that the interest points at different scales and different resolutions (pyramid) are accumulated. Empirically, it has been shown in [37] that SIFT outperforms most point detectors, and is more resilient to image deformations. Similar to STIP, SIFT has also been extended to spatio-temporal coordinates to extract space-time interest points [48]. More recently the color based SIFT method is introduced and is widely adopted [56].

3.4.2 Body Parts Detection

The parts representation can appear at configurations that may be very hard to detect due to the nonrigidity of the human performer. The ambiguities result in the loss of depth, and occlusions make this task only harder. Many methods in the literature have been extended to detect these configurations under specific constraints. One of the most commonly explored constraints is the detection of pedestrians, which can only undergo certain visual deformations. These deformations can be learned by machine learning techniques, such as adaptive boosting and support vector machines. Unlike pedestrians, other natural configurations may define a large state space that makes training a classifier harder.

A crucial step in detection of body parts is selecting an appearance model and a spatial arrangement model. Most common appearance feature is the edge information in the form of gradient magnitudes or lines fitted to image edges, which are conjectured to define the boundary of the body parts from the background. Researchers have also used color information [52] or the combination of both features. These features are used to model each body part separately [14]. Alternative to these simple appearance models, researchers have also utilized the shape context descriptor [3] and the HoG descriptor [15] to better define the part's shape. It is also not uncommon to train these appearance features using machine learning techniques to obtain a discriminative parts classifier, such as labeling them as left arm, right feet, head, etc.

Appearance models provide a means to validate the set of hypotheses corresponding to different configurations of the body parts, such as their spatial arrangements. Early work on parts configuration assumes an existing set of stick figures referred to as the key frames, which are used to compare against a proposal configuration [1]. The articulation of parts and resulting occlusions, however, make this comparison quite hard, which requires a priori knowledge about the camera configuration or motion view, such as whether the movement is parallel to the camera view or not. The prior knowledge in addition to the known configurations for certain poses simplifies the detection and labeling of parts. Another treatment of the problem is achieved by multiple camera setup, where the action is viewed by a number of cameras that have overlapping views and known camera parameters (interior and exterior camera parameters). This setup specifically removes the occlusion problem—given adequate number of views—and provides information to detect body parts in 3D. Both in 3D and 2D approaches, an important requirement is to define the state variable for the body parts. A commonly used set of state variables in 2D includes center, scale and orientation (x, y, s, α) [3, 45] of the part which is augmented to 3D by including additional state variables for orientation (x, y, z, α, β) [11]. More details can be found in [17].

There are two schools of thought on estimating the configuration of the body parts from the state variables: graph based and stochastic methods. Graph based methods assign the state variables for each body part to vertices of a graph, $G(V, E)$ which represents n different body parts. In this configuration, the edges between the vertices correspond to physical joints that provide the degrees of freedom

for various configurations [14]. The cost, $\mathscr{E}$, of a configuration can be attributed to costs incurred from the image, $\mathscr{E}_{\texttt{image}}$, and the articulation between the parts $\mathscr{E}_{\texttt{articulation}}$:

$$\mathscr{E}(L) = \mathscr{E}_{\texttt{image}}(L) + \mathscr{E}_{\texttt{articulation}}(L), \tag{3.23}$$

$$= \left(\sum_{1 \le i \le n} m_i(l_i) \right) + \left(\sum_{(v_i, v_j) \in E} d_{ij}(l_i, l_j) \right), \tag{3.24}$$

where $m_i(l_i)$ is the image based cost of placing part v_i at location l_i, and d_{ij} is the configuration cost of between parts v_i and v_j when they are, respectively, placed at locations l_i and l_j. Using this cost, the best configuration C^*, given the image, can be computed by

$$C^* = \arg\min_L \mathscr{E}_{\texttt{image}}(L) + \mathscr{E}_{\texttt{articulation}}(L), \tag{3.25}$$

where L corresponds to the set of locations (x_i, y_i) for each body part v_i. The articulation cost can be defined using various constraints. An intuitive constraint is their relative placement, such as the angle between parts, which requires connectedness between them [2]. More complicated costs can provide increased degrees of freedom. A commonly adopted joint model is spring-like connections [3], whose deformation D can be measured by

$$D_{ij}(l_i, l_j) = \left(\mathbf{T}_{ij}(l_i) - \mathbf{T}_{ji}(l_j)^\top\right) \mathtt{M}_{ij}^{-1} \left(\mathbf{T}_{ij}(l_i) - \mathbf{T}_{ji}(l_j)\right), \tag{3.26}$$

where $\mathtt{M}_{ij}$ is a diagonal weight matrix and $\mathbf{T}_{ij}$ is the transformed location of vertex i with respect to vertex j.

Given the costs associated by various configurations in the form of trees or cyclic graphs [45], the correct configuration can be computed by the Viterbi algorithm or belief propagation in polynomial time. The time for search can be further reduced if the model articulation is reduced by removing joints. For instance, pedestrians at small scales may not have clear views of both feet and hands, hence removing the hands and one of the feet would provide acceptable results with much reduced complexity.

Alternative to graph based models, stochastic approaches can assign uncertainty to possible configurations and pick the one that maps to image properties, e.g. edges or silhouette boundary, the best. Similar to the graph based methods, stochastic methods minimize (maximize) a cost (gain) function. This function is in the form of a likelihood term which measures the probability of observing a certain configuration given the image properties and the distribution of shapes for the action being tested. Some researchers assume that the body-part configuration distribution is uniform, hence can be removed from the maximum likelihood estimation (MLE). Others, on the other hand, model the distribution of body-part configurations using general kernel density estimates from the state variables of all body parts. The distribution provides an *expected* configuration or the most likely configuration, i.e. the mode of the distribution.

In a stochastic approach, researchers usually reduce the search space by assuming that the silhouette of the person is provided. We refer the reader to next section for possible approaches. The silhouette and the image provide both the shape information and the appearance model, which can be used to evaluate the configurations likelihood. Further simplification on possible configurations is usually achieved by detecting head [18] and/or hands [58]. Given configurations, the cost function takes both the shape and the appearance into consideration to measure goodness of mapping between the parts configuration and the image. For defining this measure, Lee et al. use the points on the boundary $\mathbf{b}_i : i = 1, \ldots, C$ and their distances to each of the 14 body parts $s_j : j = 1, \ldots, 14$ by $p(\mathbf{b}_i|s_j) = \exp[-d^2(\mathbf{b}_i, s_j)/\sigma^2]$, where σ defines the uncertainty of mapping [27]. In this formulation, for the parts residing inside the silhouette, no matter how small they are, the equation provides a perfect fit, hence an additional penalty term is required to overcome this problem. In their approach, the authors use the number of pixels that are inside the body configuration but outside the silhouette to penalize wrong configurations. The final likelihood of a parts configuration is then computed as the joint probability of all points on the silhouette boundary.

3.4.3 Detection of Silhouette

Alternative to detecting the body parts and their articulations is the *holistic detection* of the body. This can be achieved either from a single image or by using a sequence of frames, which will be referred to as video in the remainder of the text. In this section, we will discuss video based approaches due to their common adoption by researchers, as well as the industry. In a video acquired by a stationary camera, most regions in images contain almost identical observations due to the constancy of the scene content, except for regions where there is an activity. This observation creates redundancy which can be removed using different techniques, all of which are referred to as *background subtraction*.

An intuitive approach is to detect changes from an image that represents the static scene, empty of moving objects. In the case when the illumination remains constant, the pixels, which are projections of the static scene, will contain the same color. Hence, subtracting the image of the static scene will reveal the moving objects [21]. This simple "background model" can be improved by modeling the variation for each pixel ($I(x, y)$ with a single Gaussian: $I(x, y) \sim N(\mu(x, y), \Sigma(x, y))$ [57]. Given the model for each pixel, the incoming image pixels are labeled as *background* or *object* based on their deviation from the model, computed using Mahalanobis distance:

$$d(x, y) = \big(I(x, y) - \mu(x, y)\big)^{\top} \Sigma(\mathrm{x}, \mathrm{y})^{-1} \big(I(x, y) - \mu(x, y)\big)^2. \tag{3.27}$$

The background model can be further enhanced by introducing more involved statistical modeling, including, but not limited to, mixture of Gaussians per pixel [53], non-parametric kernel density estimates also per pixel [12], or a single domain/range

(x, y, R, G, B) kernel density estimate for the entire image [51]. The order of complexity, while increasing the accuracy, inhibits application for real-time processing. In addition to increasing the complexity of statistical background models, one can include multiple appearance features in the model to achieve invariance to changes in scene illumination. Instances of such features are the texture measures which can be in the form of image gradient [22] or filter responses [30].

The redundant scene information in a video, which corresponds to the static scene, can be conjectured to lie in a subspace that can be estimated using dimensionality reduction techniques. A simple yet straightforward approach is to compute the principal components from the images by cascading their rows into a single vector, representing the point in a multi-dimensional space. This process can be realized by decomposing the covariance matrix $\Sigma = \mathtt{B}^\top \mathtt{B}$ formed from a series of consecutive k images from a video, into its eigenspace, and selecting the most descriptive (maximum eigenvalued) eigenvectors, $\mathtt{U} = [\mathbf{u_1}\mathbf{u_2}\ldots\mathbf{u_k}]$ [40]. This decomposition process and selection of most descriptive principal components provide insensitivity to changes in the illumination. Given the new image, $\mathbf{I}$, represented as a column vector, the moving objects are detected by first projecting the image into the subspace: $\tilde{\mathbf{I}} = \mathtt{U}^\top \mathbf{I}$, and computing its difference from the back-projected image:

$$\mathbf{D} = \mathbf{I} - \mathtt{U}\tilde{\mathbf{I}}.$$

The background subtraction process results in silhouettes, which can be used to reduce the search space in detection of the parts. They also provide a means to detect the skeleton of the moving object using medial axis transform. Background subtraction, however, requires an image sequence for learning the background models, and is not suitable when multiples of performers simultaneously execute actions while partially occluding each other. The partial occlusions generate a single connected silhouette, which requires additional processing to extract individuals. While not widely adopted, detection of the performer from a single image is an alternative and is based on first extracting discriminative features of performers, followed by learning different instances of the features to test hypotheses generated at the image scale. The features are generally in the form of filter responses, such as Gabor filters, which detects oriented gradient information that lies in a vector space. This observation provides a means to apply traditional machine learning techniques for a two-class classification problem which can be stated as labeling templates as "human" or "non-human". The vector space limitations have been recently waived by computing covariance matrices from the observations which lie on a nonlinear Riemannian manifold [55]. Despite these attempts, the articulation of the human body necessitates a comprehensive dataset of postures, which in turn results in overfitting, hence poor detection performance. This limitation has resulted in the requirement to reduce the search space by limiting the detection to only upright walking pedestrians with limited articulation.

Another line of work to detect the object silhouette is by evolving a closed contour to the performers boundary. This evolution is achieved by minimizing a cost

function with the following form:

$$E(\Gamma) = \oint_0^1 E_{\texttt{internal}}(\mathbf{v}) + E_{\texttt{image}}(\mathbf{v}) + E_{\texttt{external}}(\mathbf{v})\, ds, \tag{3.28}$$

where s is the arc-length parameter, vector $\mathbf{v}$ is function of s, $E_{\texttt{internal}}$ includes regularization terms, $E_{\texttt{image}}$ matches the contour to the image, and $E_{\texttt{external}}$ facilitates additional constraints depending on the application. Particularly, internal cost includes first order and second order continuity terms, such as the local contour curvature κ, to guarantee smoothness of the resulting contour. The image based cost takes many different forms that can be *locally* or *globally* computed. Local image features traditionally use image gradient, $\nabla I(\mathbf{x})$, justified based upon the conjecture that object boundaries exhibit high gradient magnitude [8, 24]. For performers on cluttered backgrounds image gradient becomes sensitive to noise and results in poor detection. In contrast, global image features are computed inside and outside the contour, using color and texture [64]. Considering the scale difference between the object and the image size, global images features become biased and inhibit good contour localization. Alternative to using only the gradient or the global features, an image cost that combines the two using λ as the mixing parameter,

$$E_{\texttt{image}} = \lambda E_{\texttt{boundary}} + (1-\lambda) E_{\texttt{region}}, \tag{3.29}$$

provides a good balance between the two [42]. The region term in this equation may be a parametric or a non-parametric distribution of the appearance features [61], such that estimated distributions for inside ($R_{\texttt{inside}}$) and outside ($R_{\texttt{outside}}$) the contour may be used to compute the likelihood ratio:

$$L(\mathbf{x}) = \frac{p(\mathbf{x}|R_{\texttt{inside}})}{p(\mathbf{x}|R_{\texttt{outside}})}.$$

This ratio provides a measure of association for pixel $\mathbf{x}$ in the vicinity of the contour to object's silhouette or the background.

Contour based approaches evolve an initial contour to its final position. Considering the cost is minimized using gradient descent methods, initialization becomes an important step for convergence. In traditional methods, initial contour is typically placed immediately outside the object. Methods computing the cost using region based terms relax this requirement, such that a single contour is initialized either inside or outside the object or can contain both inside and outside. Alternatively, an image sequence can be used to initialize the contour based on observed changes in the image content [41]. For instance, dense optical flow (see Sect. 3.5.2 for more details) computed for each pixel will contain zero motion for static scene components, which can be used to coarsely locate the moving objects. These *coarse* regions, when used to initialize the contour, will result in a *fine* silhouette after contour evolutions.

For contour based methods, one can choose any gradient descent based minimization technique from among many alternatives. This choice is significant,

since it directly relates to the topology of the contour. Following the discussion in Sect. 3.2.1, the contour evolution, hence the minimization process, can be based on changing the positions of tuples (x_i, y_i) in an explicit representation, or on modifying the implicit level set function $\phi[\mathbf{x}, t]$ to define new zero crossings. The changes at iteration t in the level set grid are computed by

$$\phi[\mathbf{x}, t] = \left|\nabla\phi[\mathbf{x}, t-1]\right|\mathbf{n}(\mathbf{x}, t-1)\Gamma'(\mathbf{s}, \mathbf{t}-\mathbf{1}), \tag{3.30}$$

where Γ' refers to the speed of the evolution, which is based on the cost given in (3.28), and $\mathbf{n}$ is the contour normal that can be computed directly from the level set:

$$\mathbf{n}(\mathbf{x}, t-1) = \frac{\nabla\phi[\mathbf{x}, t-1]}{|\nabla\phi[\mathbf{x}, t-1]|}. \tag{3.31}$$

The level set evolution equation, which is shown to evolve iteratively using a numerical scheme, is a partial differential equation and requires high order finite differences for conserving the shape during iterations. Additionally, each evolution requires reinitializing the level set, such that the normal directions and curvature estimations are accurate. These facts suggest long iterations with small evolution at each step, which discourages their application in real-time settings. This shortcoming has heuristically been waived [28] at the cost of satisfying conservation laws, but this method has shown acceptable performance in practice.

3.5 Tracking

In context of action analysis, the goal of tracking is to locate the performer or the region containing the action in every frame of the video. There are two main strategies to tracking. The first strategy requires that the objects are detected such that their associations to object instances in the previous frame provides tracking. The alternative strategy is to define a cost function and iteratively estimate the motion parameters to estimate the object's location in the current frame. In the latter strategy, the motion model may limit the articulations an action may contain. Based on the representations introduced in Sect. 3.2, the following discussion starts with association based trackers, which is followed by different instances of minimization based approaches including:

- Flow estimation based object tracking
- Kernel tracking, and
- Articulated object tracking

For more discussion on tracking, the reader is referred to the comprehensive survey given in [62].

3.5.1 Tracking by Association

In a realistic setting where the camera continuously captures images, it is important to detect the start of an action, hence the performer or the action content at that instant. Since the start of an action is unknown, it is required to perform detection at every frame in the video. This suggests that the detected representations in the current frame can be associated with those detected in the previous frame. This association process generates trajectories, hence highlights the motion content. The association problem, however, is not always trivial due to occlusions, misdetections, entry and exit of performers in the scene. These issues can be addressed by using qualitative motion characteristics or by taking model uncertainties into account. These two different treatments, respectively, result in deterministic and stochastic handling of the association problem.

Deterministic methods define quantitative motion measures to associate the cost of assigning detected objects in consecutive frames. The motion measures are generally based on heuristics, such as

- Bounded speed, $|\mathbf{v}| = (v_x^2 + v_y^2)^{1/2}$.
- Small velocity change, $\alpha(\mathbf{v}) = \texttt{arctan}\, v_y/v_x$.
- Close objects have same velocity, $\mathbf{v}_i = \mathbf{v_j}$ for $d(l_i, l_j) < \tau$ where d, l and τ, respectively, denote distance, location and threshold.

Detected objects at two different frames provide a bipartite graph, where the vertices are the objects divided into two partitions and the edges can only exist between two vertices on two different partitions. Given this graph, the cost of such assignments is then formulated as a combinatorial optimization problem with a solution that provides one-to-one associations from among all possible correspondences. The optimal association to this combinatorial problem can be achieved using the Hungarian search [25]. Alternatively, greedy methods have also been commonly used [43]. The bipartite graph is constructed under a Markovian assumption, where the position of an object depends only on the observation in the previous frame. Missing and reappearing observations, however, pose a limitation to this assumption. At the cost of increasing assignment complexity, multiple frames can be introduced in the assignment process, such that an object at $t-1$ can be directly associated to an object at frame t. This setup generates a connected and directed graph, which is converted to a bipartite graph in [50] by converting each node into two ($+$ and $-$) nodes, where a directed edge is converted to an edge from $+$ node to a $-$ node. The resulting assignment problem can be solved using greedy methods or by Hungarian search.

Stochastic methods, in contrast to deterministic methods, take the uncertainty associated with measurements and motion model into account during estimation of the object's new state. The state of an object may be composed of its location, velocity, acceleration and the parameters of the motion model. For instance, translation-only motion contains two (t_x, t_y) parameters:

$$x^{t+1} = x^t + t_x, \tag{3.32}$$

$$y^{t+1} = y^t + t_y, \tag{3.33}$$

while affine motion contains six parameters (a, b, c, d, t_x, t_y):

$$x^{t+1} = ax^t + by^t + t_x, \tag{3.34}$$

$$y^{t+1} = cx^t + dy^t + t_y. \tag{3.35}$$

The measurements can include the appearance descriptors and the position of detected objects. For simplicity, let us assume we only consider the position as measurement, where the change in the object's position is defined by a sequence of states $\mathbf{x}^t : t = 1, \ldots, T$ computed using:

$$\mathbf{x}^t = f^t\left(\mathbf{x}^{t-1}\right) + \mathbf{w}^t, \tag{3.36}$$

where $\mathbf{w}^t$ is noise. The detected objects in the new frames are attributed to the object's state by a measurement equation $\mathbf{z}^t = h^t(\mathbf{x}^t, \mathbf{n}^t)$ where $\mathbf{n}$ is noise independent of $\mathbf{w}$. Given all the observations until $t-1$, the stochastic method estimates the posterior probability in the form of $p(\mathbf{x}^t|\mathbf{z}^1, \ldots, \mathbf{z}^t)$. The optimal solution to this problem is given by a recursive Bayesian filter, which first *predicts* the prior probability density function $p(\mathbf{x}^t|\mathbf{z}^1, \ldots, \mathbf{z}^{t-1})$, then *corrects* the observation using likelihood function $p(\mathbf{z}^t|\mathbf{x}^t)$ and estimates the posterior $p(\mathbf{x}^t|\mathbf{z}^1, \ldots, \mathbf{z}^t)$.

In the case when h and f are linear functions, and object state along with the noise $\mathbf{w}$ and $\mathbf{n}$ are Gaussian distributed, the state estimate can be computed by the Kalman filter or its different flavors, such as the extended Kalman filter. This linear model suits most pedestrian tracking settings, however, it is not suitable for tracking the body parts, where the state is not Gaussian distributed. For such cases, the estimation can be performed using Monte Carlo sampling, such as the Particle filters. For detailed the discussion on Bayesian filtering, we refer the reader to [36] and [35].

3.5.2 Flow Estimation

Flow based methods assume constancy of brightness, $I(x, y, t) = I(x + dx, y + dy, t + dt)$ in consecutive frames. This equation can be extended to Taylor series and shown to result in the optical flow equation:

$$uI_x + vI_y + I_t = 0, \tag{3.37}$$

where $u = dx/dt$, $v = dy/dt$ and subscripts denote first order derivatives. This single equation contains two unknowns and requires additional constraints and/or equations to construct an equation system. Let us assume a point-based representation with a template around it to provide a descriptor. Conjecturing that the pixels in the template move with the same optical flow, we can estimate the unknown optical flow geometrically [47] or algebraically [32]. Geometric estimation considers (3.37) as a line equation $v = au + b$ with $a = -I_x/I_y$ and $b = -I_t/I_y$. The common motion of pixels within the template results in such lines intersecting at a single position,

which is the solution to (u, v). In practice, instead of all lines intersecting at a single solution, we can expect a cluster of intersections, and the cluster center can be chosen as the solution.

Geometric approach can only be used to estimate the translational motion model. The algebraic approach, on the other hand, directly uses the equations formed from (3.37), hence the optical flow can be assumed to follow a parametric motion model other than translation. In the case of affine motion:

$$u = ax + by + t_x, \tag{3.38}$$

$$v = cx + dy + t_y. \tag{3.39}$$

Substituting these into (3.37) and assuming common motion for $(x_i, y_i) : i \leq n$ within the template will result in

$$\begin{bmatrix} x_1 I_x & y_1 I_x & I_x & x_1 I_y & y_1 I_y & I_y \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n I_x & y_n I_x & I_x & x_n I_y & y_n I_y & I_y \end{bmatrix} \begin{bmatrix} a \\ b \\ t_x \\ c \\ d \\ t_y \end{bmatrix} = \begin{bmatrix} I^t(\mathbf{x}_1) - I^{t-1}(\mathbf{x}_1) \\ \vdots \\ I^t(\mathbf{x}_n) - I^{t-1}(\mathbf{x}_n) \end{bmatrix}, \tag{3.40}$$

which can be solved using least squares adjustment. In the case of translation-only motion model, this equation simplifies to

$$\begin{bmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{bmatrix} \begin{bmatrix} t_x \\ t_y \end{bmatrix} = \begin{bmatrix} \sum I_x (I^t(\mathbf{x}_1) - I^{t-1}(\mathbf{x}_1)) \\ \sum I_y (I^t(\mathbf{x}_n) - I^{t-1}(\mathbf{x}_n)) \end{bmatrix}. \tag{3.41}$$

The left side of this equation follows the same format with the structure tensor in Harris corner detector and suggests that motion can only be computed at locations where there is significant texture, such that eigenvalues λ_i of M are significantly large.

The common treatment stated above takes a different form when the representations are more complicated. Particularly, for contour representation, all the pixels on the contour may follow considerably different motions compared to their neighborhoods, especially when the shape changes in consecutive frames. As discussed in Sect. 3.4.3, aside from the smoothness terms, the cost contains terms related to the image, as well as external constraints. Introducing (3.37) into the cost results in an evolving contour by estimating the optical flow of tuples or the level set grid [5]. The optical flow computed for contour evolution, however, is approached different from the above discussion, due to the constraint that the contour is evolved in its normal direction. This evolution constraint requires that the temporal derivative either introduces an additional cost function [5]:

$$\frac{\partial I(\mathbf{x})}{\partial t} = I_t(\mathbf{x}) \left| \nabla I(\mathbf{x}) \right|, \tag{3.42}$$

or is computed by brute force searching for a minimizer in the neighborhood N of the contour point [34]:

$$\frac{\partial I(\mathbf{x})}{\partial t} = \min_{I_t} I_t(\mathbf{x}_i) : \mathbf{x}_i \in N. \tag{3.43}$$

The former of these equations can be treated as an external cost and it can be introduced to the contour tracking equation by

$$\phi[\mathbf{x}, t] = I_t(\mathbf{x})\big|\nabla I(\mathbf{x})\big|\big|\nabla\phi[\mathbf{x}, t-1]\big|\mathbf{n}(\mathbf{x}, t-1)\Gamma'(s, t-1), \tag{3.44}$$

which in turn minimizes both costs. The evolution function given here is comparable to the cost function in (3.28) used for contour based detection and contains an additional term related to the optical flow constraint. This final cost, however, takes a much simpler form compared to the brute force search based cost due to the fact that it is only evaluated at the contour points as opposed to all the locations on the level set grid.

The most important advantage of a tracking object contour is its flexibility to handle a large variety of articulations, which is important for analyzing the action content. Due to this observation, they have been predominantly used in holistic analysis schemes such as generation of volumetric action representations.

3.5.3 Kernel Tracking

Object descriptors in the form of geometric primitive regions centered around a point define weighted spatial kernels $K(\mathbf{x})$ that may have a uniform weight or varying weight at different pixels within the kernel. The kernel function represents a convolution of the geometric primitive with the template around point $\mathbf{x}$. The motion of the kernel from one frame to the next follows a parametric model including translation, conformal transformation and affine transformation. This formulation very much resembles the flow estimation for a template discussed in the previous section, with a significant difference that the minimized cost is not based on the optical flow constraint, hence does not require brightness constancy.

A typical and commonly used approach is to represent the object's appearance by a model histogram q computed inside the kernel. Given the previous location of the object the new location is computed by minimizing a distance, such as the Bhattacharya distance $d(p) = -\log(\sum_{u=1}^{B}(p(u)q(u))^{1/2})$, between the model histogram and the candidate histogram p for B bins. The candidate histogram is estimated by initializing the kernel state $\mathbf{m}^t$ at object's previous state $\mathbf{m}^{t-1}$. Minimizing Bhattacharya distance or alternatively maximizing the Bhattacharya coefficient $\rho(p) = \sum_{u=1}^{B}(p(u)q(u))^{1/2}$ and expanding it to Taylor series suggests that the new location can be iteratively computed by estimating the likelihood ratio between the

model and candidate histograms [9]:

$$m_k = m_{k-1} + \sum_{i=1}^{n} K'(x_i - m_{k-1}) \frac{q(I^t(\mathbf{x}_i)}{p(I^t(\mathbf{x}_i)}, \tag{3.45}$$

where k is the iteration variable and K' is the derivative of the kernel function, which can be a 2D Gaussian kernel. The kernel state is traditionally defined by the centroid of the object in spatial coordinates, such that the estimation process results in the new object position. Alternatively, the state can be defined by the scale s, orientation α and position $\mathbf{x}$ of the object [59].

Kernel tracking, which contains a stable appearance model defined by q, can handle small changes in the object appearance, but does not explicitly handle occlusions or continuous changes in appearance. These changes can be included in the model by introducing noise and transient models noise process and can be estimated using an online Expectation-Maximization algorithm [23]. These additions, while increasing the complexity of the model, provide the ability to handle small occlusions and continuous appearance changes.

Despite explicit modeling of noise and transient features, kernel trackers perform poorly, or even lose tracking, in cases when the performer suddenly turns around during an action and reveals a completely different appearance, which has not been learned (such as estimation of q) before. A potential approach to overcome this limitation is to learn different views of the object offline and later use them during tracking. A particular example is to learn principal components of different object views using principal component analysis and perform tracking in the subspace by estimating the motion parameters between image at time t and its reprojected image generated from the subspace [6]. The estimation is achieved by first estimating the subspace projection parameters and using them to compute the motion parameters iteratively.

3.6 Summary

This chapter introduces techniques to detect and track important regions that define an action content. We first start the discussion with possible object representation and descriptors that defines the objects performing the actions. Later, we provide extended discussion on detection of these representations and descriptors using video. Finally, possible approaches to track the action content is presented to give necessary tools to the reader of the chapter to perform higher level action analysis introduced in the following chapters.

The approach we have taken to introduce these topics is chosen to relate to the current state of the art and is by no means an attempt to provide a detailed comparative survey.

3.7 Questions

1. Discuss problems related to skeletal representations.
2. What is the effect of camera motion on spatio-temporal representations?
3. The selection of template size plays a significant effect on their detection. Discuss the effect of increasing the template size on extraction process.
4. Show that the Harris corner detector is invariant under rotation.
5. Perform the steps to derive (3.37) using brightness constancy constraint.
6. Perform the steps to derive (3.45).
7. What is the difference between SIFT and HoG descriptors?

3.8 Glossary

- *Object representations* provide ways to present the shape and motion of the performer.
- *Interest points* have an expressive texture inside templates centered on them.
- *Silhouette extraction* provides a mask indicating the object and non-object pixels.
- *Background subtraction* is used to find moving regions in a static scene.
- *Tracking by association* provides one-to-one mapping of detected objects in the current and previous frames.
- *Contour tracking* finds the enclosing object boundary by evolving an initial contour until the cost of fitting is minimized.
- *Kernel tracking* is achieved by iteratively updating position of the region defined using a primitive geometric shape.

References

1. Akita, K.: Image sequence analysis of real world human motion. Pattern Recognit. **17**(1), 73–83 (1984)
2. Ali, A., Aggarwal, J.K.: Segmentation and recognition of continuous human activity. In: IEEE Workshop on Detection and Recognition of Events in Video, pp. 28–35 (2001)
3. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: IEEE Conf. on Computer Vision and Pattern Recognition (2009)
4. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002)
5. Bertalmio, M., Sapiro, G., Randall, G.: Morphing active contours. IEEE Trans. Pattern Anal. Mach. Intell. **22**(7), 733–737 (2000)
6. Black, M., Jepson, A.: Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. Int. J. Comput. Vis. **26**(1), 63–84 (1998)
7. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: IEEE Int. Conf. on Computer Vision (2005)
8. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. In: IEEE Int. Conf. on Computer Vision, pp. 694–699 (1995)
9. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**, 564–575 (2003)

10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
11. Ek, C., Torr, P., Lawrence, N.: Gaussian process latent variable models for human pose estimation. In: Int. Conf on Machine Learning for Multimodal Interaction (2007)
12. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: European Conf. on Computer Vision, pp. 751–767 (2000)
13. Fabbri, R., Costa, L., Torelli, J., Bruno, O.: 2D Euclidean distance transforms: a comparative survey. ACM Computing Surveys **40**(1) (2008)
14. Felzenszwalb, P.F., Huttenlocher, D.: Pictorial structures for object recognition. Int. J. Comput. Vis. **61**(1), 55–79 (2005)
15. Felzenszwalb, P.F., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Conf. on Computer Vision and Pattern Recognition (2008)
16. Freeman, W.T., Roth, M.: Orientation histograms for hand gesture recognition. In: IEEE Intl. Workshop on Automatic Face and Gesture Recognition, pp. 296–301 (1995)
17. Gavrila, D.M.: The visual analysis of human movement: A survey. Comput. Vis. Image Underst. **73**(1), 82–98 (1999)
18. Haritaoglu, I., Harwood, D., Davis, L.S.: W4: real-time surveillance of people and their activities. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 809–830 (2000)
19. Harris, C.G., Stephens, M.: A combined corner and edge detector. In: 4th Alvey Vision Conference, pp. 147–151 (1988)
20. Hartley, R., Zisserman, A.: Multiple View Geometry. Cambridge University Press, Cambridge (2000)
21. Jain, R., Nagel, H.H.: On the analysis of accumulative difference pictures from image sequences of real world scenes. IEEE Trans. Pattern Anal. Mach. Intell. **1**(2), 206–214 (1979)
22. Javed, O., Shafique, K., Shah, M.: A hierarchical approach to robust background subtraction using color and gradient information. In: IEEE Workshop on Motion and Video Computing (2002)
23. Jepson, A.D., Fleet, D.J., ElMaraghi, T.F.: Robust online appearance models for visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1296–1311 (2003)
24. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. Int. J. Comput. Vis. **1**, 321–332 (1988)
25. Kuhn, H.W.: The Hungarian method for solving the assignment problem. Nav. Res. Logist. Q. **2**, 83–97 (1955)
26. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**(2), 107–123 (2005)
27. Lee, M.W., Cohen, I., Jung, S.K.: Particle filter with analytical inference for human body tracking. In: IEEE Workshop on Motion and Video Computing (2002)
28. Li, C., Xu, C., Gui, C., Fox, M.D.: Level set evolution without reinitialization: A new variational formulation. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 430–436 (2005)
29. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
30. Liyuan, L., Maylor, L.: Integrating intensity and texture differences for robust change detection. IEEE Trans. Image Process. **11**(2), 105–112 (2002)
31. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
32. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intelligence, pp. 121–130 (1981)
33. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and Viterbi path searching. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
34. Mansouri, A.R.: Region tracking via level set PDEs without motion computation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 947–961 (2002)
35. Maskell, S., Gordon, N.: A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking. In: IEEE Target Tracking: Algorithms and Applications, vol. 2, pp. 1–15 (2001)

36. Maybeck, P.: Stochastic Models, Estimation, and Control. Mathematics in Science and Engineering, vol. 141. Elsevier, Amsterdam (1979)
37. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1615–1630 (2003)
38. Moravec, H.P.: Visual mapping by a robot rover. In: Proc. of IJCAI, pp. 598–600 (1979)
39. Niyogi, S., Adelson, E.: Analyzing gait with spatiotemporal surfaces. In: Wrks. on Nonrigid and Artic. Motion (1994)
40. Oliver, N.M., Rosario, B., Pentland, A.: A Bayesian computer vision system for modeling human interactions. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 831–843 (2000)
41. Paragios, N., Deriche, R.: Geodesic active contours and level sets for the detection and tracking of moving objects. IEEE Trans. Pattern Anal. Mach. Intell. **22**(3), 266–280 (2000)
42. Paragios, N., Deriche, R.: Geodesic active regions and level set methods for supervised texture segmentation. Int. J. Comput. Vis. **46**(3), 223–247 (2002)
43. Rangarajan, K., Shah, M.: Establishing motion correspondence. Comput. Vis. Graph. Image Process. **54**(1), 56–73 (1991)
44. Rao, C., Yilmaz, A., Shah, M.: View invariant representation and recognition of actions. Int. J. Comput. Vis. **50**(2), 203–226 (2002)
45. Ren, X., Berg, A.C., Malik, J.: Recovering human body configurations using pairwise constraints between parts. In: IEEE Int. Conf. on Computer Vision (2005)
46. Riesenfeld, R.F.: Geometric Modeling with Splines: An Introduction. CRC Press, Boca Raton (2001)
47. Schunk, B.G.: The image flow constraint equation. Comput. Vis. Graph. Image Process. **35**, 20–46 (1986)
48. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional SIFT descriptor and its application to action recognition. In: ACM Multimedia (2007)
49. Sethian, J.A.: Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics Computer Vision and Material Sciences. Cambridge University Press, Cambridge (1999)
50. Shafique, K., Shah, M.: A non-iterative greedy algorithm for multi-frame point correspondence. IEEE Trans. Pattern Anal. Mach. Intell. **27**(1), 51–65 (2005)
51. Sheikh, Y., Shah, M.: Bayesian modeling of dynamic scenes for object detection. IEEE Trans. Pattern Anal. Mach. Intell. **27**(11), 1778–1792 (2005)
52. Sigal, L., Black, M.: Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In: IEEE Conf. on Computer Vision and Pattern Recognition (2006)
53. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real time tracking. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 747–767 (2000)
54. Syeda-Mahmood, T., Vasilescu, A., Sethi, S.: Recognizing action events from multiple viewpoints. In: IEEE Workshop on Detection and Recognition of Events in Video (2001)
55. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on Riemannian manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
56. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1582–1596 (2010)
57. Wren, C.R., Azarbayejani, A., Pentland, A.: Pfinder: Real-time tracking of the human body. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 780–785 (1997)
58. Yang, M., Kriegman, D., Ahuja, N.: Detecting faces in images: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **24**(1), 34–58 (2002)
59. Yilmaz, A.: Kernel based object tracking using asymmetric kernels with adaptive scale and orientation selection. Mach. Vis. Appl. J. (2010)
60. Yilmaz, A., Shah, M.: A differential geometric approach to representing the human actions. Comput. Vis. Image Underst. **109**(3), 335–351 (2008)
61. Yilmaz, A., Li, X., Shah, M.: Contour based object tracking with occlusion handling in video acquired using mobile cameras. IEEE Trans. Pattern Anal. Mach. Intell. **26**(11), 1531–1536 (2004)

62. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. **38**(4), 13 (2006)
63. Zelnik-Manor, L.Z., Irani, M.: Event-based analysis of video. In: IEEE Conf. on Computer Vision and Pattern Recognition (2001)
64. Zhu, S.C., Yuille, A.: Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **18**(9), 884–900 (1996)

Chapter 4
Computational Visual Attention

Simone Frintrop

4.1 What Is Attention? And Do We Need Attentive Machines?

Attention is one of the key mechanisms of human perception that enables us to act efficiently in a complex world. Imagine you visit Cologne for the first time, you stroll through the streets and look around curiously. You look at the large Cologne Cathedral and at some street performers. After a while, you remember that you have to catch your train back home soon and you start actively to look for signs to the station. You have no eye for the street performers any more. But when you enter the station, you hear a fire alarm and see that people are running out of the station. Immediately you forget your waiting train and join them on their way out.

This scenario shows the complexity of human perception. Plenty of information is perceived at each instant, much more than can be processed in detail by the human brain. The ability to extract the relevant pieces of the sensory input at an early processing stage is crucial for efficient acting. Thereby, it depends on the context which part of the sensory input is relevant. When having a goal like catching a train, the signs are relevant, without an explicit goal, salient things like the street performers attract the attention. Some things or events are so salient that they even override your goals, such as the fire alarm. The mechanism to direct the processing resources to the potentially most relevant part of the sensory input is called *selective attention*. One of the most famous definitions of selective attention is from William James, a pioneering psychologist, who stated in 1890: "Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought" [12]. While the

S. Frintrop (✉)
Institute of Computer Science III, Rheinische Friedrich-Wilhelms Universität Bonn,
Römerstrasse 164, 53117 Bonn, Germany
e-mail: frintrop@iai.uni-bonn.de

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_4,

concept of attention exists for all senses, here we will concentrate on visual attention and thus on the processing of images and videos.

While it is obvious that attention is a useful concept for humans, why is it of interest for machines and which kinds of machines profit most from such a concept? To answer these questions, let us tackle two goals of attention separately. The first goal is to handle the complexity of the perceptual input. Since many visual processing tasks concerned with the recognition of arbitrary objects are NP-hard [23], an efficient solution is often not achievable. Problems arise for example if arbitrary objects of arbitrary sizes and extents shall be recognized, i.e. everything from the fly on the wall to the building in the background. The typical approach to detect objects in images is the sliding-window paradigm in which a classifier is trained to detect an object in a subregion of the image and is repeatedly applied to differently sized test windows. A mechanism to prioritize the image regions for further processing is of large interest, especially if large image databases shall be investigated or if real-time processing is desired, e.g. on autonomous mobile robots.

The second goal of attention is to support action decisions. This task is especially important for autonomous robots that act in a complex, possibly unknown environment. Even if equipped with unlimited computational power, robots still underlie similar physical constraints as humans: at one point in time, they can only navigate to one location, zoom in on one or a few regions, and grasp one or a few objects. Thus, a mechanism that selects the relevant parts of the sensory input and decides what to do next is essential. Since robots usually operate in the same environments as humans, it is reasonable to imitate the human attention system to fulfill these tasks. Furthermore, in domains as human–robot interaction, it is helpful to generate a joint focus of attention between man and machine to make sure that both communicate about the same object.[1] Having similar mechanisms for both human and robot facilitates this task.

As a conclusion, we can state that the more general a system shall be and the more complex and undefined the input data are, the more urgent the need for a prioritizing attention system that preselects the data of most potential interest.

This chapter aims to provide you with everything you must know to build a computational attention system.[2] It starts with an introduction to human perception (Sect. 4.2). This section will give you an insight to the important mechanisms in the brain that are involved in visual attention and thus provides the background knowledge that is required when working in the field of computational attention. If you are mainly interested in how to build a computational system, you might skip this section and directly jump to Sect. 4.3. This section explains how to build a bottom-up system of visual attention and how to extend such a system to perform visual search for objects. After that, we discuss different ways to evaluate attention systems (Sect. 4.4) and mention two applications of such systems in robotic contexts (Sect. 4.5). At the end of the chapter you find some useful links to Open Source code, freely accessible databases, and further readings on the topic.

[1] The social aspect of human attention is described in Chap. 8, Sect. 8.6.4.1

[2] Parts of this chapter have been published before in [5].

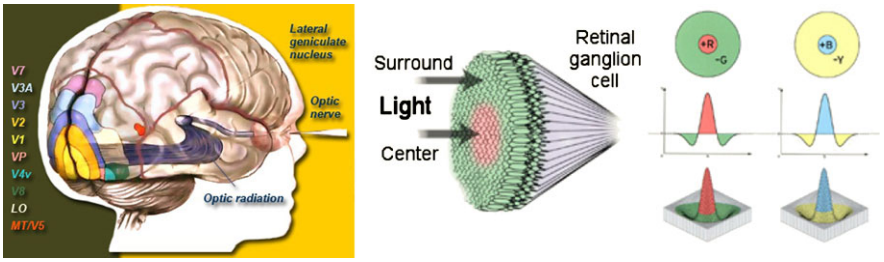


Fig. 4.1 *Left*: The human visual system (Fig. adapted from http://www.brain-maps.com/visual-fields.html). *Right*: The receptive field of a ganglion cell with center and surround and its simulation with Difference-of-Gaussian filters (Fig. adapted from [3])

4.2 Human Visual Attention

In this section, we will introduce some of the cognitive foundations of human visual attention. We start with the involved brain mechanisms, continue with several psychological concepts and evaluation methods, and finally present two influential psychological models.

4.2.1 The Human Visual System

Let us first regard some of the basic concepts of the human visual system. While being far from an exhaustive explanation, we focus on describing parts that are necessary to understand the visual processing involved in selective attention. The most important visual areas are illustrated in Fig. 4.1, left.

4.2.1.1 Eye, Retina, and Ganglion Cells

The light that enters the eye through the *pupil* passes through the *lens*, and reaches the *retina* at the back of the eye. The retina is a light-sensitive surface and is densely covered with over 100 million photoreceptor cells, *rods* and *cones*. The rods are more numerous and more sensitive to light than the cones but they are not sensitive to color. The cones provide the eye's color sensitivity: among the cones, there are three different types of color reception: long-wavelength cones (L-cones) which are sensitive primarily to the red portion of the visible spectrum, middle-wavelength cones (M-cones) sensitive to green, and short-wavelength cones (S-cones) sensitive to blue. In the center of the retina is the *fovea*, a rod-free area with very thin, densely packed cones. It is the center of the eye's sharpest vision. Because of this arrangement of cells, we perceive the small region currently fixated in a high resolution and the whole surrounding only as diffuse and coarse. This mechanism makes eye movements an essential part of perception, since they enable high resolution vision subsequently for different regions of a scene.

The photoreceptors transmit information to the so called *ganglion cells*, which combine the trichromatic (i.e. three-colored) input by subtraction and addition to determine color and luminance opponency. The receptive field of a ganglion cell, i.e. the region the cell obtains input from, is circular and separated into two areas: a center and a surround (cf. Fig. 4.1, right). There are two types of cells: *on-center cells* which are stimulated by light at the center and inhibited by light at the surround, and *off-center cells* with the opposite characteristic. Thus, on-center cells are well suited to detect bright regions on a dark background and off-center cells vice versa. Additional to the luminance contrast, there are also cells that are sensitive to red-green and to blue-yellow contrasts. The center-surround concept of visual cells can be modeled computationally with Difference-of-Gaussian filters (cf. Fig. 4.1, right) and is the basic mechanism for contrast detection in computational attention systems.

4.2.1.2 From the Optic Chiasm to V1

The visual information leaves the eye via the optic nerve and runs to the *optic chiasm*. From here, two pathways go to each brain hemisphere: the smaller one goes to the *superior colliculus* (*SC*), which is e.g. involved in the control of eye movements. The more important pathway goes to the *Lateral Geniculate Nucleus* (*LGN*) and from there to higher brain areas. The LGN consists of six main layers composed of cells that have center-surround receptive fields similar to those of retinal ganglion cells but larger and with a stronger surround. From the LGN, the visual information is transmitted to the *primary visual cortex* (*V1*) at the back of the brain.

V1 is the largest and among the best-investigated cortical areas in primates. It has the same spatial layout as the retina itself. But although spatial relationships are preserved, the densest part of the retina, the fovea, takes up a much smaller percentage of the visual field (1%) than its representation in the primary visual cortex (25%). The cells in V1 can be classified into three types: *simple cells, complex cells*, and *hypercomplex cells*. As the ganglion cells, the simple cells have an excitatory and an inhibitory region. Most of the simple cells have an elongated structure and, therefore, are orientation sensitive. Complex cells take input from many simple cells. They have larger receptive fields than the simple cells and some are sensitive to moving lines or edges. Hypercomplex cells, in turn, receive the signals from complex cells as input. These neurons are capable of detecting lines of a certain length or lines that end in a particular area.

4.2.1.3 Beyond V1: the Extrastriate Cortex and the Visual Pathways

From the primary visual cortex, a large collection of neurons sends information to higher brain areas. These areas are collectively called *extrastriate cortex*, in opposite to the striped architecture of V1. The areas belonging to the extrastriate cortex are

V2, V3, V4, the infero-temporal cortex (IT), the middle temporal area (MT or V5) and the posterior-parietal cortex (PP).[3]

Of the extrastriate areas, much less is known than of V1. One of the most important findings of the last decades was that the processing of the visual information is not serial but highly parallel. While not completely segregated, each area has a prevalence of processing certain features such as color, form (shape), or motion. Several pathways lead to different areas in the extrastriate cortex. The statements on the number of existing pathways differ: the most common belief is that there are three main pathways; one for color, one for form, and one for motion pathway, which is also responsible for depth processing [13].

The color and form pathways go through V1, V2, and V4 and end finally in IT, the area where the recognition of objects takes place. In other words, IT is concerned with the question of "what" is in a scene. Therefore, the color and form pathway together are called the *what pathway*. It is also called *ventral stream* because of its location on the ventral part of the body. The motion-depth pathway goes through V1, V2, V3, MT, and the parieto occipital area (PO) and ends finally in PP, responsible for the processing of motion and depth. Since this area is mainly concerned with the question of "where" something is in a scene, this pathway is also called the *where pathway*. Another name is *dorsal stream* because it is considered to lie dorsally.

Finally, it is worth to mention that although the processing of the visual information was described above in a feed-forward manner, it is usually bi-directional. Top-down connections from higher brain areas influence the processing and go down as far as LGN. Also lateral connections combine the different areas, for example, there are connections between V4 and MT, showing that the "what" and "where" pathway are not completely separated.

4.2.1.4 Neurobiological Correlates of Visual Attention

The mechanisms of selective attention in the human brain still belong to the open problems in the field of research on perception. Perhaps the most prominent outcome of neuro-physiological findings on visual attention is that there is no single brain area guiding the attention, but neural correlates of visual selection appear to be reflected in nearly all brain areas associated with visual processing. Attentional mechanisms are carried out by a network of anatomical areas. Important areas of this network are the posterior parietal cortex (PP), the superior colliculus (SC), the Lateral IntraParietal area (LIP), the Frontal Eye Field (FEF) and the pulvinar.

Brain areas involved in guiding eye movements are the FEF and the SC. There is also evidence that a kind of *saliency map* (i.e. a topographical representation of what is interesting—or salient—in the visual field) exists in the brain, but the opinions where it is located diverge. Some researchers locate it in the FEF, others at the LIP, the SC, at V1 or V4 (see [5] for references). Further research will be necessary to determine the tasks and interplay of the brain areas involved in the process of visual attention.

[3]The notation V1 to V5 comes from the former belief that the visual processing would be serial.

4.2.2 Psychological Concepts of Attention

Certain concepts and expressions are frequently used when investigating human visual attention and shall be introduced here.

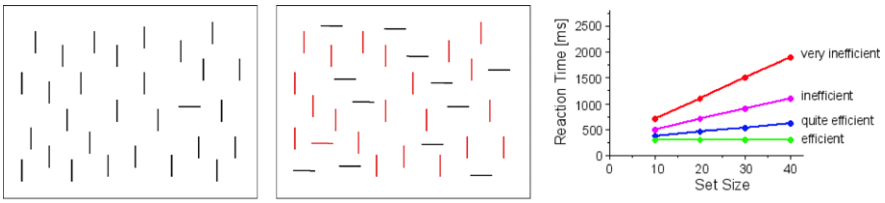
Usually, directing the focus of attention to a region of interest is associated with eye movements (*overt attention*). However, it is also possible to attend to peripheral locations of interest without moving the eyes, a phenomenon which is called *covert attention*. The allocation of attention is guided by two principles: *bottom-up and top-down factors*. Bottom-up attention (or *saliency*) is derived solely from the perceptual data. Main indicators for visual bottom-up saliency are a strong contrast of a region to its surround and the uniqueness of this region. Thus, a clown in the parliament is salient, whereas it is not particularly salient among other clowns (however, a whole group of clowns in the parliament is certainly salient!). The bottom-up influence is not voluntary suppressible: a highly salient region captures your attention regardless of the task, an effect called *attentional capture*. This effect might save your life, e.g. if an emergency bell or a fire captures your attention.

On the other hand, top-down attention is driven by cognitive factors such as preknowledge, context, expectations, and current goals. In human viewing behaviour, top-down cues always play a major role. Not only looking for the train station signs in the introductory example is an example of top-down attention, also more subtle influences like looking at food when being hungry. In psychophysics, top-down influences are often investigated by so called *cueing experiments*, in which a cue directs the attention to a target. A cue might be an arrow that points into the direction of the target, a picture of the target, or a sentence ("search for the red circle").

One of the best investigated aspects of top-down attention is *visual search*. The task is exactly what the name implies: given a target and an image, find an instance of the target in the image. Visual search is omnipresent in every-day life: finding a friend in a crowd or your keys in the living room are examples.

In psychophysical experiments, the efficiency of visual search is measured by the *reaction time* (RT) that a subject needs to detect a target among a certain number of *distractors* (the elements that differ from the target) or by the search accuracy. To measure the RT, a subject has to report a detail of the target or has to press one button if the target was detected and another if it is not present in the scene. The RT is represented as a function of set size (the number of elements in the display). The search efficiency is determined by the slopes and the intercepts of these RT $\times$ set size functions (cf. Fig. 4.2(c)). The searches vary in their efficiency: the smaller the slope of the function and the lower the value on the y-axis, the more efficient the search. Two extremes are serial and parallel search. In serial search, the reaction time increases with the number of distractors, whereas in parallel search, the slope is near zero. But note that the space of search slope functions is a continuum.

Feature searches take place in settings in which the target is distinguished from the distractors by a single basic feature (such as color or orientation) (cf. Fig. 4.2(a)). In *conjunction searches* on the other hand, the target differs by more than one feature (see Fig. 4.2(b)). While feature searches are usually fast and conjunction searches



(a) Feature search (b) Conjunction search (c) Continuum of search slopes

Fig. 4.2 (**a**) Feature search: the target (*horizontal line*) differs from the distractors (*vertical lines*) by a unique visual feature (pop-out effect). (**b**) Conjunction search: the target (*red*, *horizontal line*) differs from the distractors (*red*, *vertical* and *black*, *horizontal lines*) by a conjunction of features. (**c**) The reaction time (RT) of a visual search task is a function of set size. The efficiency is measured by the intercept and slopes of the functions (Fig. redrawn from [28])

slower, this is not always the case. Also a feature search might be slow if the difference between target and distractors is small (e.g. a small deviation in orientation). Generally, it can be said that search becomes harder as the target-distractor similarity increases and easier as distractor-distractor similarity increases. The most efficient search takes place for so called "pop-out" experiments that denote settings in which a single element immediately captures the attention of the observer. You understand easily what this means by looking at Fig. 4.2(a). Other methods to investigate visual search is by measuring accuracy or eye movements. References for further readings on this topic can be found in [7].

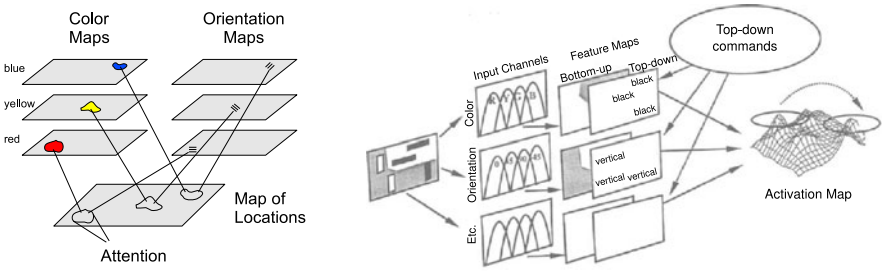
One purpose of such experiments is to study the *basic features* of human perception, that means the features that are early and pre-attentively processed in the human brain and guide visual search. Undoubted basic features are color, motion, orientation and size (including length and spatial frequency) [29]. Some other features are guessed to be basic, but there are limited data or dissenting opinions.

An interesting effect in visual search tasks are *search asymmetries*, that means the effect that a search for stimulus 'A' among distractors 'B' produces different results than a search for 'B' among 'A's. An example is that finding a tilted line among vertical distractors is easier than vice versa. An explanation is proposed by [22]: the authors claim that it is easier to find deviations among canonical (i.e. more frequently encountered in everyday life) stimuli than vice versa. Given that vertical is a canonical stimulus, the tilted line is a deviation and may be detected quickly.

4.2.3 Important Psychological Attention Models

In the field of psychology, there exists a wide variety of theories and models on visual attention. Their objective is to explain and better understand human perception. Here, we introduce two approaches which have been most influential for computational attention systems.

The *Feature Integration Theory* (*FIT*) of Treisman claims that "different features are registered early, automatically and in parallel across the visual field, while



(a) Feature integration theory (b) Guided search model

Fig. 4.3 (**a**) Model of the *Feature Integration Theory* (*FIT*) (Fig. redrawn from [20]). (**b**) The *Guided Search model* of Wolfe (Fig. adapted from [27] ©1994 Psychonomic Society)

objects are identified separately and only at a later stage, which requires focused attention" [21]. Information from the resulting *feature maps*—topographical maps that highlight conspicuities according to the respective feature—is collected in a *master map of location*. Scanning serially through this map focuses the attention on the selected scene regions and provides these data for higher perception tasks (cf. Fig. 4.3(a)). The theory was first introduced in 1980, but it was steadily modified and adapted to current research findings.

Beside FIT, the *Guided Search Model* of Wolfe is among the most influential work for computational visual attention systems [27]. The basic goal of the model is to explain and predict the results of visual search experiments. Mimicking the convention of numbered software upgrades, Wolfe has denoted successive versions of his model as Guided Search 1.0 to Guided Search 4.0. The best elaborated description of the model is available for Guided Search 2.0 [27]. The architecture of the model is depicted in Fig. 4.3(b). It shares many concepts with the FIT, but is more detailed in several aspects, which are necessary for computer implementations. An interesting point is that in addition to bottom-up saliency, the model also considers the influence of top-down information by selecting the feature type which distinguishes the target best from its distractors.

4.3 Computational Attention Systems

Computational attention systems model the principles of human selective attention and aim to select the part of the sensory input data that is most promising for further investigation. Here, we concentrate on visual attention systems that are inspired by concepts of the human visual system but are designed with an engineering objective, that means their purpose is to improve vision systems in technical applications.[4]

[4]In this chapter, we assume that the reader has basic knowledge on image processing, otherwise you find a short explanation of the basic concepts in the appendix of [5].

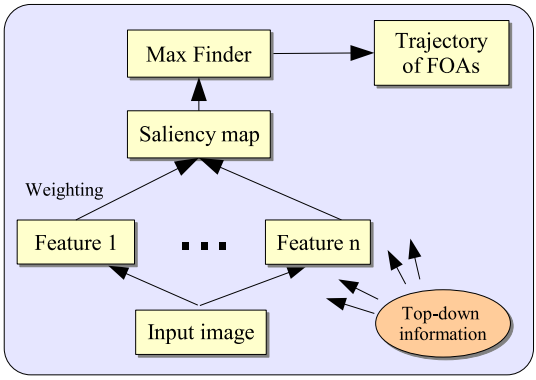


Fig. 4.4 General structure of most visual attention systems. Several features are computed in parallel and fused to a single saliency map. The maxima in the saliency map are the foci of attention (FOAs). Output is a trajectory of FOAs, ordered by decreasing saliency. Top-down cues may influence the processing on different levels

4.3.1 General Structure

Most computational attention systems have a similar structure, which is depicted in Fig. 4.4. This structure is originally adapted from psychological theories like the Feature Integration Theory and the Guided Search model (cf. Sect. 4.2.3). The main idea is to compute several features in parallel and to fuse their conspicuities in a saliency map. If top-down information is available, this can be used to influence the processing at various levels of the models. A saliency map is usually a gray-level image in which the brightness of a pixel is proportional to its saliency. The maxima in the saliency map denote the regions that are investigated by the focus of attention (FOA) in the order of decreasing saliency. This trajectory of FOAs shall resemble human eye movements. Output of a computational attention system is either the saliency map or a trajectory of focused regions.

While most attention systems share this general structure, there are different ways to implement the details. One of the best known computational attention systems is the iNVT from Itti and colleagues [11]. The VOCUS model [5] has adopted and extended several of their ideas. It is real-time capable and has a top-down mode to search for objects (visual search). Itti and Baldi presented an approach that is able to detect temporally salient regions, called *surprise theory* [9]. Bruce and Tsotsos compute saliency by determining the self-information of image regions with respect to their surround [1]. The types of top-down information that can influence an attention model are numerous and only a few have been realized in computational system. For example, the VOCUS model uses pre-knowledge about a target to weight the feature maps and perform visual search. Torralba et al. use context information about the scene to guide the gaze, e.g., to search for people on the street level of an image rather than on the sky area [19]. More abstract types of top-down cues, such as emotions and motivations, have to our knowledge not yet been integrated into computational attention systems.

In this chapter, we follow the description of the VOCUS model as representative of one of the classic approaches to compute saliency.[5] We start with introducing the bottom-up part (Sect. 4.3.2), followed by a description of the top-down visual search part (Sect. 4.3.3).

4.3.2 Bottom-up Saliency

Bottom-up saliency is usually a combination from different feature channels. The most frequently used features in visual attention systems are intensity, color, and orientation. When image sequences are processed, also motion and flicker are important. The main concept to compute saliency are contrast computations that determine the difference between a center region and a surrounding region with respect to a certain feature. These contrasts are usually computed by *center-surround filters*. Such filters are inspired by cells in the human visual system, as the ganglion cells and the simple and complex cells introduced in Sect. 4.2.1. Cells with circular receptive fields are best modeled by Difference-of-Gaussian filters (cf. Fig. 4.1, right) while cells with elongated receptive fields are best modeled by Gabor functions. In practice, the circular regions are usually approximated by rectangles.

To enable the detection of regions of different extents, the center as well as the surround vary in size. Instead of directly adapting the filter sizes, the computations are usually performed on the layers of an image pyramid.

The structure of the bottom-up part of the VOCUS attention system is shown in Fig. 4.5. Let us regard the computation of the intensity feature in more detail now to understand the concept and then extend the ideas to the other feature channels.

4.3.2.1 Intensity Channel

Given a color input image I, this image is first converted to an image I_{Lab} in the Lab (or CIELAB) color space. This space has the dimension 'L' for lightness and 'a' and 'b' for the color-opponent dimensions (cf. Fig. 4.5, bottom right); it is perceptually uniform, which means that a change of a certain amount in a color value is perceived as a change of about the same amount in human visual perception.

From I_{Lab}, a Gaussian pyramid is determined by successively smoothing the image with a Gaussian filter and subsampling it with a factor of two along each coordinate direction (see Fig. 4.6). In VOCUS, we use a 5×5 Gaussian kernel. The level of the pyramid determines the area that the center-surround filter covers: on high levels of the pyramid (fine resolution), small salient regions are detected while on low levels (coarse resolution), large regions obtain the highest response. In VOCUS, 5 pyramid levels (scales) are computed: I^s_{Lab}, $s \in \{0, \ldots, 4\}$. Level I^1_{Lab} is

[5] While the description here is essentially the same as in [5], some improvements have been made in the meantime that are included here. Differences of VOCUS from the iNVT can be found in [5].

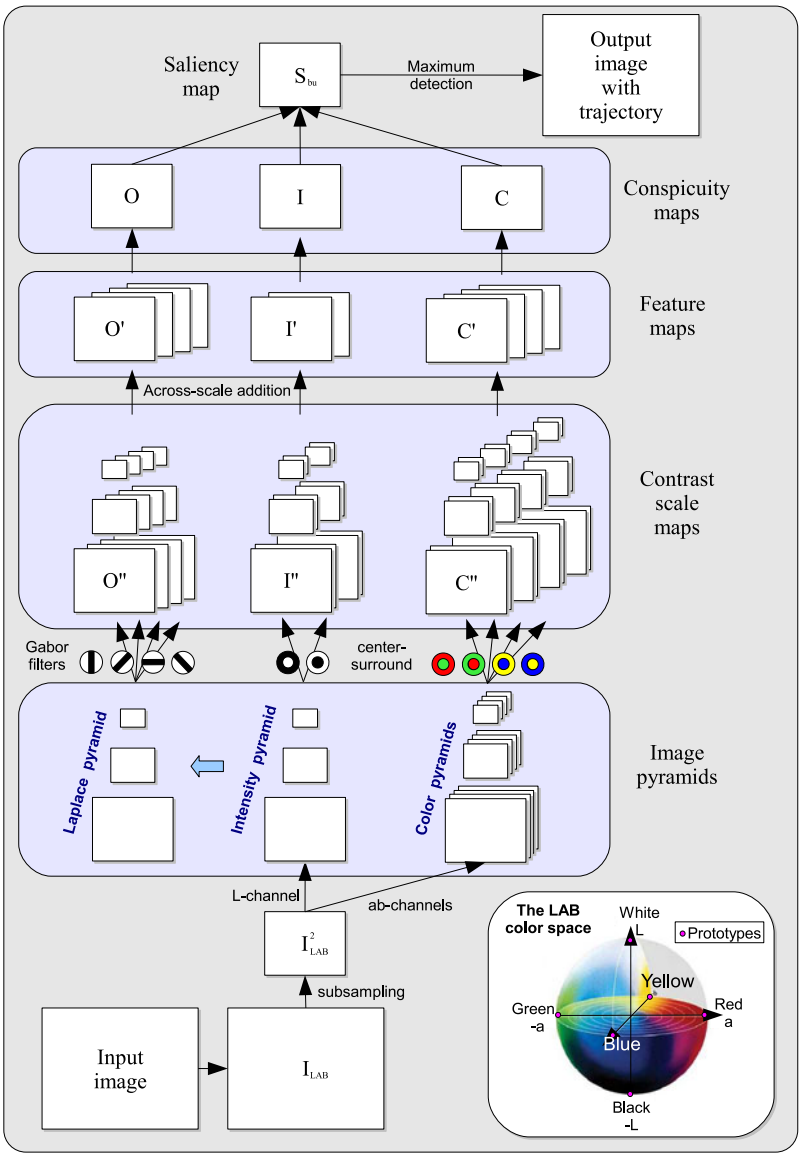


Fig. 4.5 The bottom-up saliency computation of the VOCUS attention system

only an intermediate step used for noise reduction, all computations take place on levels 2–4.[6]

[6]The number of levels that is reasonable depends on the image size, as well as on the size of the objects you want to detect. Larger images and a wide variety of possible object sizes require deeper

(a) Input image (b) Gaussian pyramid, s_0 to s_4

Fig. 4.6 (**a**) The image which serves as demonstration example throughout this chapter and (**b**) the derived Gaussian image pyramid

The intensity computations can be performed directly on the images I_L^s that originate from the 'L' channel of the LAB image. According to the human system, we determine two feature types for intensity: the on-center difference responding strongly to bright regions on a dark background, and the off-center difference vice versa. Note that it is important to treat both types separately and to not fuse them in a single map since otherwise it is not possible to detect bright-dark pop-outs, such as in Fig. 4.12. This yields 12 intensity scale maps $I''_{i,s,\sigma}$ with $i \in \{\text{(on)}, \text{(off)}\}, s \in \{2,3,4\}, \sigma \in \{3,7\}$. A pixel (x, y) in one of the on-center scale maps is thus computed as

$$\begin{aligned} I''_{on,s,\sigma}(x,y) &= center(I_L^s, x, y) - surround_\sigma\left(I_L^s, x, y\right) \\ &= I_L^s(x,y) - \frac{1}{(2\sigma+1)^2 - 1} \\ &\quad \times \left(\sum_{i=-\sigma}^{\sigma} \sum_{j=-\sigma}^{\sigma} I_L^s(x+i, y+j) - I_L^s(x,y) \right). \end{aligned} \tag{4.1}$$

The off-center maps $I''_{\text{off},s,\sigma}(x,y)$ are computed equivalently by $surround - center$. The straightforward computation of the surround value is quite costly, especially for large surrounds. To compute the surround value efficiently, it is convenient to use *integral images* [25].

The advantage of an integral image (or summed area table) is that when it is once created, the sum and mean of the pixel values of a rectangle of arbitrary size can be computed in constant time. An integral image II is an intermediate representation for the image and contains for a pixel position (x, y) the sum of all gray scale pixel values of image I above and left of (x, y), inclusive:

$$II(x,y) = \sum_{x'=0}^{x} \sum_{y'=0}^{y} I(x', y'). \tag{4.2}$$

The process is visualized in Fig. 4.7, left. The integral image can be computed recursively in one pass over the image with the help of the cumulative sum s:

$$s(x,y) = s(x, y-1) + I(x,y), \tag{4.3}$$

pyramids. The presented approach usually works well for images of up to 400 pixels in width and height in which the objects are comparatively small as in the example images of this chapter.

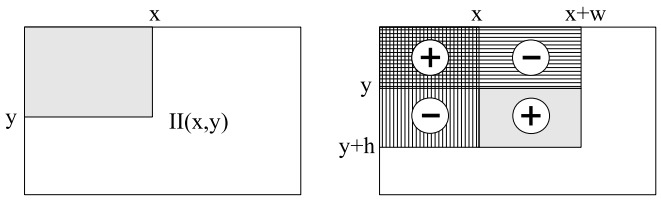


Fig. 4.7 *Left*: The integral image contains at $II(x, y)$ the sum of the pixel values in the *shaded region*. *Right*: the computation of the average value in the *shaded region* is based on four operations on the four depicted rectangles according to (4.5)

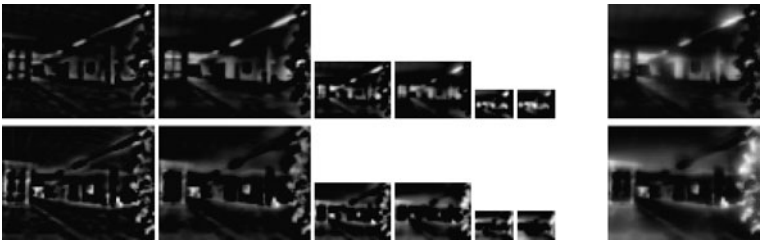


Fig. 4.8 *Left*: the 12 intensity scale maps $I''_{i,s,\sigma}$. *First row*: the *on-maps*. *Second row*: the *off-maps*. *Right*: the two intensity feature maps $I'_{(\text{on})}$ and $I'_{(\text{off})}$ resulting from the sum of the corresponding six scale maps on the *left*

$$II(x, y) = II(x - 1, y) + s(x, y) \tag{4.4}$$

with $s(x, -1) = 0$ and $II(-1, y) = 0$. This intermediate representation allows to compute the sum of the pixel values in a rectangle F using four references (see Fig. 4.7 (right)):

$$\begin{aligned} F(x, y, h, w) = {} & II(x + w - 1, y + h - 1) - II(x - 1, y + h - 1) \\ & - II(x + w - 1, y - 1) + II(x - 1, y - 1). \end{aligned} \tag{4.5}$$

The '−1' elements in the equation are required to obtain a rectangle that includes (x, y). By replacing the computation of the surround in (4.1) with the integral computation in (4.5) we obtain

$$I''_{on,s,\sigma}(x, y) = I^s_L(x, y) - \frac{F(x - \sigma, y - \sigma, 2\sigma + 1, 2\sigma + 1) - I^s_L(x, y)}{(2\sigma + 1)^2 - 1}. \tag{4.6}$$

To enable this computation, one integral image has to be computed for each of the three pyramid levels I^s_L, $s \in \{2, 3, 4\}$. This pays off since then each surround can be determined by three simple operations. The intensity scale maps I'' are depicted in Fig. 4.8, left.

The six maps for each center-surround variation are summed up by *across-scale addition*: first, all maps are resized to scale 2 whereby resizing scale i to scale $i - 1$

is done by bilinear interpolation. After resizing, the maps are added up pixel by pixel. This yields the intensity feature maps I':

$$I'_i = \bigoplus_{s,\sigma} I''_{i,s,\sigma}, \tag{4.7}$$

with $i \in \{\text{(on)}, \text{(off)}\}$, $s \in \{2, 3, 4\}$, $\sigma \in \{3, 7\}$, and $\bigoplus$ denoting the across-scale addition. The two intensity feature maps are shown in Fig. 4.8, right.

4.3.2.2 Color Channel

The color computations are performed on the two-dimensional color layer I_{ab} of the Lab image that is spanned by the axes 'a' and 'b'. Besides its resemblance to human visual perception, the Lab color space fits particularly well as a basis for an attentional color channel since the four main colors red, green, blue and yellow are at the end of the axes 'a' and 'b'. Each of the 6 ends of the axes that confine the color space serves as one prototype color, resulting in two intensity prototypes for white and black and four color prototypes for red, green, blue, and yellow.

For each color prototype, a color prototype image is computed on each of the pyramid levels 2–4. In these maps, each pixel represents the Euclidean distance to the prototype:

$$C^s_\gamma(x, y) = V_{max} - \left\| I^s_{ab}(x, y) - P_\gamma \right\|, \quad \gamma \in \{\text{R}, \text{G}, \text{B}, \text{Y}\}, \tag{4.8}$$

where V_{max} is the maximal pixel value and the prototypes P_γ are the ends of the 'a' and 'b' axes (thus, in an 8-bit image, we have $V_{max} = 255$ and $P_\text{R} = (255, 127)$, $P_\text{G} = (0, 127)$, $P_\text{B} = (127, 0)$, $P_\text{Y} = (127, 255)$). The color prototype maps show to which degree a color is represented in an image, i.e., the maps in the pyramid P_R show the "redness" of the image regions: the brightest values are at red regions and the darkest values at green regions (since green has the largest distance to red in the color space). Analogously to the intensity channel, it is also important here to separate red-green and blue-yellow in different maps to enable red-green and blue-yellow pop-outs. The four color prototype images I^2_γ are displayed in Fig. 4.9 (first row).

On these pyramids, the color contrast is computed by on-center differences, yielding $4 * 3 * 2 = 24$ color scale maps:

$$C''_{\gamma,s,\sigma} = center\big(C^s_\gamma, x, y\big) - surround_\sigma\big(C^s_\gamma, x, y\big), \tag{4.9}$$

with $\gamma \in \{\text{R}, \text{G}, \text{B}, \text{Y}\}$, $s \in \{2, 3, 4\}$, and $\sigma \in \{3, 7\}$. According to the intensity channel, the center is a pixel in the corresponding color prototype map, and the surround is computed according to (4.6). The off-center-on-surround difference is not needed, because these values are represented in the opponent color pyramid. The maps of each color are rescaled to the scale 2 and summed up into four color feature

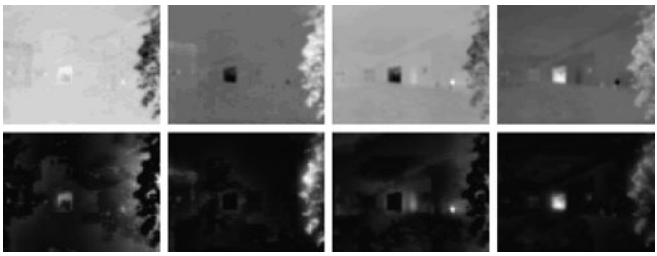


Fig. 4.9 *Top*: the color prototype images of scale s_2 for *red*, *green*, *blue*, *yellow*. *Bottom*: the corresponding color feature C'_γ maps which result after applying center-surround filters

maps C'_γ:

$$C'_\gamma = \bigoplus_{s,\sigma} C''_{\gamma,s,\sigma}. \tag{4.10}$$

Figure 4.9, bottom, shows the color feature maps for the example image.

4.3.2.3 Orientation Channel

The orientation maps are computed from *oriented pyramids*. An oriented pyramid contains one pyramid for each represented orientation (cf. Fig. 4.10, left). Each of these pyramids highlights edges with this specific orientation. To obtain the oriented pyramid, first a Laplacian Pyramid is obtained from the Gaussian pyramid I^s_L by subtracting adjacent levels of the pyramid. The orientations are computed by *Gabor filters* which respond most to bar-like features according to a specified orientation. Gabor filters, which are the product of a symmetric Gaussian with an oriented sinusoid, simulate the receptive field structure of orientation-selective cells in the primary visual cortex (cf. Sect. 4.2.1). Thus, Gabor filters replace the center-surround filters of the other channels.

Four different orientations are computed yielding $4 \times 3 = 12$ orientation scale maps $O''_{\theta,s}$, with the orientations $\theta \in \{0°, 45°, 90°, 135°\}$ and scales $s \in \{2, 3, 4\}$. The orientation scale maps $O''_{\theta,s}$ are summed up by across-scale addition for each orientation, yielding four orientation feature maps O'_θ, one for each orientation:

$$O'_\theta = \bigoplus_{s} O''_{\theta,s}, \tag{4.11}$$

The orientation feature maps for the example image are depicted in Fig. 4.10, right.

4.3.2.4 Motion Channel

If image sequences are used as input for the attention system, motion is an important additional feature. It can be computed easily by determining the optical flow field.

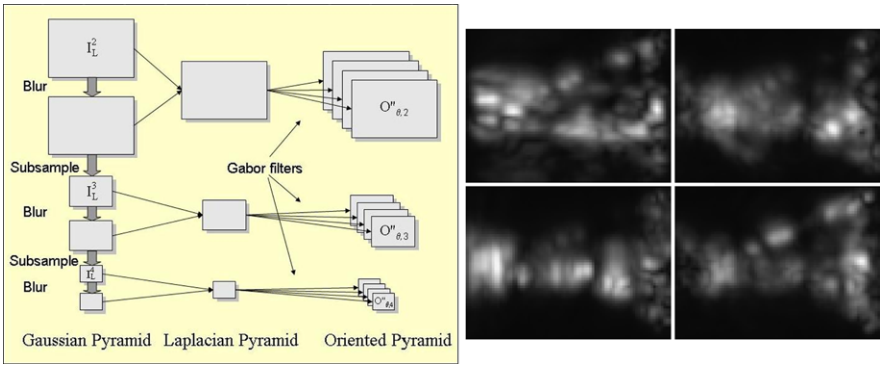


Fig. 4.10 *Left*: to obtain an oriented pyramid, a Gaussian pyramid is computed from the input image; then a Laplacian pyramid is obtained from the Gaussian pyramid by subtracting two adjacent levels and, finally, Gabor filters of four orientations are applied to each level of the Laplacian pyramid. *Right*: The four orientation feature maps O'_{0°, O'_{45°, O'_{90°, O'_{135° for the example image

Fig. 4.11 The motion feature maps M' for a scene in which a ball rolls from *right* to *left* through the image. From *left* to *right*: example frame, motion maps M'_{right}, M'_{left}, M'_{up}, M'_{down}

Here, we use a method based on total variation regularization that determines a dense optical flow field and is capable to operate in real-time [30]. If the horizontal u and the vertical v component of the optical flow are visualized as images, the center-surround filters can be applied to these images directly. By applying on- as well as off-center filters to both images, we achieve four motion maps for each scale s which we call $M''_{\vartheta,s}$, with $\vartheta = \{\text{right}, \text{left}, \text{up}, \text{down}\}$. After across-scale addition we obtain four motion feature maps,

$$M'_{\vartheta} = \bigoplus_{s} M''_{\vartheta,s}. \tag{4.12}$$

An example for a sequence in which a ball rolls from right to left through the image is displayed in Fig. 4.11. In videos, motion itself is not necessarily salient, but the contrast of the motion in the current frame to the motion (or absence of motion) in previous frames is. Itti and Baldi describe in their surprise theory how such temporal saliency can be integrated into a computational attention system [9].

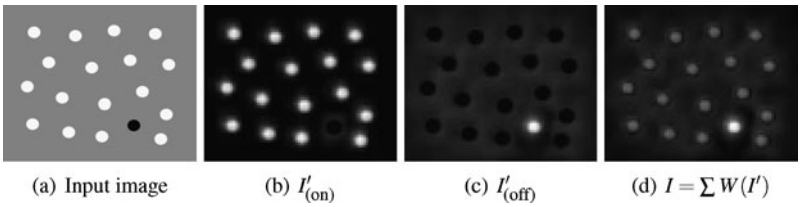


Fig. 4.12 The effect of the uniqueness weight function W (4.13). The off-center intensity feature map $I'_{(off)}$ has a higher weight than the on-center intensity feature map $I'_{(on)}$, because it contains only one strong peak. So this map has a higher influence and the region of the black dot pops out in the conspicuity map I

4.3.2.5 The Uniqueness Weight

Up to now, we have computed local contrasts for each of the feature channels. While contrast is an important aspect of salient regions, they additionally have an important property: they are rare in the image, in the best case unique. A red ball on grass is very salient, while it is much less salient among other red balls. That means, we need a measure for the uniqueness of a feature in the image. Then, we can strengthen maps with rare features and diminish the influence of maps with omnipresent features.

A simple method to determine the uniqueness of a feature is to count the number of local maxima m in a feature map X. Then, X is divided by the square root of m:

$$W(X) = X/\sqrt{m}, \tag{4.13}$$

In practice, it is useful to only consider maxima in a pre-specified range from the global maximum (in VOCUS, the threshold is 50% of the global maximum of the map). Figure 4.12 shows how the uniqueness weight enables the detection of pop-outs. Other solutions to determine the uniqueness are described in [10, 11].

4.3.2.6 Normalization

Before the feature maps can be fused, they have to be normalized. This is necessary since some channels have more maps than others. Let us first understand why this step is not trivial. The easiest solution would be to normalize all maps to a fixed range. This method comes with a problem: normalizing maps to a fixed range removes important information about the magnitude of the maps. Assume that one intensity and one orientation map belonging to an image with high intensity but low orientation contrasts are to be fused into one saliency map. The intensity map will contain very bright regions, but the orientation map will show only some moderately bright regions. Normalizing both maps to a fixed range forces the values of the orientation maps to the same range as the intensity values, ignoring that orientation is not an important feature in this case.

A similar problem occurs when dividing each map by the number of maps in this channel: imagine an image with equally strong intensity and color blobs. A color map would be divided by four, an intensity map only by two. Thus, although all blobs have the same strength, the intensity blobs would obtain a higher saliency value.

Instead, we propose the following normalization technique: To fuse the maps $\mathbf{X} = \{X_1, \ldots, X_k\}$, determine the maximum value M of all $X_i \in \mathbf{X}$ and normalize each map to the range $[0 \ldots M]$. Normalization of map X_i to the range $[0 \ldots M]$ will be denoted as $N_{[0..M]}(X_i)$ in the following.

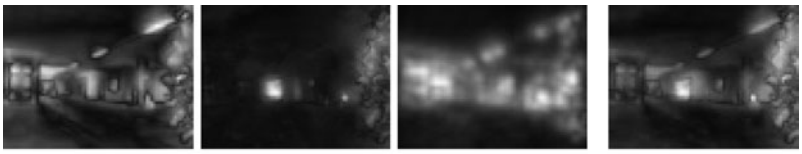
4.3.2.7 The Conspicuity Maps

The next step in the saliency computation is the generation of the *conspicuity maps*. The term conspicuity is usually used to denote feature-specific saliency. To obtain the maps, all feature maps of one feature dimension are weighted by the uniqueness weight W, normalized, and combined into one conspicuity map, yielding map I for intensity, and C for color, O for orientation, and M for motion:

$$
\begin{aligned}
&I = \sum_i N_{[0..M_i]}\big(W(I'_i)\big), \qquad M_i = maxvalue_i(I'_i),\\
&i \in \{\text{on, off}\},\\
&C = \sum_\gamma N_{[0..M_\gamma]}\big(W(C'_\gamma)\big), \qquad M_\gamma = maxvalue_\gamma(C'_\gamma),\\
&\gamma \in \{\text{R, G, B, Y}\},\\
&O = \sum_\theta N_{[0..M_\theta]}\big(W(O'_\theta)\big), \qquad M_\theta = maxvalue_\theta(O'_\theta),\\
&\theta \in \{0°, 45°, 90°, 135°\},\\
&M = \sum_\vartheta N_{[0..M_\vartheta]}\big(W(M'_\vartheta)\big), \qquad M_\vartheta = maxvalue_\vartheta(C'_\vartheta),\\
&\vartheta \in \{\text{right, left, up, down}\},
\end{aligned}
\tag{4.14}
$$

where W is the uniqueness weight, N the normalization and *maxvalue* the function that determines the maximal value from several feature maps. The conspicuity maps I, C, and O are illustrated in Fig. 4.13(a)–(c).[7]

[7] Since the input is a static image, the motion channel is empty and omitted here.



(a) Consp. map I (b) Consp. map C (c) Consp. map O (d) Saliency map S

Fig. 4.13 The three conspicuity maps for intensity, color, and orientation, and the saliency map

4.3.2.8 The Saliency Map and Focus Selection

Finally, the conspicuity maps are weighted and normalized again, and summed up to the bottom-up saliency map S:

$$S_{\text{bu}} = \sum_{X_i} N_{[0..M_C]}\big(W(X_i)\big), \quad M_C = maxvalue(I, C, O, M), \; X_i \in \{I, C, O, M\}. \tag{4.15}$$

The saliency map for our (static) example is illustrated in Fig. 4.13(d). While it is sometimes sufficient to compute the saliency map and provide it as output, it is often required to determine a trajectory of image locations which resembles eye movements. To obtain such a trajectory from the saliency map, it is common practice to determine the local maxima in the saliency map, ordered by decreasing saliency. These maxima are usually called *Focus of Attention* (*FOA*). Here, we first discuss the standard, biologically motivated approach to find FOAs, then we introduce a simple, computationally convenient solution.

The standard approach to detect FOAs in the saliency map is via a *Winner-Take-All Network* (*WTA*) (cf. Fig. 4.14) [14]. A WTA is a neural network that localizes the most salient point x_i in the saliency map. Thus, it represents a neural maximum finder. Each pixel in the saliency map gives input to a node in the input layer. Local competitions take place between neighboring units and the more active unit transmits the activity to the next layer. Thus, the activity of the maximum will reach the top of the network after $k = \log_m(n)$ time steps if there are n input units and local comparisons take place between m units. However, since it is not the value of the maximum that is of interest but the location of the maximum, a second pyramid of auxiliary units is attached to the network. It has a reversed flow of information and "marks" the path of the most active unit. An auxiliary unit is activated if it receives excitation from its main unit, as well as from the auxiliary unit at the next higher layer. The auxiliary unit y_i, corresponding to the most salient point x_i, will be activated after at most $2\log_m(n)$ time steps. On a parallel architecture with locally connected units, such as the brain, this is a fast method to determine the maximum. It is also a useful approach on a parallel computer architecture, such as a graphics processing unit (GPU). However, if implemented on a serial machine, it is more convenient to simply scan the saliency map sequentially and determine the most salient value. This is the solution chosen for VOCUS.

When the most salient point has been found, the surrounding salient region is determined by *seeded region growing*. This method starts with a seed, here the most

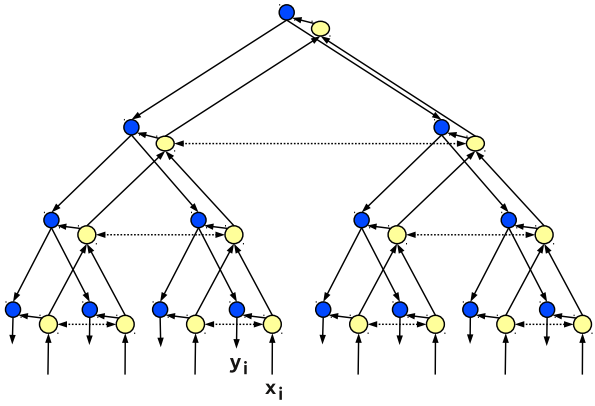


Fig. 4.14 A Winner-Take-All network (WTA) is a neural maximum finder that detects the most salient point x_i in the saliency map. Fig. redrawn from [14]

salient point, and recursively finds all neighbors with similar values within a certain range. In VOCUS, we accept all values that differ at most 25% from the value of the seed. We call the selected region *most salient region* (*MSR*). Some MSRs are shown in Fig. 4.18. For visualization, the MSR is often approximated by an ellipse (cf. Fig. 4.22).

To allow the FOA to switch to the next salient region with a WTA, a mechanism called *inhibition of return* (*IOR*) is used. It inhibits all units corresponding to the MSR by setting their value to 0. Then, the WTA activates the next salient region. If it is desired that the FOA may return to a location after a while, as it is the case in human perception, the inhibition is only active for a pre-defined time and diminishes after that. If no WTA is used, it is more convenient to directly determine all local maxima in the saliency map that exceed a certain threshold (in VOCUS, 50% of the global maximum), sort them by saliency value, and then switch the focus from one to the next. This also prevents border effects that result from inhibition when the focus returns to the borders of an inhibited region.

4.3.3 Visual Search with Top-down Cues

While bottom-up saliency is an important part of visual attention, top-down cues are even more important in many applications. Bottom-up saliency is useful if no pre-knowledge is available, but the exploitation of available pre-knowledge naturally increases the performance of every system, both biological and technical. One of the best investigated aspects of top-down knowledge is visual search. In visual search, a target shall be located in the image, e.g. a cup, a key-fob, or a book. Here, we describe the visual search mode of the VOCUS model. Learning the appearance of the target from a training image and searching for the target in a test image are both directly integrated into the previously described model. Top-down and bottom-up cues interact to achieve a joint focus of attention.

An overview of the complete algorithm for visual search is shown in Fig. 4.15.

Learning mode (input: training image and region of interest (ROI)):
compute bottom-up saliency map S_{bu}
determine most salient region (MSR) in ROI of S_{bu}
for each feature and conspicuity map X_i
compute target descriptor value v_i
Search mode (input: test image and target descriptor v):
compute bottom-up saliency map S_{bu}
compute top-down saliency map S_{td}:
compute excitation map $E = \sum_i (v_i * X_i)$ $\forall i : v_i > 1$
compute inhibition map $I = \sum_i ((1/v_i) * X_i)$ $\forall i : v_i < 1$
compute top-down saliency map $S_{\text{td}} = E - I$
compute saliency map $S = t * S_{\text{td}} + (1 - t) * S_{\text{bu}}$ with $t \in [0..1]$
determine most salient region(s) in S

Fig. 4.15 The algorithm for visual search

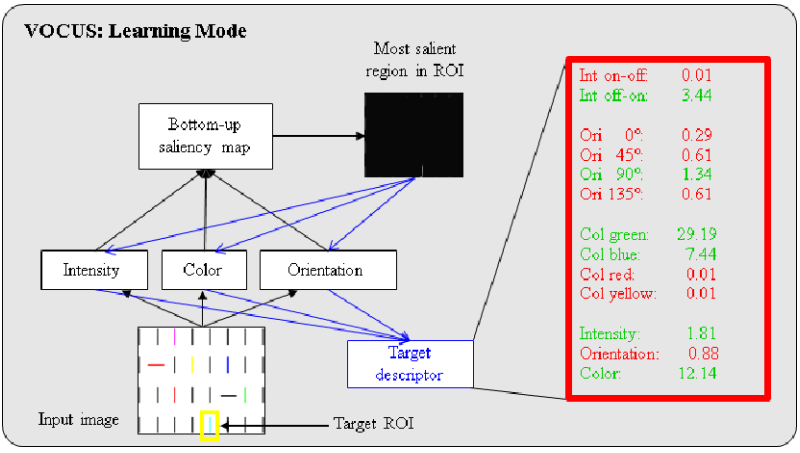


Fig. 4.16 In learning mode, VOCUS determines the *most salient region* (*MSR*) within the *region of interest* (*ROI*) (*yellow rectangle*). A target descriptor **v** is determined by the ratio of MSR vs. background for each feature and conspicuity map. Values $v_i > 1$ (*green*) are target relevant and used in search mode for excitation, values $v_i < 1$ (*red*) are used for inhibition

4.3.3.1 Learning Mode

"Learning" in our application means to determine the object properties of a specified target from one or several training images. In learning mode, the system is provided with a region of interest (ROI) containing the target object and learns which features distinguish the target best from the remainder of the image. For each feature, a value is determined that specifies to what amount the feature distinguishes the target from its background. This yields a target descriptor **v** which is used in search mode to weight the feature maps according to the search task (cf. Fig. 4.16).

The input to the system in learning mode is a training image and a ROI. The ROI is a rectangle which is usually determined manually by the user, but might also be the output of a classifier that specifies the target. Inside the ROI, the *most salient region* (*MSR*) is determined by first computing the bottom-up saliency map and, second, determining the most salient region within the ROI. This method enables the system to determine automatically what is important in a specified region and to ignore the background. Additionally, it makes the system stable since usually the same MSR is computed, regardless of the exact coordinates of the rectangle. So the system is independent of user variations in determining the rectangle manually and it is not necessary to mark the target exactly.

Next, a *target descriptor* $\mathbf{v}$ is computed. It has one entry for each feature and each conspicuity map X_i. The values v_i indicate how important a map is for detecting the target and are computed as the ratio of the mean target saliency and the mean background saliency:

$$v_i = m_{i,(MSR)} / m_{i,(X_i - MSR)}, \quad i \in \{1, \ldots, 13\}, \tag{4.16}$$

where $m_{i,(MSR)}$ denotes the mean intensity value of the pixels in the MSR in map X_i, showing how strong this map contributes to the saliency of the region of interest, and $m_{i,(X_i - MSR)}$ is the mean of the remainder of the image in map X_i, showing how strong the feature is present in the surroundings.

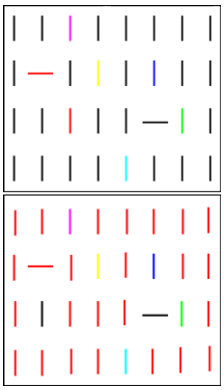
Figure 4.16 shows the target descriptor for a simple example. Values larger than 1 (green) are features that are relevant for the target while features smaller than 1 (red) are more present in the background and are used for inhibition.

Learning features of the target is important for visual search, but if these features also occur in the environment they might be of not much use. For example, if a red target is placed among red distractors it is not reasonable to consider color for visual search, although red might be the strongest feature of the target. In VOCUS, not only the target's features but also the features of the background are considered and used for inhibition. This method is supported by psychophysical experiments, showing that both excitation and inhibition of features are important in visual search. Figure 4.17 shows the effect of background information on the target descriptor.

Note that it is important that target objects are learned in their typical environment since otherwise their appearance with respect to the background cannot be represented adequately. Figure 4.18 shows some typical training images and the regions that the system determined to represent the target.

4.3.3.2 Several Training Images

Learning weights from one single training image yields good results if the target object occurs in all test images in a similar way, i.e., the background color is similar and the object always occurs in a similar orientation. These conditions often occur if the objects are fixed elements of the environment. For example, name plates or fire extinguishers within the same building are usually placed on the same kind of wall, so the background has always a similar color and intensity. Furthermore, since



Feature	target vector **v** (top)	target vector **v** (bottom)
intensity on/off	0.01	0.01
intensity off/on	9.13	13.17
orientation 0°	20.64	**29.84**
orientation 45°	1.65	1.96
orientation 90°	0.31	0.31
orientation 135°	1.65	1.96
color green	0.00	0.00
color blue	0.00	0.01
color red	**47.60**	10.29
color yellow	**36.25**	9.43
conspicuity I	4.83	6.12
conspicuity O	7.90	**11.31**
conspicuity C	**17.06**	2.44

Fig. 4.17 Effect of background information on the target vector. *Left*: the same target (*red horizontal bar*, 2nd in 2nd row) in different environments: most *vertical bars* are *black* (*top*) resp. *red* (*bottom*). *Right*: the target vectors (most important values printed in bold face). In the *upper image*, *red* is the most important feature. In the *lower image*, surrounded by red distractors, *red* is no longer the prime feature to detect the bar but orientation is (image from [5])

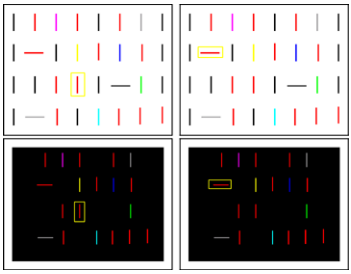
Fig. 4.18 *Top*: some training images with targets (name plate, fire extinguisher, key fob). *Bottom*: The part of the image that was marked for learning (region of interest (ROI)) and the contour of the region that was extracted for learning (most salient region (MSR)) (images from [5])

the object is fixed, its orientation does not vary and thus it makes sense to learn that fire extinguishers usually have a vertical orientation.

To automatically determine which object properties are general and to make the system robust against illumination and viewpoint changes, the target descriptor **v** can be computed from several training images by computing the average descriptor from n training images with the geometric mean:

$$v_i = \sqrt[n]{\prod_{j=1}^{n} v_{ij}}, \quad i \in \{1, \ldots, 13\}, \tag{4.17}$$

where v_{ij} is the ith feature in the jth training image. If one feature is present in the target region of some training images but absent in others, the average values will be close to 1 leading to only a low activation in the top-down map. Figure 4.19 shows



Feature	weights for red bar				
	v,b	h,b	v,d	h,d	average
int on/off	0.00	0.01	**8.34**	**9.71**	0.14
int off/on	**14.08**	**10.56**	0.01	0.04	0.42
ori 0°	1.53	**21.43**	0.49	**10.52**	3.61
ori 45°	2.66	1.89	1.99	2.10	2.14
ori 90°	**6.62**	0.36	**5.82**	0.32	1.45
ori 135°	2.66	1.89	1.99	2.10	2.14
col green	0.00	0.00	0.00	0.00	0.00
col blue	0.00	0.00	0.01	0.01	0.00
col red	**18.87**	**17.01**	**24.13**	**24.56**	**20.88**
col yellow	**16.95**	**14.87**	**21.21**	**21.66**	**18.45**
consp I	7.45	5.56	3.93	4.59	5.23
consp O	4.34	7.99	2.87	5.25	4.78
consp C	4.58	4.08	5.74	5.84	5.00

Fig. 4.19 Influence of averaging the target descriptor from several training images. *Left*: four training examples to learn *red bars* of horizontal and vertical orientation and on different backgrounds. The target is marked by the *yellow rectangle*. *Right*: The learned target descriptors. Column 2–5: the weights for a single training image (v = vertical, h = horizontal, b = bright background, d = dark background). The highest values are highlighted in bold face. Column 6: average vector. Color is the only stable feature (example from [5])

the effect of averaging target descriptors on the example of searching for red bars in different environments.

In practice, good results can be obtained by only two training images. In complicated image sets, up to four training images can be useful (see experiments in [5]). Since not each training image is equally useful, it can be preferable to select the training images automatically from a set of training images. An algorithm for this issue is described in [5].

4.3.3.3 Search Mode

In search mode, we search for a target by means of the previously learned target descriptor. The values are used to excite or inhibit the feature and conspicuity maps according to the search task. The weighted maps contribute to a top-down saliency map highlighting regions that are salient with respect to the target and inhibiting others. Figure 4.20 illustrates this procedure.

The excitation map E is the weighted sum of all feature and conspicuity maps X_i that are important for the target, namely the maps with weights greater than 1:

$$E = \sum_{i:v_i>1} (v_i * X_i). \tag{4.18}$$

The inhibition map I collects the maps in which the corresponding feature is less present in the target region than in the remainder of the image, namely the maps

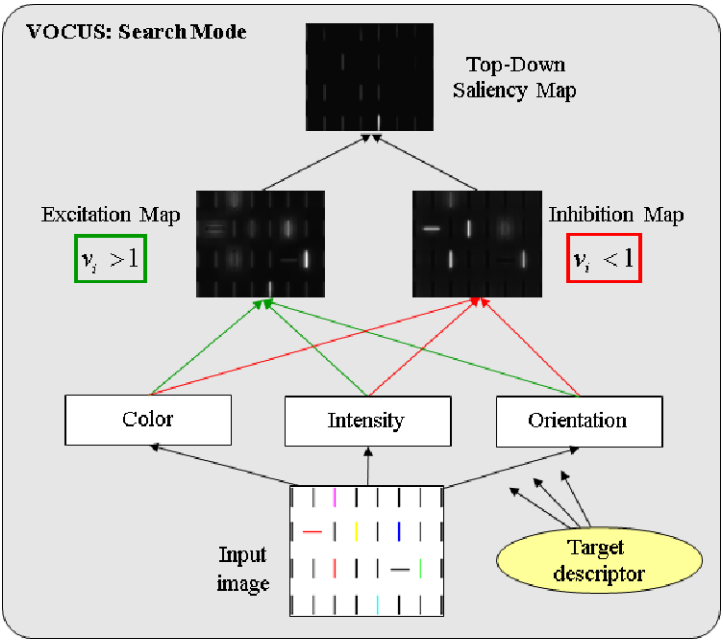


Fig. 4.20 Computation of the top-down saliency map S_{td} that results from an excitation map E and an inhibition map I. These maps result from the weighted sum of the feature and conspicuity maps, using the learned target descriptor

with weights smaller than 1:[8]

$$I = \sum_{i:v_i<1} \big((1/v_i) * X_i\big). \tag{4.19}$$

The excitation and inhibition map are not normalized to the same range since we want to preserve the differences among the maps.

The top-down map is obtained by subtracting the inhibition map from the excitation map:

$$S_{\text{td}} = E - I. \tag{4.20}$$

After subtraction, negative values are clipped to 0. Figure 4.20 shows that both, excitation and inhibition are important to find a target: when searching for the cyan vertical bar, the excitation map shows bright values for the cyan bar but the brightest region for the green bar. However, green contains also yellow which is inhibited for a cyan target. Thus in the resulting top-down map, only the cyan bar is salient.

[8] Entries with value 1 are ignored since they indicate that the mean saliency of the target region is exactly the same as the mean saliency of the surrounding; such a feature is completely useless for detecting the target. However, in practice this usually does not occur unless a feature is not present at all, e.g., color is not present in a gray-scale image and the color weights are set to 1.

If the task is pure visual search for a target, the top-down saliency map can be directly used to determine the focus of attention.[9] This is done as described in Sect. 4.3.2.8. However, if bottom-up cues shall be regarded additionally, the bottom-up and the top-down saliency maps have to be fused. This will be discussed in the next section.

4.3.3.4 Bottom-up and Top-down Cues Compete for Attention

In human perception, bottom-up and top-down cues compete for attention all the time. Depending on how engrossed in a task you are, the influences of bottom-up and top-down vary. The introductory city-visiting example illustrates this: without a clear task, the salient street performers attract your gaze. When you start to actively look for the train station, your top-down attention is focusing on street signs. Finally, the fire alarm is salient enough to override the task and captures your attention.

Consequently, it is important for a technical system to know what the overall tasks are, which one the most important task is at the moment, and how important it is. Depending on such pre-knowledge, the influence of bottom-up and top-down factors might be determined. After obtaining such a factor, the bottom-up and top-down saliency map are weighted accordingly and finally fused to a global saliency map S. To make the maps comparable, S_{td} is normalized in advance to the same range as S_{bu}:

$$S = (1 - t) * S_{\text{bu}} + t * N_{[0..M_S]} S_{\text{td}}, \quad M_S = maxvalue(S_{\text{bu}}). \tag{4.21}$$

Here, $t \in [0 \ldots 1]$ is the top-down factor that determines the amount of top-down influence. Determining t is not trivial. Probably the best solution is to learn it while performing some tasks on a real system, but this is beyond the scope of this article. Note that a simple solution for a technical system is not to fuse bottom-up and top-down saliency, but to process them independently. Bottom-up salient regions might be fed to an object recognition module that recognizes the objects, building a semantic map of the environment with object annotations, and successively improving the background knowledge of the system, while top-down cues can be used to solve the current task by searching for desired objects.

4.4 Evaluation of Computational Attention Systems

The evaluation of computational attention systems can be done from a psychophysical perspective, e.g. by comparing their results with human perception, or from a technical perspective, e.g. by measuring the success in an application.

[9] Note that in human perception, bottom-up cues always play a role and thus should be considered if similarity to human perception is desired.

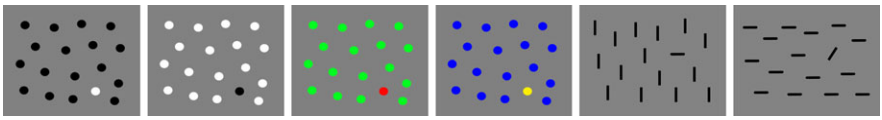


Fig. 4.21 Typical pop-out images. Attention systems should be able to detect the outliers

When considering bottom-up systems of attention, the first step is to determine whether the system is able to detect pop-outs in the dimension of the implemented features. These tests are important to ensure the basic capabilities of the system and are suitable to reveal its strengths and limitations. Thus, a system with the standard features intensity, color, and orientation should be able to detect popouts as the ones in Fig. 4.21. Hereby, the saliency of the target depends on the similarity to the distractors, the more it differs, the higher the saliency. Thus, a target that differs only slightly from the distractors might not be detected with the first fixation. This is in accordance with the psychophysical findings that the more similar target and distractors are, the slower the visual search (cf. Sect. 4.2.2)

The evaluation on artificial patterns is only the first step, testing on natural images is important too. Here, it is usually less clear which region shall be salient, top-down influences play a larger role and saliency depends stronger on the context and of preknowledge of the observer. A possibility for evaluation is to compare the output of the system with human eye movement data (see also Sect. 4.6 and Chap. 11, Sect. 11.3.2.2). Note that a computational attention system can only roughly approximate such eye movement trajectories since the top-down cues that influence human perception are hardly possible to model in such a general scenario and thus the systems usually operate in bottom-up mode. It is, however, possible to compare different attention systems based on such data.

An alternative that is recently introduced in the computer vision community is the evaluation on image databases with salient objects, manually labeled by different users [15]. Note, however, that the database in [15] contains many close-up views of objects that cover a large portion of the image, a case for which the human attention system is not designed. In contrast, the task of human attention is to direct the gaze to a small region in a complex scene which is afterwards investigated in detail. Thus, a system as the one described here is designed to operate on scene images rather than on close-up views of objects and might have to be adapted accordingly to work on the above database. A similar approach for evaluation was used by Elazary and Itti, who used 24 836 pictures of natural scenes from the LabelMe database, in which objects were manually marked and labeled by a large population of users. They found that the hot spots in the saliency map predict the locations of objects significantly above chance [4].

From a technical point-of-view it is not necessarily important that a computational attention system operates similar to human perception, as long as the outcome is useful for an application. Two applications in which attention system are applied are mentioned in Sect. 4.5. But even in these cases, a system should be able to detect outliers as in Fig. 4.21, since this belongs to the basic capabilities of visual attention systems.

Target	# test im.	av. hit number [%]
Fire extinguisher	46	1.09
Key fob	30	1.23

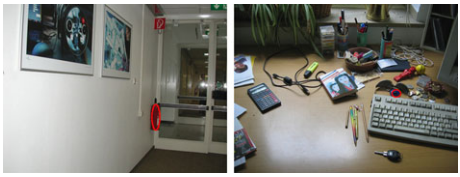


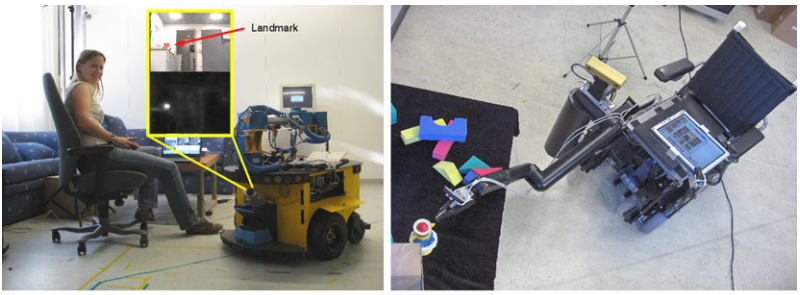
Fig. 4.22 *Top*: Average hit number of VOCUS for two targets on a set of test images. The target descriptors were computed from two training images each (examples of training images cf. Fig. 4.18). *Bottom*: Two example test images with foci of attention (*red ellipses*) (example from [5])

The evaluation of top-down systems is easier. Here, the task is clearly specified and it can be determined easily whether a target was detected, or not. Note that a top-down attention system is no object recognizer, that means it cannot decide whether an object is present in an image or not. It can simply determine locations that are likely to contain the target, usually in form of a trajectory of locations. Thus, instead of determining a detection rate, it is more reasonable to determine the *hit number*, i.e. the rank of the first FOA that is on the target. A hit number of 1 is best and means that the first focus of attention was on the target. An example of the evaluation of visual search with VOCUS is displayed in Fig. 4.22.

4.5 Applications in Computer Vision and Robotics

In the introduction, we have pointed out to the importance of attentional selection for tasks that deal with large amounts of image data. Especially in the field of autonomous mobile robots, the concept of visual attention has increasingly gained interest during the last decade. A large number of EU projects on cognitive robotics has been launched, e.g. the projects MACS, CogVis, POP, and SEARISE. In many of these projects, visual attention has been used as perception module.

We will concentrate here on two applications of visual attention systems. A broader overview can be found in [7]. The first application that we will introduce is visual robot localization. Here, a robot has to determine its position in the world by interpreting its sensor data. When a camera is used as sensor, this is usually done by detecting visual landmarks in the environment and computing the robot position based on the position estimation of the landmarks. An important property of landmarks is the re-detectability in frames that are taken from different viewpoints. Using salient regions as landmarks is a natural way of exploiting the fact that salient regions are "special" in an environment and thus, easy to re-detect. An example of a typical salient landmark is a fire extinguisher. As part of the EU project NEUROBOTICS, we have used salient visual landmarks for simultaneous localization and mapping (SLAM) [6]. This task is more difficult than pure localization since the robot initially does not know anything about its environment and has to build a map and localize itself inside the map at the same time. We have detected salient regions with VOCUS, tracked them over several frames to determine the most stable



(a) Visual SLAM (b) Object manipulation in PlayBot

Fig. 4.23 Two application scenarios for visual attention systems: (**a**) attentional landmarks for visual SLAM (simultaneous localization and mapping) at the Royal Institute of Technology (KTH) in Stockholm: robot Dumbo corrects its position estimate by re-detecting a salient landmark based on the attention system VOCUS. The *yellow rectangle* shows the currently seen frame with a landmark (*top*) and the corresponding saliency map (*bottom*) [6] (Fig. from http://www.iai.uni-bonn.de/~frintrop/research.html). (**b**) PlayBot: a visually guided robotic wheelchair for disabled children. The selective tuning model of visual attention supports the detection of objects of interest (Fig. from http://www.cse.yorku.ca/~playbot)

ones and to determine their 3D position, and stored them as landmarks in a database. At every time step, currently seen salient regions are compared with landmarks from the database to enable the robot to detect that it has returned to a previously visited location (loop closing). This is an especially important step in SLAM to correct accumulated position errors. A picture of the process is displayed in Fig. 4.23(a).

Another application is the PlayBot project, lead by Prof. John K. Tsotsos from York University, Canada [18].[10] The goal of the project is to develop a smart wheelchair to support disabled children. The wheelchair has an easily accessible user interface display, which shows pictures of places and toys. Once a task like "go to table, point to toy" is selected, the system drives to the selected location and searches for the specified toy, using mechanisms based on visual attention (see Fig. 4.23(b)).

4.6 Open Source Code, Databases, and Further Reading

This section lists some tools and references for the interested reader.

4.6.1 Open Source Code

- The iLab Neuromorphic Vision C++ Toolkit (iNVT, pronounced "invent") from the group of Laurent Itti is probably the best known and most distributed Open

[10] More on http://web.me.com/john.tsotsos/Applications/Playbot.html.

Source code for computational attention systems [11]. It includes the surprise model for temporal saliency [9] and is available at http://ilab.usc.edu/toolkit/.
- The SaliencyToolbox from Dirk B. Walther [26] is a more compact re-implementation of iNVT in MATLAB: http://www.saliencytoolbox.net/.
- The original VOCUS source code is not freely available, but a re-implementation of the bottom-up part (in C++) can be found http://sourceforge.net/projects/openvolksbot/.
- The AIM model (Attention based on Information Maximation) is an attention system based on information theory. It determines the self-information of a center region with respect to a global surround [1]. MATLAB code is available at: http://www-sop.inria.fr/members/Neil.Bruce.
- For implementing your own attention system, it is convenient to use the Open Source Computer Vision Library OpenCV that contains many basic techniques, from displaying images over computing pyramids to converting images to other color spaces: http://sourceforge.net/projects/opencvlibrary.

4.6.2 Databases

Several databases are available for testing and evaluating visual attention systems:

- Image databases of popout search arrays and natural images can be found on the websites of the iLab: http://ilab.usc.edu/imgdbs/.
- Eye tracking data from 20 test persons on 120 still images can be found on: http://www-sop.inria.fr/members/Neil.Bruce/.
- Eye-tracking data from human volunteers watching complex video stimuli are available from the CRCNS (Collaborative Research in Computational Neuroscience) data sharing website: http://crcns.org/data-sets/eye.
- The MSRA Salient Object Database contains 25.000 images with manually labeled salient objects: http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm. For a subset of 1.000 images, binary maps of the salient objects are available as ground truth: http://ivrg.epfl.ch/supplementary_material/RK_CVPR09.

4.6.3 Further Reading

More about the human visual system can be found in the books of Palmer [16] or Kandel et al. [13]. The psychology of attention and details on many psychological attention models are described in a book by Pashler [17] and in the chapter "Attention" by Bundesen & Habekost in the Handbook of Cognition [2]. A description of the social aspects of attention can be found later in this book in Chap. 8, Sect. 8.6.4.1. Wolfe has written a comprehensive article that contains everything you ever wanted to know about visual search [28]. One of the first computational models

of visual attention was introduced by Koch and Ullman in 1985 with a detailed description of the Winner-Take-All approach [14]. The basic paper that describes the widely used computational attention model by the group of Laurent Itti in a comprehensive manner is [11]. Recently, several groups have used information-theoretic approaches to determine visual saliency [1, 8, 9]. The latter also tackle the aspect of top-down saliency for object recognition by determining salient features that best distinguish a visual class from other classes [8]. Top-down information in the form of knowledge about the scene and its visual layout was used by Torralba et al. to guide visual attention to relevant parts of an image [19]. Tsotsos has recently published a book that gives an overview of attention theories and models and offers a full description of his attention model, called selective tuning model [24]. A survey on computational attention systems that aims to bridge the gap between the research on human and computational visual attention can be found in [7].

Research papers on computational attention appear on conferences and in journals of many different areas, e.g. cognitive perception, computer vision, and cognitive robotics. Important journals for cognitive aspects of attention are "Attention, Perception, and Psychophysics" and the "Journal of Vision". In the technical fields, much work can be found on workshops on cognitive systems that usually take place as satellites of big conferences, such as the "International Symposium on Attention in Cognitive Systems" at IJCAI 2011. Journal articles appear e.g. in "Computer Vision and Image Understanding" and in the "IEEE Transactions on Pattern Analysis and Machine Intelligence", or, if related to robotics, in the "IEEE Transactions on Robotics" and the "Robotics and Autonomous Systems".

4.7 Summary

Computational attention systems are inspired by human perception and aim to detect the most promising regions in images. While computational attention systems already do a good job in bottom-up saliency computation, many open questions remain in the field of top-down attention. All kinds of background knowledge about the context, the current situation, the layout of the scene, and the specification of the current task influence the visual processing in humans and should therefore also be integrated into a technical system. The more technical systems advance, the more urgent the need for preprocessing modules such as attention systems that prioritize the data and enable efficient processing with limited resources. Especially in the field of autonomous robots such a mechanism is important to facilitate the decision which actions to perform next.

4.8 Questions

1. Which objects of the following list are likely to be detected with a bottom-up attention system and which are not: a traffic sign, a glass, a large object among small ones, an apple on the table, an apple in a box full of apples?

2. You notice that the attention system detects very small salient regions on your test images. How could you adapt the attention system to detect larger objects as well? What could you do if you do not have access to the source code and you can only adapt the input image itself?
3. Why is the arithmetic mean not an adequate alternative for (4.17)? Tip: consider two training images with $v_i = 0.5$ and $v_i = 2$, respectively, for feature map i. Which value would you expect and what do you get by arithmetic/geometric mean?
4. What happens if you search for a target object with the top-down attention system in an image where the target is not present?
5. How does an attention system differ from a standard interest point detector such as the Difference of Gaussian detector or the Harris corner detector?
6. How does a top-down attention system differ from an object recognition module?

4.9 Glossary

- *Bottom-up attention*: One of the factors that guide human attention (the other is top-down attention). Bottom-up attention is purely data-driven and guides the gaze to salient regions in a scene. Indicators that attract bottom-up attention are strong contrasts and the uniqueness of a region.
- *Center-surround filters*: The main concept in visual attention systems to detect contrasts. They are inspired by on-center and off-center cells of the human visual system.
- *Saliency*: The quality of a region to stand out relative to its surround.
- *Top-down attention*: One of the factors that guide human attention (the other is bottom-up attention). Top-down attention is driven by cognitive factors such as pre-knowledge, context, expectations, motivations, and current goals. One of the best investigated areas of top-down attention is visual search.
- *Visual search*: The task to find an item in a scene. It is one of the best investigated parts of top-down attention. Visual search experiments are used frequently in cognitive sciences to investigate the human visual system.

References

1. Bruce, N.D.B., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. J. Vis. **9**(3), 1–24 (2009)
2. Bundesen, C., Habekost, T.: Attention. In: Lamberts, K., Goldstone, R. (eds.) Handbook of Cognition. Sage, London (2005)
3. Douma, M.: Color Vision and Art. Retrieved Nov 2010 from http://webexhibits.org/colorart/ganglion.html (2008)
4. Elazary, L., Itti, L.: Interesting objects are visually salient. J. Vis. **8**(3), 3 (2008)
5. Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. Lecture Notes in Artificial Intelligence (LNAI), vol. 3899. Springer, Berlin/Heidelberg (2006)

6. Frintrop, S., Jensfelt, P.: Attentional landmarks and active gaze control for visual SLAM. IEEE Trans. Robot. **24**(5) (2008). Special Issue on Visual SLAM
7. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. ACM Trans. Appl. Percept. **7**(1) (2010)
8. Gao, D., Han, S., Vasconcelos, N.: Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(6) (2009)
9. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vis. Res. **49**(10), 1295–1306 (2009)
10. Itti, L., Koch, C.: Feature combination strategies for saliency-based visual attention systems. J. Electron. Imaging **10**(1), 161–169 (2001)
11. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
12. James, W.: The Principles of Psychology. Dover, New York (1890)
13. Kandel, E.R., Schwartz, J.H., Jessell, T.M.: Essentials of Neural Science and Behavior. McGraw-Hill/Appleton & Lange, New York (1996)
14. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. **4**(4), 219–227 (1985)
15. Liu, T., Zejian, Y., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.-Y.: Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 353–367 (2009)
16. Palmer, S.E.: Vision Science: Photons to Phenomenology. MIT Press, Cambridge (1999)
17. Pashler, H.: The Psychology of Attention. MIT Press, Cambridge (1997)
18. Rotenstein, A., Andreopoulos, A., Fazl, E., Jacob, D., Robinson, M., Shubina, K., Zhu, Y., Tsotsos, J.K.: Towards the dream of intelligent, visually-guided wheelchairs. In: Proc. 2nd Int'l Conf. on Technology and Aging, Toronto, Canada, June 2007
19. Torralba, A., Oliva, A., Castelhano, M., Henderson, J.: Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. Psychol. Rev. **113**(4) (2006)
20. Treisman, A.: Preattentive processing in vision. Comput. Vis. Graph. Image Process. **31**, 156–177 (1985)
21. Treisman, A.M., Gelade, G.: A feature integration theory of attention. Cogn. Psychol. **12**, 97–136 (1980)
22. Treisman, A.M., Gormican, S.: Feature analysis in early vision: Evidence from search asymmetries. Psychol. Rev. **95**(1), 15–48 (1988)
23. Tsotsos, J.K.: A 'complexity level' analysis of vision. In: Proc. of International Conference on Computer Vision: Human and Machine Vision Workshop, London, England, June 1987
24. Tsotsos, J.K.: A Computational Perspective on Visual Attention. MIT Press, Cambridge (2011)
25. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
26. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks (2006)
27. Wolfe, J.M.: Guided search 2.0: A revised model of visual search. Psychon. Bull. Rev. **1**(2), 202–238 (1994)
28. Wolfe, J.M.: Visual search. In: Pashler, H. (ed.) Attention, pp. 13–74. Psychology Press, Hove (1998)
29. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? Nat. Rev., Neurosci. **5**, 1–7 (2004)
30. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime $TV - L^1$ optical flow. In: Proc. of the Annual Meeting of the German Assoc. for Pattern Recognition (DAGM) (2007)

Part II
Analysis of Activities

Chapter 5
Methods and Technologies for Gait Analysis

Elif Surer and Alper Kose

5.1 Introduction

Human gait is a biometric that can be used for identification of a person or for diagnostic and clinical purposes. Subsequently, gait analysis is an important assessment tool that uses physical measurements and models, including the movement of the person's centre of mass, joint kinematics, ground-reaction forces, the resultant loads, body segment energy variation and muscular work [1]. This chapter will introduce gait analysis in a clinical context. Depending on the application, the body is represented with 2-D or 3-D models. These can be pose matrices, point sets, lines or more complex models.

Low-level feature extraction, segmentation, and joint detection, and constructing 3-D structure from 2-D are generally parts of human movement analysis [2]. In this chapter, we will distinguish between marker-based and markerless techniques. In model-based markerless analysis, a model is fit to the appearance and the body is tracked through its motions. In marker-based gait analysis, the characteristic steps are placing of markers, sensing, and constructing three-dimensional trajectories from the markers, which are subsequently labelled. In order to compute joint angles from the relative marker positions of the labelled trajectories, a computer model is used [3].

This chapter gives general information on gait analysis, with the overall theory and frequently used applications. Section 5.2 summarizes how motion is measured, via marker-based or markerless motion capture methods, and inertial measurements. In Sect. 5.3 we briefly discuss force platforms and electromyography.

E. Surer (✉) · A. Kose
Department of Biomedical Sciences, University of Sassari, Sassari, Italy
e-mail: esurer@uniss.it

A. Kose
e-mail: akose@uniss.it

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_5,

5.2 Motion Measurements

The main goal of the human movement analysis is the acquisition of quantitative information about the mechanics of the musculoskeletal system while executing a motor task [1]. In order to pursue this goal, motion capture and inertial measurements are frequently used.

Motion capture acquires the data of a moving human via sensors and processes the acquired data by using a mathematical model. In motion capture, either conventional photography or optoelectronic systems are used for the acquisition of quantitative information, whereas in inertial measurements, accelerometers, gyroscopes and magnetometers are utilized in order to measure acceleration, angular velocity and magnetic field, respectively.

Brief explanations and applications of motion capture and inertial measurements are presented in the following subsections.

5.2.1 Marker-Based Motion Capture

In general, motion capture (also known as 'mocap') is classified into two: (1) marker-based techniques, (2) markerless techniques. In marker-based techniques, video-based optoelectronic systems are used and retro-reflective markers are attached on the human body. Mocap suits are also considered to be examples of the first approach. These are worn on the body and are fully equipped with sensors. In this section we will describe optoelectronic systems for marker-based mocap.

Marker-based analysis is generally performed by mounting retro-reflective markers on the subjects' bodies and reconstructing their 3-D position via video-based optoelectronic systems (Fig. 5.1). Retro-reflective markers are used together with infrared stroboscopic illumination produced by an array of light-emitting diodes (LEDs) mounted around the lens of each camera. The thresholds of the cameras can be adjusted so that only bright reflective markers are sampled and the markers are recognized via image-based methods.

If the marker is visible from at least two calibrated cameras at the same time, the 3-D position of a marker in a reference frame fixed to the laboratory (global frame—GF) can be reconstructed. Additional reference frames, linked to body segments (technical frame—TF), can be defined from the position of a GF-relative cluster of markers attached to the same body segments. Then, the pose of the TFs in the GF can be calculated. Bones are the main anchors of TFs. Even though TFs are assumed to be fixed to the underlying bone, they do not depict the anatomical properties of the body segment they are attached to. Hence, for each body segment under analysis, an additional frame—i.e. the anatomical frame (AF)—is defined.

In order to define AFs, selected anatomical landmarks (ALs) with respect to the relevant TF [4] are determined. Orientation and position in space of a body segment is the pose of an AF. By using the pose of the AFs of two adjacent body segments, the kinematics of the joint between the two body segments can be calculated.

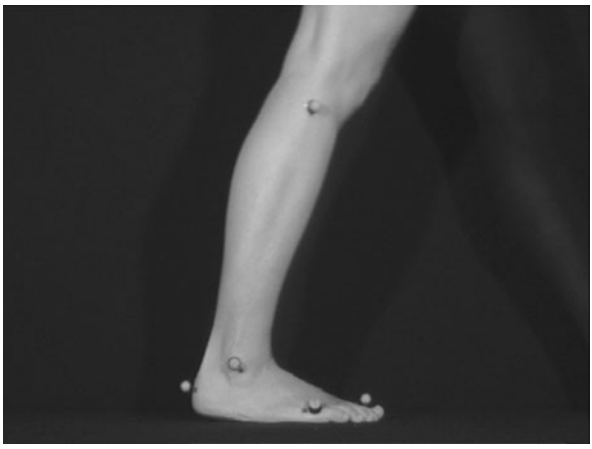


Fig. 5.1 Markers are mounted on the shank and foot complex of the subject for the analysis

5.2.1.1 Calibration of Anatomical Landmarks

Anatomical landmarks are either bony prominences or bone points of geometrical relevance, which are normally identified by examining with the hand (palpation), but they can also be identified by imaging, regression equations, or functional movements [1]. Once the ALs are identified, their location with respect to the relevant TF has to be calculated. Once calculated, reconstructing their position in the GF by simple coordinate transformations is possible. The Calibrated Anatomical System Technique (CAST) is an experimental method that uses the concept of AL calibration and allows various calibration methods to be implemented.

The calibration of AL can be performed (a) by using a marker positioned on the AL during a static acquisition, (b) by using a pointer, where a minimum of two markers are mounted with a known distance from tip, and the pointer pointing at the AL during a static acquisition, (c) by determining the centre of rotation of recorded functional movements (for joint centres, such as the hip centre), (d) by imaging of the bone and the relevant TF [5, 6].

The CAST method was recently updated by adding information on the subject-specific bone geometry. After the position of unlabelled points (UPs) located over the bone surface are determined, an initial estimation is performed. The estimation step is followed by matching a digital template-bone to the initial estimate. The updated technique is called UP-CAST and it is evaluated in terms of repeatability and accuracy [7].

5.2.1.2 Protocols

Human movement analysis makes use of the theory of multi-rigid body systems, where a number of rigid segments and adjacent segments connected by joints are used to model the human body. Certain protocols—data collection and reduction practices—have been designed in gait analysis, offering various ways of modelling

the system of rigid bodies of interest. Adopting a protocol helps in obtaining reproducible results and guides the researcher or clinician in practice. In clinical gait analysis, all model joints are rotational (either cylindrical or spherical) and AFs are defined based on this assumption.

Proposed protocols use different marker-sets to identify AFs and joint centre locations. Data acquired with different protocols can usually not be compared.

"Newington model" is the pioneering and the most commonly used protocol for gait data acquisition and reduction. Commercial applications like Plug-in Gait (PiG—Vicon Motion Systems, Oxford, UK) also use this protocol. "Servizio di Analisi della Funzione Locomotoria" developed a protocol named "SAFLo"—which differs from the Newington model in terms of segmental anatomical references and anatomical marker configurations. After that, "Calibration Anatomical System Technique" (CAST) was introduced to standardize and define references, internal anatomical landmarks and external technical markers. Later, protocols of "Laboratorio per l'Analisi del Movimento nel Bambino" (LAMB) and "Istituti Ortopedici Rizzoli Gait" were developed, of which the latter was used as the basis of the software "Total 3-D Gait" (T3Dg-Aurion s.r.l., Milan, Italy) [8, 9].

Even though there are known significant differences among these techniques, reasonable correlations are observed for most of the gait variables. Ferrari et al. compared these commonly used protocols and found out that there was good intra-protocol repeatability within the same gait cycles [8]. It is depicted that model conventions and definitions seem to be more important than the design of the relevant marker-sets. Sharing the model conventions and definitions can be adequate for worldwide data comparison in clinical gait analysis.

5.2.1.3 Errors

The estimation of 3-D points of objects from two or more images is called 'stereophotogrammetry'. There are three major sources of errors in human movement analysis performed with stereophotogrammetry.

– Instrumental errors: these errors stem from the results of both instrumental noise and volume calibration inaccuracies. They have been widely studied in the 80s and 90s [10, 11], tests for estimating them have been proposed [12]. The instrumental noise can be reduced by low pass filtering, while the volume calibration inaccuracies depend on the inadequate number of cameras and the volume calibration algorithm chosen for the application.

Direct linear transformation (DLT) algorithm [13] is mostly used to register multiple images by solving for a set of similarity relations under projective geometry, using commonly identified landmark points on each image. When the volume of interest is large, performing DLT becomes restrictive. Simultaneous multi-frame analytical calibration (SMAC) [14]—a technique based on a planar calibration object with a grid of known control points—requires the recording of the calibration object by at least two convergent cameras. SMAC allows covering larger volumes

but still for very large volumes, analytical self-calibration is more appropriate [15]. Thus, volume calibration inaccuracies can be lowered with the number of cameras and the chosen volume calibration algorithms.

The contribution of instrumental errors to the total error is considered to be very small, almost negligible.

- Soft tissue artefacts: the markers captured by the cameras can be directly attached to the skin or arranged in clusters and positioned with fixtures over a body segment and their movements cause errors. Since this error has the same frequency content as the bone movement, there is no way of distinguishing the artefact from the actual bone movement by using a filter. However, it is possible to reduce its effect on the end results in the following ways. First of all, marker locations (marker points) must be chosen so that the relative displacement is minimized. Secondly, mathematical operators can be used to estimate position and orientation of the bone from skin marker positions [16, 17]. Knowledge regarding the characteristics of the artefact movement in different body segments is necessary to manage the mentioned countermeasures against experimental artefacts.
- Anatomical landmark misplacement: The incorrect location of subcutaneous bony ALs through palpation can stem from three main factors: (1) the palpable ALs are not points but surfaces, large and irregular; (2) a soft tissue layer of variable thickness and composition covers the ALs; (3) the identification of the location of the ALs depends on which palpation procedure was used. Studies show that AL position uncertainty and the erroneous determination of AF axes may result in wrong clinical interpretations of the estimation [18].

Besides the above mentioned sources of errors, marker-based movement analysis is influenced by the markers attached to the body while the subject was moving and the need for an extended setup time for marker placement [19].

5.2.2 Markerless Motion Capture

In biomechanics, the main focus of gait analysis has been to build body models which explain the functioning of the body and to provide solutions to improve the body's movement efficiency. Obtaining joint data, analyzing the kinematics and kinetics of the movement of interest have been the common procedure [20]. As mentioned in the previous section, the most popular technique for acquiring joint data is to use markers placed on the skin, despite some drawbacks such as interference with walking and skin movement artefacts. Especially the latter problem is an important challenge for biomechanical and clinical applications, since it makes the evaluation of the underlying bone very difficult and error-prone. Also, the use of markers is intrusive, requires special hardware and cannot be used in most settings.

To overcome these limitations, markerless systems of human capture are proposed where conventional cameras can be utilized without the necessity of using special apparel or hardware [21]. Markerless motion capture ensures an important

reduction of the amount of time for setup preparation in comparison to marker-based techniques. Besides, inter-operator variability is eliminated since no specialized operator is needed to place markers on the skin [19].

Markerless techniques can be classified into model-based and model-free techniques. Model-based approaches utilize an a priori human body model and are composed of two stages: modelling and estimation.

Modelling is the building stage of a likelihood function by taking all factors into account. The likelihood function is used to determine the most plausible body model parameterization, given a set of image descriptors, in the context of the camera model and with respect to a certain matching function. The estimation stage is fitting the optimum pose in the likelihood domain designed in the modelling stage. Model-free approaches do not use an a priori human body model, but implicitly model variations in pose configuration, body shape, camera viewpoint and appearance [22].

Much of the work on motion analysis uses 3-D kinematic models and detailed estimation of 3-D motion. These techniques require multiple camera viewpoints or 3-D sensors, but motion analysis can also be operated using a single-camera input without recourse to 3-D motion [23].

Motion capture from a single camera is a difficult task; data acquisition is very simple, but derivation of a motion is a computer vision challenge that focusses on interference as much as movement [24].

5.2.2.1 Cardboard Models

A significant example of 2-D model-based markerless techniques is the work of Howe et al. [24], in which 3-D motion from single camera is reconstructed using a learning-based approach based on the information learned from a labelled training set. In a Cardboard model, body parts are modelled as rectangular patches—with a total of 34 parameters—connected to each other. First, these parameters are initialized in the first frame by hand, by overlaying a model onto the 2-D image of the first frame, and then the joints and body parts are tracked in 2-D video. This tracking process returns the coordinates of each limb for each frame which are combined with a prior model of the human motion to estimate the body's motion in 3-D. In each frame, the algorithm infers the correct depth of each point by using the 2-D positions of the tracked body points. To do so, a training set of 3-D human motion examples is used and a Bayesian framework is adapted to compute prior probabilities of different 3-D motions. The results show that it is possible to obtain good results (as measured by the distance to ground truth) with this method.

Ju et al. [25] use a cardboard model to define the human body as a set of connected planar patches and to approximate the limbs as planar regions (Fig. 5.2). The main assumption behind this model is that the motions of the limb planes are the same at the points of articulation.

The image motion of a rigid planar patch of the scene is described by the following eight-parameter model:

$$u(x, y) = a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy, \tag{5.1}$$

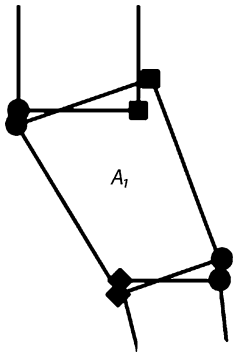


Fig. 5.2 A "chain" structure of limbs, similar to the chain structure used in [25]

$$v(x, y) = a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2, \tag{5.2}$$

where $\mathbf{a} = [a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7]$ is the vector of parameters to be estimated, and $\mathbf{u}(\mathbf{x}, \mathbf{a}) = [u(x, y), v(x, y)]^T$ are the components of the optical flow at image point $\mathbf{x} = (x, y)$. The coordinates (x, y) are defined in reference to a particular point, which can be the center of a patch or a point of articulation.

To simplify the optical flow equation, brightness is assumed to be constant for a given patch. By solving the optical flow equation and minimizing the total energy, the motions of all the patches (a_s) are estimated, where s denotes patch index.

After the estimation of the absolute motions, articulated motions need to be estimated. To do so, the motions of limbs relative to their preceding (parent) patches are recovered with the following formula:

$$\mathbf{u}(\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1}), \mathbf{a}_s^r) = \mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1}), \tag{5.3}$$

where $\mathbf{a}_s^r$ is the relative motion of patch s, $\mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1})$ is the relative displacement at point $\mathbf{x}$, and $\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}_{s-1})$ is the new location of point $\mathbf{x}$ under motion $\mathbf{a}_{s-1}$. A planar motion has eight parameters, so four different points of patch s are enough in order to solve for $\mathbf{a}_s^r$ given in (5.3). Then, two corners of each patch—which are the articulated points with the next patch—are tracked by using the estimation of articulated motions. For each frame, using the articulation motion estimation, the location of each patch is found and updated within a "chain" approach [25].

5.2.2.2 Tracking

The main principle of a probabilistic framework for tracking is maintaining a time-evolving probability distribution of the tracker state. Cham and Rehg applied a probabilistic multiple-hypothesis framework to figure tracking, which is the first application of this formulation in movement analysis [23]. To generate a mode-based representation for the probability distribution of the tracker state, the algorithm has

to recover these modes in each time-frame. The proposed algorithm here may be modularized in a way compatible with Bayes' Rule:

$$p(x_t|Z_t) = kp(z_t|x_t)p(x_t|Z_{t-1}),$$

where x_t is the tracker state at time t, z_t is the observed data point, Z_t is the collection of past image observations (i.e. z_τ for $\tau = 0, \ldots, t$), and k is a normalization constant. Furthermore, z_t is assumed to be conditionally independent of Z_{t-1} given x_t.

The stages of the algorithm at each time-frame are "

1. Generating the new prior density $p(x_t|Z_{t-1})$ by passing the modes of $p(x_{t-1}|Z_{t-1})$ through a Kalman filter prediction step.
2. Computing the likelihood by:
   (a) Creating initial hypothesis seeds by sampling the distribution of $p(x_t|Z_{t-1})$.
   (b) Refining the hypotheses through differential state-space search to obtain the modes of the likelihood $p(z_t|x_t)$.
   (c) Measuring the local statistics associated with each likelihood mode using perturbation analysis.
3. Computing the posterior density $p(x_t|Z_t)$ via Bayes' Rule (1), then updating and selecting the set of modes."

Then, Scaled Prismatic Models (SPM)—a class of 2-D kinematic models—are used to model the human body. These models impose 2-D constraints on the figure motion with an underlying 3-D kinematic model. Each link in the model corresponds to the image form of a rigid segment with a 3-D kinematic chain. This correspondence is very flexible, so that the changes of a joint in the underlying 3-D model can easily be represented by translations and rotations. The tracking problem starts with an initial state and consists of estimating a vector of SPM parameters for the figure in each frame of a video sequence [23].

SPM is an example of an explicit human body model. When no such model is available, a direct correspondence between image observation and pose must be established. Also, when the tracker loses the tracked person, it needs to be re-initialized. Model-free algorithms do not suffer from (re)initialization problems and can be used for initialization of model-based pose estimation approaches [22].

Mori and Malik estimate body pose and configuration in 3-D space by locating the joint points in a single 2-D image containing a human figure [26]. First, a number of exemplar views of the human body in different configurations and viewpoints with respect to the camera, are stored. Each of the stored views are manually marked at the body joints and labelled for future use. Then, the input figure is matched to each stored view using a shape context matching method with a kinematic chain-based deformation model. By extracting external and internal contours of an object, shape contexts are employed to encode the edges.

5.2.2.3 Matching

The problem of estimating the keypoints in the test image, while an exemplar (with labelled keypoints) is given, is cast as one of deformable matching. The exemplar is deformed into the shape of the test image and during the deformation, a matching score is computed in order to measure similarity between the exemplar and the test image.

In the study [26], a shape is represented by a discrete set of n points $P = \{p_1, \ldots, p_n\}$, $p_i \in \Re^2$ from the internal and external contours on the shape.

The deformable matching process contains three steps. Given sample points on the exemplar and test image:

1. Find correspondences between sample points of exemplar and test image.
2. Calculate the deformation of exemplar.
3. Apply deformation to exemplar sample points.

In order to estimate a deformation, the source of the sample point—i.e. which kinematic chain segment each sample point belongs—should be determined. For this purpose, hand-labelled keypoints, which automatically assign hundreds of sample points to segments, are used. This is done by finding minimum distance of each sample point to bone-line, which is the line segment connecting the keypoints at the segment ends, for arm and leg segments. After the correspondence step is over, the locations of the body joints are transferred to the test shape. Given the 2-D joint locations, the 3-D body configuration and pose are estimated using an algorithm proposed by Taylor [27] which uses point correspondences in a single image. In the estimation step, the stored exemplars are deformed in order to match the image observation. The most likely 2-D joint estimate is found by enforcing 2-D image distance consistency between body parts. This technique can be applied to each frame of a video sequence so that tracking recognition becomes repeatable for every frame [26].

In another 2-D model-free markerless application, Elgammal and Lee derive 3-D poses directly from human silhouettes extracted from a single camera [28]. The objective is to recover the intrinsic body configuration, viewpoint and reconstruct the silhouette and detect and discard outliers from the visual input. To recover intrinsic body configurations from the silhouette, manifolds are learned from the visual input and subsequently mappings are learned from manifolds to visual input and 3-D poses. The experiments demonstrate that the model can be learned from the data of one person and can easily be generalized for recovering the body configurations of other people. Thus, the interpolation of 3-D poses is possible even if they are not part of the training data [28].

5.2.2.4 3-D Model-Based Approaches

The 3-D motion of humans is not determined thoroughly when the observation is limited to a single camera. To overcome this limitation, 3-D markerless techniques

Fig. 5.3 A body model similar to the model used in [29]

are developed. An important study exemplifying 3-D model-based markerless technique is the study by Bottino and Laurentini [29]. They present a technique in order to reconstruct unconstrained motion from 3-D reconstruction of multiple-view images. First, views of the human body are recorded with different cameras and 2-D silhouette of the human body is extracted from each of the images. Then, a volumetric description is recovered by intersecting the cones derived from the corresponding silhouette. This step is called the volume intersection and gives the final voxel (volumetric pixel) representation. Finally, a model of the human body is fitted to the extracted volume (Fig. 5.3). Model fitting is done by minimizing a distance function between the volume and the model with a search through the 32 dimensional space of pose parameters.

Deutscher and Reid propose a generic tracking technique which does not necessitate special preparations of subjects or restrictive assumptions [21]. It is possible to search high dimensional configurations by using a modified particle filter without any assumptions. The idea of annealing is adapted to perform a particle based stochastic search. The adapted algorithm is called annealed particle filtering and is capable of recovering full articulated body motion. The articulated model of the human body is built around a kinematic chain, where each limb is fleshed out using conic sections with elliptical cross-sections. This model presents computational simplicity and compact representation. The results show that this new technique leads to robust tracking even in complex and difficult sequences of movements [21].

Corazza et al. also use the idea of annealing to employ a markerless technique, which uses visual hull reconstruction and an a priori model of the subject [19]. Visual hull of an object is the convex approximation of the volume occupied by the subject, and can be approximated by volume intersection. After the visual hull reconstruction by projection of the subject's silhouettes from each of the cameras, model matching to the visual hull is performed with adapted fast simulated annealing approach. Tracking capability of this approach is evaluated in a virtual environment and the results captured in a gait laboratory are compared to validate this approach in a clinical environment. This technique offers great potential, since it

is based on the entire shape, instead of a small number of points, which makes it perform well, even when the camera resolutions are low [19].

Bregler and Malik [30] demonstrate a motion estimation technique that is able to extract high degree-of-freedom articulated human body configurations from complex video sequences using exponential maps and twist motions. The product of exponential maps and twist motions and their integration into differential motion estimation is a significant parameterization. By this way, the pose of each body segment is defined with respect to its "parent" segment which is attached through a revolute joint.

Chu et al. [31] propose an approach in which underlying nonlinear axes (or skeleton curve) from a volume of a human subject are used. Human volumes are captured with multiple cameras and the kinematic posture is estimated by using skeleton curves. These curves are used to automatically produce kinematic motion. To do so, nonlinear spherical shells (NSS) are used for extracting the skeleton point features that are linked to the underlying axis of the human. The procedure for NSS consists of three steps: (1) volume is transformed into an intrinsic space using the Isomap algorithm, (2) pose-independent volume is divided so that principal curves are found in intrinsic space, (3) a skeleton curve for the point volume is produced, and the kinematic posture of the human subject is determined. This technique is fast and accurate enough to be applied to all frames in a motion and accomplishes posture definition without an a priori model. Also, it can be used as the initialization step of the marker-based techniques.

In another application, Grauman et al. present an image-based approach to infer 3-D structure parameters [32]. Probabilistic shape and structure models are created by using a probability density of multi-view silhouette images with known 3-D structure parameters. This model is merged with a model of the observation uncertainty of the silhouettes seen in each camera in order to compute Bayesian estimate of structure parameters. Hence, this is a study where an image-based statistical shape model is used in order to infer the 3-D structure. Also, by using a computer graphics model of articulated bodies, a database of views augmented with the known 3-D feature locations are formed in order to learn the image-based models from known 3-D shape models. This synthetic training set removes the necessity of labelled real data. The main strength of the approach lies within the use of a probabilistic multi-view shape model to restrict the object shape and its possible configurations.

A summary of the above mentioned markerless applications are given in Table 5.1.

Markerless techniques have great potential in terms of proposing an alternative, easy and low-cost solution. Nevertheless, the use of markerless techniques to capture human movement for biomechanical or clinical applications has been restricted by the complexity of acquiring accurate 3-D kinematics. The problem of estimating the free motion of the human body is under-constrained when compared with marker-based systems. Although many computer vision approaches offer a great potential for markerless motion capture, they have not been validated for biomechanical applications. Existing approaches should be assessed thoroughly in order to address biomechanical applications. Besides, simple or general models of a human body

Table 5.1 Summary of the markerless applications described in this section

Reference	Main Body Representation	Application
Howe et al. [24]	Cardboard model—14 body parts	2-D model-based
Ju et al. [25]	Connected planar patches	2-D model-based
Cham and Rehg [23]	Scaled Prismatic Model	2-D model-based
Mori and Malik [26]	–	2-D model-free
Elgammal and Lee [28]	–	2-D model-free
Bottino and Laurentini [29]	Silhouette and volumetric description	3-D model-based
Deutscher and Reid [21]	Articulated body model	3-D model-based
Corazza et al. [19]	Visual hull	3-D model-based
Bregler and Malik [30]	Articulated body model	3-D model-based
Chu et al. [31]	–	3-D model-free
Grauman et al. [32]	–	3-D model-free

are often used for enhancing computational performance, but biomechanical and clinical applications require detailed and accurate representation of 3-D joint mechanics [33]. To sum up, different methodologies should be combined in order to provide a solution to use prior knowledge in a more effective way [22, 34].

5.2.3 Inertial Measurements

Using inertial and magnetic sensors for body tracking is a relatively new technology. They are independent of an artificially generated source (i.e. sourceless), so they are free from range limitations seen in cameras (e.g. the recently introduced Kinect sensor has a ranging limit of 1.2–3.5 m, similarly, most 3-D sensors operate in well-calibrated distances) and interference problems (e.g. illumination effects).

Low-cost, small size microelectro-mechanical systems (MEMS) sensors are used in the production of wrist-watch-sized inertial/magnetic sensor modules, which make it possible to track orientation in real time. Also, placing these sensor modules to each of the major limb segments of human body makes it possible to independently estimate the orientation of each segment relative to an earth-fixed reference frame. It is also possible to compile the human model from these independent limb segments without knowing their relative orientation [35].

Inertial measurement units generally consist of three different sensors: accelerometers, gyroscopes and magnetometers. Their descriptions and application areas are briefly explained in the following sections.

5.2.3.1 Accelerometers

Accelerometers are devices which measure the applied acceleration along an axis. Although there exist different transducers for this purpose (piezoelectric crystals,

piezoresistive sensors, servo force balance transducers, electronic piezoelectronic sensors, etc.), the main theory behind the accelerometers is a spring mass system. The response of the small mass within the system (a force to the spring) is used in order to calculate the applied acceleration.

Using accelerometers provides a practical and low-cost method for monitoring human movements. They are used to measure physical activity levels, for movement identification and classification, and to monitor movements such as gait, sit-to-stand, postural sways and falls (see Chap. 12).

A uni-axial accelerometer records accelerations in a single direction, while a triaxial accelerometer operates on three orthogonal axes and provides the measurements on each axis. To measure body parts, accelerometers are placed on the body part whose movement is being studied. To measure whole body movements, multiple instruments are used [36].

Activity recognition from accelerometer data is a very active topic in research. In the study of Bao and Intille [37], subjects wear 5 bi-axial accelerometers on different body parts while performing activities such as walking, sitting, standing still, bicycling etc. Data extracted from accelerometers are used in order to train a set of classifiers to discriminate between types of activities [38]. The fact that most modern mobile phones are equipped with accelerometers creates new application drives.

5.2.3.2 Gyroscopes

Gyroscope is a device consisting of a vibrating element merged with a sensing element, functioning as a Coriolis sensor. The Coriolis effect is an evident force that manifests itself in a rotating reference frame and it is proportional to the angular rate of rotation.

The gyroscope provides angular velocity measurements. Joint angles are derived by the integration of angular velocity, but data obtained can be distorted by offsets and drifts. Alternatively, gyroscopes are used to measure angular velocity without being affected by gravity and linear acceleration.

Their low current consumption makes gyroscopes appropriate for ambulatory monitoring. Aminian et al. propose such a system for the estimation of spatio-temporal parameters during long periods of walking [39]. The values of gait parameters are computed from the angular velocity of lower limbs by using wavelet transform. The validation of measurements was assessed using foot pressure sensors as a baseline and data were gathered from young and elderly subjects to calculate the accuracy of the proposed system in a broad range for each gait parameter. The proposed method seems to be a significant monitoring tool for several reasons. It enables measurements of gait features during a long period of walking. The portability of the system makes it possible to be used in other settings than a gait laboratory, and to obtain information regarding the real performance of the subjects [39].

Tong and Granat investigate the usage of uni-axial gyroscopes to develop a basic portable gait analysis system [40]. Gyroscopes are attached on the shank and thigh

segments' skin surface and the angular velocity for each segment is recorded. Using the segment angular velocities, segment inclinations and knee angle are derived.

5.2.3.3 Magnetometers

Magnetometer is a device which measures the strength and direction of the magnetic field in its locality. In general, magnetometers are combined with accelerometers and gyroscopes in biomechanics applications to increase the reliability of the system and to make the definition in the global reference frame possible. To do so, they are integrated into Magnetic, Angular Rate and Gravity (MARG) sensor modules.

Bachmann et al. design a MARG sensor module in order to measure the three degrees of freedom orientations in real time without singularities[1] [41]. Each MARG sensor module contains orthogonally mounted micro-machined rate sensors, accelerometers and magnetometers for a total of nine sensor components. The MARG sensor requirements are derived from the necessities of human body motion tracking. The design goal behind the MARG sensor is being able to measure three degrees of freedom rotational motions without singularities, to be sourceless (not depending on a generated signal source) and to have a suitable form factor, i.e., it should not encumber a human subject when the sensor units are attached.

Design and implementation of MARG sensors demonstrate that all sensor components are linear within the intended operating conditions. Besides being used in human body tracking, these sensor units have important applications in teleoperation, virtual reality and entertainment as well [41].

Marins et al. present an extended Kalman filter for real-time estimation of rigid body orientation using MARG sensors [42]. The filter represents rotations using quaternions rather than Euler angles, in order to eliminate the singularities. The linearity of the Kalman filter reduces the computational time, and the orientation estimation is assessed in real-time [42].

Yun and Bachmann also design a quaternion-based Kalman filter by preprocessing the accelerometer and magnetometer data using the single-frame QUaternion ESTimator (QUEST) algorithm [43]. QUEST represents the positioning of a rigid body relative to a fixed coordinate system. The quaternion produced by the QUEST algorithm is provided as input to the Kalman filter along with angular rate data. When compared with previous approaches, this preprocessing step significantly reduces the complexity of the filter design. Filter performance is validated in experiments and the results are very promising. Even when there are delays, the algorithm manages to handle the dynamic errors [35].

[1] If the external earth magnetic field is in alignment with any of the sensor axes, the rotation rate about that axis cannot be determined. This is called a singularity.

5.3 Force Platforms and Electromyography

A force platform (or a force plate), an equipment with either strain-gauge or piezoelectric transducers, is broadly used in gait analysis. Force platforms are fixed in the ground and they record the force between the ground and the plantar surface of the foot—i.e. ground reaction forces. In general, force plates provide a three-dimensional description of the ground reaction force. The output signals show three components of the force (vertical, lateral and fore-aft), two coordinates of the center of pressure, and the moments about the vertical axis.

By using the ground reaction force data, the resultant forces and moments acting at the joints of the subject's lower extremities—ankles, knees, and hips—are calculated [44]. Force plates are practical in determining the toe-off and heel-strike phases of the gait cycle by observing the ground reaction forces.

The disadvantages of using force plates while interpreting forces are: (1) they should be built on the walkway; (2) the number of different contact surfaces to be measured is limited, (3) during one measurement cycle, only one foot is measured.

Electromyography (EMG) is a significant technique used in biomechanics in order to study muscle function and dysfunction by recording motor unit activities of muscles with surface electrodes.

Monitoring primary form of EMG data (raw data) is essential for detecting signal artefacts, typically caused by cable and wire movements which affect the input impedance and friction at the electrode-skin interface. Oscilloscopes are used in order to monitor the raw EMG signals.

EMG is used for analyzing the muscle functions, muscle tensions and biofeedback [45]. Three important applications of surface EMG signals are: (1) initiation of muscle activation, (2) force generation by a muscle, and (3) measurement of the fatigue within a muscle. Measuring the force contribution of muscles is important in modelling the segments of the musculoskeletal system. Also, measurement of fatigue has the potential to predict the beginning of contractile fatigue, which is associated with exercise intolerance in patients with chronic obstructive pulmonary disease (COPD) [46, 47].

Crosstalk—the signal recorded over one muscle that is actually generated by a nearby muscle—is an important problem with the EMG. Crosstalk is affected by the electrode system used, layer thickness and conductivity of the skin. Mesin et al. provide a mathematical description of the muscle fibre anatomy to simulate and reduce crosstalk [48].

5.4 Summary

This chapter gives a brief introduction on the background theory and current applications of the gait analysis in a clinical context. Motion measurements are handled under the sections of motion capture and inertial measurements. Motion capture section focuses on the marker-based and markerless techniques, brief background information, algorithms and applications. The inertial measurements section summarizes

accelerometers, gyroscopes and magnetometers with their technical properties and application areas. Finally, force measurements and muscle activity measurements are briefly introduced.

5.5 Questions

(1) Explain what CAST stands for and how it works.
(2) What are the limitations of marker-based techniques?
(3) What are the contributions of markerless techniques?
(4) What is a model-based markerless technique?
(5) What are the limitations of markerless techniques?
(6) What are the inertial sensors?
(7) What are the possible applications of force plates?
(8) What are the possible applications of EMG in biomechanics?

5.6 Glossary

- *Direct Linear Transform*: It is an algorithm which solves a set of variables by using similarities. In this text, it is used to describe a calibration algorithm where a calibration object with control points is used.
- *Simultaneous Multi-frame Analytical Calibration*: This is a self-calibration technique based on a planar calibration object with a grid of known control points. It requires the recording of the calibration object by at least two convergent cameras.
- *Calibrated Anatomical System Technique*: It is a methodology that describes anatomical landmark calibration.
- *Scaled Prismatic Models*: They are a class of 2D kinematic models which enforce 2D constraints consistent with the core 3D model.
- *Magnetic Angular Rate and Gravity*: They are hybrid instrumental measurement units which are composed of accelerometer, gyroscope and magnetometer.
- *QUaternion ESTimator*: It is an algorithm that is used to estimate single-frame quaternion which is a representative of the movement of a rigid body in a fixed coordinate system.

References

1. Cappozzo, A., Della Croce, U., Leardini, A., Chiari, L.: Human movement analysis using stereophotogrammetry Part 1: theoretical background. Gait Posture **21**, 186–196 (2005)
2. Aggarwal, J., Cai, Q.: Human motion analysis: a review. Comput. Vis. Image Underst. **73**, 428–440 (1999)
3. Davis, R.B., Õunpuu, S., Tyburski, D., Gage, J.R.: A gait analysis data collection and reduction technique. Hum. Mov. Sci. **10**, 575–587 (1991)

4. Cappozzo, A., Catani, F., Leardini, A., Benedetti, M.G., Della Croce, U.: Position and orientation in space of bones during movement: experimental artefacts. Clin. Biomech. **11**(2), 90–100 (1995)
5. Cappozzo, A., Catani, F., Della Croce, U., Leardini, A.: Position and orientation of bones during movement: anatomical frame definition and determination. Clin. Biomech. **10**(4), 171–178 (1995)
6. Benedetti, M.G., Catani, F., Leardini, A., Pignotti, E., Giannini, S.: Data management in gait analysis for clinical applications. Clin. Biomech. **13**(3), 204–215 (1998)
7. Donati, M., Camomilla, V., Vannozzi, G., Cappozzo, A.: Enhanced anatomical calibration in human movement analysis. Gait Posture **26**, 179–185 (2007)
8. Ferrari, A., et al.: Quantitative comparison of five current protocols in gait analysis. Gait Posture **28**, 207–216 (2008)
9. Baker, R.: Gait analysis methods in rehabilitation. J. NeuroEng. Rehabil. **3**(4), 1–10 (2006)
10. Fioretti, S., Jetto, L.: Accurate derivate estimation from noisy data: a state space approach. Int. J. Syst. Sci. **20**, 33–53 (1989)
11. Chen, L., Armstrong, C.W., Raftopoulos, D.D.: An investigation on the accuracy of three-dimensional space reconstruction using the direct linear transformation technique. J. Biomech. **27**(4), 493–500 (1994)
12. Della Croce, U., Cappozzo, A.: A spot check for estimating stereophotogrammetric errors. Med. Biol. Eng. Comput. **38**(3), 260–266 (2000)
13. Abdel-Aziz, Y.I., Karara, H.M.: Direct linear transformation into object space coordinates in close range photogrammetry. In: Proceedings of the ASP Symposium on Close-Range Photogram, Urbana, IL, pp. 1–18 (1971)
14. Woltring, H.J.: Planar control in multi-camera calibration for three dimensional gait studies. J. Biomech. **13**(1), 39–48 (1980)
15. Chiari, L., Della Croce, U., Leardini, A., Cappozzo, A.: Human movement analysis using stereophotogrammetry Part 2: Instrumental errors. Gait Posture **21**, 197–211 (2005)
16. Lucchetti, L., Cappozzo, A., Cappello, A., Della Croce, U.: Skin movement artefact assessment and compensation in the estimation of knee-joint kinematics. J. Biomech. **31**, 977–984 (1998)
17. Alexander, E., Andriacchi, T.P.: Correcting for deformation in skin-based marker systems. Journal of Biomechanics **34**, 355–361
18. Della Croce, U., Leardini, A., Chiari, L., Cappozzo, A.: Human movement analysis using stereophotogrammetry Part 4: assessment of anatomical landmarks misplacement and its effects on joint kinematics. Gait Posture **21**, 226–237 (2005)
19. Corazza, S., Mündermann, L., Chaudhari, A.M., Demattio, T., Cobelli, C., Andriacchi, T.P.: A markerless motion capture system to study musculoskeletal biomechanics: visual hull and simulated annealing approach. Ann. Biomed. Eng. **34**(6), 1019–1029 (2006)
20. Gavrila, D.M.: The visual analysis of human movement: a survey. Comput. Vis. Image Underst. **73**(1), 82–98 (1999)
21. Deutscher, J., Reid, A.: Articulated body motion capture by annealed particle filtering. In: Proceedings of Computer Vision and Pattern Recognition, South Carolina, pp. 126–133 (2000)
22. Poppe, R.: Vision-based human motion analysis: an overview. Comput. Vis. Image Underst. **108**, 4–18 (2007)
23. Cham, T.J., Rehg, J.M.: A multiple hypothesis approach to figure tracking. In: Proceedings of Computer Vision and Pattern Recognition, Ft. Collins, CO, pp. 239–245 (1999)
24. Howe, N.R., Leventon, M.E., Freeman, W.T.: Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. Advances in Neural Information Processing Systems, vol. 12. MIT Press, Cambridge (2000)
25. Ju, S., Black, M.J., Yacoob, Y.: Cardboard people: a parametrized model of articulated image motion. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, Killington, pp. 38–44 (1996)

26. Mori, G., Malik, J.: Recovering 3D human body configurations using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **28**(7), 1052–1062 (2006)
27. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. Comput. Vis. Image Underst. **80**, 349–363 (2000)
28. Elgammal, A., Lee, C.S.: Inferring 3D body pose from silhouettes using activity manifold learning. In: Proceedings of Computer Vision and Pattern Recognition, Washington DC, USA, pp. 681–688 (2004)
29. Bottino, A., Laurentini, A.: A silhouette based technique for the reconstruction of human movement. Comput. Vis. Image Underst. **83**, 79–95 (2001)
30. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: Proceedings of Computer Vision and Pattern Recognition, Santa Barbara, CA, pp. 8–15 (1998)
31. Chu, C.W., Jenkins, O.C., Matari, M.J.: Towards model-free markerless motion capture. In: Proceedings of Computer Vision and Pattern Recognition, Madison, WI, pp. 475–482 (2003)
32. Grauman, K., Shakhnarovich, G., Darrell, T.: Inferring 3D structure with a statistical image-based shape model. In: Proceedings of International Conference on Computer Vision, Nice, France, vol. 1, pp. 641–647 (2003)
33. Mündermann, L., Corazza, S., Andriacchi, T.P.: The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. J. Neuroeng. Rehabil. (2006)
34. Bray, J.: Markerless based human motion capture: a survey. Department of Systems Engineering, Brunel University (2001)
35. Yun, X., Bachmann, E.: Design, implementation, and experimental results of a quaternion-based Kalman filter for human body motion tracking. IEEE Trans. Robot. **22**(6), 1216–1227 (2006)
36. Mathie, M.J., Celler, B.G., Lovell, N.H., Coster, A.C.F.: Classification of basic daily movements using a triaxial accelerometer. Med. Biol. Eng. Comput. **42**, 679–687 (2004)
37. Bao, L., Intille, S.: Activity recognition from user-annotated acceleration data. In: Proceedings of the 2nd International Conference on Pervasive Computing, Vienna, Austria, pp. 1–17 (2004)
38. Aminian, K., Robert, P.H., Buchser, E.E., Rutschmann, B., Hayoz, D., Deparion, M.: Physical activity monitoring based on accelerometry: validation and comparison with video observation. Med. Biol. Eng. Comput. **37**(3), 304–308 (1999)
39. Aminian, K., Najafi, B., Büla, C., Leyvraz, P.F., Robert, P.H.: Spatio-temporal parameters of gait measured by an ambulatory system using miniature gyroscopes. J. Biomech. **35**, 689–699 (2002)
40. Tong, K., Granat, M.H.: A practical gait analysis system using gyroscopes. Med. Eng. Phys. **21**, 87–94 (1999)
41. Bachmann, E.R., Yun, X., Mickinney, D., McGhee, R.B., Zyda, M.J.: Design and Implementation of MARG sensors for 3-DOF orientation measurement of rigid bodies. In: Proceedings of the 2003 IEEE International Conference On Robotics & Automation, Taipei, Taiwan, September 14–19 2003
42. Marins, J.L., Yun, X., Bachmann, E.R., McGhee, R.B., Zyda, M.J.: An extended Kalman filter for quaternion-based orientation estimation using MARG sensors. In: Proceedings of the 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems, Maui, Hawaii, USA, pp. 2003–2011 (2001)
43. Shuster, M.D., Oh, S.D.: Three-axis attitude determination from vector observations. J. Guid. Control **4**(1), 70–77 (1981)
44. Vaughan, C.L., Davis, B.L., Jeremy, C.O.C.: Dynamics of Human Gait. Kiboho, Cape Town (1999)
45. Soderberg, G.L., Cook, T.M.: Electromyography in biomechanics. Phys. Ther. **64**, 1813–1820 (1984)

46. De Luca, C.J.: The use of surface electromyography in biomechanics. J. Appl. Biomech. **13**, 135–163 (1997)
47. Saey, D., Côté, C.H., Mador, J., Laviolette, L., Leblanc, P., Jobin, J., Maltais, F.: Assessment of muscle fatigue during exercise in chronic obstructive pulmonary disease. Muscle Nerve **34**, 62–71 (2006)
48. Mesin, L., Smith, S., Hugo, S., Viljoen, S., Hanekom, T.: Effect of spatial filtering on crosstalk reduction in surface EMG recordings. Med. Eng. Phys. **31**, 374–383 (2009)

Chapter 6
Hand Gesture Analysis

Cem Keskin, Oya Aran, and Lale Akarun

6.1 Hand Gestures in Human Communication

Webster's dictionary defines a gesture as: (1) "a movement usually of the body or limbs that expresses or emphasizes an idea, sentiment, or attitude"; (2) "the use of motions of the limbs or body as a means of expression" [25]. Most gestures are performed with the hand, but the face and the body also play an important role during gesturing. Specifically, hand gestures are formed by the shape and movement of the hand, as well as its position with respect to other body parts.

Gestures are used in many aspects of human communication. They can be used to consciously or unconsciously accompany speech, or to communicate in environments where speaking is hard or impossible. In a more structured way, they are used to form the sign languages of the hearing-impaired people. With the progress on Human–Computer Interaction (HCI), gestures have found a new area of usage. Systems that enable the use of computer programs with hand gestures, such as operating system control, games, and virtual reality applications, have been developed.

C. Keskin (✉) · L. Akarun
Computer Engineering Department, Boğaziçi University, Istanbul, Turkey
e-mail: keskinc@cmpe.boun.edu.tr

L. Akarun
e-mail: akarun@boun.edu.tr

O. Aran
Idiap Research Institute, Martigny, Switzerland
e-mail: oya.aran@idiap.ch

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_6,

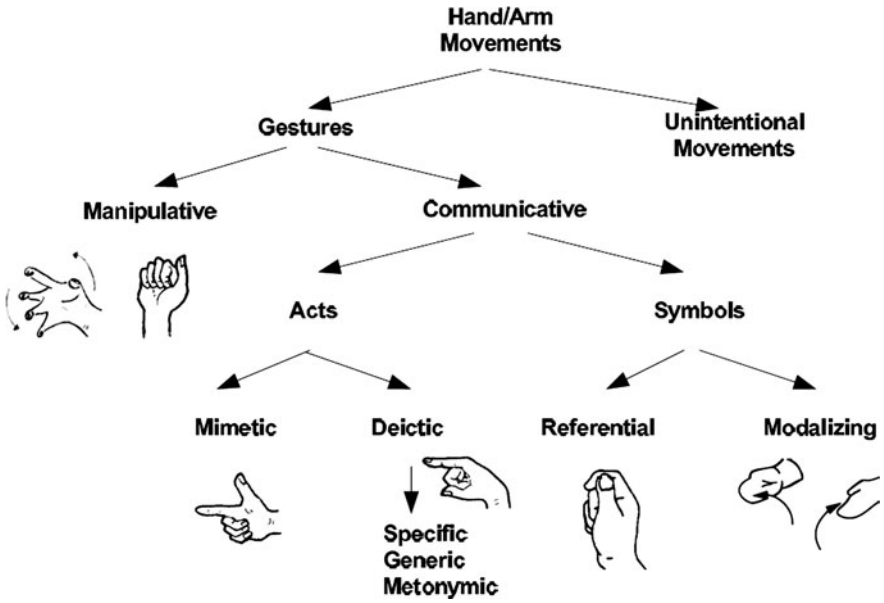


Fig. 6.1 A taxonomy of hand gestures for HCI

6.1.1 Taxonomy of Hand Gestures

Although hand gesture recognition for HCI is a relatively recent research area, the research on hand gestures used in human–human communication is well developed. Several taxonomies are presented in the literature by considering different aspects of gestures. For instance, hand gestures can be classified with respect to their independence, such as autonomous gestures, and gesticulation, which are gestures used together with another means of communication [18]. In [24], gestures are classified into three groups: *Iconic gestures*, which are used to display objects, spatial relations, and actions; *metaphoric gestures*, which explain a concept; *beats*, which are rhythmic beating of fingers, hands or arms. Another set of gesture categories consists of symbolic gestures, which are conventional, context-independent expressions; deictic gestures, which point to entities; iconic gestures; and pantomimic gestures, which imitate using an object. In [36], gestures are classified into four groups: conversational, controlling (pointing to an object), manipulative (moving and handling a virtual object) and communicative (sign language gestures, HCI commands). A similar categorization is given in [30], which views the gestures in terms of the relation between the intended interpretation and the abstraction of the movement. In [29], this taxonomy of gestures is accepted as the most appropriate one for HCI purposes. An extended version of this taxonomy is given in Fig. 6.1, and explained in the following sections.

6.1.2 Hand Gestures Accompanying Speech

Hand gestures are frequently used in human to human communication, either alone or together with speech. There is considerable evidence that hand gestures are produced unconsciously along with speech in many situations and enhance the content of accompanying speech. It is also known that even when the listener cannot see the hands of the speaker or there is no listener at all, hand gestures are produced.

The hand/ arm movements during conversation can be classified into two groups: intended or unintended. Although unintended hand movements must also be taken into account in order to realize human–computer interaction as natural as human-human interaction, current research on gesture recognition focuses on intended gestures, which are used for either communication or manipulation purposes. Manipulative gestures are used to act on objects, such as rotation and grasping, whereas communicative gestures have an inherent communicational purpose. In a natural environment, they are usually accompanied by speech.

Communicative gestures can be *acts* or *symbols*. Symbol gestures are generally used in a linguistic role with a short motion as in sign language. In most cases, the symbol itself is not directly related to the meaning and these gestures have a predetermined convention. Symbols consist of two types of gestures, namely *referential* and *modalizing* gestures. Referential gestures are used to refer to an object or a concept independently. For example, rubbing the index finger and the thumb in a circular fashion independently refers to money. Modalizing gestures are used with some other means of communication, such as speech. For example, the sentence "I saw a fish, it was **this** big." is only meaningful with the gesture of the speaker.

Unlike symbol gestures, act gestures are directly related to the intended interpretation. Such movements are classified as either *mimetic* or *deictic*. Mimetic gestures usually mimic a concept or object. For example, a smoker going through the motion of "lighting up" with a cigarette in his mouth indicates that he needs a light. Deictic gestures, or pointing gestures, are used for pointing to objects.

Another type of gesture that has significantly different characteristics than the rest is the beat gesture [24]. Beat gestures consist of short and quick movements of the hand or fingers along with the rhythm of speech. A typical beat gesture is the quick and rhythmic motion of the hand up and down, or back and forth. Beats are mainly used to create emphasis and grab attention.

With the exception of beat gestures, each gesture starts, continues for some interval and ends. This is not only valid for dynamic gestures that include both spatial and temporal components, but also for static gestures that only contain spatial components. A gesture is constituted in three phases: preparation, stroke, and retraction or recovery [24]. In the preparation phase, the hand is oriented for the gesture. The stroke phase is the phase of the actual gesture. Finally, in the retraction phase, the hand returns to the rest position. The preparation and the stroke phases constitute a gesture phrase and together with the recovery phase, they constitute a gesture unit [19].

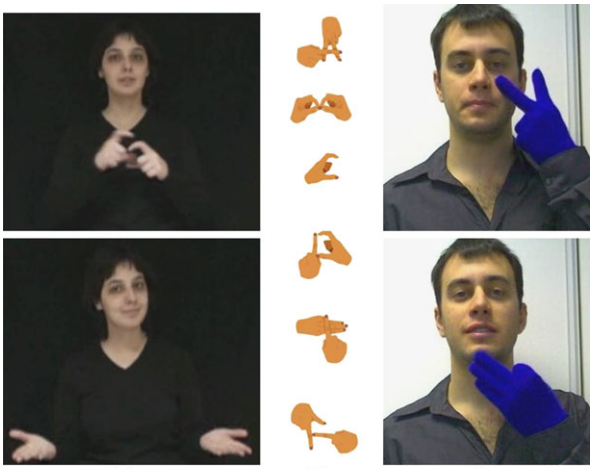


Fig. 6.2 Examples from (**a**) Turkish sign language, (**b**) Turkish finger spelling, and (**c**) French cued speech

6.1.3 Hand Gestures in Hearing Impaired Communication

Sign languages are the natural communication media of hearing-impaired people. Like the spoken languages, they emerge spontaneously and evolve naturally among hearing-impaired communities.

The signs are perceived visually and produced alone or simultaneously, by use of hand shapes, hand motion, and hand location (manual signs), as well as facial expressions, head motion, and body posture (non-manual signs). Sign languages have both sequential and parallel nature, since signs come one after the other, showing a sequential behavior. However, each sign may contain parallel actions of hands, face, head or body. Apart from differences in production and perception, sign languages contain phonology, morphology, semantics, and syntax like spoken languages [32]. Figure 6.2(a) shows an example sign from Turkish sign language (Türk İşaret Dili—TID).

Apart from sign languages, there are other means of hearing-impaired communication: finger spelling and cued speech. Finger spelling is a method of visually spelling words by using certain hand gestures for each letter. It is an important part of sign languages, and it can be used to represent words which have no sign equivalent, to emphasize or clarify concepts, or when teaching or learning a sign language. Figure 6.2(b) shows some examples from the finger-spelling alphabet of TID.

Cued speech is a mode of communication based on the phonemes and properties of spoken languages [8]. It uses both lip shapes and hand gestures to represent the phonemes. The aim of cued speech is to overcome the problems of lip-reading and to enable a full understanding of spoken languages. Cued speech replaces invisible articulators that participate in the production of the sound (vocal cords, tongue, and jaw) by hand gestures, while keeping visible articulators (lips). Basically, it complements the lip-reading by various hand gestures, so that phonemes which have similar lip shapes can be differentiated. Figure 6.2(c) shows an example word from French cued speech.

6.1.4 Hand Gestures in Human–Computer Interaction

HCI using hand gestures is a new area, enabled by advances in computer technologies and popularized by science fiction movies. Especially, with the recent introduction of cheap depth sensors, natural interaction based interfaces are expected to be used increasingly in daily life. Hand gestures are expected to replace remote controls, mice and keyboards at least for simpler tasks.

Both manipulative and communicative gestures are crucial for hand gesture based HCI. Communicative gestures are used to give high level commands to the system (e.g. change the channel on TV, choose a tool in a modeling application, start the car), and manipulative gestures are used to *tune* the system (e.g. set the volume level, brightness, application parameters). The selection of actual gestures depends on the application or system. Currently, gestures used for HCI systems are heavily influenced by the gestures used for multi-touch screens. However, these are not necessarily the most natural gestures, especially because hand gestures in the air are more tiring to perform.

Hand gestures characteristically possess more degrees of freedom than conventional input devices. Each joint angle, and any configuration of those angles, as well as the absolute or relative location of the hand can be used to communicate intents. This suggests that an HCI system will benefit more from hand gestures, if the system or application is specifically designed to make use of this higher degree of freedom. For instance, the effectiveness of a regular mouse is limited for a 3D modeling tool, whereas the 3D nature of hand gestures is naturally suitable to view and manipulate information in 3D. However, hand gesture based systems are not as precise as other input devices they aim to replace at the moment.

Required level of accuracy for hand gestures is different for each system. For some gestures, the joint angles for all the fingers may be needed with high accuracy, whereas for some other gestures, only the motion of the hand is important. The exact direction of the hand motion may not be important for some gestures, but it should be analyzed for many of the manipulative gestures. For deictic gestures, the hand posture is always the same; there is generally no hand motion, and what is important is the direction of the pointing finger.

The rest of this chapter covers common approaches and state-of-the-art techniques for hand gesture recognition. First, a general framework for recognizing and acting on hand gestures is given in Sect. 6.2. Hand pose estimation techniques are briefly introduced in Sect. 6.3. Common graphical models used for recognition are explained and analyzed in Sect. 6.4, along with techniques to tackle the gesture spotting problem. Finally, several application examples are given in Sect. 6.5.

6.2 Hand Gesture Recognition Framework

A general framework for vision based hand gesture recognition is given in Fig. 6.3. It starts with a camera system retrieving images, which are then preprocessed to

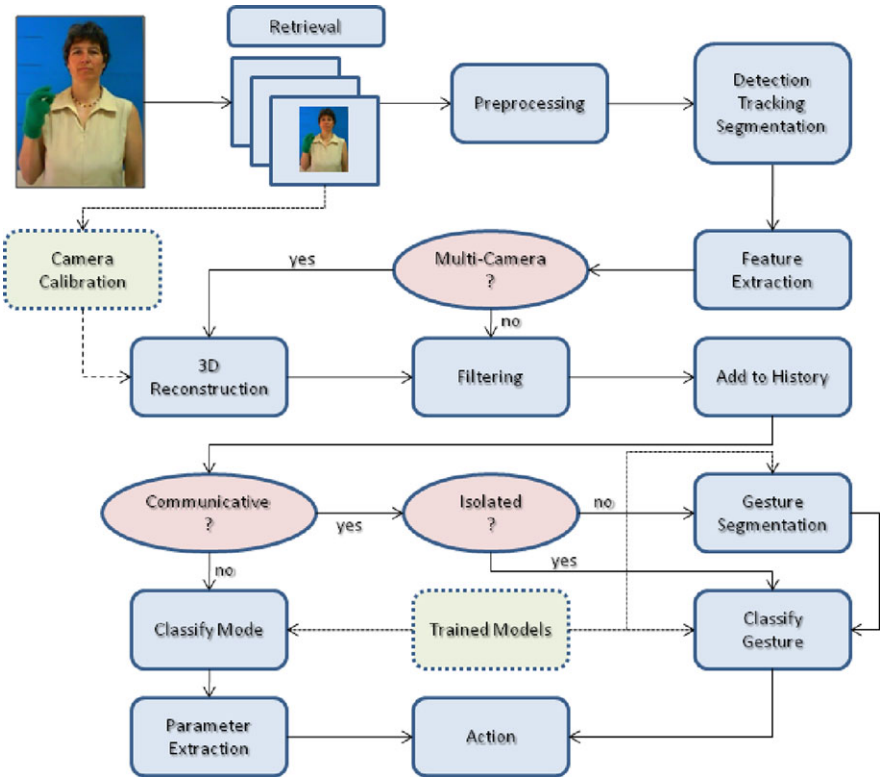


Fig. 6.3 A general framework for recognizing manipulative and communicative hand gestures

simplify the hand detection and segmentation tasks. The preprocessing step may involve methods like noise reduction, edge detection, color space conversion and color segmentation.

Although object detection and segmentation problems have attracted considerable interest in the field, and sophisticated algorithms have been developed [13], detection, tracking, and segmentation of the hand can be a very complicated problem depending on the environment. For instance, skin color can be used to detect a hand in front of a cluttered background, but not in front of a face or another body part. Distinguishing the hand from the face involves complex algorithms that are presently not suitable for real-time applications. Therefore, a common solution is to make use of colored markers (see Fig. 6.4(a)). Certain assumptions about the background can be made, such as assuming that there is a single skin colored blob, or that the hands are always the fastest objects in the scene.

In natural settings, a marker cannot be used and there are several challenges such as unconstrained clothing, changing speakers, changing backgrounds, frequent contacts and occlusions between the hands and the face, low resolution, and motion blur (see Figs. 6.4(b) and 6.4(c)). These challenges make the hand detection and tracking a difficult problem during unconstrained, natural communication, and sophisticated

Fig. 6.4 Hand segmentation examples (**a**) with colored markers in each hand, (**b**) without markers, and (**c**) examples of hand–hand, hand–face occlusion

algorithms, such as probabilistic tracking algorithms [10], are required to solve this task.

Probabilistic tracking algorithms make multiple hypotheses for the object at a given time and estimate the object position by a combination of these hypotheses [10]. Particle Filter (PF) is a well-known example and is well suited to applications where the dynamics of the tracked object are not well-defined. Particle filters estimate the probability of the object state, given the initial state and the observations, using the sequential Bayes method. Conditional density propagation algorithm is a simple implementation of the PF and is proposed for the object tracking problem [14].

In the *feature extraction* phase, the segmented hand image and its difference from the previous image are used to form a feature vector that describes the shape, motion, and location of the hand. There are several approaches to hand shape modeling, which are briefly explained in Sect. 6.3.

Although some gesture recognition systems restrict themselves to a 2D plane and a small 2D gesture vocabulary, 3D information is necessary to handle a larger gesture vocabulary or sign language. Stereo or other special 3D cameras can be used to retrieve 3D information about the scene directly. Otherwise, a 3D description of the hand can be reconstructed using a multiple camera system, which need to be *calibrated*. There are numerous methods for camera calibration [7]. Once the camera matrices are known, the task of estimating the 3D location of the hand reduces to solving a simple linear system [7]. Figure 6.5 shows an example 3D reconstruction from two cameras placed in front of the gesturer.

A filtering step is necessary, as the retrieved images are characteristically noisy, and the extracted feature vectors are also subject to noise. The sensor quality, lighting conditions and algorithm robustness, as well as the variability of the gesturer's performance contribute to this noise, and most systems need to filter the produced feature vectors.

Kalman filter offers a solution for filtering noise. Figure 6.6 shows an example of hand trajectory filtering with a Kalman filter. It is common for vision systems with a high sampling rate to assume linear motion between frames. Under this assumption, the hand motion dynamics can be described by a partially observable stochastic process with linear dynamics and linear observations, which is a suitable framework for a Kalman filter [16]. A simple application of the Kalman filter can be found in [28].

Gesture *recognition* methods used for manipulative and communicative gestures differ significantly. *Manipulative gestures* are frequently chosen such that the ges-

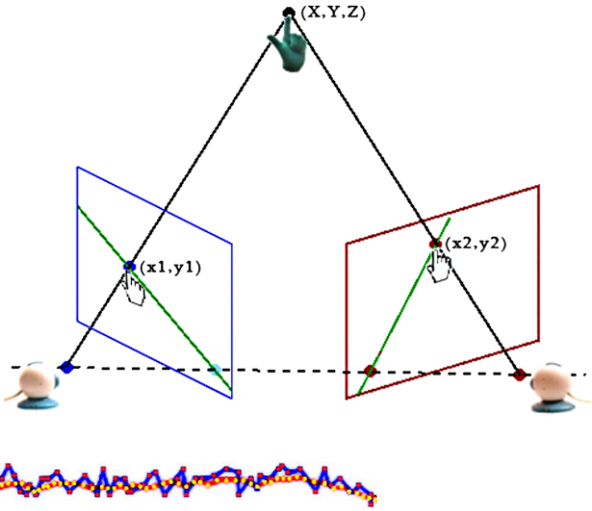


Fig. 6.5 3D reconstruction from two front-facing cameras

Fig. 6.6 Example of hand trajectory filtering with Kalman filter. The *blue straight line* and *dotted red line* show the extracted and filtered hand trajectory, respectively

turers keep their hands in a certain shape indicating a *mode*, and move in a certain manner, indicating a parameter change related to that mode. Hence, manipulative gesture type can be recognized using a single frame, and manipulation parameters are extracted from the change in location or shape of the hand.

Having a *reject class* is important for manipulative gestures. Without a reject class, the system is forced to make a recognition at each step, choosing the most likely gesture each time instant and interpreting the motion of the hand as a parameter change.

Communicative gestures usually involve both motion and shape, and more importantly, they require a much longer memory. This requirement not only stems from the longer duration of these gestures, but also from the fact that the system might need to remember the previous gestures to be able to resolve ambiguities. This is especially true for sign language recognition systems, where the meaning associated with a gesture depends on the context, i.e. on previous gestures.

Applications like sign language interpreters cannot restrict the gesturer to signify start and end points of each sign. Likewise, it is unnatural for gesturers using a HCI application to have to prove their intent each time a gesture is performed. Hence, systems that support continuous gestures have to *spot* and *segment* gestures first (see Fig. 6.7). These methods will be explained in more detail in Sect. 6.4.3.

6.3 Hand Pose Estimation

In vision based systems, all the information regarding the actual 3D hand shape is contained in the 2D projected image of the hand. For some applications, the exact

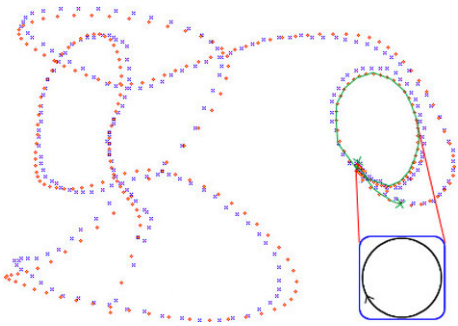


Fig. 6.7 Gesture spotting example

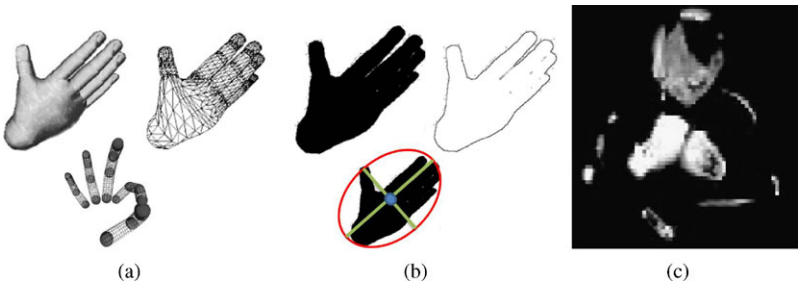


Fig. 6.8 Hand shape feature examples. (**a**) High level shape modeling with 3D hand models, (**b**) low level shape modeling with appearance-based models, and (**c**) low level modeling with motion history images

configuration of fingers is important, which consists of all the joint angles of the hand skeleton. Other applications such as finger spelling, make use of predetermined hand shapes and try to determine the class label of the hand shape, instead of every single angle. Finally, many applications view the hand as a simple blob and try to estimate its certain properties, such as orientation, size and location. In order to infer high level information about the hand, vision based systems must rely on features extracted from the hand image at each frame.

6.3.1 Modeling Hand Shape

The hand can be described either by a high level 3D hand model, or by a low level appearance based model. 3D hand models make use of a priori knowledge about the hand. In the case of a skeletal hand model, the system attempts to estimate joint angles and global orientation directly by minimizing the difference between the 2D projection of the flexible 3D model and the 2D hand image with respect to the model parameters [31]. Alternatively, a voxel model can be reconstructed, which can be reconstructed from multiple silhouette images in order to estimate the joint angles indirectly [34]. Figure 6.8(a) shows several high level feature examples.

Variational segmentation methods can also estimate high-level parameters via an energy minimization technique. These methods regard the general problem of region segmentation, object tracking and 3D interpretation as an optimization problem, where some energy measure that is usually a combination of region and boundary functionals is minimized [3, 17, 26].

Appearance based models are used to relate the image of the hand to its actual posture. The centroid of the hand and location of finger tips are among simple low-level features describing such models (see Fig. 6.8(b)).

The most common low level features are the image moments. Hu moments are invariant under translation, changes in scale, and rotation [12], but they are neither complete, nor independent [11]. Zernike moments, on the other hand, can be used to reconstruct the original image up to the required level of accuracy, and are also rotation, scale and translation invariant [21]. A similar approach is based on principal component analysis (PCA) of the extracted hand images, which provides an efficient representation of the hand using a small number of features that can be used to reconstruct the original image approximately. This is called the *eigenhand* representation of the hand [5, 9], inspired by the analogous eigenface method used for face recognition.

Motion energy images and motion history images are accumulated images that are calculated over a limited history [6]. Motion energy images are union of all the connected regions that show significant change over the given time interval, whereas motion history images decrease the effect of older frames gradually (see Fig. 6.8(c)). Unlike other features, these features describe the hand motion as well, and are used to classify gestures directly. More sophisticated methods for gesture classification will be explained in more detail in Sect. 6.4.

6.3.2 Hand Shape Classification

Hand shape classification is the task of determining the class label, based on an analysis of the projected hand image. A segmented hand image contains very high dimensional and redundant data, which are not directly suitable for classification. Instead, the features mentioned in Sect. 6.3.1 are used to classify hand shapes. Neural networks, support vector machines, Gaussian mixture models, decision trees and RBF classifiers are some of the standard methods that can be used. Using most discriminating features (MDF) is also a good choice, since it is a supervised method that aims to maximize the distance of hand shape classes in the projected subspace [9]. Another method is elastic graph matching, which represents hands with labeled graphs that have Gabor filters attached to the nodes [33]. The graph is superposed on the hand image; the node locations code for pose, and filter responses code for appearance.

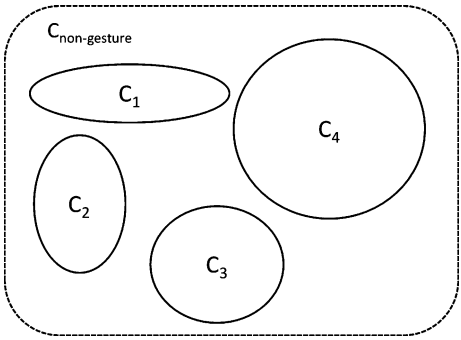


Fig. 6.9 Distribution of class labels over parameter space. Note that the class label corresponding to non-gestures is not bounded

6.4 Hand Gesture Recognition

Gesture signals are of the form $\mathbf{X} = (x_1, \ldots, x_N)$, where each x_t is a D dimensional vector corresponding to extracted features at frame t. Hand gesture classification is the task of automatically assigning $\mathbf{X}$ a class label c from a predefined set $\mathbf{C} = (c_1, \ldots, c_K)$.

It is simpler to classify a sequence $\mathbf{X}^k$ belonging to class c_k, if it is isolated from other gestures, i.e. if its start and end points in time are known. This type of classification is called *isolated gesture recognition*. However, in most cases, the gesture sequence $\mathbf{X}$ is a continuous stream of unknown length, starting at a certain time t_0. Sequences corresponding to gestures are embedded in this stream, starting at frame t_s and ending at frame t_e. In between such gestures, $\mathbf{X}$ may contain feature vector sequences that correspond to unknown gestures, unintentional hand movements or co-articulation artifacts. Such a stream has the following form:

$$\mathbf{X} = (x_{t_0}, \ldots, x_{t_s}, \ldots, x_{t_e}, \ldots, x_T), \tag{6.1}$$

where the index T corresponds to the current time frame. In this case, the sequence $\mathbf{X}^k$ must be *spotted* first. Thus, the system not only needs to distinguish between different gesture classes, but also between *gestures* and *non-gestures*. This type of classification is called *continuous gesture recognition*, and is considerably harder than isolated gesture recognition, since it requires modeling of non-gestures, and the determination of the start and end points of each gesture. Distribution of class labels over a 2D parameter space is depicted in Fig. 6.9. Unlike other gesture classes, the distribution of the non-gesture class over the parameter space is unbounded and directly depends on other gesture classes. An analogy can be drawn between non-gesture and *non-speech* in this manner, which also must be distinguished from actual speech for continuous speech recognition. While *garbage*, or *filler* models are used for non-speech, *threshold* models will be used for non-gesture modeling.

Continuous gesture recognition is especially important for sign language recognition and human–computer interaction applications, as it allows the user to perform each gesture in a seamless manner.

In isolated gesture recognition, the system is provided with a sequence $\mathbf{X}^k$ each time a gesture is performed. Similarly, in the continuous case, the system extracts

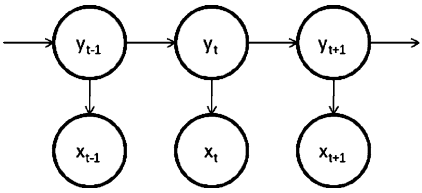


Fig. 6.10 Graphical model of HMM

candidate sequences $\mathbf{X}^k$, which may belong to a predefined gesture class. In either case, the classification problem can be thought of as estimating the class label c^* that satisfies the following equation:

$$c^* = \arg\max_c \mathrm{P}\left(c|\mathbf{X}^k\right). \tag{6.2}$$

Using Bayes' rule, we know that

$$\mathrm{P}\left(c|\mathbf{X}^k\right) = \frac{\mathrm{P}(\mathbf{X}^k|c)\mathrm{P}(c)}{\mathrm{P}(\mathbf{X}^k)} \propto \mathrm{P}\left(c, \mathbf{X}^k\right). \tag{6.3}$$

Here, $\mathrm{P}(\mathbf{X}^k)$ is constant. Therefore, both $\mathrm{P}(c|\mathbf{X}^k)$ and $\mathrm{P}(c, \mathbf{X}^k)$ can be used for classification. In this respect, modeling a gesture means estimating these probabilities. Since $\mathbf{X}$ is sequential and can be arbitrarily long, it is more convenient to use graphical models for classification.

6.4.1 Modeling Hand Gestures with Graphical Models

Graphical models introduce a network of hidden variables $\mathbf{Y}$ and their relations in the form of a graph, which can be used to explain an observation sequence $\mathbf{X}$. The values $\mathrm{P}(c|\mathbf{X})$ and $\mathrm{P}(c, \mathbf{X})$ are determined by estimating $\mathrm{P}(c, \mathbf{Y}|\mathbf{X})$ and $\mathrm{P}(c, \mathbf{Y}, \mathbf{X})$, and then by marginalizing over $\mathbf{Y}$. See Chap. 2 for a detailed explanation of graphical models in general. In the rest of this chapter, it is assumed that the reader has a basic understanding of graphical models.

Bayesian networks model $\mathrm{P}(\mathbf{X}, \mathbf{Y}, c)$, i.e. the joint probability of the observation sequence, the hidden variables and the class label. Since $\mathrm{P}(\mathbf{X}, \mathbf{Y}, c)$ is equal to $\mathrm{P}(\mathbf{X}, \mathbf{Y}|c)\mathrm{P}(c)$, there can either be a large network that models the joint density of all the random variables involved, or separate networks for each gesture, each modeling the joint density of the observations and hidden states conditioned on the class label. These networks can also *generate* sample observation sequences, since $\mathrm{P}(\mathbf{X}|c)$ is implicitly modeled. Hence, they are called *generative* models. The simplest type of Bayesian network is the hidden Markov model (HMM). The graphical model of HMM is depicted in Fig. 6.10.

The ability of generative models to generate samples is not required for classification. Instead, Markov random fields can be used to attack the problem by directly modeling the conditional probability $\mathrm{P}(c|\mathbf{X})$. Such models are called *discriminative*

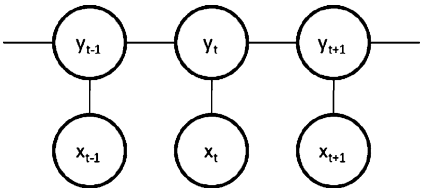


Fig. 6.11 Graphical model of CRF

models, which learn to distinguish between class labels in the parameter space, instead of learning their distribution densities. The simplest type of Markov random field is the conditional random field (CRF), which is the discriminative counterpart of HMM. The graphical model of CRF is given in Fig. 6.11.

CRFs do not model $P(c, \mathbf{Y}|\mathbf{X})$ as expected. Instead, CRFs model $P(\mathbf{Y}|\mathbf{X})$, where each class label is represented by a single node in the network. CRFs model the high level relation of gestures, or *inter-class dynamics*, conditioned on the observations. Hence, CRFs *label* sequences instead of classifying them, forming a sequence of class labels the same size as $\mathbf{X}$. For isolated gesture recognition, the most frequently occurring label can be assigned to the sequence. HMMs, on the other hand, can represent each gesture with several hidden variables, effectively modeling the *intra-class dynamics*.

Intra-class dynamics of gestures are especially important, when the spatio-temporal variability of gestures is high (e.g. when multiple performers are involved), or when the observation features are similar for some gestures. In these cases, duration modeling may be essential to distinguish between gestures. Here, duration is the time spent in a hidden state, which directly affects the final trajectory of the gesture. If the variance of the duration distribution is too large, the trajectory is easily deformed, whereas if it is too low, the model cannot represent the temporal variability of the gesture well enough.

In the case of HMMs, durations of regimes are implicitly modeled with geometric distribution (see Exercise 5). On the other hand, CRFs do not model durations, as they do not model intra-dynamics.

6.4.2 HMM and CRF Variants Used in Gesture Recognition

HMM and CRF are the simplest of all graphical networks. They have several weaknesses, which limit their usage for certain applications. Therefore, several variants have been proposed, which remedy these weaknesses.

6.4.2.1 Hidden Conditional Random Fields

The most severe weakness of CRFs is their inability to model intra-class dynamics. On the other hand, HMMs can effectively learn substructures, but unlike CRFs, they cannot be trained discriminatively. A solution is to augment the CRFs with a number

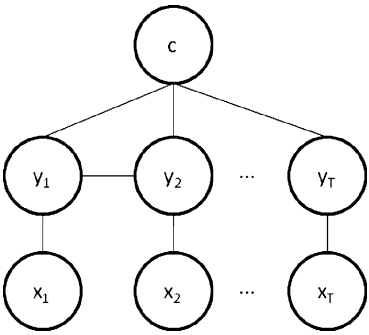


Fig. 6.12 Graphical model of HCRF

of hidden states for the class labels. The resulting Markov random field is called the hidden CRF (HCRF), which incorporates a single class label and several hidden states [35]. The graphical model of HCRF is given in Fig. 6.12. In the figure, x_t are the features, y_t are the hidden states, and C is the class label.

Since HCRFs are discriminative models, either *one-vs.-all* models are used, where different HCRF models learn to distinguish a single gesture from other gestures, or a single *multi-class* model can be used, which can learn to map the observations (or their features) to the class labels. Multi-class HCRF models are shown to outperform CRFs, HMMs, and one-vs.-all HCRFs, which can be attributed to their ability of jointly learning the best discriminative structure [35].

6.4.2.2 Latent Dynamic Conditional Random Fields

While HCRF provides a solution to the weakness of CRF, it also introduces a new weakness. HCRFs employ only a single class variable and therefore, they cannot model high level relationship between gestures, i.e. inter-class dynamics, which is crucial for tasks such as sign language recognition. Latent dynamic CRFs (LDCRF) attempt to combine the strong points of CRFs and HCRFs, i.e. to capture both intra- and inter-class dynamics of gestures [27]. This is achieved by extending the HCRF to incorporate a stream of class labels, which are associated with a disjoint set of hidden states. Moreover, since LDCRF models include a class label per observation, they can be used for recognition on sequences that are not segmented. Thus, they can naturally be used for continuous gesture recognition. The graphical model of LDCRF is given in Fig. 6.13.

6.4.2.3 Input Output Hidden Markov Models

LDCRF does not suffer from the same weaknesses as HMM, CRF, and HCRF, and it is shown to outperform these models [27]. Still, some applications may require the model to be generative (for instance, when the gesture modeled needs to be visualized), or the durations of substructures to be modeled with a certain distribution. The

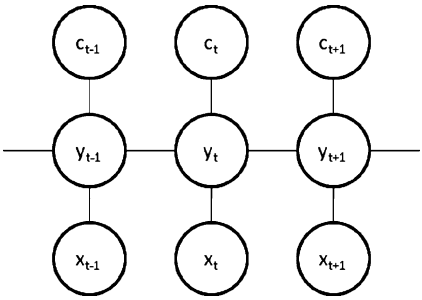


Fig. 6.13 Graphical model of LDCRF

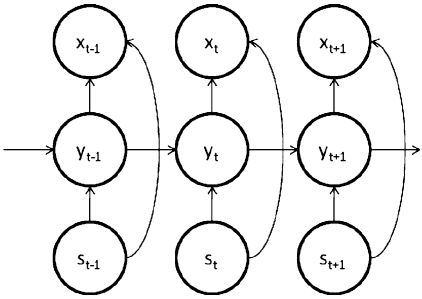


Fig. 6.14 Graphical model of IOHMM

input–output HMM (IOHMM) is a Bayesian network similar to HMM, which conditions the state transition probability $\mathrm{P}(y_{t+1}|y_t)$ and emission probability $\mathrm{P}(x_t|y_t)$ on an external input sequence $\mathbf{S} = (\mathbf{s_1}, \ldots, \mathbf{s_T})$ [4]:

$$\mathrm{P}(y_{t+1} = j|y_t = i, \mathbf{s_t}) = \varphi_{ij,t}, \tag{6.4}$$

$$\mathrm{P}(x_t = \mathbf{m}|y_t = i, \mathbf{s_t}) = \eta_{i,t}(\mathbf{m}). \tag{6.5}$$

Here, $\varphi_{ij,t}$ and $\eta_{i,t}(\mathbf{m})$ are called *local models*, and can be any function such as radial basis functions and neural networks. As in the case of CRF variants, the feature vectors forming $\mathbf{S}$ are not assumed to be independent. The corresponding graphical model is given in Fig. 6.14.

Model selection for IOHMMs is not simple. The input sequence $\mathbf{S}$, the output sequence $\mathbf{X}$ and the local models need to be chosen carefully. IOHMMs will use $\mathbf{S}$ discriminatively to determine next states and emissions, while the system retains the ability to generate samples from $\mathbf{X}$, if a sample of $\mathbf{S}$ is provided. Also, local models can be as simple as binary switches or counters (for instance for explicit duration modeling), or as complex as multi-layered perceptrons with several hidden nodes. $\mathbf{X}$ can even be chosen to be the class labels. In this case IOHMMs behave like LDCRFs, mapping the input sequence to a class label sequence, while capturing the intrinsic dynamics with hidden nodes. To see some different approaches, see [15, 20, 23].

Table 6.1 Certain properties of several common graphical models

Model	Intra-class dynamics	Inter-class dynamics	Duration modeling	Generative	Discriminative
HMM	✓		geometric	✓	
CRF		✓			✓
HCRF	✓				✓
LDCRF	✓	✓			✓
IOHMM	✓	✓	any	✓	✓

6.4.2.4 Comparison of Graphical Models

Table 6.1 lists the important properties of HMMs, CRFs, HCRFs, LDCRFs and IOHMMs. The second and third columns list whether the model is capable of modeling intra-class and inter-class dynamics. The fourth column lists the type of distribution used for duration modeling, if any. The fifth and sixth columns show whether the model is generative or discriminative.

Table 6.1 helps to assess a model's suitability for a given application. Classification speed is also an important measure for real-time applications. As HMMs are simpler, they are also faster than the rest of the models and therefore, they are still widely used. The speed of the other models directly depends on their complexity. The window size w affects CRFs, HCRFs and LDCRFs, whereas the number of hidden nodes affect HMMs, HCRFs, LDCRFs and IOHMMs. Moreover, the speed of IOHMMs depends on the complexity of their local models. Naturally, there is a trade-off between speed and accuracy, and the target hardware should be considered while choosing a model.

The suitability of the models for continuous gesture recognition may also be important. Each model has a different approach to gesture spotting and non-gesture modeling.

6.4.3 Continuous Gesture Recognition

As mentioned at the start of Sect. 6.4, spotting gestures in a continuous stream usually requires distinguishing between gestures and non-gestures, and modeling non-gestures is not as straightforward as modeling gestures. In the rest of this section, common approaches to the continuous gesture recognition task will be explained with some examples.

6.4.3.1 HMM Based Methods

HMM based gesture recognition systems usually model each gesture with a single HMM and use these to evaluate candidate sequences. Then, the posterior likeli-

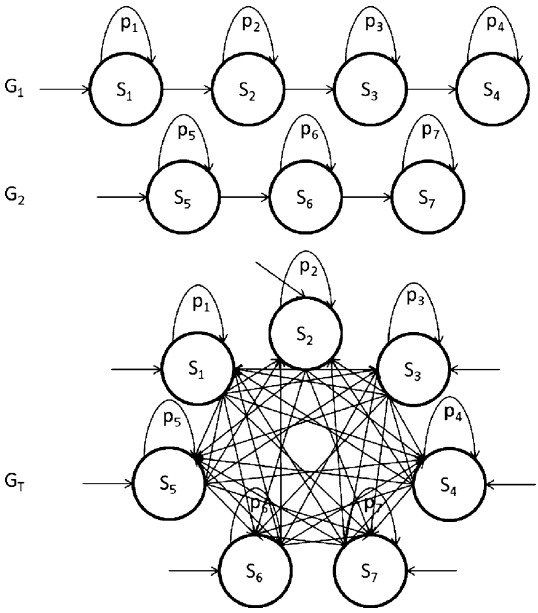


Fig. 6.15 Construction of an adaptive threshold model for HMM based frameworks

hoods of class labels are compared to determine the gesture type. Therefore, the non-gesture model should act like a threshold, which the likelihood of other models must exceed.

Lee and Kim proposed an adaptive threshold model for HMM-based frameworks, which can be constructed from the trained models in the system [22]. Construction of a threshold model for a system of two gestures is given in Fig. 6.15. Here, the threshold model G_T consists of a copy of each state in other HMMs. Each copy retains its self-transition probability p_i, and divides the rest of the probability mass $1 - p_i$ evenly among transitions to every other state. Emission probabilities are also retained. Finally, each state in the model is equally likely to be the initial state. If this model is the most likely to produce the candidate sequence, it is labeled as a non-gesture.

The idea behind this model is that the threshold model is similar to every gesture in the system, but ignores their intrinsic dynamics, allowing the actual gesture model to attain higher likelihoods, when evaluating a sample from its own class.

6.4.3.2 CRF Based Methods

The same idea can be adapted to CRFs to construct a CRF based threshold model [37]. Like every other gesture in a CRF based framework, the threshold model is represented with a single class label. First, the CRF is trained with the gestures without considering the threshold model. Then, the new class label corresponding to the threshold model is constructed, and every other state is linked to this state in the network. Rejection occurs if after applying the Viterbi algorithm the state variable

is assigned this new label. For a detailed explanation of the construction process, see [37].

6.4.3.3 IOHMM Based Methods

As IOHMMs are quite similar to HMMs, an adaptive threshold can be constructed from the states of IOHMM in the same manner. The only difference is that the self-transition probability p_t of a state is not fixed and has to be calculated at each time step.

In another approach, a threshold model with only a few hidden states, but with very complex local models is trained with all the positive samples, instead of being constructed. The rest of the IOHMMs are trained starting with a single hidden state until their parameters converge. Then their complexity is gradually increased until they are more likely to produce their dataset than the threshold model. This also solves the model selection problem [20].

6.5 Applications

In this section we will give two hand gesture recognition application examples. The first application is SignTutor, which teaches sign language, evaluates performances and gives feedback for improvements. The second application is Sign Tracking and Recognition System (STARS), which enables the user to control third party applications with 2D or 3D hand gestures.

6.5.1 SignTutor: An Interactive System for Sign Language Tutoring

SignTutor is an interactive system, which automatically evaluates users' signing and gives multimodal feedbacks to guide them to improve their signing [1]. SignTutor allows users to practice instructed signs and to receive feedback on their performance. The system automatically evaluates sign instances by multimodal analysis of the hand and head gestures, being one of the first systems that combines manual and non-manual information together for sign recognition.

6.5.1.1 System and Modules

SignTutor aims to teach the basics of the sign language interactively. The advantage of SignTutor is that it automatically evaluates the student's signing and enables auto-evaluation via visual feedback and information about the goodness of the performed

Fig. 6.16 SignTutor GUI: training, practice, information, synthesis panels

sign, through various feedback modalities: a text message, the recorded video of the user, the video of the segmented hands and/or an animation on an avatar.

One of the key factors of SignTutor is that it integrates hand motion and shape analysis together with head motion analysis to recognize signs that include both hand gestures and head movements.

Figure 6.16 shows the graphical user interface of Sign Tutor. The system follows three steps for teaching a new sign: training, practice and feedback. In the training phase, the users select a sign from the list of possible signs and watch the corresponding video until they are ready to practice. In the practice phase, users are asked to perform the selected sign and their performance is recorded by a single webcam. SignTutor analyzes the hand motion, hand shape and head motion in the recorded video and compares it with the selected sign, to give feedback to the user.

The SignTutor system consists of a face and hand detector stage, followed by the analysis stage, and the final sign classification stage. The critical part of SignTutor is the analysis and recognition sub-system which receives the camera input, detects and tracks the hand, extracts features and classifies the sign.

For each hand, four hand motion features (position and velocity in vertical and horizontal coordinates), and 19 hand shape features are extracted. On top of these, the relative position of the hands with respect to the face center of mass for each hand, normalized by the face height and width, are also extracted. The classification module receives all features to train the HMM models. Each HMM is a continuous four-state left-to-right model and is trained for each sign, using the Baum–Welch algorithm.

At the classification phase, a sequential likelihood fusion method for combining manual and non-manual parts of the sign is used [1]. The strategy uses the fact that there may be similar signs which differ slightly and cannot be classified accurately in an "all signs" classifier. The sequential fusion method is based on two successive classification steps: In the first step, an inter-cluster classification is carried out, and

Table 6.2 System accuracy. Signer-independent test results

	Sbj1	Sbj2	Sbj3	Sbj4	Sbj5	Sbj6	Sbj7	Sbj8	Average
$\mathbf{HMM}_M$	66.32	77.9	60.00	71.58	57.9	81.05	52.63	70.53	67.24
$\mathbf{HMM}_{M\&N}$	73.68	91.58	71.58	81.05	62.11	81.05	65.26	77.89	75.53
$\mathbf{HMM}_{M\&N} \Rightarrow \mathbf{HMM}_N$	85.26	76.84	77.89	89.47	63.16	80.00	88.42	75.79	**79.61**

in the second step intra-cluster classifications are performed. Potential sign clusters which are similar in manual gestures, but differ in non-manual signals, are automatically discovered. The sequential likelihood fusion idea is further extended in [2] via a belief formalism to detect the level of uncertainty in the decisions of the first stage classifier and to determine the sign clusters for the second stage.

6.5.1.2 Evaluation

Three different methods are compared to evaluate the classification accuracy of the system: (1) Classification by using only manual information (HMM_M), (2) Feature fusion on manual and non-manual information ($\text{HMM}_{M\&N}$), and (3) Sequential fusion, the two-tier cluster-based sequential fusion strategy ($\text{HMM}_{M\&N} \Rightarrow \text{HMM}_N$).

The accuracy is reported on a signer-independent protocol, with leave-one-subject-out cross validation. The results are given in Table 6.2.

The need for the usage of the head features can be deduced from the high increase of the overall accuracy with the contributions of non-manual features. With $\text{HMM}_{M\&N}$, the accuracy increases to 75.5% as compared to the accuracy of HMM_M, 67.2%. Sequential fusion methodology increases the accuracy by an additional 4% in comparison to the feature level fusion.

6.5.2 STARS: Sign Tracking and Recognition System Using IOHMMs

STARS is an IOHMM-based framework aimed to enable a system to seamlessly replace mice and keyboard actions with hand gestures to control generic PC applications. Users can manipulate target applications by replacing the mouse with manipulative hand gestures, and by giving high level commands to the target application with communicative gestures. In this section, we give a brief overview of the system. More details can be found in [20].

6.5.2.1 System and Modules

STARS allows training of 2D and 3D communicative gestures and automatically distinguishes these from manipulative gestures and unintentional hand motions in

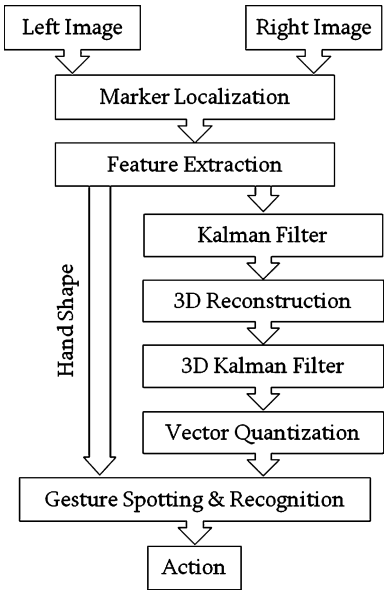


Fig. 6.17 Block diagram of STARS

continuous streams in real time. The system does not rely on special hardware and requires a single webcam for 2D gestures, and two webcams for 3D gestures. The system expects the user to wear smoothly colored gloves. Gestures are modeled using both hand motion and hand shape, without using predefined shape templates, allowing users to define gestures with no restrictions on the complexity. Upon recognition of a gesture, STARS triggers a sequence of operating system events associated with the gesture, thus controlling the target application.

STARS does not assume specific signals to indicate the start or end of gestures. The user can freely perform communicative gestures among a stream of manipulative gestures or unintentional movements, which are continuously tracked to spot communicative gestures. The rest of the stream is interpreted as manipulative gestures according to the context.

The flowchart of STARS is given in Fig. 6.17. The system estimates the color of the marker in the HSV color space via motion detection methods in the marker registration phase. The cameras continuously retrieve stereo images of the scene and a color based region growing method is used to detect and localize the marker.

As motion features, the centroids of the hand images are extracted from each frame. These are smoothed via a Kalman filter to form a 2D trajectory for the hand. The 3D trajectory is then reconstructed from the filtered 2D trajectories and the camera calibration matrices using a least squares approach. The camera calibration matrices are automatically estimated via a calibration tool. The resulting 3D trajectory is then smoothed via a 3D Kalman filter to eliminate reconstruction noise.

The shape descriptors extracted in the feature extraction phase are used to form the input sequence to train and evaluate the IOHMMs. The system continuously tries to spot communicative gestures among the observation sequence using the input

Table 6.3 Comparison of recognition success rates with different methods

Method	MoG-HMM	G-HMM	IOHMM	HCRF-3S	HCRF-7S	HCRF-12S
Dataset 1	84.6%	75.0%	**97.6%**	92.8%	95.1%	96.9%
Dataset 2	–	–	**94.1%**	89.8%	93.6%	**94.1%**

sequence using an adaptive threshold model. If it recognizes a gesture, it fires the corresponding event listed in the configuration file and retrieves the next frames from the cameras.

IOHMMs used in STARS have left–right architecture. The input sequence is chosen to be the hand shape, as it consists of high dimensional continuous data. The hand shape information consists of seven Hu moments and orientations extracted from each hand image. Orientation is the angle of the major axis of the hand, which can be calculated using image moments. A variable indicating normalized time is also included in the input sequence.

Hand motion information consists of the spherical angles of the 3D velocity vector of the hand in each frame, which is quantized into 15 symbols. Since both observations and the states are discrete, a classifier is chosen as the local model that takes a real valued vector as input and produces the likelihood for discrete states. STARS employs MLP classifiers as local models, which are suitable for this scenario.

The outputs of the MLPs correspond to the transition or emission probabilities of a single state and therefore should be normalized, which can be ensured via a softmax function at the output layer.

6.5.2.2 Evaluation

To test the recognition efficiency of IOHMMs, two pre-segmented communicative gesture datasets are used. A total of 1735 samples were performed by three users. The first dataset consists of ten isolated gestures that differ both in motion and in appearance. The second dataset consists of the same gestures, and also includes ten new gestures that are analogous to the first ten gestures in motion, but are different in appearance. 5×2 cross validation method is used to estimate optimal system parameters, as well as recognition and spotting efficiencies. For each cross validation test, half of the samples are used for training and the rest is used for testing.

The recognition rate of the IOHMM-based framework is compared to purely generative HMMs and purely discriminative HCRFs. MoG-HMM models observations with mixture of Gaussians and G-HMM models them with a single Gaussian. HCRF-xS uses x states for the HCRF, where x can be 3, 7 or 12. The comparison of recognition rates is given in Table 6.3. Since Mog-HMM and G-HMM perform considerably worse than IOHMMs and HCRFs even on the simpler Dataset 1, they are not tested on Dataset 2.

6.6 Summary

Hand gestures are important components of body language and are widely used both consciously and unconsciously in human communication. With the ongoing development of gesture recognition techniques, gestures are now also a part of human–computer communication. In this chapter, following a detailed description of gestures and how they are used in human communication, we present a general framework that covers main aspects of gesture recognition, applicable to many different gesture types. The framework covers each step necessary for performing gesture recognition: initial preprocessing steps, tracking, feature extraction, and recognition. We give an overview of state-of-the-art gesture recognition and spotting techniques for both isolated and continuous gesture recognition, with a focus on graphical models.

With the advancement of camera and sensor technology, human movements are captured in a more precise and robust way, making some of the steps that are described in the framework straightforward. The future challenges of gesture recognition mainly exist in two dimensions. First, there is still a lot of room for improvement on gesture modeling techniques to increase the accuracy and robustness of recognition systems. Second, to be able to use these systems in people's daily lives, a more intelligent analysis is required to differentiate between a person's gesturing to address the system and gesturing during a conversation.

6.7 Questions

1. What is a static gesture? What is a dynamic gesture?
2. In what kind of settings can the 3D coordinates of a hand trajectory be extracted?
3. Can the basic particle filter tracking approach be used to track multiple identical objects? Discuss several extensions to the particle filter for tracking the two hands and the face in natural settings.
4. Discuss advantages and disadvantages of using high level or low level shape features for pose estimation.
5. Prove that HMMs implicitly model durations with geometric distribution.
6. What is a real-time system? For the presented gesture recognition models (HMM, CRF, their variants, etc.), discuss their applicability to a real-time system.
7. Suppose you have a robust and accurate gesture recognition system that works with a single camera. Think about a potential application that can use this system.

6.8 Glossary

- *Communicative gestures*: Gestures that have an inherent communicational purpose, usually accompanied by speech.

- *Continuous gesture recognition*: The system attempts to recognize a continuous stream of gestures. The gestures are performed continuously, one after the other, with possible pauses or out-of-vocabulary gestures in between.
- *Discriminative model*: Models that learn to condition their parameters on features of observations. These models learn to separate different classes, and cannot be used to synthesize new samples.
- *Generative model*: Models that can be used to synthesize new samples from the class. These models learn the probability density of observations conditioned on the class label.
- *Gesture spotting*: The segmentation of gestures from a continuous stream of samples. The system finds the start and end point of each known gesture (in-vocabulary gesture) performed in the stream.
- *Isolated gesture recognition*: The system knows when a gesture starts and ends, and attempts to recognize the gesture performed in this interval.
- *Manipulative gestures*: Gestures that are used to act on objects, such as rotation, grasping, etc.

References

1. Aran, O., Ari, I., Benoit, A., Campr, P., Carrillo, A.H., Fanard, F.-X., Akarun, L., Caplier, A., Sankur, B.: Signtutor: An interactive system for sign language tutoring. IEEE Multimed. **16**(1), 81–93 (2009)
2. Aran, O., Burger, T., Caplier, A., Akarun, L.: A belief-based sequential fusion approach for fusing manual and non-manual signs. Pattern Recognit. **42**(5), 812–822 (2009)
3. Aubert, G., Barlaud, M., Faugeras, O., Jehan-Besson, S.: Image segmentation using active contours: Calculus of variations or shape gradients. SIAM J. Appl. Math. **63**(6), 2128–2154 (2003)
4. Bengio, Y., Frasconi, P.: Input-output HMM's for sequence processing. IEEE Trans. Neural Netw. **7**(5), 1231–1249 (1996)
5. Black, M.J., Jepson, A.D.: Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In: Proceedings of the 4th European Conference on Computer Vision (ECCV '96), vol. I, pp. 329–342. Springer, London (1996)
6. Bobick, A., Davis, J.: Real-time recognition of activity using temporal templates. In: Proceedings of the Workshop on Applications of Computer Vision (1996)
7. Clarke, T.A., Fryer, J.G.: The development of camera calibration methods and models. Photogramm. Rec. **16**(91), 51–66 (1998)
8. Cornett, R.O.: Cued speech. Am. Ann. Deaf **112**, 3–13 (1967)
9. Cui, Y., Swets, D.L., Weng, J.J.: Learning-based hand sign recognition using shoslif-m. In: Proceedings of the Fifth International Conference on Computer Vision, ICCV '95, p. 631, Washington, DC, USA. IEEE Comput. Soc., Los Alamitos (1995)
10. Doucet, A., De Freitas, N., Gordon, N. (eds.): Sequential Monte Carlo Methods in Practice. Springer, Berlin (2001)
11. Flusser, J., Suk, T.: Rotation moment invariants for recognition of symmetric objects. IEEE Trans. Image Process. **15**, 3784–3790 (2006)
12. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Trans. Inf. Theory **IT-8**, 179–187 (1962)
13. Hu, W., Tieniu, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern. **34**, 334–352 (2004)

14. Isard, M., Blake, A.: Condensation—conditional density propagation for visual tracking. Int. J. Comput. Vis. **26**(1), 5–28 (1998)
15. Just, A., Bernier, O., Marcel, S.: HMM and IOHMM for the recognition of mono- and bi-manual 3D hand gestures. IDIAP-RR 39, IDIAP, 2004
16. Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Basic Eng. **82**, 35–45 (1960)
17. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. Int. J. Comput. Vis. **1**(4), 321–331 (1988)
18. Kendon, A.: Current issues in the study of gesture. In: Nespoulous, J.L., Peron, P., Lecours, A.R. (eds.) The Biological Foundations of Gestures: Motor and Semiotic Aspects, pp. 23–47. Erlbaum, Hillsdale (1986)
19. Kendon, A.: Gesture. Cambridge (2004)
20. Keskin, C., Akarun, L.: STARS: Sign tracking and recognition system using input–output HMMs. Pattern Recognit. Lett. **30**, 1086–1095 (2009)
21. Khotanzad, A., Hong, Y.H.: Invariant image recognition by Zernike moments. IEEE Trans. Pattern Anal. Mach. Intell. **12**, 489–497 (1990)
22. Lee, H.-K., Kim, J.H.: An HMM-based threshold model approach for gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. **21**(10), 961–973 (1999)
23. Marcel, S., Bernier, O., Viallet, J.-E., Collobert, D.: Hand gesture recognition using input-output hidden Markov models. In: FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2008, p. 456, Washington, DC, USA. IEEE Comput. Soc., Los Alamitos (2000)
24. McNeill, D., Levy, E.: Conceptual representations in language activity and gesture. In: Jarvella, R., Klein, W. (eds.) Speech, Place, and Action. Wiley, New York (1982)
25. Mish, F.C.: The Merriam-Webster Dictionary. Merriam-Webster, Chicago (1997)
26. Mitiche, A., Sekkati, H.: Optical flow 3d segmentation and interpretation: A variational method with active curve evolution and level sets. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1818–1829 (2006)
27. Morency, L.-P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2007)
28. Oka, K., Sato, Y., Koike, H.: Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems. In: Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Washington D.C., US, p. 429 (2002)
29. Pavlovic, V., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 677–695 (1997)
30. Quek, F.K.H.: Eyes in the interface. Image Vis. Comput. **13**(6), 511–525 (1995)
31. Riviere, J., Guitton, P.: Real time model based tracking using silhouette features. In: Proceedings of RFIA, Toulouse, France (2004)
32. Stokoe, W.C.: Sign language structure: An outline of the visual communication systems of the American deaf. Stud. Linguist., Occas. Pap. **8** (1960)
33. Triesch, J., von der Malsburg, C.: Classification of hand postures against complex backgrounds using elastic graph matching. Image Vis. Comput. **20**(13–14), 937–943 (2002)
34. Ueda, E., Matsumoto, Y., Imai, M., Ogasawara, T.: A hand-pose estimation for vision-based human interfaces. IEEE Trans. Ind. Electron. **50**(4), 676–684 (2003)
35. Wang, S.B., Quattoni, A., Morency, L.-P., Demirdjian, D.: Hidden conditional random fields for gesture recognition. In: CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp. 1521–1527. IEEE Comput. Soc., Los Alamitos (2006)
36. Wu, Y., Huang, T.S.: Hand modeling, analysis, and recognition for vision based human computer interaction. IEEE Signal Process. Mag. **21**(1), 51–60 (2001)
37. Yang, H.-D., Sclaroff, S., Lee, S.-W.: Sign language spotting with a threshold model based on conditional random fields. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 1264–1277 (2009)

Chapter 7
Semantics of Human Behavior in Image Sequences

Nataliya Shapovalova, Carles Fernández, F. Xavier Roca, and Jordi Gonzàlez

7.1 Introduction

Images, video, or multimedia are words that currently sound familiar to the majority of people. An enormous amount of video is daily produced by surveillance systems or broadcast companies, but also by travelers who want to keep the memories of new places visited. Considering such an amount of multimedia data, its analysis, processing, indexing and retrieval is a truly challenging task.

From this point of view, automatic image and video scene understanding is of importance. The main task of scene understanding is to give a semantic interpretation to observed images and video. In other words, scene understanding tries to bridge the semantic gap between the low-level representation of images and videos and the high-level, natural language description that a human would give about them [49]. In our work, only those scenes including humans will be considered, as they are by far the most common ones in the studied domains. Nevertheless, emphasis will be on the interaction of these humans with their environment, since such a global approach will be proven to provide more information than separate analysis.

N. Shapovalova (✉) · C. Fernández · F.X. Roca · J. Gonzàlez
Departament de Ciències de la Computació and Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain
e-mail: shapovalova@cvc.uab.es

C. Fernández
e-mail: perno@cvc.uab.es

F.X. Roca
e-mail: xavir@cvc.uab.es

J. Gonzàlez
e-mail: poal@cvc.uab.es

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_7,

Human Event Understanding (HEU) is important in a number of domains including video surveillance, image and video search, and human–computer interaction, among others [35]. The importance of these fields is briefly discussed here.

Video surveillance is a growing research area that has recently gained exceptional importance due to increased security demands in public locations. CCTV cameras populate airports, railway stations, shopping malls, and subways. Understanding and interpreting video scenes obtained from these cameras would greatly assist in the detection of dangerous situations in real time and, in consequence, increase security. However, automated surveillance still has a long way to go, and in most of the cases the analysis is conducted by human operators. Fully automatic surveillance systems would facilitate the understanding of dangerous occurrences, in order to prevent undesired consequences [46].

Image and video search have grown in importance during the last years. The enormous amount of multimedia data owned by broadcast companies and produced on the Internet demand advanced techniques of image and video retrieval. Automatic indexing and annotation of video and image data would significantly ease the search of multimedia contents according to the needs of the user.

As computers and smart devices are being progressively involved in our everyday life, the field of human–computer interaction is facing new challenges day by day. Human–computer interaction has evolved from the use of rigid peripherals like keyboard or mouse toward a new form of smart interaction through the recognition of basic actions and gestures, e.g. via Wii and Kinect sensors. New types of human-like interaction intend to be more transparent and natural to the users.

This chapter proposes a novel framework for the understanding of human events in their context, which benefits many applications in the aforementioned domains. The chapter is organized as follows: in Sect. 7.2 we analyze the current state-of-the-art on scene understanding, while focusing on the main techniques and taxonomies for HEU. Section 7.3 shows an application for activity recognition from multiple cues and provides automatic behavior interpretation by means of Situation Graph Tree (SGTs). Finally, the main ideas of this work are summarized and some final conclusions are drawn.

7.2 State of the Art

HEU has become an active and growing research area due to its inherent complexity and increasing number of potential applications. Different aspects of HEU have already been covered, among them human motion analysis, activity recognition, and behavior modeling. In this section we analyze two main paradigms of HEU, which follow bottom-up and top-down approaches, respectively. In addition, a taxonomy of human events will be discussed.

7.2.1 Bottom-up Approaches in Behavior Understanding

Bottom-up approaches analyze and interpret human behavior based on low-level features of the video or image scene. While many works take into consideration uniquely human motion and appearance, recent work in behavior understanding go toward the analysis of humans in their context. In appearance-based approaches, the appearance and the shape or pose of a person are the most common cues for human behavior understanding. Appearance is usually presented using the Bag of Words (BoW) technique, which has been proved to be successful for object and scene recognition. Pose information is usually extracted from the shape of human bodies, and template matching techniques are used to compare human poses. Delaitre et al. [9] analyze local and global approaches for the application of BoW to action recognition from still images. They show that action recognition performance can be improved by combining human and context appearance representations. The BoW technique was extended to histogram of oriented rectangles [19] in order to capture spatial distributions of visual words, and to extract these features only from the silhouette of a person. For describing the shape of a human body, Dalal and Triggs introduced the Histogram of Oriented Gradients (HOG) technique [8], which produced excellent results in human detection and was therefore extended to action recognition. Ikizler-Cinbis et al. [21] apply template matching based on the HOG descriptor to learn human actions from Internet images. Wang et al. [52] apply a deformable template matching scheme to cluster human actions based on the shape of their poses. Appearance-based approaches are particularly important when motion features are not available, e.g. in still images, or not significant, e.g. in actions like *reading*.

The analysis of human motion is also important for the understanding of human behavior. Direct motion recognition techniques attempt to infer the behavior only from the motion cues, and disregarding the information of the body about appearance and pose. Optical flow and trajectories are the main features extracted from motion. Optical flow is a low-level feature calculated from the motion of individual pixels, while trajectories are usually computed for the whole human body. Polana and Nelson [43] use repetitive motion as a strong cue to discriminate between different activities. Noceti et al. [40] accomplish behavior understanding by analyzing trajectories over long-time observations, to discover behavior patterns in a particular scenario.

The mixed analysis of static and motion cues is also employed in many works. The temporal template matching algorithm introduced by Bobick et al. [5] combines shape and motion cues to capture different movements in aerobics exercises. The desire to combine motion and appearance features led to an extension of detectors and descriptors of spatial salient points for the spatiotemporal domain [32].

While human motion and appearance provide important clues for behavior understanding, very often they are not sufficient for analyzing and modeling complex human behavior, like riding a horse, phoning or even going shopping. In these cases, contextual knowledge such as the type of scene (indoor, outdoor, supermarket) and

the objects in that scene (horse, phone, goods) should be exploited. Scene classes and their correlation with human activities have been studied in [32]. It is illustrated that incorporating scene information increases the recognition rate of human activities. Kjellström et al. [23] and Gupta et al. [17] combine activity recognition and object recognition into one framework, which significantly improves the final output. Alternatively, Li and Fei-Fei [27] try to answer the 3 W's: *what?* (event label), *where?* (scene label) and *who?* (list of objects). They build a generic model that incorporates different levels of information such as event/activity, scene, and object knowledge.

7.2.2 Top-down Modeling of Behavior Understanding

Algorithms for detection and tracking have been greatly improved during the last years, and although there are still many issues to cope with—e.g., appearance variability, long-term occlusions, high-dimensional motion, crowded scenes—, robust solutions have been already provided that capture the motion properties of the objects in dynamic and complex environments [44, 45]. But to understand scenes involving humans and showing semantic developments, we need to consider the notion of *event*. An event is regarded as a conceptual description summarizing the contents of a development, that description being closely related to real world knowledge.

The recognition of events in video sequences has been extensively tackled by the research community, ranging from simple actions like walking or running [39] to complex, long-term, multi-agent events [25]. The recognition of complex events and behaviors is becoming more and more a hot topic of the literature in this field. Three main approaches are generally followed toward the recognition of non-basic events: pattern recognition methods, state models, and semantic models.

First of all, the modeling formalisms used include many diverse techniques for pattern recognition and classification, such as neural networks and self-organizing maps [55], K-nearest neighbors (kNN) [33], boosting [50], support vector machines (SVM) [39], or probabilistic or stochastic context-free grammars (CFG) [22, 36]. In addition, the statistical modeling of Markov processes is tackled using state models, such as hidden Markov Models (HMM) [41, 53], Bayesian networks (BN) [18], or dynamic Bayesian networks (DBN) [1]. These have been often used when pursuing the recognition of actions and activities.

Nevertheless, the high complexity found in the domain of video sequences stresses the need to employ more explicit semantic models. The interpretation of activities depends strongly on the locations where the events occur, e.g., traffic scenes, airports, banks, or border controls in the case of surveillance, which can be efficiently exploited by means of conceptual models. Therefore, it is reasonable to make use of domain knowledge in order to deal with uncertainty and evaluate context-specific behaviors. Thus, a series of tools based on symbolic approaches have been proposed to define the domain of events appearing in selected environments, e.g., those based on conceptual graphs or conditional networks.

Starting from the early use of Finite State Automata [18, 20] and similar improved symbolic graphs [31], researchers have increased the expressivity of the models, so that they can manifest precise spatial, temporal, and logical constraints. Such constraints have ended up complementing each other in multivariate analysis, e.g., by means of temporal constraint satisfaction solvers applied over symbolic networks [51]. More recently, Nagel and Gerber [38] proposed a framework that combines SGTs with Fuzzy Metric Temporal Horn Logic (FMTL) reasoning, in order to generate descriptions of observed occurrences in traffic scenarios. Extensions of Petri Nets have also been a common approach to model multi-agent interactions, and used as well for human activity detection [2]. Petri Nets are graphical models represented by a directed bipartite graph that contains nodes drawn as circles (places) and bars or boxes (transitions), in which the state of the net gets to be defined by the number of tokens found in each place. Petri Nets enforce temporal and spatial distance relations as transition enabling rules, and can be also used to model concurrency and partial ordering among sub-events. Some other recent approaches have employed symbolic networks combined with rule-based temporal constraints, e.g. for activity monitoring applications [15].

7.2.3 Taxonomy of Human Events

The automatic understanding of human behavior has been addressed by many authors [4, 14, 16, 18, 37], who have typically decomposed this problem into different levels of knowledge. Typically, taxonomies of human events are constructed in order to capture and organize these levels and facilitate procedural solutions.

Nagel in his work on machine perception of motion presented a taxonomy including five stages: *change*, *event*, *verb*, *episode*, and *history*, where a *change* refers to a discernible difference in a sequence; an *event* is a change that is considered as a primitive of a more complex description; a *verb* defines some activity; an *episode* is a complex motion which may involve several actions; and a *history* is an extended sequence of related activities [37]. This taxonomy is oriented to provide a high-level description in natural language.

Alternatively, [4] approached the same problem from the point of view of motion, using another taxonomy: *movement*, *activity*, and *action*. This taxonomy reflects the analysis of human motion from the levels of semantics required for interpretation: *movement* is the most atomic primitive, requiring no semantic knowledge to be incorporated; a movement is often addressed using geometric techniques. An *activity* refers to the sequence of movements or states, where the only real knowledge required is the statistics of the sequence. *Actions* are larger scale events, which typically include interaction with the environment and causal relationships. Bobick's taxonomy is useful from a low-level perspective, although more complex analysis is required to bridge perception and cognition.

Another approach is presented in [18]. Their terminology distinguishes between *single* and *complex events*. A *single event* refers to anything that can be captured from a single frame, e.g. static postures, actions, or gestures. A *complex event* corresponds to a linearly ordered time sequence of simple events or other complex events. In addition, *single/multiple thread events* are distinguished according to the number of actors they involve. This approach is suitable for generic solutions, although it does not help to develop explicit implementations.

In [16], the concept of human event is decomposed into *movement*, *action*, *activity*, and *behavior*. The emphasis of this taxonomy is on temporal information of human motion. A *movement* represents a change of human posture between consecutive frames. An *action* is a temporal series of human movements which can be denoted with a verb label such as running, jumping, turning left, laughing. An *activity* is a sequence of two or more human actions, plus the transition between them, e.g. the path followed by a human in the scene. A *behavior* refers to one or more activities which acquire their meaning in a specific context, like crossing the road or giving a concert. The main advantage of the presented terminology is that it fills the gap between high-level semantic description of [37] and low-level motion description of [4]. However, this taxonomy does not provide a scene description terminology; therefore it is impossible to capture interactions between humans and environment.

The work presented by [14] solves this problem by defining an ontological organization of concepts that includes entities—comprising agents, objects, and locations—and events. The taxonomy of events has three semantic levels: *status*, i.e., actions or gestures that can be independently analyzed for agents in isolation, e.g. running or turning; *contextualized events*, which considers interactions among agents, objects, and locations, e.g. picking up an object, meeting someone; and *behavior interpretations*, which use prior domain assumptions to interpret complex interactions over time, like abandoning objects, stealing, giving way, or waiting to cross.

In our work we extend the taxonomy of [16] with the concept of *entity* from [13], reusing this concept as a low-level unit in our taxonomy of events. The resulting taxonomy is then *entity*, *movement*, *action*, *activity*, and *behavior*, where we have the following.

- *Entity* is defined as something that has its own distinct existence (physical or abstract); it can be perceived, known, or inferred. Moreover, we can discriminate between several types of entities:
  - Active entities (agents, actors): entities that can move by themselves; among them are humans, cars, animals.
  - Passive entities (objects): entities which can only be moved by active entities; among them chair, guitar, bicycle.
  - Background entities (context): global environment, location where the agent is observed; among them indoor, outdoor, road, sky, sea.
- *Movement* is a change of agent pose and/or location between consecutive frames; no semantic knowledge is required.

- *Action* is a semantically defined set of movements of the agent, e.g. running, jumping. Low-level semantics is used; interpretation is restricted to the agent itself.
- *Activity* is defined as an action of the agent in a particular context, where all the entities of the scene are taken into account. Here we consider not running, but playing football or kicking the ball, and not waving hands, but catching a ball. In other words, this level of semantics can be defined as all the semantics that can be extracted from the current context.
- *Behavior* is a reaction of an agent to the particular *situations*, i.e., longer observations in context. We want to understand *why* the agent is doing something, so we should acknowledge past situations (the cause of the action) or the final objective of the agent. For example, for an agent hitting a vending machine we may imply that s/he wants to retrieve something that got stuck in the machine (knowing the past), or that s/he wants to damage the machine to steal something (objective). The key point of interpretations at this level is not only context, but also temporal and causal relationships. This level of understanding goes from image understanding to video understanding.

7.2.4 Human Event Datasets and Benchmarks

There are many public datasets for human event recognition and detection. The main idea of these datasets is to provide images or image sequences (video) and ground truth (e.g. event labels for every image frame) in order to evaluate performance of human event recognition algorithm. The seven popular datasets, which are currently used by most of the approaches are: KTH [47], Weizmann [3], Hollywood 2 [24], UCF50 [29], TRECVID [48], Sports [17], and PASCAL [11] (see Table 7.1). They can be categorized according to different criteria. First, whether the dataset contains temporal information (KTH, Weizmann, Hollywood 2, UCF50, TRECVID, PASCAL) or not (Sports, PASCAL). In other words, human event recognition can be done from still images or from videos. Second, there are simplistic (KTH, Weizmann) and realistic datasets (Hollywood 2, UCF50, TRECVID, Sports, PASCAL). In simplistic datasets the human event is staged by actors in a controlled environment. In realistic datasets, the event is captured from the real world. It is evident that it is quite easy to detect and recognize human event in the simplistic datasets, while dealing with realistic datasets is quite challenging. Finally, it is important that the dataset not only has images/videos and ground truth (KTH, Weizmann, Hollywood 2, UCF50, Sports), but also evaluation metrics (TRECVID, PASCAL). The advantage of datasets with evaluation metrics is that they allow one not only to evaluate the performance of a particular algorithm, but also to do it in such a manner that it can easily be compared with other algorithms. Examples of human events from these datasets are illustrated in Fig. 7.1.

Table 7.1 Comparison of main datasets in Human Event Recognition

	Temporal information	Evaluation metrics	Number of events	Examples of events	Source of data
KTH [47]	✓	✗	6	walking, jogging, running, boxing, hand waving, hand clapping	Staged by actors
Weizmann [3]	✓	✗	10	run, walk, skip, jumping-jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, gallopsideways, wave-two-hands, wave-one-hand, bend	Staged by actors
Hollywood 2 [24]	✓	✗	12	AnswerPhone, DriveCar, Eat, FightPerson, GetOutCar, HandShake, HugPerson, Kiss, Run, SitDown, SitUp, StandUp.	Hollywood movies
UCF50 [29]	✓	✗	50	Biking, Diving, Fencing, Playing Guitar, Horse Race, Military Parade, TaiChi, Walking with a dog, etc.	YouTube videos
TRECVID [48]	✓	✓	7	PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, Pointing	Surveillance video
Sports [17]	✗	✗	6	batting (cricket), bowling (cricket), serve (tennis), forehand (tennis), serve (volleyball), shot (croquet)	Internet
PASCAL [11]	✗	✓	9	phoning, playing a musical instrument, reading, riding a bicycle or motorcycle, riding a horse, running, taking a photograph, using a computer, walking	Flickr

7.3 Methodology

7.3.1 Architecture

The problem of HEU is not trivial, since it requires analyzing both the humans and the context where these humans are observed. The interaction of entities in the

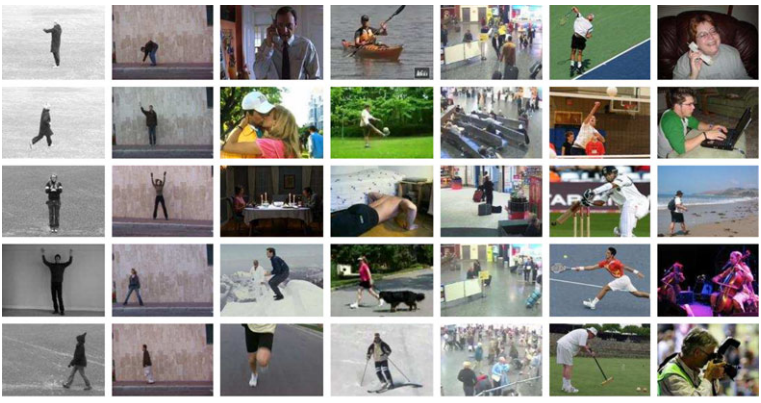


Fig. 7.1 Examples of human events from different datasets: (**a**) KTH [47], (**b**) Weizmann [3], (**c**) Hollywood [24], (**d**) UCF50 [29], (**e**) TRECVID [48], (**f**) Sports [17], and (**g**) PASCAL [11]

scene plays a crucial role in many cases, so a proper analysis and fusion of different sources of knowledge may be required:

- Motion cue: achieved by tracking and optical flow analysis
- Pose cue: achieved with pose estimation and reconstruction
- Interaction cue: achieved by analysis of human interaction with entities in the scene (interaction with objects and holistic scene analysis)

As we can observe in Fig. 7.2, general HEU demands analysis at different levels of semantics. Initially, given a video scene, salient features are extracted. At the same time, basic motion information is captured using frame differences. Then, entity detection, holistic scene analysis and motion tracking are performed. At this stage, basic scene knowledge like passive and background entities needs to be found. Next, given detection and tracking, complete information about active entities is extracted such as motion patterns and pose, which usually define an action model. At the following level, the interactions between scene entities are analyzed and human activity patterns are inferred.

A final, higher-level interpretation is accomplished in a top-down fashion using SGTs. These predefined models organize the possible world states as a tree of situations ranging from the most generic root (e.g., *an agent is active*) to the most particularized leaves (e.g., *theft detected*). At each level of the tree, a directed graph connects a starting situation to temporally subsequent ones, e.g., *walk-to-object*→*pick-up-object*→*carry-object*. Each situation evaluates specific conditions based on current visual observations such as positions, velocities and actions, and asserted situations define a traversal within the tree as we navigate to the most concrete interpretation for each time step. In the next section we illustrate this framework on an example that uses parts of the framework.

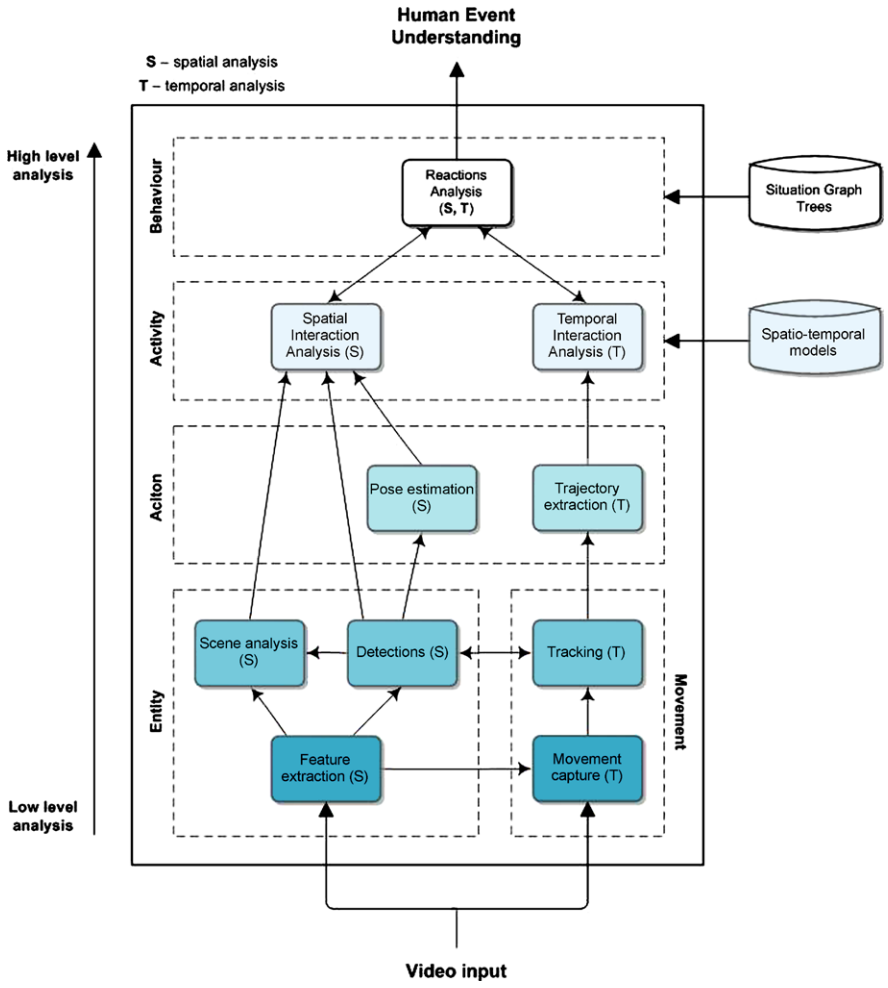


Fig. 7.2 Human event understanding

7.3.2 Activity Recognition

Activity recognition is an intermediate process in HEU that requires analysis of pose, scene, interactions, and motion. In this example we concentrate on activity recognition only from still images, discarding temporal information.

The framework of activity recognition is illustrated in Fig. 7.3(a) and is an important part of the global architecture. The proposed method has been applied to the Pascal VOC Action Classification dataset, which provides images with different human events and bounding boxes of humans performing an activity. We used the bounding box information for pose estimation, instead of using results from human detection. The typical input to the framework is shown in Fig. 7.3(b).

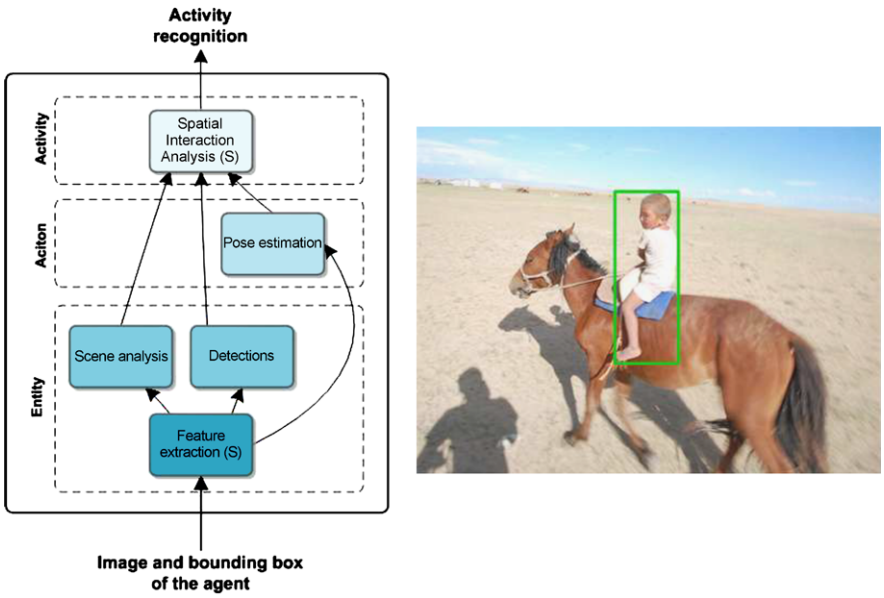


Fig. 7.3 (**a**) Framework for the activity recognition and (**b**) input to the framework

7.3.2.1 Entity Level

Feature Extraction and Image Representation

The low-level image analysis includes extraction of two kind of features: (i) appearance, and (ii) shape. The global scene *appearance* is represented by the Bag of Words (BoW) approach [6] based on Scale Invariant Feature Transform (SIFT) [30]. In order to capture *shape* information we apply a Histogram of Oriented Gradients (HOG) descriptor [8]. In the following, we will give more details about these three techniques.

Bag of Words

The main idea of BoW approach is to represent the image as a set of unordered elementary visual features (so-called words).

Constructing the BoW image representation involves the following steps [6]: (i) automatically detecting keypoints (salient regions in the image), (ii) computing local descriptors over these keypoints, (iii) quantizing the descriptors into words to form the visual vocabulary, (iv) and finding the occurrences in the image of each specific word in the vocabulary to build the histogram of words. Figure 7.4 schematically describes the four steps involved in the definition of the BoW model.

In our work we use a combination of sparse (*corner detector*, *blob detector*) and dense (*regular grid based*) detectors. Sparse detectors allow us to capture the most

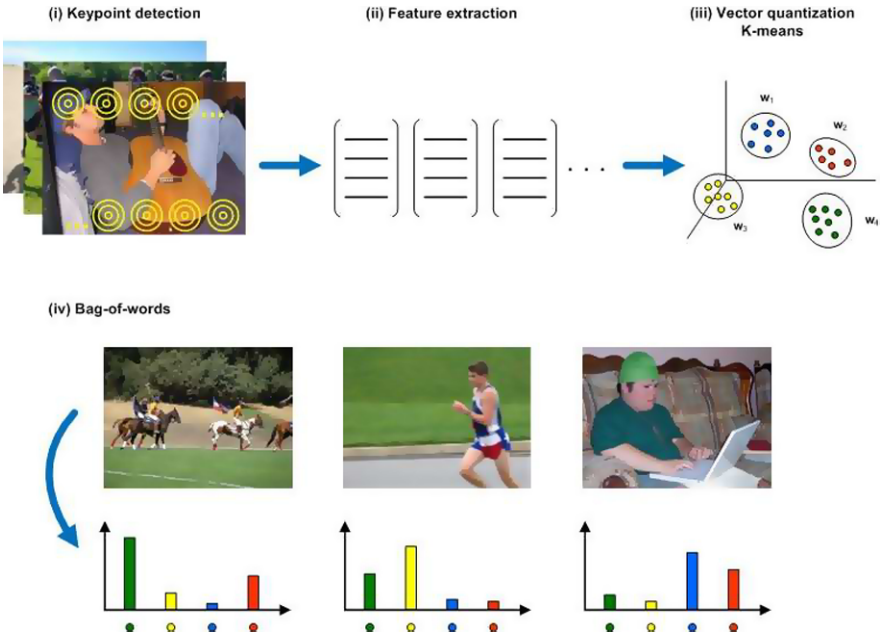


Fig. 7.4 The four steps to compute a BoW: (i–iii) obtaining the visual vocabulary by clustering feature vectors, and (iv) representing the image as a histogram of words

prominent salient points of the image; for example, the Harris–Laplacian detector (for corners) [34] and Laplacian-of-Gaussian detector (for blobs) [28] are commonly applied. On the other hand, the advantage of using a grid based detector is that it captures information in areas where sparse detectors are not able to extract enough keypoints. It is usually implemented by projecting a regular grid on the image and therefore segmenting the image into cells which are considered as image keypoints. Moreover, grids of different scales are used to enhance keypoint detection.

The next step is to represent all the keypoints using a local descriptor; for example, using SIFT as a local descriptor gives us the advantage of having local feature vectors which are invariant to image translation, scaling, and rotation. Therefore, given an image I, a collection of local feature vectors $X_I = [x_1, x_2, \ldots, x_M]$ is extracted, where M is the number of detected keypoints.

The process of building the vocabulary includes two main stages. From a set of images a pool of local feature vectors is first extracted, and then clustered into K number of words $W = [w_1, w_2, \ldots, w_K]$. For clustering, standard K-means technique is widely used, where each feature vector belongs to the cluster with the nearest mean:

$$W = \underset{W}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{x_j \in S_{w_k}} \|x_j - w_k\|^2, \tag{7.1}$$

and w_k is the mean of all the feature points that belong to the cluster S_{w_k}. The cluster centers are randomly initialized, and iteratively refined.

Lastly, local feature vectors X_I are assigned to the nearest word w_k from the vocabulary W with cluster membership indicators $U = [u_1, \ldots, u_M]$ [54]:

$$U = \underset{U}{\operatorname{argmin}} \sum_{m=1}^{M} \|x_m - u_m W\|^2, \tag{7.2}$$

where u_i is such that only one element of u_i is nonzero, and $|u_i| = 1$. The index of the nonzero element in u_m indicates which cluster the vector x_m belongs to. Finally, a representation of the image I is computed as a histogram of words H_I:

$$H_I = \frac{1}{M} \sum_{m=1}^{M} u_m. \tag{7.3}$$

Scale Invariant Feature Transform

SIFT is a widely used algorithm for describing keypoints of the image in such a manner that the output keypoint descriptor is highly distinctive and invariant to rotation, scale, illumination and 3D viewpoint variations [30]. SIFT properly describes the distribution of the gradient over an image patch centered in the detected keypoint.

First of all, it is important to determine the scale and the orientation of the keypoint. The scale s of the keypoint is defined at the keypoint detection stage. In order to find orientation, the following steps should be done in the Gaussian smoothed image $I_s(x, y)$ at a scale s, so that all computations are performed in a scale-invariant manner. Given the image sample $I_s(x, y)$, the gradient is computed using a 1D centered mask $[-1\,0\,1]$:

$$\begin{aligned} g_x(x, y) &= I_s(x + 1, y) - I_s(x - 1, y) \quad \forall x, y, \\ g_y(x, y) &= I_s(x, y + 1) - I_s(x, y - 1) \quad \forall x, y, \end{aligned} \tag{7.4}$$

where $g_x(x, y)$ and $g_y(x, y)$ denotes the x and y components of the image gradient.

Then the magnitude $m(x, y)$ and the orientation $\theta(x, y)$ of the gradient are computed as

$$\begin{aligned} m(x, y) &= \sqrt{g_x(x, y)^2 + g_y(x, y)^2}, \\ \theta(x, y) &= \tan^{-1} \frac{g_y(x, y)}{g_x(x, y)}. \end{aligned} \tag{7.5}$$

Next, the region around the keypoint is weighted with a Gaussian window and a corresponding magnitude, and a histogram of the region with 36 orientation bins is created. As a result, the peaks in the orientation histogram indicate the keypoint orientation.

Once the scale and orientation have been selected, the feature descriptor of the keypoint is built. First the image gradient magnitudes and orientations are sampled

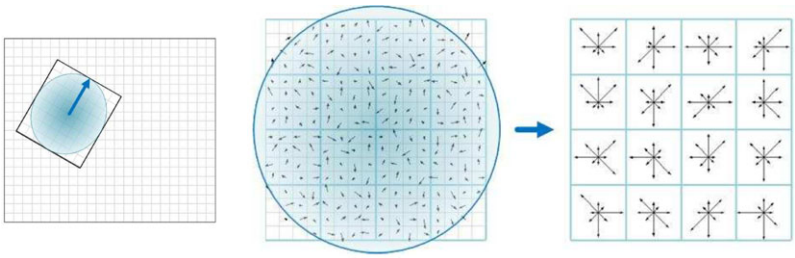


Fig. 7.5 Building a SIFT descriptor: (**a**) keypoint scale and rotation, (**b**) image gradient and final keypoint descriptor

around the location of keypoint in the given scale. In order to achieve orientation invariance, the coordinates of the descriptor and the gradient orientations are rotated relative to the keypoint orientation (see Fig. 7.5(a)). Just like before, the region around the keypoint is weighted by the gradient magnitude and by a Gaussian window. Then, the region is separated into four sub-regions, and for every sub-region the histogram with 8 bins for orientation is created (see Fig. 7.5(b)). Subsequently, in order to form a descriptor, these histograms are concatenated. This leads to a $4 \times 4 \times 8 = 128$ dimensional descriptor vector. Finally, the descriptor vector is normalized to unit length in order to make it invariant to affine illumination changes. To reduce the impact of non linear illumination changes, high values in the descriptor vector are thresholded, and the vector is again normalized.

Histograms of Oriented Gradients

The HOG feature is revealed to be very effective for object class detection tasks [8]. While BoW represents the image without considering spatial information, HOG is used when it is important to take into account spatial relationships between image regions. Since HOG allows to capture shape and structure, representing a region of interest (e.g. person, object) with HOG is more informative when compared to representing the whole image.

The computation of HOG is as follows. First, for each sub-sampled image $I(x, y)$, we compute the image gradient and then the magnitude $m(x, y)$ and orientation $\theta(x, y)$ of the gradient according to (7.4) and (7.5).

After that, a weighted histogram of orientations is computed for a certain square region that is called a cell. The histogram is computed by summing up the orientation of each pixel of the cell weighted by its gradient magnitude. The histogram of each cell is usually smoothed using trilinear interpolation, both in space and orientation. In order to account for changes in illumination and contrast, gradient strengths are locally normalized, which requires grouping the cells together into larger, spatially-connected blocks. The HOG descriptor is then found as the vector of the components of the normalized cell histograms given all of the block regions. These blocks typically overlap, meaning that each cell contributes more than once

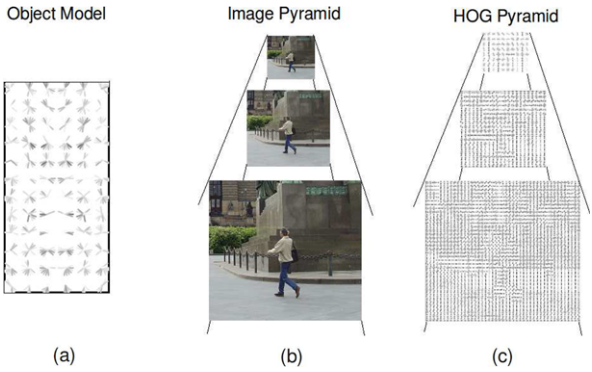


Fig. 7.6 Sliding windows components

to the final descriptor. Normalization of the block can be done using L1 or L2 norm:

$$\begin{aligned} \text{L1-norm:} \quad & v = \frac{v}{\|v\|_1 + \varepsilon}, \\ \text{L2-norm:} \quad & v = \frac{v}{\sqrt{\|v\|_2^2 + \varepsilon^2}}, \end{aligned} \tag{7.6}$$

where ε is a very small constant to prevent division by zero.

The use of orientation histograms over image gradients allows us to capture local contour information, which is the most characteristic information of a local shape. Translations and rotations do not influence HOG as long as they are smaller than the local spatial and orientation bin size, respectively. Finally, local contrast normalization makes the descriptor invariant to affine illumination changes which greatly improves detections in challenging lighting conditions.

Detections

In our work we consider the scene to be a configuration of objects that can be perceived in the scene. Therefore, the observation module is presented with an object detector. The applied object detector is based on the Recursive Coarse-to-Fine Localization (RFCL) [12, 42]. RCFL is an approach to speed up a sliding windows method. In sliding windows (see Fig. 7.6) an object model is scanned over a pyramid of features representing an image. The pyramid of feature is a set of matrices $F_s(x, y)$, where each element is an f-dimensional feature vector. Each matrix F_s is built from a smoothed and sub-sampled version $I_s(x, y)$ of the original image at a certain scale s. The object model for a linear classifier is an $h \times w$ matrix $M(x, y)$, where each elements is an f-dimensional weight vector. The response D_s, or score, of the object model centered at position (x, y) and scale s is defined as

$$D_s(x, y) = \sum_{\hat{x}, \hat{y}} M(\hat{x}, \hat{y}) \cdot F_s(\hat{x} + x - w/2, \hat{y} + y - h/2), \tag{7.7}$$

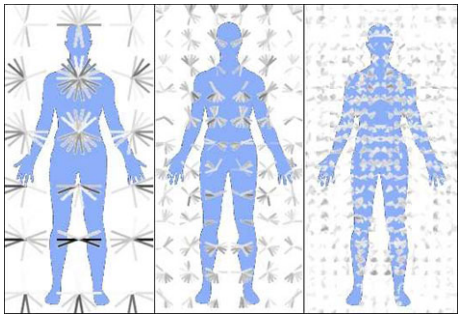


Fig. 7.7 HOG pyramid model M for the class 'person'. The low-resolution features ($d = 0$) give a general coarse representation of the human silhouette, while the high resolution ($d = 2$) focuses more on details

where $\hat{x} \in 0, 1, \ldots, w-1$, $\hat{y} \in 0, 1, \ldots, h-1$. In this way D_s is a pyramid of matrices of the same size as F_s, but where each element is a scalar that represents the response of the object model for the corresponding position and scale.

RCFL extends the sliding windows approach, and the object is searched not only in different scales, but also at different resolutions, from coarse to fine. The final score of the detector is now the sum of partial scores, one for each resolution. For this reason, the object model is a dyadic pyramid composed of l levels, where each level d is a matrix M_d of weight vectors. Initially, SVM is used for learning an object model M. An example of a 3-level pyramid model for the class 'person' in shown in Fig. 7.7, while an example of recursive localization refinement is shown in Fig. 7.8.

The computation of the partial score R_s^d for a resolution level d of the object model pyramid at a position (x, y) and scale s of the pyramid of features is then

$$R_s^d(x, y) = \sum_{\hat{x}_d, \hat{y}_d} M_d(\hat{x}_d, \hat{y}_d) \cdot F_s^d\big(\hat{x}_d + (x - w/2)s^d, \hat{y} + (y - h/2)2^d\big), \quad (7.8)$$

where $\hat{x}_d \in 0, 1, \ldots, w2^d - 1$, $\hat{y}_d \in 0, 1, \ldots, h2^d - 1$. When $d = 0$ this is exactly (7.7).

For each F_d^s, the search space is split into adjacent neighborhoods $\Delta(x, y)$. The neighborhood represents all the locations where an object can be found; therefore in RCFL the number of hypotheses corresponds to the number of neighborhoods. While locating the object, firstly, the position of the object Π for each (x, y) and scale s is defined as the location that maximizes the partial score R_s^0 over the neighborhood on the coarse level (resolution level $d = 0$):

$$\Pi_s^0(x, y) = \underset{(\hat{x}, \hat{y}) \in \Delta(x, y)}{\operatorname{argmax}} \; R_s^0(\hat{x}, \hat{y}). \quad (7.9)$$

Secondly, the optimal position at levels $d > 0$ is recursively defined as a refinement of the position at $d - 1$:

$$\Pi_s^d(x, y) = \underset{(\hat{x}, \hat{y}) \in \Delta(2\Pi_s^{d-1}(x, y))}{\operatorname{argmax}} \; R_s^d(\hat{x}, \hat{y}). \quad (7.10)$$

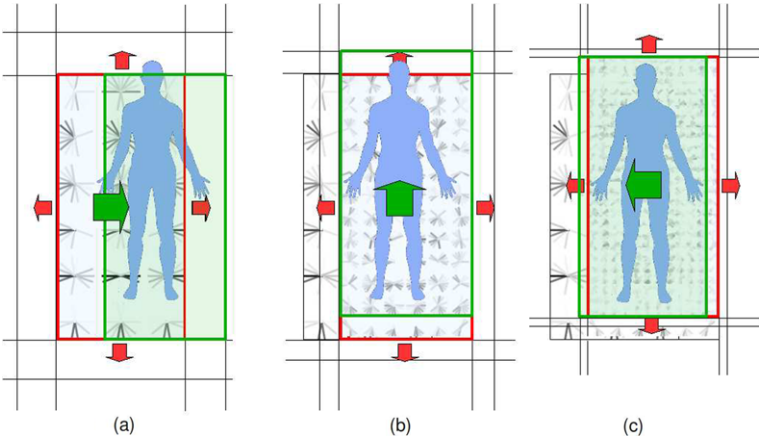


Fig. 7.8 Example of RCFL for detection. In (**a**), at a certain position (x, y) and scale s (*red box*) of the pyramid of features F, the best location $\Pi_s^0(x, y)$ (*green box*) for the low-resolution model of the object M_0 is searched in the local neighborhood $\Delta(x, y)$. In (**b**), the same procedure is repeated for the next resolution level, using as center of the neighborhood the best location computed at low resolution $\Pi_s^0(x, y)$. The process is recursively repeated for all feature resolution levels. In (**c**), the location obtained at the finest resolution $\Pi_s^2(x, y)$ is the location of the final detection

Finally, the total detection score D_s of the object at position (x, y) and scale s can be calculated as:

$$D_s(x, y) = \sum_d R_s^d\left(\Pi_s^d(x, y)\right). \tag{7.11}$$

The coordinates of the bounding box of the object are obtained from the finest level of object refinement. The final output of the detection is a set of detected objects O_D, defined by their class, location and probability. For more details on the detection algorithm see [42].

Scene Analysis

Scene analysis is accomplished using a low-level appearance-based image representation and a BoW approach. However, the main disadvantage of the standard BoW approach is that the information about spatial distribution of local feature vectors is lost in the final representation of the image. To overcome this problem, a spatial pyramid technique is usually applied [26].

The spatial pyramid that is used in our work is illustrated in Fig. 7.9. It has two levels: the zero-level includes the entire background region except the bounding box, and the first-level consists of three horizontal bars, which are defined by the foreground (bounding box). We used horizontal bars rather than a coarser grid, so that the image representation contains a histogram describing the center of the image, which usually contains a great deal of discriminative information. Histograms

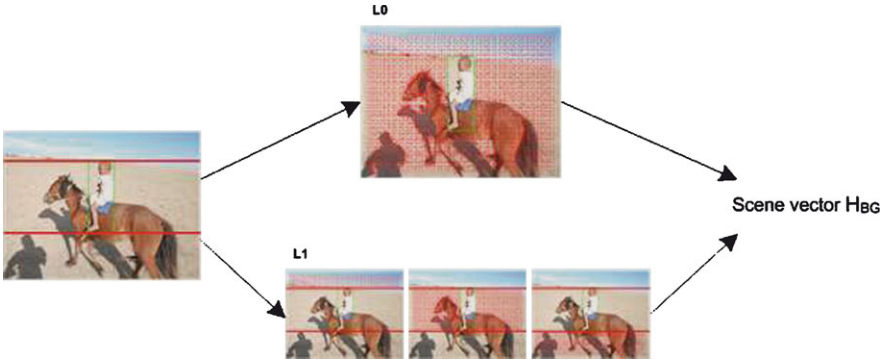


Fig. 7.9 Scene model H_{BG}. The final representation of the scene is a concatenation of the histograms from zero-level and first-level of pyramid

are computed from the keypoints found in each cell of these two levels, normalized, and then concatenated. Therefore, the background of the image I can be represented with a concatenation of histograms of all the pyramid levels:

$$H_{BG} = \frac{1}{2}[H_{BG_{L0}}\ H_{BG_{L1}}], \tag{7.12}$$

where $H_{BG_{L0}}$ and $H_{BG_{L1}}$ are the histograms of zero and first levels of the background pyramid, respectively.

In our implementation, we use 1000 words in the vocabulary, obtaining a total of $1000 + 1000 \times 3 = 4000$ bins in the background histogram H_{BG}.

7.3.2.2 Action Level: Pose Estimation

The information about the location of a human agent is obtained from a bounding box, provided in the dataset as ground truth. Pose estimation is achieved by fusing knowledge about the local appearance and the local shape of the human (Fig. 7.10).

The appearance of a human pose, H_{PA}, is computed straightforwardly in the area within the bounding box, using the BoW technique. A local shape representation of the human pose, H_{PS}, is obtained using Pyramid of HOG features (PHOG) [7], an extension of HOG [8]. In this approach, an image is divided into multiple grids at different levels of resolution; a HOG histogram is computed for each level and a final PHOG vector is the concatenation of all HOG histograms (see Fig. 7.10).

A final human pose model results from the concatenation of the appearance (H_{PA}) and shape (H_{PS}) representations:

$$H_P = \frac{1}{2}[H_{PA}\ H_{PS}]. \tag{7.13}$$

In our implementation, the dimensionality of H_{PA} is 500, since it is equal to the number of words in the vocabulary. For the H_{PS}, the HOG descriptor is discretized

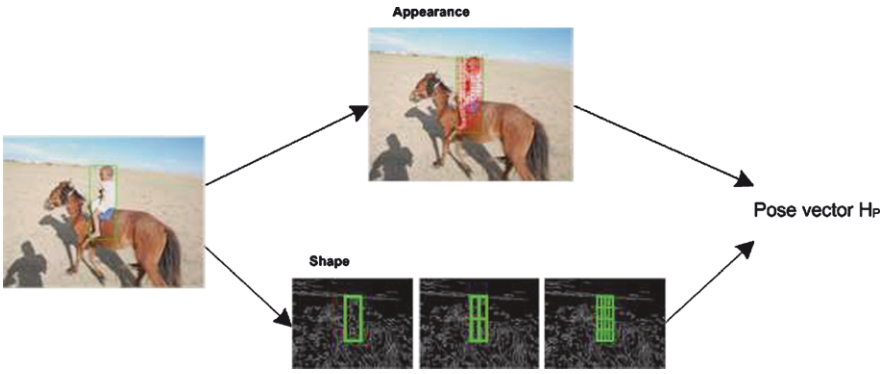


Fig. 7.10 Pose model H_p. The appearance model is computed using a BoW approach, and the shape model is represented with a PHOG feature vector

into 8 orientation bins and a 3-level pyramid is applied. The resulting PHOG descriptor is a $8 \times (2^0)^2 + 8 \times (2^1)^2 + 8 \times (2^2)^2 + 8 \times (2^3)^2 = 680$ dimensional vector. Therefore, the final histogram H_P has $500 + 680 = 1180$ bins in total.

7.3.2.3 Activity Level: Spatial Interaction

To handle spatial interactions at the activity level we combine two interaction models, see Fig. 7.11: (i) a local interaction model, and (ii) a global interaction model (adapted from [10]).

Local Interaction Model

The local interaction model helps to analyze the interactions between a human and the objects that are being manipulated by the human. The computation of the local interaction model is done by applying the BoW approach over the neighborhood of the human. The neighborhood is defined by the Mahalanobis distance from the center of the bounding box. This way, the local interaction of the image I can be represented with a histogram H_{LI} using a BoW approach that takes into account those feature vectors X_I that belong to a neighborhood ξ of the bounding box:

$$1 - \xi < \sqrt{(X_I - \mu)' S^{-1} (X_I - \mu)} < 1 + \xi, \tag{7.14}$$

$$S = \begin{bmatrix} (\frac{w}{2})^2 & 0 \\ 0 & (\frac{h}{2})^2 \end{bmatrix}, \tag{7.15}$$

where μ, h, and w are the center point, height, and width of the bounding box, respectively, and the neighborhood ξ is set to 0.5.

In our implementation of the local interaction model, the dimensionality of the H_{LI} is 1500, since it is the number of visual words in used vocabulary.

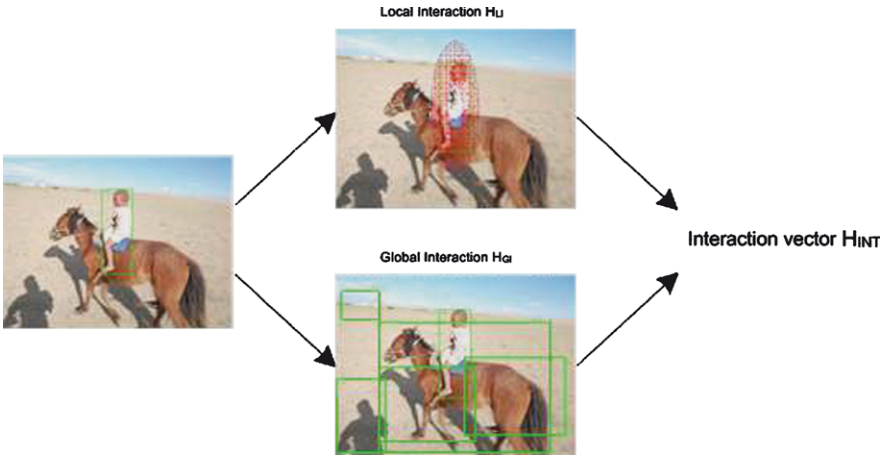


Fig. 7.11 Interaction model H_{INT}, as a combination of a local interaction and a global interaction model

Global Interaction Model

A basic description of actions in a scene can be done using information about the types of objects that are found in the scene by the object detector (see Sect. 7.3.2.1). Given N_O the number of object detections $O = [O_1, O_2, \ldots, O_{N_O}]$ in the image I, object occurrence can be represented as a histogram H_O:

$$H_O = \sum_{i=1}^{N_O} P_i u_i, \tag{7.16}$$

where u_i is such that only one element of u_i is nonzero, and $|u_i| = 1$. The index of the only nonzero element in u_i indicates the class of the object O_i with probability P_i.

This kind of information is necessary to quickly estimate the activity in the scene; for example, by observing a human, a racket, a ball, and a net in the image, we could propose a hypothesis about *playing tennis*. Moreover, if there is no necessary object in the scene (human, racket, ball, net), we can almost be sure that the *playing tennis* activity is not being performed. Therefore, using co-occurrence of objects, we could reject the most evident negative examples of not performing a particular activity. To prove or reject other hypotheses we need a more advanced model, which takes into account spatial distribution of objects in the scene.

The latter model can be obtained by analyzing the interactions across all the objects in the scene. The interaction between two objects i and j can be represented by a spatial interaction feature d_{ij}, which bins the relative location of the detection windows of i and j into one of the canonical semantic relations including *ontop*, *above*, *below*, *overlapping*, *next-to*, *near*, and *far* (see Fig. 7.12). Hence d_{ij} is a sparse binary vector of length $D = 7$ with a 1 for the kth element when the kth

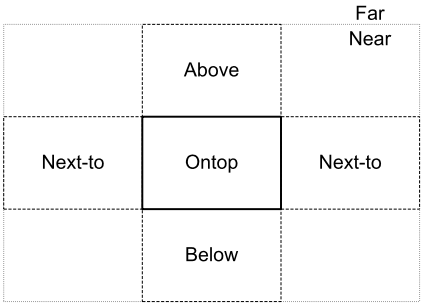


Fig. 7.12 A visualization of the spatial histogram feature d_{ij} from [10]. Seven spatial relationships are considered: *ontop*, *above*, *below*, *overlapping* (not shown), *next-to*, *near*, and *far*

relation is satisfied between the current pair of windows. Subsequently, every image I can be represented with an interaction matrix H_I. Every element $h_{I_{kl}}$ of the matrix H_I represents the spatial interaction between classes k and l:

$$h_{I_{kl}} = \sum_{i=1}^{N_{O^k}} \sum_{j=1}^{N_{O^l}} d\left(O_i^k, O_j^l\right) min\left(P_i^k, P_j^l\right), \tag{7.17}$$

where O^k and O^l are the detections of objects of classes k and l, correspondingly.

Therefore, the global interactions model H_{GI} is represented as the concatenation of the histograms H_O and H_I; the dimensionality of H_{GI} is $20 + 20 \times 20 \times 7 = 2820$, since we have 20 classes and seven possible spatial relationships. The final spatial interaction model H_{INT} is defined as the concatenation of the local and global interaction models, H_{LI} and H_{GI}:

$$H_{INT} = \frac{1}{2}[H_{LI}\, H_{GI}]. \tag{7.18}$$

Correspondingly, the dimensionality of the H_{INT} is $1500 + 2820 = 4320$, since it is a concatenation of H_{LI} and H_{GI}.

7.3.2.4 Classification

In this stage, the image histograms are classified using a Support Vector Machine (SVM) classifier, which was trained and tested using the respective image sets, as shown in Fig. 7.13. Moreover, a histogram intersection kernel is used to introduce non-linearity to the decision functions [26].

In order to fuse the multiple available representations of the image, namely H_P, H_{BG}, and H_{INT}, a concatenation of histograms and further L1-normalization have been applied.

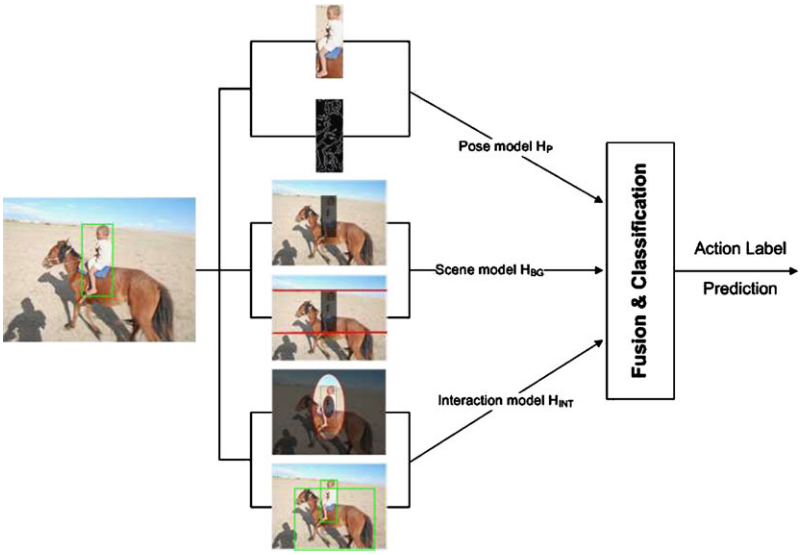


Fig. 7.13 Classification process

7.3.2.5 Experiments and Results

Dataset

We tested the presented approach on the dataset provided by the Pascal VOC Challenge 2010 [11]. The main feature of this dataset is that a person is indicated by a bounding box, and each person has been annotated with the activities they are performing from the following set: phoning, playing a musical instrument, reading, riding a bicycle or motorcycle, riding a horse, running, taking a photograph, using a computer, or walking. In addition, the images are not fully annotated—only 'person' objects forming part of the training and test sets are annotated. Moreover, actions are not mutually exclusive, e.g. a person may simultaneously be walking and phoning. The dataset contains 454 images and 608 bounding boxes in the training set and 454 images and 613 bounding boxes in the test set. Some images from the dataset are shown in Fig. 7.14.

To train the spatial interaction model based on object detections we used 20 object classes: *aeroplane*, *bicycle*, *bird*, *boat*, *bus*, *car*, *cat*, *chair*, *cow*, *dog*, *dining table*, *horse*, *motorbike*, *person*, *potted plant*, *sheep*, *sofa*, *train*, and *tv/monitor*. The object models were trained over the Pascal 2010 dataset images, using the included toolbox for object detection.

Evaluation of the Results and Discussion

To evaluate the importance of context and interactions in Human Event Understanding (HEU), three main experiments were conducted: action recognition (i) using

Fig. 7.14 Images belonging to different activity classes in the Pascal VOC Challenge 2010

Table 7.2 Average precision results on the Pascal Action Dataset using multiple cues

	H_P	H_P & H_{BG}	H_P & H_{BG} & H_{INT}
Walking	**67.0**	64.0	62.0
Running	75.3	75.4	**76.9**
Phoning	**45.8**	42.0	45.5
Playing instrument	45.6	**55.6**	54.5
Taking photo	22.4	28.6	**32.9**
Reading	27.0	25.8	**31.7**
Riding bike	64.5	65.4	**75.2**
Riding horse	72.8	87.6	**88.1**
Using PC	48.9	62.6	**64.1**
Average	52.1	56.3	**59.0**

only pose models, (ii) using pose models and scene analysis, and (iii) using pose model, scene analysis, and spatial interaction models (see Table 7.2). A selection of correctly classified and misclassified examples are illustrated in Figs. 7.15 and 7.16. The complexity of the dataset is that there are *simple actions* (e.g. walking, running), *actions with unknown objects* (e.g. phoning, playing an instrument, taking a photo, reading), and *actions with known objects* (e.g. riding a bike, riding a horse, using a PC). The evaluation of results is accomplished computing precision-recall curves and average precision measures.

As we can see from Table 7.2, for simple actions like *walking* or *running*, pose information is the most important cue. The minor improvements for the *running* class can be explained by the fact that *running* is usually observed outdoor in groups of people, while *walking* does not have such a pattern and can be performed both indoor and outdoor. Thus, when adding both context and interaction information, the recognition rate decreases.

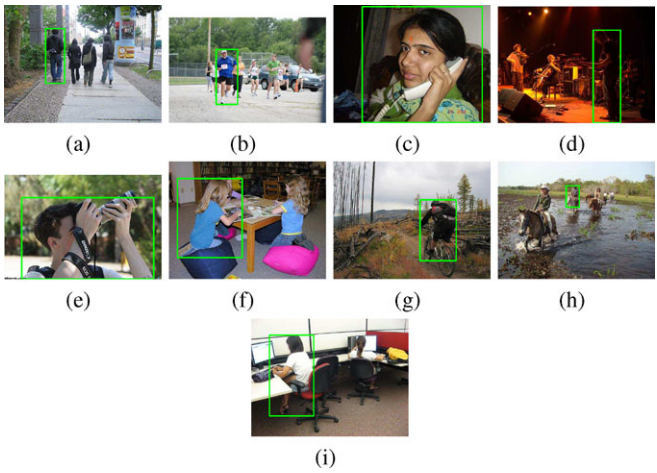


Fig. 7.15 Correctly classified examples of (**a**) walking, (**b**) running, (**c**) phoning, (**d**) playing an instrument, (**e**) taking a photo, (**f**) reading, (**g**) riding a bike, (**h**) riding a horse, (**f**) using a PC

Fig. 7.16 Examples misclassified as (**a**) walking, (**b**) running, (**c**) phoning, (**d**) playing an instrument, (**e**) taking a photo, (**f**) reading, (**g**) riding a bike, (**h**) riding a horse, (**i**) using a PC

Next, for those actions including interactions with unknown objects there is no single solution. The results for *phoning* are better when the pose model is used alone. This has two explanations: (i) the typical pose is discriminative enough for this action, and (ii) the bounding box containing the human usually occupies almost the whole image, so there is not much room for the context and objects in the scene. An action like *playing an instrument* improves significantly with a scene model, since that activity often means "playing in a concert" with quite particular and distinguishable context, e.g. cluttered dark indoor scenes. Even though we can observe the increase of performance for *taking a photo*, its recognition rate is still very low due to the significant variations in appearance, as well as the compact sizes of current cameras, which do not have discriminative features. The recognition results for the *reading* class significantly increase when object interaction models are added,

as *reading* is usually observed in an indoor environment, where many objects like sofas, chairs, or tables can be detected.

Finally, actions like *riding a bike, riding a horse, using a PC* improve significantly (13.5% in average per class) when we use a complete model consisting of pose, scene, and interaction, compared to the results based on a pose model only. This shows the particular importance of using context and spatial object interactions for activity recognition.

7.3.3 Behavior Modeling

As stated in the introduction, we distinguish among different objects of interest within a scene, namely *active entities*, like pedestrians and vehicles; *passive entities*, for movable objects like bags or chairs; and *background entities* for relevant static objects that define a context, like benches, and also interesting locations of the scenario, such as sidewalks, crosswalks, or waiting regions. The activities detected in the scene relate the interactions between active entities (agents) and passive entities, in the context defined by a series of background objects.

Nevertheless, a further step is required in order to incorporate a temporal dimension for a better understanding of behaviors. For that reason, information about temporal events provided by tracking systems has to be also incorporated to the analysis. This section proposes a methodology for extracting patterns of behavior using a deterministic rule-based approach that integrates appearance and motion cues. The motion-based detection and capture of interesting objects within image sequences are accomplished by segmentation and tracking procedures that capture the motion information of the scene from a single static camera (see [44, 45] for further information). As a result, a series of quantitative measurements over time is provided for each relevant moving object, such as positions, velocities, and orientations.

Although we could express the observed facts in a strictly quantitative way, e.g. treating the speed of a given target as 23 km/h, the use of fuzzy prototypical concepts allows us to evaluate facts in more generic and vague terms that can be better described by semantic models. Then, it would be recommendable to consider that this target may have *low*, *medium*, or *high* speed depending on the context of this observation, also to deal with the inherent uncertainty of the assertions, and to better relate and categorize the situations we observe. Thus, a common approach for conceptual modeling requires managing the conversion from quantitative to qualitative information.

First, spatiotemporal data are represented by means of logical predicates, created for each frame of the video sequence, in which numerical information is represented by its membership to predefined fuzzy functions. For example, a *zero*, *low*, *medium* or *high* tag can be assigned, depending on the instantaneous velocity value (V) for an agent (see Fig. 7.17). Apart from categorizing instantaneous facts, a conceptual scenario model also enables us to situate agents and objects in meaningful regions of the recorded location, e.g. crosswalk, sidewalk, or waiting zones. Finally, we have

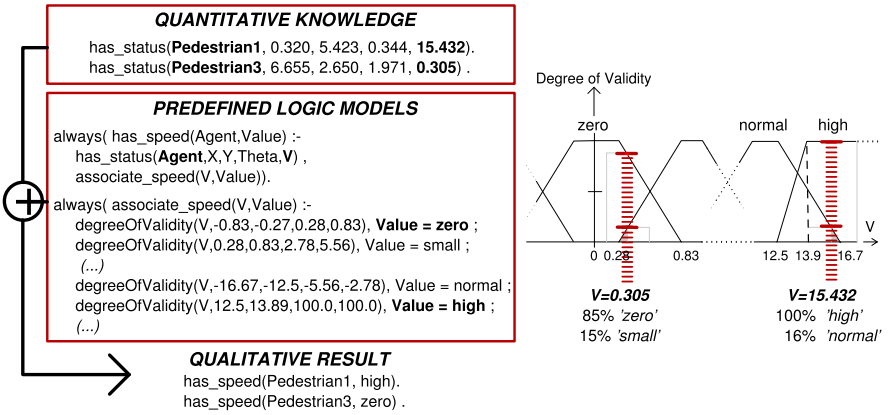


Fig. 7.17 Conversion from quantitative to qualitative values. The numerical value of velocity for an agent (last field of has_status) at a time step is linked to the most probable membership of the has_speed fuzzy function

selected a Fuzzy Metric Temporal Horn Logic (FMTL) framework to integrate the asserted facts and interpret the situations according to general motion-based models. This reasoning engine is similar to Prolog, but extended to take into account both geometric and temporal considerations.

A large collection of basic geometric facts results from this conceptualization, including information relative to positions and velocities, which needs to be subsequently filtered. Our particular aim is to detect admissible sequences of occurrences that contextualize geometric and temporal information about the scene, and will allow us to interpret the situation of a given person. For instance, a sequence in which a person walks by a sidewalk and stops in front of a crosswalk probably means that this person is waiting to cross.

Situation Graph Tree (SGTs) are the specific tool used to build these models [14, 38]. An SGT is a hierarchical classification tool used to describe behavior of agents in terms of situations they can be in. These trees contain a priori knowledge about the admissible sequences of occurrences in a defined domain, connecting each possible state (i.e. situation) by means of prediction and specialization edges. When a set of conditions is asserted for one of the situations, a new high-level reaction predicate is generated. SGTs have been successfully deployed to cover a large range of applications including HEU [14], road traffic analysis [37], video partitioning and annotation [13], visual scene detection and augmented reality.

The semantic knowledge related to any agent at a given point in time is contained in a *situation scheme*, the basic component of a SGT (see Fig. 7.18). A situation scheme can be seen as a semantic function that evaluates an input containing a series of conditions—the so-called *state predicates*—, and generates logic outputs at a higher level—the *reaction predicates*—once all the conditions are asserted. Here, the reaction predicate is a `note` method that generates an interpretation for the current situation in a language-oriented form, with fields related to thematic roles such as locations, agents and rigid objects.

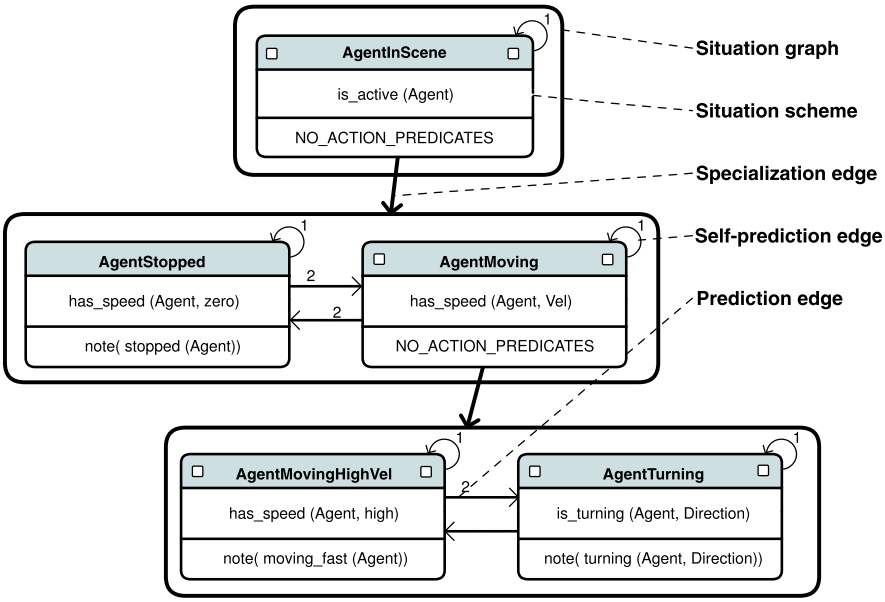


Fig. 7.18 Naive example of a SGT, depicting its components. Specialization edges particularize a general situation scheme with one of the situations within its child situation graph, if more information is available. Prediction edges indicate the situations available from the current state for the following time step; in particular, self-prediction edges hold a persistent state

The temporal dimension is also tackled by the SGTs. As seen in Fig. 7.18, the situation schemes are distributed along the tree-like structure by means of three possible directional connections: *particularization*, *prediction*, and *self-prediction edges*, with a number on each edge indicating the order of evaluation. Particularization (or specialization) edges refine a general situation with more particularized ones, once its conditions have been asserted. On the other hand, prediction edges inform about the following admissible states within a situation graph from a given state, including the maintenance of the current state by means of self-prediction edges. Thus, the conjunction of these edges allow for defining a map of admissible paths through the set of considered situations. Figure 7.19 shows a part of an SGT that aims at identifying situations such as an abandoned object or a theft.

7.4 Summary

HEU is still an open problem today. Its particular solutions are used in numerous applications such as video surveillance, video and image search, or human–computer interaction, among others. In these domains, our main hypothesis has been that to correctly conduct HEU, it is required to model the relationships between the humans and the environments where the events are happening. Specifically, human

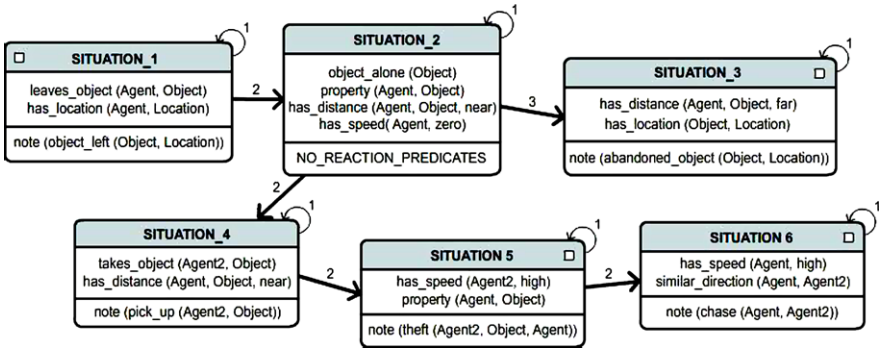


Fig. 7.19 SGTs are used to model situations and behaviors as predefined sequences of basic events. The example of Situation Graph shown, part of a SGT, allows for complex inferences such as abandoned objects, chasing or thefts, by means of high-level `note` predicates

events can be understood by combining the knowledge of the global scene and objects nearby together with the detected human poses and their analyzed motion. In this chapter, we (i) inferred from the related work a proper general framework for the analysis of human events, (ii) analyzed the importance of interactions and context in activity recognition, and (iii) proposed a top-down modeling scheme to incorporate appearance, context, and motion information toward the assessment of complex human behavior patterns.

Our general methodology for HEU has been developed based on a novel taxonomy comprised by the terms *entity*, *movement*, *action*, *activity*, and *behavior*. This proposal covers the whole process of scene understanding, ranging from scene observation (i.e., perception) to scene interpretation (i.e., cognition). For this purpose, different levels of semantic interpretation have been established to bridge the gap between pixels and video hermeneutics. The final scheme has been designed as a combination of bottom-up inference of context and a top-down integration of appearance, motion, and scene interactions using a logical reasoner (FMTL) over temporal-contextual models (SGT). New trends on HEU suggest that a plausible path to follow involves to understand scene context by means of an integrative combination of multiple cues like the one proposed.

The following sentences summarize some of the important ideas about HEU that can be extracted from this chapter.

- **Context.** While traditional approaches in HEU use only information about the human itself, recent trends show that it is much more beneficial to analyze a human behavior in context.
- **Multilevel.** Different semantic levels of HEU require considerations of context at different levels to carry out a proper analysis.
- **Taxonomies.** In complex problems involving sequential stages, each one having characteristic properties, such as HEU, taxonomies are useful for capturing and organizing the knowledge and facilitating further procedural analyzes.

- **Semantic models**. Prior models, be they deterministic or probabilistic, ease in great measure the formalization of complex properties and contexts, although they require human experts to model such constraints.

7.5 Questions

- What is human behavior and which subclasses of it are there? Why do we need to differentiate among them?
- Is it necessary to include appearance categories in a taxonomy of human events? If so, why? If not, why not? Think about examples of appearance detections which might be fundamental for the recognition of human events in context.
- Why do some particular types of human behavior require mainly pose information? Which types of behavior require motion and interaction models for their proper understanding?
- Why do we need to combine bottom-up inference and top-down modeling to accomplish HEU?
- Propose a minimal conceptual model consisting of a set of FMTL rules and a SGT, which is capable of, for example, detecting people falling down from a horse in image sequences.

7.6 Glossary

- *Bag of Words* is a simplifying assumption used in many domains including computer vision. In this model, a image (or a region of interest) is represented as an unordered collection of visual features ("words"), disregarding their spatial location [6].
- *Histogram of Oriented Gradients* is a descriptor that encodes information about distribution of local intensity gradients of the region of interest (object, image, human, etc.) [8].
- *Pyramid of HOG* is a concatenation of HOG calculated on different resolution levels over the region of interest [7].
- *Fuzzy Metric Temporal Horn Logic* is a form of logic in which conventional formalisms are extended by a temporal and a fuzzy component. The first one permits to represent, and reason about, propositions qualified in terms of time; the last one deals with uncertain or partial information, by allowing degrees of truth or falsehood [38].
- *Situation Graph Tree* is a deterministic model that explicitly represents and combines the specialization, temporal, and semantic relationships of its constituent conceptual predicates. SGTs are used to describe the behavior of an agent in terms of situations he can be in [14, 38].
- *Human Event Understanding* is a framework to guide the assessment and evaluation of human behavior in image sequences, which takes into account spatial, temporal, and contextual properties of the observed scene.

Acknowledgements We gratefully acknowledge Marco Pedersoli in providing the detection module. This work was initially supported by the EU Project FP6 HERMES IST-027110 and VIDI-Video IST-045547. Also, the authors acknowledge the support of the Spanish Research Programs Consolider-Ingenio 2010: MIPRCV (CSD200700018); Avanza I+D ViCoMo (TSI-020400-2009-133); CENIT-IMAGENIO 2010 SEGUR@; along with the Spanish projects TIN2009-14501-C02-01 and TIN2009-14501-C02-02.

References

1. Al-Hames, M., Rigoll, G.: A multi-modal mixed-state dynamic Bayesian network for robust meeting event recognition from disturbed data. In: IEEE International Conference on Multimedia and Expo (ICME 2005), pp. 45–48 (2005)
2. Albanese, M., Chellappa, R., Moscato, V., Picariello, A., Subrahmanian, V.S., Turaga, P., Udrea, O.: A constrained probabilistic Petri Net framework for human activity detection in video. IEEE Trans. Multimed. **10**(6), 982–996 (2008)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: International Conference on Computer Vision (2005)
4. Bobick, A.F.: Movement, activity and action: the role of knowledge in the perception of motion. Philos. Trans. R. Soc. Lond. B, Biol. Sci. **352**(1358), 1257 (1997)
5. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2002)
6. Bosch, A., Munoz, X., Marti, R.: Which is the best way to organize/classify images by content? Image Vis. Comput. **25**(6), 778–791 (2007)
7. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: ACM International Conference on Image and Video Retrieval (2007)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893, San Diego (2005)
9. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: Proceedings of the British Machine Vision Conference, Aberystwyth, UK (2010)
10. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: International Conference on Computer Vision (2009)
11. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results (2010)
12. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intell. **32** (2010)
13. Fernández, C., Baiget, P., Roca, F.X., Gonzàlez, J.: Interpretation of complex situations in a cognitive surveillance framework. Signal Process. Image Commun. **23**(7), 554–569 (2008)
14. Fernández, C., Baiget, P., Roca, F.X., Gonzàlez, J.: Determining the best suited semantic events for cognitive surveillance. Expert Syst. Appl. **38**(4), 4068–4079 (2011)
15. Fusier, F., Valentin, V., Brémond, F., Thonnat, M., Borg, M., Thirde, D., Ferryman, J.: Video understanding for complex activity recognition. Mach. Vis. Appl. **18**(3), 167–188 (2007)
16. Gonzàlez, J.: Human sequence evaluation: the key-frame approach. PhD thesis, UAB, Spain (2004). http://www.cvc.uab.es/~poal/hse/hse.htm
17. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 1775–1789 (2009)
18. Hongeng, S., Nevatia, R.: Multi-agent event recognition. In: International Conference on Computer Vision, pp. 84–93 (2001)
19. Ikizler, N., Duygulu, P.I.: Histogram of oriented rectangles: A new pose descriptor for human action recognition. Image Vis. Comput. **27**(10), 1515–1526 (2009)

20. Ikizler, N., Forsyth, D.A.: Searching video for complex activities with finite state models. In: CVPR (2007)
21. Ikizler-Cinbis, N., Cinbis, R.G., Sclaroff, S.: Learning actions from the web. In: International Conference on Computer Vision (2009)
22. Kitani, K.M., Sato, Y., Sugimoto, A.: Recovering the basic structure of human activities from noisy video-based symbol strings. Int. J. Pattern Recognit. Artif. Intell. **22**(8), 1621–1646 (2008)
23. Kjellström, H., Romero, J., Martínez, D., Kragić, D.: Simultaneous visual recognition of manipulation actions and manipulated objects. In: European Conference on Computer Vision, pp. 336–349 (2008)
24. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska (2008)
25. Laxton, B., Lim, J., Kriegman, D.: Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007 (CVPR'07), pp. 1–8 (2007)
26. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, pp. 2169–2178 (2006)
27. Li, L.-J., Fei-Fei, L.: What, where and who? classifying event by scene and object recognition. In: International Conference on Computer Vision (2007)
28. Lindeberg, T.: Feature detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 77–116 (1998)
29. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: IEEE Conference on Computer Vision and Pattern Recognition, Florida, USA (2009)
30. Lowe, D.G.: Object recognition from local scale-invariant features. In: International Conference on Computer Vision, Kerkyra, Greece, p. 1150 (1999)
31. Mahajan, D., Kwatra, N., Jain, S., Kalra, P., Banerjee, S.: A framework for activity recognition and detection of unusual activities. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing. Citeseer, University Park (2004)
32. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition, Florida, USA (2009)
33. Masoud, O., Papanikolopoulos, N.: A method for human action recognition. Image Vis. Comput. **21**(8), 729–743 (2003)
34. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. Int. J. Comput. Vis. **60**(1), 63–86 (2004)
35. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2), 90–126 (2006)
36. Moore, D., Essa, I.: Recognizing multitasked activities from video using stochastic context-free grammar. In: Proceedings of the National Conference on Artificial Intelligence, pp. 770–776 (2002)
37. Nagel, H.H.: From image sequences towards conceptual descriptions. Image Vis. Comput. **6**(2), 59–74 (1988)
38. Nagel, H.H., Gerber, R.: Representation of occurrences for road vehicle traffic. AI Mag. **172**(4–5), 351–391 (2008)
39. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vis. **79**(3), 299–318 (2008)
40. Noceti, N., Santoro, M., Odone, F., Disi, V.D.: String-based spectral clustering for understanding human behaviours. In: Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences, pp. 19–27 (2008)
41. Oliver, N., Rosario, B., Pentland, A.: A Bayesian computer vision system for modeling human interactions. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 831 (2000)
42. Pedersoli, M., Gonzàlez, J., Bagdanov, A.D., Roca, F.X.: Recursive coarse-to-fine localization for fast object detection. In: European Conference on Computer Vision (2010)

43. Polana, R., Nelson, R.C.: Detection and recognition of periodic, nonrigid motion. Int. J. Comput. Vis. **23**(3), 261–282 (1997)
44. Roth, D., Koller-Meier, E., Van Gool, L.: Multi-object tracking evaluated on sparse events. Multimed. Tools Appl. 1–19 (September 2009), online
45. Rowe, D., Rius, I., Gonzàlez, J., Villanueva, J.J.: Improving tracking by handling occlusions. In: 3rd ICAPR. LNCS, vol. 2, pp. 384–393. Springer, Berlin (2005)
46. Saxena, S., Brémond, F., Thonnat, M., Ma, R.: Crowd behavior recognition for video surveillance. In: Advanced Concepts for Intelligent Vision Systems, pp. 970–981 (2008)
47. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: International Conference on Pattern Recognition, Cambridge, UK (2004)
48. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVID. In: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (2006)
49. 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)
50. Smith, P., da Vitoria, N., Shah, M.: Temporal boost for event recognition. In: 10th IEEE International Conference on Computer Vision, October 2005
51. Vu, V.T., Brémond, F., Thonnat, M.: Automatic video interpretation: A recognition algorithm for temporal scenarios based on pre-compiled scenario models. Comput. Vis. Syst. 523–533 (2003)
52. Wang, Y., Jiang, H., Drew, M.S., Li, Z.N., Mori, G.: Unsupervised discovery of action classes. In: IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, pp. 1654–1661 (2006)
53. Xiang, T., Gong, S.: Beyond tracking: modelling activity and understanding behaviour. Int. J. Comput. Vis. **67**(1), 21–51 (2006)
54. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, Florida, USA (2009)
55. Zheng, H., Wang, H., Black, N.: Human activity detection in smart home environment with self-adaptive neural networks. In: Proceedings of the IEEE International Conference on Networking, Sensing and Control (ICNSC), pp. 1505–1510, April 2008

Part III
Social and Affective Behaviors

Chapter 8
Social Signals: A Psychological Perspective

Isabella Poggi and Francesca D'Errico

8.1 Introduction

Along the whole twentieth century, a large part of psychology was devoted to explicate and measure intelligence, meant as the set of cognitive skills (memory, inference, insight, reasoning) that allow humans to solve problems and adapt to environment. Only at the end of the century, after the growth of the cognitive over the behaviorist approach and the rejoining of emotion and cognition, studies by Salovey and Mayer [152], Damasio [53] and Goleman [81] showed that cognitive processes like decision making cannot do without the contribution of emotional processes and introduced the notion of "emotional intelligence", the capacity of expressing one's emotions, understanding others' emotions and being empathic with them, as a great part of a human's capacity for adaptation. At the beginning of the third millennium the notion of "social intelligence" was finally proposed: a set of skills that include understanding of other people's feelings, seeing things from their point of view (Goleman [82]) and giving them effective responses (Gardner [79]), but also Machiavellian intelligence, the capacity to understand what others want to better manipulate them. A notion taken up as particularly relevant in managerial psychology, and viewed as a weapon for leadership, so much as to be seen as "the science of success" (Albrecht [2]). An important part of Social Intelligence is the delivery and comprehension of Social Signals, the signals that inform about an ongoing interaction, or a social relationship, an attitude taken or an emotion felt toward another person.

I. Poggi · F. D'Errico (✉)
Roma Tre University, Rome, Italy
e-mail: fderrico@uniroma3.it

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_8,

8.2 The Dawn of Social Signals. From Computer Science to Social Psychology and Back

In 2007 Alex Pentland, working at MIT in the areas of computational social science and organization engineering, first introduced the notion of "social signal processing" and applied signal processing techniques to what he called "social signals", the nonverbal aspects of communication including interactive and conversational dynamics, to predict the outcomes of a negotiation or of a speed date within its very first minutes (Pentland [120]; Curhan and Pentland [52]). Later, he contended that prosodic emphasis, mirroring, conversational turn-taking, and activity level, are "honest signals" that allow to predict the outcomes of speed dates, negotiations, and other types of social interaction (Pentland [121]). These works were primarily based on findings of social psychology concerning the role of nonverbal behavior on conversation and social interaction. A first input was an intriguing study by Ambady and Rosenthal [5], "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness", which demonstrated that judgments of personality are based on very "thin slices" of nonverbal behavior, thus reinforcing Asch's hypothesis about the strong human capacity to form an accurate impression of people as a precondition to social life [10]. A silent video of 30 seconds was a strong predictor of teacher's evaluations at the end of semester; but even when replicated with a shorter video of six seconds, results confirmed the "thin slice" hypothesis.

The idea of "social signals" was also inspired by studies on the role of nonverbal behavior in impression formation. For example, smile and rapid body movement were judged as extraversion cues (Kenny et al. [103]), while physical attractiveness predicted sociability and social competence (Eagly et al. [66]). Similarly, within research on self-presentation (DePaulo [57]) nonverbal behavior was seen as a cue to impression management often intentionally driven. As pointed out by Goffman [80], "tie-signs" are used by people in personal relationships to make clear or to reinforce the nature of a relationship: a man gazes or orients his body toward his interlocutor more frequently if the interlocutor is a female than a male, and in the former case he does so more often during an intimate than an impersonal conversation.

Other contributions come from studies on facial expression (Ekman and Friesen [69]; Ekman et al. [70]) and on gestures and other modalities (Rosenthal et al. [150]). Ekman [71] distinguished facial movements into *emotional signals* and *conversational signals*; the former, not totally under voluntary control, leave room for the detection of deception through *leakage cues* and *deception cues*: in leakage, due to cultural or social norms an unfelt facial expression is only superimposed to the expression of the really felt emotion, thus revealing the information concealed; while deception cues are micro displays so brief that they are very difficult to detect (DePaulo et al. [58]).

Much of the research above highlighted the relevance of automatic processes that are not under conscious control (Bargh [13, 14]) but nonetheless heavily determine social perception and social cognition.

Based on these works, since 2009, the European Network of Excellence SSPNet has put the bases for a new research field in "Social Signal Processing", concerning the cognitive modeling, analysis and synthesis of social signals.

Although much empirical research has been done in social psychology, apart from some first work in the SSPNet (Cowie et al. [33]; Vinciarelli et al. [173]; Poggi and D'Errico [133, 137]), no clear definition of "social signals" has been given yet. But to set a new research field, a clear definition is needed of what is inside and what is outside the field. In this chapter we provide a definition of social signals and present some recent studies in this area, taking the perspective of a model of mind and social action in terms of goals and beliefs.

8.3 Social Cognition. How Others Are Represented in Our Mind and Our Brain

A central area of research to understand the processes that allow the perception, memorization and representation of social signals in the human mind is "Social Cognition" a psychological approach that studies how people interpret and attribute meaning to one's own and others' behaviors. The pioneering work of Bartlett [18] on *schemes* pointed out a central axiom in the social cognition approach: the representational nature of our knowledge. He demonstrated that memory reconstructs its stored events, since it is oriented to making memories coherent with reference schemes. But also categories like status, role, human groups, at least in their default working process, are uncontrollable and unintentional, because they respond to *laws of cognitive economy* [75]. Perception, memorization, judgment follow an automatic path, in absence of "motivation" and "opportunity" of time and resources; categories and schemes influence information on the basis of accessibility in terms of past experience or primacy effect (i.e., what is seen/heard first: Higgings and Rholes [92]); and, as demonstrated by Asch [10], to have a coherent description of a person we organize impressions as a whole, starting from few first elements. Finally, categorization—the clustering of different elements on the basis of one shared condition—simplifies social perception and social judgment by making external stimuli more accessible and triggering sets of information focused on particular objects, interrelated and organized in schemes (Fiske and Taylor [76]). Even the discriminative behavior toward outgroup members is generally triggered on the basis of automatic activation of belongingness categories (Bargh [15]).

Research in neuroscience has investigated the neural underpinning of social cognition by demonstrating that processes involved in social perception and behavior, like perception of conspecifics, memory and behavior concerning others, activate different neural systems from the perception of objects (Adolphs [1]). And the discovery of mirror neurons—the neurons activated not only by one's motor action, but also by the perception of action in a conspecific (Rizzolatti and Arbib [149])—demonstrated how humans are programmed for empathy (Gallese [78]), the representation of self and others (Uddin et al. [169]), learning from others through

imitation (Meltzoff and Decety [112]; Iacoboni [94]), and joint action (Vesper et al. [171]). A further demonstration of their importance comes from studies that suspect impairment of these structures in autism (Williams et al. [178]; Oberman et al. [118]), where the capacity for a Theory of mind (a representation of the other's mental states, like emotions, goals and beliefs) is disrupted (Baron-Cohen [17]; Gallese [78]).

In this work we adopt the view of social cognition put forward by a model of mind and social action in terms of goals and beliefs.

8.4 A Goal and Belief Model of Mind and Social Action

In the model of mind, social action and communication designed by Castelfranchi and Parisi [42], Conte and Castelfranchi [50], Castelfranchi and Poggi [43] and Poggi [128, 129], the life of any natural or artificial, individual or collective system consists of pursuing *goals*, i.e., regulating states that, as long as they are perceived by the system as not being realized in the world, trigger its action. To realize a goal a system performs plans, i.e. hierarchical structures where all actions are means for goals and possibly superordinate goals (supergoals). This requires internal resources (*beliefs* and *action* capacities) and external resources (material resources, world conditions). When a system lacks the power to achieve goals on its own due to lack of necessary resources, it may *depend* on another system endowed with those resources, and need the other to *adopt* its goals, i.e., help to achieve them. The social device of *adoption*—the fact that a system pursues another's goal as its own—multiplies the power to achieve goals for all systems. Further, a system may need to *influence* another (induce it to pursue some goals), for its own sake (e.g., a master giving orders to his slave) or for the sake of the influenced system (a father giving advice to his child), or both (two colleagues trying to persuade each other to find a common solution).

Beliefs are an essential resource to achieve goals. A belief is information about some state of the external world or of the system (like "it is sunny now" or "I am hungry"), represented in a sensorimotor or conceptual format, not necessarily in a conscious way. Beliefs are necessary to choose the goals to pursue, assess preconditions and adequate plans; hence the necessity to acquire, process, store and use beliefs, for all systems, including animals and machines, but more so for humans. Humans and higher animals acquire beliefs through *perception*, *signification* and *communication*, and process and store them in long-term memory, connecting them in belief networks through links that denote time, space, part-of relations, class-example relations, cause, goal, condition, and thus generating *inferences*, i.e., new beliefs drawn on the basis of others that have been acquired through perception or retrieved from memory. The difference between perception and the other sources of beliefs is that while in the former the information drawn is, so to speak, not distinct from the perceptual information (from seeing *clouds* I come to know there are *clouds*), in inference, signification and communication there is a splitting between the information drawn and its source: from mother saying *Stay home*, I understand

mama wants me home (communication), from seeing *clouds* I understand *rain soon* (signification, meaning), from *rain* I predict *mama won't let me out* (inference).

In signification and communication, from a perceivable stimulus, a "signal", we draw a "meaning". A meaning can be seen as a privileged and pre-determined inference: when a perceivable stimulus has generally given rise to the same inference, this has become linked in memory to that perceivable stimulus and is systematically drawn from it, thanks to a stable connection shared by a group through biological coding or cultural learning.

8.5 Social Signals. A Definition

We define a signal as any perceivable stimulus from which a system can draw some meaning, and we distinguish informative from communicative signals: a *communicative signal* is a perceivable stimulus produced by an animate system (a self-regulated Sender) having the goal to provide information to another system (Addressee); an *informative signal* is one from which some system (Receiver) draws some meaning without necessarily the intention, or even the existence, of a Sender intending to convey it.

We define signals as "social signals" on the basis of their meaning, i.e., their concerning "social facts". A *social signal* is "*a communicative or informative signal which, either directly or indirectly, provides information about "social facts", that is, about social interactions, social attitudes, social relations and social emotions* (Poggi and D'Errico [133, 136]).

8.5.1 Social and Non-social, Informative and Communicative Signals

Suppose in the mountains you see on the ground some splinters from the horns of a Big Horn; if you understand there has been a contest between two big horns, this is for you a *social informative signal*: *social* because it concerns a social interaction between Big Horns, and *informative* because the Big Horn did not have the goal to inform you of the contest. But if you simply see the footprints of one Big Horn, and hence you can tell he crossed the wood, this is an *informative non-social* signal. Again, if during a class break I see that some children are close to each other forming a circle, but one is slightly outside the circle, I might predict that this child is somewhat isolated from the group, possibly at risk of being bullied, though no one in the group wants to communicate to each other or to that child that he is somewhat isolated. The children's spatial location is a signal informing, not communicating on purpose, about some social relation.

Now suppose you observe mimicry between two teenagers talking together: each inadvertently imitates the other's movements and posture; a signal of syntonization.

For you, as an external observer, this is an *informative social signal*, since they are not communicating this to you. Between them, instead, this counts as a *communicative social signal*, because through mimicry they communicate their reciprocal affinity to each other.

8.5.2 Communicative Signals and Their Communicative Goals

In communicative signals (both social and non-social ones), the goal of conveying information is not necessarily a conscious intention, i.e., a deliberate and aware communicative goal of an individual, as is generally the case, in humans, for verbal language or some codified gestures. Animals' signals are governed by biological goals of communicating, but also in humans the goal to convey information may be an unconscious goal, or even a social end or a biological function. Some examples: if I have a bad opinion of you because you offended me, I can insult you deliberately; in this case my goal of communicating my negative evaluation of you is conscious: I not only want to offend, but I also know I want to offend. An *intention* is a conscious goal, i.e., one not only represented, but also meta-represented in my mind, while an unconscious goal is one I have but I do not know (I somehow conceal from myself) I have it. Take this case by Ekman and Friesen [69]: a student is being treated in an offensive way by her professor, to whom she cannot obviously show anger; so she is forced to answer him in a polite way. But at the same time on her knee she is extending her middle finger in an obscene insulting gesture that contradicts her hypocritical politeness. Here, if at a conscious level she does not want to offend, at the same time she does have the goal to express her anger and to insult in turn; but if due to her family education she is a very polite girl, she might not even be aware of this latter goal. If this is the case, her extended middle finger is a communicative social signal, but one triggered by an unconscious goal. Like mimicry, the imitation of one's interlocutor's movements generally occurs without awareness.

Some communicative social signals are governed by social ends, that is, goals not primarily of an individual, but of the society: for instance those that convey information about social roles and social identities, such as uniforms or status symbols. A cop's uniform tells us, on behalf of the whole society, that the one who wears it plays a particular social role; a Ferrari tells us its driver belongs to a group of very rich people. Other communicative social signals governed by biological goals are those of sexual identity (a man's beard), sexual readiness (a stickleback's reddened abdomen, a woman's pupil dilation), and some emotional signals like blushing (see Sect. 8.7.4.1).

In general, many entities and events can work as social informative and communicative signals: individual and collective actions (a letter of complaint, a strike); morphological features, either transitory (blushing that expresses shame) or permanent (a woman's breast as a signal of sexual identity); objects (a Ferrari or a uniform), but also combinations of actions: simultaneous (many people applauding

at the same time) or sequential (mimicry). Of course, while individual actions are generally driven by individual conscious goals of communicating, morphological features are ruled by biological ones, and objects often by social ends. For actions of more people, those that are pre- and co-ordinated, like a strike, may still have a conscious goal, thus being *communicative social signals*; but if thousands of people click on the same video on Youtube, this is an *informative social signal*.

This has important consequences as to their reliability. With deliberate *communicative signals* we come to know something from someone else, who potentially might want to mislead; on the other hand, involuntary signals, that is, *informative signals* and *communicative signals* are not under conscious control; they may be subject to misunderstanding, but in principle not to cheating. (Unless, of course, they are only apparently involuntary, that is, produced by someone deliberately but pretending not to be deliberate.)

8.5.3 Direct and Indirect Social Signals

Sometimes information about social facts is not conveyed explicitly but in an indirect way. Here we contrast the *literal* to the *indirect* meaning of signals. According to Poggi [128], in communicative signals of whatever modality (gesture, facial expression, gaze, posture, physical contact), the relation between signal and meaning may be either *"codified"* or *"creative"*. The former case implies a stable connection in long-term memory between a specific perceivable stimulus and the corresponding belief, with a list of these signal-meaning pairs making a "lexicon" (for example a lexicon of gestures, head movements, gaze items) possibly similar to the mental lexicon of words for a verbal language. In the latter case, the signal—meaning link is not represented once and for all, but can be deduced on the basis of systematic rules: for instance, an iconic gesture is constructed (and a meaning is drawn from it) on the basis of the similarity between shape and movements of hands and the content referred to by the gesture. This meaning, whether codified or creative, is the "literal meaning" of a signal. But when the signal is produced (and understood) in context, information coming from context may combine with that literal meaning and, through inferential processes, give rise to further, "indirect" meanings, which differ across contexts.

Some examples of indirect meanings. A is presently depressed, and her state of sadness and depression is clearly signaled by her facial expression (Cohn [48]). This is a communicative signal, but not a "social" one, because sadness is not a "social emotion" in itself (whereas, for instance, being "sorry-for" someone might be one). Yet, from her depression, B might infer that A does not want to talk with B. In this case the signal of depression is an indirect social signal of interaction. Again: take a teacher, who, in total good faith, overhelps her pupil, i.e., helps him to complete a task he could well complete by himself (D'Errico et al. [61, 132]); the pupil might finally infer a negative evaluation of his own skills. This is then an *indirect informative social signal*. But take (another real example) an amateur orchestra conductor

during a concert rehearsal; the concert sponsor, who is the habitual conductor of that orchestra and does not trust the amateur conductor's skills, fearing a potential failure, wants to convey the orchestra players that they should not follow him because he is not a good conductor, and staying behind the amateur conductor he makes conducting gestures too. While the teacher's case was an informative social signal, not sent on purpose, this is an *indirect communicative social signal*, since the inferences drawn by the Addressees (the players) are intended by the Sender (the sponsor).

8.6 Modalities of Social Signals

Humans produce Social Signals in all modalities, words, prosody and intonation, gestures (McNeill [110]; Kendon [102]), posture (Condon and Ogston [49]), head movements (Cerrato [44]), facial expression; gaze (Argyle and Cook [7]; Poggi [128, 129]) physical contact and spatial behavior (Montagu [115]; Hall [87]), and by their sequential and simultaneous combination make multimodal "discourses" (Poggi [128]). Let us overview studies in these modalities.

8.6.1 Verbal and Vocal Features

An obvious case of social signals are words and sentences, mainly those expressing social acts, feelings and evaluations. Research in the detection of linguistic social signals includes "sentiment analysis" (see for example Wilson et al. [180]; de Rosis and Novielli [55]) and the analysis of "subjectivity" (Wilson and Hofer [179]) which, after distinguishing objective from subjective utterances, those expressing positive and negative "private states" (opinions, beliefs, sentiments, emotions, evaluations, uncertainties, speculations), describe them in terms of the state of an experiencer holding an attitude toward a target (Wiebe [176, 177]). "Sentiment analysis" aims at recognizing the viewpoint that underlies a text, by classifying the polarity (positive/negative) of its words and sentences, possibly by means of thesauri or semantic dictionaries such as WordNet,[1] and by measuring their frequency. A first relevant issue here is to take into account the *valence shifters*: modifiers that change the intensity of a term (intensifiers and diminishers) or its orientation (negations). It is different, of course, to say "Jane is nice" or "Jane is *very* nice" or "Jane is *not* nice". Another issue is the attribution of the mental states mentioned to its source: if I say "I like Jane", the subjective state is felt by me, while if I say "I know you like Jane" the state of liking is attributed not to me but to you, which can be detected thanks to verbs of saying or subordinate clauses. But purely syntactic or lexical cues are not sufficient to capture more subtle ways to express opinions or evaluations: the analysis of context, for example of previous sentences in a debate or an interview, is necessary to draw the right inferences (de Rosis and Novielli [55]).

[1] http://wordnet.princeton.edu/

The analysis of linguistic subjectivity has been used in marketing to detect customers' orientations, in persuasive natural language processing (Guerini et al. [85]) but also in dialogues to detect role distribution (Wilson and Hofer [179]).

Another important domain of Social signal processing are the prosodic aspects of speech, which include temporal features, like pauses, vowel length, speech rhythm, articulation rate, but also acoustic features such as pitch, vocal intensity and voice quality.

The first important exploitation of signal processing techniques in this domain were Pentland's studies, which from correlations between participants' acoustic features detected activity level, influence and mirroring, which signal particular role relations and interaction outcomes. Pioneering studies in psychology (Scherer [154]; Scherer and Scherer [155]) on the acoustic cues to personality features recently gave rise to signal processing studies in which prosodic features like pitch, formants, energy and speaking rate were used to predict personality (Mohammadi et al. [114]). Following the seminal work of Sacks et al. [151] and Duncan and Fiske [65] in conversation analysis, which showed how turn-taking behavior is a cue to the kind of social interaction, in a discussion, automatic analysis of conversations has recently shown how the relations between turns of different speakers (their smooth alternation vs. their overlapping) can tell us something about the recognition of roles in the discussion—who is the protagonist, the attacker, the supporter—(Vinciarelli [172]), the identification of dominant individuals (Jayagopi et al. [96]), and of fragments of conflictual interactions.

For example, when speakers start to talk faster, the general loudness of conversation increases, and when the turns of two participants overlap, especially if the turn overlapping lasts longer than expected, all this will tell you the conversation is becoming conflictual. Recent work in automatic conflict detection, beside taking into account turn-taking flow, identifies steady conversation periods, built on the duration of continuous slots of speech or silence, thus capturing the attitude of some participant to take the turn even when the interlocutor has not finished speaking (Pesarin et al. [123]). Results show that speech overlapping, especially if it lasts longer, clearly discriminates between conflictual and non-conflictual discussions.

8.6.2 Gestures

A "communicative gesture", or simply "gesture", is any movement of hands, arms or shoulder produced by a Sender to convey some meaning (Poggi [126]). On the signal side, any gesture can be analyzed in terms of a set of parameters: its handshape, location (the place over or around the Sender's body where the gesture is produced), orientation of palm and metacarp, and movement, including direction and trajectory, but also velocity, tension, repetition, fluidity (the so-called "expressivity parameters" of Hartmann et al. [89]), which generally impress emotional nuances to the gesture. Gestures are a very powerful means of expression and communication; they not only convey many types of meaning, but also contribute in the elaboration

and phrasing of thought (McNeill [110]). Types of gestures differ in terms of various criteria.

For instance, they can be "creative" (created on the spot) versus "codified" (steadily stored in memory, like words in a lexicon). We make a creative gesture when, depicting the shape or imitating the movements of an object, we produce an "iconic" gesture that represents meanings by reproducing their connected images. Typical codified gestures are "symbolic gestures" that convey the meaning of specific words and sentences, and have a shared verbal translation in a given culture. Gestures can also be either "motivated" (iconic or natural) or "arbitrary". A gesture is "iconic" when there is relation of similarity between form and meaning (e.g., *beating hands like wings* to mean "bird", or mimicking a cat climbing on a drainpipe), and "natural" when the relation is one of mechanic determinism (the gesture of elation of *shaking up arms*, determined by the physiological arousal of this emotion). A gesture is "arbitrary" if you cannot guess its meaning from its form. Most "symbolic gestures" are arbitrary, since, being codified in memory, they can afford not being iconic.

Another criterion to distinguish gestures is a semantic one. Like all signals, gestures may convey three types of meaning (Poggi [128]): information about the World (concrete and abstract entities and events, and their properties and relations), the Sender's Identity (sex, age, cultural roots, ethnicity, personality), and the Sender's Mind (his/her beliefs, goals and emotions concerning ongoing discourse). For example, among symbolic Italian gestures about the World, the gesture of extended index fingers, with palms down, getting closer to each other, paraphrased as *"se l'intendono"* (they have an understanding with each other) or *"c'e' del tenero"* (they are lovers) may be viewed as a "social signal", since it concerns a social relationship, while *extended little finger up*, "thin", denoting a physical feature of a person, is not a social signal. *Right fist beating on left palm*, which means "stubborn", a personality trait of not being easy to persuade, is "social". Gestures informing on the Sender's identity, like *fist raised up* (= I am a communist), to claim one's belonging to a political or ideological group; or *hand on heart*, a self-presentation of one's positive moral identity, may be viewed as "social gestures". Within information on the Sender's mind, *pulling back flat hands with palms forward*, which means "I apologize", and then greeting gestures, handing something, or showing a seat, communicate the performative part of a social act. *Raising a hand* to ask for the turn, or *pointing with open hand* to solicit someone to speak fulfill turn-taking functions. Among "creative" gestures, those invented on the spot and used during discourse, good candidates to be "social" gestures are those pointed out by Bavelas and Gerwing [19]: "interactive gestures" (for instance, pointing at the Interlocutor to acknowledge his suggestion), and "collaborative gestures" (like indicating the fold of a virtual origami depicted by your interlocutor in the air): they are "a dialogic event, created by the joint actions of the participants" (Bavelas and Gerwing [19], p. 292). But the "social" information contained in a gesture need not necessarily lie in its overall meaning, but also simply in some of its parameters. Suppose I am giving directions to my husband about where to find a tobacco shop, but I am angry at him for his smoking too much: the direction of my gesture tells him where

the tobacconist is, but its jerky movement might tell him my anger; the handshape and direction of movement inform about the world, but the expressive parameter of velocity and non-fluidity is a "social signal" of a social emotion.

One more case of a "social gesture". When a speaker makes gestures to illustrate his narrative, the interlocutor sometimes produces the same gestures of the narrator or different ones that anticipate ongoing narration, to show he is following and even predicting subsequent development. These are simply iconic gestures, concerning shapes or movements, but their use is "social", since by them the interlocutor gives a *backchannel*, thus helping social interaction.

8.6.3 Head movements

Head movements have been studied in a marginal way compared to facial expression and gestures (Heylen [91]), although they have lots of semantic, narrative and discursive functions in the process of communication. McClave [109] analyzed head movements in relation to their lateral movement and their orientation, by attributing to the former meanings of intensification, inclusion and representation of uncertainty, and to the latter the function of locating referents in an abstract space. According to Kendon [101], the head shake is anticipated or postponed to negation, for rhetorical purposes. Dittmann and Llewellyn [62], Hadar et al. [86] and Cerrato [44] focused on the interpersonal function represented by the synchrony voice-nod, which corresponds to the wish of the listener to speak or the wish of the speaker for feedback (Dittmann and Llewellyn [62]; Cerrato [44]), and measured the duration of the nod in relation to speech rate, while Boholm and Allwood [27] considered the functions of head nod and head shake repetition. Heylen [91] and Cerrato [44, 45] consider the nod as a backchannel signal that indicates acceptance, agreement and submission; maybe for this reason, in the context of power and gender communication, it is more frequent for women than men and for low than high status people (Hegen-Larsen et al. [90]). Cerrato [44] pointed out the most frequent head nod functions (giving continuation, giving agreement, requesting feedback and focus) and showed that a nod of "giving continuation" (0.40 msec) is briefer than one "giving acceptance" (0.60 msec), while Hadar et al. [86] classified the "synchrony" movements as having low amplitude and short duration compared to "anticipation" movements, of low frequency and large amplitude. Beside analyzing the head nod from the point of view of the signal, recent research emphasizes the effects of nodding in increasing agreement (Wells and Petty [175]) and changing attitude. Briñol and Petty [31] demonstrated that if one nods, as opposed to shaking his head, while another speaker is delivering a persuasive message, the headnod brings about a more persuasive effect than a headshake, since it increases confidence in oneself, becoming an internal cue to the validity of the message heard (*self-validation hypothesis*).

Poggi, D'Errico, Vincze [144] define a nod as a vertical head movement in which the head, after a slight tilt up, bends downward and then goes back to its starting

point. It is a holophrastic signal, since it conveys the meaning not of a single word but of a whole communicative act, including performative and propositional content (Poggi [128, 130]), which can be paraphrased, depending on the context, as "I confirm", "I agree", "I thank you" or other. This means that it is a polysemic signal, i.e., one corresponding to two or more meanings that are not completely unrelated but share some common semantic element.

Based on the analysis of 150 nods from the "Canal 9", a corpus of political debates available on the portal of the European Network of Excellence SSPNet,[2] a typology of nods was outlined, with each type characterized by subtle cues in other modalities (like parallel smile or blink behaviors) and by the context of production.

Some nods are produced by the present Speaker, some by the Addressee and some by a Third Listener, a simple bystander whom the present speaker is not addressing. For the Interlocutor's nods, those produced while the present Speaker is talking are nods of backchannel, while for those after the Speaker has finished talking the meaning depends on the speech act performed by the Speaker in the previous turn. So, like for *yes*, a *nod* following a *yes/no* question counts as a confirmation of the Speaker's hypothesis, one following an assessment or a proposal conveys agreement or approval; it is a permission after a permission request, submission after an order, *acknowledgement* or *thanks* after a prosocial act like an offer, and finally conveys a *back-agreement* (when the Listener agrees with what he had previously thought) or a *processing* nod (something like scanning the steps of one's reasoning). Also the Third Listener's nods convey a *confirmation* after an information, and *agreement* after an evaluative opinion.

The Speaker's nods, performed while holding the turn, include two broad families, with two semantic cores, respectively, of importance and of confirmation. Within the former, we may nod to emphasize what we are saying (*emphasis*), meaning "this part of my sentence or discourse is particularly important": the head goes up and down but also slightly forward, in correspondence with a stressed syllable and gaze toward the Interlocutor. Particular cases of emphasis are when we nod in correspondence of all the stressed syllables of our sentence (*baton*), and while listing (*list*), to convey we are mentioning items of the same list. Other nods of the Speaker are linked to confirmation. A nod while looking at the Interlocutor and frowning, or with oblique head, slightly tilted sidewise (*interrogative nod*) is a request for confirmation, just as a *"yes?"* with interrogative intonation, and can be used also as a rhetorical question (*Rhetorical interrogative nod*). Finally, a nod with an interrogative expression may ask confirmation that the other is following (*backchannel request*).

8.6.4 Gaze

Beside studies concerning vision (Yarbus [181]) gaze has been investigated, in psychology and communication research, as to its role in face-to-face interaction

[2] http://sspnet.eu/

(Kendon [99]; Argyle and Cook [7]; Goodwin [83]; Bavelas and Gerwing [19]; Allwood et al. [4]), in narration and persuasion (Poggi et al. [141]; Heylen [91]; Poggi and Vincze [139]) and as to evolutionary differences in gaze between humans and apes (Tomasello et al. [163]).

8.6.4.1 Gaze and Social Attention

Most studies on gaze focus on the genuinely social and interactive functions of gaze direction. Gazing at a person signals attention to the other, but also solicits shared attention (Argyle and Cook [7]; Trevarthen [168]). When a child wants to attract his mother's attention to something, he alternatively gazes at her and at his object of interest. That autistic children tend not to gaze to their interlocutor has led to think that mutual gaze is linked to the construction of the Theory of Mind (Baron-Cohen [16]) the capacity to imagine the other's mental states.

If gazing at another is the core of social attention, specific uses of gaze direction regulate face-to-face interaction, working as an important set of signals for turn-taking and backchannel: by looking at the interlocutor we ask him to follow or to provide feedback; by averting gaze we tell we are retrieving words, and by doing so we keep the floor; while if we are the interlocutor, by gazing at the speaker we assure him of our interest and attention.

These findings gave rise to gaze tracking studies to detect mutual attention, interaction regulation, and early symptoms of autism (Boraston and Blakemore [28]).

In the field of Embodied Agents, the role of gaze direction, especially if compared to face and body direction, has been stressed as a sign of interest (Peters et al. [124]), and its use in face-to-face interaction and backchannel has been simulated in Virtual Agents (Cassell [39]; Bevacqua et al. [25]).

8.6.4.2 Lexicon and Parameters of Gaze

Beside establishing shared attention and the setting for interaction, gaze conveys specific meanings. Eibl-Eibesfeldt [67] and Ekman [71] analyzed some conversational, emotional and syntactic functions of the eyebrows; Sign Language scholars (Baker-Schenk [12]) studied the syntactic and semantic role of gaze in ASL (American Sign Language) and LIS (Italian Sign Language). For the Hearing (non-deaf) people, the repertoire of gaze meanings was investigated by Kreidlin's "Oculesics" [105], and Poggi [128] proposed to write a lexicon of gaze, arguing it is a communicative system as complex and sophisticated as facial expression or gesture can be. In gaze, the signals—the perceivable stimuli an interlocutor can see and interpret by attributing them some meaning—are morphological features and muscular actions exhibited in the eye region, which includes *eyebrows*, *eyelids*, *eyelashes*, *eyes* and *eye-sockets*. The meanings are imagistic or conceptual representations linked to those signals.

The parameters proposed by Poggi [128] to analyze the signals of gaze are:

1. movements of the eyebrows (e.g., eyebrow frowning means worry or concentration, eyebrow raising, perplexity or surprise)
2. position, tension and movement of the eyelids (in hate one lowers upper eyelids and raises lower eyelids with tension; in boredom upper eyelids are lowered but relaxed)
3. various aspects of the eyes: humidity (bright eyes in joy or enthusiasm), reddening (bloodshot eyes in rage), pupil dilation (a cue to sexual arousal); focusing (staring out into space while thinking), direction of the iris with respect to Speaker's head direction and to Interlocutor (which allows us to point at things or persons by using eyes, like we do with gestures)
4. size of eye sockets (expressing tiredness)
5. duration of movements (a defying gaze focuses longer over the other's eyes).

On the meaning side (Poggi [128]), fragments of a lexicon of gaze were described, within which several uses of gaze convey "social" meanings. Ethnicity for instance is conveyed by *eyelid shape*; *bright eyes* reveal aspects of personality. Some gaze items communicate the performative part of our sentence (*you stare at the Interlocutor* to request for attention), others, turn-taking moves (*you gaze at present Speaker* to take the floor) and feedback (*frowning* expresses incomprehension or disagreement, raising eyebrows with half-open eyes, perplexity). The meanings of specific gaze signals have been investigated by empirical and observational studies. In a study on the degrees of aperture of upper eyelids (wide-open, half-open, half-closed) and lower eyelids (lowered, default, raised) it was found that the *half-open* upper eyelids convey de-activation, referred to a physical (*sleepy, exhausted*), cognitive (*how boring*) or emotional state (*sad, I am sorry*, *I couldn't care less*), but the combination with *raised lower eyelids* adds a component of effort (*I am trying to remember*, *I am about to cry*) (Poggi, D'Errico, Spagnolo [142]).

Within eye-closing behaviors, the blink, a rapid closing of the eyes that in general has a bare physiological function of keeping standard eye humidity, may convey agreement, and accompanying or substituting a nod (Vincze and Poggi [174]). Yet, blink rate has also been found to be a cue to deception: people while deceiving are concentrated and do not blink, but when the deception is over they blink more frequently to compensate (Leal and Vrij [106]).

On the other hand, the wink, a rapid asymmetrical closing of a single eye, conveys allusion and complicity. It addresses only one specific Addressee, with whom the Sender feels syntonization and complicity, while excluding others: it then appears furtive and allusive, implying inclusion of Sender and Addressee in the same group, and exclusion of a third party (Vincze and Poggi [174]). Sometimes, mainly if accompanied by a smile, it conveys *playful complicity* and can be exhibited also (or mainly) to the third party, thus making it clear that the Sender is just kidding. The *warning* wink, instead, which warns a confederate (the Addressee) about something concerning an "enemy", must exclude the enemy and hence be concealed from him, to be perceived only by the confederate.

8.6.5 Posture

Postures are complex and multifaceted social signals expressing interpersonal attitude, relations and emotions. In their description various parameters must be taken into account: arms (open, closed, on hips), trunk (backward, forward, and lateral), head (downward, upward and lateral), and legs (open, crossed, extended).

Concerning their meaning, postures have been studied by Mehrabian [111] and Argyle [6] mainly in relation to two main dimensions: status and affiliation. High status is expressed through space, by enlarging the body (rising up to full height, wide legs) and through a relaxed body (leaning, sitting and asymmetric postures). In this sense, according to Argyle, posture corresponds to a reflection of established hierarchy: lower status people tend to be more tense, nervous and aggressive because they have to achieve material and symbolic resources.

From a relational point of view Scheflen [153] and Kendon [100] analyzed the similarity and orientation of postures: when two interactants share the same centre of attention, posture similarity seems to be a reliable signal of quality of relation, and it even induces more positive relations.

Mehrabian [111] identified postures of attraction and intimacy describing proximity and forward inclination, gaze and orientation toward the interlocutor. He observed that women tend to be more intimate than men; this tendency seems to be correlated to the patterns of low status persons, confirming that gender behaviors and signals are akin to status signals (D'Errico [59]).

Posture, when dynamically considered, has a relevant function in the management of turn taking: posture shift is correlated with topic shift (Condon and Ogston [49]) and with a "situational" change, for example in a temporary change of rights and obligations between the interactants (Blom and Gumperz [26]), and they are more frequent at the start of turns (48%) independent of the discourse structure (Cassell et al. [38]).

Finally, emotions are expressed by postures. Lowered shoulders, for example, signal depression (Tomkins [164]; Badler et al. [11]): "cognitive" emotions like interest and boredom are clearly identified by Bull [34] who described the bored posture as head downward or left-rightward (leaning on one hand) and extended legs. In shame one lowers head, while in pride head is raised and bust erected (see below).

According to Kleinsmith and Bianchi-Berthouze [104], posture is a very reliable cue to affect: considering nine categories of particular emotions (angry, confused, fearful, happy, interested, relaxed, sad, startled and surprised) they found that, in particular, extension of the body (lateral, frontal and vertical), body torsion, inclination of the head and shoulders in 70% of cases allow to assess the classical dimensions of emotions of valence, arousal, potency and avoidance.

8.6.6 Proxemics and Touch

Other important social signals can be found in proxemics (Hall [87]), physical contact, and posture. In his seminal work, Hall [87] introduced the notion of Proxemics,

the set of rules that regulate the use of space and distance between people. He found that people keep different distances with their interlocutors in their face-to-face interaction, depending on the kind of social relationship they entertain with them. He distinguished intimate distance (0–45 cm) from personal (45–120 cm), social (120–360 cm), and public distance (360–750 cm), but he also found out that the distance considered acceptable for these different face-to-face interactions is highly determined by culture. People from Mediterranean and African cultures, for example, compared to English or Scandinavian people, tend to speak so close to each other as to sense each other's smell, and often tend to touch each other while talking. Since people tend to conform to these rules, their spatial behavior can be taken both as a communicative social signal of the social relationship they want to entertain with the interlocutor, and as an informative signal of their cultural roots.

Strictly connected with spatial behavior is physical contact between interactants. "Haptic" communication concerns the signals perceived through the sense of touch. One of the first to develop in humans, this sense provides an important route of communication between mother and child and the basis for a strong attachment bond (Bowlby [30]), and later in adult life being touched by other accepted people or by oneself gives a person a sense of reassurance (Montagu [115]). In this, touch may be seen as a "social signal" par excellence.

Within the acts of touch performed by a person on another, some are not aimed to communicate but to grasp (e.g., grabbing the arm of a thief that has just stolen your videocamera), to sense (a blind touching a person to sense who he is), or to feel (like in erotic intercourse). But in other cases, touch is communicative (Kreidlin [105]), and it is possible to find out a lexicon—a set of correspondence rules between specific acts of touch and their meanings—and a "phonology" ("haptology")—a set of parameters of the act of touch (Poggi [128, 129]). In fact, depending on the way one touches the other and the touching and touched body part, different acts of touch convey different meanings (a slap tells you something very different from a caress), and various communicative acts (request and offer of help and of affect, proposal, sharing), hence establishing various kinds of social relationships with the Addressee: affiliative, friendly, protective, aggressive and so forth. Moreover, touch signals are subject to specific norms of use, varying across cultures, as to who may touch whom, and where, based on their social relations. Also in this case, then, whose and which part of the body is touching and touched may work as communicative and/or as informative signals.

8.7 Social Facts and Their Signals

We have defined as "social" the signals concerning "social facts", namely *social interactions, social attitudes, social relations and social emotions*. Let us explore these contents and the social signals that inform or communicate about them.

8.7.1 Social Interaction

A social interaction is an event in which two or more agents perform reciprocal social actions, that is, actions in which the goal of one participant is directed to the other by considering him as an autonomous agent, one regulated by one's own goals. A football game, a surgery operation, a string quartet, a fight, a sexual intercourse, a school class are social interactions. But since, as demonstrated by Nass and Steuer [116], also computers are seen as "social actors", a dialogue between an Embodied Agent and a User, or one between two robots, also fits the definition. Generally social interactions are or require communication; all require synchronization, i.e. mutual reactions between interaction participants, and negotiation of the participants' roles.

Some typical social signals exchanged during a communicative interaction are those for turn-taking and backchannel. The turn taking system is a set of rules to state who is to speak, and compliance to it is conveyed by nonverbal signals: you may ask for turn by handraising, mouth opening, gaze direction, variation of vocal intensity (Duncan and Fiske [65]; Goodwin [83]; Thòrisson [162]; Allwood et al. [4]). But the exploitation or even the violation of turn-taking rules is by itself a social signal: turn interruption may be an informative or communicative signal of aggressiveness, turn overlapping an informative signal of a competitive interaction. Yet, as all signals, also these must be interpreted while taking other elements into account: if you know the one who interrupts is a very close friend of the interrupted, he might be simply completing the other's sentence, thus giving a signal of high syntonization.

Another set of behaviors enabling smooth interaction are *backchannel* signals, through which the interlocutor communicates to the present speaker if he is listening, following, understanding (Yngve [182]; Allwood et al. [3]), possibly believing, finding interesting, and agreeing (Poggi [131]; Bevacqua [24]), by making use of hesitations, interjections, fillers, affect bursts (Jucker [97]; Bazzanella [21]; James [95]; Poggi [130]; Schröder [156]; Schröder et al. [157]) head movements (Heylen [91]) and smiles (Goodwin [83]; Bavelas et al. [20]; Chovil [46]). But also in this case, giving backchannel is not only telling the other if you are following: it may be in itself an indirect signal that you accept the other, you are empathic with him or her: you care.

In group interaction, the interactants come to assume spontaneous or institutionalized roles that are instrumental in group functioning and to pursuing the group's goals. These roles can be reflected by the verbal messages exchanged during interaction. According to Benne and Sheats [22], based on the particular statements they tend to use, participants can be attributed the functional roles of *harmonizer*, *encouraging*, *compromiser*, correlated to the so-called group task roles of *elaborator*, *coordinator* and *orienter*. In this interactional framework, even if these authors do not indicate any normative duties for the achievement of the group goals, they identify also negative functional roles such as *dominator*, *aggressor* and *recognition seeker*. As mentioned above, work is presently being done in automatic role detection.

8.7.2 Social Attitudes

A social attitude has been defined in classical Social Psychology as the tendency of a person to behave in a certain way toward another person or a group. Social attitudes include cognitive elements like beliefs, evaluations, opinions, but also emotions, which all determine and are determined by preferences and intentions (Fishbein and Ajzen [74]). Here we overview some studies on the signals used to persuade and those to convey agreement.

8.7.2.1 Persuasion

Persuasion is a communicative action aimed at changing people's attitudes, that is, at influencing their tendency to action by changing their opinions: we persuade as we influence another to do or not to do something by inducing him to conclude that what we propose is good for him. Since classical rhetoric (Aristotle 360 B.C. [9]; Cicero 40 B.C. [47]; Quintilian [146]), the study of persuasion has been a major topic, in social psychology and media studies (Petty and Cacioppo [125]; Fishbein and Ajzen [74]), linguistics, argumentation (Perelman and Olbrechts-Tyteca [122]; Toulmin [165]; van Eemeren and Grootendorst [170]), and cognitive science (Castelfranchi and Guerini [41]; de Rosis et al. [56], Miceli et al. [113]).

In the last decade Fogg [77], investigating the role of computers as persuasive social actors, opened the field of *Captology* (acronym of Computer As Persuasive Technology), based on the assertion that also technologies persuade by giving a variety of social cues that elicit social responses from their human users. Fogg applied powerful persuasive strategies to human–computer interaction by using psychological cues based on attractiveness, similarity and reciprocation principles, and he found many communalities between persuasive technologies and human-human persuasion.

Persuasion as Influence Over the Other's Goals

In the model we adopt (Poggi [127]), a persuader A aims at influencing a persuadee B, i.e., at increasing or decreasing the likeliness for B to pursue some goal G_A proposed by A. To pursue G_A, B must believe it is a valuable goal, because it is a means for some other goal G_B that B already has, and/or because it makes B feel some positive emotion or prevents some negative one. Emotions have a high motivating power because they trigger goals (Miceli et al. [113]). Among other forms of influence—from education to threat and promise, from manipulation to the use of force—persuasion is characterized by three features: 1. communication: in trying to induce B to pursue G_A, A makes clear to B he wants to do so; 2. freedom: A leaves B free of pursuing G_A or not (thus differing from threat); 3. disinterest: A tries to convince B that G_A is in the interest of B since it is a means for some goal G_B that B has.

To persuade B, A can use the strategies highlighted by Aristotle (360 B.C.): *logos* (the logical arguments that support the desirability of G_A and the means-end link between G_A and G_B); *pathos* (the positive emotions B might feel, or the negative he might avoid by achieving G_A); and *ethos*: in Aristotle's terms, the character of the Persuader; in our terms, *"ethos-competence"*, A's intellectual credibility, his having the skills necessary for goal choice and planning, and *"ethos-benevolence"*, his moral reliability: the fact that A does not want to hurt or cheat B, or to act in his own concern.

We persuade by producing multimodal persuasive discourses, i.e. sequential and/or simultaneous combinations of communicative acts in various modalities. For example, in a pre-election discourse, all the sentences, gestures, face and body movements of a politician, at various levels of awareness, through their direct and indirect meanings pursue *logos*, *ethos* or *pathos* strategies, trying to convey "I ask you to vote for me". But in this case, is only the combination of all signals "persuasive", or can we say that some words, or some gestures, intonations, gaze items, are in themselves persuasive?

Persuasive Gestures, Persuasive Gaze

Recent work (Poggi and Pelachaud [138]; Poggi and Vincze [140]), analyzing multimodal communication in political persuasive discourses by Italian and French politicians, found out that only rarely are some words, gestures or gaze items "persuasive" by their very meaning—for example, a gesture of incitation, or a word of encouragement; rather, we may call "persuasive" some uses of gaze, some gestures or sometimes simply some parameters of their expressivity, which convey "persuasive information", i.e., information that is salient in persuasive discourse. First, information relevant to pursue a *logos* strategy:

1. *Importance* of the goal proposed by the persuader, borne by gestures or gaze items conveying performatives of incitation or request for attention, like beats or eyebrow raisings that convey emphasis, but also by the irregularity or discontinuity of gesture movements that capture attention.
2. *Evaluation*. Persuading implies inducing positive evaluations of objects, persons, events, so all words, gestures, and other signals mentioning evaluations have a potentially persuasive import.
3. *Certainty*. Persuading implies convincing, i.e. making someone believe, with a high degree of certainty, what goals are worth to pursue (their value, importance) and how to pursue them (means-end relationships). Gestures with a meaning of high certainty, like the *ring* (thumb and index making a circular shape going down and up; Kendon [102]) that conveys precision and commitment to what one is saying, or a *small frown*, which means "I am serious, not kidding", may be persuasive. Yet, one may also indirectly convey *certainty* by pursuing an *ethos* strategy, e.g. showing self-confidence about what one is saying by exhibiting an *easy posture* or a *fluid speech rhythm*.

Other meanings that bear on an *ethos* strategy are:

4. *Sender's benevolence and competence*. To be persuaded we do not only evaluate the goals proposed or the means to achieve them, but the persuader: the Sender's *ethos*, which encompasses his *benevolence*—his taking care of our goals—and *competence*—his having the skills to do so. A gesture driven by an *ethos benevolence* strategy, namely, showing one's moral reliability, quite frequent in political communication, is *hand on heart* (Serenari [158]), generally meaning "I am noble, fair, reliable". A gesture evoking *ethos competence* strategy is one by the Italian politician Silvio Berlusconi who, in talking of quite technical things concerning taxes, *rotates his right hand curve open, palm to left, rightward twice*, meaning that he is passing over such technicalities, possibly difficult for the audience; his relaxed curve movement indirectly communicates how smart he is, talking of such difficult things easily and unconstrained. This projects an image of competence.

When exploiting a *pathos* strategy, the persuader mentions or evokes emotions:

5. *Emotions*: expressing an emotion may induce it by emotional contagion and hence trigger the desired goal. The Italian politician Romano Prodi, while talking about his country, *moves his forearm with short and jerky movements of high power and velocity* to convey his pride of being Italian and transmit it to the audience, to induce the goal of voting for him.

Based on these principles, the persuasive use of gesture and gaze was investigated in some fragments of pre-electoral interviews in Italy in 1994 and 2006 (Achille Occhetto and Romano Prodi) and in France in 2007 (Ségolène Royal). In the annotation scheme used for the analysis, beside a transcription of the verbal context, each gesture or gaze item was described in terms of its parameters and a verbal paraphrase of its literal and possibly indirect meaning was classified in terms of a semantic taxonomy (Information on the World, the Sender's Identity and the Sender's Mind, Poggi [128]), and in terms of the persuasive strategy pursued: *logos*, *pathos*, *ethos benevolence*, or *ethos competence*.

For example, Ségolène Royal, while talking of the top managers who spoil the enterprises like Mr. Forgeat, *looks at the Interviewer* Arlette Chabot, with a *fixed gaze* which means "I am severe, I do not let you avert gaze": this conveys information about Royal's personality, her being serious and determined, aimed at a strategy of *ethos competence*, possibly indirectly implying she is one who struggles against injustice: one more information on her *ethos*, but on the moral side, *benevolence*. Then Royal, while *leaning her head on the left*, *looks at the Interviewer obliquely* and *with half-closed eyelids*, an expression of anger and indignation: information about her emotion, which she possibly wants to induce in the audience, thus pursuing a *pathos* strategy.

By computing gesture and gaze items in the fragments analyzed, you can single out patterns of persuasive strategies in the subject observed. From the studies above it resulted that the Italian politician Achille Occhetto has a higher percentage of persuasive gestures than Prodi out of the total of communicative gestures (Occhetto 20 out of 24, 83%, Prodi 34 out of 49, 69%), also because Prodi sometimes uses iconic gestures that convey Information on the World and have no persuasive import

except for some in the expressivity parameters. Further, Occhetto relies much more on *pathos* than on *logos* gestures (30% vs. 5%) while Prodi uses the two strategies in a more balanced way, but with a preference for *logos* (23% vs. 12%). In both most gestures (65%) pursue an *ethos* strategy, and both tend to project an image of competence more than one of benevolence, but more so for Prodi (50% vs. 15%) than for Occhetto (45% vs. 20%).

The differences in the patterns of persuasive gesture and gaze of the politicians under analysis are coherent not only with the argumentative structure of the fragments analysed, but also with the politicians' general persuasive style, as well as with their political history. For example, since in the fragment analyzed, Occhetto is attacking his opponent Berlusconi from an ethical point of view, he aims to project an ethically valuable image of himself; Prodi instead is describing his program and thus wants to project the image of one able to carry it on in an effective way. In terms of political strategies, compared to Prodi, a centre-left politician coming from a former catholic party, the communist Occhetto has a higher need to show his image of benevolence.

8.7.2.2 Signals of Agreement

When persuasion, our attempt to change the other's opinion and tendency to action, succeeds, the other finally agrees with us. But what is agreement, and what are its signals? How can one catch not only clear-cut but also subtle cases of agreement and disagreement, expressed, directly or indirectly, in words, gesture, intonation (Ogden [119]), face, gaze, head movements, posture?

From a cognitive and social point of view, agreement occurs when there is a relation of identity, similarity or congruence between the mental states of two or more persons. Yet, different from an act of confirmation, a communicative act of agreement—and its underlying cognitive state—may not occur about "factual" beliefs, i.e. about simply informative speech acts (I cannot agree after a question like *"Did Napoleon die in 1821?"* nor after a statement like *"Napoleon died in 1821"*, unless someone challenges this as not a factual belief but a questionable statement), but only about speech acts like a proposal, an assessment (i.e., the expression of some evaluation) or the expression of an opinion (for instance, after a sentence like *"I think that Napoleon was a great man"* or *"I propose that all teachers give a home assignment about Napoleon"*).

An opinion is a "subjective" belief, that is, one we know is not necessarily shared, and have no empirical evidence for, firstly because it is not about something that can be perceived by senses. It is a belief we draw concerning some entity or event by considering it from our particular "point of view", somehow determined by our beliefs and goals (Poggi et al. [143]). And since this opinion may be our point of view not only about facts, but also about goals or evaluation, we can also agree about another's proposal, assessment, or an opinion thereof.

An evaluation is a subjective belief concerning how much some entity or event has or gives the power to achieve some goal, and it may typically be the object of an

opinion. A proposal is a requestive speech act (one asking to pursue some goal), in which 1. the goal "proposed" is in the interest also of the Addressee, 2. the Sender does not intend to make use of power over the Addressee, who is thus free to pursue the proposed goal or not; 3. acceptance implies that the Addressee approves the proposal, i.e. s/he also believes it is functional to her/his goals too.

In conclusion, agreement is an internal mental state of some agent B: B's assumption about his having the same opinion as another agent A. This assumption may be communicated by B to A or to others through an "expression of agreement", i.e. a simple or complex social communicative act, composed by verbal and/or body signals.

An observational study on the political debates in the "Canal 9" corpus (Poggi et al. [143]) found out that during a debate a participant may express agreement in at least three ways:

1. by *discourse*: one or more sentences that express an opinion similar or congruent with one previously expressed by another participant;
2. by verbal expressions containing specific words, *verbal agreement markers*, like "*ok*", "*I agree*", "*oui*" or others:
3. by *body agreement markers* like *nods*, *smiles*, *gestures* or *gaze signals*.

Agreement is expressed *by discourse* when one or more sentences of a participant either literally repeat or rephrase an opinion expressed earlier by another participant. But it can also be expressed only by *agreement markers*: words or constructions with a semantic content of agreement, like (in French) *d'accord* (ok), *oui* (yes), *vous avez absolument raison* (you are absolutely right), *nous sommes d'accord* (we agree), *je vous rejoins* (I join you [in believing x]), *effectivement* (in fact), *(bien) évidemment* (obviously), *tout-à-fait* (absolutely).

A frequent *agreement marker* is *"d'accord"* (= ok; I agree), which though, to mean agreement, must be used in a performative way, that is, as meant by the same person who is speaking, not as another's reported agreement (e.g., *"(Je suis) d'accord"* as opposed to *"Il est d'accord"*). In other cases, *d'accord* counts more as an acknowledgement of what another Speaker has just said: a backchannel signal rather than real agreement.

Typical signals of agreement are *smiles*, *eyebrow raisings* and *nods* (see Bousmalis et al. [29] for a survey), but also some gestures: for instance, *moving right hand forward*, as if presenting and showing what the other is saying as a good example of what you also think; or, again, *raising hands with open palms up*, which means "this is evident", while another is speaking, to underline one totally agrees with what he is saying. Eye and mouth behaviors may convey agreement too. Within gaze signals, the *closing of the eyelids* is relevant: when agreeing with the present Speaker, the Interlocutor's nods are often accompanied by *rapid blinks*, or by *wide open eyes*, usually with *raised eyebrows*, which emphasize the extent to which one agrees with the other: eyelid behavior as a nod intensifier. Yet, *blinks* as cues of agreement can appear also by themselves, without being accompanied by a *nod*. As to mouth behavior, agreement is often conveyed by *smiling*, but sometimes also by lip pressing which, if accompanying a nod, emphasizes agreement.

From a semantic point of view, these signals may convey True, Indirect or Apparent agreement. Sometimes, in fact, a bodily or verbal expression of agreement can be taken at face value (*true agreement*), in either stronger or weaker forms (*enhanced* and *unwilling agreement*). We talk of *enhanced agreement* when people do not simply communicate that they share the other's opinion, but provide additional arguments or emphasize their verbal agreement by *smile* or *eyebrow raising*. One may express *unwilling agreement*, instead, by admitting the other is right, but only by nonverbal signals (e.g. by *stepping back* or *lowering head*) to keep a low level of commitment and minimize the self-humiliation implied in acknowledging one was wrong. On the other hand, sometimes agreement is expressed indirectly: no apparent agreement marker is produced, but substantive agreement can be inferred from the global meaning of what is literally communicated by words or body signals. Other times, finally, agreement is only local, partial or hypocritical, while it actually masks indirect disagreement (*apparent agreement*): this is the *"Yes, but..."* strategy, which makes use of various stratagems, like uttering an agreement marker (e.g., the French *effectivement* = in fact) with a *suspensive intonation* that announces reversing the polarity from agreement to disagreement, or using expressions (like *je dois dire* = I must say) that limit the scope of one's agreement.

8.7.3 Social Relationships

A social relationship is a relation of interdependency of goals between two or more persons: one in which the pursuit, achievement or thwarting of a goal of one determines or is determined by the pursuit, achievement or thwarting of a goal of the other (Lewin [107]; Festinger et al. [73]; Byrne [36]).

Types of social relationship have been distinguished in terms of criteria like public vs. private, cooperation vs. competition, presence vs. absence of sexual relations, social-emotional support oriented vs. task oriented (Berscheid and Reiss [23]). Within group relationships, some studies concern the definition and description of mechanisms of power, dominance, and leverage (Castelfranchi [40]; Lewis [108]), their change and enhancement through alliance, influence and reputation (Conte and Paolucci [51]), their interaction with gender relations, and the nature of leadership.

Typical signals revealing social relationships include the manner of greeting (saying *hello* signals the wish for a positive social relation, saluting signals belonging to a specific group like the army, etc.), the manner of conversing (e.g., formal allocutives like addressing someone as *professor* to signal submission), mimicry and the display of typical group behavior (signaling the existence or wish of a positive social relation), spatial positioning and gaze direction (e.g., making a circle around a certain person, or gazing at her more frequently, indicates her as the group leader: Argyle and Cook [7]), physical contact (that one touches another, and the way one does, may indicate affective or power relations: Hall [87]; Poggi [128]).

For group relationships, both deliberate and unaware signals, like dress or haircut, vs. regional accent and mimicry, reveal felt or wished group belonging. Em-

blems on clothes, elaborate hair, or objects like a crown or a huge desk in the office reveal status or role in the group (Hinde [93]; Halliday [88]).

To provide a blow-up on the complexity of social signals in this domain, let us focus on the relation of dominance.

8.7.3.1 Dominance and Its Signals

The notion of *dominance* reflects different research approaches and is sometimes confused with notions like *status* or *power*. In the sociological perspective, *status* is a hierarchical position in a group or organization, determined by native (e.g., gender or ethnicity) or gained characteristics (e.g., skill in work). In social psychology, according to the *expectation states theory* (Ridgeway [148]) at the interpersonal level people form *expectations of status*, evaluative beliefs about positive or negative competences associated to this nominal feature, and at the personal level they have *expectations of performance*, which anticipate the contribution needed for a specific task. In the *social identity theory* (Tajfel [160]), the awareness of belonging to a social group is a central part of the concept of self, with associated emotional, motivational, behavioral responses; so people tend to evaluate the *stability* and *legitimacy* of status differences to decide what cognitive strategy is useful in their condition: re-categorization, social creativity, individual or collective mobility across the hierarchy (Tajfel and Turner [161]).

Power is defined as "the ability to influence or to control other persons or groups" (Ellyson and Dovidio [72]). Status may well be a condition for power in this sense, but does not necessarily imply attitude change and control, and it is focused not on personal competence, but on a nominal or structural position in a social group or institution.

Dominance might be seen as a combination of status and power since it is defined as "ability to influence or control others", but it also involves *groupness*, since it concerns power relationships within a relatively enduring social organization (Ellyson and Dovidio [72]). As to its roots, some authors view dominance as a personality trait (Pratto et al. [145]), stressing its being a steady feature of an individual, others propose a situational view: dominance as gained from time to time depending on the context (Aries et al. [8]; Burgoon and Dunbar [35]; Pratto et al. [145]).

In recent literature, for the *social dominance theory* (Pratto et al. [145]) one possible explanation of discrimination phenomena is the psychological construct of *social dominance orientation* (SDO), i.e. the personal preference for hierarchical relationships between social groups. The degree of social dominance is determined by group membership because members of more powerful groups are more dominant than less powerful ones (e.g., men more dominant than women); further, SDO is a way to maintain social hierarchies, since people with high levels of social dominance tend to legitimate racism, nationalism and conservatism.

The *dyadic power theory* (Dunbar et al. [64]) proposes the notion of *interpersonal dominance* as "a relationally-based communication strategy dependent on the context and motives of the individuals involved" (Burgoon and Dunbar [35]), viewing dominance as a dynamic combination of personal and contextual characteristics,

based on a relational model, according to which the influence or control of powerful individuals depends on the submission or acquiescence of others. From this perspective much research has focused on the verbal and nonverbal indicators of dominance (for ample reviews see Ridgeway [147]; Argyle [6]; Dunbar and Burgoon [63]; Dunbar et al. [64]).

Signals of dominance in various modalities have been explored. Within studies on gaze, Keating and Bai [98] demonstrated that in Western cultures lowered eyebrows are perceived as a strong signal of dominance. Argyle [6] pointed out that the dominant person gazes less and during interaction reduces the amount of gaze and breaks mutual gaze first. Yet, in close relationships the dominant person has a more expressive face, he looks more than the less dominant and shows higher *visual dominance*, i.e. higher looking while speaking than while listening (Ellyson and Dovidio, [72]; Dunbar and Burgoon [63]). As to hand movements, the dominant person uses more gestures, and within them, more illustrators than adaptors (Dunbar and Burgoon [63]). Since illustrators are the gestures that accompany speech by adding information of an imagistic kind, while adaptors (Ekman and Friesen [68]) are hand movements onto one's own body or objects performed by the speaker to feel more at ease or to reassure oneself, a lower use of adaptors gives an impression of relaxation and confidence. Posture and spatial behavior are salient in the expression of pride (Tracy and Robins [167]), where expanded postures are typical especially in males (Cashdan [37]). In vocal behavior, dominance passes through speech intensity, tempo and pitch (Ridgeway [147]; Gregory and Webster [84]), but also through turn taking management (Jayagopi et al. [96]): perception of dominance is strictly connected to amount of speaking (Stein and Heller [159]), topic introduction (Brooke and Ng [32]), frequency and maintenance of turns, and interruptions (Ng and Bradac [117]).

8.7.3.2 Blatant and Subtle Dominance Strategies

Starting from a definition of dominance as "the fact that an Agent has more power than another" (not necessarily as a stable trait, but also in a specific context), Poggi and D'Errico [137] analyzed signals of dominance in political debates. They singled out various "dominance strategies", sets of behaviors that all bear a message of dominance, "I have more power than you", but each conveying, in one or more modalities, directly or indirectly, a specific message, attitude or image. They distinguish blatant from subtle strategies. Among the former, one is *aggressiveness*, which includes:

a. *imperiousness*, expressed by requestive communicative acts and deontic words like *must*, *ought to*, *necessarily*, hence conveying the message "I give you commands, → I can afford to do so → I have power over you"
b. *judgment*, expressed by insults and evaluative words like *praiseworthy* of *filthy*, frowning and facial expressions of severity, which tells "I can judge you, so I have power over you"

c. *invasion* of the other's space and time territory, performed by loud voice, ample gestures, turn interruption and overlapping
d. *norm violation*, like to go on speaking when the moderator gives the turn to another participant: violation of generally accepted rules implicitly conveys the idea that one is so strong as to be above rules.

Another blatant dominance strategy, mainly used from a down-to-up position, is *defiance*, conveyed by expressions of pride like fixed stare and erected posture, which communicates: "you are not stronger than I, I will finally gain power over you".

Among "subtle" dominance strategies, one is *touchiness*, i.e., to show one feels offended even for slightly negative evaluative words. Being touchy means to have a low threshold for feeling offended, and you feel offended when you think that some communicative or non-communicative action caused a blow to your image. Since there is a somehow a direct correlation between severity of an offense and power of the offended one, to show you are powerful and worth respect you simply need to show that you feel offended for things that would not be so serious for other people.

Another subtle strategy is *victimhood*. Playing the victim implies that others unduly did wrong to you, so you are entitled to retaliate and to claim your rights. On the other hand, in *haughtiness*, the Sender wants to convey his superiority, but not through boasting, rather through a prig and didactic attitude, as if others were all children or stupid; by *explaining things clearly*, using gestures like the *"ring"* (a circular shape made by thumb and index fingertips touching each other) that evoke precision and seriousness; sitting down with *trunk backward*, as if withdrawing from the other to avoid contact; *half-closed eyelids*, which by conveying relaxation mean "I need not worry about you"; in sum, conveying the other is so inferior that you do not bother about him at all.

Also *ridiculization* and *irony* are dominance strategies. Laughter is an emotional expression triggered by surprise and then relief, caused by an incongruous event that leaves you in a suspension but then turns out to be not dangerous, so the previous worry results in relief and in a sense of superiority over the event or its cause. Thus, one who *laughs at another* feels (and shows himself to be) superior to him, while the other feels impotent—he does not even have the power to scare or worry anyone!—and abased. Also irony, to the extent to which it is a way of teasing, of making fun of another, is a dominance strategy in which aggressiveness is masked by the elegance of a rhetorical figure.

Finally, *easiness* (expressed by a *loose and relaxed posture*) conveys "I am satisfied, I do not depend on you, you have no power over me"; *carelessness* typically entails *not gazing at the opponent*, as if he did not exist—the worst of insults! In *assertiveness* and *calm strength*, *low voice*, *relaxed and fluid gestures* convey self-confidence and no fear of the other, hence superiority.

8.7.4 Social Emotions

Emotions are an adaptive device that monitors the state of our most important goals: they are multifaceted subjective states, encompassing internal feelings and cognitive, physiological, expressive, motivational aspects, which are triggered any time an important adaptive goal is, or is likely to be, achieved or thwarted.

Within human emotions we may distinguish "individual" ones and three types of "social" emotions (Poggi [128]). First, those that are felt *toward* someone else; in this sense, while happiness and sadness are individual emotions, admiration, envy, contempt, compassion are social ones: I cannot admire without admiring someone, I cannot envy or contemn but someone, while I can be happy or sad myself. Second, some emotions are "social" in that they are very easily *transmitted* from one person to another: like enthusiasm, panic, or anxiety. A third set are the so-called "*self-conscious* emotions" (Lewis [108]), like shame, pride, embarrassment, which we feel when our own image or self-image, an important part of our social identity, is at stake, and thus concern and heavily determine our relations with others.

We define as *image* (Castelfranchi and Poggi [43]) the set of evaluative and non-evaluative beliefs others conceive of about us, and as *self-image* the evaluative and non-evaluative beliefs that we have about ourselves. Functional to our interaction with other people, we form a "goal of image", the set of standards against which we want to be evaluated positively by others, and a "goal of self-image", those against which we want to evaluate ourselves positively. These are very important goals for our individual and social life: we have a high self-esteem when we evaluate ourselves as positively as we wish to, and a good level of esteem by others when others evaluate us positively. Being esteemed by others is important to have good relationships with them, to be accepted in the community and obtain their help and cooperation. On the other hand, a high self-esteem is functional to be so self-confident as to confront challenging goals, and to be autonomous, not too dependent on others' help.

Given their importance, any time the goals of image or self-image are at stake, the "self-conscious" emotions are triggered. Two such emotions are shame and pride. We feel shame when our goals of image and/or self-image are (or are likely to be) thwarted, and pride when they are achieved.

8.7.4.1 Shame and the Multimodal Discourse of Blush

Shame is a negative emotion we feel when our goal of image or of self-image—our desire of eliciting positive evaluation from others or ourselves—is certainly or probably thwarted. According to Castelfranchi and Poggi [43], we are ashamed when we feel we have fallen short of a norm or value that we share with our group, or anyway one with respect to which we want to live up. So we can feel shame both before others and before ourselves. Suppose I pretend to be a good pianist, and while playing in front of my friends I make a mistake; I may be ashamed before them if I think they realized my fault, but I may also feel shame only before myself because,

even if they are not so skilled as to realize my subtle fault, I did; and I want to be perfect for and before myself. When a standard becomes part of our self-image, we are sincerely sorry any time we fall short of it; but if we share it with our group, our fault might lead the group to reject us and close social relations with us, so we feel shame also before others. In this case, feeling and showing our shame is a way to apologize, to tell others: "Yes, I transgressed this norm or value, but I did not do so on purpose, I still share this norm with you; so refrain from aggressing and rejecting me, accept me again in the group". Based on these assumptions, Castelfranchi and Poggi [43], different from Darwin [54], who viewed blushing as a mere side-issue of self-oriented attention, argue that the feeling of shame is a sort of internal self-punishment, while its external expression is a communicative signal, namely an apology, a request for forgiveness by one's group. In fact, the communicative display of shame includes three communicative signals:

1. a person S *blushes*, i.e., his/her face reddens
2. S *lowers his/her eyes*
3. S *lowers his/her head*

Signals 2. and 3. are actions, resulting in the typical posture of shame, but signal 1. is a morphological transitory feature. So while the latter two are to some extent under voluntary control, the former is an involuntary, even, counter-voluntary signal (as already pointed out by Darwin [54]): so much so that if you blush and you realize it, you would like not to blush (if only because blushing unmasks you have something to be ashamed of), and this makes you blush even more! But also the actions of avoiding gaze and lowering head function serve to acknowledge one's faults or shortcomings and to apologize for them, to block the group's aggression and prevent rejection.

According to Castelfranchi and Poggi [43], the three signals make up a multimodal discourse, where each conveys its specific meaning and all converge toward a global meaning. The *blush* (signal 1.), making face reddened as one of a baby, might be seen as communicating "I am like a baby", which carries the inference "I am inadequate", "I was/did the wrong way, like a baby", thus publicly acknowledging one's inadequacy and inferiority. Signal 2., *lowering eyes*, conveys "I give up looking at you", and since looking is a way to get power over things or people, it means "I give up my power of control over you", which again means "I am inferior". Signal 3., *lowering head*, shows one is smaller, again acknowledging one's inferiority, and giving up any defiant attitude. But acknowledging one's inferiority and giving up defiance, conveyed by lowering face and eyes, and acknowledging one's inadequacy, conveyed by the blush, communicate that one shares the missed value, that one cares about the other's judgment. All of this communicates "I am one of you", therefore "Do not attack me".

8.7.4.2 Pride

You feel pride when, due to an action (e.g. you run faster than others), a *property* (you are stubborn, you have long dark hair), or simply an *event* (your party has won

the elections), your goal of image and/or of self-image is fulfilled, that is, when you evaluate yourself, or believe to be evaluated by others, very positively with respect to some goals that make up part of your goal of image or self-image. The action, property or event must be due to yourself, or anyway be an important *part of your identity*. You can be proud of your son because you see what he is or does as something stemming from you, or be proud of the good climate of your country just because you feel it as *your* country.

Sometimes one feels *"proud of"* something not only before oneself but also because the positive event, property or action enhances one's own image before others. Before your colleagues in a foreign college, you can be proud of your country winning the football championship, since this gives you the image of one belonging to a champion country.

But if the goal of image is sometimes a condition to feel proud of something, is it always a necessary condition? In this, pride and shame are symmetrical. You are sincerely ashamed before others only if you also feel shame before yourself (Castelfranchi and Poggi [43]), that is, only if the value you are evaluated against makes part not only of your goal of image before others but also of the image you want to have of yourself. If you do not share some value (say, to be a very macho man) but this is not a relevant value for your own self-image even if others evaluate you against it, you do not feel shame if you do not look very macho to others. And if you happen to look so, you will not feel proud of it.

Beside being an emotion—a short transitory state—, pride can also be viewed as a more enduring state; a personality trait. A "proud" person is one who attributes a high value to his goal of self-image, mainly to his self-image as an autonomous person, one not dependent on anyone else. In fact, there are two sides of autonomy: self-sufficiency and self-regulation. You are self-sufficient when you possess all the (material and mental) resources you need to achieve your goals by yourself, that is, when you do not depend on others' help. And you are self-regulated when you can decide which goals to pursue, when and how, by yourself: in a word, when you are free. These two sides of autonomy are strictly connected: if you are self-sufficient (you have all the resources you need), you can afford self-regulation (you have the right to be free).

Three types of pride can be distinguished: superiority, arrogance, and dignity pride. In *superiority pride* the proud person pretends he is superior to the other, for instance because he has won over him. In *dignity pride*, he only claims to being at the same level as the other, not inferior to him: he wants to be acknowledged for his dignity as a human, and credited for his right to freedom, autonomy and self-regulation. In *arrogance pride*, finally, the *proud* person is, at the start, on the "down" side of the power comparison: he has less power than the other, but wants to challenge, to defy his power, and communicates he does have more power than the other.

The emotion of pride is expressed by a multimodal pattern of body signals: a small smile, expanded posture, head tilted backward, and arms extended out from the body, possibly with hands on hips (Tracy and Robins [166]). *Smile*, which is in general a signal of happiness, in this case conveys a positive feeling due to one's

goal of image or self-image being achieved; the *expanded posture*, enlarging the person's body, conveys dominance, superiority, but also makes one more visible; in fact, one who is proud of something may want to exhibit his merits. *Expanding chest* might be seen as making reference to oneself, to one's own *identity*. *Head tilted back* is a way to look taller, to symbolically communicate one is superior, but it also induces to *look down on the other*, thus, symmetrically, communicating the other's inferiority.

But if these are the expressive signals of pride in general, do different combinations of signals distinguish the three types of pride? Two studies have been carried on to test this hypothesis (Poggi and D'Errico [134]).

In an observational study a qualitative analysis was conducted on pride expressions in political debates. Results indicate that dignity pride is characterized by *head tilted upward*, but also by signals of worry and anger like a *frown* or vertical wrinkles on the forehead (AU4), *rapid and nervous gestures*, *high intensity of voice*, eyes fixed to interlocutor and *no smile*; all signaling seriousness of the proud person's request to have one's dignity acknowledged.

Superiority pride is characterized by *low rhythm and intensity of voice* that signal the absence of worry (if you are superior you have nothing to fear or worry about from the other), and sometimes by *gazing away from the Interlocutor* (he is so inferior that he does not even deserve your gaze or your attention). Arrogance pride is characterized by a *large smile*, quite close to a scornful laughter; *expanded chest*, *head tilted back*, and *gaze fixed to the Interlocutor*, which convey challenge and defiance, and *provocative*, possibly *insulting words*. The whole pattern conveys that the proud person does not fear the Interlocutor, even if he is presently superior to him.

Based on this and previous studies an experimental study on the expression of the three types of pride tested the following hypotheses (D'Errico and Poggi [60]; Poggi and D'Errico [135]): it was expected that a frown and absence of smile characterize *Dignity pride*, asymmetrical eyebrows and no smile, *Superiority pride*, and absence of frown and presence of smile, *Arrogance pride*. A bifactorial 3 × 2 *between subjects* study was designed with two independent variables (eyebrow position—frown, no frown, asymmetrical eyebrows, and smile—present or absent), and three dependent variables (detection of dignity, superiority or arrogance pride). A multiple choice questionnaire was submitted to 58 subjects (females, range 18–32 years old, mean age 22) where each of 6 facial expressions, constructed by crossing the eyebrow and smile variables, were attributed meanings pointing to dignity, superiority, or arrogance pride. Results show that asymmetrical eyebrows without smile were interpreted as either superiority pride or dignity pride, while frown with smile was mainly attributed a meaning of dignity beside other positive meanings as *I am resolute*, *I want to humiliate you* and *I won*.

In general the frown is primarily interpreted as dignity pride, while the asymmetrical eyebrows orient subjects to an interpretation of superiority, and no frown to arrogance. The smile, probably interpreted as ironic, mainly points to the choice *I will win over you*, confirming its characterizing arrogance. The absence of smile instead is associated to dignity (*I don't submit to you*) and to superiority pride.

8.7.4.3 Enthusiasm

Enthusiasm is a "social emotion", not in that it is "felt toward" someone else, but in that it typically tends to be "socialized", that is, transmitted to others through contagion. It belongs to the family of happiness, being an intensely positive emotion, felt for the achievement of a very important goal, but it differs from happiness, exultance or elation, both for the goal at stake and for the time it is felt (Poggi [128, 129]). On the one hand, enthusiasm is only felt about goals that are in some way great, important, worth to pursue: for example, in activities that entail novelty and creativity (like creating a new musical group, or founding a newspaper), or for goals of equity and altruism (like fighting for your ideas or defending noble causes). On the other hand, enthusiasm tends to be felt not so much after the achievement, but during the very pursuit of a goal. The football players feel exultance when the game is over and they have won, but enthusiasm at the moment of a goal. This first achievement triggers a set of proprioceptive sensations typical of high activation: heart beat acceleration, a sense of energy, well-being, good mood, heat, excitation; as you feel enthusiasm you cannot stand still, you want to talk, to hop up and down, to make uncontrolled movements, speak loud, sometimes shout (Poggi [128, 129]).

That such internal energy is felt not when the final objective, but an intermediate goal of the plan is achieved, is functional to sustain the goal pursuit: this tough partial success makes you believe that "you can", you have the internal capacities to achieve your goal; achieving the intermediate step makes you feel confident that you will attain the final objective. This enhances your sense of self-efficacy; whereas in trust and hope you rely, respectively, upon other people or world conditions, with enthusiasm you have a feeling of omnipotence: coherent with its etymology "en theòn", which means: "(to have) a God inside"!

This self-attribution of power has two effects. First, you believe that achieving the final goal does not depend on world conditions but on your own action: you feel more responsible. Second, since for decision making rules choosing which goals to pursue in part depends on how likely it is that you can achieve them, if you believe you can, you will strive with particular persistency. Thus the high arousal of enthusiasm can trigger the physiological resources necessary for goal pursuit.

The function of this emotion is then to work as the "gasoline of motivation". The great physiological activation sustained by it fosters physical and mental resources and gives higher persistency, induces self-confidence and renews motivation, providing new energy for action.

8.8 Summary

Social signals are physical stimuli, either produced by a Sender or simply perceived by a Receiver, which provide information about social interactions and social relationships, social attitude and social emotions. Studies in cognitive and social psychology and neuroscience, along with research in computer science, signal processing and computer graphics may contribute to the sensing and interpretation of social

signals and to their simulation in Embodied Agents, in view of building systems for Human–Computer Interaction, Ambient Intelligence, Persuasive Technology and Affective Computing.

8.9 Questions

1. Provide examples of the following signals: one informative social signal, one informative non-social signal, one communicative social signal and one communicative non-social signal.
2. Provide two examples of indirect social signals.
3. Find out two real examples social signals (either direct or indirect, either communicative or informative) for each of the following modalities: head movements, gaze, facial expression, gestures.
4. Choose a fragment of 2–3 minutes from a political debate in which one or more participants express dominance strategies, and describe all the verbal and vocal features, gaze items, head movements, facial expression, gestures, posture, use of space and possible physical contact, through which these dominance strategies are expressed (see Sect. 8.7.3).

8.10 Glossary

- *Agreement*: Relation of identity, similarity or congruence between the mental states of two or more persons. It concerns informative speech acts (proposal, an assessment), not "factual" beliefs. Agreement may be communicated through verbal discourse and by verbal and nonverbal markers.
- *Attitude*: Tendency of a person to behave in a certain way toward another person or a group. Social attitudes include cognitive elements like beliefs, evaluations, opinions, but also emotions that all determine and are determined by preferences and intentions.
- *Emotion*: Adaptive device that monitors the state of our most important goals: they are multifaceted subjective states, encompassing internal feelings and cognitive, physiological, expressive, motivational states, which are triggered any time an important adaptive goal is, or is likely to be, achieved or thwarted.
- *Interaction*: Event in which two or more biological or artificial agents perform reciprocal social actions, that is, actions in which the goal of one participant is directed to the other by considering him as an autonomous agent, one regulated by one's own goals.
- *Opinion*: A subjective belief, that is, one we know is not necessarily shared, and have no empirical evidence for, which stems from considering some entity or event from a particular point of view, somehow determined by our beliefs and goals.

- *Persuasion*: A communicative action aimed at changing people's attitudes, that is, at influencing their tendency to action by changing their opinions: we persuade as we influence another to do or not to do something by inducing him to conclude that what we propose is good for him.
- *Social relationship*: Interdependency of goals between two or more persons: the pursuit, achievement or thwarting of a goal of one determines or is determined by the pursuit, achievement or thwarting of a goal of the other.
- *Social signal*: A communicative or informative signal which, either directly or indirectly, provides information about "social facts", that is, about social interactions, social attitudes, social relations and social emotions.
- *Signal*: A perceivable stimulus from which a privileged inference can be drawn, which thus becomes its meaning: a new belief different from the perceivable stimulus but linked to it in the mind of one or more agents.
- *Informative signal*: Signal that a receiver can interpret, that is, from which he can draw some meaning, even if no sender had the goal to have someone believe this meaning.
- *Communicative signal*: Signal emitted by a sender with the goal (conscious, unconscious, tacit, social or biological) of having a receiver come to believe some meaning.
- *Direct signal*: Signal that holds a systematic relation with a meaning. The relation may be creative (invented on the spot) or codified (stably represented in memory, like in a lexicon). The meaning linked to the signal in a shared way is its literal, or direct, meaning.
- *Indirect signal*: Signal for which a further meaning can be drawn from the "literal meaning", through inferences based on contextual or previously shared information, which thus may differ across contexts.

Acknowledgements This research is supported by the Seventh Framework Program, European Network of Excellence SSPNet (Social Signal Processing Network), Grant Agreement No. 231287.

References

1. Adolphs, R.: Social cognition and the human brain. Trends Cogn. Sci. **3**(12), 469–479 (1999)
2. Albrecht, K.: Social Intelligence: The New Science of Success. Wiley, Berlin (2005)
3. Allwood, J., Nivre, J., Ahlsen, E.: On the semantics and pragmatics of linguistic feedback. J. Semant. **9**(1), 1–26 (1992)
4. Allwood, J., Ahlsén, E., Lund, J., Sundqvist, J.: Multimodality in own communication management. In: Allwood, J., Dorriots, B., Nicholson, S. (eds.) Multimodal Communication. Proceedings from the Second Nordic Conference on Multimodal Communication. Gothenburg Papers in Theoretical Linguistics. Department of Linguistics, Gothenburg University, Gothenburg (2005)
5. Ambady, N., Rosenthal, R.: Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. J. Pers. Soc. Psychol. **64**, 431–441 (1993)
6. Argyle, M.: Bodily Communication, 2nd edn. Methuen, New York (1988)
7. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press, Cambridge (1976)

8. Aries, E.J., Gold, C., Weigel, R.H.: Dispositional and situational influences on dominance behavior in small groups. J. Pers. Soc. Psychol. **44**, 779–786 (1983)
9. Aristotle: Retorica. Laterza, Bari (1973)
10. Asch, S.E.: Forming impressions of personality. J. Abnorm. Soc. Psychol. **41**, 258–290 (1946)
11. Badler, N., Chi, D., Chopra Kullar, S.: Virtual human animation based on movement observation and cognitive behavior models. In: Computer Animation Conference, pp. 128–137 (1999)
12. Baker-Schenk, C.: Nonmanual behaviors in sign languages: methodological concerns and recent findings. In: Stokoe, W., Volterra, V. (eds.) SRL 1983, Sign Language Research. Linstock Press, Silver Spring (1985)
13. Bargh, J.A.: Auto-motives: Preconscious determinants of thought and behavior. In: Higgins, E.T., Sorrentino, R.M. (eds.) Handbook of Motivation and Cognition. Guilford, New York (1990)
14. Bargh, J.A.: The four horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In: Wyer, R.S. Jr., Srull, T.K. (eds.) Handbook of Social Cognition. Erlbaum, Hillsdale (1994)
15. Bargh, J.A.: Social Psychology and the Unconscious: The Automaticity of the Higher Mental Processes. Psychology Press, Philadelphia (2006)
16. Baron-Cohen, S.: Precursors to a theory of mind: Understanding attention in others. In: Whiten, A. (ed.) Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading. Basil Blackwell, Oxford (1991)
17. Baron-Cohen, S.: Mindblindness: an Essay on Autism and Theory of Mind. MIT Press, Cambridge (1995)
18. Bartlett, F.C.: Remembering: A Study in Experimental and Social Psychology. University Press, Cambridge (1932)
19. Bavelas, J., Gerwing, J.: Conversational hand gestures and facial displays in face-to-face dialogue. In: Fiedler, K. (ed.) Social Communication. Psychology Press, New York (2007)
20. Bavelas, J., Chovil, N., Lawrie, D., Wade, A.: Interactive gestures. Discourse Process. **15**(4), 469–489 (1992)
21. Bazzanella, C.: Le Facce del Parlare. La Nuova Italia, Firenze (1994)
22. Benne, K.D., Sheats, P.: Functional roles of group members. J. Soc. Issues **4**, 41–49 (1948)
23. Berscheid, E., Reiss, H.: Attraction and close relationships. In: Gilbert, D., Fiske, S., Lindzey, G. (eds.) Handbook of Social Psychology. McGraw-Hill, New York (1997)
24. Bevacqua, E.: Computational model of listener behavior for embodied conversational agents. Ph.D. dissertation, University of Paris 8 (2009)
25. Bevacqua, E., Mancini, M., Niewiadomski, R., Pelachaud, C.: An expressive ECA showing complex emotions. In: Language, Speech and Gesture for Expressive Characters, AISB 2007, Newcastle, UK (2007)
26. Blom, J.-P., Gumperz, J.: Social meaning in linguistic structures: code switching in northern Norway. In: Gumperz, J., Hymes, D. (eds.) Directions in Sociolinguistics: The Ethnography of Communication, pp. 407–434. Holt, Rinehart & Winston, New York (1972)
27. Boholm, M., Allwood, J.: Repeated head movements, their function and relation to speech. In: Calzolari, N., et al. (ed.) Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10), Valetta, Malta, May 19–21 (2010). European Language Resources Association (ELRA) (Workshop on Multimodal Corpora)
28. Boraston, Z., Blakemore, S.J.: The application of eye-tracking technology in the study of autism. J. Physiol. **581**(3), 893–898 (2007)
29. Bousmalis, K., Mehu, M., Pantic, M.: Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, vol. II, pp. 121–129 (2009)
30. Bowlby, J.: Attachment and Loss. Hogarth Press, London (1969)
31. Briñol, P., Petty, R.E.: Overt head movement and persuasion: a self-validation analysis. J. Pers. Soc. Psychol. **84**(6), 1123–1139 (2003)

32. Brooke, M.E., Ng, S.H.: Language and social influence in small conversation groups. J. Lang. Soc. Psychol. **5**, 201–210 (1986)
33. Brunet, P.M., Donnan, H., McKeown, G., Douglas-Cowie, E., Cowie, R.: Social signal processing: What are the relevant variables? And in what ways do they relate? In: Proceedings of the IEEE International Workshop on Social Signal Processing, pp. 1–6 (2009)
34. Bull, E.P.: Posture and Gesture. Pergamon, Oxford (1987)
35. Burgoon, J.K., Dunbar, N.E.: An interactionist perspective on dominance submission: Interpersonal dominance as a dynamic, situationally contingent social skill. Commun. Monogr. **67**, 96–121 (2000)
36. Byrne, D.: The Attraction Paradigm. Academic Press, New York (1971)
37. Cashdan, E.: Smiles, speech, and body posture: How women and men display sociometric status and power. J. Nonverbal Behav. **22**(4), 209–228 (1998)
38. Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., Rich, C.: Non-verbal cues for discourse structure. In: Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics, Toulouse, France, pp. 106–115 (2001)
39. Cassell, J.: Nudge nudge wink wink: elements of face-to-face-conversation for embodied conversational agents. In: Cassell, J. et al. (eds.) Embodied Conversational Agents. MIT Press, Cambridge (2000)
40. Castelfranchi, C.: Social power: a missed point. In: Demazeau, Y., Mueller, J. (eds.) Decentralized AI. Elsevier, North-Holland (1990)
41. Castelfranchi, C., Guerini, M.: Is it a promise or a threat? Pragmat. Cogn. **15**(2), 277–311 (2007)
42. Castelfranchi, C., Parisi, D.: Linguaggio, Conoscenze e Scopi. Il Mulino, Bologna (1980)
43. Castelfranchi, C., Poggi, I.: Blushing as a discourse: Was Darwin wrong. In: Crozier, R. (ed.) Shyness and Embarrassment. Perspectives from Social Psychology. Cambridge University Press, New York (1990)
44. Cerrato, L.: Linguistic functions of head nods. In: Allwood, J., Dorriots, B. (eds.) Proc. from The Second Nordic Conference on Multi-modal Communication. Gothenburg Papers in Theoretical Linguistics, vol. 92. Gothenburg University, Gothenburg (2005)
45. Cerrato, L.: Investigating communicative feedback phenomena across languages and modalities. Dissertation, KTH, Stockholm (2007)
46. Chovil, N.: Discourse-oriented facial displays in conversation. Res. Lang. Soc. Interact. **25**, 163–194 (1992)
47. Cicero, M.T.: De Oratore (55 B.C.)
48. Cohn, J.F.: Social signal processing in depression. In: Proceedings of SSPW 2010—Social signal Processing Workshop. ACM/Sheridan Press, New York (2010). doi:978-1-4503-0174-9/10/10
49. Condon, W.S., Ogston, W.D.: Speech and body motion synchrony of the speaker-hearer, the perception of language. In: Horton, D.H., Jenkins, J.J. (eds.) The Perception of Language. Academic Press, New York (1971)
50. Conte, R., Castelfranchi, C.: Cognitive and Social Action. University College, London (1995)
51. Conte, R., Paolucci, M.: Reputation in Artificial Societies. Social Beliefs for Social Order. Kluwer, Boston (2002)
52. Curhan, J., Pentland, A.: Thin slices of negotiation: predicting outcomes from conversational dynamics within the first five minutes. J. Appl. Psychol. **92**, 802–811 (2007)
53. Damasio, AR: Descartes' Error: Emotion, rationality and the human brain. Putnam, New York (1994)
54. Darwin, C.: The Expression of the Emotions in Man and Animals. Appleton, New York (1872)
55. de Rosis, F., Novielli, N.: From language to thought: inferring opinions and beliefs from verbal behavior. In: Proceeding of Mindful Environments Workshop in the Scope of AISB (2007)

56. de Rosis, F., Pelachaud, C., Poggi, I.: Transcultural believability in embodied agents: a matter of consistent adaptation. In: Trappl, R., Payr, S. (eds.) Agent Culture: Designing virtual characters for a multi-cultural world. Kluwer Academic, Dordrecht (2004)
57. DePaulo, B.M.: Non Verbal behavior and self-presentation. Psychol. Bull. **111**, 203–243 (1992)
58. DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., Cooper, H.: Cues to deception. Psychol. Bull. **129**, 74–118 (2003)
59. D'Errico, F.: Power relations and gender differences in helping behaviour and representation. PhD dissertation, University of Bari (2006)
60. D'Errico, F., Poggi, I.: Gaze, mouth and body of pride. The multimodal expression of a social emotion. In: Allwood, J. (ed.) Proceedings of Nordic Multimodal Communication (2011, forthcoming)
61. D'Errico, F., Leone, G., Poggi, I.: Types of help in the teacher's multimodal behavior. In: Salah, A.A., et al. (ed.) HBU 2010. LNCS, vol. 6219. Springer, Heidelberg (2010)
62. Dittmann, A.T., Llewellyn, L.G.: Relationship between vocalizations and head nods as listener responses. J. Pers. Soc. Psychol. **9**(1), 79–84 (1968)
63. Dunbar, N.E., Burgoon, J.K.: Perceptions of power and interactional dominance in interpersonal relationships. J. Soc. Pers. Relatsh. **22**, 231–257 (2005)
64. Dunbar, N.E., Bippus, AM, Young, S.L.: Interpersonal dominance in relational conflict: a view from dyadic power theory. Interpersona **2**(1), 1–33 (2008)
65. Duncan, S., Fiske, D.: Face-to-Face Interaction: Research, Methods, and Theory. Erlbaum, Hillsdale (1977)
66. Eagly, A.H., Ashmore, R.D., Makhaijani, M.D., Longo, L.C.: What is beautiful is good, but... A meta-analytic review of research on physical attractiveness stereotype. Psychol. Bull. **110**, 109–128 (1991)
67. Eibl-Eibesfeldt, I.: Similarities and differences between cultures in expressive movements. In: Hinde, R. (ed.) Non-verbal Communication. Cambridge University Press, London (1972)
68. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. Psychiatry **32**, 88–106 (1969)
69. Ekman, P., Friesen, W.V.: Unmasking the Face. A Guide to Recognizing Emotions from Facial Clues. Prentice-Hall, Englewood Cliffs (1975)
70. Ekman, P., Friesen, W., Hager, J.: Facial Action Coding System (FACS). In: A Human Face, Salt Lake City, USA (2002)
71. Ekman, P.: About brows: Emotional and conversational signals. In: von Cranach, M. et al. (eds.) Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium. Cambridge University Press, Cambridge (1979)
72. Ellyson, S.L., Dovidio, J.F.: Power, dominance, and nonverbal behavior: Basic concepts and nonverbal behavior. In: Ellyson, S.L., Dovidio, J.F. (eds.) Power, Dominance, and Nonverbal Behavior. Springer, New York (1985)
73. Festinger, L., Schachter, S., Back, K.: Social Pressures in Informal Groups: A Study of Human Factors in Housing. Stanford University Press, Palo Alto (1950)
74. Fishbein, M., Ajzen, I.: Attitude, Intention and Behavior. An Introduction to Theory and Research. Addison-Wesley, Reading (1975)
75. Fiske, S.T., Taylor, S.E.: Social Cognition, 2nd edn. McGraw-Hill, New York (1991)
76. Fiske, S.T., Taylor, S.E.: Social Cognition: From Brains to Culture. McGraw-Hill, New York (2008)
77. Fogg, B.J.: Persuasive Technology: Using Computers to Change What We Think and Do. Kaufmann, Los Altos (2002)
78. Gallese, V.: Intentional attunement: a neurophysiological perspective on social cognition and its disruption in autism. Brain Res. **1079**, 15–24 (2006)
79. Gardner, H.: Frames of Mind. The Theory of Multiple Intelligences. Basic Books, New York (1983/2003)

80. Goffman, E.: Relations in Public: Microstudies of the Public Order. Harper & Row, New York (1971)
81. Goleman, D.P.: Emotional Intelligence: Why It Can Matter More Than IQ for Character, Health and Lifelong Achievement. Bantam Books, New York (1995)
82. Goleman, D.P.: Social Intelligence. Hutchinson, London (2006)
83. Goodwin, C.: Conversational Organization. Interaction Between Speakers and Hearers. Academic Press, New York (1981)
84. Gregory, S.W., Webster, S.: A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. J. Pers. Soc. Psychol. **70**, 1231–1240 (1996)
85. Guerini, M., Strapparava, C., Stock, O.: CORPS: a corpus of tagged political speeches for persuasive communication processing. J. Inf. Technol. Polit. **5**(1), 19–32 (2008)
86. Hadar, U., Steiner, T., Rose, C.F.: Head movement during listening turns in conversation. J. Nonverbal Behav. **9**, 214–228 (1995)
87. Hall, R.: The Hidden Dimension. Doubleday, New York (1966)
88. Halliday, M.: Il Linguaggio Come Semiotica Sociale. Zanichelli, Bologna (1983)
89. Hartmann, B., Mancini, M., Pelachaud, C.: Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. In: Computer Animation, pp. 111–119 (2002)
90. Hegen-Larsen, M., Cunningham, S.J., Carrico, A., Pergram, A.M.: To nod or not to nod: an observational study of nonverbal communication and status in female and male college students. In: Psychology of Women Quarterly, vol. 28, pp. 358–361 (2004)
91. Heylen, D.: Challenges Ahead. Head movements and other social acts in conversations. In: Proceedings of the 5th International Conference on Interactive Virtual Agents, Kos, Greece (2005)
92. Higgins, E.T., Rholes, W.S.: Impression formation and role fulfillment: a "holistic reference" approach. J. Exp. Soc. Psychol. **12**, 422–435 (1976)
93. Hinde, R.: La Comunicazione Non-verbale nell'Uomo. Laterza, Bari (1977)
94. Iacoboni, M.: Neural mechanisms of imitation. Curr. Opin. Neurobiol. **15**, 632–637 (2005)
95. James, D.: The syntax and semantics of some English interjections. Ph.D. dissertation, University of Michigan (1974)
96. Jayagopi, D.B., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance ingroup conversations using nonverbal activity cues. IEEE Trans. Audio Speech Lang. Process. **17**, 3 (2009)
97. Jucker, A.: The discourse marker well: a relevance theoretical account. J. Pragmat. **19**(5), 435–452 (1993)
98. Keating, C.F., Bai, D.L.: Children's attributes of social dominance from facial cues. Child Dev. **57**, 1269–1276 (1986)
99. Kendon, A.: Some functions of gaze direction in social interaction. Acta Psychol. **26**, 1–47 (1967)
100. Kendon, A.: Movement coordination in social interaction: some examples described. Acta Psychol. **332**, 1–25 (1970)
101. Kendon, A.: Some uses of the head shake. Gesture **2**(2), 147–182 (2002)
102. Kendon, A.: Gesture. Visible Action as Utterance. Cambridge University Press, Cambridge (2004)
103. Kenny, D.A., Horner, C., Kashy, D.A., Chu, L.: Consensus at zero acquaintance: replication, behavioral cues, and stability. J. Pers. Soc. Psychol. **62**, 88–97 (1992)
104. Kleinsmith, A., Bianchi-Berthouze, N.: Recognizing affective dimensions from body posture. In: Proceedings of the International Conference of Affective Computing and Intelligent Interaction. LNCS, vol. 4738, pp. 48–58 (2007)
105. Kreidlin, G.: The comparative semantics and pragmatics of Russian nonverbal acts of touching and Russian verbs of touching. In: Rector, M., Poggi, I., Trigo, N. (eds.) Gestures. Meaning and Use. Edicoes Universidade Fernando Pessoa, Oporto (2003)
106. Leal, S., Vrij, A.: Blinking during and after lying. J. Nonverbal Behav. **32**, 187–194 (2008)
107. Lewin, K.: Group decision and social change. In: Newcomb, T., Hartley, E. (eds.) Readings in Social Psychology, New York (1947)

108. Lewis, M.: Self-conscious emotions: Embarrassment, pride, shame, and guilt. In: Lewis, M., Haviland-Jones, J. (eds.) Handbook of Emotions. Guilford, New York (2000)
109. McClave, E.Z.: Linguistic functions of head movements in the context of speech. J. Pragmat. **32**, 855–878 (2000)
110. McNeill, D.: Hand and Mind. University of Chicago Press, Chicago (1992)
111. Mehrabian, A.: Nonverbal Communication. Aldine-Atherton, Chicago (1972)
112. Meltzoff, A.N., Decety, J.: What imitation tells us about social cognition. A rapprochement between developmental psychology and cognitive neuroscience. Philos. Trans. R. Soc. Lond. **358**, 491–500 (2003)
113. Miceli, M., Poggi, I., de Rosis, F.: Emotional and nonemotional persuasion. Appl. Artif. Intell. **20**(10), 849–879 (2006). Special Issue on Natural Argumentation
114. Mohammadi, G., Vinciarelli, A., Mortillaro, M.: The voice of personality: mapping nonverbal vocal behavior into trait attributions. In: Proceedings of ACM Multimedia Workshop on Social Signal Processing, pp. 17–20 (2010)
115. Montagu, A.: Touching: the Human Significance of the Skin. Columbia University Press, New York (1971)
116. Nass, C., Steuer, J.: Voices, boxes and sources of messages: computers and social actors. Hum. Commun. Res. **19**(4), 504–527 (1993)
117. Ng, S.H., Bradac, J.J.X.: Power in Language. Sage, Thousand Oaks (1986)
118. Oberman, L.M., Hubbard, E.M., McCleery, J.P., Altschuler, E.L., Ramachandran, V.S., Pineda, J.A.: EEG evidence for mirror neuron dysfunction in autism spectrum disorders. Cogn. Brain Res. **24**, 190–198 (2005)
119. Ogden, R.: Phonetics and social action in agreements and disagreements. J. Pragmat. **38**, 1752–1775 (2006)
120. Pentland, A.: Social signal processing. IEEE Signal Process. Mag. **24**(4), 108–111 (2007)
121. Pentland, A.: Honest Signals. How Do They Shape Our World. MIT Press, Cambridge (2008)
122. Perelman, C., Olbrechts-Tyteca, L.: The New Rhetoric: A Treatise on Argumentation. University of Notre Dame Press, Notre Dame (1969)
123. Pesarin, A., Cristani, M., Murino, V., Vinciarelli, A.: Conversation analysis at work: detection of conflict in competitive discussions through automatic turn-organization analysis. In: Social Signals. Theory and Application. (2011, forthcoming). Special Issue of "Cognitive Processing"
124. Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., Poggi, I.: A model of attention and interest using gaze behavior. In: Panayiotopoulos, T., et al. (eds.) IVA 2005. LNCS, vol. 3661, pp. 229–240. Springer, Heidelberg (2005)
125. Petty, R., Cacioppo, J.T.: The elaboration likelihood model of persuasion. In: Berkowitz, L. (ed.) Advances in Experimental Social Psychology, vol. 19, pp. 123–205. Academic Press, New York (1986)
126. Poggi, I.: From a typology of gestures to a procedure for gesture production. In: Wachsmuth, I., Sowa, T. (eds.) Gesture and Sign Language in Human–Computer Interaction, pp. 158–168. Springer, Berlin (2002)
127. Poggi, I.: The goals of persuasion. Pragmat. Cogn. **13**, 298–335 (2005)
128. Poggi, I.: Mind, Hands, Face and Body. A Goal and Belief View of Multimodal Communication. Weidler, Berlin (2007)
129. Poggi, I.: Enthusiasm and its contagion: nature and function. In: Proceedings of ACII 2007, pp. 410–421 (2007)
130. Poggi, I.: The language of interjections. In: COST 2102 School, pp. 170–186 (2008)
131. Poggi, I., D'Errico, F.: The mental ingredients of Bitterness. J. Multimodal User Interfaces **3**, 79–86 (2009). doi:10.1007/s12193-009-0021-9
132. Poggi, I., D'Errico, F.: Social signals and the action—cognition loop. The case of overhelp and evaluation. In: Proceeding of IEEE International Conference on Affective Computing and Intelligent Interaction, New York, pp. 106–113 (2009). doi:10.1109/ACII.2009.5349468

133. Poggi, I., D'Errico, F.: Cognitive modelling of human social signals. In: Proceedings of SSPW 2010—Social Signal Processing Workshop. ACM/Sheridan Press, New York (2010). doi:978-1-4503-0174-9/10/10
134. Poggi, I., D'Errico, F.: Pride and its expression in political debates. In: Paglieri, F., Tummolini, L., Falcone, R., Miceli, M. (eds.) The Goals of Cognition. Festschfit for Cristiano Castelfranchi. London College Publications, London (2011, forthcoming)
135. Poggi, I., D'Errico, F.: Types of pride and their expression. In: Esposito, A. (ed.) Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue. LNCS, pp. 445–459. Springer, Heidelberg (2011)
136. Poggi, I., D'Errico, F.: Defining social signals. In: Poggi, I., D'Errico, F., Vinciarelli, A. (eds.) Special Issue of Cognitive Processing on Social Signals. From Theory to Applications (2011, forthcoming)
137. Poggi, I., D'Errico, F.: Dominance in political debates. In: Salah, A.A., et al. (eds.) HBU. LNCS, vol. 6219. Springer, Heidelberg (2010)
138. Poggi, I., Pelachaud, C.: Persuasive gestures and the expressivity of ECAs. In: Wachsmuth, I., Lenzen, M., Knoblich, G. (eds.) Embodied Communication in Humans and Machines. Oxford University Press, Oxford (2008)
139. Poggi, I., Vincze, L.: Persuasive gaze in political discourse. In: Proceedings of the Symposium on Persuasive Agents. AISB, Aberdeen (2008)
140. Poggi, I., Vincze, L.: Gesture, gaze and persuasive strategies in political discourse. In: Kipp, M. et al. (eds.) Multimodal Corpora. From Models of Natural Interaction to Systems and Applications. Lecture Notes in Computer Science, vol. 5509, pp. 73–92. Springer, Berlin (2009)
141. Poggi, I., Pelachaud, C., de Rosis, F.: Eye communication in a conversational 3D synthetic agent. AI Commun. **13**, 169–181 (2000)
142. Poggi, I., D'Errico, F., Spagnolo, A.: The embodied morphemes of gaze. In: Kopp, S., Wachsmuth, I. (eds.) Gesture in Embodied Communication and Human–Computer Interaction. LNAI, vol. 5934. Springer, Berlin (2010). doi:10.1007/978-3-642-12553-9_4
143. Poggi, I., D'Errico, F., Vincze, L.: Agreement and its multimodal communication in debates. A qualitative analysis. Cogn. Comput. (2010). doi:10.1007/s12559-010-9068-x
144. Poggi, I., D'Errico, F., Vincze, L.: Types of nods. The polysemy of a social signal. In: Calzolari, N., et al. (eds.) Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Malta, 19–21 May 2010. European Language Resources Association (ELRA)
145. Pratto, E., Sidanius, J., Stallworth, L.M., Malle, B.F.: Social dominance orientation: A personality variable predicting social and political attitudes. J. Pers. Soc. Psychol. **67**, 741–763 (1994)
146. Quintilianus, M.F.: Institutio Oratoria. Le Monnier, Firenze (95)
147. Ridgeway, C.L.: Nonverbal behavior, dominance, and the basis of status in task groups. Am. Sociol. Rev. **52**, 683–694 (1987)
148. Ridgeway, C.L.: Gender status and leadership. J. Soc. Issues **57**, 637–655 (2001)
149. Rizzolatti, G., Arbib, M.A.: Language within our grasp. Trends Neurosci. **21**, 188–194 (1998)
150. Rosenthal, R., Archer, D., Koivumaki, J.H., Di Matteo, M.R., Rogers, P.: The language without words. Psychol. Today 44–50 (1974)
151. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn taking for conversation. Language **50**, 696–735 (1974)
152. Salovey, P., Mayer, J.D.: Emotional intelligence. Imagin. Cogn. Pers. **9**, 185–211 (1990)
153. Scheflen, A.E.: The significance of posture in communication systems. Psychiatry **26**, 316–331 (1964)
154. Scherer, K.R.: Personality markers in speech. In: Scherer, K.R., Giles, H. (eds.) Social Markers in Speech. Cambridge University Press, Cambridge (1979)
155. Scherer, K.R., Scherer, U.: Speech behavior and personality. In: Darby, J. (ed.) Speech Evaluation in Psychiatry, pp. 171–187. Grune & Stratton, New York (1981)

156. Schröder, M.: Experimental study of affect bursts. Speech Commun. **40**(1–2), 99–116 (2003). Special Issue Speech and Emotion
157. Schröder, M., Heylen, D., Poggi, I.: Perception of non-verbal emotional listener feedback. In: Proceedings of Speech Prosody, Dresden, Germany (2006)
158. Serenari, M.: Examples from the Berlin dictionary of everyday gestures. In: Rector, M., Poggi, I., Trigo, N. (eds.) Gestures. Meaning and Use. Edicoes Universidade Fernando Pessoa, Porto (2003)
159. Stein, R.T., Heller, T.: An empirical analysis of the correlations between leadership status and participation rate reported in literature. J. Pers. Soc. Psychol. **37**, 1993–2003 (1979)
160. Tajfel, H.: Interindividual behaviour and intergroup behaviour. In: Tajfel, H. (ed.) Differentation Between Social Groups: Studies in the Social Psychology of Intergroup Relations. Academic Press, London (1978)
161. Tajfel, H., Turner, J.C.: The social identity theory of intergroup behaviour. In: Worchel, S., Austin, W.G. (eds.) Psychology of Intergroup Behavior. Erlbaum, Hillsdale (1986)
162. Thòrisson, K.: Natural turn-taking needs no manual. In: Granstrom, I.K.B., House, D. (eds.) Multimodality in Language and Speech Systems. Kluwer Academic, Dordrecht (2002)
163. Tomasello, M., Hare, B., Lehmann, H., Call, J.: Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. J. Hum. Evol. **52**(3), 314–320 (2007)
164. Tomkins, S.: Affect, Imagery, Consciousness. Springer, New York (1962)
165. Toulmin, S.: The Uses of Argument. Cambridge University Press, Cambridge/New York (1958)
166. Tracy, J.L., Robins, R.W.: Show your pride: Evidence for a discrete emotion expression. Psychol. Sci. **15**, 194–197 (2004)
167. Tracy, J.L., Robins, R.W.: The prototypical pride expression: Development of a nonverbal behavioral coding system. Emotion **7**, 789–801 (2007)
168. Trevarthen, C.: Communication and cooperation in early infancy: a description of primary intersubjectivity. In: Bullowa, M. (ed.) Before Speech. Cambridge University Press, Cambridge (1979)
169. Uddin, L.Q., Davies, M.S., Scott, A.A., Zaidel, E., Bookheimer, SY, et al.: Neural basis of self and other representation in autism: an fMRI study of self-face recognition. PLoS ONE **3**(10), e3526 (2008). doi:10.1371/journal.pone.0003526
170. van Eemeren, F.H., Grootendorst, R.: Argumentation, Communication and Fallacies—a Pragma-Dialectical Perspective. Erlbaum, Hillsdale (1992)
171. Vesper, C., Butterfill, S., Knöblich, G., Sebanz, N.: A minimal architecture for joint action. Neural Netw. **23**, 998–1003 (2010)
172. Vinciarelli, A.: Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. IEEE Trans. Multimed. **9**(9), 1215–1226 (2007)
173. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009)
174. Vincze, L., Poggi, I.: Close your eyes and communicate. In: Proceedings of Teorie e Trascrizione—Trascrizione e Teoria, Bolzano (2011, forthcoming)
175. Wells, G.L., Petty, R.E.: The effects of overt head movements on persuasion: compatibility and incompatibility of responses. Basic Appl. Soc. Psychol. **1**, 219–230 (1980)
176. Wiebe, J.: Tracking point of view in narrative. Comput. Linguist. **20**(2), 233–287 (1994)
177. Wiebe, J., Wilson, T., Cardie, C. : Annotating expressions of opinions and emotions in language. Lang. Resour. Eval. **39**(2/3), 164–210 (2005)
178. Williams, J.H.G., Whiten, A., Suddendorf, T., Perrett, D.I.: Imitation mirror neurons, and autism. Neurosci. Biobehav. Rev. **25**, 287–295 (2001)

179. Wilson, T., Hofer, G.: Using linguistic and vocal expressiveness in social role recognition. In: Proceedings of the International Conference on Intelligent User Interfaces, IUI (2011)
180. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT-EMNLP, Vancouver (2005)
181. Yarbus, A.L.: Eye Movements and Vision. Plenum, New York (1967)
182. Yngve, V.: On getting a word in edgewise. In: Papers from the Sixth Regional Meeting, pp. 567–578. Chicago Linguistic Society, Chicago (1970)

Chapter 9
Voice and Speech Analysis in Search of States and Traits

Björn Schuller

9.1 Vocal Behaviour Analysis—an Introduction

It is the aim of this chapter to introduce the analysis of vocal behaviour and more general paralinguistics in speech and language. By 'voice' we refer to the acoustic properties of a speakers' voice—this will be dealt with in Sect. 9.2. By 'speech' we refer more generally to spoken language in the sense of added linguistics—dealt with in Sect. 9.3. Obviously, the introduced methods of linguistic analysis can also be applied to written text, albeit with slightly different pre-processing. Also, models trained on written text may differ insofar as spoken language is often grammatically different and possesses more fragments of words, etc.

9.1.1 A Short Motivation

Paralinguistic speech and language analysis, i.e., the analysis of consciously or unconsciously expressed non-verbal elements of communication, is constantly developing into a major field of speech analysis, as new human–machine interaction and media retrieval systems advance over sheer speech recognition.

The additional information over 'what' is being said bears high potential for improved interaction or retrieval of speech files. By such information, social competence is provided to systems that can react more human-like or provide more human-like information. In addition, this information can also help to better recognise 'what' is being said, as acoustic and linguistic models can be adapted to differ-

B. Schuller (✉)
Institute for Human–Machine Communication, Technische Universität München, 80290 Munich, Germany
e-mail: schuller@tum.de

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_9,

ent speaker states and traits and non-verbal outbursts are not confused with linguistic entities [52]. A number of such paralinguistic phenomena are next given.

9.1.2 From Affection to Zest

One can broadly divide the multifaceted field of paralinguistics into speaker states and speaker traits and vocal behaviour. Speaker *states* thereby deal with states changing over time, such as affection and intimacy [5], deception [14], emotion [7], interest [41], intoxication [35], sleepiness [24], health state [19], and stress [21] or zest, while the speaker *traits* identify permanent speaker characteristics such as age and gender [46], height [32], likeability [57], or personality [31]. Vocal behaviour additionally comprises non-linguistic vocal outbursts like sighs and yawns [34], laughs [10], cries [33], hesitations and consent [41], and coughs [30]. We next deal with the principle of how to computationally analyse any of these automatically.

9.1.3 Principle

Here we share a unified perspective on the computationally 'intelligent' analysis of speech as a general pattern recognition paradigm.

Figure 9.1 gives an overview of the typical steps in such a system. The dotted lines indicate the training or learning phase that is usually carried out once before using such a system in practice. It can, however, re-occur during application in the case of online or unsupervised and semi-supervised adaptation. Interestingly, the information is partly well suited for online learning based on user feedback, as user (dis-)satisfaction or similar states and affirmative vocalisations can be used to adapt models accordingly.

The building blocks of a voice and speech analysis system are:

Pre-processing usually deals with enhancement of signal properties of interest from input speech. Such speech may be coming from a capture device like an A/D converter in a live setting, or from offline databases of stored audio files for training and evaluation purposes. Such enhancement includes de-reverberation and noise suppression, e.g., by exploitation of multiple microphones, or separation of multiple speakers by blind source separation.

Feature Extraction deals with the reduction of information to the relevant characteristics of the problem to be investigated in the sense of a canonical representation and will be dealt with in more detail—separately for acoustic and linguistic features.

Classification/Regression assigns the actual label to an unknown test instance. In the case of classification, discrete labels such as Ekman's 'big six' emotion classes (anger, disgust, fear, happiness, sadness, and surprise) or, e.g., binary low/high

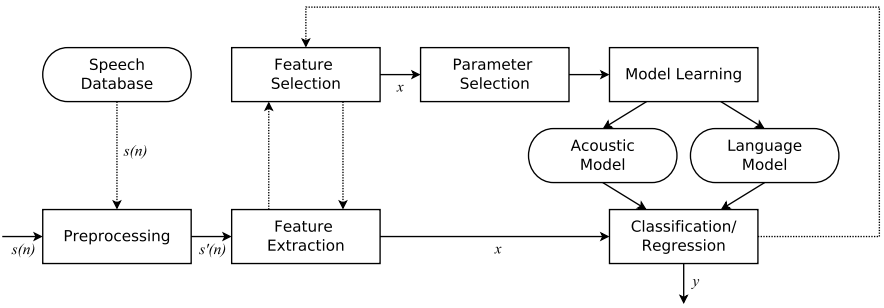


Fig. 9.1 Analysis of voice and speech—an overview. *Dotted lines* indicate the training phase

labels per each of the 'big five' personality dimensions (openness, conscientiousness, extraversion, agreeableness, and neuroticism—"OCEAN") are decided for. In the case of regression, the output is a continuous value like a speaker's height in centimetre or age in years, or—in the case of emotion—dimensions like potency, arousal, and valence, typically ranging from -1 to $+1$. We will discuss the frequently encountered machine learning algorithms in the field later on.

Speech Databases comprise the stored audio of exemplary speech for model learning and testing. In addition, a transcription of the spoken content may be given and the labelling of the problem at hand, such as speaker emotion, age, or personality. Usually, one wishes for adequate data in the sense of natural data rather than elicited or acted in ideal conditions, excluding disruptive influence or well-described and targeted noise or reverberation, a high total amount—which is rarely given. Further, data should ideally include a large number of speakers, a meaningful categorisation, which is usually non-trivial in this field (cf. the emotion categories vs. dimensions), a reliable annotation either by the speaker herself or a higher number of annotators to avoid skewness, additional perception tests by independent labellers to provide a comparison of human performance on the task, balanced distribution of instances among classes or the dimensional continuum, knowledge of the prior distribution, high diversity of speakers' ages, gender, ethnicity, language, etc., and high spoken content variation. Finally, one wishes for well defined test, development, and training partitions without prototypical selection of 'friendly cases' for classification [49], free availability of the data, and well-documented meta-data.

Model Learning is the actual training phase in which the classifier or regressor model is built, based on labelled data. There are classifiers or regressors that do not need this phase—so called lazy learners—as they only decide at run-time by training instances' properties which class to choose, e.g., by the training instance with shortest distance in the feature space to test instances [20]. However, these are seldom used, as they typically do not lead to sufficient accuracy in the rather complex task of speech analysis.

Feature Selection decides which features actually to keep in the feature space. This may be of interest if a new task, e.g., estimation of a speaker's weight from acoustic properties, is not well known. In such a case, a multiplicity of features can be

'brute-forced', as will be shown. From these, the ones well suited for the task at hand can be kept.

Parameter Selection fine 'tunes' the learning algorithm. Indeed, the performance of a machine learning algorithm can be significantly influenced by optimal or sub-optimal parametrisation. As for the feature selection, it is crucial not to 'tune' on speech instances used for evaluation as obviously this would lead to overestimation of performance.

Acoustic Models consist of the learnt dependencies between acoustic observations and classes, or continuous values in the case of regression, stored as binary or text files.

Language Models resemble acoustic models—yet, they store the learnt dependencies of linguistic observations and according assignments.

9.2 'Voice'—the Acoustic Analysis

In this section we will be dealing with the acoustic properties of the voice ignoring 'what' is being said and entirely focusing on 'how' it is said (cf. also Chap. 10). For this analysis, we will first need to chunk the audio stream (for an example see Fig. 9.2a) before extracting features for these chunks and then proceed with the selection of relevant features before the classification/regression, and 'fine tuning'.

9.2.1 Chunking

Already for the annotation of human behaviour that changes over time, one mostly needs to 'chunk' the speech, which is often stored as a single file ranging over several seconds up to hours, into 'units of analysis'. These chunks may be based on the 'quasi-stationarity' of the signal, as given by single frames obtained by applying a window function to the signal—typically having a length of some 10–30 ms and applied every 10 ms as the window often has a softening character at its ends, or larger units of constant duration. Most frequently, though, 'turns' are analysed that are based on speech onset until offset of one speaker in conversations. Onset and offset of speech are thereby often determined by a simple signal energy-based hysteresis, i.e., for a given minimum time, the speech pause energy level has to be exceeded to determine a speech onset and vice versa. While being an objective measure which is somewhat easy to obtain automatically, such turns may highly vary in length.

Alternatives are either pragmatic units like time slices, or proportions of longer units obtained by subdivision into parts of relative or absolute equal length, or 'meaningful' units with varying lengths, such as syllables, words, phrases. In [6], the word as the smallest possible, meaningful unit, is favoured for the analysis of emotion in speech, and in [50] it is shown that stressed syllables alone can be on a par with words as far as classification performance is concerned.

One may assume that units that are more connected to the task of analysis will become important in future research. In addition, incremental processing will be of increasing interest. Such incremental processing means providing an online estimate after the onset, updated continuously until the offset—this is often referred to as 'gating'. Additionally, one may want to decide for the optimal unit in a multimodal context if for example also video or physiological information is analysed that typically investigates different units, but shall be fused in a synergistic manner. In fact, this problem can already arise to a certain extent when we want to fuse acoustic and linguistic information. Even for processing exclusively acoustic information, consideration of several temporal units at the same time may be interesting, to benefit from shorter frames in the case of spectral characteristics, but larger 'supra-segmental' units in the case of prosodic features, i.e., features dealing with intonation, stress, and rhythm, such as speaker's pitch.

9.2.2 Acoustic Feature Extraction

Arguably the most important step in the automated recognition of speaker states, traits and vocal behaviour is the extraction of features that are relevant for the task at hand and providing a compact representation of the problem.

Let us divide features into groups in the following to provide a comprehensive overview. While there is no unique classification into such groups, the most basic distinction is technology driven: The main groups are at first *acoustic* and *linguistic* features.

Depending on the type of affective state of vocal behaviour one aims to analyse, different weights will be given to these. To give an obvious example, linguistic features are of limited interest when assessing non-verbal vocal outbursts such as laughter, sighs, etc. However, investigating a speaker's emotion or personality, they bear high potential.

In the past, the common focus was put on prosodic features, more specifically on pitch, duration and intensity, and less frequently on voice quality features as harmonics-to-noise ratio (HNR), jitter, or shimmer. Segmental, spectral features modelling formants, or cepstral features (MFCC) are also often found in the literature. More details of these features will be given later.

Until recently, a comparably small feature set (around 20 to 30 features) has usually been employed. The recent success of systematically generated static feature vectors is probably justified by the supra-segmental nature of most paralinguistic phenomena. These features are derived by projection of the low-level descriptors (LLD, for examples see Fig. 9.2b–f) on single scalar values by descriptive statistical functionals, such as lower order moments or extrema. As an alternative, LLD features can be modelled directly. In general, these LLD calculate a value per speech 'frame' with a typical frame rate of 100 frames per second (fps, cf. Sect. 9.2).

The large number of LLD and functionals has recently promoted the extraction of very large feature vectors (brute-force extraction), up to many thousands of features

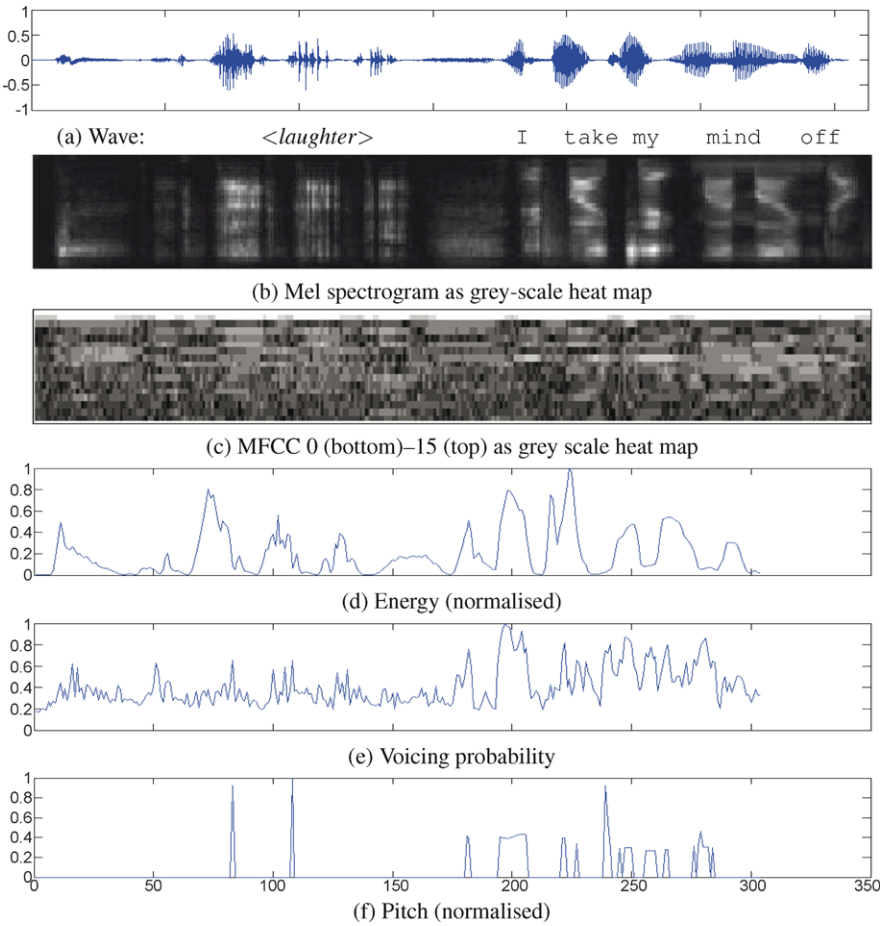


Fig. 9.2 Exemplary speech wave form over time in ms: laughter (0.0–150 ms) followed by "I take my mind off" taken from the SAL database (male speaker) and selected low-level descriptors

obtained either by analytical feature generation or, in a few studies, by evolutionary generation (note that a similar development can be found in vision analysis, where large amounts of features are produced and then reduced). Such brute-forcing also often includes hierarchical functional application (e.g., mean of maxima) to better cope with statistical outliers.

However, also expert-based hand-crafted features still play their role, as these are lately often crafted with more emphasis put on details hard to find by sheer brute-forcing such as perceptually more adequate ones, or more complex features such as articulatory ones, for instance, (de-)centralisation of vowels (i.e., how exact and constant are vowels articulated). This can thus also be expected as a trend in future acoustic feature computation.

Let us now introduce the groups of features.

Intensity features usually model the loudness of a sound as perceived by the human ear, based on the amplitude, whereby different types of normalisation are applied. Often, however, simply the frame energy is calculated for simplification, as human loudness perception requires a complex model respecting effects of duration and pitch of sound. As the intensity of a stimulus increases, the hearing sensation grows logarithmically (decibel scale). It is further well-known that sound perception also depends on the spectral distribution. The loudness contour is thus the sequence of short-term loudness values extracted on a frame-by-frame basis.

The basics of *pitch* extraction have largely remained the same over the years; nearly all Pitch Detection Algorithms (PDA) are built using frame-based analysis: The speech signal is broken into overlapping frames and a pitch value is inferred from each segment mostly by the maximum in the autocorrelation function (ACF) in its manifold variants and derivatives such as Average Magnitude Difference Function (AMDF). AMDF substitutes the search of a maximum by a minimum search, as instead of multiplication of the signal with itself, a subtraction is considered for improved efficiency. Often, the Linear Predictive Coding (LPC) residual or a band-pass filtered version is used over the original signal to exclude other influences from the vocal tract position. Pitch can also be determined in the time signal which allows for analysis of micro-perturbations, but is usually more error-prone. Pitch features are often made perceptually more adequate by logarithmic/semitone transformation, or normalisation with respect to some (speaker-specific) baseline. Pitch extraction is error-prone itself, which may influence recognition performance of the actual target problem [4]. However, the influence is rather small, at least for the current state-of-the-art in modelling pitch features.

Voice quality is a complicated issue in itself, since there are many different measures of voice quality [28], mostly clinical in origin and mostly evaluated for constant vowels only. Other, less well-known voice quality features were intended towards normal speech from the outset, e.g., those modelling 'irregular phonation', cf. [3]. Noise-to-Harmonic Ratio, jitter (micro-perturbation of pitch), shimmer (micro-perturbation of energy), and further micro-prosodic events are measures of the quality of the speech signal. Although they depend in part on other LLDs such as pitch and energy, they reflect peculiar voice quality properties such as breathiness or harshness.

The *spectrum* is characterised by formants (spectral maxima depending on the vocal tract position) modelling spoken content, especially the lower ones. Higher formants also represent speaker characteristics. Each one is fully represented by position, amplitude and bandwidth. The estimation of formant frequencies and bandwidths can be based on LPC or on cepstral analysis. A number of further spectral features can be computed either directly from a spectral transform such as by Fast Fourier Transform or the LPC spectrum, such as centroid, flux, and roll-off. Furthermore, the long term average spectrum over a unit can be employed: this averages out formant information, giving general spectral trends.

The *cepstrum*, i.e., the inverse spectral transform of the logarithm of the spectrum, emphasises changes or periodicity in the spectrum, while being relatively robust against noise. Its basic unit is frequency. Mel-Frequency Cepstral Coefficients (MFCCs)—as homomorphic transform with equidistant band-pass filters on

the Mel-scale—tend to strongly depend on the spoken content. Yet, they have been proven beneficial in practically any speech processing task. MFCC are calculated based on the Fourier transform of a speech frame. Next, overlapping windows—usually of triangular shape and equidistant on the Mel scale—are used for mapping the powers of the obtained spectrogram onto the Mel scale to model human frequency resolution. Next, the logarithms of the powers are taken per such Mel frequency filter band—the idea at this point is to decouple the vocal tract transfer function from the excitation signal of human sound production. Then, the Discrete Cosine Transform (DCT) of the list of mel log powers is taken for decorrelation (other transforms are often used as well) to finally obtain the MFCCs as the amplitudes of the resulting DCT spectrum.

Perceptual Linear Predictive (PLP) coefficients and MFCCs are extremely similar, as they both correspond to a short-term spectrum smoothing—the former by an autoregressive model, the latter by the cepstrum—and to an approximation of the auditory system by filter-bank-based methods. At the same time, PLP coefficients are also an improvement of LPC by using the perceptually based Bark filter bank. Variants such as Mel Frequency Bands (MFB) that do not decorrelate features as a final step are also found in this particular field.

Wavelets give a short-term multi-resolution analysis of time, energy and frequencies in a speech signal. Compared to similar parametric representations they are able to minimise the time-frequency uncertainty.

Duration features model temporal aspects. Relative positions on the time axis of base contours like energy and pitch such as maxima or on/offset positions do not strictly represent energy and pitch, but duration—because they are measured in seconds, and because they are often highly correlated with duration features. By that, they can be distinguished according to the way they are extracted: Those that represent temporal aspects of other acoustic base contours, and those that exclusively represent the 'duration' of higher phonological units, like phonemes, syllables, words, pauses, or utterances. Duration values are usually correlated with linguistic features: For instance, function words are shorter on average, content words are longer: This information can be used for classification, no matter whether the signal is encoded in linguistic, or acoustic (i.e., duration) features.

Subsequent to the LLD extraction, a number of operators and functionals can be applied to obtain feature vectors of equal size from each LLD. Functionals provide a sort of normalisation over time: LLD associated with words (and other units) have different lengths, depending on the duration of each word and on the dimension of the window step; with the usage of functionals, we obtain one feature vector per chunk, with a constant number of elements that can be modelled by a static classifier or regressor. This cascade procedure, namely LLD extraction followed by functional application, has two major advantages: Features derived from longer time intervals can be used to normalise local ones, and the overall number of features might be opportunely shrunk or expanded with respect to the number of initial LLDs [48].

More intelligent brute-forcing can be obtained by search masks and by a broader selection of functionals and parameters. In this way, an expert's experience can be combined with the freedom of exploration taken by an automatic generation.

Before functionals are applied, LLDs can be filtered or (perceptually) transformed, and first or second derivatives are often calculated and end up as additional LLDs. Functionals can range from statistical ones to curve fitting methods. The most popular statistical functionals cover the first four moments (mean, standard deviation, skewness and kurtosis), higher order statistics (extreme values and their temporal information), quartiles, amplitude ranges, zero-crossing rates, roll-on/-off, on/offsets and higher level analysis. Curve fitting methods (mainly linear) produce regression coefficients, such as the slope of linear regression, and regression errors (such as the mean square errors between the regression curve and the original LLD). A comprehensive list of functionals adopted so far in this field can be found in [7].

Figure 9.3 provides an overview of the commonly used features and the principle of their brute-forcing in several layers.

As a typical example, we can have a look at the 'large' feature set of the public open source toolkit openSMILE [16] that is frequently used in the field: Acoustic feature vectors of 6.552 dimensions are extracted as 39 functionals of 56 acoustic LLDs, including first and second order delta regression coefficients: Table 9.1 shows the LLDs to which the statistical functionals are applied that are summarised in Table 9.2 to map a time series of variable length onto a static feature vector as described above.

9.2.3 Feature Selection

To improve reliability and performance, but also to obtain more efficient models in terms of processing speed and memory requirements, one usually has to select a subset of features that best describe the audio analysis task. A multiplicity of feature selection strategies have been employed, e.g., for recognition of emotion or personality, but even for non-linguistic vocalisations, different types of features are often considered and selected.

Ideally, feature selection methods should not only reveal single (or groups of) most relevant attributes, but also decorrelate the feature space. Wrapper-based selection—that is employing a target classifier's accuracy or regressor's cross-correlation as optimisation criterion in 'closed loop'—is widely used to tailor the feature set in match with the machine learning algorithm. However, even for relatively small data sets, exhaustive selection considering any permutation of features is still not affordable. Therefore, the search in the feature space must employ some more restrictive, and thus less optimal, strategies. Probably the most common procedure chosen is the *sequential forward search*—a hill climbing selection starting with an empty set and sequentially adding best features; as this search function is prone to nesting, an additional floating option should be added: At each step one or more features are deleted and it is checked if others are more suited.

Apart from wrappers, less computationally expensive 'filter' or 'open loop' methods are frequently used if repeated selection is necessary, such as information theoretic filters and correlation-based analysis.

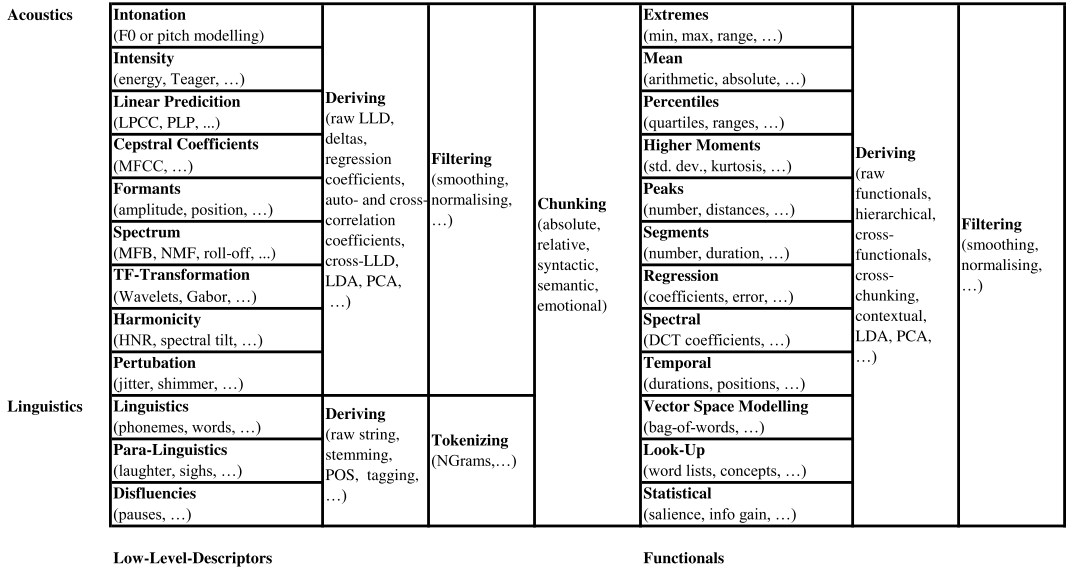


Fig. 9.3 Overview on features commonly used for acoustic and linguistic emotion recognition. Abbreviations: Linear Prediction Cepstral Coefficients (LPCC), Mel Frequency Bands (MFB)

Table 9.1 33 exemplary typical Low-Level Descriptors (LLD)

Feature Group	Features in Group
Raw Signal	Zero-crossing-rate
Signal energy	Logarithmic
Pitch	Fundamental frequency F_0 in Hz via Cepstrum and Autocorrelation (ACF). Exponentially smoothed F_0 envelope.
Voice Quality	Probability of voicing ($\frac{ACF(T_0)}{ACF(0)}$)
Spectral	Energy in bands 0–250 Hz, 0–650 Hz, 250–650 Hz, 1–4 kHz 25%, 50%, 75%, 90% roll-off point, centroid, flux, and rel. pos. max./min.
Mel-spectrum	Band 1–26
Cepstral	MFCC 0–12

Table 9.2 39 exemplary functionals as typically applied to LLD contours

Functionals	#
Respective rel. position of max./min. value	2
Range (max.-min.)	1
Max. and min. value—arithmetic mean	2
Arithmetic mean, Quadratic mean, Centroid	3
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, Mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95% and 98% percentile	2
Std. deviation, variance, kurtosis, skewness	4
Zero-crossing rate	1
# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks—overall arth. mean	4
Linear regression coefficients and error	4
Quadratic regression coefficients and error	5

There are, however, also classifiers and regressors with 'embedded' selection, such as Decision Trees or Ridge Regression.

As a refinement, *hierarchical* approaches to feature selection try to optimise the feature set not globally for all target classes, but for groups of them, mainly couples.

Apart from genuine selection of features, the *reduction* (i.e. feature extraction) of the feature space is often considered to reduce the complexity and number of free parameters to be learnt for the machine learning algorithms while benefiting from all original feature information. This is achieved by mapping of the input space onto a less dimensional target space, while keeping as much information as possible.

Principal Component Analysis (PCA) and Linear or Heteroscedastic Discriminant Analysis (LDA) are the most common techniques.

While PCA is an unsupervised feature reduction method and thus is often sub-optimal for more complex problems, LDA is a supervised feature reduction method which searches for the linear transformation that maximises the ratio of the determinants of the between-class covariance matrix and the within-class covariance matrix, i.e., it is a discriminative method as the name indicates.

In fact, none of these methods is optimal: There is no straight forward way of knowing the optimal target space size—typically the variance covered is a decisive measure. Further, a certain degree of normal distribution is expected, and LDA additionally demands linear separability of the input space. PCA and LDA are also not very appropriate for feature mining, as the original features are not retained after the transformation.

Finally, Independent Component Analysis (ICA) and Non-negative Matrix Factorization (NMF) [25] can be named. ICA maps the feature space onto an orthogonal space and the target features have the attractive property of being statistically independent. NMF is a recent alternative to PCA in which the data and components have to be non-negative. NMF is at present mainly employed for large linguistic feature sets.

Also, it seems important to mention that there is a high danger of over-adaptation to the data that features are selected upon. As a counter-measure, it seems wise to address feature importance across databases [15].

9.2.4 Classification and Regression

A number of factors motivate consideration of diverse machine learning algorithms, the most important being tolerance to high dimensionality, capability of exploiting sparse data, and handling of skewed classes. In addition, more general considerations such as the ability to solve non-linear problems, discriminative learning, self-learning of relevant features, high generalisation, on-line adaptation, handling of missing data, efficiency with respect to computational and memory costs in training and recognition, etc. can play a decisive role. Further, one may wish for human-readable learnt models, provision of meaningful confidence measures and handling of input uncertainties (features like pitch are not determined flawlessly—here an algorithm may also consider a certainty measure in addition to the predicted pitch value) for optimal integration in a system context.

As previously mentioned, we can basically differentiate between classifiers that decide for discrete classes and regressors that estimate a continuous value in the sense of a function learner. However, practically any classifier can be turned into a regressor and vice versa, although the result would not necessarily be as efficient for this task as for its 'native' task. Classification using regression methods can for example be obtained by having each class binarised and one regression model built for each class value. The other way round, a regression scheme can be realised by

using any classifier on a copy of the data where the continuous 'class' is discretised. The predicted value is the expected value of the mean class value for each discretised interval, based on the predicted probabilities for each interval [58] (also see 'squashing' in Chap. 1).

The problem of a high dimensional feature set is usually better addressed by feature selection and elimination before actual classification takes place. Popular classifiers such as Linear Discriminant Classifiers (LDCs) and k-Nearest Neighbor (kNN) classifiers have been used since the very first studies. However, they suffer from the increasing number of features that leads to regions of the feature space where data are very sparse ('curse of dimensionality'). Classifiers such as kNN that divide the feature space into cells are affected by the curse of dimensionality and are sensitive to outliers. A natural extension of LDCs are Support Vector Machines (SVM): they combine discriminative learning and solving of non-linear problems by a Kernel-based transformation of the feature space. While they may not always lead to the best result, they provide good generalisation properties, and can be seen as a sort of state-of-the-art classifier (or regressor, as the related Support Vector Regression allows for handling of continuous problem descriptions).

Small data sets are, in general, best handled by discriminative classifiers. The most used non-linear discriminative classifiers apart from SVM are likely to be Artificial Neural Networks (ANNs) and decision trees. Decision hyperplanes learnt with ANN might become very complex and depend on the topology of the network (number of neurons), on the learning algorithm (usually a derivation of the well-known Backpropagation algorithm) and on the activity rules. For this reason, ANNs are less robust to over-fitting, and require greater amounts of data to be trained on. The recent incorporation of a long–short-term memory function seems to be a promising future direction [60] that may raise their popularity. Also, multi-task learning is well established, which may be of particular interest in this field to, e.g., assess emotion and personality in one pass, benefiting from mutual dependencies.

Decision trees are also characterised by the property of handling non-linearly separable data; moreover, they are less of a 'black box' compared to SVM or neural networks, since they are based on simple recursive splits (i.e., questions) of the data. These binary questions are very readable, especially if the tree has been adequately pruned. As accuracy degrades in case of irrelevant features or noisy patterns, Random Forests (RF) can be employed: They consist of an ensemble of trees, each one accounting for random, small subsets of the input features obtained by sampling with replacement. They are practically insensitive to the curse of dimensionality, while, at the same time, still providing all the benefits of classification trees.

As many paralinguistic tasks (such as emotion) are not evenly distributed among classes in databases, balancing of the training instances with respect to instances per class is often a necessary step before classification [43]. The balancing of the output space can be addressed either by considering proper class weights (e.g., priors), or by resampling, i.e., (random) up- or down-sampling. Class priors are implicitly taken into account by discriminative classifiers.

As explained above, applying functionals to LLD is done for obtaining the same number of features for different lengths of units such as turns or words. Dy-

namic classifiers like Hidden Markov Models, Dynamic Bayesian Networks or simple Dynamic Time Warp allow to skip this step in the computation by implicitly warping observed feature sequences over time. Among dynamic classifiers, Hidden Markov Models (HMM) have been used widely. The performance of static modelling through functionals is often reported as superior [43], as paralinguistic tasks are apparently better modelled on a time-scale above frame level; note that a combination of static features such as minimum, maximum, onset, offset, duration, regression, etc. implicitly shape contour dynamics as well. A possibility to use static classifiers for frame-level feature processing is further given by multi-instance learning techniques, where a time series of unknown length is handled by SVM or similar techniques. Also, a combination of static and dynamic processing may help improve overall accuracy [55].

Ensembles of classifiers combine their individual strengths, and might improve training stability. There exists a number of different approaches to combine classifiers. Popular are methods based on majority voting such as *Bagging*, *Boosting* and other variants (e.g., *MultiBoosting*). More powerful, however, is the combination of diverse classifiers by the introduction of a meta-classifier that learns 'which classifier to trust when' and is trained only on the output of 'base-level' classifiers, known as *Stacking*. If confidences are provided on lower level, they can be exploited as well. Still, the gain over single strong classifiers such as SVM may not justify the extra computational costs [37].

In line with the different models to describe the named problems, e.g., by classes or continuous dimensions, also different approaches towards classification are needed: As real-life application is not limited to prototypical cases, also *detection* as opposed to classification can be expected as an alternative paradigm: 'Out-of-vocabulary' classes need to be handled as well (as an example, imagine the emotions anger and neutral having been trained, but in the recognition phase joy appears), and apart from the easiest solution of introducing a garbage class [43], detection allows for more flexibility. Detection is thereby defined by inheriting a rejection threshold. In this respect, *confidence measurements* should be mentioned, which are, however, not sufficiently explored, yet.

9.2.5 Parameter Tuning

Apart from the selection of features, a crucial factor in optimisation of performance is the 'fine tuning' of classifiers' parameters on a development partition of the training data. Typically such parameters comprise the exponent of polynomial Kernels for Support Vector Machines or the number of nearest neighbors in k nearest neighbor classification, etc. While these can be optimised by equidistant scanning of the parameter space, more efficient methods exist, of which grid search is the most frequently encountered in the field (e.g., [23]). Grid search is a greedy algorithm that first performs a rough search over the values and then narrows down on promising areas in terms of best accuracy for a classifier or cross-correlation for a regressor

in a recursive manner. Obviously, just as for the selection of features, such searches do not necessarily lead to the global optimum, if the search is not exhaustive. In addition, they also may differ drastically for different databases, depending on their size and complexity. Thus, again cross-corpus parameter tuning may help find more generally valid sets than considering just intra-database variation.

9.3 'Speech'—the (Non-)linguistic Analysis

As said, in this chapter speech stands for spoken text, i.e., the analysis of textual cues. Apart from the analysis of linguistic content, non-linguistic vocal outbursts as laughter are dealt with.

9.3.1 Analysis of Linguistic Content

Spoken or written text provides cues on emotion, personality or further states and traits. This is usually reflected in the usage of certain words or grammatical alterations, which means in turn, in the usage of specific higher semantic and pragmatic entities. A number of approaches exists for this analysis: keyword spotting [12], rule-based modelling [26], Semantic Trees [61], Latent Semantic Analysis [17], World-knowledge-Modelling, Key-Phrase-Spotting, String Kernels [40], and Bayesian Networks [8]. Contextual and pragmatic information has been modelled as well, e.g., dialogue acts [26], or system and user performance [1]. Two methods seem to be predominant, presumably because they are shallow representations of linguistic knowledge and have already been frequently employed in automatic speech processing: (*class-based*) *N-Grams* and *Bag of Words* (*vector space modelling*), cf. [38].

N-Grams and Class-based N-Grams are commonly used for general language modelling. Thereby the posterior probability of a (class of a) word is given by its predecessors from left to right within a sequence of N words. For recognition of a target problem such as emotion or personality, the probability of each target class is determined per N-gram of an utterance. In addition, word-class-based N-grams can be used as well, to better cope with data sparseness. For the example of emotion recognition, due to data sparseness mostly uni-grams ($N = 1$) have been applied so far, besides bi-grams ($N = 2$) and trigrams ($N = 3$) [2]. The actual target class is calculated by the posterior probability of the class given the actual word(s) by maximum likelihood or a-posteriori estimation. An extension of N-Grams which copes with data sparseness even better is *Character N-Grams*; in this case larger histories can be used.

Bag of Words is a well-known numerical representation form of texts in automatic document categorisation [22]. It has been successfully ported to recognise sentiments or emotion [38] and can equivalently be used for other target problems. In this approach each word in the vocabulary adds a dimension to a linguistic vector

representing the term frequency within the actual utterance. Note that easily very large feature spaces may occur, which usually require intelligent reduction. The logarithm of frequency is often used; this value is further better normalised by the length of the utterance and by the overall (log)frequency within the training corpus.

In addition, exploitation of on-line knowledge sources without domain specific model training has recently become an interesting alternative or addition [42]—e.g., to cope with out-of-vocabulary events. The largely related fields of opinion mining and sentiment analysis in text bear interesting alternatives and variants of methods.

Although we are considering the analysis from spoken text, only few results for paralinguistic speaker state and trait recognition rely on automatic speech recognition (ASR) output [45] rather than on manual annotation of the data. As ASR of affective speech itself is a challenge [52], this step is likely to introduce errors. To some extent errors deriving from ASR and human transcription can be eliminated by soft-string-matching such as tolerating a number of deletions, insertions, or substitutions of characters.

To reduce the complexity for the analysis, *stopping* is usually used. This resembles elimination of irrelevant words. The traditional approach towards stopping is an expert-based list of words, e.g., of function words. Yet, even for an expert it seems hard to judge which words can be of importance in view of the target problem. Data-driven approaches like salience or information gain based reduction are popular. Another often highly effective way is stopping words that do not exceed a general minimum frequency of occurrence in the training corpus.

Tokenisation, i.e., chunking of the continuous text string similar to chunking of the acoustic stream above, can be obtained by mapping the text onto word classes: *Stemming* is the clustering of morphological variants of a word (such as “fight”, “fights”, “fought”, “fighting”, etc.) by its stem into a *lexeme*. This reduces the number of entries in the vocabulary, while at the same time providing more training instances per class. Thereby also words that were not seen in the training can be mapped upon their representative morphological variant, for instance by (Iterated) Lovins or Porter stemmers that are based on suffix lists and rules. Part-of-Speech (POS) tagging is a very compact approach where classes such as nouns, verbs, adjectives, particles, or more detailed sub-classes are modelled [51]. POS tagging and stemming have been studied thoroughly [40].

Also *sememes*, i.e., semantic units represented by lexemes, can be clustered into higher semantic concepts such as generally positive or negative terms [7]. In addition, non-linguistic vocalisations can easily be integrated into the vocabulary [41].

9.3.2 Analysis of Non-linguistic Content

While non-linguistic events such as laughter can be modelled as an extra type of feature stream or information, a very simple way is to include them in the string of linguistic events. On the positive side, this can put events like laughter in direct

relation with the words. This may, however, disrupt linguistically meaningful sequences of words. Alternatively, frequencies of occurrences normalised to time or even functionals applied to occurrences are alternative solutions.

9.3.3 (Non-)linguistic Feature Extraction

While non-linguistic events can be recognised directly in-line with speech as by an Automatic Speech Recogniser, it seems noteworthy to mention that one can also use brute-forced features as described above. Interestingly, little to no difference is reported for these two types of representation [41]. The incorporation into a speech recogniser has the advantage that speech is recognised with integration of higher-level knowledge as coming from the language model. However, if non-linguistic vocalisations are modelled on their own, a richer feature representation can be used that may unnecessarily increase space complexity for speech recognition. Furthermore, in case of non-linguistic vocalisations such as laughter, these may also appear 'blended' with speech, as in the case of 'speech-laughter', i.e., laughter while actually speaking words. This cannot easily be handled in-line with ASR, as the ASR engine typically would have to decide for phonetically meaningful units or laughter.

9.3.4 Classification and Regression

In principle, any of the formerly discussed learning algorithms can be used for linguistic analysis, as well. However, different ones may be typically preferred owing to the slightly different characteristics of linguistic features. In particular, statistical algorithms and Kernel machines such as Support Vector Machines are popular. Noteworthy, there are also specific algorithms that may operate directly on string input such as the String Kernels [27] for Support Vector Machines.

9.4 Data, Benchmarks, and Application Examples

In this section, let us first have a look at some typical databases focusing on affective speaker states. Next, two examples of systems that analyse vocal behaviour on different levels will shortly be described.

9.4.1 Frequently Encountered Data-Sets and Their Benchmarks

As benchmark databases, nine most frequently encountered databases that span a range from acted over induced to spontaneous affect portrayals are presented, focusing in particular on affective speaker states. For better comparability of obtained

performances among corpora, the diverse affect groups are additionally mapped onto the two most popular axes in the dimensional emotion model as in [47]: arousal (i.e., passive ("−") vs. active ("+")) and valence (i.e., negative ("−") vs. positive ("+")). Note that these mappings are not straight forward—here we will favour better balance among target classes. Let us further discretise into the four quadrants (q) 1–4 of the arousal-valence plane for continuous labelled corpora. In the following, each set is shortly introduced, including the mapping to binary arousal/valence by "+" and "−" per emotion and its number of instances in parentheses. Note that the emotions are referred to as in the original database descriptions.

The *Danish Emotional Speech* (DES) database [13] is professionally acted and contains nine sentences, two isolated words, and chunks that are located between two silent segments of two passages of fluent text. Affective states contain angry (+/−, 85), happy (+/+, 86), neutral (−/+, 85), sadness (−/−, 84), and surprise (+/+, 79).

The *Berlin Emotional Speech Database* (EMOD) [9] features professional actors speaking ten emotionally undefined sentences. 494 phrases are commonly used: angry (+/−, 127), boredom (−/−, 79), disgust (−/−, 38), fear (+/−, 55), happy (+/+, 64), neutral (−/+, 78), and sadness (−/−, 53).

The *eNTERFACE* (eNTER) [29] corpus consists of recordings of subjects from 14 nations speaking pre-defined spoken content in English. The subjects listened to six successive short stories eliciting a particular emotion out of angry (+/−, 215), disgust (−/−, 215), fear (+/−, 215), happy (+/+, 207), sadness (−/−, 210), and surprise (+/+, 215).

The *Airplane Behavior Corpus* (ABC) [39] is based on induced mood by pre-recorded announcements of a vacation (return) flight, consisting of 13 and 10 scenes. It contains aggressive (+/−, 95), cheerful (+/+, 105), intoxicated (+/−, 33), nervous (+/−, 93), neutral (−/+, 79), and tired (−/−, 25) speech.

The *Speech Under Simulated and Actual Stress* (SUSAS) database [21] serves as a first reference for spontaneous recordings. Speech is additionally partly masked by field noise in the chosen speech samples of actual stress. SUSAS content is restricted to 35 English air-commands in the speaker states of high stress (+/−, 1202), medium stress (+/−, 1276), neutral (−/+, 701), and scream (+/−, 414).

The *Audiovisual Interest Corpus* (AVIC) [41] consists of spontaneous speech and natural emotion. In its scenario setup, a product presenter leads subjects through a commercial presentation. AVIC is labelled in "levels of interest" (loi) 1–3 having loi1 (−/−, 553), loi2 (+/+, 2279), and loi3 (+/+, 170).

The *Belfast Sensitive Artificial Listener* (SAL) data are part of the HUMAINE database. The subset used—as in [59]—has an average length of 20 minutes per speaker of natural human-SAL conversations. The data have been labelled continuously in real time with respect to valence and activation, using a system based on FEELtrace [11]. The annotations were normalised to zero-mean globally and scaled so that 98% of all values are in the range from -1 to $+1$. The 25 recordings have been split into turns using energy based Voice Activity Detection. Labels for each obtained turn are computed by averaging over the complete turn. Per quadrant the samples are: q1 (+/+, 459), q2 (−/+, 320), q3 (−/−, 564), and q4 (+/−, 349).

Table 9.3 Overview on the selected corpora (E/D/G: English/German/Danish, act/ind/nat: acted/induced/natural, Lab: labellers, Rec: recording environment, f/m: (fe-)male subjects). Speaker-independent recognition performance benchmarks are provided by weighted (WA) and unweighted (UA) average accuracy. * indicates results obtained by Support Vector Machines if these had outperformed Deep Neural Networks as taken in all other cases

Corpus	Speech	# All	h:mm	# m	# f	# Lab	Rec	kHz	# All		# Arousal		# Valence	
									UA	WA	UA	WA	UA	WA
ABC	G fixed act	430	1:15	4	4	3	studio	16	56.1	61.5	69.3	80.6	79.6	79.0
AVIC	E free nat	3002	1:47	11	10	4	studio	44	59.9	79.1	75.6	85.3	75.2	85.5
DES	D fixed act	419	0:28	2	2	–	studio	20	59.9*	60.1*	90.0	90.3	71.7	73.7
EMOD	G fixed act	494	0:22	5	5	–	studio	16	84.6*	85.6*	97.6	97.4	82.2	87.5
eNTER	E fixed ind	1277	1:00	34	8	2	studio	16	72.5*	72.4*	78.1	79.3*	78.6*	80.2*
SAL	E free nat	1692	1:41	2	2	4	studio	16	35.9	34.3	65.1	66.4	57.7	53.0
Smart	G free nat	3823	7:08	32	47	3	noisy	16	25.0	59.5	55.2	79.2	52.2	89.4
SUSAS	E fixed nat	3593	1:01	4	3	–	noisy	8	61.4*	56.5*	68.2	83.3	74.4	75.0
VAM	G free nat	946	0:47	15	32	6/17	noisy	16	39.3	68.0	78.4	77.1	52.4	92.3

The *SmartKom* (Smart) [53] corpus consists of Wizard-Of-Oz dialogues. For evaluations, the dialogues recorded during a public environment technical scenario are used. It is structured into sessions which contain one recording of approximately 4.5 min length with one person, and labelled as anger/irritation (+/−, 220), helplessness (+/−, 161), joy/gratification (+/+, 284), neutral (−/+, 2179), pondering/reflection (−/+, 643), surprise (+/+, 70), and unidentifiable episodes (−/+, 266).

Finally, the *Vera-Am-Mittag* (VAM) corpus [18] consists of recordings taken from a German TV talk show. The audio recordings were manually segmented to the utterance level, whereas each utterance contains at least one phrase. The labelling is based on a discrete five point scale for each of the valence, activation, and dominance dimensions. Samples among quadrants are q1 (+/+, 21), q2 (−/+, 50), q3 (−/−, 451), and q4 (+/−, 424).

Further details on the corpora are summarised in Table 9.3 and found in [44]. Looking at the table, some striking facts become evident: most notably, the high sparseness of data with these sets typically providing only one hour of speech from only around 10 subjects. In related fields as ASR, several hundreds of hours of speech and subjects are typically contained. This is one of the major problems in this field at the moment. In addition, one sees that often such data are rather acted as opposed to natural and that the linguistic content is often restricted to pre-defined phrases or words. Obviously, this is rather an annotation challenge, as emotional speech data per se would be available.

To provide an impression on typical performances in the field, the last columns of Table 9.3 provide weighted (WA) and unweighted (UA) accuracy of speaker-independent recognition by feature reduction with Deep Neural Networks and subsequent distance classification (DNN) or Support Vector Machines on the original

space (SVM), with the 'large' standard feature set of openSMILE introduced in Sect. 9.2. Such speaker independence is obtained by partitioning the data in a 'leave-one-speaker-out', or—for databases with many speakers, here starting at more than 10—'leave-one-speaker-group-out' cross-validation manner. This cross-validation is very popular in this field, as it allows to test on all instances of the very limited resources. The accuracy of the classifier that produced the higher result on development data is presented, each, in the table. Balancing of the training partition is used to cope with the imbalance of instances in the training set among affective states. More details are found in [54]. If we now look at these numbers, it seems clear that acted data are considerably easier to recognise automatically owing to their often exaggerated display. Naturally, this is more true in the case where the verbal content is limited. Another interesting but typical fact is that arousal is usually recognised more reliably. To better handle valence, one would best integrate linguistic feature information.

To conclude this chapter, let us now have a look at two examples of voice and speech analysis systems that are currently used in practice and that investigate a number of different issues in contrast to the above named problems.

9.4.2 Human-to-Human Conversation Analysis

The AVIC corpus as introduced above and as used in the INTERSPEECH Paralinguistic Challenge [46] provides a good example of vocal behaviour analysis in natural human conversational speech. In [41] an analysis of non-linguistic vocalisations and speaker's interest is shown, based on these non-linguistic vocalisations and further acoustic and linguistic features as introduced above. The acoustic feature space consists of a brute-force large space with subsequent feature selection with the classifier in the loop and SVM for classification. Linguistic features are the described Bag of Words, integrated directly into the feature vector. Non-linguistics are recognised in a separate recognition pass by the same basis of acoustic features but optimised for this task. The occurrence of non-linguistics is simply added to the linguistic feature string.

To demonstrate efficiency over weighted and unweighted accuracies like we presented in Table 9.3, 40 participants interacted with a virtual product and company tour that took participants' interest into account to change topic in case of their boredom. Three variants were used: topic change after a fixed time, with fully automatic interest recognition with this system or by a human Wizard-of-Oz. The question "Did you think the system was taking into account your interest?" was positively answered by 35% in the first case (no interest recognition), by 63% in the second case (fully automatic interest recognition) and by 84% in the last case (human interest recognition) nicely demonstrating that the technology seems to be generally working, but that there is also still headroom for improvement to reach human-like performance.

In our next example, let us switch to human–computer conversation.

9.4.3 Human-to-Agent Conversation Analysis

In the European SEMAINE project, a Sensitive Artificial Listener (SAL)—a multimodal dialogue system with the social interaction skills needed for a sustained conversation with a human user—was built [36]. Such a system demands for on-line incremental emotion recognition, in order to select responses as early as possible. In SEMAINE, the user's affective state and non-verbal behaviour are the major factors for deciding upon agent actions. Therefore it is essential to obtain a fast estimate of the user's affective state as soon as the user starts speaking, and refine the estimate as more data are available (for example, in [56] 350 ms are suggested for human-like back-channelling in certain situations). Moreover, the system needs to know how reliable the affect dimension predictions are, in order to identify salient parts of highly affective speech reliably, in order to choose appropriate actions. The verbal dialogue capabilities of the system are very limited on purpose. They are basically limited to agreement/disagreement, emotionally relevant keywords, and changing characters (see below for more information on the four different SEMAINE characters/personalities).

In the SEMAINE system, which is freely available as release for research and tutoring,[1] *Feature extractors* analyse low-level audio and video signals, and provide feature vectors periodically (10 ms) to the *analysers*, which process the low-level features and produce a representation of the current user state, in terms of epistemic-affective states (emotion, interest, etc.). Since, automatic speech recognition or emotion recognition might benefit from the dialogue context or user profiles at a higher level, *interpreter* components are contained in the system to address this issue. A typical and obvious example is the 'turn-taking interpreter', which decides when it is time for the agent to take the turn. These are examples—the SEMAINE API goes beyond these capabilities [36].

The next group of components is a set of *action proposers* which produce agent action candidates independently from one another. The action proposers take their input mainly from the user, dialog, and agent state. As for the voice and speech analysis, the free open source openSMILE[2] [16] module extracts state of the art features stemming from the large feature set described in Sect. 9.2 for voice activity detection, prosody analysis, keyword spotting, non-linguistic vocalisation detection, and an acoustic emotion recognition module. Prosodic features, which are used by other SEMAINE components (e.g., for turn-taking decisions), include pitch contour, energy/loudness, and per pseudo-syllable pitch direction estimates. Classification and regression are based on on-line Long Short-Term Memory (LSTM) Recurrent Neural Networks.

The SEMAINE keyword spotter detects a set of 176 keywords (including the non-linguistic vocalisations 'breathing', 'laughing', and 'sighing' handled in-line) which are relevant for the dialogue management and for linguistic emotion recognition. As system responses have to be prepared already before the user has finished

[1] http://semaine.sourceforge.net/

[2] http://www.openaudio.eu

speaking, the keyword spotter operates incrementally. The acoustic feature extractor extracts large sets of acoustic features used for recognition of the user's affective state (5 continuous dimensions: arousal, expectation, intensity, power, and valence, and 3 'levels of interest': bored, neutral, interested) incrementally in real time with regression models trained on the SEMAINE database. High dimensional acoustic feature vectors are concatenated with linguistic Bag of Words vectors, which are computed from the keyword spotter output. An incremental segmentation scheme is applied to the continuous audio input: analysis is conducted over windows of up to five seconds length, which are shifted forward in time with a step of two seconds, thus producing an estimate of the user's affective state every two seconds. The same acoustic feature set as for the 5 dimensional affect recognition is used in models trained on the AVIC corpus, as described in Sect. 9.4.

9.5 Summary

This chapter introduces the principles of analysis of acoustic and linguistic properties of the voice and speech for the assessment of speaker states, traits, and vocal behaviour such as laughter. While voice and speech analysis follows the general pattern recognition paradigm, one of its major peculiarities might be the choice of features. In particular, brute-forcing of rather large feature spaces and subsequent selection are common procedure. Further, the type of features—either low-level descriptors that provide a value per short frames of speech (usually around 100 per second), or functionals per larger units of time—decide on the type of classifier or regressor. Owing to the diversity of tasks reaching from emotion to personality or laughter, different machine learning algorithms are preferred and used. Features and parameters of these learning algorithms can be fine tuned to the problem and data at hand, yet this comes at the risk of over-adaptation.

Another main peculiarity is the ambiguity of ground truth due to the often very subjective nature of labelling and to the fact that models for the description of tasks like emotion or personality prediction are usually non-trivial. Finally, one of the most decisive limiting factors is typically the ever-present lack of data—in particular of natural data of a multiplicity of speakers and languages and cultures. However, reasonably functioning accuracies independent of speakers can already be provided, allowing for first systems to be operated 'in the wild'. At the same time, further research will usually be needed to achieve human-like performance for cross-database and task operation potentially in the presence of noise and reverberation.

9.6 Questions

1. Discuss the difference between speaker states and traits and list at least three examples for each of these two.
2. Name at least five ideal conditions for a collection of speech and voice data.

3. Describe the chain of processing for the analysis of speech and voice including each block and its function.
4. Explain the difference between Low-Level Descriptors and functionals and name at least five examples for each of these two.
5. Which units for chunking exist and why is chunking needed?
6. Name at least five ideal conditions for a classification or regression algorithm.
7. How can linguistic information be incorporated in the analysis process? Name at least two alternative strategies and describe their principle.

9.7 Glossary

- *Chunking*: Segmentation of the audio stream into units of analysis.
- *Low-Level Descriptor*: Time series of extracted feature values—typically on frame level.
- *Functional*: Projection of a function onto a scalar value by statistical or other functions.
- *Regression*: Mapping of a feature input vector onto a real-valued output instead of discrete classes as in classification.
- *Prosody*: Rhythm, stress, and intonation of speech.
- *N-Gram*: Subsequence (e.g., words or characters) from a given sequence (e.g., turns or words) with n consecutive items.
- *Bag of Words*: Representation of text (e.g., of a speaker turn) as numerical feature representation (e.g., per word or N-Gram of words) without modelling of order of units.
- *Wizard-of-Oz* (*experiment*): The Wizard-of-Oz simulates an autonomous system by a human response during an experiment, for example to test new technology and its acceptance before it actually exists or to allow for data collection.
- *Arousal*: Physiological/psychological state of being (re-)active.
- *Valence*: Here used to categorise emotion as 'positive' (e.g., joy) or 'negative' (e.g., anger).
- *Non-Linguistic* (*event*): Describes vocal outbursts of non-linguistic character such as laughter or sigh.
- *Pitch*: Perceived frequency of sound (here speech) as opposed to the fundamental frequency—perception can vary according to the intensity, duration, and frequency of the stimuli.
- *Keyword Spotter*: Automatic Speech Recogniser that focuses on the highly robust detection of selected words within a speech or general audio stream.
- *Lexeme*: In linguistics, this roughly subsumes a number of forms (such as flexions) of a single word (such as *speak*, *speaks*, *spoken* as forms of the lexeme SPEAK).

References

1. Ai, H., Litman, D., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A.: Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In: Proc. Interspeech, Pittsburgh, PA, USA, pp. 797–800 (2006)
2. Ang, J., Dhillon, R., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: Proc. Interspeech, Denver, CO, USA, pp. 2037–2040 (2002)
3. Batliner, A., Steidl, S., Nöth, E.: Laryngealizations and emotions: how many babushkas? In: Proc. of the International Workshop on Paralinguistic Speech—Between Models and Data (ParaLing'07), Saarbrücken, Germany, pp. 17–22 (2007)
4. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: The impact of F0 extraction errors on the classification of prominence and emotion. In: Proc. ICPhS, Saarbrücken, Germany, pp. 2201–2204 (2007)
5. Batliner, A., Schuller, B., Schaeffler, S., Steidl, S.: Mothers, adults, children, pets—towards the acoustics of intimacy. In: Proc. ICASSP, Las Vegas, NV, pp. 4497–4500 (2008)
6. Batliner, A., Seppi, D., Steidl, S., Schuller, B.: Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. Adv. Hum.-Comput. Interact. **2010**, 782802 (2010), 15 pages
7. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Amir, N.: Whodunnit—searching for the most important feature types signalling emotional user states in speech. Comput. Speech Lang. **25**, 4–28 (2011)
8. Breese, J., Ball, G.: Modeling emotional state and personality for conversational agents. Technical Report MS-TR-98-41, Microsoft (1998)
9. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of German emotional speech. In: Proc. Interspeech, Lisbon, Portugal, pp. 1517–1520 (2005)
10. Campbell, N., Kashioka, H., Ohara, R.: No laughing matter. In: Proc. Interspeech, Lisbon, Portugal, pp. 465–468 (2005)
11. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: Feeltrace: An instrument for recording perceived emotion in real time. In: Proc. of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, pp. 19–24 (2000)
12. Elliott, C.: The affective reasoner: a process model of emotions in a multi-agent system. PhD thesis, Dissertation, Northwestern University (1992)
13. Engbert, I.S., Hansen, A.V.: Documentation of the Danish emotional speech database DES. Technical report, Center for PersonKommunikation, Aalborg University, Denmark (2007). http://cpk.auc.dk/~tb/speech/Emotions/. Last visited 11/13/2007
14. Enos, F., Shriberg, E., Graciarena, M., Hirschberg, J., Stolcke, A.: Detecting deception using critical segments. In: Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), Antwerp, Belgium, pp. 2281–2284 (2007)
15. Eyben, F., Batliner, A., Schuller, B., Seppi, D., Steidl, S.: Cross-corpus classification of realistic emotions—some pilot experiments. In: Proc. 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Valetta, pp. 77–82 (2010)
16. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE—the Munich versatile and fast open-source audio feature extractor. In: Proc. ACM Multimedia, Florence, Italy, pp. 1459–1462 (2010)
17. Goertzel, B., Silverman, K., Hartley, C., Bugaj, S., Ross, M.: The baby webmind project. In: Proc. of The Annual Conference of The Society for the Study of Artificial Intelligence and the Simulation of Behavior (AISB) (2000)
18. Grimm, M., Kroschel, K., Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database. In: Proc. of the IEEE International Conference on Multimedia and Expo (ICME), Hannover, Germany, pp. 865–868 (2008)

19. Haderlein, T., Nöth, E., Toy, H., Batliner, A., Schuster, M., Eysholdt, U., Hornegger, J., Rosanowski, F.: Automatic evaluation of prosodic features of tracheoesophageal substitute voice. Eur. Arch. Oto-Rhino-Laryngol. **264**(11), 1315–1321 (2007)
20. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11** (2009)
21. Hansen, J.H.L., Bou-Ghazale, S.: Getting started with SUSAS: a Speech Under Simulated and Actual Stress database. In: Proc. EUROSPEECH-97, vol. 4, Rhodes, Greece, pp. 1743–1746 (1997)
22. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) Proc. of ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, pp. 137–142. Springer, Heidelberg (1998)
23. Kao, Y.-H., Lee, L.-S.: Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In: Proc. ICLSP, pp. 1814–1817 (2006)
24. Krajewski, J., Kröger, B.: Using prosodic and spectral characteristics for sleepiness detection. In: Eighth Annual Conf. Int. Speech Communication Association, pp. 1841–1844 (2007)
25. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
26. Litman, D., Forbes, K.: Recognizing emotions from student speech in tutoring dialogues. In: Proc. ASRU, Virgin Island, USA, pp. 25–30 (2003)
27. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. J. Mach. Learn. Res. **2**, 419–444 (2002)
28. Lugger, M., Yang, B.: Psychological motivated multi-stage emotion classification exploiting voice quality features. In: Mihelic, F., Zibert, J. (eds.) Speech Recognition, p. 1. In-Tech, Rijeka (2008)
29. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE'05 audio-visual emotion database. In: IEEE Workshop on Multimedia Database Management (2006)
30. Matos, S., Birring, S.S., Pavord, I.D., Evans, D.H.: Detection of cough signals in continuous audio recordings using hidden Markov models. In: IEEE Trans. Biomedical Engineering, pp. 1078–1108 (2006)
31. Mohammadi, G., Vinciarelli, A., Mortillaro, M.: The voice of personality: mapping nonverbal vocal behavior into trait attributions. In: Proc. SSPW, Firenze, Italy, pp. 17–20 (2010)
32. Mporas, I., Ganchev, T.: Estimation of unknown speakers' height from speech. Int. J. Speech Technol. **12**(4), 149–160 (2009)
33. Pal, P., Iyer, A.N., Yantorno, R.E.: Emotion detection from infant facial expressions and cries. In: Proc. ICASSP, Toulouse, France, pp. 809–812 (2006)
34. Russell, J.A., Bachorowski, J.A., Fernandez-Dols, J.M.: Facial and vocal expressions of emotion. Annual Review of Psychology, 329–349 (2003)
35. Schiel, F., Heinrich, C.: Laying the foundation for in-car alcohol detection by speech. In: Proc. INTERSPEECH 2010, Brighton, UK, pp. 983–986 (2009)
36. Schröder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C., Schuller, B.: Towards responsive sensitive artificial listeners. In: Proc. 4th Intern. Workshop on Human-Computer Conversation, Bellagio, Italy (2008)
37. Schuller, B., Jiménez Villar, R., Rigoll, G., Lang, M.: Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In: Proc. ICASSP, vol. I, Philadelphia, PA, USA, pp. 325–328 (2005)
38. Schuller, B., Müller, R., Lang, M., Rigoll, G.: Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensemble. In: Proc. Interspeech, Lisbon, Portugal, pp. 805–808 (2005)
39. Schuller, B., Wimmer, M., Arsic, D., Rigoll, G., Radig, B.: Audiovisual behaviour modeling by combined feature spaces. In: Proc. ICASSP, Honolulu, HY, pp. 733–736 (2007)
40. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Emotion Recognition from Speech: Putting ASR in the Loop. In: Proc. ICASSP, Taipei, Taiwan, pp. 4585–4588. IEEE Press, New York (2009)

41. Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H.: Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image Vis. Comput. **27**, 1760–1774 (2009). Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior
42. Schuller, B., Schenk, J., Rigoll, G., Knaup, T.: The "godfather" vs. "chaos": comparing linguistic analysis based on online knowledge sources and bags-of-n-grams for movie review valence estimation. In: Proc. International Conference on Document Analysis and Recognition, Barcelona, Spain, pp. 858–862 (2009)
43. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 Emotion Challenge. In: Proc. Interspeech, Brighton, UK, pp. 312–315 (2009)
44. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.: Acoustic emotion recognition: a benchmark comparison of performances. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Merano, pp. 552–557 (2009)
45. Schuller, B., Metze, F., Steidl, S., Batliner, A., Eyben, F., Polzehl, T.: Late fusion of individual engines for improved recognition of negative emotions in speech—learning vs. democratic vote. In: Proc. 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, pp. 5230–5233 (2010)
46. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: The INTERSPEECH 2010 Paralinguistic Challenge. In: Proc. INTERSPEECH 2010, Makuhari, Japan, pp. 2794–2797 (2010)
47. Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.: Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans. Affect. Comput. **1**, 119–132 (2010)
48. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge (2011). Speech Communication—Special Issue on Sensing Emotion and Affect—Facing Realism in Speech Processing
49. Seppi, D., Batliner, A., Schuller, B., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Aharonson, V.: Patterns, prototypes, performance: classifying emotional user states. In: Proc. Interspeech, Brisbane, Australia, pp. 601–604 (2008)
50. Seppi, D., Batliner, A., Steidl, S., Schuller, B., Nöth, E.: Word accent and emotion. In: Proc. Speech Prosody 2010, Chicago, IL (2010)
51. Steidl, S.: Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Logos, Berlin (2009). PhD thesis, FAU Erlangen-Nuremberg
52. Steidl, S., Batliner, A., Seppi, D., Schuller, B.: On the impact of children's emotional speech on acoustic and language models. EURASIP J. Audio Speech Music Process. **2010**, 783954 (2010). 14 pages, doi:10.1155/2010/783954
53. Steininger, S., Schiel, F., Dioubina, O., Raubold, S.: Development of user-state conventions for the multimodal corpus in smartkom. In: Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation, Las Palmas, Spain, pp. 33–37 (2002)
54. Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., Schuller, B.: Deep neural networks for acoustic emotion recognition: raising the benchmarks. In: Proc. ICASSP, Prague, Czech Republic (2011)
55. Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G.: Combining frame and turn-level information for robust recognition of emotions within speech. In: Proc. Interspeech, Antwerp, Belgium, pp. 2249–2252 (2007)
56. Ward, N., Tsukahara, W.: Prosodic features which cue backchannel responses in English and Japanese. J. Pragmat. **32**, 1177–1207 (2000)
57. Weiss, B., Burkhardt, F.: Voice attributes affecting likability perception. In: Proc. INTERSPEECH, Makuhari, Japan (2010)
58. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
59. Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R.: Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies. In: Proc. Interspeech, Brisbane, Australia, pp. 597–600 (2008)

60. Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G.: Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE J. Sel. Top. Signal Process. **4**, 867–881 (2010). Special Issue on "Speech Processing for Natural Interaction with Intelligent Environments"
61. Zhe, X., Boucouvalas, A.C.: Text-to-emotion engine for real time internet communication. In: Proc. of the International Symposium on Communication Systems, Networks, and DSPs, Staffordshire University, pp. 164–168 (2002)

Chapter 10
Continuous Analysis of Affect from Voice and Face

Hatice Gunes, Mihalis A. Nicolaou, and Maja Pantic

10.1 Introduction

Human affective behavior is multimodal, continuous and complex. Despite major advances within the affective computing research field, modeling, analyzing, interpreting and responding to human affective behavior still remains a challenge for automated systems as affect and emotions are complex constructs, with fuzzy boundaries and with substantial individual differences in expression and experience [7]. Therefore, affective and behavioral computing researchers have recently invested increased effort in exploring how to best model, analyze and interpret the subtlety, complexity and continuity (represented along a continuum e.g., from -1 to $+1$) of affective behavior in terms of latent dimensions (e.g., arousal, power and valence) and appraisals, rather than in terms of a small number of discrete emotion categories (e.g., happiness and sadness). This chapter aims to (i) give a brief overview of the existing efforts and the major accomplishments in modeling and analysis of emotional expressions in dimensional and continuous space while focusing on open issues and new challenges in the field, and (ii) introduce a representative approach for

H. Gunes (✉)
Queen Mary University of London, London, UK
e-mail: haticeg@ieee.org

M.A. Nicolaou · M. Pantic
Imperial College, London, UK

M.A. Nicolaou
e-mail: mihalis@imperial.ac.uk

M. Pantic
e-mail: m.pantic@imperial.ac.uk

M. Pantic
University of Twente, Twente, The Netherlands

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_10,

multimodal continuous analysis of affect from voice and face, and provide experimental results using the audiovisual Sensitive Artificial Listener (SAL) Database of natural interactions. The chapter concludes by posing a number of questions that highlight the significant issues in the field, and by extracting potential answers to these questions from the relevant literature.

The chapter is organized as follows. Section 10.2 describes theories of emotion, Sect. 10.3 provides details on the affect dimensions employed in the literature as well as how emotions are perceived from visual, audio and physiological modalities. Section 10.4 summarizes how current technology has been developed, in terms of data acquisition and annotation, and automatic analysis of affect in continuous space by bringing forth a number of issues that need to be taken into account when applying a dimensional approach to emotion recognition, namely, determining the duration of emotions for automatic analysis, modeling the intensity of emotions, determining the baseline, dealing with high inter-subject expression variation, defining optimal strategies for fusion of multiple cues and modalities, and identifying appropriate machine learning techniques and evaluation measures. Section 10.5 presents our representative system that fuses vocal and facial expression cues for dimensional and continuous prediction of emotions in valence and arousal space by employing the bidirectional Long Short-Term Memory neural networks (BLSTM-NN), and introduces an output-associative fusion framework that incorporates correlations between the emotion dimensions to further improve continuous affect prediction. Section 10.6 concludes the chapter.

10.2 Affect in Dimensional Space

Emotions and affect are researched in various scientific disciplines such as neuroscience, psychology, and cognitive sciences. Development of automatic affect analyzers depends significantly on the progress in the aforementioned sciences. Hence, we start our analysis by exploring the background in emotion theory, perception and recognition.

According to research in psychology, three major approaches to affect modeling can be distinguished [31]: categorical, dimensional, and appraisal-based approach. The categorical approach claims that there exist a small number of emotions that are basic, hard-wired in our brain, and recognized universally (e.g. [18]). This theory on universality and interpretation of affective nonverbal expressions in terms of basic emotion categories has been the most commonly adopted approach in research on automatic measurement of human affect. However, a number of researchers have shown that in everyday interactions people exhibit non-basic, subtle and rather complex affective states like thinking, embarrassment or depression. Such subtle and complex affective states can be expressed via dozens of anatomically possible facial and bodily expressions, audio or physiological signals. Therefore, a single label (or any small number of discrete classes) may not reflect the complexity of the affective state conveyed by such rich sources of information [82]. Hence, a number of researchers advocate the use of dimensional description of human affect, where

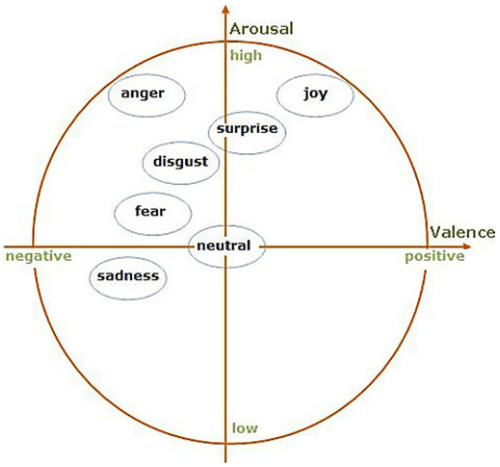


Fig. 10.1 Russell's valence-arousal space. The figure is by courtesy of [77]

affective states are not independent from one another; rather, they are related to one another in a systematic manner (see, e.g., [31, 82, 86]). It is not surprising, therefore, that automatic affect sensing and recognition researchers have recently started exploring how to model, analyze and interpret the subtlety, complexity and continuity (represented along a continuum from -1 to $+1$, without discretization) of affective behavior in terms of latent dimensions, rather than in terms of a small number of discrete emotion categories.

The most widely used dimensional model is a circular configuration called *Circumplex of Affect* (see Fig. 10.1) introduced by Russell [82]. This model is based on the hypothesis that each basic emotion represents a bipolar entity being a part of the same emotional continuum. The proposed poles are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant), as illustrated in Fig. 10.1. Another well-accepted and commonly used dimensional description is the 3D emotional space of pleasure—displeasure, arousal—nonarousal and dominance—submissiveness [63], at times referred to as the *PAD emotion space* [48] or as *emotional primitives* [19].

Scherer and colleagues introduced another set of psychological models, referred to as componential models of emotion, which are based on the appraisal theory [25, 31, 86]. In the appraisal-based approach emotions are generated through continuous, recursive subjective evaluation of both our own internal state and the state of the outside world (relevant concerns/needs) [25, 27, 31, 86]. Despite pioneering efforts of Scherer and colleagues (e.g., [84]), how to use the appraisal-based approach for automatic measurement of affect is an open research question as this approach requires complex, multicomponential and sophisticated measurements of change. One possibility is to reduce the appraisal models to dimensional models (e.g., 2D space of arousal-valence).

Ortony and colleagues proposed a computationally tractable model of the cognitive basis of emotion elicitation, known as OCC [71]. OCC is now established as a standard (cognitive appraisal) model for emotions, and has mostly been used in affect synthesis (in embodied conversational agent design, e.g. [4]).

Each approach, categorical or dimensional, has its advantages and disadvantages. In the categorical approach, where each affective display is classified into a single category, complex mental states, affective state or blended emotions may be too difficult to handle [108]. Instead, in dimensional approach, observers can indicate their impression of each stimulus on several continuous scales. Despite exhibiting such advantages, dimensional approach has received a number of criticisms. Firstly, the usefulness of these approaches has been challenged by discrete emotions theorists, such as Silvan Tomkins, Paul Ekman, and Carroll Izard, who argued that the reduction of emotion space to two or three dimensions is extreme and resulting in loss of information. Secondly, while some basic emotions proposed by Ekman, such as happiness or sadness, seem to fit well in the dimensional space, some basic emotions become indistinguishable (e.g., fear and anger), and some emotions may lie outside the space (e.g., surprise). It also remains unclear how to determine the position of other affect-related states such as confusion. Note, however, that arousal and valence are not claimed to be the only dimensions or to be sufficient to differentiate equally between all emotions. Nonetheless, they have already proven to be useful in several domains (e.g., affective content analysis [107]).

10.3 Affect Dimensions and Signals

An individual's inner emotional state may become apparent by subjective experiences (how the person feels), internal/inward expressions (bio signals), and external/outward expressions (audio/visual signals). However, these may be incongruent, depending on the context (e.g., feeling angry and not expressing it outwardly).

The contemporary theories of emotion and affect consider appraisal as the most significant component when defining and studying emotional experiences [81], and at the same time acknowledge that emotion is not just appraisal but a complex multifaceted experience that consists of the following stages (in order of occurrence):

1. *Cognitive Appraisal*. Only events that have significance for our goals, concerns, values, needs, or well-being elicit emotion.
2. *Subjective feelings*. The appraisal is accompanied by feelings that are good or bad, pleasant or unpleasant, calm or aroused.
3. *Physiological arousal*. Emotions are accompanied by autonomic nervous system activity.
4. *Expressive behaviors*. Emotions are communicated through facial and bodily expressions, postural and voice changes.
5. *Action tendencies*. Emotions carry behavioral intentions, and the readiness to act in certain ways.

This multifaceted aspect of affect poses a true challenge to automatic sensing and analysis. Therefore, to be able to deal with these challenges, affect research scientists have ended up making a number of assumptions and simplifications while studying emotions [7, 72]. These assumptions can be listed as follows.

1. *Emotions are on or off at any particular point in time*. This assumption has implications on most data annotation procedures where raters label a user's expressed emotion as one of the basic emotion categories or a specific point in a dimensional space. The main issue with this assumption is that the boundaries for defining the expressed emotion as on or off are usually not clear.
2. *Emotion is a state that the subject does not try to actively change or alleviate*. This is a common assumption during the data acquisition process where the subjects are assumed to have a simple response to the provided stimulus (e.g., while watching a clip or interacting with an interface). However, such simple passive responses do not usually hold during daily human–computer interactions. People generally regulate their affective states caused by various interactions (e.g., an office user logging into Facebook to alleviate his boredom).
3. *Emotion is not affected by situation or context*. This assumption pertains to most of the past research work on automatic affect recognition where emotions have been mostly investigated in laboratory settings, outside of a social context. However, some emotional expressions are displayed only during certain context (e.g., pain).

Affect research scientists have made the following simplifications while studying emotions [7, 72]:

1. *Emotions do occur in asynchronous communication* (e.g., via a prerecorded video/sound from a sender to a receiver). This simplification does not hold in reality as human nonverbal expressive communication occurs mostly face-to-face.
2. *Interpersonal emotions do arise from communications with strangers* (e.g., laboratory studies where people end up communicating with people they do not know). This simplification is unrealistic as people tend to be less expressive with people they do not know on an interpersonal level. Therefore, an automatic system designed using such communicative settings is expected to be much less sensitive to its user's realistic expressions.

Overall, these assumptions and simplifications are far from reality. However, they have paved the initial but crucial way for automatic affect recognizers that attempt to analyze both the felt (e.g., [9, 10, 59]) and the internally or the externally expressed (e.g., [50, 54]) emotions.

10.3.1 Affect Dimensions

Despite the existence of various emotion models described in Sect. 10.2, in automatic measurement of dimensional and continuous affect, valence (how positive or negative the affect is), activation (how excited or apathetic the affect is), power (the sense of control over the affect), and expectation (the degree of anticipating or being taken unaware) appear to make up the four most important affect dimensions [25]. Although ideally the intensity dimension could be derived from the other dimensions, to guarantee a complete description of affective coloring, some researchers

include intensity (how far a person is away from a state of pure, cool rationality) as the fifth dimension (e.g., [62]). Solidarity, antagonism and agreement have also been in the list of dimensions investigated [13]. Overall, search for optimal low-dimensional representation of affect remains open [25].

10.3.2 Visual Signals

Facial actions (e.g., pulling eyebrows up) and facial expressions (e.g., producing a smile), and to a much lesser extent bodily postures (e.g., head bent backwards and arms raised forwards and upwards) and expressions (e.g., head nod), form the widely known and used visual signals for automatic affect measurement. Dimensional models are considered important in this task as a single label may not reflect the complexity of the affective state conveyed by a facial expression, body posture or gesture. Ekman and Friesen [17] considered expressing discrete emotion categories via face, and communicating dimensions of affect via body as more plausible.

A number of researchers have investigated how to map various visual signals onto emotion dimensions. For instance, Russell [82] mapped the facial expressions to various positions on the two-dimensional plane of arousal-valence, while Cowie et al. [13] investigated the emotional and communicative significance of head nods and shakes in terms of arousal and valence dimensions, together with dimensional representation of solidarity, antagonism and agreement.

Although in a stricter sense not seen as part of the visual modality, motion capture systems have also been utilized for recording the relationship between body posture and affect dimensions (e.g., [57, 58]). For instance, Kleinsmith et al. [58] identified that scaling, arousal, valence, and action tendency were the affective dimensions used by human observers when discriminating between postures. They also reported that low-level posture features such as orientation (e.g., orientation of shoulder axis) and distance (e.g., distance between left elbow and right shoulder) appear to help in effectively discriminating between the affective dimensions [57, 58].

10.3.3 Audio Signals

Audio signals convey affective information through explicit (linguistic) messages, and implicit (acoustic and prosodic) messages that reflect the way the words are spoken. There exist a number of works focusing on how to map audio expression to dimensional models. Cowie et al. used valence-activation space (similar to valence-arousal) to model and assess affect from speech [11, 12]. Scherer and colleagues have also proposed how to judge emotional effects on vocal expression, using the appraisal-based theory [31].

In terms of affect recognition from audio signals the most reliable finding is that pitch appears to be an index into arousal [7]. Another well-accepted finding is

that mean of the fundamental frequency (F0), mean intensity, speech rate, as well as pitch range [46], "blaring" timbre [14] and high-frequency energy [85] are positively correlated with the arousal dimension. Shorter pauses and inter-breath stretches are indicative of higher activation [99].

There is relatively less evidence on the relationship between certain acoustic parameters and other affect dimensions such as valence and power. Vowel duration and power dimension in general, and lower F0 and high power in particular, appear to have correlations. Positive valence seems to correspond to a faster speaking rate, less high-frequency energy, low pitch and large pitch range [85] and longer vowel durations. A detailed literature summary on these can be found in [87] and [88].

10.3.4 Bio Signals

The bio signals used for automatic measurement of affect are galvanic skin response that increases linearly with a person's level of arousal [9], electromyography (frequency of muscle tension) that is correlated with negatively valenced emotions [41], heart rate that increases with negatively valenced emotions such as fear, heart rate variability that indicates a state of relaxation or mental stress, and respiration rate (how deep and fast the breath is) that becomes irregular with more aroused emotions like anger or fear [9, 41].

Measurements recorded over various parts of the brain including the amygdala also enable observation of the emotions felt [79]. For instance, approach or withdrawal response to a stimulus is known to be linked to the activation of the left or right frontal cortex, respectively.

A number of studies also suggest that there exists a correlation between increased blood perfusion in the orbital muscles and stress levels for human beings. This periorbital perfusion can be quantified through the processing of thermal video (e.g., [102]).

10.4 Overview of the Current Technology

This section provides a brief summary of the current technology by describing how affective data are acquired and annotated, and how affect analysis in continuous space is achieved.

10.4.1 Data Acquisition and Annotation

Cameras are used for acquisition of face and bodily expressions, microphones are used for recording audio signals, and thermal (infrared) cameras are used for recording blood flow and changes in skin temperature. 3D affective body postures or

gestures can alternatively be recorded by utilizing motion capture systems (e.g., [57, 58]). In such scenarios, the actor is dressed in a suit with a number of markers on the joints and body segments, while each gesture is captured by a number of cameras and represented by consecutive frames describing the position of the markers in the 3D space. This is illustrated in Fig. 10.2 (second and third rows).

In the bio signal research context, the subject being recorded usually wears a headband or a cap on which electrodes are mounted, a clip sensor, or touch type electrodes (see Fig. 10.2, last row). The subject is then stimulated with emotionally-evocative images or sounds. Acquiring affect data without subjects' knowledge is strongly discouraged and the current trend is to record spontaneous data in more constrained conditions such as an interview (e.g., [10]) or interaction (e.g., [62]) setting, where subjects are still aware of placement of the sensors and their locations.

Annotation of the affect data is usually done separately for each modality, assuming independency between the modalities. A major challenge is the fact that there is no coding scheme that is agreed upon and used by all researchers in the field that can accommodate all possible communicative cues and modalities. In general, the Feeltrace annotation tool is used for annotating the external expressions (audio and visual signals) with continuous traces (impressions) in the dimensional space. Feeltrace allows coders to watch the audiovisual recordings and move their cursor, within the 2-dimensional emotion space (valence and arousal) confined to $[-1, +1]$, to rate their impression about the emotional state of the subject [11] (see the illustration in Fig. 10.3(a)). For annotating the internal expressions (bio signals), the level of valence and arousal is usually extracted from subjective experiences (subjects' own responses) (e.g., [59, 79]) due to the fact that feelings, induced by an image or sound, can be very different from subject to subject. The Self Assessment Mannequin (SAM) [60], illustrated in Fig. 10.3(b), is the most widely used means for self assessment.

When discretized dimensional annotation is adopted (as opposed to continuous one), researchers seem to use different intensity levels: either a ten-point Likert scale (e.g., 0-low arousal, 9-high arousal) or a range between -1.0 and 1.0 (divided into a number of levels) [37]. The final annotation is usually calculated as the mean of the observers' ratings. However, whether this is the best way of obtaining ground-truth labels of emotional data is still being discussed. Overall, individual coders may vary in their appraisal of what is happening in the scene, in their judgment of the emotional behavior of the target individual, in their understanding of the terms 'positive emotion' and 'negative emotion' and in their movement of the computer mouse to translate their rating into a point on the onscreen scale. Furthermore, recent findings in dynamic emotional behavior coding indicate that the temporal pattern of ratings appears similar across cultures but that there exist significant differences in the intensity levels at which participants from different cultural backgrounds rate the emotional behaviors [96]. Therefore, how to obtain and use rich emotional data annotations, from multiple and multi-cultural raters, needs serious consideration.

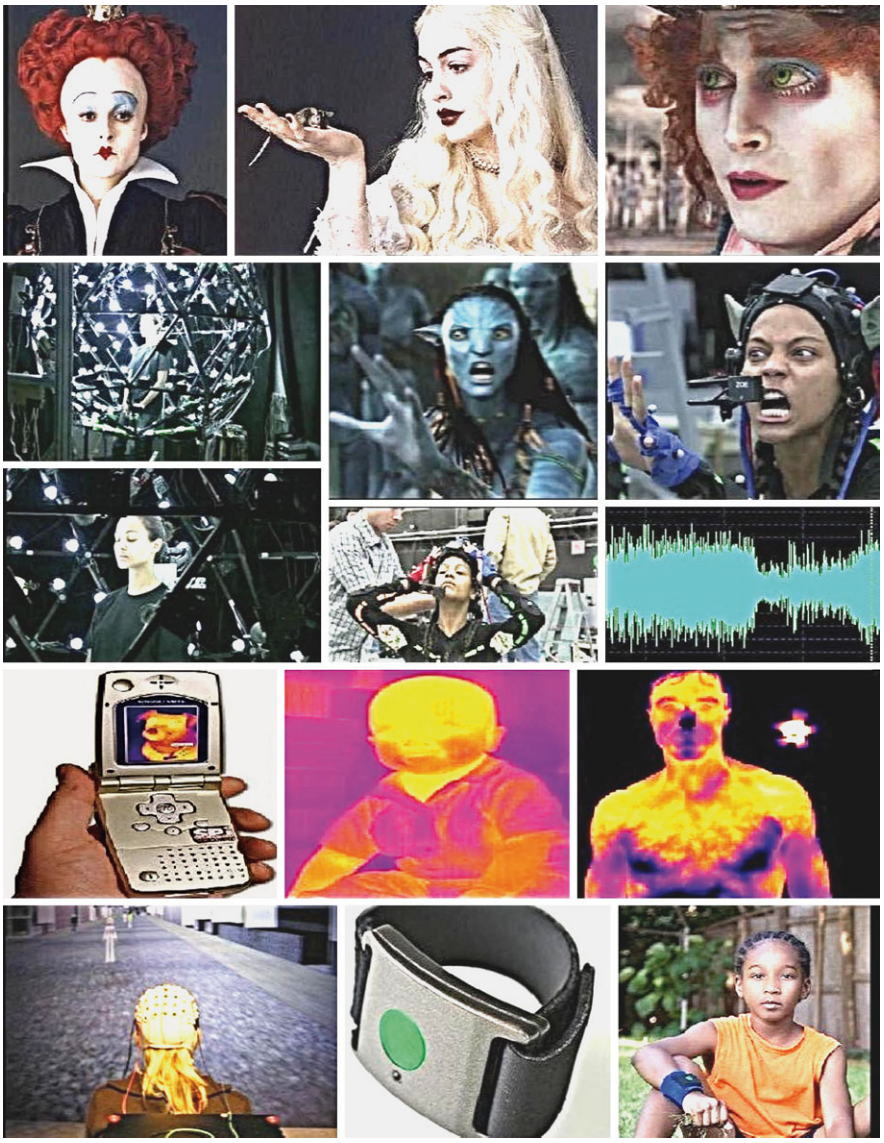


Fig. 10.2 Examples of sensors used in multimodal affective data acquisition: (*1st row*) camera for visible imagery (face and body), (*2nd & 3rd rows*) facial and body motion capture, and audio signals (used for animation and rendering), (*4th row*) infrared camera for thermal imagery, and (*5th row*) various means for recording bio signals (brain signals, heart and respiration rate, etc.)

10.4.2 Automatic Dimensional Affect Prediction and Recognition

After affect data have been acquired and annotated, representative and relevant features need to be extracted prior to the automatic measurement of affect in dimen-

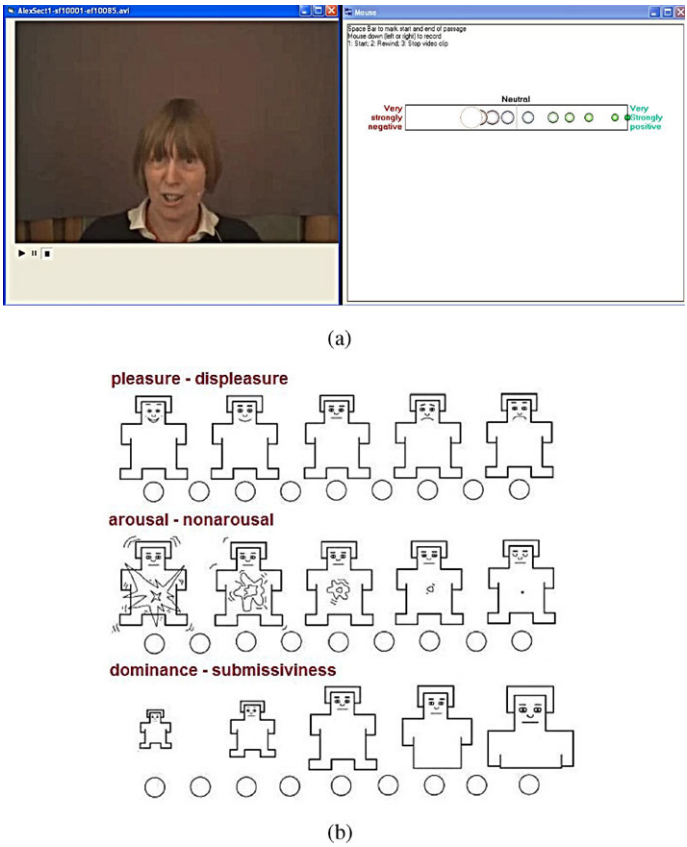


Fig. 10.3 Illustration of (**a**) the Feeltrace annotation tool [11], and (**b**) the Self Assessment Mannequin (SAM) [60]

sional and continuous space. The feature extraction techniques used for each communicative source are similar to the previous works (reviewed in [40]) adopting a categorical approach to affect recognition.

In dimensional affect analysis emotions are represented along a continuum. Considering this, systems that target automatic dimensional affect measurement should be able to predict the emotions continuously. However, most of the automatic recognition systems tend to simplify the problem by quantizing the continuous labels into a finite number of discrete levels. Hence, the most commonly employed strategy in automatic dimensional affect prediction is to reduce the continuous prediction problem to a two-class recognition problem (positive vs. negative or active vs. passive classification; e.g., [66, 92]) or a four-class recognition problem (classification into the quadrants of 2D V-A space; e.g., [8, 26, 29, 47, 106]).

For example, Kleinsmith and Bianchi-Berthouze discriminate between high–low, high–neutral and low–neutral affective dimensions [57], while Wöllmer et al. quantize the V-A dimensions of the SAL database into either 4 or 7 levels, and then

use Conditional Random Fields (CRFs) to predict the quantized labels [105]. Attempts for discriminating between more coarse categories, such as positive vs. negative [66], and active vs. passive [8] have also been attempted. Of these, Caridakis et al. [8] uses the SAL database, combining auditive and visual modalities. Nicolaou et al. focus on audiovisual classification of spontaneous affect into negative or positive emotion categories using facial expression, shoulder and audio cues, and utilizing 2- and 3-chain coupled Hidden Markov Models and likelihood space classification to fuse multiple cues and modalities [66]. Kanluan et al. combine audio and visual cues for affect recognition in V-A space by fusing facial expression and audio cues, using Support Vector Machines for Regression (SVR) and late fusion with a weighted linear combination [50]. The labels used have been discretized on a 5-point scale in the range of $[-1, +1]$ for each emotion dimension. The work presented in [106] utilizes a hierarchical dynamic Bayesian network combined with BLSTM-NN performing regression and quantizing the results into four quadrants (after training).

As far as actual continuous dimensional affect prediction (without quantization) is concerned, there exist a number of methods that deal exclusively with speech (i.e., [33, 105, 106]). The work by Wöllmer et al. uses the SAL Database and Long Short-Term Memory neural networks and Support Vector Machines for Regression (SVR) [105]. Grimm and Kroschel use the Vera am Mittag database [35] and SVRs, and compare their performance to that of the distance-based fuzzy k-Nearest Neighbor and rule-based fuzzy-logic estimators [33]. The work by Espinosa et al. also use the Vera am Mittag database [35] and examine the importance of different groups of speech acoustic features in the estimation of continuous PAD dimensions [19].

Currently, there are also a number of works focusing on dimensional and continuous prediction of emotions from the visual modality [39, 56, 69]. The work by Gunes and Pantic focuses on dimensional prediction of emotions from spontaneous conversational head gestures by mapping the amount and direction of head motion, and occurrences of head nods and shakes into arousal, expectation, intensity, power and valence level of the observed subject using SVRs [39]. Kipp and Martin in [56] investigated (without performing automatic prediction) how basic gestural form features (e.g., preference for using left/right hand, hand shape, palm orientation, etc.) are related to the single PAD dimensions of emotion. The work by Nicolaou et al. focuses on dimensional and continuous prediction of emotions from naturalistic facial expressions within an Output-Associative Relevance Vector Machine (RVM) regression framework by learning non-linear input and output dependencies inherent in the affective data [69].

More recent works focus on dimensional and continuous prediction of emotions from multiple modalities. For instance, Eyben et al. [21] propose a string-based approach for fusing the behavioral events from visual and auditive modalities (i.e., facial action units, head nods and shakes, and verbal and nonverbal audio cues) to predict human affect in a continuous dimensional space (in terms of arousal, expectation, intensity, power and valence dimensions). Although automatic affect analyzers based on physiology end up using multiple signal sources, explicit fusion of multimodal data for continuous modeling of affect utilizing dimensional models of emotion is still relatively unexplored. For instance, Khalili and Moradi propose

multimodal fusion of brain and peripheral signals for automatic recognition of three emotion categories (positively excited, negatively excited and calm) [52]. Their results show that, for the task at hand, EEG signals seem to perform better than other physiological signals, and nonlinear features lead to better understanding of the felt emotions. Another representative approach is that of Gilroy et al. [28] that propose a dimensional multimodal fusion scheme based on the power-arousal-PAD space to support detection and integration of spontaneous affective behavior of users (in terms of audio, video and attention events) experiencing arts and entertainment. Unlike many other multimodal approaches (e.g., [8, 50, 66]), the ground truth in this work is obtained by measuring Galvanic Skin Response (GSR) as an independent measure of arousal.

For further details on the aforementioned systems, as well as on systems that deal with dimensional affect recognition from a single modality or cue, the reader is referred to [37, 38, 109].

10.4.3 Challenges and Prospects

The summary provided in the previous section reflects that automatic dimensional affect recognition is still in its pioneering stage [34, 37, 38, 91, 105]. There are a number of challenges which need to be taken into account when applying a dimensional approach to affect prediction and advancing the current state of the art.

The interpretation accuracy of expressions and physiological responses in terms of continuous emotions is very challenging. While visual signals appear to be better for interpreting valence, audio signals seem to be better for interpreting arousal [33, 68, 100, 105]. A thorough comparison between all modalities would indeed provide a better understanding of which emotion dimensions are better predicted from which modalities (or cues).

Achieving inter-observer agreement is one of the most challenging issues in dimension-based affect modeling and analysis. To date, researchers have mostly chosen to use self-assessments (subjective experiences, e.g. [41]) or the mean (within a predefined range of values) of the observers' ratings (e.g. [57]). Although it is difficult to self-assess arousal, it has been reported that using classes generated from self-assessment of emotions facilitate greater accuracy in recognition (e.g., [9]). This finding results from a study on automatic analysis of physiological signals in terms of A-V emotion space. It remains unclear whether the same holds independently of the utilized modalities and cues. Modeling inter-observer agreement levels within automatic affect analyzers and finding which signals better correlate with self assessment and which ones better correlate with independent observer assessment remain unexplored.

The window size to be used to achieve optimal affect prediction is another issue that the existing literature does not provide a unique answer to. Current affect analyzers employ various window sizes depending on the modality, e.g., 2–6 seconds for speech, 3–15 seconds for bio signals [54]. For instance, when measuring

affect from heart rate signals, analysis should not be done on epochs of less than a minute [6]. A time window of 50 s appears to be also necessary to accurately monitor mental stress in realistic settings [83]. There is no consensus on how the efficiency of such a choice should be evaluated. On one hand achieving real-time affect prediction requires a small window size to be used for analysis (i.e., a few seconds, e.g. [10]), while on the other hand obtaining a reliable prediction accuracy requires long(er)-term monitoring [6, 83]. For instance, Chanel et al. [10] conducted short-term analysis of emotions (i.e., time segments of 8 s) in valence and arousal space using EEG and peripheral signals in a self-induction paradigm. They reported large differences in accuracy between the EEG and peripheral features which may be due to the fact that the 8 s length of trials may be too short for a complete activation of peripheral signals while it may be sufficient for EEG signals.

Measuring the intensity of expressed emotion appears to be modality dependent. The way the intensity of an emotion is apparent from physiological data may be different from the way it is apparent from visual data. Moreover, little attention has been paid so far to whether there are definite boundaries along the affect continuum to distinguish between various levels or intensities. Currently intensity is measured by quantizing the affect dimensions into arbitrary number of levels such as neutral, low and high (e.g., [57, 59, 105]). Separate models are then built to discriminate between pairs of affective dimension levels, for instance, low vs. high, low vs. neutral, etc. Generalizing intensity analysis across different subjects is a challenge yet to be researched as different subjects express different levels of emotions in the same situation. Moreover, recent research findings indicate that there also exist significant differences in the intensity levels at which coders from different cultural backgrounds rate emotional behaviors [96].

The Baseline problem is another major challenge in the field. For physiological signals (bio signals) this refers to the problem of finding a condition against which changes in measured physiological signals can be compared (a state of calmness) [65]. For the audio modality this is usually achieved by segmenting the recordings into turns using energy based voice activity detection and processing each turn separately (e.g., [105]). For visual modality the aim is to find a frame in which the subject is expressionless and against which changes in subject's motion, pose, and appearance can be compared. This is achieved by manually segmenting the recordings, or by constraining the recordings to have the first frame containing a neutral expression (see, e.g., [66, 67, 75]). Yet, as pointed out by Levenson in [61], emotion is rarely superimposed upon a prior state of *rest*; instead, emotion occurs most typically when the organism is in some prior activation. Hence, enforcing existence of expressionless state in each recording or manually segmenting recordings so that each segment contains a baseline expression are strong, unrealistic constrains. This remains a great challenge in automatic analysis, which typically relies on existence of a baseline for analysis and processing of affective information.

Generalization capability of automatic affect analyzers across subjects is still a challenge in the field. Kulic and Croft [59] reported that for bio signal based affect measurement, subjects seem to vary not only in terms of response amplitude and duration, but for some modalities, a number of subjects show no response at all.

This makes generalization over unseen subjects a very difficult problem. A common way of measuring affect from bio signals is doing it for each participant separately (without computing baseline), e.g. [10]. When it comes to other modalities, most of the works in the field report mainly on subject-dependent dimensional affect measurement and recognition due to limited number of subjects and limited amount of data (e.g., [39, 68, 69, 105]).

Modality fusion refers to combining and integrating all incoming unimodal events into a single representation of the affect expressed by the user. When it comes to integrating multiple modalities, the major issues are: (i) when to integrate the modalities (at what abstraction level to do the fusion), (ii) how to integrate the modalities (which criteria to use), (iii) how to deal with the increased number of features due to fusion, (iv) how to deal with the asynchrony between the modalities (e.g., if video is recorded at 25 Hz, audio is recorded at 48 kHz while EEG is recorded at 256–512 Hz), and (v) how to proceed with fusion when there is conflicting information conveyed by the modalities. Typically, multimodal data fusion is either done at the feature level (in a maximum likelihood estimation manner) or at the decision level (when most of the joint statistical properties may have been lost). Feature-level fusion is obtained by concatenating all the features from multiple cues into one feature vector which is then fed into a machine learning technique. In the decision-level data fusion, the input coming from each modality/cue is modeled independently, and these single-cue and single-modality based recognition results are combined in the end. Since humans display multi-cue and multimodal expressions in a complementary and redundant manner, the assumption of conditional independence between modalities and cues in decision-level fusion can result in loss of information (i.e. mutual correlation between the modalities). Therefore, model-level fusion has been proposed as an alternative approach for fusing multimodal affect data (e.g., [75]). Despite such efforts in the discrete affect recognition field (reviewed in [40, 109]), these issues remain yet to be explored for dimensional and continuous affect prediction.

Machine learning techniques used for dimensional and continuous affect measurement should be able to produce continuous values for the target dimensions. Overall, there is no agreement on how to model dimensional affect space (continuous vs. quantized) and which machine learning technique is better suited for automatic, multimodal, continuous affect analysis using a dimensional representation. Recognition of quantized dimensional labels is obtained via classification while continuous prediction is achieved by regression. Conditional Random Fields (CRF) and Support Vector Machines (SVM) have mostly been used for quantized dimensional affect recognition tasks (e.g., [105]). Some of the schemes that have been explored for the task of prediction are Support Vector Machines for Regression (SVR) (e.g., [39]) and Long Short-Term Memory Recurrent Networks (LSTM-RNN). The design of emotion-specific classification schemes that can handle multimodal and spontaneous data is one of the most important issues in the field. In accordance with this, Kim and Andre propose a novel scheme of emotion-specific multilevel dichotomous classification (EMDC) using the property of the dichotomous categorization in the 2D emotion model and the fact that arousal classification

yields a higher correct classification ratio than valence classification (or direct multiclass classification) [55]. They apply this scheme on classification of four emotions (positive/high arousal, negative/high arousal, negative/low arousal and positive/low arousal) from physiological signals recorded while subjects were listening to music. How to create such emotion-specific schemes for dimensional and continuous prediction of emotions from other modalities and cues should be investigated further.

Evaluation measures applicable to categorical affect recognition are not directly applicable to dimensional approaches. Using the Mean Squared Error (MSE) between the predicted and the actual values of arousal and valence, instead of the recognition rate (i.e., percentage of correctly classified instances) is the most commonly used measure by related work in the literature (e.g., [50, 105]). However, using MSE might not be the best way to evaluate the performance of dimensional approaches to automatic affect measurement and prediction. Therefore, the correlation coefficient that evaluates whether the model has managed to capture patterns inhibited in the data at hand is also employed by several studies (e.g., [50, 67]) together with MSE. Overall, however, how to obtain optimal evaluation metrics for continuous and dimensional emotion prediction remains an open research issue [37]. Generally speaking, the performance of an automatic analyzer can be modeled and evaluated in an *intrinsic* and an *extrinsic* manner (as proposed for face recognition in [103]). The intrinsic performance and its evaluation depend on the intrinsic components such as the dataset chosen for the experiments and the machine learning algorithms (and their parameters) utilized for prediction. The extrinsic performance and evaluation instead depend on the extrinsic factors such as (temporal/spatial) resolution of the multimodal data and recording conditions (e.g., illumination, occlusions, noise, etc.). Future research in continuous affect prediction should analyze the relevance and prospects of the aforementioned performance components, and how they could be applied to continuous prediction of affect.

10.4.4 Applications

Various applications have been using the dimensional (both quantized and continuous) representation and prediction of emotions, ranging from human–computer (e.g., Sensitive Talking Heads [45], Sensitive Artificial Listeners [89, 90], spatial attention analysis [95], arts installations [104]) and human–robot interaction (e.g., humanoid robotics [5, 51]), clinical and biomedical studies (e.g., stress/pain monitoring [36, 64, 101], autism-related assistive technology), learning and driving environments (e.g., episodic learning [22], affect analysis in the car [20]), multimedia (e.g, video content representation and retrieval [53, 98] and personalized affective video retrieval [97]), and entertainment technology (e.g., gaming [80]). These indicate that affective computing has matured enough to have a presence and measurable impact in our lives. There are also spin off companies emerging out of collaborative research at well-known universities (e.g., Affectiva [1] established by R. Picard and colleagues of MIT Media Lab).

10.5 A Representative System: Continuous Analysis of Affect from Voice and Face

The review provided in the previous sections indicates that currently there is a shift toward subtle, continuous, and context-specific interpretations of affective displays recorded in naturalistic settings, and toward multimodal analysis and recognition of human affect. Converging with this shift, in this section we present a representative approach that: (i) fuses facial expression and audio cues for dimensional and continuous prediction of emotions in valence and arousal space, (ii) employs the bidirectional Long Short-Term Memory neural networks (BLSTM-NNs) for the prediction task, and (iii) introduces an output-associative fusion framework that incorporates correlations between the emotion dimensions to further improve continuous prediction of affect.

The section starts with the description of the naturalistic database used in the experimental studies. Next, data pre-processing, audio and facial feature extraction and tracking procedures, as well as the affect prediction process are explained.

10.5.1 Dataset

We use the Sensitive Artificial Listener Database (SAL-DB) [16] that contains spontaneous data collected with the aim of capturing the audiovisual interaction between a human and an operator undertaking the role of a SAL character (e.g., an avatar). The SAL characters intend to engage the user in a conversation by paying attention to the user's emotions and nonverbal expressions. Each character has its own emotionally defined personality: Poppy is happy, Obadiah is gloomy, Spike is angry, and Prudence is pragmatic. During an interaction, the characters attempt to create an emotional workout for the user by drawing her/him toward their dominant emotion, through a combination of verbal and nonverbal expressions.

The SAL database contains audiovisual sequences recorded at a video rate of 25 fps (352×288 pixels) and at an audio rate of 16 kHz. The recordings were made in a lab setting, using one camera, a uniform background and constant lighting conditions. The SAL data have been annotated manually. Although there are approximately 10 hours of footage available in the SAL database, V-A annotations have only been obtained for two female and two male subjects. We used this portion for our experiments.

10.5.2 Data Pre-processing and Segmentation

The data pre-processing and segmentation stage consists of (i) determining ground truth by maximizing inter-coder agreement, (ii) detecting frames that capture the transition *to* and *from* an emotional state, and (iii) automatic segmentation of spontaneous audiovisual data. We provide a brief summary of these in the following sections. For a detailed description of these procedures the reader is referred to [67].

10.5.2.1 Annotation Pre-processing

The SAL data have been annotated by a set of coders who provided continuous annotations with respect to valence and arousal dimensions using the Feeltrace annotation tool [11], as explained in Sect. 10.4.1. Feeltrace allows coders to watch the audiovisual recordings and move their cursor, within the 2-dimensional emotion space (valence and arousal) confined to $[-1, +1]$, to rate their impression about the emotional state of the subject.

Annotation pre-processing involves dealing with the issue of missing values (interpolation), grouping the annotations that correspond to one video frame together (binning), determining normalization procedures (normalization) and extracting statistics from the data in order to obtain segments with a baseline and high inter-coder agreement (statistics and metrics).

Interpolation In order to deal with the issue of missing values, similar to other works reporting on data annotated in continuous dimensional spaces (e.g., [105]), we interpolated the actual annotations at hand. We used piecewise cubic interpolation as it preserves the monotonicity and the shape of the data.

Binning Binning refers to grouping and storing the annotations together. As a first step the measurements of each coder c are binned separately. Since we aim at segmenting video files, we generate bins which are equivalent to one video frame f. This is equivalent to a bin of 0.04 seconds (SAL-DB was recorded at a rate of 25 frames/s). The fields with no annotation are assigned a 'not a number' (NaN) identifier.

Normalization The A-V measurements for each coder are not in total agreement, mostly due to the variance in human coders' perception and interpretation of emotional expressions. Thus, in order to deem the annotations comparable, we need to normalize the data. We experimented with various normalization techniques. After extracting the videos and inspecting the superimposed ground-truth plots, we opted for local normalization (normalizing each coder file for each session). This helps us avoid propagating noise in cases where one of the coders is in large disagreement with the rest (where a coder has a very low correlation with respect to the rest of the coders). Locally normalizing to zero mean produces the smallest mean squared error (MSE) both for valence (0.046) and arousal (0.0551) dimensions.

Statistics and Metrics We extract two useful statistics from the annotations: correlation and agreement. We start the analysis by constructing vectors of pairs of coders that correspond to each video session, e.g., when we have a video session where four coders have provided annotations, this gives rise to six pairs. For each of these pairs we extract the correlation coefficient between the valence (val) values of each pair, as well as the level of agreement in emotion classification in terms of positive or negative. We define the agreement metric by

$$AGR = \frac{\sum_{f=0}^{n} e(c_i(f).val, c_j(f).val)}{|frames|}, \tag{10.1}$$

where $c_i(f).val$ stands for the valence value annotated by coder c_i at frame f. Function e is defined as

$$e(i, j) = \begin{cases} 1 & \text{if } (sign(i) = sign(j)), \\ 0 & \text{else.} \end{cases}$$

In these calculations we do not consider the NaN values to avoid negatively affecting the results. After these metrics are calculated for each pair, each coder is assigned the average of the results of all pairs that the coder has participated in. We choose the Pearson's Correlation (COR) as the metric to be used in the automatic segmentation process as it appears to be stricter than agreement (AGR) providing better comparison amongst the coders.

10.5.2.2 Automatic Segmentation

The segmentation stage consists of producing negative and positive audiovisual segments with a temporal window that contains an offset before and after (i.e., the baseline) the displayed expression. For instance, for capturing negative emotional states, if we assume that the transition *from* non-negative *to* negative emotional state occurs at time t (in seconds), we would have a window of $[t - 1, t, t', t' + 1]$ where t' seconds is when the emotional state of the subject turns to non-negative again. The procedure is completely analogous for positive emotional states.

Detecting and Matching Crossovers For an input coder c, the crossing over from one emotional state to the other is detected by examining the valence values and identifying the points where the sign changes. Here a modified version of the sign function is used, it returns 1 for values that are higher than 0 (a value of 0 valence is never encountered in the annotations), -1 for values that are less than zero, and 0 for NaN values. We accumulate all crossover points for each coder, and return the set of crossovers *to-a-positive* and *to-a-negative* emotional state. The set of crossovers is then used for matching crossovers across coders. For instance, if a session has annotations from four coders, the frame (f) where each coder detects the crossover is not the same for all coders (for the session in question). Thus, we have to allow an offset for the matching process. This procedure searches the crossovers detected by the coders and then accepts the matches where there is less than the predefined offset (time) difference between the detections. When a match is found, we remove the matched crossovers and continue with the rest. The existence of different combinations of crossovers which may match using the predefined offset poses an issue. By examining the available datasets, we decided to maximize the number of coders participating in a matched crossover set rather than minimizing the temporal distances between the participating coders. The motivations for this decision are as follows: (i) if more coders agree on the crossover, the reliability of the ground truth produced will be higher, and (ii) the offset amongst the resulting matches is on average quite small (<0.5 s) when considering only the number of participating coders. We disregard cases where only one coder detects a crossover due to lack of agreement between coders.

Segmentation Driven by Matched Crossovers In order to illustrate how the crossover frame decision (for each member of the set) is made, let us assume that for *to-a-negative* transition a coder detects a crossover at frame 2, while the other coder detects a crossover at frame 4. If the frames are averaged to the nearest integer, then we can assume that the crossover happens at frame 3. In this case we have only 2 coders agreeing, we use the *correlation* metric in order to weight their decision and determine the crossover point. This provides a measurement of the relative importance of the annotations for each coder and propagates information from the other two coders not participating in the match. In order to capture 0.5 s before the transition window, the number of frames corresponding to the predefined offset are subtracted from the *start frame*. The ground-truth values for valence are retrieved by incrementing the initial frame number where each crossover was detected by the coders. Again, following the previous example, this means that we consider frame 2 of coder 1 and frame 4 of coder 2 to provide ground-truth values for frame 3 (the average of 2 and 4). This gives us an averaged valence value. Then, the frame 4 valence value (ground truth) would be the combination of frame 3 of coder 1 and frame 5 of coder 2. The procedure of determining combined average values continues until the valence value crosses again to a *non-negative* valence value. The endpoint of the audiovisual segment is then set to the frame including the offset after crossing back to a *non-negative* valence value. The ground truth of the audiovisual segment consists of the arousal and valence (A-V) values calculated.

Typically, an automatically produced segment or clip consists of a single interaction of the subject with the avatar (operator), starting with the final seconds of the avatar speaking, continuing with the subject responding (and thus reacting and expressing an emotional state audiovisually) and concluding where the avatar starts responding.

10.5.3 Feature Extraction

Our audio features include Mel-frequency Cepstrum Coefficients (MFCC) [49] and prosody features (the energy of the signal, the Root Mean Squared Energy and the pitch obtained by using a Praat pitch estimator [74]). Mel-frequency Cepstrum (MFC) is a representation of the spectrum of an audio sample which is mapped onto the nonlinear mel-scale of frequency to better approximate the human auditory system's response. The MFCC coefficients collectively make up the MFC for the specific audio segment. We used six cepstrum coefficients, thus obtaining six MFCC and six MFCC-Delta features for each audio frame. We have essentially used the typical set of features used for automatic affect recognition (e.g., [75]). Along with pitch, energy and RMS energy, we obtained a set of features with dimensionality $d = 15$ per audio frame. Note that we used a 0.04 second window with a 50% overlap (i.e. first frame 0–0.04, second from 0.02–0.06 and so on) in order to obtain a double frame rate for audio (50 Hz) compared to that of video (25 fps). This is an effective and straightforward way to synchronise the audio and video streams (similarly to [75]).

Fig. 10.4 Examples of the data at hand from the SAL database along with the extracted 20 points, used as features for the facial expression cues

To capture the facial motion displayed during a spontaneous expression we track 20 facial feature points (FFP), as illustrated in Fig. 10.4. These points are the corners of the eyebrows (4 points), eyes (8 points), nose (3 points), the mouth (4 points) and the chin (1 point). To track these facial points we used the Patras–Pantic particle filtering tracking scheme [73]. For each video segment containing n frames, we obtain a set of n vectors containing 2D coordinates of the 20 points tracked in n frames ($Tr_f = \{Tr_{f1} \ldots Tr_{f20}\}$ with dimensions $n * 20 * 2$).

10.5.4 Dimensional Affect Prediction

This section describes how dimensional affect prediction from voice and face is achieved using the Bidirectional Long Short-Term Memory Neural Networks (BLSTM-NN). It first focuses on single-cue prediction from voice or face, and then introduces the model-level and output-associative fusion using the BLSTM-NNs.

10.5.4.1 Bidirectional Long Short-Term Memory Neural Networks

The traditional Recurrent Neural Networks (RNN) are unable to learn temporal dependencies longer than a few time steps due to the vanishing gradient problem [42, 43]. LSTM Neural Networks (LSTM-NNs) were introduced by Graves and Schmidhuber [32] in order to overcome this issue. The LSTM structure introduces recurrently connected memory blocks instead of traditional neural network nodes

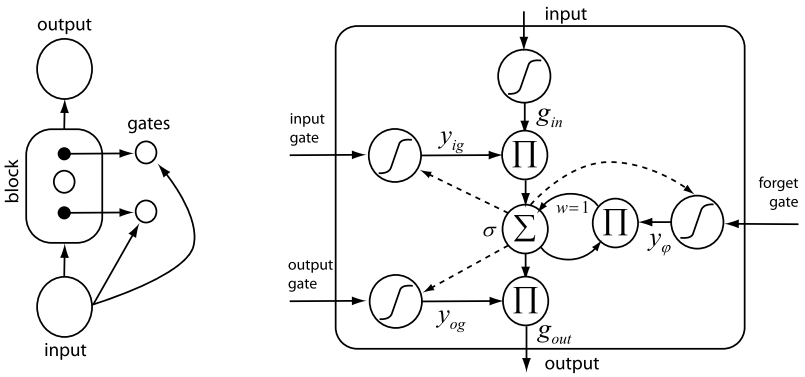


Fig. 10.5 Illustration of (**a**) the simplest LSTM network, with a single input, a single output, and a single memory block in place of the hidden unit, and (**b**) a typical implementation of an LSTM block, with multiplication units (Π), an addition unit (Σ) maintaining the cell state and typically non-linear squashing function units

(Fig. 10.5(a)). Each memory block contains memory cells and a set of multiplicative gates. In its simplest form, a memory block contains one memory cell.

As can be seen from Fig. 10.5(b), there are three types of gates: the input, output and forget gates. These gates are estimated during the training phase of an LSTM-NN.

The input, output and forget gates can be thought of as providing write, read and reset access to what is called a cell state (σ), which represents temporal network information. This can be seen from examining the state updates at time t:

$$\sigma(t) = y_{\phi}(t)\sigma(t-1) + y_{ig}(t)g_{in}(t).$$

The next state $\sigma(t)$ is defined as the sum of the forget gate at time t ($y_{\phi}(t)$) multiplied by the previous state, $\sigma(t-1)$ and the squashed input to the cell $g_{in}(t)$ multiplied by the input gate $y_{ig}(t)$. Thus, the forget gate can reset the state of the network, i.e. when $y_{\phi} \approx 0$ then the next state does not depend on the previous one:

$$\sigma(t) \approx y_{ig}(t)g_{in}(t).$$

This is similar when the input gate is near zero. Then, the next state depends only on the previous state and the forget gate value. The output of the cell is the cell state, as regulated by the value of the output gate (Fig. 10.5(b)). This configuration enforces constant error flow and overcomes the vanishing gradient problem.

In addition, traditional RNNs process input in a temporal order, thus learning input patterns by relating only to past context. Bidirectional RNNs (BRNNs) [3, 94] instead modify the learning procedure to overcome the latter issue of the past and future context: they present each of the training sequences in a forward and a backward order (to two different recurrent networks, respectively, which are connected to a common output layer). In this way, the BRNN is aware of both future and

past events in relation to the current timestep. The concept is directly expanded for LSTMs, referred to as Bidirectional Long Short-Term Memory neural networks (BLSTM-NN). BLSTM-NN have been shown to outperform unidirectional LSTM-NN for speech processing (e.g., [32]) and have been used for many learning tasks. They have been successfully applied to continuous emotion prediction from speech (e.g., [105, 106]) proving that modeling the sequential inputs and long range temporal dependencies appear to be beneficial for the task of automatic emotion prediction.

10.5.4.2 Single-Cue Prediction

The first step in continuous affect prediction task consists of prediction based on single cues. Let $\mathscr{D} = \{V, A\}$ represent the set of emotion dimensions, $\mathscr{C}$ the set of cues consisting of the facial expressions, shoulder movement and audio cues. Given a set of input features $\mathbf{x_c} = [\mathbf{x_{1_c}}, \ldots, \mathbf{x_{n_c}}]$ where n is the training sequence length and $c \in \mathscr{C}$, we train a machine learning technique f_d, in order to predict the relevant dimension output, $\mathbf{y_d} = [y_1, \ldots, y_n]$, $d \in \mathscr{D}$.

$$f_d : \mathbf{x} \mapsto y_d. \tag{10.2}$$

This step provides us with a set of predictions for each machine learning technique, and each relevant dimension employed.

10.5.4.3 Model-Level Fusion

As already explained in Sect. 10.4.2, since humans display multi-cue and multimodal expressions in a complementary and redundant manner, the assumption of conditional independence between modalities and cues in decision-level fusion can result in loss of information (i.e. mutual correlation between the modalities). Therefore, we opt for model-level fusion of the continuous predictions as this has the potential of capturing correlations and structures embedded in the continuous output of the predictors/regressors (from different sets of cues). This is illustrated in Fig. 10.6(a).

More specifically, during model-level fusion, a function learns to map predictions to a dimension d from the set of cues as follows:

$$f_{mlf} : f_d(\mathbf{x_1}) \times \cdots \times f_d(\mathbf{x_m}) \mapsto y_d, \tag{10.3}$$

where m is the total number of fused cues.

10.5.4.4 Output-Associative Fusion

In the previous section, we have treated the prediction of valence or arousal as a 1D regression problem. However, psychological evidence shows that valence and

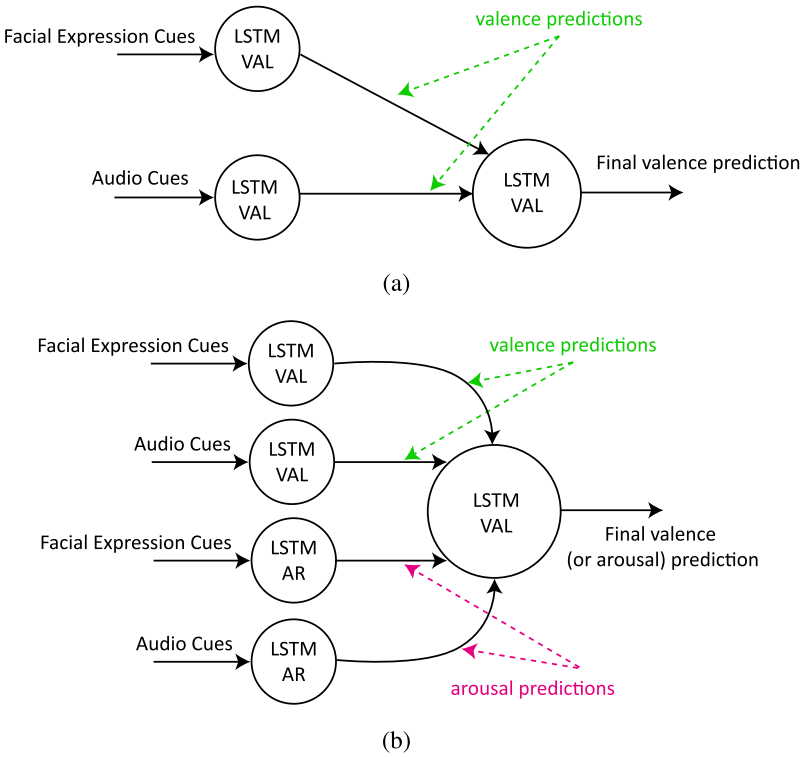


Fig. 10.6 Illustration of (**a**) model-level fusion and (**b**) output-associative fusion using facial expression and audio cues. Model-level fusion combines valence predictions from facial expression and audio cues by using a third network for the final valence prediction. Output-associative fusion combines both valence and arousal values predicted from facial expression and audio cues, again by using a third network, which outputs the final prediction.

arousal dimensions are correlated [2, 70, 107]. In order to exploit these correlations and patterns, we propose a framework capable of learning the dependencies that exist amongst the predicted dimensional values.

Given the setting described in Sect. 10.5.4.2, this framework learns to map the outputs of the intermediate predictors (each BLSTM-NN as defined in (10.2)) onto a higher (and final) level of prediction by incorporating cross-dimensional (output) dependencies (see Fig. 10.6(b)). This method, which we call *output-associative fusion*, can be represented by a function f_{oaf}:

$$f_{oaf} : f_{Ar}(\mathbf{x_1}) \times f_{Val}(\mathbf{x_1}) \times \cdots \times f_{Ar}(\mathbf{x_m}) \times f_{Val}(\mathbf{x_m}) \mapsto y_d . \tag{10.4}$$

As a result, the final output, taking advantage of the temporal and bidirectional characteristics of the regressors (BLSTM-NNs), depends not only on the entire sequence of input features $\mathbf{x_i}$ but also on the entire sequence of intermediate output predictions $\mathbf{f_d}$ of both dimensions (see Fig. 10.6(b)).

Table 10.1 Single-cue prediction results for valence and arousal dimensions

Dimension	Modality	RMSE	COR	SAGR
Arousal	Voice	0.240	0.586	0.764
	Face	0.250	0.493	0.681
Valence	Voice	0.220	0.444	0.648
	Face	0.170	0.712	0.841

10.5.5 Experiments and Analysis

10.5.5.1 Experimental Setup

Prior to experimentation, all features have been normalized to the range of $[-1, +1]$, except for the audio features which have been found to perform better with z-normalization (i.e., normalizing to mean $= 0$ and standard deviation $= 1$).

As the main evaluation metrics we choose to use the root mean squared error (RMSE) that evaluates the root of the prediction by taking into account the squared error of the prediction from the ground truth, the correlation (COR) that provides an evaluation of the linear relationship between the prediction and the ground truth, and subsequently, an evaluation of whether the model has managed to capture linear structural patterns inhibited in the data at hand, and the sign agreement metric (SAGR) that measures the agreement level of the prediction with the ground truth by assessing the valence dimension as being positive $(+)$ or negative $(-)$, and the arousal dimension as being active $(+)$ or passive $(-)$.

For validation purposes we use a subset of the SAL-DB that consists of 134 audiovisual segments (a total of 30,042 video frames) obtained by the automatic segmentation procedure (proposed in [67]). As V-A annotations have only been provided for two female and two male subjects, for our experiments we employ *subject-dependent leave-one-sequence-out cross-validation*. More specifically, the evaluation consists of 134 folds where at each fold one sequence is left out for testing and the other 133 sequences are used for training. The prediction results are then averaged over 134 folds.

The parameter optimization for BLSTM-NNs refers to mainly determining the topology of the network along with the number of epochs, momentum and learning rate.

10.5.5.2 Results and Analysis

Single-cue results are presented in Table 10.1, while results obtained from fusion are presented in Table 10.2.

We initiate our analysis with the single-cue results (Table 10.1) and the valence dimension. Various automatic dimensional emotion prediction and recognition studies have shown that arousal can be much better predicted than valence using audio

Table 10.2 Results for output-associative fusion (AOF) and model-level fusion (MLF). The best results are obtained by employing output-associative fusion (shown in bold)

Dimension	OAF			MLF		
	RMSE	COR	SAGR	RMSE	COR	SAGR
Arousal	**0.220**	**0.628**	0.800	0.230	0.605	0.800
Coders	0.145	0.870	0.840	0.145	0.870	0.840
Valence	**0.160**	**0.760**	**0.892**	0.170	0.748	0.856
Coders	0.141	0.850	0.860	0.141	0.850	0.860

cues (e.g., [33, 68, 100, 105]). Our experimental results also support these findings indicating that the visual cues appear more informative for predicting the valence dimension. The facial expression cues provide a higher correlation with the ground truth (COR = 0.71) compared to the audio cues (COR = 0.44). This fact is also confirmed by the RMSE and SAGR metrics. The facial expression cues also provide higher SAGR (0.84), indicating that the predictor was accurate in predicting an emotional state as positive or negative for 84% of the frames. For prediction of the arousal dimension the audio cues appear to be superior to the visual cues. More specifically, audio cues provide COR = 0.59, whereas the facial expression cues provide COR = 0.49.

Fusing facial and audio cues using model-level fusion outperforms the single-cue prediction results. Model-level fusion appears to be much better for predicting the valence dimension rather than the arousal dimension. This is mainly due to the fact that the single-cue predictors for valence dimension perform better, thus containing more correct temporal dependencies and structural characteristics (while the weaker arousal predictors contain fewer of these dependencies). Model-level fusion also reconfirms that visual cues are more informative for valence dimension than the audio cues. Finally, the newly proposed output-associative fusion provides the best results, outperforming both single-cue analysis and model-level fusion results. We denote that the performance increase of output-associative fusion is higher for the arousal dimension (compared to the valence dimension). This could be justified by the fact that the single-cue predictors for valence perform better than for arousal (Table 10.1) and thus, more correct valence patterns are passed onto the output-associative fusion framework. An example of the output-associative valence and arousal prediction from face and audio is shown in Fig. 10.7.

Based on the experimental results provided in Tables 10.1–10.2, we conclude the following.

- Facial expression cues are better suited to the task of continuous valence prediction compared to audio cues. For arousal dimension, instead, the audio cues appear to perform better. This is in accordance with the previous findings in the literature.
- The inherent temporal and structured nature of continuous affective data appears to be highly suitable for predictors that can model temporal dependencies and relate temporally distant events. To evaluate the performance of such frameworks,

the use of not only the RMSE but also the correlation coefficient appears to be very important. Furthermore, the use of other emotion-specific metrics, such as the SAGR (used in this work), is also desirable as they contain valuable information regarding emotion-specific aspects of the predictions.
- As confirmed by the psychological theory, valence and arousal are correlated. Such correlations appear to exist in our data where fusing predictions from both valence and arousal dimensions (output-associative fusion) improves the results compared to using predictions from either valence or arousal dimension alone (as in the model-level fusion case).
- In general, audiovisual data appear to be more useful for predicting valence than for predicting arousal. While arousal is better predicted by using audio features alone, valence is better predicted by using audiovisual data.

Overall, our output-associative fusion framework (i) achieves RMSE $= 0.160$, COR ≈ 0.760 and SAGR ≈ 0.900 for the valence dimension, compared to the human coder (inter-coder) RMSE ≈ 0.141, COR ≈ 0.850, and SAGR ≈ 0.860, and (ii) provides RMSE $= 0.220$, COR ≈ 0.628 and SAGR ≈ 0.800 for the arousal dimension, compared to the human coder (inter-coder) RMSE ≈ 0.145, COR ≈ 0.870 and SAGR ≈ 0.840.

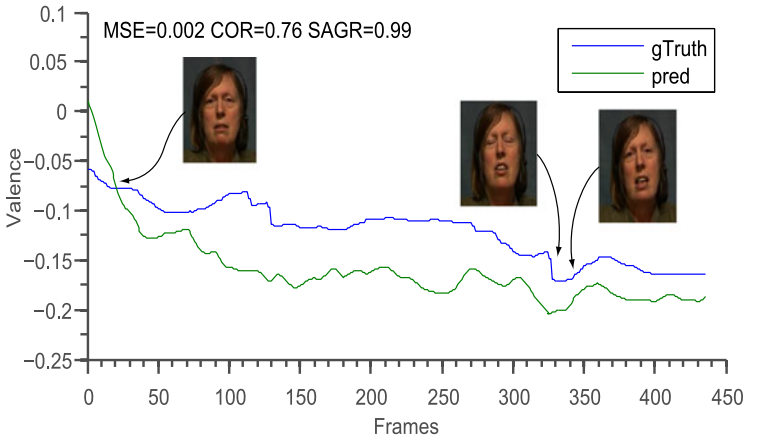
In our experiments we employed a subject-dependent leave-one-sequence-out cross-validation procedure due to the small number of annotated data available. As spontaneous expressions appear to have somewhat person-dependent characteristics, subject-independent experimentation is likely to be more challenging and affect our prediction results.

10.6 Concluding Remarks

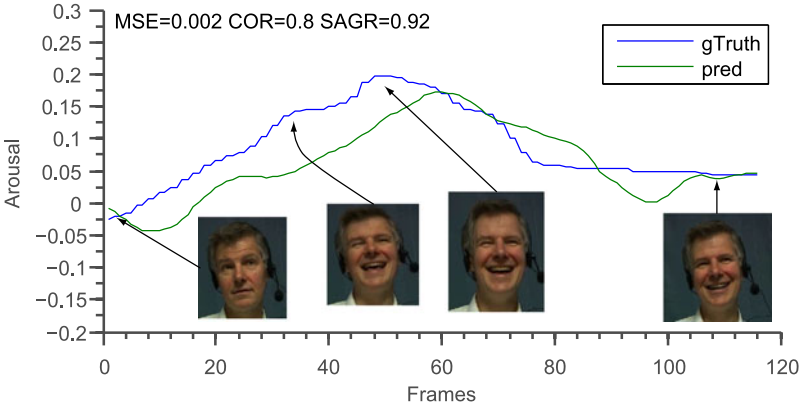
The review provided in this chapter suggests that the automatic affect sensing field has slowly started shifting from categorical (and discrete) affect recognition to dimensional (and continuous) affect prediction to be able to capture the complexity of affect expressed in naturalistic settings. There is a growing research interest driven by various advances and demands (e.g., real-time representation and analysis of naturalistic and continuous human affective behavior for emotion-related disorders like autism), and funded by various research projects (e.g., European Union FP 7, SEMAINE[1]). To date, despite the existence of a number of dimensional emotion models, the two-dimensional model of arousal and valence appears to be the most widely used model in automatic measurement of affect from audio, visual and bio signals.

The current automatic measurement technology has already started dealing with spontaneous data obtained in less-controlled environments using various sensing devices, and exploring a number of machine learning techniques and evaluation measures. However, naturalistic settings pose many challenges to continuous affect

[1] http://www.semaine-project.eu



(a)



(b)

Fig. 10.7 Valence and arousal ground truth (gTruth) compared to predictions (pred) from output-associative fusion of facial expressions and audio cues

sensing and prediction (e.g., when subjects are not restricted in terms of mobility, the level of noise in all recorded signals tends to increase), as well as affect synthesis and generation. As a consequence, a number of issues that should be addressed in order to advance the field remain unclear. These have been summarized and discussed in this chapter.

As summarized in Sect. 10.4.2 and reviewed in [37], to date, only a few systems have actually achieved dimensional affect prediction from multiple modalities. Overall, existing systems use different training/testing datasets (which differ in the way affect is elicited and annotated), they differ in the underlying affect model (i.e., target affect categories), as well as in the employed modality or combination of

modalities, and the applied evaluation method. As a consequence, it remains unclear which recognition and prediction method is suitable for dimensional affect prediction from which modalities and cues. These challenges should be addressed in order to advance the field while identifying the importance, as well as the feasibility, of the following issues:

1. *Among the available remotely observable and remotely unobservable modalities, which ones should be used for automatic dimensional affect prediction? Should we investigate the innate priority among the modalities to be preferred for each affect dimension? Does this depend on the context (who the subject is, where she is, what her current task is, and when the observed behavior has been shown)?*

   Continuous long-term monitoring of bio signals (e.g., autonomic nervous system) appears to be particularly useful and usable for health care applications (e.g., stress and pain monitoring, autism-related assistive technology). Using bio signals for automatic measurement is especially important for applications where people do not easily express themselves outwardly with facial and bodily expressions (e.g., people with autism spectrum disorders) [24]. As stated before, various automatic dimensional emotion prediction and recognition studies have shown that arousal can be much better predicted than valence using audio cues (e.g., [33, 68, 100, 105]). For the valence dimension instead, visual cues (e.g., facial expressions and shoulder movements) appear to perform better [68]. Whether such conclusions hold for different contexts and different data remains to be evaluated. Another significant research finding is that when multiple modalities are available during data annotation, both speed and accuracy of judgments increase when the modalities are expressing the same emotion [15]. How such findings should be incorporated into automatic dimensional affect predictors remains to be researched further.
2. *When labeling emotions, which signals better correlate with self assessment and which ones correlate with independent observer assessment?*

   When acquiring and annotating emotional data, there exist individual differences in emotional response, as well as individual differences in the use of rating scales. We have mentioned some of these differences before, in Sect. 10.4.1. Research also shows that affective state labeling is significantly affected by factors such as familiarity of the person and context of the interaction [44]. Even if the emotive patterns to be labeled are fairly similar, human perception is biased by context and prior experience. Moreover, Feldman presented evidence that when individuals are shown emotional stimulus, they differ in their attention to valence and arousal dimensions [23]. We have also mentioned cross-cultural intensity differences in labeling emotional behaviors [96]. If such issues are ignored and the ratings provided by the human annotators are simply averaged, the measure obtained may be useful in certain experimental contexts but it will be insensitive to individual variations in subjective experience. More specifically, this will imply having a scale that assumes that individual differences are unimportant or nonexistent. An implication of this view is that for an ideal representation of a subject's affective state, labeling schemes and rating scales should be clearly defined (e.g., by making the subjective distances between adjacent numbers on every portion

of the scale equal) and contextualized (e.g., holding the environmental cues constant), both self assessment and external observer assessment (preferably from observers who are familiar with the user to be assessed) should be obtained and used, and culture-related issues should be taken into consideration.

3. *How does the baseline problem affect prediction? Is an objective basis (e.g., a frame with an expressionless display) strictly needed prior to computing the dimensional affect values? If so, how can this be obtained in a fully automatic manner from naturalistic data?*

   Determining the baseline in naturalistic affective displays is challenging even for human observers. This is particularly the case for the visual modality which constitutes of varying head pose and head gestures (like nods and shakes), speech-related facial actions, and blended facial expressions. The implications for automatic analysis can initially be addressed by training predictors that predict baseline (or neutrality) for each cue and modality separately.

4. *How should intensity be modeled for dimensional and continuous affect prediction? Should the aim be personalizing systems for each subject, or creating systems that are expected to generalize across subjects?*

   Modeling the intensity of emotions should be based on the task-dependent environment and target user group. A common way of measuring affect from bio signals is doing it for each participant separately (without computing baseline), e.g., [10]. Similarly to the recent works on automatic affect prediction from the audio or the visual cues (e.g., [69]), better insight may be obtained by comparing subject-dependent vs. subject-independent prediction results. Customizing the automatic predictors to specific user needs is usually desired and advantageous.

5. *In a continuous affect space, how should duration of affect be defined? How can this be incorporated in automated systems? Will focusing on shorter or longer observations affect the accuracy of the measurement process?*

   Similarly to modeling the emotional intensity level, determining the affect duration should be done based on the task-dependent environment and target user group. Focusing on shorter or longer durations appears to have an effect on the prediction accuracy. Achieving real-time affect prediction requires a small window size to be used for analysis (i.e., a few seconds, e.g., [10]), while on the other hand obtaining a reliable prediction accuracy requires long(er)-term monitoring [6, 83]. Therefore, analysis duration should be determined as a trade-off between reliable prediction accuracy and real-time requirements of the automatic system.

Finding comprehensive and thorough answers to the questions posed above, and fully exploring the terrain of the dimensional and continuous affect prediction, depends on all relevant research fields (engineering, computer science, psychology, neuroscience, and cognitive sciences) stepping out of their labs, working side-by-side together on real-life applications, and sharing the experience and the insight acquired on the way, to make affect research tangible for realistic settings and lay people [76]. Pioneering projects representing such inter-disciplinary efforts have already started emerging, ranging, for instance, from publishing compiled books of related work (e.g., [30]) and organizing emotion recognition challenges (e.g., INTERSPEECH 2010 Paralinguistic Challenge featuring the affect sub-challenge

with a focus on dimensional affect [93]) to projects as varied as affective human-embodied conversational agent interaction (e.g., European Union FP 7 SEMAINE [89, 90]), and affect sensing for autism (e.g., [76, 78]).

10.7 Summary

Human affective behavior is multimodal, continuous and complex. Despite major advances within the affective computing research field, modeling, analyzing, interpreting and responding to human affective behavior still remains a challenge for automated systems as affect and emotions are complex constructs, with fuzzy boundaries and with substantial individual differences in expression and experience [7]. Therefore, affective and behavioral computing researchers have recently invested increased effort in exploring how to best model, analyze and interpret the subtlety, complexity and continuity (represented along a continuum e.g., from -1 to $+1$) of affective behavior in terms of latent dimensions (e.g., arousal, power and valence) and appraisals, rather than in terms of a small number of discrete emotion categories (e.g., happiness and sadness). This chapter aimed to (i) give a brief overview of the existing efforts and the major accomplishments in modeling and analysis of emotional expressions in dimensional and continuous space while focusing on open issues and new challenges in the field, and (ii) introduce a representative approach for multimodal continuous analysis of affect from voice and face, and provide experimental results using the audiovisual Sensitive Artificial Listener (SAL) Database of natural interactions. The chapter concluded by posing a number of questions that highlight the significant issues in the field, and by extracting potential answers to these questions from the relevant literature.

10.8 Questions

1. What are the major approaches used for affect modeling and representation? How do they differ from each other?
2. Why has the dimensional affect representation gained interest?
3. What are the dimensions used for representing emotions?
4. Affect research scientists usually make a number of assumptions and simplifications while studying emotions. What are these assumptions and simplifications? What implications do they have?
5. How is human affect sensed and measured? What are the signals measured for analyzing human affect?
6. How are affective data acquired and annotated?
7. What is the current state of the art in automatic affect prediction and recognition?
8. What are the challenges faced in automatic dimensional affect recognition?
9. List a number of applications that use the dimensional representation of emotions.

10. What features are extracted to represent an audio-visual affective sequence? How are the audio and video streams synchronized?
11. What is a Bidirectional Long Short-Term Memory Neural Network? How does it differ from a traditional Recurrent Neural Network?
12. What is meant by the statement 'valence and arousal dimensions are correlated'? What implications does this have on automatic affect prediction?
13. What is output-associative fusion? How does it compare to model-level fusion?
14. How are the root mean squared error, correlation, and sign agreement used for evaluating the automatic prediction of emotions?

10.9 Glossary

- *Categorical description of affect*: Hypothesizes that there exist a small number of emotion categories (i.e., anger, disgust, fear, happiness, sadness and surprise) that are basic, hard-wired in our brain, and recognized universally (e.g. [18]).
- *Dimensional description of affect*: Hypothesizes that affective states are not independent from one another; rather, they are related to one another in a systematic manner.
- *Circumplex Model of Affect*: A circular configuration introduced by Russell [82], based on the hypothesis that each basic emotion represents a bipolar entity being a part of the same emotional continuum.
- *PAD emotion space*: The three dimensional description of emotion in terms of pleasure–displeasure, arousal–nonarousal and dominance–submissiveness [63].
- *Dimensional and continuous affect prediction*: Analyzing and inferring the subtlety, complexity and continuity of affective behavior in terms of latent dimensions (e.g., valence and arousal) by representing it along a continuum (e.g., from -1 to $+1$) without discretization.
- *Long Short-Term Memory neural network*: A Bidirectional Recurrent Neural Network that consists of recurrently connected memory blocks, and uses input, output and forget gates to represent and learn the temporal information and dependencies.
- *Output-associative fusion*: A fusion approach that uses multi-layered prediction, i.e. the initial features extracted from each modality are used for intermediate (output) prediction, and these are further used for a higher (and final) level of prediction (by incorporating cross-dimensional dependencies).

Acknowledgements This work has been funded by EU [FP7/2007-2013] Grant agreement No. 211486 (SEMAINE) and the ERC Starting Grant agreement No. ERC-2007-StG-203143 (MAHNOB).

References

1. Affectiva's homepage: http://www.affectiva.com/ (2011)

2. Alvarado, N.: Arousal and valence in the direct scaling of emotional response to film clips. Motiv. Emot. **21**, 323–348 (1997)
3. Baldi, P., Brunak, S., Frasconi, P., Pollastri, G., Soda, G.: Exploiting the past and the future in protein secondary structure prediction. Bioinformatics **15**, 937–946 (1999)
4. Bartneck, C.: Integrating the occ model of emotions in embodied characters. In: Proc. of the Workshop on Virtual Conversational Characters, pp. 39–48 (2002)
5. Beck, A., Canamero, L., Bard, K.A.: Towards an affect space for robots to display emotional body language. In: Proc. IEEE Int. Symp. in Robot and Human Interactive Communication, pp. 464–469 (2010)
6. Berntson, G.G., Bigger, J.T., Eckberg, D.L., Grossman, P., Kaufmann, P.G., Malik, M., Nagaraja, H.N., Porges, S.W., Saul, J.P., Stone, P.H., van der Molen, M.W.: Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology **34**(6), 623 (1997)
7. Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010)
8. Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., Karpouzis, K.: Modeling naturalistic affective states via facial and vocal expressions recognition. In: Proc. of ACM Int. Conf. on Multimodal Interfaces, pp. 146–154 (2006)
9. Chanel, G., Ansari-Asl, K., Pun, T.: Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In: Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics, pp. 2662–2667, October 2007
10. Chanel, G., Kierkels, J.J.M., Soleymani, M., Pun, T.: Short-term emotion assessment in a recall paradigm. Int. J. Hum.-Comput. Stud. **67**(8), 607–627 (2009)
11. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schroder, M.: Feeltrace: An instrument for recording perceived emotion in real time. In: Proc. of ISCA Workshop on Speech and Emotion, pp. 19–24 (2000)
12. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human–computer interaction. IEEE Signal Process. Mag. **18**, 33–80 (2001)
13. Cowie, R., Gunes, H., McKeown, G., Vaclau-Schneider, L., Armstrong, J., Douglas-Cowie, E.: The emotional and communicative significance of head nods and shakes in a naturalistic database. In: Proc. of LREC Int. Workshop on Emotion, pp. 42–46 (2010)
14. Davitz, J.: Auditory correlates of vocal expression of emotional feeling. In: The Communication of Emotional Meaning, pp. 101–112. McGraw-Hill, New York (1964)
15. de Gelder, B., Vroomen, J.: The perception of emotions by ear and by eye. Cogn. Emot. **23**, 289–311 (2000)
16. Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, L., McRorie, M., Martin, L. Jean-Claude, Devillers, J.-C., Abrilian, A., Batliner, S., Noam, A., Karpouzis, K.: The HUMAINE database: addressing the needs of the affective computing community. In: Proc. of the Second Int. Conf. on Affective Computing and Intelligent Interaction, pp. 488–500 (2007)
17. Ekman, P., Friesen, W.V.: Head and body cues in the judgment of emotion: A reformulation. Percept. Mot. Skills **24**, 711–724 (1967)
18. Ekman, P., Friesen, W.V.: Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. Prentice Hall, New Jersey (1975)
19. Espinosa, H.P., Garcia, C.A.R., Pineda, L.V.: Features selection for primitives estimation on emotional speech. In: Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing, pp. 5138–5141 (2010)
20. Eyben, F., Wöllmer, M., Poitschke, T., Schuller, B., Blaschke, C., Färber, B., Nguyen-Thien, N.: Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. Adv. Hum.-Comput. Interact. **2010**, 263593 (2010), 17 pages
21. Eyben, F., Wöllmer, M., Valstar, M., Gunes, H., Schuller, B., Pantic, M.: String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (2011)

22. Faghihi, U., Fournier-Viger, P., Nkambou, R., Poirier, P., Mayers, A.: How emotional mechanism helps episodic learning in a cognitive agent. In: Proc. IEEE Symp. on Intelligent Agents, pp. 23–30 (2009)
23. Feldman, L.: Valence focus and arousal focus: Individual differences in the structure of affective experience. J. Pers. Soc. Psychol. **69**, 153–166 (1995)
24. Fletcher, R., Dobson, K., Goodwin, M.S., Eydgahi, H., Wilder-Smith, O., Fernholz, D., Kuboyama, Y., Hedman, E., Poh, M.Z., Picard, R.W.: iCalm: Wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. IEEE Tran. on Information Technology in Biomedicine **14**(2), 215
25. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.: The world of emotion is not two-dimensional. Psychol. Sci. **18**, 1050–1057 (2007)
26. Fragopanagos, N., Taylor, J.G.: Emotion recognition in human–computer interaction. Neural Netw. **18**(4), 389–405 (2005)
27. Frijda, N.H.: The Emotions. Cambridge University Press, Cambridge (1986)
28. Gilroy, S.W., Cavazza, M., Niiranen, M., Andre, E., Vogt, T., Urbain, J., Benayoun, M., Seichter, H., Billinghurst, M.: Pad-based multimodal affective fusion. In: Proc. Int. Conf. on Affective Computing and Intelligent Interaction Workshops, pp. 1–8 (2009)
29. Glowinski, D., Camurri, A., Volpe, G., Dael, N., Scherer, K.: Technique for automatic emotion recognition by body gesture analysis. In: Proc. of Computer Vision and Pattern Recognition Workshops, pp. 1–6 (2008)
30. Gökçay, D., Yıldırım, G.: Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives. IGI Global, Hershey (2011)
31. Grandjean, D., Sander, D., Scherer, K.R.: Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. Conscious. Cogn. **17**(2), 484–495 (2008)
32. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**, 602–610 (2005)
33. Grimm, M., Kroschel, K.: Emotion estimation in speech using a 3d emotion space concept. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop, pp. 381–385 (2005)
34. Grimm, M., Mower, E., Kroschel, K., Narayanan, S.: Primitives based estimation and evaluation of emotions in speech. Speech Commun. **49**, 787–800 (2007)
35. Grimm, M., Kroschel, K., Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database. In: ICME, pp. 865–868. IEEE Press, New York (2008)
36. Grundlehner, B., Brown, L., Penders, J., Gyselinckx, B.: The design and analysis of a real-time, continuous arousal monitor. In: Proc. Int. Workshop on Wearable and Implantable Body Sensor Networks, pp. 156–161 (2009)
37. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. Int. J. Synth. Emot. **1**(1), 68–99 (2010)
38. Gunes, H., Pantic, M.: Automatic measurement of affect in dimensional and continuous spaces: Why, what, and how. In: Proc. of Measuring Behavior, pp. 122–126 (2010)
39. Gunes, H., Pantic, M.: Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In: Proc. of International Conference on Intelligent Virtual Agents, pp. 371–377 (2010)
40. Gunes, H., Piccardi, M., Pantic, M.: Affective computing: focus on emotion expression, synthesis, and recognition. In: Or, J. (ed.) From the Lab to the Real World: Affect Recognition using Multiple Cues and Modalities, pp. 185–218. I-Tech Education and Publishing, Vienna (2008)
41. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: First steps towards an automatic system. In: LNCS, vol. 3068, pp. 36–48 (2004)
42. Hochreiter, S.: Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München (1991)
43. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **6**(2), 107–116 (1998)

44. Hoque, M.E., El Kaliouby, R., Picard, R.W.: When human coders (and machines) disagree on the meaning of facial affect in spontaneous videos. In: Proc. of Intelligent Virtual Agents, pp. 337–343 (2009)
45. Huang, T.S., Hasegawa-Johnson, M.A., Chu, S.M., Zeng, Z., Tang, H.: Sensitive talking heads. IEEE Signal Process. Mag. **26**, 67–72 (2009)
46. Huttar, G.L.: Relations between prosodic variables and emotions in normal American english utterances. J. Speech Hear. Res. **11**, 481–487 (1968)
47. Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis, T., Karpouzis, K., Kollias, S.: Emotion recognition through facial expression analysis based on a neurofuzzy method. Neural Netw. **18**(4), 423–435 (2005)
48. Jia, J., Zhang, S., Meng, F., Wang, Y., Cai, L.: Emotional audio-visual speech synthesis based on PAD. IEEE Trans. Audio Speech Lang. Process. **PP**(9), 1 (2010)
49. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 2nd edn. Prentice-Hall, New York (2008)
50. Kanluan, I., Grimm, M., Kroschel, K.: Audio-visual emotion recognition using an emotion recognition space concept. In: Proc. of the 16th European Signal Processing Conference (2008)
51. Karg, M., Schwimmbeck, M., Kühnlenz, K., Buss, M.: Towards mapping emotive gait patterns from human to robot. In: Proc. IEEE Int. Symp. in Robot and Human Interactive Communication, pp. 258–263 (2010)
52. Khalili, Z., Moradi, M.H.: Emotion recognition system using brain and peripheral signals: Using correlation dimension to improve the results of EEG. In: Proc. Int. Joint Conf. on Neural Networks, pp. 1571–1575 (2009)
53. Kierkels, J.J.M., Soleymani, M., Pun, T.: Queries and tags in affect-based multimedia retrieval. In: Proc. IEEE Int. Conf. on Multimedia and Expo, pp. 1436–1439 (2009)
54. Kim, J.: Robust speech recognition and understanding. In: Grimm, M., Kroschel, K. (eds.) Bimodal Emotion Recognition using Speech and Physiological Changes, pp. 265–280. I-Tech Education and Publishing, Vienna (2007)
55. Kim, J., Andre, E.: Emotion recognition based on physiological changes in music listening. IEEE Trans. Pattern Anal. Mach. Intell. **30**(12), 2067–2083 (2008)
56. Kipp, M., Martin, J.-C.: Gesture and emotion: Can basic gestural form features discriminate emotions? In: Proc. Int. Conf. on Affective Computing and Intelligent Interaction Workshops, pp. 1–8 (2009)
57. Kleinsmith, A., Bianchi-Berthouze, N.: Recognizing affective dimensions from body posture. In: Proc. of the Int. Conf. on Affective Computing and Intelligent Interaction, pp. 48–58 (2007)
58. Kleinsmith, A., De Silva, P.R., Bianchi-Berthouze, N.: Recognizing emotion from postures: Cross–cultural differences in user modeling. In: Proc. of the Conf. on User Modeling, pp. 50–59 (2005)
59. Kulic, D., Croft, E.A.: Affective state estimation for human-robot interaction. IEEE Trans. Robot. **23**(5), 991–1000 (2007)
60. Lang, P.J.: The Cognitive Psychophysiology of Emotion: Anxiety and the Anxiety Disorders. Erlbaum, Hillside (1985)
61. Levenson, R.: Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. In: Social Psychophysiology and Emotion: Theory and Clinical Applications, pp. 17–42 (1988)
62. McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: Proc. of IEEE Int'l Conf. Multimedia, Expo (ICME'10), pp. 1079–1084, July 2010
63. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Curr. Psychol. **14**, 261–292 (1996)
64. Mihelj, M., Novak, D., Munih, M.: Emotion-aware system for upper extremity rehabilitation. In: Proc. Int. Conf. on Virtual Rehabilitation, pp. 160–165 (2009)

65. Nakasone, A., Prendinger, H., Ishizuka, M.: Emotion recognition from electromyography and skin conductance. In: Proc. of the 5th International Workshop on Biosignal Interpretation, pp. 219–222 (2005)
66. Nicolaou, M.A., Gunes, H., Pantic, M.: Audio-visual classification and fusion of spontaneous affective data in likelihood space. In: Proc. of IEEE Int. Conf. on Pattern Recognition, pp. 3695–3699 (2010)
67. Nicolaou, M.A., Gunes, H., Pantic, M.: Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In: Proc. of LREC Int. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, pp. 43–48 (2010)
68. Nicolaou, M.A., Gunes, H., Pantic, M.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence–arousal space. IEEE Trans. Affect. Comput. **2**(2), 92–105 (2011)
69. Nicolaou, M.A., Gunes, H., Pantic, M.: Output-associative RVM regression for dimensional and continuous emotion prediction. In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (2011)
70. Oliveira, A.M., Teixeira, M.P., Fonseca, I.B., Oliveira, M.: Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity. In: Proc. of the 22nd Annual Meeting of the Int. Society for Psychophysics, pp. 245–250 (2006)
71. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1988)
72. Parkinson, B.: Ideas and Realities of Emotion. Routledge, London (1995)
73. Patras, I., Pantic, M.: Particle filtering with factorized likelihoods for tracking facial features. In: Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 97–102 (2004)
74. Paul, B.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the Institute of Phonetic Sciences, pp. 97–110 (1993)
75. Petridis, S., Gunes, H., Kaltwang, S., Pantic, M.: Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. In: Proc. of ACM Int. Conf. on Multimodal Interfaces, pp. 23–30 (2009)
76. Picard, R.W.: Emotion research by the people, for the people. Emotion Review **2**(3), 250–254
77. Plutchik, R., Conte, H.R.: Circumplex Models of Personality and Emotions. APA, Washington (1997)
78. Poh, M.Z., Swenson, N.C., Picard, R.W.: A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Trans. Inf. Technol. Biomed. **57**(5), 1243–1252 (2010)
79. Pun, T., Alecu, T.I., Chanel, G., Kronegg, J., Voloshynovskiy, S.: Brain–computer interaction research at the Computer Vision and Multimedia Laboratory, University of Geneva. IEEE Trans. Neural Syst. Rehabil. Eng. **14**, 210–213 (2006)
80. Rehm, M., Wissner, M.: Gamble-a multiuser game with an embodied conversational agent. In: Lecture Notes in Computer Science, vol. 3711, pp. 180–191 (2005)
81. Roseman, I.J.: Cognitive determinants of emotion: A structural theory. In: Shaver, P. (ed.) Review of Personality & Social Psychology, Beverly Hills, CA, vol. 5, pp. 11–36. Sage, Thousand Oaks (1984)
82. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)
83. Salahuddin, L., Cho, J., Jeong, M.G., Kim, D.: Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In: Proc. of the IEEE 29th International Conference of the EMBS, pp. 39–48 (2007)
84. Sander, D., Grandjean, D., Scherer, K.R.: A systems approach to appraisal mechanisms in emotion. Neural Netw. **18**(4), 317–352 (2005)
85. Scherer, K.R., Oshinsky, J.S.: Cue utilization in emotion attribution from auditory stimuli. Motiv. Emot. **1**, 331–346 (1977)

86. Scherer, K.R., Schorr, A., Johnstone, T.: Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, Oxford/New York (2001)
87. Schröder, M.: Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis. Ph.D. dissertation, Univ. of Saarland, Germany (2003)
88. Schröder, M., Heylen, D., Poggi, I.: Perception of non-verbal emotional listener feedback. In: Hoffmann, R., Mixdorff, H. (eds.) Speech Prosody, pp. 1–4 (2006)
89. Schröder, M., Bevacqua, E., Eyben, F., Gunes, H., Heylen, D., Maat, M., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., Sevin, E., Valstar, M., Wöllmer, M.: A demonstration of audiovisual sensitive artificial listeners. In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction, vol. 1, pp. 263–264 (2009)
90. Schröder, M., Pammi, S., Gunes, H., Pantic, M., Valstar, M., Cowie, R., McKeown, G., Heylen, D., ter Maat, M., Eyben, F., Schuller, B., Wöllmer, M., Bevacqua, E., Pelachaud, C., de Sevin, E.: Come and have an emotional workout with sensitive artificial listeners! In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition (2011)
91. Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H.: Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image Vis. Comput. **27**, 1760–1774 (2009)
92. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.: Acoustic emotion recognition: A benchmark comparison of performances. In: Proc. of Automatic Speech Recognition and Understanding Workshop, pp. 552–557 (2009)
93. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: The INTERSPEECH 2010 paralinguistic challenge. In: Proc. INTERSPEECH, pp. 2794–2797 (2010)
94. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**, 2673–2681 (1997)
95. Shen, X., Fu, X., Xuan, Y.: Do different emotional valences have same effects on spatial attention. In: Proc. of Int. Conf. on Natural Computation, vol. 4, pp. 1989–1993 (2010)
96. Sneddon, I., McKeown, G., McRorie, M., Vukicevic, T.: Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour. PLoS ONE **6**, e14679–e14679 (2011)
97. Soleymani, M., Davis, J., Pun, T.: A collaborative personalized affective video retrieval system. In: Proc. Int. Conf. on Affective Computing and Intelligent Interaction and Workshops, pp. 1–2 (2009)
98. Sun, K., Yu, J., Huang, Y., Hu, X.: An improved valence-arousal emotion space for video affective content representation and recognition. In: Proc. IEEE Int. Conf. on Multimedia and Expo, pp. 566–569 (2009)
99. Trouvain, J., Barry, W.J.: The prosody of excitement in horse race commentaries. In: Proc. ISCA Workshop Speech Emotion, pp. 86–91 (2000)
100. Truong, K.P., van Leeuwen, D.A. Neerincx, M.A. de Jong, F.M.G.: Arousal and valence prediction in spontaneous emotional speech: Felt versus perceived emotion. In: Proc. INTERSPEECH, pp. 2027–2030 (2009)
101. Tsai, T.-C., Chen, J.-J., Lo, W.-C.: Design and implementation of mobile personal emotion monitoring system. In: Proc. Int. Conf. on Mobile Data Management: Systems, Services and Middleware, pp. 430–435 (2009)
102. Tsiamyrtzis, P., Dowdall, J., Shastri, D., Pavlidis, I.T., Frank, M.G., Ekman, P.: Imaging facial physiology for the detection of deceit. Int. J. Comput. Vis. (2007)
103. Wang, P., Ji, Q.: Performance modeling and prediction of face recognition systems. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1566–1573 (2006)
104. Wassermann, K.C., Eng, K., Verschure, P.F.M.J.: Live soundscape composition based on synthetic emotions. IEEE Multimed. **10**, 82–90 (2003)
105. Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R.: Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies. In: Proc. INTERSPEECH, pp. 597–600 (2008)

106. Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G.: Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. IEEE J. Sel. Top. Signal Process. **4**(5), 867–881 (2010)
107. Yang, Y.-H., Lin, Y.-C., Su, Y.-F., Chen, H.H.: Music emotion classification: A regression approach. In: Proc. of IEEE Int. Conf. on Multimedia and Expo, pp. 208–211 (2007)
108. Yu, C., Aoki, P.M., Woodruff, A.: Detecting user engagement in everyday conversations. In: Proc. of 8th Int. Conf. on Spoken Language Processing (2004)
109. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 39–58 (2009)

Chapter 11
Analysis of Group Conversations: Modeling Social Verticality

Oya Aran and Daniel Gatica-Perez

11.1 Introduction

Social interaction is a fundamental aspect of human life and is also a key research area in psychology and cognitive science. Social psychologists have been researching the dimensions of social interaction for decades and found out that a variety of social communicative cues strongly determine social behavior and interaction outcomes. Many of these cues are consciously produced, in the form of spoken language. However, besides the spoken words, human interaction also involves nonverbal elements, which are extensively and often unconsciously used in human communication. The nonverbal information is conveyed as wordless messages, in parallel to the spoken words, through aural cues (voice quality, speaking style, intonation) and also through visual cues (gestures, body posture, facial expression, and gaze) [31]. These cues can be used to predict human behavior, personality, and social relations. It has been shown that, in many social situations, humans can correctly interpret the nonverbal cues and can predict behavioral outcomes with high accuracy, when exposed to short segments or "thin slices" of expressive behavior [1]. The length of these thin slices can change from a few seconds to several minutes depending on different situations.

Computational analysis of social interaction, in particular of face-to-face group conversations is an emerging field of research in several communities such as human–computer interaction, machine learning, speech and language processing, and computer vision [20, 38]. Close connection with other disciplines including psychology and linguistics also exist in order to understand what kind of verbal

O. Aran (✉) · D. Gatica-Perez
Idiap Research Institute, Martigny, Switzerland
e-mail: oya.aran@idiap.ch

D. Gatica-Perez
e-mail: gatica@idiap.ch

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_11,

and nonverbal signals are used in diverse social situations to infer human behavior. The ultimate aim is to develop computational systems that can automatically infer human behavior by observing a group conversation via sensing devices such as cameras and microphones. Besides the value for several social sciences, the motivation behind the research on automatic sensing, analysis, and interpretation of social behavior has several dimensions. These systems could open doors to a number of relevant applications that support interaction and communication. These include tools that improve collective decision making and that support self-assessment, training, and education, with possible example applications such as automatic meeting evaluators, trainers, and automatic personal coaches for self learning. Moreover, not only supporting human interaction, these systems can also support a natural human–robot or human–computer interaction, by designing socially aware systems [38], i.e. by enabling a robot to understand the social context around it and to act accordingly.

In this chapter we focus on one aspect of social relations and interactions: social verticality. Social verticality is one of the many dimensions of human relations and refers to the structure of interpersonal relations positioned in a low-to-high continuum, stating a kind of social hierarchy among people [22]. It relates to power, status, dominance, leadership, and other related concepts. The vertical dimension is in contrast to the horizontal dimension, which is the affective and socio-emotional dimension that describes the emotional closeness of human relations. Instead, vertical dimension describes how each person is positioned in the group, e.g. as higher status/lower status. We present computational models for the analysis of social verticality through nonverbal cues in small groups.

The next section gives definitions and a brief summary of the psychological and cognitive aspects of the display and perception social verticality during human interactions. Section 11.3 describes computational methods and Sect. 11.4 presents four case studies. A summary of the chapter, acknowledgments, end-of-chapter questions and a small glossary can be found in the remaining sections.

11.2 Social Verticality in Human Interaction and Nonverbal Behavior

Social verticality constructs, such as power, status, and dominance, are related to each other with important differences in their definitions. For example, power indicates “the capacity to produce intended effects, and in particular, the ability to influence the behavior of another person” [17] (p. 208), without implying any respect or prestige. As power is defined as an ability, it is not always exercised. On the other hand, dominance can be defined as “a personality trait involving the motive to control others, the self perception of oneself as controlling others, and/or as a behavioral outcome with a success on controlling others” [22] (p. 898), and as a result, it is “necessarily manifest” [17] (p. 208). Dominance can be seen as a “behavioral manifestation of the relational construct of power” [17] (p. 208).

Status relates to both power and dominance and is defined as an "ascribed or achieved quality implying respect and privilege, does not necessarily include the ability to control others" [22] (p. 898). Leadership is another related construct, which can be defined as the ability of motivating a group of people to pursue a common goal. Thus, leadership is related to the end result, not just a manifested act. Among various types of leadership types, "emergent leadership", for example, is where the leader arises from a group of equal status people [46].

Dominance is one of the fundamental dimensions of social interaction. It is signaled via both verbal and nonverbal cues. The nonverbal cues include vocalic ones such as speaking time [45], loudness, pitch, vocal control, turns, and interruptions [17] and kinesic ones such as gesturing, posture, facial expressions, and gaze [16]. Dominant people are in general more active both vocally and kinesically, with an impression of relaxation and confidence [22]. It has been shown that they also have a higher visual dominance ratio (looking-while speaking to looking-while-listening ratio), i.e. they look at others more while speaking and less while listening [16].

In a study that investigated the relationship between the leadership style and sociable and aggressive dominance, it is found out that there is a higher correlation between leadership and sociable dominance [28]. Sociably dominant people look at others more while speaking and use more gestures, receiving more frequent and longer-lasting glances from the group; whereas aggressively dominant people interrupt more, and they look at others less while listening.

Social verticality in a group can also be defined by roles that constitute a hierarchy-like structure, such as physician/patient, manager/employee, teacher/learner, interviewer/interviewee, where one part has more expertise than the other, in terms of knowledge or rank. In [45], it is shown that the association between speaking time and dominance is higher for both dominant and high status people. It is important to note that not all the role distributions of a group present a vertical relationship, i.e. a hierarchical relation. In this chapter, we mainly refer to the roles that have a vertical dimension. We also consider the roles that are defined based on the psychological behavior (i.e. functional roles) of the participants in a group that partly shows a hierarchical structure.

The relationship between the vertical constructs and personality traits is also of interest to social psychologists. Personality is defined as a collection of consistent behavioral and emotional traits that characterize a person [18]. While personality addresses stable and consistent behavior of a person, the social verticality constructs address the behavior of a person in a group which may not be consistent across time, relations and situations [22]. Nevertheless, verticality constructs are closely related with the personality traits of the individuals, as personality strongly influences verticality in a social relationship of peers. For example, it is shown that cognitive ability and the personality traits of extroversion and openness to experience are predictive of emergent leadership behavior [29].

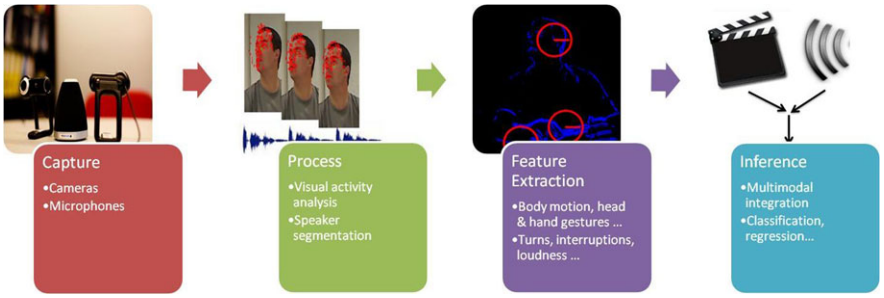


Fig. 11.1 The functional blocks of social interaction analysis

11.3 Automatic Analysis of Social Verticality from Nonverbal Features

The aim of the automatic analysis of group interaction is to infer and possibly predict aspects of the underlying social context, including both individual attributes and interactions with other people in the group. As a specific case, the concept of social verticality imposes a social hierarchy on the group and requires a special interest.

In this chapter we concentrate on three social constructs of social verticality, dominance, roles, and leadership, and explain the functional blocks, as shown in Fig. 11.1, that are required for analysis. Table 11.1 presents an overview of recent works selected from the literature and Table 11.2 summarizes several datasets used for social verticality analysis.

In the next sections, we mainly focus on audio and visual sensors as the two primary sources of information. A brief discussion on other sensors to capture social behavior is given in Sect. 11.3.2.3.

11.3.1 Processing

Although it is known that nonverbal communication involves both the aural and visual modalities, most initial works in the literature on computational analysis of group interactions have largely focused on audio features. This is partly related to the data capture technology. Reasonable quality audio capture systems were available earlier than video capture systems. Overall, video capture is more problematic than audio capture; video is quite sensitive to environmental conditions, and requires adequate resolution and frame rates. A second dimension is related to the challenges of video processing in natural conversations. One last dimension is related to privacy. People are in general less willing to have their video recorded compared to audio.

It is also important to note that social interaction capture systems should not use sensors that affect the interaction. The subjects should be able to act naturally without any distraction caused by the sensors that are used. For instance, to record audio,

Table 11.1 Related works on the analysis of social verticality

	Task	Audio features	Video features	Fusion	Inference
Dominance					
[24]	most/least dominant person	speaking turn	NA	NA	rule-based
[2]	most/least dominant person	speaking turn	visual activity, audio-visual activity	score level	rule-based
[12]	correlations with dominance	prosodic	NA	NA	correlations
[27]	most/least dominant person	speaking turn, prosodic	visual activity	feature level	rule-based supervised -SVM
[26]	most dominant/ high status person	speaking turn	visual activity, visual attention	NA	rule based
[42]	dominance level (as high, normal, low)	speaking turn (manual)	NA	NA	supervised -SVM
Roles					
[41]	role patterns	speaking turn	NA	NA	rule-based supervised -Influence Model
[43]	role recognition	speaking turn	NA	feature level	supervised -MAP, ML est. -Simulated ann.
[19]	role recognition	speaking turn, verbal	NA	score level	supervised -ML estimate -Boosting
[15]	role recognition	speaking turn	visual activity	feature level	supervised -SVM -HMM -Influence Model
[50]	role recognition	speaking turn	visual activity	feature level	supervised -SVM
[6]	role identification	speaking turn, verbal	NA	NA	supervised -Boosting
Leadership					
[44]	emergent leadership	speaking turn	NA	feature level	rule-based
[48]	leadership	prosodic	visual activity, gestures	feature level	rule-based
[25]	group conversational patterns	speaking turn	NA	feature level	unsupervised -topic models

Table 11.2 Selected databases used for social verticality analysis

Dataset Name	Task	Details	Length	References
DOME (part 1,2)	Dominance	Meetings, a subset of the AMI corpus, publicly available	∼10 hours	[2, 3]
DOME (part 1)	Dominance	Meetings, a subset of the AMI corpus, publicly available	∼5 hours	[12, 24, 26, 27]
–	Dominance, Roles	Meetings from "The Apprentice" TV show	90 minutes	[41]
–	Roles	News, talk shows from Swiss radio	∼46 hours	[43]
AMI	Roles	Meetings, publicly available	∼46 hours	[19, 43]
Survival	Roles	Meetings on the mission survival task	∼5 hours	[15, 40, 50]
ELEA	Leadership	Meetings of newly formed groups	∼10 hours	[44]

a distant microphone array device should be preferred to head-set microphones that are attached to the people. Regarding the cameras, while it is true that people might be distracted or feel self-conscious at the beginning of an interaction, when they are in a natural environment and focused on an engaging task, they rapidly tend to forget about the presence of cameras and act naturally.

11.3.1.1 Audio Processing

The key audio processing step is to segment each speaker's data in the conversation such that it allows robust further processing to extract features for each of the participants separately. This process is called speaker diarization, a combination of speaker segmentation (finding speaker change points) and speaker clustering (grouping segments for each speaker). Speaker diarization is an active topic, not only in social interaction analysis, but in speech processing in general. Here we refer to its applications in social interaction analysis very briefly.

As the audio capture methodology, one can use different setups ranging from one single microphone [24], to microphone arrays [44] and head-set microphones [2, 27]. Each of these setups have different noise levels and require different levels of processing for speaker diarization. Head-set microphones provide lower levels of noise, however, there is still a need for speaker diarization, since voices of other participants also exist in the recordings. In [24], the authors used a single audio source and investigated the performance of speaker diarization and dominance estimation under different conditions. Their results show that dominance estimation is robust to diarization noise in the single audio source case. For recording three-four people meetings, Sanchez-Cortes et al. [44] used a commercial microphone array,

which provides the speaker diarization output along with the audio recordings. The speaker diarization output is used for estimating the emergent leader in the group.

11.3.1.2 Video Processing

Once the camera input is received, the visual activity of each participant in the meeting needs to be processed. The level of processing depends on the features that will be extracted and ranges from face detection to motion detection, from face tracking to skin blob tracking and body parts tracking (see Chap. 3 for more details).

Face detection algorithms in general are designed to detect either frontal or profile faces. Their performance is affected especially if there are out-of-plane rotations. During group conversations, even if the participants are sitting around a table and are stationary, as a part of the interaction, they may frequently move their heads, and gaze at each other. Thus, a face detection algorithm alone is not sufficient to extract the positions of the faces during an interaction. Face tracking algorithms (using Kalman filter, particle filter, etc.) could be applied for better results. On top of face tracking, if the body movements, such as hand gestures and body postures, are also of interest, advanced techniques to track body parts or just the skin-colored parts can be applied. A lower level of processing, such as motion detection, can be applied if the interest is on the general visual activity of the participant and not on the individual body parts' activity. More detail on these techniques is given in Sect. 11.3.2.2.

11.3.2 Feature Extraction

The descriptors of a social interaction and social behavior can be categorized with respect to the sensor used, but also with respect to whether the feature is extracted from a single participant's activity or from an interaction that involves more than one person. We will call the former type of features as "independent features" and the latter as "relational features". Moreover, one can extract features that represent the overall interaction of the whole meeting, instead of the participants one by one, which we refer as "meeting features". The following sections present a sensor based categorization of both independent, relational, and meeting features.

The features presented below are generally extracted from thin slices of meetings, summarizing independent and group behavior for that meeting segment. Research in social psychology has shown that by examining only brief observations of expressive behavior, humans are able to predict behavioral outcomes [1]. The current research in computational social behavior analysis uses these conclusions and investigates whether computational methods are also able to predict similar outcomes by applying thin slice based processing. The length of the slices and how these slices are to be processed should be determined with respect to the

task and are open questions. The whole meeting duration can be used as one single segment, or the entire segment can be described as an accumulation of shorter slices.

11.3.2.1 Audio Nonverbal Features

In this section, we present the audio nonverbal features in two groups: speaking turn features and prosodic features. The term "speaking turn features" refers to audio nonverbal features that are extracted based on the speaking status of the participants and their turn taking behavior. The prosodic features on the other hand, represent the rhythm, stress, and intonation of speech.

Speaking Turn Features

Speaking turn features are frequently used in social interaction analysis for two main reasons. First, given the speaker diarization output, they are easy to calculate, with very low computational complexity. Second, despite their simplicity, they are very successful in many social tasks, supported by both social psychology and social computing research [20, 45].

Speaking turn features can be categorized as independent and relational. Independent features describe the speaking activity for one participant. These include speaking length, number of speaking turns, turn duration statistics (average, min, max, etc.). A turn is defined as one continuous segment where a participant starts and ends her/his speech. The relational features describe the interaction of one participant with other participants in the group. These include interruptions, overlapped speech, turn taking order (i.e. who speaks after whom) and also centrality features. Centrality features are relational features that represent the relative position of each participant in the group. We can represent the interaction of a group as a graph, by taking the nodes as the participants and the edges as the indication of how one person relates to others. The edges can be connected to several other relational features, such as "who interrupts whom", "who speaks after whom", etc. One can assign weights to the edges, representing the strength of the relation. Based on the definition of the relational features, the edges can be directed, in which case they are called arcs. Once the graph is formed, centrality measures can be calculated in various ways, such as indegree and outdegree of each node, closeness to other nodes (with weights defined as distances), etc.

The interaction patterns between participants can also be defined via Social Affiliation Networks (SAN) and used to represent the relationships between the roles [43]. A SAN is a graph that encodes "who interacts with whom and when". The two kinds of nodes in a SAN refer to the actors and the events, respectively. In [43], the events are defined as the segments from the recordings, and the participants are linked to the events if they talk during the corresponding segment. The assumption in this representation is that if the roles influence the structure of the interaction, similar interaction patterns should correspond to the same roles.

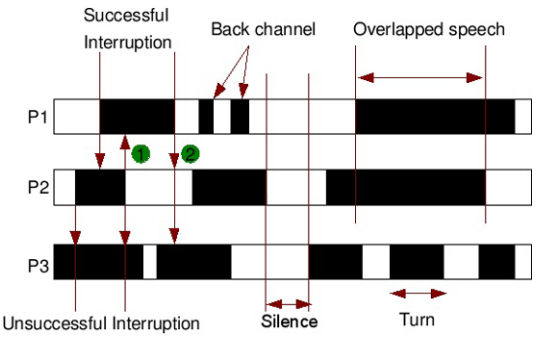


Fig. 11.2 Speaking turn audio nonverbal features

Speaking turn features that describe the whole meeting include the amounts of silence, overlapped speech and non-overlapped speech, among others. Accumulated statistics of all participants can also be used as meeting features (total number of turns, interruptions, etc.).

Figure 11.2 shows the illustration of a conversation between three participants. Each line represents the timeline for one participant and black segments indicate that the participant is speaking. Each black segment is a turn. The overlapped speech and silence segments are indicated. The interruptions and backchannels are also represented. The automatic detection of successful interruptions can be done in several ways if the verbal information is to be omitted. One definition can be made with respect to the interruptee's point of view (indicated with 1 in Fig. 11.2): "P1 started speaking when P2 was already speaking and P2's turn ended before P1's". Another definition uses the interrupter's point of view (indicated with 2 in Fig. 11.2): "P1 started speaking when P2 was already speaking and when P1's turn ended P2 was not speaking anymore". If we follow the first definition, P1 successfully interrupts P3 as well, however, with the second definition, it is an unsuccessful interruption. By definition, an interruption occurs between two participants. In that case, to calculate the number of interruptions for one person, all possible pairings with that person should be considered. As an alternative, interruptions that affect the whole group can be extracted [2].

All of these features need to be normalized with respect to the meeting duration (and with respect to the number of participants) before they are used in inference.

Prosodic Features

Other than speaking turn features, prosodic nonverbal features are also indicators of social behavior and used recently in several tasks such as dominance estimation. Features like pitch, energy, rhythm, or spectral features like formants, bandwidths, spectrum intensity can be extracted as independent features for each participant. Recently Charfuelan et al. [12] investigated the correlations between various prosodic features and dominance. Their results show that the most dominant person tends to speak louder and the least dominant person tends to speak softer than average.

In [48], prosodic features from audio such as loudness and spectral features are fused with visual features to estimate leadership in musical performances. In [26, 27], the speaking energy is used as a prosodic feature, together with other speaking turn features, for dominance estimation tasks. However, the results show that the speaking turn audio features, such as total speaking length and number of turns, can outperform speaking energy in predictive power.

11.3.2.2 Visual Nonverbal Features

Visual Activity

Most of the works in the nonverbal communication literature extract low-level visual features based on global image motion or geometric image primitives. In part, this approach can be feasible as there are no clearly defined hand shapes and hand trajectories for the visual nonverbal features of social verticality. Image and motion based approaches either assume that the background is stationary and any detected motion will indicate participant's visual activity, or find the skin-colored regions or faces and calculate the motion for these parts only.

In [27], two types of visual information, extracted from the compressed domain [49], are used for modeling dominance: the motion vector magnitude and the residual coding bit rate. While the motion vector magnitude reflects the global motion, the residual coding bit rate provides the local motion, such as lip movement on the face or finger movement on the hands. These two types of information can be used as indicators of the visual activity of the participant, either alone, or as a combination. In [2, 27], by thresholding the motion information, the authors extracted a binary vector in which zeros indicate no-motion segments and ones indicate the segments with motion. A number of higher level visual features are extracted from this binary vector, including the length, turns and interruptions of visual activity, similar to the audio speaking turn features explained in Sect. 11.3.2.1. Moreover, audio-visual versions of these features can also be extracted, by looking at the joint speaking and visual activity behavior. For example, in [2], the authors extracted visual attention features while speaking to estimate the most/least dominant persons in a group.

Another method is to use the motion history templates [9] for the detection and understanding of visual activity. In [13], motion history images are calculated for skin-colored regions and the amount of fidgeting, defined as "a condition of restlessness as manifested by nervous moments", is measured, by applying empirically determined thresholds. These features are used for the recognition of functional roles [15, 40, 50].

Visual Attention and Gaze

When and how much people look at each other during a conversation is a clear indicator of many social constructs. For example, dominant people often look at

others more while speaking and less while listening. And the ratio between these two measures, defined as the "Visual Dominance Ratio (VDR)", is considered as one of the classic measures of dominance [16]. Moreover, receiving more visual attention from other participants is an indicator of dominance.

For the automatic estimation of gaze and visual Focus of Attention (FOA), one can either use eye gaze or head pose. Although eye gaze is a more reliable source, with the current technology, this can be only implemented via eye trackers or high resolution cameras focused on each participants face area. Alternatively, the head pose can be estimated as an approximation to the actual eye gaze [5, 21, 36]. In a natural conversation environment, the focus target needs to be defined. Other than the participants in the conversation, there can be other targets such as the table, laptops, presentation screen, board, etc. In [5], an input–output hidden Markov model is used to detect the FOA of the group participants. In their work, the authors propose to recognize the FOA of all participants jointly and introduce a context dependent interaction model. Their model achieves around 10% performance increase when compared to using independent models for each participant. More details on FOA estimation can be found in Chap. 4.

Once the FOA of participants for each time frame is extracted, several measures can be defined as indicators of dominance. These include received visual attention, given visual attention (looking at others), and the VDR [23, 26]. These features, which are initially defined for dyadic conversations, can be generalized to the multi-party case by accumulating all possible pairwise participant combinations. It is important to note that VDR is by definition a multimodal cue, as it considers the FOA of a participant with respect to speaking status. For VDR, one needs to calculate "looking-while-speaking" and "looking-while-listening" measures. The "looking-while-speaking" case is trivial, however, "looking-while-listening" can be defined as "looking-while-not-speaking" or "looking-while-someone-else-is-speaking". In [26], the authors define two variants of VDR following these two definitions, and use them for dominance and status estimation.

Gestures and Facial Expressions

Despite the progress in computer vision to analyze structured gestures (e.g. hand gesture recognition, sign language recognition, gait recognition, etc.), the use of more accurate models of visual nonverbal communication has been largely unexplored. The main challenge is the lack of clearly defined gestures for visual nonverbal features. Another challenge is the uncontrolled experimental setup. Contrary to the natural conversational environment that is required for social interaction analysis, most of the developed gesture recognition algorithms require controlled environments and restrict people to perform gestures in a certain way.

To the authors' knowledge, the use of specific hand gestures, other than extracting general hand activity, has not been applied to the automatic analysis of social verticality. In one study [48], expressive gestures for musical performance, such as motion fluency, impulsiveness, directness, are used as features for leadership estimation. Head and body gestures have been used in related tasks in social interaction

analysis. While most of the works focus on the face area and head gestures, there are a few studies that also use the body motion. In [10], for estimating the participant status, the authors extract features from the face area and estimate yes-no head movements and also the global body movement. In [37], for an addressing task, to respond to questions such as "who responds to whom, when and how?", the authors extract features from the head area using a discrete wavelet transform to estimate head gestures such as nodding, shaking and tilt. A magnetic sensor is used in this study to capture the head motion.

Facial expressions are also strong indicators of social behavior. Despite progress on the automatic analysis of facial expressions, a widely studied topic in recent years, they are not as widely used in social interaction analysis. The main reason behind this is that facial expression analysis requires high resolution recordings of the facial region. However, most of the databases used for social interaction analysis use upper body or full body recordings of the participants and the captured facial region in these recordings are not good enough to perform high-level automatic expression analysis. Among the few works, in [32], the participants' smiling status is extracted during an interaction.

11.3.2.3 Other Sensors

The increasing use of mobile devices in people's daily lives introduced the opportunity to researchers to use these devices as capture devices. Mobile devices can record vast amounts of data from people's daily interaction via built-in audio and video sensors, but also via other sensors such as accelerometers to measure body movement, bluetooth or radio signals to measure proximity between two devices, and several others [34, 39].

The main challenge of using mobile devices for social behavior capture is limited computational resources. To overcome this difficulty, special equipment to collect social behavioral data can be developed. The sociometric badge is an example of such devices: it collects and analyzes social behavioral data. It allows voice capture, infrared (IR) transmission and reception and is capable of extracting features that can further be used for social behavior analysis [30], in real-time.

11.3.3 Inference

This section presents four groups of methods that have been used to infer social verticality:

1. *Rule-based approaches*
2. *Unsupervised approaches*
3. *Supervised approaches*
4. *Temporal approaches*

In the first approach, a decision with regards to the social verticality concept, for instance dominance, is made via rules defined using expert knowledge, without the need of training data. In the second and third approaches, unsupervised or supervised machine learning techniques are used, respectively. Their main difference is in the availability of labeled training data. Finally, in the fourth approach, the entire temporal dynamics of interaction is taken into account.

11.3.3.1 Rule-Based Approaches

For some social behavioral concepts, it is shown that people use several nonverbal cues more frequently or less frequently, with respect to other people in the group. For example, according to social psychology, dominant people often speak more, move more, or grab the floor more often [17, 22], so if someone speaks the most or moves the most, he/she is more likely to be perceived as dominant over the other people in the meeting. Following this information, one can assume that the nonverbal cues defined above are positively correlated with dominance and define a rule-based estimator on each related nonverbal feature [2, 27]. Similarly, other rule-based estimators can be defined for other social tasks, based on sociological and psychological aspects. Here we give an example of a rule-based estimator for dominance.

To estimate the most dominant person in meeting i, using feature f, the rule is defined as:

$$MD_i = \arg\max_p \left(f_p^i\right), \quad p \in \{1, 2, \ldots, P\}, \tag{11.1}$$

where p is the participant number, f_p^i is the value of the feature for that participant in meeting i, and P is the number of participants. The least dominant person can be estimated similarly using the following rule:

$$LD_i = \arg\min_p \left(f_p^i\right), \quad p \in \{1, 2, \ldots, P\}. \tag{11.2}$$

The main advantage of these rule based estimators is that they do not require any training data and they are very fast to compute. On the other hand, the major disadvantage is that they only allow the use of a single feature and cannot directly utilize the power of combining multiple features. By definition, the rule-based estimator is limited to a single feature. In the next section, we explain several approaches to perform fusion using the rule-based estimator.

Multimodal Fusion via Rule-Based Estimator

Although speaking length alone is a good estimator of dominance, there are other displays of dominance as well, such as the visual activity, which provides complementary information. Thus, different features representing different aspects of dominance could be fused together to obtain a better estimator. We can define a

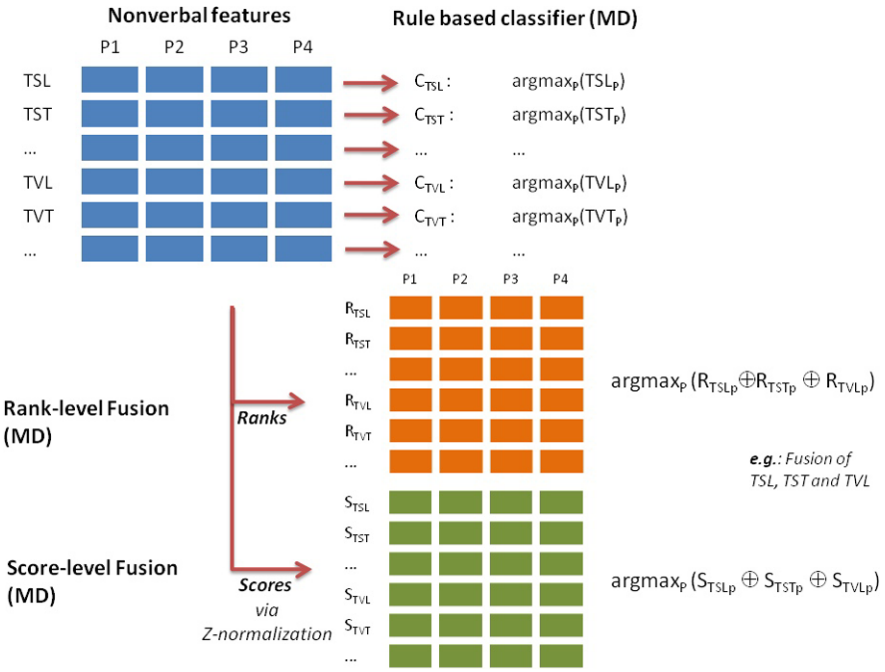


Fig. 11.3 Dominance estimation with rule based classifier and rank and score level fusion. The rank and score information is calculated from the extracted nonverbal features for each participant

rule-based estimator on each feature as an independent classifier and apply fixed combination rules on the decisions of these classifiers. Two different fusion architectures are presented in this section: score level and rank level fusion [33]. An overview of the fusion architectures is shown in Fig. 11.3.

For *Score Level Fusion*, each classifier should provide scores, representing the support of the classifier for each class. The scores of each classifier are then combined by simple arithmetic combination rules such as sum, product, etc. The scores to be combined should be in the same range, so a score normalization should be performed prior to fusion.

As the actual feature values are positively correlated with dominance, they can be used as the scores of the rule-based classifier, defined on that nonverbal feature, following a normalization step. z-normalization can be used to normalize the features for each meeting:

$$\hat{f}_p^i = \left(f_p^i - \mu_{fi}\right)/(\sigma_{fi}), \quad \forall p \in 1, \ldots, P, \tag{11.3}$$

where $\hat{f}_p^i$ and f_p^i are the values of the feature f for participant p in meeting i, z-normalized and prior to normalization, respectively. μ_{fi} and σ_{fi} are the mean and the standard deviation over all participants. The score level fusion can then be performed by using an arithmetic combination rule. For meeting i, this would mean

using feature combination $\mathcal{C}$, combining the scores for each participant (e.g. with sum rule) and selecting the participant with the highest total score:

$$S_i^{\mathcal{C}} = \arg\max_p \left(\sum_{f \in \mathcal{C}} \hat{f}_p^i \right), \quad \mathcal{C} \subseteq \mathcal{F}, \tag{11.4}$$

where $\mathcal{F}$ is the set of all features.

Rank Level Fusion is a direct extension of the rule-based estimator. Instead of selecting the participant with the maximum feature value, the participants are ranked and the rank information is used to fuse different estimators based on different features. The ranks for each participant are summed up and the one with the highest total rank is selected as the most dominant. For meeting i, using feature combination $\mathcal{C}$, the most dominant participant is selected by

$$R_i^{\mathcal{C}} = \arg\max_p \left(\sum_{f \in \mathcal{C}} r_{f_p}^i \right), \quad \mathcal{C} \subseteq \mathcal{F}, \tag{11.5}$$

where $r_{f_p}^i$ is the rank of participant p using feature f in meeting i. In case of ties, the selection can be performed based on the z-normalized scores.

11.3.3.2 Unsupervised Approaches

Unsupervised approaches can be applied to analyze social behavior data to discover and to differentiate patterns of certain behavior types. The motivation behind using unsupervised approaches is that they decrease the dependency to labeled training data. Given the difficulty of collecting data annotations for social interactions, this is a huge opportunity. Moreover, for problems with none or vague class descriptions, unsupervised approaches provide better models. Although there is always a trade-off between the performance and the amount of labeled training data, efficient unsupervised (or semi-supervised) techniques can be developed that would result in low performance degradation by using none or a very small amount of training data, when compared to using huge amounts of data.

Among the many diverse methods, we present here topic models, in particular Latent Dirichlet Allocation (LDA) [8], as an example model to discover social patterns. Topic models are probabilistic generative models that are proposed to analyze the content of documents. Although topic models were originally used in text modeling, they are capable of modeling any collection of discrete data. The patterns are discovered based on word co-occurrence. In topic models, each document is viewed as a mixture of topics, where topics are distributions over words. A word is defined as a basic unit of the discrete data. The probability of a word w in a document, assuming the document is generated from a convex combination of T topics, is given as

$$p(w_i) = \sum_{t=1}^{T} p(w_i | z_i = t) p(z_i = t), \tag{11.6}$$

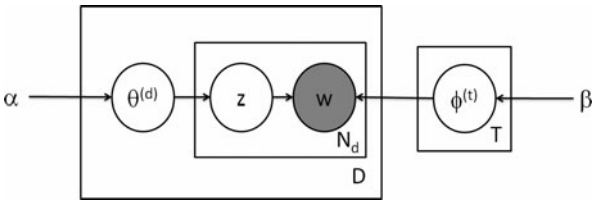


Fig. 11.4 Latent Dirichlet allocation model.

where z_i is a latent variable indicating the topics from which the word w_i may be drawn.

Assume that a document d is a bag of N_d words, and a corpus is a collection of D documents, with a total of N words (i.e. $N = \sum N_d$), and a vocabulary size of V. Let $\phi^{(t)} = p(w|z = t)$ refer to the multinomial word distribution for each topic t, and $\theta^{(d)} = p(z)$ refer to the topic distribution for each document. ϕ indicates which words are important for each topic, and θ indicates which topics are important for each document.

LDA [8] model assumes a Dirichlet prior both on the topic and word distributions ($p(\theta)$ and $p(\phi)$ are Dirichlet with hyperparameters α and β, respectively) to provide a complete generative model for documents. As the Dirichlet distribution is a conjugate prior to the multinomial, its usage simplifies the statistical inference problem and allows variational inference methods to be used. Then, the joint distribution of the set of all words in a given document is given by

$$p(z, w, \theta, \phi|\alpha, \beta) = \prod_{i=1}^{N} p(w_i|z_i, \phi)p(z_i|\theta)p(\theta|\alpha)p(\phi|\beta), \tag{11.7}$$

where z_i is the topic assignment of the ith word w_i.

The graphical model for LDA is shown in Fig. 11.4. The shaded variables indicate the observed variables, whereas the unshaded ones indicate the unobserved/latent variables. In the case of LDA, words are the only observed variables.

The objective of LDA inference is to estimate the word distribution for each topic $\phi^{(t)} = p(w|z = t)$, and the topic distribution for each document $\theta^{(d)} = p(z)$, given a training corpus and the parameters α, β, and T. The posterior distribution over z for a given document can be calculated by marginalizing over θ and ϕ, using Gibbs sampling. More details of Gibbs sampling for LDA inference can be found in [47].

As an example of the application of topic models to social verticality problems, we present a case study in Sect. 11.4.4 [25]. In this work, analogous to the bag-of-words approach in a text collection, bag-of-nonverbal patterns are defined to represent the group nonverbal behavior, for modeling group conversational patterns. In this context, the documents are the meetings, the topics are the conversational patterns, and the words are low-level nonverbal features, calculated from thin slices of small group meetings.

11.3.3.3 Supervised Approaches

Supervised approaches, including support vector machines, boosting methods, and naive Bayes, are frequently used in tasks like role recognition [6, 15, 19, 50] and dominance estimation [27, 42]. The details of these models are not given in this chapter, as they are well known models. Interested readers may refer to above mentioned references. In this section, we focus on two issues of using supervised models for social verticality problems. The first is on how to formulate the given problem as a supervised learning task, and the second is on how to obtain reliable labels for using during training from noisy and subjective annotations.

Depending on the task, the supervised learning problem can be formulated as a regression problem (e.g. if the leadership score of a participant is in question), as a binary classification problem (e.g. whether the person is dominant or not), or as a multiclass classification problem (e.g. assigning a role to each participant, among multiple role definitions). From the supervised learning point of view, one interesting problem is the estimation of the most dominant (or similarly the least dominant) person in a meeting. The trained supervised model needs to select exactly one participant from among the all participants in the meeting. In [27], the authors employed a binary classification approach to discriminate between the 'most' dominant participants and the rest, in each meeting. They trained a two-class SVM, and for each test participant in a meeting, the SVM scores were calculated with respect to the distance to the class boundary. With this formulation, the participant that has the highest score receives the 'most dominant' label, generating exactly one most dominant person per meeting. An alternative approach would be to define the problem as a regression problem and assign a dominance score to each participant. Then, the participant receiving the highest score could be selected as the most dominant person.

One of the challenges of using supervised models in social verticality problems is the need for a labeled training dataset, as obtaining these labels is not trivial for most of the social behavior estimation tasks, for which there is no "true label". As a result, the labels need to be collected from human annotators. However, when the question at hand is the existence of a social construct, even human judgments can differ, given the fact that a single correct answer does not necessarily exist. To cope with this variability, multiple annotators are used to annotate social behavior data. A common approach is to use majority agreement of annotators as the ground-truth labels. However, majority agreement has its disadvantages. It discards data points for which the annotators do not have an agreement. Furthermore, it weighs each annotator equally, without considering their different levels of expertise. Other than using majority voting, several other approaches in diverse domains are proposed to model multiple human judgments to estimate the underlying true label. In the field of social computing, as the only example so far, Chittaranjan et al. proposed an Expectation-Maximization (EM) based approach that uses annotations, and also the annotator confidences to model the ground truth [14].

11.3.3.4 Temporal Modeling

Instead of modeling a meeting as a whole, modeling the temporal evolution of the interaction in the meeting could reveal further properties of the interaction, enabling a better analysis. Dynamic Bayesian networks, in particular Hidden Markov Models (HMM) and its variants, are the most popular temporal models used in interaction analysis (see Chap. 2).

A straightforward idea is to model each participant in a meeting with one HMM and then to compute all the combinations of interacting states between these chains. However, this approach results in a high number of states, exponential with the number of chains. An alternative approach would be to use coupled HMMs or N-chain coupled HMMs. However both approaches require large number of parameters,

As an alternative to these models, in [4, 7], the *influence model* is presented and used for analyzing the interaction in groups and various social constructs such as roles [15, 41], and dominance [7].

The influence model is proposed as a generative model for describing the connections between many Markov chains. The parametrization of the model allows the representation of the influence of each chain on the others. The advantage of the influence model with respect to HMMs or coupled HMMs is that it models interacting chains while still keeping the model tractable. Figure 11.5 shows the graphical models of coupled HMM and the influence model.

The graphical model for the influence model and for the generalized N-chain coupled HMM are identical, with one very important simplification [7]. In the influence model, the probability of being at state i at time t is approximated by the pairwise conditional probabilities instead of modeling them jointly:

$$P\left(S_t^i | S_{t-1}^1, \ldots, S_{t-1}^N\right) = \sum_j \alpha_{ij} P\left(S_t^i | S_{t-1}^j\right), \tag{11.8}$$

where α_{ij} indicates the influence of chain i on chain j. Although this pairwise modeling limits the capability of the model, it allows tractability and scalability. The details of the model and the EM algorithm for learning influence model parameters can be found in [7].

11.4 Case Studies

This section presents example studies on automatic analysis of social verticality, for four different social constructs: dominance, emergent leadership, roles, and leadership styles.

11.4.1 Dominance Estimation

In this section we report a study that explores ways to combine audio and visual information to estimate the most and least dominant person in small group interactions. More details of this work can be found in [2].

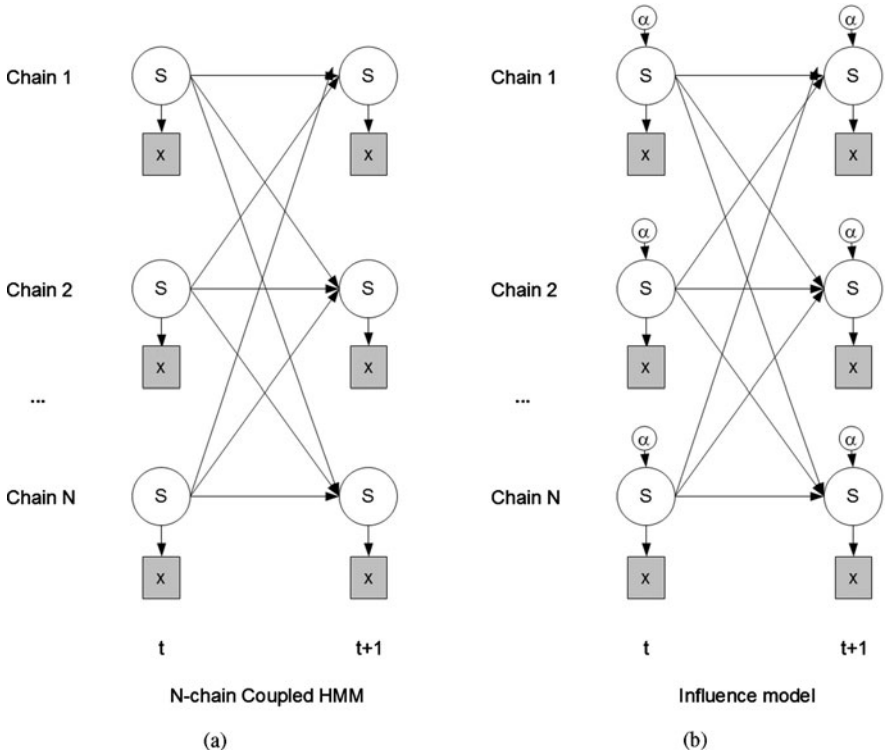


Fig. 11.5 Graphical models for (**a**) coupled HMM, and (**b**) influence model with hidden states

11.4.1.1 Task Definition and Data

Given a meeting, a small group conversation, this study focuses on finding the Most Dominant (MD) and Least Dominant (LD) participants in the group.

The data used in this study are publicly available as the DOME corpus [3]. The DOME corpus contains 125 five-minute meeting segments selected from the Augmented Multi-party Interaction (AMI) corpus [11]. Each meeting has four participants, and is recorded with multiple cameras and microphones. The total length of the DOME corpus corresponds to more than 10 hours of recordings. Each meeting segment in the DOME corpus is annotated by three annotators. The annotators ranked the participants according to their level of perceived dominance. Then the agreement (full and majority agreement on most and least dominant person) between the annotators for each meeting is assessed. Following this procedure, two annotated meeting datasets for each task are obtained (see Table 11.3).

Table 11.3 Number of meetings with full and majority agreement in DOME corpus

	Full	Maj		Full	Maj
Most dominant	67	121	Least dominant	71	117

11.4.1.2 Features and Model

Social psychology research states that dominance is displayed via audio nonverbal cues such as the speaking time, number of turns and interruptions, pitch, as well as visual cues such as visual activity, expressions and gaze [22, 31]. Based on these studies, several audio and visual features can be extracted as descriptors of some of the above cues.

For audio nonverbal features, speaking turn features such as speaking time, number of turns and interruptions are considered. The audio recordings from the close-talk microphones are processed for each participant and their speech activity, in the form of binary speaking status, is extracted. The following speaking turn features are used in this study: Total Speaking Length (TSL), Total Speaking Turns (TST), TST without Short Utterances (TSTwoSU), Total Successful Interruptions (TSI), and Average Speaker Turn Duration (AST).

Visual activity based nonverbal features are extracted from the close-up camera that captures the face and the upper body of each participant. The amount of motion in the skin-colored regions are calculated using compressed domain processing (see Sect. 11.3.2.2), in the form of binary visual activity information for each participant. Visual activity (-V-) equivalents of the above given audio features are extracted as visual nonverbal features.

Furthermore, in addition to the above audio-only and video-only features, a set of multimodal features is defined, which represent the audio-visual (-AV-) activity jointly. The visual activity of the person is measured only while speaking, and audio visual equivalents of the audio-only and video-only features are extracted.

Dominance estimation is performed by a rule-based estimator. The fusion of audio and visual nonverbal features is done via rank and score level multimodal fusion, using the rule-based estimator. The details of these techniques are presented in Sect. 11.3.3.1.

11.4.1.3 Experiments and Results

The experiments are performed on Full and Maj datasets for MD and LD tasks on the DOME corpus (see Table 11.3). The accuracy is calculated as follows: it is assumed that the estimation is correct with weight one, if it matches the agreement. If there is a tie, and one of the tied results is correct, a weight is assigned, which is the reciprocal of the number of ties.

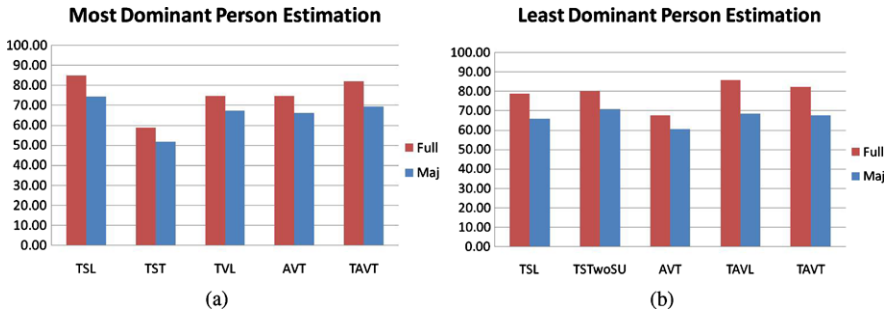


Fig. 11.6 Single feature accuracy for selected audio and visual nonverbal features for (**a**) MD task and (**b**) LD task

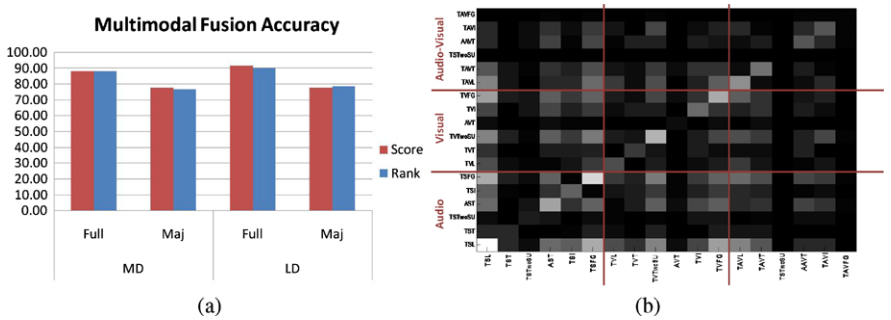


Fig. 11.7 (**a**) Multimodal fusion accuracy for MD and LD tasks. (**b**) Pairwise feature frequencies in best combinations for MD task. *Lighter colors* indicate higher frequency

The classification accuracies for selected single nonverbal features are shown in Fig. 11.6. For the MD task, the best results are obtained with TSL (85.07% and 74.38%) and for the LD task, with Total Audio-Visual Length (TAVL) (85.92%) and TSTwoSU (70.94%), on Full and Maj datasets, respectively.

To find the best combination of nonverbal features in multimodal fusion, an exhaustive search is performed: all feature combinations are evaluated and the best one that combines fewest number of features is reported. The classification accuracies for the best combinations are shown in Fig. 11.7(a). The results show that one can achieve ~3% increase on MD task and ~7% on LD task using rank or score level fusion. It is important to note that there is more than one combination that gives the highest result. Figure 11.7(b) shows the pairwise feature frequencies in best combinations for MD task.

As dominance is displayed multimodally, via audio and visual features, automatic methods should utilize multimodal fusion techniques for estimation of dominance. The results above show that the visual information is complementary to audio, and multimodal fusion is needed to achieve better performance.

11.4.2 Identifying Emergent Leaders

The second study focuses on automatically identifying the emergent leader in small groups. More details of this work can be found in [44].

11.4.2.1 Task Definition and Data

Two main questions are asked in the context of predicting emergent leadership in small groups based on automatic sensing. The first question deals with the existence of a correlation between how the emergent leader is perceived and her/his nonverbal behavior. The second question is whether one can predict emergent leadership using automatically extracted acoustic nonverbal features.

To study emergence of leadership, an audio-visual corpus is collected: The Emergent LEAder data corpus (ELEA) includes audio-visual recordings of groups performing a 'winter survival task' [29], and also questionnaires filled by each group member before and after the interaction. The winter survival task focuses on ranking a list of items in order to survive an airplane crash in winter [29]. The groups are composed of previously unacquainted people. The questionnaires ask participants about themselves and also about the other group members, to evaluate their leadership skills and personality. Several variables are computed from the questionnaires, indicating the perceived leadership, perceived dominance, dominance rank, and perceived competence.

11.4.2.2 Features and Model

The audio recordings of ELEA corpus are collected with a microphone array, which creates automatic speaker segmentations along with the audio recording. This results in a binary segmentation for each participant, indicating the binary speaking status. From this binary segmentation, speaking turn audio features are extracted as audio nonverbal features. The features include the speaking length (TSL), turns (TST and TSTf), average turn duration (TSTD) and interruptions (TSI and TSIf). For turns and interruptions, the filtered versions (TSTf, TSIf) consider only the turns longer than two seconds. Two different definitions of interruptions are used (TSI^1 and TSI^2) (see Sect. 11.3.2.1 and Fig. 11.2) for both the filtered and non-filtered versions. The rule-based estimator, presented in Sect. 11.3.3.1, is used for automatic identification of the emergent leader. The variables from questionnaires were used as a ground truth for evaluation purposes.

11.4.2.3 Experiments and Results

Correlation Between the Questionnaires and the Nonverbal Features Table 11.4 shows Pearson correlation values between questionnaire outputs and nonverbal features. There is a correlation between several nonverbal features and perceived leadership, suggesting that emergent leadership perception has a connection

Table 11.4 Correlation values between variables from questionnaires and nonverbal acoustic features

	Perc. leadership	Perc. dominance	Ranked dominance	Perc. competence
TSL	0.51	0.46	0.49	0.28
TSTD	0.44	0.39	0.40	0.19
TSTf	0.60	0.60	0.53	0.27
TSIf	0.62	0.60	0.54	0.26

Table 11.5 Accuracy (%) of individual features on predicting emergent leadership

	TSL	TSTD	TST	TSTf	TSI^1	$TSIf^1$	TSI^2	$TSIf^2$
Plead	60	70	35	65	50	65	55	70

to the person who talks the most, has more turns, and interrupts the most. Furthermore, several nonverbal features also have correlation with perceived or ranked dominance. The correlations with perceived competence is relatively low.

Automatic Inference Table 11.5 shows the accuracy using single features, where the best accuracy for perceived leadership is achieved using $TSIf^2$ and TSTD with 70%, followed by TSTf and $TSIf^1$ with 65%. The accuracy is calculated as in the previous case study: it is assumed that the estimation is correct with weight one, if it matches the agreement. If there is a tie, and one of the tied results is correct, a weight is assigned, which is the reciprocal of the number of ties.

Score level fusion (see Sect. 11.3.3.1) is applied to combine different acoustic nonverbal features. For the estimation of perceived leadership, a 10% increase in the accuracy is observed, achieving an accuracy of 80%, via the combination of TSTD and TSI features.

This study, summarized from [44], is a first attempt to automatically identify the emergent leader in small groups. Although the collected corpus is currently quite limited, several observations can be made. First there are correlations between the perceived leadership and automatically extracted acoustic nonverbal features. The emergent leader was perceived by his/her peers as a dominant person, who talks the most, and has more turns and interruptions. An accuracy up to 80% is obtained to identify the emergent leader using a combination of nonverbal features.

11.4.3 Recognizing Functional Roles

The third case study attempts to recognize functional roles in meetings using the influence model. More details can be found in [15, 40, 50].

11.4.3.1 Task Definition and Data

Given a small group meeting, the task in this study is to identify the role of each participant. The roles are defined based on the Functional Role Coding Scheme (FRCS), in two complementary areas. The *task area* includes the roles related to the tasks and the expertise of the meeting participants. These include 'orienter', 'giver', 'seeker', and 'follower' roles. The *socio-emotional area* roles, i.e. 'attacker', 'protagonist', 'supporter', and 'neutral', are related to the relationships of the group members.

The Mission Survival Corpus is used as the meeting corpus [40], which includes eight four-people meetings, recorded with microphones and cameras. The annotations for the roles are done by one annotator, by considering the participant's behavior every five seconds. As a result, instead of assigning one role for each participant for the entire meeting, a thin slice based approach is used. This coding scheme assumes that the participants can have different roles throughout the meeting.

11.4.3.2 Features and Model

As features, the authors use automatically extracted speech and visual activity features. The speech recorded from close-talk microphones is automatically segmented for each participant and speaking/non-speaking status is used as speech activity features. The number of simultaneous speakers is also used as a feature. As visual activity features, the amount of fidgeting (i.e. the amount of energy) for hands and body is used [13].

The influence model is proposed as a suitable approach to model the group interaction (see Sect. 11.3.3.4), as it can model complex and highly structured interacting processes. To model a meeting with the influence model, two processes per participant are used: one for the task roles, and another for the socio-emotional roles. The latent states of the models are the role classes.

11.4.3.3 Experiments and Results

The performance of the influence model is compared with two other models: SVM and HMM. For each of the models, the training is performed with half of the available meeting data, using two fold cross-validation. The feature vector for each participant is composed of all extracted audio-visual features. For SVM, the feature vectors of each participant is concatenated and a single feature vector is composed.

The role recognition accuracies for each model is presented in Table 11.6, as reported in [15]. The SVM suffers from the curse of dimensionality and overfitting. The influence model achieves the highest accuracy, as it handles the curse of dimensionality by modeling each participant with different processes. Although the HMM handles the curse of dimensionality using the same approach, as there is no interaction between the processes, the recognition accuracy is lower. Another advantage of using the influence model is its flexibility: it is adaptable to different-sized groups.

Table 11.6 Role recognition accuracies (%) of the Influence model, HMM, and SVM

	Task roles	Social roles	Overall
Influence model	75	75	75
HMM	60	70	65
SVM	–	–	70

11.4.4 Discovering Leadership Styles in Group Conversations

As the last example, we present a study that differs from the previously presented ones, in the sense that it aims to model the group as a whole, instead of modeling individuals. More details of this study can be found in [25].

11.4.4.1 Task Definition and Data

The addressed problem in this study is to automatically discover group conversational patterns from nonverbal features, extracted from brief observations of interaction. Specifically, following the definition in [35], the group conversations can be grouped in three categories: autocratic groups, in which the decisions are determined by the leader; participative groups, in which the leader encourages group discussion and decision making; and free-rein groups, in which the group has complete freedom to decide without leader participation. The study uses a subset of the AMI corpus [11], corresponding to 17 hours of meetings. Part of this subset is annotated by human annotators and used for assessment of the group conversation type.

11.4.4.2 Features and Model

In this study, a novel descriptor of interaction slices—a bag of group nonverbal patterns is described, which captures the behavior of the group as a whole, and its leader's position in the group. The discovery of group interaction patterns is done in an unsupervised way, using principled probabilistic topic modeling.

Analogous to how topics are inferred from a text collection, by representing documents in a corpus as histograms of words, group dynamics can be characterized via bag-of group nonverbal patterns (bag-of-NVPs). The bag-of-NVPs are produced from low-level speaking turn audio nonverbal features, calculated from thin slices of small group meetings. The low-level features include individual and group speaking features such as speaking length, turns, interruptions, backchannels, overlapped/non-overlapped speech, and silence. These low-level features are quantized to generate the bag-of-NVPs. There are two types of bag-of-NVPs: generic group patterns and leadership patterns. Generic group patterns describe the group as a whole without using the identity information. The leadership patterns describe the leader in the group. A diagram showing the features is given in Fig. 11.8(a). Once the bag-of-NVPs are produced, the mining of group patterns is done using

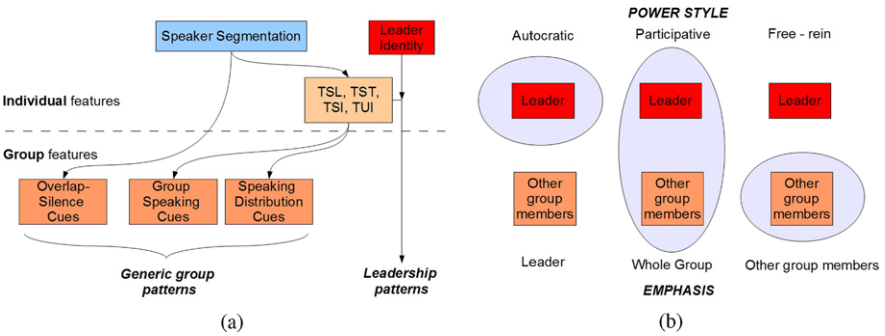


Fig. 11.8 (**a**) Extracted features to characterize individual and group behavior. (**b**) Leadership styles by Lewin et al. [35]

latent Dirichlet allocation topic model (see Sect. 11.3.3.2), to discover topics by considering the co-occurrence of word patterns.

11.4.4.3 Experiments and Results

The experiments are performed on a subset of meetings selected from the AMI corpus [11]. Effect of different time scales and different combinations of bag features are analyzed. The evaluation is done via comparison with human annotations.

On different time scales, the authors observed that the group interactions look more like a monologue at finer time scales (e.g., 1 minute) and like a discussion at coarser time scales (e.g., 5 minute). Also, successful interruptions are not very common at fine time scales.

The LDA based discovery approach is applied to discover three topics. The results on 5-minute scale show that the three discovered topics resemble three classic leadership styles of Lewin et al. [35], as illustrated in Fig. 11.8(b). In comparison to the human annotators, the accuracy of the model for autocratic, participative and free-rein classes are 62.5%, 100%, and 75%, respectively. This suggests that the discovered topics are indeed meaningful. The LDA experiments are repeated for a 2-minute scale as well. The distribution of the topics found with the 2-minute scale is more balanced than the topics at 5-minute scale, indicating that at longer time intervals, the interaction styles are captured more strongly.

11.5 Summary

This chapter focuses on the computational analysis of social verticality. Social verticality refers to the vertical dimension of social interactions, in which the participants of the group position themselves in a hierarchical-like structure. We presented a brief summary of main nonverbal features that humans display and perceive during social interactions that represent social verticality constructs such as dominance,

power, status, and leadership. As the main sources of these nonverbal features are audio and video, we described processing and feature extraction techniques for these modalities. Different inference approaches, such as rule-based, unsupervised, supervised, and temporal are also discussed with examples from the literature. We also present a non-exhaustive survey on the computational approaches for modeling social verticality. In the last section of this chapter we presented four case studies on dominance estimation, identifying emergent leadership, role recognition, and discovering leadership styles in group conversations as examples to the techniques discussed in the chapter.

The future dimensions of this field lie in all the functional blocks that are presented in the chapter, with the inference block being the core challenge. Developments on new sensor technologies will result in better capture of social behavior of humans. On top of this, the current research on tracking human movements should be further extended to cover human behavior in natural settings. Features that better represent nonverbal social behavior should be investigated in close contact with social psychology research. The inference models lie at the core of social behavior analysis. Flexible models that can handle dynamic groups with varying numbers of participants are needed, applicable to different settings to estimate and model social constructs that relate to individuals, as well as to their group behavior.

11.6 Questions

1. What is the difference between verbal and nonverbal communication?
2. What are the differences between power, status, and dominance?
3. What kind of audio nonverbal features can be extracted for social verticality analysis?
4. What kind of visual nonverbal features can be extracted for social verticality analysis?
5. What are the techniques that can be used to fuse different modalities for social verticality analysis?
6. What is thin slice based modeling?
7. Discuss what kind of models can be used in a meeting scenario or how the standard models can be modified when the number of participants in a group vary, i.e. the dataset contains meetings with different number of participants.
8. Explain the differences between the influence model and coupled HMM.

11.7 Glossary

- *Emergent leadership*: The leader who arises from an interacting group and has a base of power arising from followers rather than from a higher authority.
- *Influence model*: A representation to model the dynamics between interacting processes.

- *Thin slice*: The smallest segment from which, when exposed, humans can correctly predict behavioral outcomes with high accuracy by interpreting nonverbal cues.
- *Topic models*: A statistical model for discovering the hidden topics that occur in a collection of documents.
- *Visual dominance ratio*: Looking-while-speaking to looking-while-listening ratio.

Acknowledgements This work is supported by the EU FP7 Marie Curie Intra-European Fellowship project "Automatic Analysis of Group Conversations via Visual Cues in nonverbal Communication" (NOVICOM), and by the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (IM2) and by the Sinergia project on "Sensing and Analysing Organizational Nonverbal Behaviour" (SONVB). The authors would like to thank Dinesh Jayagopi, Dairazalia Sanchez-Cortes, and Gokul Chittaranjan for their contributions to several studies presented in this chapter.

References

1. Ambady, N., Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychol. Bull. **111**, 256–274 (1992)
2. Aran, O., Gatica-Perez, D.: Fusing audio-visual nonverbal cues to detect dominant people in small group conversations. In: 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey (2010)
3. Aran, O., Hung, H., Gatica-Perez, D.: A multimodal corpus for studying dominance in small group conversations. In: LREC Workshop on Multimodal Corpora, Malta (LREC MMC'10) (2010)
4. Asavathiratham, C., Roy, S., Lesieutre, B., Verghese, G.: The influence model. IEEE Control Syst. Mag. **21**, 52–64 (2001)
5. Ba, S.O., Odobez, J.-M.: Multiperson visual focus of attention from head pose and meeting contextual cues. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 101–116 (2011)
6. Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S.: The rules behind roles: Identifying speaker role in radio broadcasts. In: 17th National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pp. 679–684. AAAI, Washington (2000)
7. Basu, S., Choudhury, T., Clarkson, B., Pentland, A.: Learning human interactions with the influence model. Technical report, MIT Media Lab, Cambridge, MA (2001)
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
9. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 257–267 (2001)
10. Campbell, N., Douxchamps, D.: Processing image and audio information for recognising discourse participation status through features of face and voice. In: INTERSPEECH 2007, pp. 730–733 (2007)
11. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The AMI meeting corpus: A pre-announcement. In: Workshop Mach. Learn. for Multimodal Interaction (MLMI'05), Edinburgh, UK, pp. 28–39 (2005)
12. Charfuelan, M., Schröder, M., Steiner, I.: Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings. In: Interspeech 2010, Makuhari, Japan, Sept. (2010)

13. Chippendale, P.: Towards automatic body language annotation. In: 7th International Conference on Automatic Face and Gesture Recognition (FG '06), Washington, DC (2006)
14. Chittaranjan, G., Aran, O., Gatica-Perez, D.: Exploiting observers' judgments for multimodal nonverbal group interaction analysis. In: 9th IEEE Conference on Automatic Face and Gesture Recognition, Santa Barbara, CA (2011)
15. Dong, W., Lepri, B., Cappelletti, A., Pentland, A.S., Pianesi, F., Zancanaro, M.: Using the influence model to recognize functional roles in meetings. In: 9th International Conference on Multimodal Interfaces (ICMI'07), pp. 271–278 (2007)
16. Dovidio, J.F., Ellyson, S.L.: Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. Soc. Psychol. Q. **45**(2), 106–113 (1982)
17. Dunbar, N.E., Burgoon, J.K.: Perceptions of power and interactional dominance in interpersonal relationships. J. Soc. Pers. Relatsh. **22**(2), 207–233 (2005)
18. Funder, D.C.: Personality. Annu. Rev. Psychol. **52**, 197–221 (2001)
19. Garg, N.P., Favre, S., Salamin, H., Hakkani Tür, D., Vinciarelli, A.: Role recognition for meeting participants: an approach based on lexical information and social network analysis. In: 16th ACM International Conference on Multimedia (MM'08), pp. 693–696 (2008)
20. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: a review. Image Vis. Comput. **27**(12), 1775–1787 (2009)
21. Gorga, S., Otsuka, K.: Conversation scene analysis based on dynamic bayesian network and image-based gaze detection. In: ICMI-MLMI 2010 (2010)
22. Hall, J.A., Coats, E.J., Smith LeBeau, L.: Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. Psychol. Bull. **1313**(6), 898–924 (2005)
23. Hung, H., Jayagopi, D.B., Ba, S., Gatica-Perez, D., Odobez, J.-M.: Investigating automatic dominance estimation in groups from visual attention and speaking activity. In: Int. Conf. on Multimodal Interfaces (ICMI), Chania, Greece (2008)
24. Hung, H., Huang, Y., Friedland, G., Gatica-Perez, D.: Estimating dominance in multi-party meetings using speaker diarization. IEEE Trans. Audio Speech Lang. Process. **19**(4), 847–860 (2011)
25. Jayagopi, D.B., Gatica-Perez, D.: Mining group nonverbal conversational patterns using probabilistic topic models. IEEE Trans. Multimed. (2010)
26. Jayagopi, D.B., Ba, S., Odobez, J.-M., Gatica-Perez, D.: Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In: Int. Conf. on Multimodal Interfaces (ICMI), Special Session on Social Signal Processing, Chania, Greece (2008)
27. Jayagopi, D.B., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations from nonverbal activity cues. IEEE Trans. Audio Speech Lang. Process. **17**(3), 501–513 (2009). Special Issue on Multimodal Processing for Speech-based Interactions
28. Kalma, A.K., Visser, L., Peeters, A.: Sociable and aggressive dominance: Personality differences in leadership style? Leadersh. Q. **4**(1), 45–64 (1993)
29. Kickul, J., Neuman, G.: Emergent leadership behaviours: The function of personality and cognitive ability in determining teamwork performance and KSAs. J. Bus. Psychol. **15**(1) (2000)
30. Kim, T., Chang, A., Pentland, A.: Meeting mediator: Enhancing group collaboration with sociometric feedback. In: ACM Conference on Computer Supported Collaborative Work, San Diego, CA, pp. 457–466 (2008)
31. Knapp, M.L., Hall, J.A.: Nonverbal Communication in Human Interaction, 7th edn. Wadsworth, Belmont (2009)
32. Kumano, S., Otsuka, K., Mikami, D., Yamato, J.: Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings. In: 11th International Conference on Multimodal interfaces (ICMI'09), ICMI-MLMI '09, pp. 99–106 (2009)
33. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley, New York (2004)
34. Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. IEEE Commun. Mag. **48**, 140–150 (2010)

35. Lewin, K., Lippit, R., White, R.K.: Patterns of aggressive behavior in experimentally created social climates. J. Soc. Psychol. **10**, 271–301 (1939)
36. Otsuka, K., Yamato, J., Takemae, Y., Murase, H.: Conversation scene analysis with dynamic Bayesian network based on visual head tracking. In: ICME 2006 (2006)
37. Otsuka, K., Sawada, H., Yamato, J.: Automatic inference of cross-modal nonverbal interactions in multiparty conversations. In: ACM 9th Int. Conf. Multimodal Interfaces (ICMI2007), pp. 255–262 (2007)
38. Pentland, A.: Socially aware computation and communication. Computer **38**(3), 33–40 (2005)
39. Pentland, A.: Honest Signals: How They Shape Our World. MIT Press, Cambridge (2008)
40. Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A.: A multimodal annotated corpus of consensus decision making meetings. Lang. Resour. Eval. **41**, 409–429 (2007)
41. Raducanu, B., Gatica-Perez, D.: Inferring competitive role patterns in reality TV show through nonverbal analysis. Multimed. Tools Appl. 1–20 (2010)
42. Rienks, R.J., Heylen, D.: Automatic dominance detection in meetings using easily detectable features. In: Workshop Mach. Learn. for Multimodal Interaction (MLMI'05), Edinburgh, UK (2005)
43. Salamin, H., Favre, S., Vinciarelli, A.: Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. IEEE Trans. Multimed. **11**(7), 1373–1380 (2009)
44. Sanchez-Cortes, D., Aran, O., Schmid-Mast, M., Gatica-Perez, D.: Identifying emergent leadership in small groups using nonverbal communicative cues. In: 12th International Conference on Multimodal Interfaces (ICMI'10), Beijing, China (2010)
45. Schmid-Mast, M.: Dominance as expressed and inferred through speaking time: A meta-analysis. Hum. Commun. Res. **28**(3), 420–450 (2002)
46. Stein, R.T.: Identifying emergent leaders from verbal and nonverbal communications. J. Pers. Soc. Psychol. **32**(1), 125–135 (1975)
47. Steyvers, M., Griffiths, T.: Probabilistic Topic Models. Erlbaum, Hillsdale (2007)
48. Varni, G., Volpe, G., Camurri, A.: A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. IEEE Trans. Multimed. **12**(6), 576–590 (2010)
49. Yeo, C., Ahammad, P., Ramchandran, K., Sastry, S.S.: High-speed action recognition and localization in compressed domain videos. IEEE Trans. Circuits Syst. Video Technol. **18**(8), 1006–1015 (2008)
50. Zancanaro, M., Lepri, B., Pianesi, F.: Automatic detection of group functional roles in face to face interactions. In: 8th International Conference on Multimodal Interfaces (ICMI'06), pp. 28–34 (2006)

Part IV
Selected Applications

Chapter 12
Activity Monitoring Systems in Health Care

Ben Kröse, Tim van Oosterhout, and Tim van Kasteren

12.1 Introduction

The current quality of medicine and living conditions, combined with the decreasing number of births makes the average age of the world population increase at a rapid pace. Asia and Europe are the two regions where a significant number of countries face severe population aging in the near future. As a consequence of this, the cost of health care is expected to grow enormously in the coming years. To keep these costs limited, we need better possibilities for self management and independent aging. One of these solutions is to use technologies that assist people to be independent. The assistive technology can offer physical help, cognitive help or social help.

Assistive health systems need accurate assessments of the health and wellbeing of a patient. For this, sensing systems are needed that monitor the patient. A first class of monitoring systems monitor vital signs directly using special sensors (for example heart rate sensors or blood sugar sensors). A second class of systems monitor the health state indirectly, by measuring the activities performed by the patient, using sensors either mounted on the patient or mounted in the environment. In this chapter we focus on the second class, monitoring the activity behavior of the patient.

B. Kröse (✉)
University of Amsterdam, Amsterdam, The Netherlands
e-mail: b.j.a.krose@uva.nl

B. Kröse · T. van Oosterhout
Amsterdam University of Applied Science, Amsterdam, The Netherlands

T. van Oosterhout
e-mail: T.J.M.van.Oosterhout@hva.nl

T. van Kasteren
Boğaziçi University, Bebek, Istanbul, Turkey
e-mail: tim0306@gmail.com

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_12,

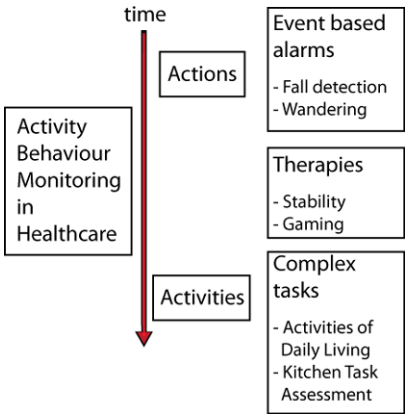


Fig. 12.1 In activity behavior often a distinction between actions and activities is made. Also, in healthcare monitoring, simple actions, as well as complicated activities are worth monitoring

Human activities may have complex forms. In this chapter we make a distinction between 'action' and 'activity', as is used in surveys on behavior analysis from video [80]. An action is a simple human motion pattern usually in the order of a couple of seconds. An activity is a more complex motion pattern, typically of longer duration. For example, picking up a glass is considered an action, but having lunch is an activity. In behavior monitoring systems for health care, both actions and activities are indicative of the health of the patients. Figure 12.1 gives an overview of the different types of activities relevant for health care and their complexity.

Action monitoring is usually related to the safety of the patient and often causes an intervention in the form of an alarm. Currently, many systems are designed that focus on fall detection, either in hospital environments [13] or home environments. Also unattended wandering is a serious problem, particularly for cognitively impaired older adults. Elopement from locked dementia units is a major safety concern in long-term care facilities. An automatic system for detecting such wandering actions in a home setting is a valuable tool to assist informal caregivers.

More elaborate actions and activities are monitored in systems for automatic rehabilitation or therapy. With the introduction of minimally invasive surgery (MIS), the recovery time of patients has been shortened significantly. This has led to a shift of post-operative care from hospital to home environment, where the latter can be enhanced with monitoring systems. In games and computer driven physical exercises, more complicated gestures and activities need to be measured. Similarly, cognitive training requires monitoring of more complex activities. An example is The Kitchen Task Assessment (KTA), which is a functional measure that records the level of cognitive support required by a person with Senile Dementia of the Alzheimer's Type (SDAT) to complete a cooking task successfully.

A large group of monitoring systems focus on recognizing activities of daily living (ADL); a more complex set of activities performed on a daily basis, such as sleeping, toileting and cooking. The list of ADL was set up by Katz [48], and registering how well ADL are performed over time is a commonly used method in healthcare for monitoring the wellbeing of a person, in particular of elderly. Healthcare professionals measure ADL manually by visiting the home of an elderly per-

son and observing them in performing activities. Measuring of ADL is proposed by Garrod et al. [30] for the assessments of chronic obstructive pulmonary disease (COPD). ADL were also studied by Kurz et al. [52] to make a categorization of dementia patients. Irregular sleep patterns, changes in the frequency of toilet use and an increase in the duration of time it takes to complete an ADL are all important indicators of physical and cognitive health disorders [35, 76]. When applied on a large scale, it is expected that the wealth of information will be extremely useful to evidence based nursing, a form of nursing relying on scientific data. It is expected that such an approach makes it possible to treat certain diseases proactively, before any real damage is done [23].

Virone et al. [88] measure activities of a longer duration than ADL. In their study they monitor residents' cyclic physical activity inside a home environment using wireless passive sensors. The so-called "circadian activity rhythms" (or CARs) describe the measurement of this in-home rhythmic behavioral activity that the resident engages in the habitat. CARs are influenced by social rhythms, but also interact with the biological rhythm of the person. Deviations indicating anomalies were detected, and seemed to be correlated with observations by professional caregivers about the monitored residents.

This chapter focuses on systems for the automatic monitoring, classification and detection of the activities relevant for healthcare. In Sect. 12.2 we describe the sensors that are currently used, both in commercial systems and in research. In Sect. 12.3 we focus on the recognition of simple actions, for applications like alarm and therapy. In Sect. 12.4 we describe the recognition of more complex activities. Finally we give some insights in the acceptance and privacy issues of such systems.

12.2 Sensing Systems

The current state of sensor, processing and communication technologies, combined with relatively low priced hardware make it possible to equip a living environment with systems of communicating sensor devices. In Fig. 12.2 a setup is sketched of a home equipped with many types of sensing and communication devices, connected to the outside world with a broadband Internet connection.

The nature of the sensors is an important aspect for the acceptance and for the performance of an activity recognition system. Because sensing takes place inside someone's private house, it is important to evaluate how intrusive the user experiences the sensors. For example, a camera is considered as an intrusive device by many users. Also, some sensors need to be worn on the body, which might be considered inconvenient by the user. Therefore often monitoring systems are proposed that consist of networks of simple (less-intrusive) sensors mounted in the home. In the following we present some of the sensing systems used in health care activity monitoring. Figure 12.3 gives an overview of the different types of sensing systems.

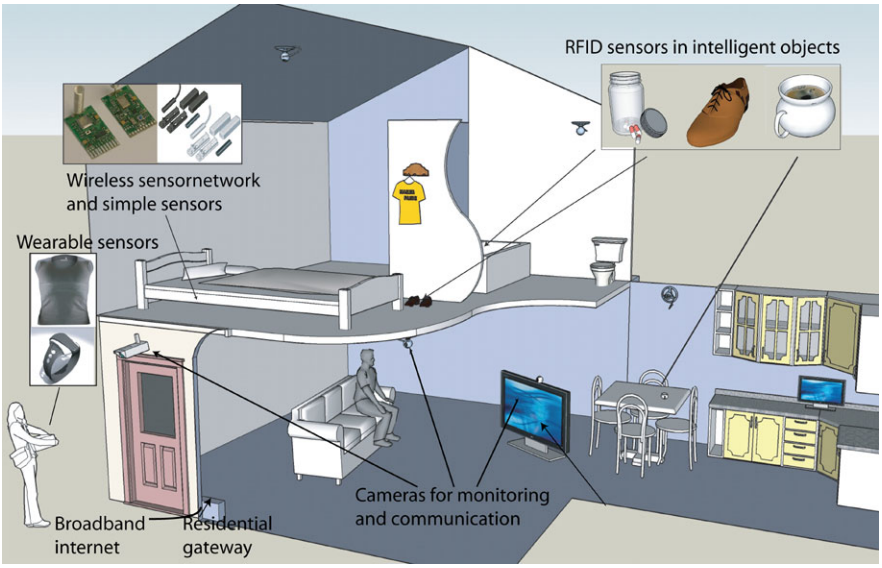


Fig. 12.2 Many sensors can be used for health monitoring and communication

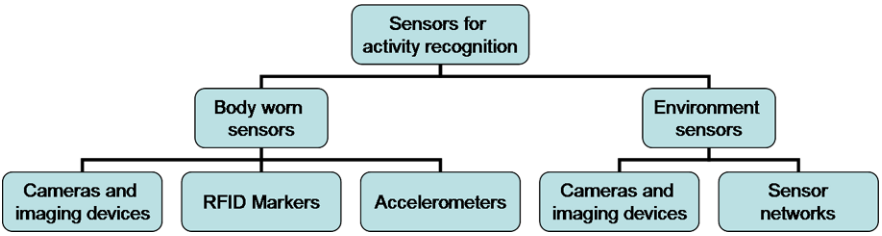


Fig. 12.3 Overview of the types of sensors for activity monitoring

12.2.1 Body Worn Sensors

Sensors that are worn by the user can be used for activity recognition very well. Commonly used sensors are radio frequency identification (RFID) tags, but also accelerometers, tilt meters, gyroscopes, microphones and cameras are used.

RFID is a technology for reading information from a distance, from the so-called RFID-tags. Passive tags extract energy from the radio frequency signal emitted by an RFID reader and use that energy to send a stream of information back. The amount and contents of information can vary, but usually it contains at least a unique identification string. Active tags are equipped with a battery used as a source of power for the tag's circuitry and antenna, which makes it possible to read the tags from a much larger distance than passive tags [22]. Traditionally, RFID tags in healthcare are mainly found in hospitals and nursing homes for tracking equipment and patients [70]. For these systems, the patients have to wear a bracelet with the RFID tag

Fig. 12.4 Three commercial wearable devices. From *left* to *right*: Fall detector (detecting relative angle), elopement detector (based on radio frequency), elopement detector (based on infrared sensing)

and can be identified when they are in the vicinity of a reader. When an RFID tag is given sensing capabilities, the line between RFID and sensor networks becomes blurred. Many active and semi-active tags have incorporated sensors into their design, allowing them to take sensor readings and transmit them to a reader at a later time. They are not quite sensor network nodes, because they lack the capacity to communicate with one another through a cooperatively formed ad hoc network, but they are beyond simple RFID storage tags [39]. In this way, RFID is converging with sensor networking technology. Intel used RFID technology to develop a sensing method used specifically for activity recognition. Their product, named the iBracelet is a RFID reader in the form of a bracelet which the user wears on a single hand or on both hands. By tagging a large number of objects in the house with passive RFID tags, the iBracelet is able to observe which objects the users are holding in their hands. Objects are very indicative of the action a user is performing, therefore making activity recognition possible [27].

For activity recognition, accelerometers that are worn by the human are used frequently. Sometimes they are combined with RFID readers, to measure both the objects used and the person's movement. This way, both sensors can compensate for each other's shortcomings, resulting in better activity recognition performance [42, 75]. A combination with wearable cameras has also been reported [20]. Currently, mobile phones are used for processing the sensor data [34].

Wearable accelerometers are also used for fall detection [18]. Most of the commercial wearable fall detectors use accelerometers or inclinometers. Some examples are shown in Fig. 12.4.

12.2.2 Wireless Sensor Networks

Although body worn sensors are powerful devices for activity measurement, the disadvantage is that the user has to carry them around, or wear the clothes they are inserted in, all the time. There are quite a number of situations where this is not the case, for example people with dementia forgetting them, or users that for example

do not carry the device when they go to the toilet at night. Environment mounted sensors come in two flavors: (wireless) networks of simple sensors or cameras.

A wireless sensor network consists of a collection of small network nodes. Each node is capable of performing some processing, gathering sensor measurements and communicating wirelessly with other nodes in the network [73]. Nodes are designed to be as small as possible and therefore usually have a working memory of only several kilobytes. Specifically engineered operating systems have been developed for dealing with such conditions, such as TinyOS [38]. Although the term 'wireless sensor network' is broadly applicable, here we use the term to describe a network of simple sensors that give binary output and which are installed in fixed locations.

The nodes generally run on batteries and therefore a lot of research is devoted to energy efficiency. The communication between these nodes typically requires little bandwidth and is relatively insensitive to latency, so that energy efficient communication protocols are possible [82]. It is also possible to save power by shutting down parts of the node when these are not in use [71]. The use of ad hoc routing protocols allows the nodes to dynamically form a temporary network without any pre-existing network infrastructure or centralized administration [16]. Such routing protocols also allow further power saving schemes, by shutting down nodes strategically and avoiding them in the network route [100].

A large variety of simple sensors can be incorporated in the network nodes. Since these nodes are installed in fixed locations they are typically used in a house setting or in offices. Sensors used include: contact switches for open-close states of doors and cupboards; pressure mats to measure sitting on a couch or lying in bed; mercury contacts for the movement of objects such as drawers; passive infrared sensors to detect motion in a specific area; float sensors to measure the toilet being flushed; temperature sensors to measure when the stove is used; humidity sensors to measure when the shower is used [84, 97] and accelerometers to detect when a large object is used [58, 77]. Basically any kind of sensor can be combined with a network node and its output is often converted to a binary format. For example, in work by Fogarty et al. microphones were attached to water pipes in the basement to record whether water was flowing. Such an approach allows easy installation by avoiding the need for plumbing [28].

Wireless sensor networks can be used to determine whether an object is used. However, in contrast to RFID tags, it is difficult to equip small objects, such as a tooth brush or a dinner plate, with a sensing node. This limits the observation abilities of these networks and results in more ambiguous sensor readings. For example, in a house setting, the network can be used to observe that a cupboard is opened, but not to observe which item is taken from the cupboard.

Nonetheless, the use of wireless sensor networks in a home setting offers many advantages compared to other sensing modalities. First, the majority of the sensors can be installed out of sight of users, therefore limiting the intrusiveness of the sensors. Second, the data recorded are anonymous and contain very limited privacy-sensitive information. Third, installation of the sensors can be done quickly without any need for the installation of power and network cables.

12.2.3 Visual Sensors

Video cameras in health care are currently mainly applied for remote visual monitoring or communication with a physician or care giver. Cameras provide rich data that are very informative of the activities. However, automatic activity recognition from video data is hard. There are many surveys available discussing action and activity recognition from image data [1, 31, 62, 80]. Depending on the application, cameras may need to give the full pose information, only the location of the patient, or just a general motion picture. For coverage of the entire living space, multiple cameras and fish eye lens cameras are needed. In work by Duong et al. [26] multiple cameras installed in the corners of the room observe a person performing activities. The authors only use the position of the person as an indicator of its activities. A similar approach uses a camera with a top down viewpoint, making it easier to divide the image into squared regions of interest and to detect the location of a person.

Instead of using location, the object a person is using is also a good indicator of the action a person performs. Typically a single camera is focused on a particular area of interest where activities are performed, such as the sink in a bathroom or the kitchen cooking area. In the work by Wu et al. [98] image data are processed to detect which objects a person uses. The detected objects are used to recognize activities such as making tea or taking medicine. Messing et al. [60] use various salient visual features extracted from the image and calculate the velocity of these features to track the movement and position of the hands. These data allow them to distinguish between actions such as drinking from a cup and pealing a banana.

Some approaches for activity recognition are based on 2.5D or 3D data. This can be derived from multiple cameras, stereo cameras or time-of-flight (TOF) cameras [33]. The first can generate full 3D data providing a 360° view of the scene, the accuracy and density of which depends on the number of cameras used and their relative placement, while the latter two can only generate a range image (sometimes called 2.5D). When TOF cameras are used, no color information is available. In all three cases however, shape information is obtained which requires different processing from the standard 2D color images, but enables some features to be more easily detected. For example, foreground detection may be made more robust if apart from the color information also range information is used [33]. Additional methods can be applied to identify the foreground object as a human being, for instance by skin-tone recognition [53], head detection [69, 101] or pose estimation [24]. In Fig. 12.5 an example is shown from a fall, detected with a stereo system in our lab. Since stereo and TOF cameras only have one viewpoint, occlusions are as problematic in these methods as they are in 2D methods [33, 37]. These difficulties can be overcome by using any of the well known remedies from the 2D domain. Voxel methods based on multiple cameras are less sensitive to occlusions, especially as the number of cameras increases and if the cameras are positioned correctly, as one or more of the other views can make up for the occlusion in another viewpoint [9].

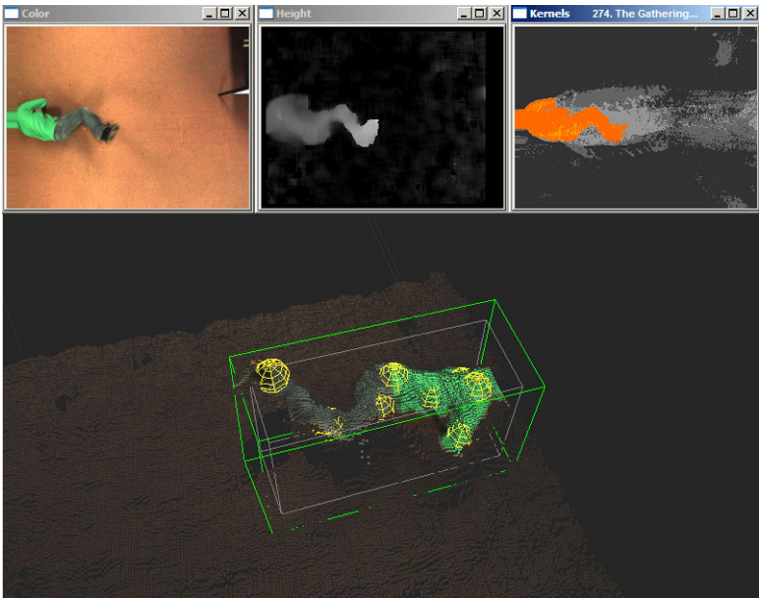


Fig. 12.5 Results from fall detection in our lab, using a stereo camera

12.3 Monitoring of Simple Actions

In this section we describe work on the monitoring of simple actions. These are motion patterns that typically last a couple of seconds and can be of repetitive nature. In health care applications the recognition of such simple motion patterns is often associated with alarm situations. One of the most relevant examples is fall detection: falls and their consequences are amongst the most common health problems for people of advanced age [67]. Another example that is highly relevant is wandering behavior, in particular elopement of older adults with dementia. An alarm has to be given if people move into regions where they are not allowed. Another category of systems that monitor simple actions is that of therapy and rehabilitation. In this field there is an increasing interest to combine action recognition with game elements for therapy.

12.3.1 Fall Detection

Much research on fall detection has been carried out on systems using wearable accelerometers [18, 47]. The classification of the (noisy) data is essential in order to prevent false alarms. Karantonis et al. [47] use a rule based system in which the average signal magnitude is thresholded to classify a fall. A similar method was used by Bourke et al. [15], which employed wearable gyroscopes instead of

accelerometers. Commercial fall detection systems are mostly based on these principles.

The problem with wearable systems is that users forget to put them on, or place them incorrectly on the body, which causes false alarms. Therefore, environment sensors are a popular topic of research. Some systems use audio obtained by microphones as input, but most of the research is carried out using cameras. A number of methods are based on the assumption that a fall results in inactivity, while others determine the dynamics of the observed person to classify a fall.

With the inactivity measure, not the fall itself, but the result of the fall is being looked at. Inactivity in itself is relatively easy to determine. The problem then does not lie in determining inactivity, but in determining 'irregular' inactivity. Approaches have been presented to learn the difference between 'regular' (reading a book for example) or 'irregular' (the result of a fall) inactivity [66, 94], or these can be manually inserted into the system in form of rules [78]. This way, the system can distinguish areas where inactivity is normal, like lying on a sofa or bed, and where it is not normal, like lying on the floor. Yet these methods are sensitive to light fluctuations. Recently, work is being done for fall detection on the 3D modeling of people in a room [7], the use of dynamics [29, 72] and fusion of visual input with audio data [78].

Without estimating the pose, an observed person can be judged to be standing upright or lying down by considering the distance of their centroid to the ground plane. A fall can be concluded if the centroid goes below a certain threshold [17, 33, 46]. More advanced methods use the vertical volume distribution [8], or the principal component of the point or voxel distribution and determine its angle with the ground plane [37]. The principal component can be combined with centroid height to create absolute and in-between states [7]. Instead of analysing the full pose, the focus can be directed at certain body parts. For instance the velocity and direction of the head movement can indicate a fall [69, 101]. Furthermore, instead of the principal axis of the entire silhouette, the estimation of pose enables the judgment of the more salient angle of the torso [45].

The certainty of classifying a potential fall can be increased by looking for a period of inactivity following the event [7, 33, 45]. To prevent false alarms for locations where inactivity is normal, such as a couch or a bed, such a method should be equipped with a notion of where these inactivity regions are and refrain from detecting falls there [66]. Location-based activity histograms can provide further context to evaluate whether or not a certain duration of inactivity in a particular location is normal [45]. Table 12.1 gives an overview on the different methods and sensing systems used for fall detection.

Systems for fall detection are essential in ambient assisted living facilities, and such systems are most desired by elderly who live on their own. However, prevention of a fall is even a larger challenge. Therefore it is important to assess the stability of the user, and coach him or her to keep the stability intact. In Sect. 12.3.3 we will briefly discuss technologies for this.

Table 12.1 Comparison of fall detection methods

Reference	Sensors	Recognition method
Zhang et al. [99]	Single far-field microphone	SVM based on GMM supervectors
Karantonis et al. [47]	Triaxial accelerometer	Thresholding on energy
Bourke et al. [15]	Bi-axial gyroscope sensor	Threshold on acceleration
Alemdar et al. [3]	Accelerometers and camera	Thresholding of features
Nait-Charif et al. [66]	Overhead wide angle camera	Spatial context and tracking
Jansen et al. [46]	TOF camera	Centroid trajectory
Diraco et al. [24]	TOF camera	Pose classification
Yu et al. [101]	Stereo cameras	Head motion
Hazelhoff et al. [37]	Stereo cameras	Principal component orientation
Auvinet et al. [8]	Multiple cameras	Vertical volume distribution ratio
Anderson et al. [7]	Multiple cameras	Fuzzy logic on centroid and axis
Sixsmith and Johnson [72]	Infrared heat camera	Downward motion
Fu et al. [29]	Temporal contrast camera	Motion event rate and height

12.3.2 Wandering and Elopement

Wandering is a commonly observed behavior among older adults with Alzheimer's disease (AD) and other types of dementia. When wandering around becomes wandering away, older adults with dementia are at a high risk of injury. Commercial systems for indoor are available based on radio frequency (RF) or infrared based wearable devices. For outdoor, GPS based systems are available. However, the problem is that people with dementia often forget to take these devices with them. Ambient sensors and cameras have been proposed to track people and warn the caregivers if they move into non-allowed areas. For example, ultrasonic sensors and pressure mats have been used by Biswas et al. [13]. Tracking on the basis of a distributed camera system is described by Chen et al. [19]. Tagless tracking may be a problem if many users are in the building, and tagging may be better solution.

12.3.3 Prevention and Therapy

The human motion pattern can also be used as a diagnostic tool. For example, gait characteristics are reported to be correlated with the physical condition of elderly, and a change in the gait profile over time may also indicate that a person is more at risk of falling. In [91] the gait is quantified by measuring step time and step length using a voxel representation derived from two cameras.

The 'sit to stand' behavior (measuring how people get out of a chair) is used by Allin et al. [5] as an indication of balance. The user is observed with a number of cameras and 3D features are derived from silhouettes. The same behavior was

Table 12.2 Different approaches for pose and gait monitoring

Reference	Sensors	Recognition method
Wang et al. [91]	Cameras	Voxel reconstruction and step time estimation
Allin et al. [5]	Cameras	Blob detection and 3D features
Ali et al. [4]	Accelerometer	Spectral clustering in subspace of motion features
Wilson et al. [96]	Accelerometer and gyro	Visualization of 3D arm pose
Hamel et al. [36]	Oximeter, accelerometers, instrumented soles, respiratory belt	Visualization of body angles, weight-bearing, respiration

studied by Kerr et al. [51], which used accelerometers mounted on the trunk and the knee. A more general study on transition of activities was studied by Ali et al. [4]. The authors use the 'e-AR' (Ear worm activity recognition) sensor that generates a feature vector derived from accelerometer data, and study the transitions between the classes that are found with a Bayesian classifier. The transitions are found by projection of the sensor data on a low-dimensional manifold using the Isomap algorithm and applying a graph clustering method.

Another important area for activity monitoring is post-operational care and rehabilitation. A body sensor network in the form of an ear-mounted device has been applied by Aziz et al. [10] to monitor the activities of patients recovering from abdominal surgery. Features derived from accelerometers and heart rate were used for this. Body area networks were also used in stroke rehabilitation systems [96] and for tele-rehabilitation for geriatric patients [36]. In Table 12.2 an overview on different methods is given.

12.4 Recognition of More Complex Activities

Activities consist of a complex combination of actions and systems that recognize activities from sensor data are now an active topic of research. Some systems are dedicated to recognizing a limited number of activities such as toileting and sleeping behavior [43]. But the majority of work focuses on the recognition of a large variety of ADL such as preparing breakfast, washing dishes and other kitchen activities [25, 59, 84] or ironing, vacuuming and other housekeeping activities [75].

The recognition methods may be divided into two major streams: logically based theories [14, 50] and probabilistic methods [26, 83, 95]. Also the sensing modalities vary. For the analysis of a task in the kitchen, sometimes cameras are used [61] or tags on objects combined with cupboard sensors and pressure mats are employed [14, 41].

In our group we performed activity recognition in three different houses [87]. We mounted wireless sensor networks with simple sensors like reed switches to measure open/close states of doors and cupboards; pressure mats to measure sitting

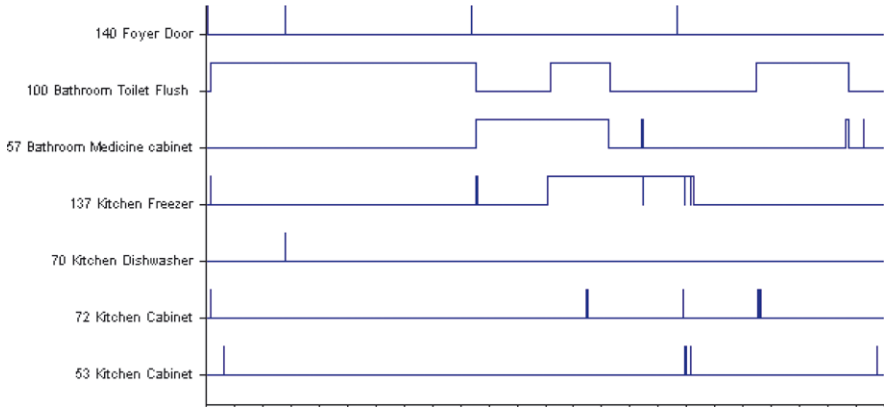


Fig. 12.6 A typical pattern of the sensors from one of our test sites

on a couch or lying in bed; mercury contacts for movement of objects (e.g., drawers); passive infrared (PIR) sensors to detect motion in a specific area; float sensors to measure the toilet being flushed; temperature sensors to measure the use of the stove or shower. This gives a sequence of binary sensor data $\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$ as depicted in Fig. 12.6. In our activity recognition problem, we wish to infer the corresponding sequence of class labels $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$, where a class label indicates the activity (sleeping, toileting, etc.).

In a series of experiments we compared generative models (HMM) with a discriminative model (CRF). Generative models, like hidden Markov models and dynamic Bayesian networks, deal with this problem by explicitly modeling the relations between the observations and the class labels. More specifically, in generative models we express the dependencies among variables ($\mathbf{x}$ and $\mathbf{y}$). For example, the Markov assumption states that the current state y_t depends only on the previous state y_{t-1}. As a result we can express our belief in y_t based only on y_{t-1} and ignore all the other variables (i.e. x_t, y_{t-2}, etc.) These dependencies, or rather the independence assumptions with respect to the other variables, therefore, greatly reduce the number of parameters that specify the model [11]. However, a violation of dependencies, in the case there do exist dependencies in the actual data that we do not model, can strongly affect the performance of the model [81].

Discriminative models, on the other hand, are more robust in dealing with violations of dependency assumptions. The idea is that since observations are always given during an inference, there is no need to model them explicitly. Instead, discriminative approaches directly model the discriminative boundary between the different class labels [11]. The advantage of these approaches is that we can incorporate all sorts of rich overlapping features and the model will find a set of parameters for discriminating the classes even if any there exist violations of independence assumptions.

Conditional random fields (CRF) are temporal discriminative probabilistic models that have this property. They were first applied in text recognition problems

where features such as capitalization of a word and the presence of particular suffixes significantly improved performance. Since then they have been applied to a variety of domains such as gesture recognition [63] and the activity recognition of robots in a game setting [81]. They have also been applied to human activity recognition from video, in which primitive actions such as 'go from fridge to stove' are recognized in a lab-like kitchen setup [79]. In an outdoor setting, CRFs have been applied to activity recognition from GPS data, where activities such as 'going to work' and 'visiting a friend' were distinguished [54].

In our experiment, the performance measure was computed on the basis of the classification of ten different activities. Our experiments show three things. First, that CRFs are more sensitive to overfitting on a dominant class than HMMs. Second, that the use of raw sensor data gives bad results and that a preprocessing is needed; in our case using the change points in the sensor values give the best results. And third, that differences in the layout of houses and the way a dataset was annotated can greatly affect the performance in activity recognition. Recently we have focused on semi-hidden Markov models [86] and 'transfer-learning' [85].

Several commercial systems exist that focus on long-term monitoring using ADL. Such systems generally use motion sensors to track inhabitants inside their own homes [55]. The use of motion sensors does not allow for much diversity in observing an inhabitant, and therefore, the location of the sensor that is triggered is generally an important indicator of the activity that is performed. Other commercial health monitoring systems rely on the recording of physiological measurements such as heart rhythm and blood pressure. Examples are Intel's Health Guide [93] or Bosch's Health Buddy [92]. These systems prompt users to take measurements and provide a user friendly interface for doing so. The measurements can be made available to physicians for analysis and can be important indicators when something is wrong. Including automatically recognized ADL can play an important role for such systems, since the ADL information will provide additional contextual clues that a physician can use for interpreting the physiological data.

12.5 Visualization, Coaching and Communication

The previous chapters described systems for the observation and monitoring of human activities. What can we do with this information? Visualization of activity related information can further help caregivers in analyzing the behavior of individuals. In the work by Wang and Skubic, data obtained from motion sensors were visualized using a density map [90]. The density map shows the amount of motion registered by the sensors over time using a color coding. By presenting the sensor data of several days in a single density map, certain lifestyle trends become clearly visible. The authors discuss several case studies which show how the density maps can be used to highlight changes in the behavior of an individual over time. Aipperspach et al. showcased the potential of a more detailed visualization representation by studying the use of portable electronic devices in the home [2].

They used radio based location tracking technology to track the location of inhabitants and their laptops. The sensor data were visualized using a map of the house in which frequently visited areas were highlighted using colored regions. This allowed them to study where and when users typically use their laptops in the home.

Information about activity behavior can also be used for coaching of the users. One type of coaching is cognitive coaching. Activity recognition systems are used to assist people suffering from dementia, who tend to forget certain steps while performing an activity. For example, when making coffee they might install a coffee filter, but forget to add coffee powder. An activity recognition system can assist these people by recognizing the activity they are performing and reminding them which action to perform to complete the activity. Audio cues can be used to guide the person in performing any missing steps [40, 61] or a display can be used to show images of the actions that need to be performed [57, 64].

Also physical coaching can make use of the information. Persuasive technology motivates people to change their behavior, such as leading a healthier life style. One way is to provide users with well-timed information when they have to make decisions concerning their health. For example, to reduce the chances of obesity, a system can provide diet suggestions when it detects the user is preparing dinner. The appropriate timing of such a message is crucial to the success of such a system [44]. Another approach is to use a reward when the user is living a healthy life style. For example, in a study by Consolvo et al. [21], users were given an exercise program and a mobile phone showing the image of a virtual garden. If they spent enough time performing exercises from the exercise program, they received a visual reward in the form of flowers appearing in the virtual garden. The amount of time spent on exercises by users of the persuasive system was compared to participants that did not use the system. Participants using the system were shown to spend significantly more time on exercises than participants that did not use the system.

Features of activities can also be communicated to informal care givers. Work by Mynatt et al. used a digital photo frame to communicate activity data to family members [65]. Their setup involves two houses. One house is equipped with sensors, from which activity information is automatically recognized, and the other house is supplied with a photo frame. The photo frame consists of a picture of the individual whose activities are recorded, and uses several icons to display information about the activities. Field trials found that participants used the photo frame as a form of reassurance that everything is all right at the other person's home. Participants also reported that they used the photo frame to initiate phone conversations, because the photo frame displayed something of interest such as an increase in activity. Another example of such a communication device is the 'SnowGlobe' [89], depicted in Fig. 12.7.

12.6 Acceptance and Impact

Much of what has been published on barriers to the use of health technologies has focused on the technology or infrastructure, relationships within the health delivery

Fig. 12.7 Activity based communication: the SnowGlobe changes its appearance as a function of the activities in the elders home

system, or costs and reimbursements [32, 74]. Little has been published about the perceived needs or preferences, barriers and beliefs about health technology from the senior's perspective, especially minority seniors who have lower education, less computer literacy and more disabilities compared to the general population. Bertera et al. [12] report a study on acceptance of possible health technology. Ambient sensors and audio-visual communication with a doctor or a nurse, especially when a medical emergency occurred, scored high in acceptance, whereas the use of cameras raised concerns. Also the study carried out by Alwan et al. [6] showed that monitoring systems are generally accepted. There was a positive change in the perceived quality of life for some, but not all, of the participants after three months of monitoring. Also using ambient sensors, Virone et al. [88] showed that there was a correlation between the (deviations in the) behavior patterns of the users and the notes of the professional caregivers.

Acceptance was also studied for wearable systems. In [49] a study is reported that provides a nationally representative sample of consumer attitudes in United States on the topic of RFID medical informatics. It appears that public opposition to RFID technology is not widespread, and in fact there is enthusiasm for some applications. Evidence also suggests that attachment of RFID devices to the body is not viewed as objectionable by much of the public. Specifically, placement of RFID-based medical informatics devices on the arm by tape vs. as part of one's mobile phone does not seem to affect acceptability judgments except in a small percentage of the sample. On the other hand, wearable systems have the disadvantage that elderly do not wear them all the time or place them incorrectly on the body, as mentioned before.

The use of cameras is still an issue of debate. Bertera et al. [12] report that elderly do not accept cameras "... that allowed a nurse to check on them with a camera when they were unwell". In our own study on cameras for tele-health [68] we found a more detailed view: if an alarm system detects a fall, the elders do not mind that a camera is used. Also in a study by Londei et al. [56] it was shown that 96% of the elderly in the study were favorable or partially favorable to intelligent video systems for fall detection.

Automatic behavior analysis has great potential in health care applications, but the developed systems should take into account usability issues like acceptance and privacy.

12.7 Summary

With the aging population and the foreseen shortage of care personnel, there is an increasing interest in health technology. It has been shown that the health condition is strongly related to personal and environmental factors such as participation, activity and body functions. This chapter focuses on activity monitoring in a home setting for health care purposes. First the most current sensing systems are described. We make a distinction between wearable sensors like accelerometers or RFID tags, and ambient, environment mounted sensors, like cameras, motion detectors or pressure mats. An overview is given of the type of sensor and the application area.

In the second part of the chapter we focus on the activity behavior. We distinguish between simple actions, of limited time duration, and more complex activities, which may take longer. Several approaches for the recognition of simple actions are discussed, focusing on fall detection, wandering detection and therapies. After that, the recognition of more complex activities is discussed. A number of applications for the care givers are presented. The chapter concludes with a section on acceptance and privacy.

12.8 Questions

1. Give some more examples of actions and activities that you may think are relevant for health care.
2. What are the disadvantages of body worn sensors and what are the advantages?
3. Can we send a videostream from a surveillance camera over a Zigbee network?
4. How can multiple cameras looking at the same scene be calibrated?

12.9 Glossary

- *ADL*: Activities of daily living is a term used in health care to refer to daily self-care activities. Two types of ADL are distinguished: Basic ADL, necessary for fundamental functioning (bathing, clothing, etc.) and Instrumental ADL, such as shopping, managing money. There are several evaluation tools, such as the Katz ADL scale and the Lawton IADL scale.
- *RFID*: Radio frequency identification is a technology that uses communication via radio waves to exchange data between a reader and an electronic tag attached to an object, for the purpose of identification and tracking.

- *TOF camera*: A Time of Flight camera is a camera system that creates distance data with help of the time-of-flight (TOF) principle. Usually they work on a pulsed laser and a custom imaging integrated circuit with a fast counter behind every pixel, or by modulating the outgoing beam with an RF carrier, then measuring the phase shift of that carrier on the receive side.

Acknowledgements The research reported in this paper was supported by the Foundation Innovation Alliance SIA with funding from the Dutch Ministry of Education, Culture and Science (OCW), in the framework of the 'Smart Systems for Smart Services' project, and through the Pieken in de Delta-program by the Ministry of Economic Affairs and the cities of Utrecht and Lelystad and the provinces of Utrecht, Noord-Holland and Flevoland in the framework of the 'Zorgen voor Morgen' project.

References

1. Aggarwal, J., Park, S.: Human motion: modeling and recognition of actions and interactions. In: 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings, pp. 640–647. IEEE Press, New York (2004)
2. Aipperspach, R.J., Woodruff, A., Anderson, K., Hooker, B.: Maps of our lives: Sensing people and objects together in the home. Technical Report UCB/EECS-2005-22, EECS Department, University of California, Berkeley, November 30 2005. http://www.eecs.berkeley.edu/Pubs/TechRpts/2005/EECS-2005-22.html
3. Alemdar, H.Ö., Yavuz, G.R., Özen, M.O., Kara, Y.E., Incel, Ö.D., Akarun, L., Ersoy, C.: Multi-modal fall detection within the WeCare framework. In: Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks, pp. 436–437. ACM, New York (2010)
4. Ali, R., Atallah, L., Lo, B., Yang, G.Z.: Transitional activity recognition with manifold embedding. In: Proc. of BSN09, vol. 1 (2009)
5. Allin, S., Mihailidis, A.: Sit to stand detection and analysis. In: AI in Eldercare: New Solutions to Old Problems: Papers from the AAAI Fall Symposium (2008)
6. Alwan, M., Dalal, S., Mack, D., Kell, S., Turner, B., Leachtenauer, J., Felder, R.: Impact of monitoring technology in assisted living: outcome pilot. IEEE Trans. Inf. Technol. Biomed. **10**(1), 192–198 (2006)
7. Anderson, D., Luke, R.H., Keller, J.M., Skubic, M., Rantz, M., Aud, M.: Linguistic summarization of video for fall detection using voxel person and fuzzy logic. Comput. Vis. Image Underst. **113**(1), 80–89 (2009)
8. Auvinet, E., Multon, F., St-Arnaud, A., Rousseau, J., Meunier, J.: Fall detection using body volume reconstruction and vertical repartition analysis. In: Image and Signal Processing, pp. 376–383 (2010)
9. Auvinet, E., Multon, F., Saint-Arnaud, A., Rousseau, J., Meunier, J.: Fall detection with multiple cameras: An occlusion-resistant method based on 3D silhouette vertical distribution. IEEE Trans. Inf. Technol. Biomed. **15**(2), 290–300 (2011)
10. Aziz, O., Lo, B., King, R., Darzi, A., Yang, G.Z.: Pervasive body sensor network: an approach to monitoring the post-operative surgical patient. In: International Workshop on Wearable and Implantable Body Sensor Networks, 2006. BSN 2006, pp. 4–18. IEEE Press, New York (2006)
11. Barber, D.: Machine learning. A probabilistic approach (2006)
12. Bertera, E.M., Tran, B.Q., Wuertz, E.M., Bonner, A.: A study of the receptivity to telecare technology in a community-based elderly minority population. J. Telemed. Telecare **13**(7), 327 (2007)

13. Biswas, J., Zhang, D., Qiao, G., Foo, V., Qiang, Q., Philip, Y.: A system for activity monitoring and patient tracking in a smart hospital. In: Proceedings of 4th International Conference on Smart Home and Health Telematic
14. Bouchard, B., Giroux, S., Bouzouane, A.: A keyhole plan recognition model for Alzheimer's patients: first results. Appl. Artif. Intell. **21**(7), 623–658 (2007)
15. Bourke, A.K., O'Brien, J.V., Lyons, G.M.: Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm. Gait Posture **26**(2), 194–199 (2007)
16. Broch, J., Maltz, D.A., Johnson, D.B., Hu, Y.-C., Jetcheva, J.: A performance comparison of multi-hop wireless ad hoc network routing protocols. In: MobiCom '98: Proceedings of the 4th Annual ACM/IEEE International Conference on Mobile Computing and Networking, pp. 85–97. ACM, New York (1998). doi:10.1145/288235.288256
17. Canas, J.M., Marugán, S., Marrón, M., Garcia, J.: Visual fall detection for intelligent spaces. In: IEEE International Symposium on Intelligent Signal Processing (WISP 2009), pp. 157–162. IEEE Press, New York (2009)
18. Chen, J., Kwong, K., Chang, D., Luk, J., Bajcsy, R.: Wearable sensors for reliable fall detection. In: 27th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (IEEE-EMBS 2005), pp. 3551–3554. IEEE Press, New York (2005)
19. Chen, D., Bharucha, A.J., Wactlar, H.D.: Intelligent video monitoring to improve safety of older persons. In: 29th Annual International Conference of the Engineering in Medicine and Biology Society (EMBS 2007), pp. 3814–3817. IEEE Press, New York (2007)
20. Cho, Y., Nam, Y., Choi, Y.J., Cho, W.D.: SmartBuckle: human activity recognition using a 3-axis accelerometer and a wearable camera. In: Proceedings of the 2nd International Workshop on Systems and Networking Support for Health Care and Assisted Living Environments, pp. 1–3. ACM, New York (2008)
21. Consolvo, S., McDonald, D.W., Toscos, T., Chen, M.Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., Smith, I., Landay, J.A.: Activity sensing in the wild: a field trial of ubifit garden. In: CHI '08: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 1797–1806. ACM, New York (2008). doi:10.1145/1357054.1357335
22. Das, R.: RFID explained. IDTechEX White Paper (2005)
23. DiCenso, A., Cullum, N., Ciliska, D.: Implementing evidence-based nursing: some misconceptions. Evid.-Based Nurs. **1**(2), 38 (1998)
24. Diraco, G., Leone, A., Siciliano, P.: An active vision system for fall detection and posture recognition in elderly healthcare. In: Design, Automation & Test in Europe Conference & Exhibition (DATE 2010), pp. 1536–1541. IEEE Press, New York (2010)
25. Duong, T.V., Bui, H.H., Phung, D.Q., Venkatesh, S.: Activity recognition and abnormality detection with the switching hidden semi-Markov model. In: CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 838–845. IEEE Comput. Soc., Washington (2005). doi:10.1109/CVPR.2005.61
26. Duong, T.V., Bui, H.H., Phung, D.Q., Venkatesh, S.: Activity recognition and abnormality detection with the switching hidden semi-Markov model. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 838–845. IEEE Press, New York (2005)
27. Fishkin, K.P., Philipose, M., Rea, A.: Hands-on RFID: Wireless wearables for detecting use of objects. In: Proceedings of the Ninth IEEE International Symposium on Wearable Computers, pp. 38–41. IEEE Press, New York (2005)
28. Fogarty, J., Au, C., Hudson, S.E.: Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition. In: UIST '06: Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology, pp. 91–100. ACM, New York (2006). doi:10.1145/1166253.1166269
29. Fu, Z., Culurciello, E., Lichtsteiner, P., Delbruck, T.: Fall detection using an address-event temporal contrast vision sensor. In: IEEE International Symposium on Circuits and Systems (ISCAS 2008), pp. 424–427. IEEE Press, New York (2008)

30. Garrod, R., Bestall, J., Paul, E., Wedzicha, J., Jones, P.: Development and validation of a standardized measure of activity of daily living in patients with severe COPD: the London Chest Activity of Daily Living scale (LCADL). Respir. Med. **94**(6), 589–596 (2000)
31. Gavrila, D.M.: The visual analysis of human movement: a survey. Comput. Vis. Image Underst. **73**(1), 82–98 (1999)
32. Goldman, J., Hudson, Z.: Perspective: virtually exposed: privacy and e-health. Health Aff. **19**(6), 140 (2000)
33. Grassi, M., Lombardi, A., Rescio, G., Malcovati, P., Malfatti, M., Gonzo, L., Leone, A., Diraco, G., Distante, C., Siciliano, P., et al.: A hardware-software framework for high-reliability people fall detection. In: Sensors, 2008 IEEE, pp. 1328–1331 (2008)
34. Győrbíró, N., Fábián, Á., Hományi, G.: An activity recognition system for mobile phones. Mob. Netw. Appl. **14**(1), 82–91 (2009)
35. Haigh, K.Z., Yanco, H.: Automation as caregiver: A survey of issues and technologies. In: Proceedings of the AAAI-02 Workshop "Automation as Caregiver", pp. 39–53 (2002). AAAI Technical Report WS-02-02
36. Hamel, M., Fontaine, R., Boissy, P.: In-home telerehabilitation for geriatric patients. IEEE Eng. Med. Biol. Mag. **27**(4), 29–37 (2008)
37. Hazelhoff, L., Han, J., de With, P.H.N.: Video-based fall detection in the home using principal component analysis. In: Advanced Concepts for Intelligent Vision Systems: 10th International Conference (ACIVS 2008), Juan-les-Pins, France, October 20–24, 2008, p. 298. Springer, New York (2008)
38. Hill, J., Szewczyk, R., Woo, A., Hollar, S., Culler, D., Pister, K.: System architecture directions for networked sensors. ACM SIGPLAN Not. **35**(11), 93–104 (2000). doi:10.1145/356989.356998
39. Ho, L., Moh, M., Walker, Z., Hamada, T., Su, C.F.: A prototype on RFID and sensor networks for elder healthcare: progress report. In: Proceedings of the 2005 ACM SIGCOMM Workshop on Experimental Approaches to Wireless Network Design and Analysis, pp. 70–75. ACM, New York (2005)
40. Hoey, J., Poupart, P., Boutilier, C., Mihailidis, A.: POMDP models for assistive technology. In: Proc. AAAI Fall Symposium on Caring Machines: AI in Eldercare (2005)
41. Hoey, J., Poupart, P., Bertoldi, A., Craig, T., Boutilier, C., Mihailidis, A.: Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process. Comput. Vis. Image Underst. **114**(5), 503–519 (2010)
42. Hong, Y.J., Kim, I.J., Ahn, S.C., Kim, H.G.: Mobile health monitoring system based on activity recognition using accelerometer. Simul. Model. Pract. Theory **18**(4), 446–455 (2010)
43. Hori, T., Nishida, Y., Murakami, S.: Pervasive sensor system for evidence based nursing care support. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 1680–1685 (2006)
44. Intille, S.S.: A new research challenge: persuasive technology to motivate healthy aging. IEEE Trans. Inf. Technol. Biomed. **8**(3), 235–237 (2004)
45. Jansen, B., Deklerck, R.: Context aware inactivity recognition for visual fall detection. In: Pervasive Health Conference and Workshops, 2006, pp. 1–4. IEEE Press, New York (2007)
46. Jansen, B., Temmermans, F., Deklerck, R.: 3D human pose recognition for home monitoring of elderly. In: 29th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBS 2007), pp. 4049–4051. IEEE Press, New York (2007)
47. Karantonis, D.M., Narayanan, M.R., Mathie, M., Lovell, N.H., Celler, B.G.: Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. IEEE Trans. Inf. Technol. Biomed. **10**(1), 156–167 (2006)
48. Katz, S.: Assessing self-maintenance: Activities of daily living, mobility, and instrumental activities of daily living. J. Am. Geriatr. Soc. **31**(12), 721–726 (1983)
49. Katz, J.E., Rice, R.E.: Public views of mobile medical devices and services: A US national survey of consumer sentiments towards RFID healthcare technology. Int. J. Med. Inform. **78**(2), 104–114 (2009)

50. Kautz, H., Arnstein, L., Borriello, G., Etzioni, O., Fox, D.: An overview of the assisted cognition project. In: AAAI-2002 Workshop on Automation as Caregiver: The Role of Intelligent Technology in Elder Care, pp. 60–65 (2002)
51. Kerr, K., White, J., Barr, D., Mollan, R.: Analysis of the sit-stand-sit movement cycle in normal subjects. Clin. Biomech. **12**(4), 236–245 (1997)
52. Kurz, X., Scuvee-Moreau, J., Rive, B., Dresse, A.: A new approach to the qualitative evaluation of functional disability in dementia. Int. J. Geriatr. Psychiatry **18**(11), 1050–1055 (2003)
53. Kwolek, B.: Face tracking system based on color, stereovision and elliptical shape features. In: IEEE Conference on Advanced Video and Signal Based Surveillance, p. 21. IEEE Comput. Soc., Los Alamitos (2003). doi:10.1109/AVSS.2003.1217897
54. Liao, L., Fox, D., Kautz, H.: Extracting places and activities from gps traces using hierarchical conditional random fields. Int. J. Robot. Res. **26**(1), 119 (2007)
55. Living independently—quietcare system. www.livingindependently.com
56. Londei, S.T., Rousseau, J., Ducharme, F., St-Arnaud, A., Meunier, J., Saint-Arnaud, J., Giroux, F.: An intelligent videomonitoring system for fall detection at home: perceptions of elderly people. J. Telemed. Telecare **15**(8), 383 (2009)
57. LoPresti, E.F., Mihailidis, A., Kirsch, N.: Assistive technology for cognitive rehabilitation: State of the art. Neuropsychol. Rehabil. **14**(1–2), 5–39 (2004)
58. Marin-Perianu, M., Lombriser, C., Amft, O., Havinga, P., Tröster, G.: Distributed activity recognition with fuzzy-enabled wireless sensor networks. In: DCOSS '08: Proceedings of the 4th IEEE International Conference on Distributed Computing in Sensor Systems, pp. 296–313. Springer, Berlin (2008). doi:10.1007/978-3-540-69170-9_20
59. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: ICCV '09: Proceedings of the Twelfth IEEE International Conference on Computer Vision. IEEE Comput. Soc., Washington (2009)
60. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: IEEE 12th International Conference on Computer Vision 2009, pp. 104–111. IEEE Press, New York (2010)
61. Mihailidis, A., Barbenel, J.C., Fernie, G.: The efficacy of an intelligent cognitive orthosis to facilitate handwashing by persons with moderate to severe dementia. Neuropsychol. Rehabil. **14**(1–2), 135–171 (2004)
62. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2–3), 90–126 (2006)
63. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press, New York (2007)
64. Mynatt, E.D., Essa, I., Rogers, W.: Increasing the opportunities for aging in place. In: Proceedings of the 2000 Conference on Universal Usability, pp. 65–71. ACM, New York (2000)
65. Mynatt, E.D., Rowan, J., Craighill, S., Jacobs, A.: Digital family portraits: supporting peace of mind for extended family members. In: CHI, pp. 333–340 (2001). http://doi.acm.org/10.1145/365024.365126
66. Nait-Charif, H., McKenna, S.J.: Activity summarisation and fall detection in a supportive home environment. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), pp. 323–326 (2004)
67. Nevitt, M.C., Cummings, S.R., Hudes, E.S.: Risk factors for injurious falls: a prospective study. J. Gerontol. **46**(5), 164 (1991)
68. Rijnboutt, J., Evers, V., Kröse, B.: Cliënten willen meer controle over de camera. In: ICT en Zorg, pp. 30–32 (2010) (In Dutch)
69. Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Monocular 3d head tracking to detect falls of elderly people. In: 28th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBS'06), pp. 6384–6387. IEEE Press, New York (2008)
70. Sangwan, R., Qiu, R., Jessen, D.: Using RFID tags for tracking patients, charts and medical equipment within an integrated health delivery network. In: Proc. IEEE Networking, Sensing and Control, pp. 1070–1074. IEEE Press, New York (2005)

71. Sinha, A., Chandrakasan, A.: Dynamic power management in wireless sensor networks. IEEE Des. Test Comput. **18**(2), 62–74 (2001). doi:10.1109/54.914626
72. Sixsmith, A., Johnson, N.: A smart sensor to detect the falls of the elderly. IEEE Pervasive Comput. 42–47 (2004)
73. Sohrabi, K., Gao, J., Ailawadhi, V., Pottie, G.J.: Protocols for self-organization of a wireless sensor network. IEEE Pers. Commun. **7**(5), 16–27 (2000)
74. Song, W.J., Son, S.H., Choi, M., Kang, M.: Privacy and security control architecture for ubiquitous RFID healthcare system in wireless sensor networks. In: IEEE Int. Conf. Consumer Electronics, Digest of Technical Papers, pp. 239–240. IEEE Press, New York (2006)
75. Stikic, M., Huynh, T., Van Laerhoven, K., Schiele, B.: ADL recognition based on the combination of RFID and accelerometer sensing. In: Second International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2008), pp. 258–263. IEEE Press, New York (2008)
76. Tam, T., Dolan, A., Boger, J., Mihailidis, A.: An intelligent emergency response system: Preliminary development and testing of a functional health monitoring system. Gerontechnology **4**, 209–222 (2006)
77. Tapia, E.M., Intille, S.S., Lopez, L., Larson, K.: The design of a portable kit of wireless sensors for naturalistic data collection. In: Proceedings of the 4th International Conference on Pervasive Computing. Lecture Notes in Computer Science, vol. 3968, pp. 117–134. Springer, Berlin (2006)
78. Töreyin, B.U., Dedeoğlu, Y., Çetin, A.E.: HMM based falling person detection using both audio and video. In: Computer Vision in Human-Computer Interaction, pp. 211–220 (2005)
79. Truyen, T.T., Phung, D.Q., Bui, H.H., Venkatesh, S.: Hierarchical semi-Markov conditional random fields for recursive sequential data. In: Neural Information Processing Systems (NIPS) (2008)
80. Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O.: Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Technol. **18**(11), 1473–1488 (2008)
81. Vail, D.L., Veloso, M.M., Lafferty, J.D.: Conditional random fields for activity recognition. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 1–8. ACM, New York (2007)
82. van Dam, T., Langendoen, K.: An adaptive energy-efficient mac protocol for wireless sensor networks. In: SenSys '03: Proceedings of the 1st International Conference on Embedded Networked Sensor Systems, pp. 171–180. ACM, New York (2003). doi:10.1145/958491.958512
83. van Kasteren, T.L.M., Noulas, A., Englebienne, G., Kröse, B.: Accurate activity recognition in a home setting. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 1–9. ACM, New York (2008)
84. van Kasteren, T.L.M., Noulas, A., Englebienne, G., Kröse, B.J.A.: Accurate activity recognition in a home setting. In: UbiComp '08: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 1–9. ACM, New York (2008). doi:10.1145/1409635.1409637
85. van Kasteren, T.L.M., Englebienne, G., Kröse, B.: Transferring knowledge of activity recognition across sensor networks. IEEE Pervasive Comput. 283–300 (2010)
86. van Kasteren, T.L.M., Englebienne, G., Kröse, B.J.A.: Activity recognition using semi-Markov models on real world smart home datasets. J. Ambient Intell. Smart Environ. **2**(3), 311–325 (2010)
87. van Kasteren, T.L.M., Englebienne, G., Kröse, B.J.A.: An activity monitoring system for elderly care using generative and discriminative models. Pers. Ubiquitous Comput. **14**(6), 489–498 (2010)
88. Virone, G., Alwan, M., Dalal, S., Kell, S.W., Turner, B., Stankovic, J.A., Felder, R.: Behavioral patterns of older adults in assisted living. IEEE Trans. Inf. Technol. Biomed. **12**(3), 387–398 (2008)
89. Visser, T., Vastenburg, M., Keyson, D.: SnowGlobe: the development of a prototype awareness system for longitudinal field studies. In: Proceedings of the 8th ACM Conference on Designing Interactive Systems, pp. 426–429. ACM, New York (2010)
90. Wang, S., Skubic, M.: Density map visualization from motion sensors for monitoring activity level. In: 4th IET International Conference on Intelligent Environments (2008)

91. Wang, F., Stone, E., Dai, W., Skubic, M., Keller, J.: Gait analysis and validation using voxel data. In: Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC 2009), pp. 6127–6130. IEEE Press, New York (2009)
92. Website: Bosch health buddy. http://www.healthbuddy.com/
93. Website: Intel health guide. http://www.intel.com/healthcare/ps/healthguide/
94. Williams, A., Ganesan, D., Hanson, A.: Aging in place: fall detection and localization in a distributed smart camera network. In: Proceedings of the 15th International Conference on Multimedia, pp. 892–901. ACM, New York (2007)
95. Wilson, D.H., Consolvo, S., Fishkin, K.P., Philipose, M.: Current practices for in-home monitoring of elders' activities of daily living: A study of case managers. Technical report, Intel Research Seattle (2005)
96. Wilson, S., Davies, R., Stone, T., Hammerton, J., Ware, P., Mawson, S., Harris, N., Eccleston, C., Zheng, H., Black, N., et al.: Developing a telemonitoring system for stroke rehabilitation. Contemp. Ergon. **2007**, 505 (2007)
97. Wren, C.R., Tapia, E.M.: Toward scalable activity recognition for sensor networks. In: LoCa (2006)
98. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.M.: A scalable approach to activity recognition based on object use. In: IEEE 11th International Conference on Computer Vision (ICCV 2007), pp. 1–8. IEEE Press, New York (2007)
99. Zhuang, X., Huang, J., Potamianos, G., Hasegawa-Johnson, M.: Acoustic fall detection using Gaussian mixture models and GMM supervectors. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 69–72 (2009). doi:10.1109/ICASSP.2009.4959522
100. Xu, Y., Heidemann, J., Estrin, D.: Geography-informed energy conservation for ad hoc routing. In: MobiCom '01: Proceedings of the 7th Annual International Conference on Mobile Computing and Networking, pp. 70–84. ACM, New York (2001). doi:10.1145/381677.381685
101. Yu, M., Naqvi, S.M., Chambers, J.: Fall detection in the elderly by head tracking. In: IEEE/SP 15th Workshop on Statistical Signal Processing (SSP'09), pp. 357–360. IEEE Press, New York (2009)

Chapter 13
Behavioral, Cognitive and Virtual Biometrics

Roman V. Yampolskiy

13.1 Introduction to Behavioral Biometrics[1]

With the proliferation of computers in our everyday lives need for reliable computer security steadily increases. Biometric technologies provide user friendly and reliable control methodology for access to computer systems, networks and workplaces [1–3]. Biometric methods uniquely identify persons based on intrinsic physical or behavioral characteristics. Most research is aimed at studying well established physical biometrics such as fingerprint [4] or iris scans [5]. Behavioral biometrics systems are usually less established, and only those which are in large part based on muscle control such as keystrokes, gait or signature are well analyzed [6–11]. We define behavioral biometrics as any quantifiable actions of a person. Such actions may not be unique to the person and may take a different amount of time to be exhibited by different individuals. Biometric systems begin by enrolling individuals in the system, essentially introducing them to the security system and collecting personal data necessary for future authentication.

Behavioral biometrics provide a number of advantages over traditional biometric technologies. They can be collected non-obtrusively or even without the knowledge of the user. Collection of behavioral data often does not require any special hardware and is thus very cost effective. While most behavioral biometrics are not unique

[1]This chapter is based on numerous previous surveys and in particular expands on work in "Behavioral Biometrics: a Survey and Classification." by R. Yampolskiy and V. Govindaraju, which appeared in the International Journal of Biometrics, 1(1), 81–113 and Taxonomy of Behavioral Biometrics by same authors, a chapter in Liang Wang and Xin Geng (Eds.), Behavioral Biometrics for Human Identification: Intelligent Applications, pp. 1–43, 2009. Republished with permission of copyright holders IGI global and Inderscience.

R.V. Yampolskiy (✉)
University of Louisville, Louisville, KY, USA
e-mail: roman.yampolskiy@louisville.edu

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_13,

enough to provide reliable human identification (recognition) they have been shown to provide sufficiently high accuracy identity verification.

In accomplishing their everyday tasks, human beings employ different strategies, use different styles and apply unique skills and knowledge. One of the defining characteristics of a behavioral biometric is the incorporation of time dimension as a part of the behavioral signature. The measured behavior has a beginning, duration, and an end [12]. Behavioral biometrics researchers attempt to quantify behavioral traits exhibited by users and use resulting feature profiles to successfully verify identity [13]. In this section we present an overview of most established behavioral biometrics.

Behavioral biometrics can be classified into five categories based on the type of information being collected about the user. The first category is made up of authorship-based biometrics, which are based on examining a piece of text or a drawing produced by a person. Verification is accomplished by observing style peculiarities typical to the author of the work being examined, such as the used vocabulary, punctuation or brush strokes.

The second category consists of Human-Computer Interaction (HCI)-based biometrics [14]. In their everyday interaction with computers, humans employ different strategies, use different styles and apply unique abilities and knowledge. Researchers attempt to quantify such traits and use resulting feature profiles to successfully verify identity. HCI-based biometrics can be further subdivided into additional categories, first one consisting of human interaction with input devices such as keyboards, computer mice, and haptics, which is about registering inherent, distinctive and consistent muscle actions [15]. The second group consists of HCI-based behavioral biometrics, which measure advanced human behavior such as strategy, knowledge or skill exhibited by the user during interaction with different software.

The third group is closely related to the second one and is the set of the indirect HCI-based biometrics which are the events that can be obtained by monitoring user's HCI behaviors indirectly via observable low-level actions of computer software [16]. Those include system call traces [17], audit logs [18], program execution traces [19], registry access [20], storage activity [21], call-stack data analysis [22] and system calls [23, 24]. Such low-level events are produced unintentionally by the user during interaction with different software items. The same HCI-based biometrics are sometimes known to different researchers under different names. Intrusion Detection Systems (IDS) based on system calls or audit logs are often classified as utilizing program execution traces and those based on call-stack data as based on system calls. The confusion is probably related to the fact that a lot of interdependency exists between different indirect behavioral biometrics and they are frequently used in combinations to improve accuracy of the system being developed. For example system calls and program counter data may be combined in the same behavioral signature, or audit logs may contain information about system calls. Because they are indirect measures of behavior, they are outside of the scope of the current discussion and will not be evaluated in any detail in this chapter. The interested reader is encouraged to read the survey of indirect behavioral biometrics [16] for additional information.

The fourth and probably the best researched category of behavioral biometrics relies on motor-skills of the users to accomplish verification [25]. Motor skill is an ability of a human being to utilize muscles. Muscle movements rely upon the proper functioning of the brain, skeleton, joints, and nervous system and so motor skills indirectly reflect the quality of functioning of such systems, making person verification possible. Most motor skills are learned, not inherited, with disabilities having potential to affect the development of motor skills. We adopt here a definition for motor-skill-based behavioral biometrics, a.k.a. *kinetics*, as those biometrics which are based on innate, unique and stable muscle actions of the user while performing a particular task [26].

The fifth and final category consists of purely behavioral biometrics. Purely behavioral biometrics are those which measure human behavior directly (not concentrating on measurements of body parts) or intrinsic, inimitable and lasting muscle actions, such as the way an individual walks, types or even grips a tool [26]. Humans utilize different strategies, skills and knowledge during performance of mentally demanding tasks. Purely behavioral biometrics quantify such behavioral traits and make successful identity verification a possibility.

The present chapter additionally looks at Behavioral Passwords, Biosignals and Virtual Biometrics, such as avatar representations of the user. All of the authentication approaches reviewed in this chapter share a number of characteristics and so can be analyzed as a group using the seven properties of good biometrics presented by Jain et al. [5, 27]. It is a good idea to check them before declaring some characteristics suitable for the automated recognition of individuals.

- **Universality** Behavioral biometrics are dependent on specific abilities possessed by different people to a different degree (or not at all) and so in a general population, the universality of behavioral biometrics is very low. But since behavioral biometrics are only applied in a specific domain, the actual universality of behavioral biometrics is a 100%.
- **Uniqueness** Since only a small set of different approaches to performing any task exists, uniqueness of behavioral biometrics is relatively low. Number of existing writing styles, different game strategies and varying preferences are only sufficient for user verification, and not identification, unless the set of users is extremely small [28].
- **Permanence** Behavioral biometrics exhibit a low degree of permanence, as they measure behavior which changes with time as the person learns advanced techniques and faster ways of accomplishing tasks. However, this problem of concept drift is addressed in behavior-based intrusion detection research, and systems are developed capable of adjusting to the changing behavior of the users [29, 30].
- **Collectability** Collecting behavioral biometrics is relatively easy and unobtrusive to the user. In some instances the user may not even be aware that data collection is taking place. The process of data collection is fully automated and is of very low cost.
- **Performance** The identification accuracy of most behavioral biometrics is low, particularly as the number of users in the database becomes large. However verification accuracy is very good for some behavioral biometrics.

- **Acceptability** Since behavioral biometric characteristics can be collected without user participation, they enjoy a high degree of acceptability, but might be objected to for ethical or privacy reasons.
- **Circumvention** It is relatively difficult to get around behavioral biometric systems as they require intimate knowledge of someone else's behavior, but once such knowledge is available, fabrication might be very straightforward [31]. This is why it is extremely important to keep the collected behavioral profiles securely encrypted.

13.2 Description of Behavioral Biometrics

Table 13.1 shows majority of behavioral biometrics covered in this chapter, classified according to the five categories outlined in the previous section [32]. Many of the reviewed biometrics are cross-listed in multiple categories due to their dependence on multiple behavioral attributes. In addition, enrolment time and verification time (D = days, H = hours, M = Minutes, S = Seconds) of the listed biometrics are provided, as well as any hardware required for the collection of the biometric characteristic data. Out of all the listed behavioral biometrics only two are believed to be useful not just for person verification, but also for reliable large scale person identification. Those are: signature/handwriting and speech. Other behavioral biometrics may be used for identification purposes but are not reliable enough to be employed in that capacity in real-world applications.

Presented next are short overviews of the most researched behavioral biometrics listed in alphabetical order [32]. Figure 13.1 provides a visual overview of some of the presented behavioral biometrics.

13.2.1 Avatar Representation

With the advent of virtual communities such as Second Life, a lot of modern social interactions take place in cyber-worlds. In such interactions users are represented by virtual characters known as Avatars, which they design based on personal preferences. Recent work by Yampolskiy et al. [33–37] has shown that visual and behavioral aspects of avatars could be profiled for the purpose of user verification or identification. It is interesting to note that some biometric methods came very close to avatar development and intelligent robots/software authentication on a number of different instances. For example, in 1998, M.J. Lyons and his colleagues published a report: "*Avatar Creation using Automatic Face Recognition*", where authors discussed specific steps and processing techniques that need to be taken in order for an avatar to be created almost automatically from the human face [38]. In fact, the process described in the above article is essentially the process of biometric synthesis, conceptualized and generalized in the book devoted specifically to this subject [39].

Table 13.1 Classification and properties of behavioral biometrics [32]

Classification of the Various Types of Behavioral Biometrics	Authorship	Direct Human Computer Interaction		Motor Skill	Purely Behavioral	Properties of Behavioral Biometrics			
		Input Device Interaction Based	Software Interaction Based			Enrolment time	Verification time	Identification	Required Hardware
Avatar Representation					●	M	M	N	Computer
Biometric Sketch	●				●	M	S	N	Mouse
Blinking				●		M	S	N	Camera
Calling Behavior					●	D	D	N	Phone
Car Driving Style					●	H	M	N	Car Sensors
Center of Gravity				●		M	S	N	Shoe Sensors
Command Line Lexicon			●		●	H	H	N	Computer
Credit Card Use					●	D	D	N	Credit Card
Dynamic Facial Features				●		M	S	N	Camera
Email Behavior	●		●		●	D	M	N	Computer
Finger Pressure				●		M	S	N	Pressure Sensor
Floor Pressure				●		M	S	N	Floor Sensor
Gait/Stride				●		M	S	N	Camera
Game Strategy			●		●	H	H	N	Computer
Gaze/Eye Tracking					●	M	S	Y	Eye Tracker
Handgrip				●		M	S	N	Gun Sensors
Haptic		●		●		M	M	N	Haptic

Table 13.1 (Continued)

Classification of the Various Types of Behavioral Biometrics	Authorship	Direct Human Computer Interaction		Motor Skill	Purely Behavioral	Properties of Behavioral Biometrics			
		Input Device Interaction Based	Software Interaction Based			Enrolment time	Verification time	Identification	Required Hardware
Human Shadows				●		M	S	N	Camera
Keystroke Dynamics		●		●		M	S	N	Keyboard
Lip Movement				●		M	S	N	Camera
Mouse Dynamics		●		●		M	S	N	Mouse
Motion of Fingers				●		M	S	N	Camera
Painting Style	●				●	D	D	N	Scanner
Programming Style	●		●		●	H	H	N	Computer
Signature/Handwriting				●		M	S	Y	Stylus
Shirt Term Memory					●	M	M	Y	Mouse
Tapping				●		M	S	N	Sensor
Text Authorship	●				●	H	M	N	Computer
Visual Scan					●	M	M	Y	Mouse
Voice/Speech/Singing				●		M	S	Y	Microphone

Users of virtual words have also noted that avatars often take on the characteristics of their creators, and not only their facial characteristics, but also body shape, accessories and clothes.

But what about other, less obvious resemblances, such as manner of communication, responses to various situations, nature of work, style of house, leisure/recreational activities, time of appearing in virtual world, etc.? All of the above encompasses behavioral characteristics that can be exploited by the fusion of biometric-based techniques, with methodology tailored to specifics of the virtual world. Such behavioral characteristics, as the author of this chapter would postulate, are even less likely to change than the avatar's facial appearance and clothes during virtual world sessions, as users typically invest a lot of time and money in the creation of a consistent virtual image, but would not so easily change their patterns of behavior.

13.2.2 Biometric Sketch

Bromme et al. [40, 41] proposed a biometric sketch authentication method based on sketch recognition and a user's personal knowledge about the drawing's content. The system directs a user to create a simple sketch, for example of three circles, and each user is free to do so in any way he pleases. Because a large number of different combinations exist for combining multiple simple structural shapes, sketches of different users are sufficiently unique to provide accurate authentication. The approach measures a user's knowledge about the sketch, which is only available to the previously authenticated user. Such features as the sketch's location and relative position of different primitives are taken as the profile of the sketch. Finally a V-go Password requests a user to perform the simulation of simple actions, such as mixing a cocktail using a graphical interface, with the assumption that all users have a personal approach to bartending [42].

13.2.3 Blinking

Westeyn et al. [43], Westeyn and Starner [44] have developed a system for identifying users by analyzing voluntary song-based blink patterns. During the enrolment phase the user looks at the system's camera and blinks to the beat of a song he has previously chosen, producing a so-called "blinkprint". During the verification phase, the user's blinking is compared to the database of the stored blinked patterns to determine which song is being blinked and as a result user identification is possible. In addition to the blink pattern itself, supplementary features can also be extracted, such as: time between blinks, how long the eye is held closed at each blink, and other physical characteristics the eye undergoes while blinking. Based on those additional features, it was shown to be feasible to distinguish users blinking the same exact pattern and not just a secretly selected song.

13.2.4 Calling Behavior

With the proliferation of the mobile cellular phone networks, communication companies are faced with increasing amount of fraudulent calling activity. In order to automatically detect theft of service, many companies are turning to behavioral user profiling with the hopes of detecting unusual calling patterns to be able to stop fraud at an earliest possible time. Typical systems work by generating a user calling profile, which consists of usage indicators such as: date and time of the call, duration, called ID, called number, cost of call, number of calls to a local destination, number of calls to mobile destinations, number of calls to international destinations and the total statistics about the calls for the day [45]. Grosser et al. [46] have shown that neural networks can be successfully applied to such a feature vector for the purpose of fraud detection. Fawcett et al. [47] developed a rule-learning program to uncover indicators of fraudulent behavior from a large database of customer transactions.

13.2.5 Car Driving Style

People tend to operate vehicles in very different ways; some drivers are safe and slow, others are much more aggressive and often speed and tailgate. As a result, driving behavior can be successfully treated as a behavioral biometric. Erdogan et al. [48] have shown that by analyzing pressure readings from the accelerator and brake pedals in kilogram force per square centimeter, the vehicle speed in revolutions per minute, and steering angle within the range of -720 to $+720$ degrees, it is possible to achieve genuine versus impostor driver authentication. Gaussian mixture modeling was used to process the resulting feature vectors, after some initial smoothing and subsampling of the driving signal. Liu et al. [49] in their work on prediction of driver behavior have demonstrated that inclusion of the driver's visual scanning behavior can further enhance accuracy of the driver behavior model. Once fully developed, driver recognition can be used for car personalization, theft prevention, as well as for detection of drunk or sleepy drivers. With so many potential benefits from this technology, research in driver behavior modeling is not solely limited to the biometrics community [50, 51].

13.2.6 Center of Gravity

Porwik et al. [52] have proposed a system based on analysis of the motion of the human body's gravity center. By utilizing specially designed shoe soles with sensors and asking 15 volunteers to engage in some stationary movement (without lifting their feet) they were able to collect time series data about the subjects' center of gravity.

13.2.7 Command Line Lexicon

A popular approach to the construction of behavior-based intrusion detection systems is based on profiling the set of commands utilized by the user in the process of interaction with the operating system. A frequent target of such research is the UNIX operating system, probably due to it having mostly command line nature. Users differ greatly in their level of familiarity with the command set and all the possible arguments which can be applied to individual commands. Regardless of how well a user knows the set of available commands, most are fairly consistent in their choice of commands used to accomplish a particular task.

A user profile typically consists of a list of used commands together with corresponding frequency counts, and lists of arguments to the commands. Data collection process is often time consuming, since as many as 15,000 individual commands need to be collected for the system to achieve a high degree of accuracy [53, 54]. Additional information about the session may also be included in the profile, such as the login host and login time, which help to improve accuracy of the user profile, as it is likely that users perform different actions on different hosts [55]. Overall, this line of research is extremely popular, but recently a shift has been made toward user profiling in a graphical environment such as Windows, as most users prefer the convenience of a Graphical User Interface (GUI).

13.2.8 Credit Card Use

Data mining techniques are frequently used in detection of credit card fraud. Looking out for statistical outliers such as unusual transactions, payments to far away geographical locations or simultaneous use of a card at multiple locations can all be signs of a stolen account. Outliers are considerably different from the remainder of the data points and can be detected by using discordancy tests. Approaches for fraud related outlier detection are based on distance, density, projection, and distribution analysis methods. A generalized approach to finding outliers is to assume a known statistical distribution for the data and to evaluate the deviation of samples from the distribution. Brause et al. [56] have used symbolic and analog number data to detect credit card fraud. Such transaction information as account number, transaction type, credit card type, merchant ID, merchant address, etc. were used in their rule-based model. They have also shown that analog data alone cannot serve as a satisfying source for detection of fraudulent transactions.

13.2.9 Dynamic Facial Features

Pamudurthy et al. [57] proposed a dynamic approach to face recognition based on dynamic instead of static facial features. They track the motion of skin pores on

the face during a facial expression and obtain a vector field that characterizes the deformation of the face. In the training process, two high-resolution images of an individual, one with a neutral expression and the other with a facial expression, like a subtle smile, are taken to obtain the deformation field [58].

Smile recognition research in particular is a subfield of dynamic facial feature recognition currently gaining in prominence [59]. The existing systems rely on probing the characteristic pattern of muscles beneath the skin of the user's face. Two images of a person in quick progression are taken, with subjects smiling for the camera in the second sample. An analysis is later performed of how the skin around the subject's mouth moves between the two images. This movement is controlled by the pattern of muscles under the skin, and is not affected by the presence of make-up or the degree to which the subject smiles [58]. Other researchers have done research in this area under such names as: Facial Behavior [60] and Facial Actions [61].

13.2.10 Email Behavior

Email sending behavior is not the same for all individuals. Some people work at night and send dozens of emails to many different addresses; others only check mail in the morning and only correspond with one or two people. All these peculiarities can be used to create a behavioral profile which can serve as a behavioral biometric characteristic for an individual. Length of the emails, time of the day the mail is sent, how frequently inbox is emptied and of course the recipients' addresses among other variables can all be combined to create a baseline feature vector for the person's email behavior. Some work in using email behavior modeling was done by Stolfo et al. [62, 63]. They have investigated the possibility of detecting virus propagation via email by observing abnormalities in the email sending behavior, such as unusual clique of recipients for the same email. For example sending the same email to your girlfriend and your boss is not an everyday occurrence.

De Vel et al. [64] have applied authorship identification techniques to determine the likely author of an email message. Alongside the typical features used in text authorship identification, authors also used some email specific structural features such as: use of a greeting, farewell acknowledgment, signature, number of attachments, position of re-quoted text within the message body, HTML tag frequency distribution and total number of HTML tags. Overall, almost 200 features are used in the experiment, but some frequently cited features used in text authorship determination are not appropriate in the domain of email messages due to the shorter average size of such communications.

13.2.11 Finger Pressure

Many modern mobile devices are equipped with touchpad devices capable of detecting pressure. Saevanee et al. [65] have proposed utilizing finger pressure as a

Table 13.2 Floor pressure biometric—accuracy rates comparison

Features	Recognition Rate	False Accept Rate	Researchers	Year
Pressure profile over footsteps	50%	–	Addlesee et al. [67]	1997
Trajectories of center of pressure	64%	5.8%	Jung et al. [68]	2003
Pressure over the entire floor area	76.9%	11.6%	Pirttikangas et al. [69]	2003
Stride length, stride cadence, heel-to-toe ratio	80%	–	Middleton et al. [70]	2005
Compensated foot centers	92.8	–	Yun et al. [71]	2003
Points from pressure profile	93%		Orr et al. [72]	2000
Patterns of footsteps	92%	–	Yoon et al. [73]	2005
Pressure and time features	81.9%	–	Suutala et al. [74]	2008
Mean pressure and stride length	92.3%	6.79%	Qian et al. [66]	2010

behavioral biometric and have achieved an impressive 99% accuracy rate. In their experiments they combine finger pressure defined as the force applied over the finger position with keystroke dynamics. Specifically Saevanee et al. consider the pressing area not as a single point, but a group of multiple points on the pad. They utilize the average value over these multiple pressing points to produce a representative feature vector: $FP_i = [P_{i,1}, p_{i,2}, \ldots, P_{i,10}]$ where $P_{i,j}$ denotes the average value of finger pressure values at the round i of digit j [65]. They were able to achieve an Equal Error Rate value of 1% for a group of 10 test subjects and an accuracy rate of 99%.

13.2.12 Floor Pressure

While walking, people exert pressure on the floor surface, which could be analyzed and used for personal authentication. Different types of floor sensors could be used to collect floor pressure data; for example load cells, pressure mats, force sensitive resistor mats and switch sensors have been experimented with. Qian et al. [66] used a large high-resolution pressure sensing floor to capture a 1D pressure profile and 2D position trajectories for both feet. They later separate data from the two feet and from those trajectories corresponding to the centers of pressure, extract features such as the mean pressure and stride length. Extracted features were classified with a Fisher's linear discriminant classifier. Table 13.2 summarizes accuracy rates obtained by different researchers of the floor pressure biometric [66].

13.2.13 Gaze/Eye Tracking

While viewing an image, a person goes through a sequence of eye fixation-saccade events necessary to build up a perception of a scene. The spatial and temporal pat-

terns associated with eye tracking are widely varied between different people and could be used to produce a visual attention map of each individual. The position of gaze locations could be produced deliberately or subconsciously during viewing behavior and a simple webcam or a sophisticated eye-tracker device could be utilized in the data collection process. For a given image, Maeder and Fookes [75] analyze gaze data sampled at 15 fps using a spatial clustering algorithm to extract any fixations with an approximate viewing time of 1.0 secs and a tolerance of 0.3 secs. The approach could be combined with a standard PIN-like authentication mechanism by mapping locations on pre-labeled regions of the image [76].

13.2.14 Gait/Stride

Gait is one of the best researched muscle control-based biometrics [77–79]; it is a complex spatio-temporal motor-control behavior which allows biometric recognition of individuals at a distance, usually from captured video. Gait is subject to significant variations based on the changes in a person's body weight, waddling during pregnancy, injuries of extremities or of the brain, or due to intoxication [27]. Typical features include: amount of arm swing, rhythm of the walker, bounce, length of steps, vertical distance between head and foot, distance between head and pelvis, maximum distance between the left and right foot [80].

13.2.15 Game Strategy

Yampolskiy et al. [81–83] proposed a system for verification of online poker players based on a behavioral profile, which represents a statistical model of player's strategy. The profile consists of frequency measures indicating range of cards considered by the player at all stages of the game. It also measures how aggressive the player is via such variables as percentages of re-raised hands. The profile is actually human readable, meaning that a poker expert can analyze and understand the strategy employed by the player from observing his or her behavioral profile [84]. For example just by knowing the percentage of hands a particular player chooses to play pre-flop, it is possible to determine which cards are being played with high degree of accuracy.

Ramon et al. [85] have demonstrated the possibility of identifying Go players based on their style of game play. They analyzed a number of Go specific features such as type of opening moves, how early such moves are made and total number of liberties in the formed groups. They also speculated that the decision tree approach they have developed can be applied to other games such as Chess or Checkers.

In [86], Jansen et al. report about their research in chess strategy inference from game records. In particular, they were able to surmise good estimates of the weights used in the evaluation function of computer chess players, and later applied same

techniques to human grandmasters. Their approach is aimed at predicting future moves made by the players, but the opponent model created with some additional processing can be utilized for opponent identification or at least verification. This can be achieved by comparing new moves made by the player with predicted ones from models for different players and using the achieved accuracy scores as an indication of which profile models which player.

13.2.16 Handgrip

Developed mostly for gun control applications, grip-pattern recognition approach assumes that users hold the gun in a sufficiently unique way to permit user verification to take place. By incorporating a hardware sensor array in the gun's butt, Kauffman et al. [87, 88] were able to get resistance measurements in as many as 44×44 points which are used in the creation of a feature vector. Obtained pressure points are taken as pixels in the pressure pattern image used as input for verification algorithm based on a likelihood-ratio classifier for Gaussian probability densities [87]. Experiments showed that more experienced gun users tended to be more accurately verified as compared to first time subjects.

13.2.17 Haptic

Haptic systems are computer input/output devices, which can provide us with information about direction, pressure, force, angle, speed, and position of user's interactions [89, 90]. Because so much information is available on the user's performance, a high degree of accuracy can be expected from a haptic-based biometrics system. Orozco et al. [89, 90] have created a simple haptic application built on an elastic membrane surface, in which the user is required to navigate a stylus through the maze. The maze has gummy walls and a stretchy floor. The application collects data about the ability of the user to navigate the maze, such as reaction time to release from sticky wall, the route, the velocity, and the pressure applied to the floor. The individual user profiles are made up of such information as 3D world location of the pen, average speed, mean velocity, mean standard deviation, navigation style, angular turns and rounded turns. In a separate experiment Orozco et al. [91] implement a virtual mobile phone application where the user interacts through a haptic pen to simulate making a phone call via a touch pad. The keystroke duration, pen's position, and exerted force are used as the raw features collected for user profiling.

13.2.18 Human Shadows

Shadow biometrics rely on the use of shadows and shadow dynamics for behavior-based recognition of persons. In the above-the-head imagery taken from a large

distance, direct recognition of humans is not always possible due to limited view of the observation angle. Shadows provide additional information, which may be sufficient for person identification. Shadows have a larger observable area and reflect well the underlining gait dynamics, making biometric authentication possible. Potential features of shadow biometrics include: shadow area, parameters for a triangular model formed by extremities of head and the feet, parameters for a pentagonal model formed by the head, two hands and two feet, correction for the position of the light source (usually the sun), and dynamic features such as amplitude and periodicity of movement and deviation from regularity [92]. Iwashita et al. [93] extracted gait features from manually selected shadows and obtained a Correct Classification Rate of over 95%. With automatic shadow area selection, accuracy dropped to 90%.

13.2.19 Keystroke Dynamics

Typing patterns are characteristic to each person, some people are experienced typists utilizing the touch-typing method, and others utilize the hunt-and-peck approach which uses only two fingers. Those differences make verification of people based on their typing patterns a proven possibility, and some reports suggest identification is also possible [94]. For verification a small typing sample such as the input of user's password is sufficient, but for recognition a large amount of keystroke data are needed and identification is based on comparisons with the profiles of all other existing users already in the system.

Keystroke features are based on time durations between the keystrokes, interkey strokes and dwell time, which is the time a key is pressed down, overall typing speed, frequency of errors (use of backspace), use of numpad, the order in which the user presses shift key to get capital letters and possibly the force with which keys are hit for specially equipped keyboards [27, 94]. Keystroke dynamics are probably the most researched type of HCI-based biometric characteristics, with novel research taking place in different languages, for long text samples, and for email authorship identification.

In a similar fashion, Bella et al. [95] have studied finger movements of skilled piano players. They have recorded finger motion from skilled pianists while playing a musical keyboard. Pianists' finger motion and speed with which keys are struck were analyzed using functional data analysis methods. Movement velocity and acceleration were consistent for the participants and in multiple musical contexts. Accurate pianist classification was achieved by training a neural network classifier using velocity/acceleration trajectories preceding key presses. Gamboa et al. [96] have used keystroke dynamics in a system they called Webbiometrics, which used web interaction for user verification.

13.2.20 Lip Movement

This approach, originally based on the visual speech reading technology, attempts to generate a model representing the lip dynamics produced by a person during speech. User verification is based on how close the generated model fits observed lip movement. Such models are typically constructed around spatio-temporal lip features. First the lip region needs to be isolated from the video feed, and then significant features of lip contours are extracted, typically from edges and gradients. Lip features include: the mouth opening or closing, skin around the lips, mouth width, upper/lower lip width, lip opening height/width, distance between horizontal lip line and upper lip [97, 98]. Typically, lip dynamics is utilized as a part of a multimodal biometric system, usually combined with speaker recognition-based authentication [99–102], but stand-alone usage is also possible [103].

13.2.21 Mouse Dynamics

By monitoring all mouse actions produced by the user during an interaction with the Graphical User Interface (GUI), a unique profile can be generated which can be used for user re-authentication [23]. Mouse actions of interest include general movement, drag and drop, point and click, and stillness. From those, a set of features can be extracted, for example average speed against the distance traveled, and average speed against the movement direction [104, 105]. Pusara et al. [23] describe a feature extraction approach in which they split the mouse event data into mouse wheel movements, clicks, menu and toolbar clicks. Click data are further subdivided into single- and double-click data.

Gamboa et al. [106, 107] have tried to improve the accuracy of mouse-dynamics-based biometrics by restricting the domain of data collection to an online game instead of a more general GUI environment. As a result, the applicability of their results is somewhat restricted, and the methodology is more intrusive to the user. The system requires around 10–15 minutes of devoted game play instead of seamless data collection during normal user-computer interaction. As far as the extracted features go, x and y coordinates of the mouse, horizontal velocity, vertical velocity, tangential velocity, tangential acceleration, tangential jerk and angular velocity are utilized with respect to the mouse strokes to create a unique user profile.

13.2.22 Motion of Fingers

Nishiuchi et al. [108] have proposed a method of identifying individuals using the bending motion of fingers. The person being authenticated moves a finger over a solid color background. This allows for easy finger detection and post-processing, which involves calculation of the curvature defined as a value indicating the level

of bending at each point on a curve, or a curved surface. By extracting edge pixels from a binary image of the forefinger they were able to calculate curvature involved in the motion of the finger. The correlation coefficient is used for the evaluation of the curvature profiles. In their experimental setup, Nishuchi et al. use seven angles of the forefinger (15° to 45° in 5° increments) and utilize six test subjects. With the decision threshold set at 0.970, the False Reject Rate is 0%.

13.2.23 Painting Style

Just like authorship of literary works can be attributed based on the writer's style, so can works of art be accredited based on the style of the drawing. In particular the subtle pen and brush strokes characteristic of a particular painter can be profiled. Lyu et al. [109] developed a technique for performing a multi-scale, multi-orientation painting scan decomposition. This decomposition changes the basis from functions maximally localized in space to one in which the basis functions are also localized in orientation and scale. By constructing a compact model of the statistics from such a function, it is possible to detect consistencies or inconsistencies between paintings and drawings supposedly produced by the same painter.

13.2.24 Programming Style

With the increasing number of viruses, worms, and Trojan horses, it is often useful in a forensic investigation to be able to identify an author of such malware programs based on the analysis of the source code. It is also valuable for the purposes of software debugging and maintenance to know who the original author of a certain code fragment was. Spafford et al. [110] have analyzed a number of features potentially useful for the identification of software authorship. In case only the executable code is available for analysis, data structures and applied algorithms can be profiled, as well as any remaining compiler and system information, observed programming skill level, knowledge of the operating system and choice of the system calls. Additionally, use of predefined functions and provisions for error handling are not the same for different programmers.

In case the original source files are available, a large number of additional identifying features become accessible such as: chosen programming language, code formatting style, type of code editor, special macros, style of comments, variable names, spelling and grammar, use of language features such as choice of loop structures, the ratio of global to local variables, temporary coding structures, and finally types of mistakes observable in the code. Software metrics such as the number of lines of code per function, comment-to-code ratio and function complexity may also be introduced [110].

13.2.25 Signature/Handwriting

Signature verification is a widely accepted methodology for confirming identity [111–114]. Two distinct approaches to signature verification are traditionally recognized based on the data collection approach, they are: on-line and off-line signature verification, also known as static and dynamic approaches [115]. In the off-line signature verification, the image of the signature is obtained using a scanning device, possibly some time after the signing took place. With on-line signature verification, special hardware is used to capture dynamics of the signature; typically pressure sensitive pens in combination with digitizing tablets are utilized. Because on-line data acquisition methodology obtains features not available in the off-line mode, dynamic signature verification is more reliable [116].

With on-line signature verification, in addition to the trajectory coordinates of the signature, other features like pressure at pen tip, acceleration and pen-tilt can be collected. In general, signature related features can be classified into two groups: global and local. Global features include: signing speed, signature bounding box, Fourier descriptors of the signature's trajectory, number of strokes, and signing flow. Local features describe specific sample points in the signature and relationship between such points. For example, the distance and curvature changes between two successive points may be analyzed, as well as x and y offsets relative to the first point on the signature trajectory, and critical points of the signature trajectory [116, 117].

Signature-based user verification is a particular type of general handwriting-based biometric authentication. Unlike with signatures, handwriting-based user verification/recognition is content independent, which makes the process somewhat more complicated [118–120]. Each person's handwriting is seen as having a specific texture. The spatial frequency and orientation contents represent the features of each texture [121]. Since handwriting provides a much more substantial biometric characteristic sample in comparison to a signature, respective verification accuracy can be much greater.

13.2.26 Short Term Memory

Human sensory input storage receives visual snapshots from the eyes and stores them briefly in visual cortex to allow for the analysis of the perceived data. Such analysis involves retrieval of important information from points of interest and filtering out of inapplicable information. Short Term Memory (STM) could be characterized in terms of size and decay, which for visual information is estimated to be 17 letters and an average time of 200 ms. It has been shown in multiple experiments that the information retrieval capability from STM is different from one subject to the other [122, 123]. Hamdy et al. [123] proposed an approach for measuring STM time indirectly via analysis of mouse movements under stressful conditions while performing a cognitive task. Extracted features included traveled distance decrease rate as the one-key occurrence increased and fly time improvement rate as the one-key occurrence increased.

13.2.27 Soft Behavioral Biometrics

Jain et al. [124, 125] define soft biometrics as: "... traits as characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals". They further state that soft biometric traits can either be continuous, such as height or weight, or discrete, such as gender or ethnicity. Authors propose expanding the definition to include soft behavioral biometrics, which also can be grouped into continuous and discrete types. For instance, continuous soft behavioral biometric traits can include measurements produced by various standardized tests (some of the most popular such tests are IQ test for intelligence, and verbal sections of SAT, GRE, GMAT for language abilities). Discrete soft behavioral biometrics are skills which a particular person either has or does not have. Examples of such include the ability to speak a particular foreign language, knowledge of how to fly a plane, or to ride a motorcycle, etc.

While such soft behavioral biometrics are not sufficient for identification or verification of individuals, they can be combined with other biometric approaches to increase system accuracy. They can also be used in certain situations to reject an individual's verification claim. For example in a case of academic cheating, a significantly fluctuating score on a repeatedly taken standardized test can be used to suspect that not the same person answered all the questions on a given test [126].

13.2.28 Tapping

Henderson et al. [127, 128] have studied the idea of tapping recognition, based on the idea that you are able to recognize who is knocking on your door. They concentrated on the waveform properties of the pulses which result from tapping a polymer thick-film sensor on a smart card. Produced pressure pulses are further processed to extract useful features such as: pulse height, pulse duration, and the duration of the first inter-pulse interval. The recognition algorithm utilized in this research has been initially developed for processing of keyboard dynamics, which is a somewhat similar technology of recognizing tapping with respect to keyboard keys.

13.2.29 Text Authorship

Email and source code authorship identification represent application and improvement of techniques developed in a broader field of text authorship determination. Written text and spoken word (once transcribed) can be analyzed in terms of vocabulary and style to determine authorship. In order to do so, a linguistic profile needs to be established. Many linguistic features can be profiled, such as: lexical patterns, syntax, semantics, pragmatics, information content or item distribution through a text [129]. Stematatos et al. [130], in their analysis of modern Greek texts, proposed

using such text descriptors as: sentence count, word count, punctuation mark count, noun phrase count, word included in noun phrase count prepositional phrase count, word included in prepositional phrase count and keyword count. The overall area of authorship attribution is very promising, with a lot of ongoing research [131–133].

13.2.30 Visual Scan/Search and Detection

This novel biometric is based on the human visual system. In our daily life, as we examine signs, advertisements or websites for a specific piece of information, we discriminate a target of interest from surrounding distracters. The idea is to properly measure the average inspection time of an individual and use that information in a behavioral signature [122]. The investigators demonstrated an approach for measuring Visual Scan time indirectly via examining mouse movements under stressful conditions. Obtained features included: speed, fly time and distance traveled. In combination with Short Term Memory cognitive factor this biometrics has been able to achieve Equal Error Rate of 3.88% on a dataset of 275 test subjects [123].

13.2.31 Voice/Speech/Singing

Speaker identification is one of the best researched biometric technologies [134–136]. Verification is based on information about the speaker's anatomical structure conveyed in amplitude spectrum, with the location and size of spectral peaks related to the vocal tract shape and the pitch striations related to the glottal source of the user [80]. Speaker identification systems can be classified based on the freedom of what is spoken [137]:

- **Fixed text:** The speaker says a particular word selected at enrolment.
- **Text dependent:** The speaker is prompted by the system to say a particular phrase.
- **Text independent:** The speaker is free to say anything he wants, verification accuracy typically improves with larger amount of spoken text.

  Feature extraction is applied to the normalized amplitude of the input signal, which is further decomposed into several band-pass frequency channels. A frequently extracted feature is the logarithm of the Fourier Transform of the voice signal in each band, along with features of pitch, tone, cadence, and shape of the larynx [27]. Accuracy of voice-based biometrics systems can be increased by inclusion of visual speech (lip dynamics) [99–102] and incorporation of soft behavioral biometrics such as accent [138, 139]. Recently some research has been aimed at expanding the developed technology to singer recognition for the purposes of music database management [140] and to laughter recognition. Currently, laughter-recognition software is rather crude and cannot accurately distinguish between different people [58, 59].

Some of the presented approaches are not sufficiently unique, permanent, easily collectable or difficult to circumvent, but they can be seen as behavioral counterparts of "soft" physical biometrics well recognized in the field. Soft biometrics are also not strong enough to be a backbone of a standalone biometric security system, but are nonetheless valuable in improving accuracy of multimodal systems. Likewise, we believe that multimodal behavior-based biometric systems will be able to take advantage of many of the technologies presented in our survey and therefore it is important to include them to make our survey as comprehensive and as useful as possible to the largest number of researchers and developers. For example, game strategy alone may not be sufficient for person identification, but combined with keyboard dynamics and mouse movements, it might be sufficiently discriminative. Also, as breakthroughs are made in the field of behavioral biometrics, it is likely that some of the described technologies will become easier to collect and harder to circumvent.

Practically none of the behavioral biometrics are strong enough for person identification, and they are only useful for verification purposes. So, the assumption is always made that we are dealing with a cooperating subject who wishes to positively verify his identity. For all behaviors, even for low level ones, an un-cooperating subject can completely change his behavior in order to avoid being successfully profiled by the security system. This is an inherent limitation of most behavioral biometric systems.

13.3 Biological Signals as a Behavioral Biometrics

Because behavioral biometrics is a new and still developing field, even a basic concept as what qualifies as a behavioral biometric is still not universally accepted. In our detailed survey we have chosen to only cover approaches in which the behavior in question is under full or at least partial control of the individual exhibiting it. In this section, we present a number of approaches which have been classified as behavioral biometrics by other researchers in the field [141] and which as a rule are not under the full control of the subject.

A number of biological signals have been classified as behavioral biometrics in recent literature [141–143]. Numerous examples include the electrocardiogram (ECG), the electroencephalogram (EEG), and the electrooculogram (EOG) as well as some emerging technologies, like Brain-Computer Interface (BCI), and Electroencephalogram Interface (EEGI), NHCI (Neural Human-Computer Interface) and NI (Neural Interface) [144]. In addition to electrical activity, neural activity also generates other types of signals, for example magnetic and metabolic signals, that could be utilized in a BCI. Magnetic activity is recordable with magnetoencephalography (MEG), brain metabolic activity as mirrored by changes in blood flow can be measured with positron emission tomography (PET), and functional magnetic resonance imaging (fMRI) [142]. There are also invasive BCI signal recording methods such as implanted electrodes [143]. Table 13.3 provides known results for biosignal-based security systems. The following explanations are meant to increase the understanding of non-professionals regarding biosignals.

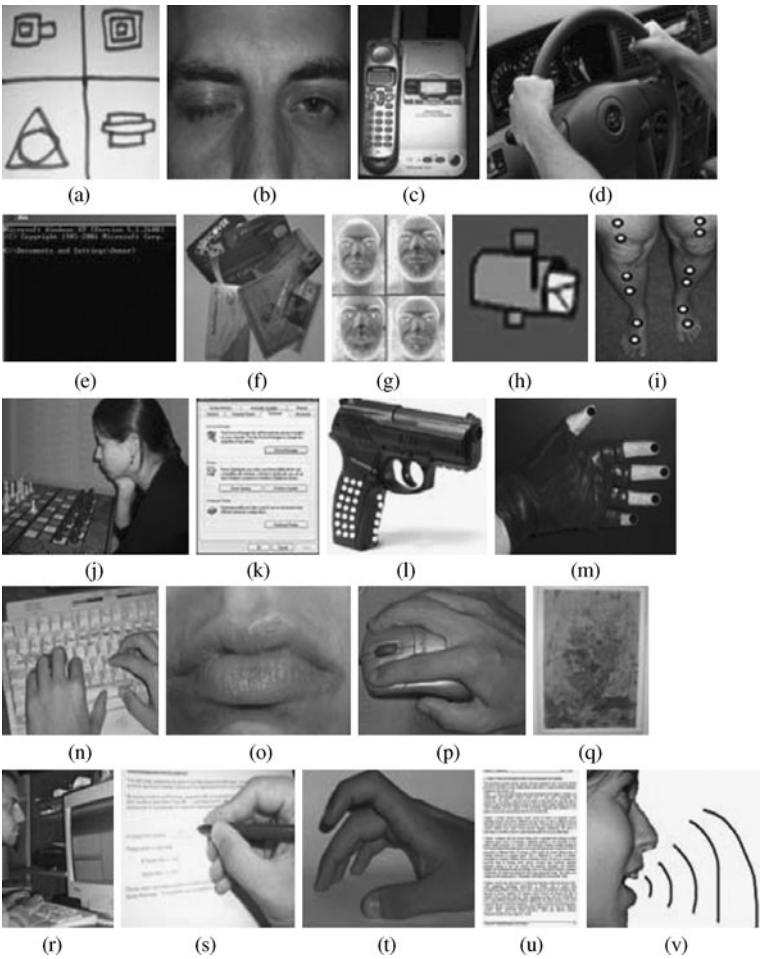


Fig. 13.1 Examples of Behavioral Biometrics: (**a**) Biometric Sketch, (**b**) Blinking, (**c**) Calling, (**d**) Car Driving, (**e**) Command Line Lexicon, (**f**) Credit Card Use, (**g**) Dynamic Facial Features, (**h**) Email, (**i**) Gait, (**j**) Game Strategy, (**k**) GUI Interaction, (**l**) Handgrip, (**m**) Haptic, (**n**) Keystrokes, (**o**) Lip Movement, (**p**) Mouse Dynamics, (**q**) Painting Style, (**r**) Programming Style, (**s**) Signature, (**t**) Tapping, (**u**) Text Authorship, (**v**) Voice [32]. Image used with permission from Interscience Publishers Ltd. © 2008

- **EEG [Electro Encephalo Gram]**: a graph of the brain's electrical activity versus time. The electrical activity is a result of electrical impulses traversing the neurons in the brain. Numerous studies demonstrate that the brainwave pattern of every individual is unique and that the EEG can be used for biometric identification [142]. The EEG signal changes with variation in types of cognitive activities. The signal itself can be isolated from the background noise through a series of filters. The idea behind this approach is to associate a particular EEG signature with a particular set of thoughts, such as recorded during human–computer inter-

Table 13.3 Accuracy rates for utilized biosignal-based biometrics

Biosignal	Publication	Accuracy Rate
PhonoCardioGram (PCG)	[150]	96%
ElectroCardioGram (ECG)	[141]	100%
ElectroEncephaloGram (EEG)	[145]	80–100%
PassThoughts	[143]	90%

action [141]. Correct classification of individual in the accuracy range of 80% to 100% has been achieved in recent experiments [145].

- **ECG/EKG [Electro Cardio/Kardio Gram]**: a graph of the heart's electrical activity versus time. The electrical activity is a result of the electric current flowing in both heart muscles and the neuronal network within the heart. Both EEG and ECG/EKG are extrapolations of the actual electric signals. A series of sensors are positioned over the heart and pick up the electrical signals produced by various regions of the heart during the pumping cycle. The recording of the heartbeat generates a unique and reliable profile for any particular individual. Recent experiments provide sufficient evidence to suggest that it is a highly discriminative biometric modality in some cases near 100% accurate [141, 146].
- **GSR [Galvanic Skin Response]**: a measure of the skin's resistance/conductance. This is affected by how moist the skin is (varying with the amount of sweat) and since the sweat glands are controlled by the nervous system this is an indirect measure of neuronal activity [147].
- **fTCD [functional Trans-Cranial Doppler]**: any Doppler scan uses ultrasound waves to visualize underlying organs or tissue. In the case of fTCD, it is used to visualize the brain. However, since it is functional, it visualizes the brain over time and shows variations with varying levels of activity (this type is called dynamic imaging).
- **Odor:** animals, for example dogs, are perfectly capable of recognizing people based on odor. Idea behind this type of authentication is to create an Electronic Nose (ENose) capable of sniffing out a person's identity. The ENose consists of a collection of sensors, each one serving as a receptor for a particular odor. Once a significant number of odors can be profiled by the system, it becomes an interesting pattern recognition problem to match odor-prints to people. This is a promising line of research and is still in the early stages of development, with no functional systems available on the market [148].
- **EP [Evoked Potential]**: measures brain's electrical activity generated from actively stimulating the patient. The stimulus in this case is mostly artificial.
- **ERP [Event Related Potential]**: also measures brain's electrical activity resulting from a stimulus. However, the stimulus in this case is some actual event rather than just an artificial factor.
- **PCG [PhonoCardioGram]:** essentially a recording of a cardiac sound, this biosignal has been successfully utilized in biometric identification systems after undergoing frequency analysis. Berittelli et al. [149] have demonstrated biometric applicability of PCG on a database of 20 subjects and obtained a FRR of 5.0%

and a FAR of 2.2%. The main advantage of using heart sound as a biometric is that it cannot be easily spoofed as compared to other, particularly non-physical biometric modalities. Preliminary results show that with optimally selected parameters, an identification rate of up to 96% is achievable for a small database of seven persons [150]. The heart beat is known as the Inherent Liveness Biometric because "The way the human heart beats" characteristic is only valid for a living person [151].

- **BVP [Blood Volume Pulse]:** uses photoplethysmography to detect the blood pressure in the extremities by applying a light source and measuring the light reflected by the skin. As blood is forced through the peripheral vessels by the heart, it produces engorgement of the vessels, thereby modifying the amount of light to the photosensor, which could be recorded as a waveform [152].
- **PassThoughts:** Thorpe et al. proposed using Brain Computer Interface (BCI) technology to have a user directly transmit his thoughts to a computer. The system extracts entropy from a user's brain signal upon reading a thought. The brain signals are processed in an accurate and repeatable way, providing a changeable authentication method. The potential size of the space of a PassThoughts system is not clear at this point, but likely to be very large, due to the lack of bounds on what composes a thought [143].

13.3.1 Behavioral Passwords

While behavioral passwords are not the same as biometrics, they are authentication methods based on preferences and psychological predispositions of people, and so clearly fall under computer analysis of human behavior. Therefore we include a short overview of the state of the art in this chapter, due to Yampolskiy [153].

13.3.1.1 Text-Based Behavioral Passwords

Text-Based passwords can be subdivided into syntactic, semantic and one-time methods. The classical passwords and passphrases are examples of syntactic methods, in which a user is expected to memorize a sequence of characters or words. The sequence can either be generated for the user, or user selected [42]. The problem is that a user's ability to memorize complicated or multiple passwords is limited, and so authentication may present problems for the user. Alternatively, easy to remember passwords are also easy to guess and so provide a low level of security. Some researchers present methods which might be easier for users to remember. For example, the Check-Off Password System (COPS) [154] allows users to enter characters in any order and therefore the users can choose to remember their password in many different ways. Each user is assigned eight different characters selected from the sixteen most commonly used letters. The user may use any character more than once to form words which are easy to remember and so it is claimed that COPS provides an advantage over regular passwords.

Semantic or cognitive passwords typically work by asking the user some questions and treating the user's answer as the key to the authentication mechanism. One approach described by Renaud [42] relies on asking the user clarifying questions until the answer matches the one expected by the system. An alternative technique provided a set of questionnaires, asking users to answer some fact-based or opinion-based questions [155]. These approaches are not very user-friendly, as it might take a long time for the user to arrive at the desired answer, and since users are very sensitive to the time component of an authentication protocol, the cognitive-based methods are not expected to become widely popular.

13.3.1.2 Graphics-Based Behavioral Passwords

Graphical passwords are designed to take advantage of human visual memory capabilities, which are far superior to our ability to remember textual information. Two main types of graphical passwords are currently in use: recognition-based and position-based, respectively. In recognition-based systems, users must identify images they have previously seen among new graphics.

Probably, the most well known recognition-based graphical authentication system is called Passfaces [156, 157]. It relies on the ease with which people recognize familiar faces. During enrollment, a user is presented with a set of faces from which a subset is selected, which the user is asked to memorize. During authentication, a screen with nine faces is presented to the user, with one of the faces being from his Passface set. User has to select a face, which is familiar from the enrollment step. This process is repeated five times, resulting in a relatively small space of 59,050 possible face combinations. Obviously, this is not sufficient if the system is open to an exhaustive search.

Another authentication system, Déjà Vu, is based on random art images. User is asked to choose five images as his pass set and during authentication needs to select his pass set from a challenge set of 25 pictures. Since the pictures used are completely random and are generated by a computer program, it is next to impossible to share a Déjà Vu password with others. Preliminary research shows that users prefer real photographs to random art images and that the enrollment phase is more time consuming than that of alphanumeric passwords [158].

The two systems mentioned above are probably representative of many other similar recognition-based graphical authentication systems currently in existence. Visual Identification Protocol [42, 159], Picture Password [160], and Picture-Pins [161] are all reliant on exploiting the users' good visual memory and power of recall to easily authenticate users by making them pick familiar images from a large set of graphics. A non-visual but also a sensory recall-based authentication approach utilizing music is presented in the work of Gibson et al. [162] on the Musipass.

The remaining authentication approaches presented in this review are graphical position-based systems. A typical position-based approach is presented in Pass-Points, a system based on having the user select points of interest within a single image. The number of points is not limited and so a relatively large search space protects against any attempt to guess a PassPoints authentication sequence [163, 164].

This is similar to the methodology used in the original patent for graphical passwords obtained by Blonder in 1996 [165].

An alternative to having a user select a portion of an image is to have a user input a simple drawing into a predefined grid space. This approach is attempted in [166] with a system called Passdoodles and also in [167, 168] with a system called Draw-a-Secret. Finally, a V-go Password requests a user to perform simulation of simple actions such as mixing a cocktail using a graphical interface [42].

There is also a separate area of research targeting development of password reminder cues based on different psychological and behavioral prompts. Primary examples of such cue eliciting systems are Inkblot cues [169–172] and Handwriting reminders [173]. Inkblot-based systems attempt to assist users in better recalling their passwords by providing implicit information, which users associate with their password. The idea is based on the concept of a Rorschach test, in which subject's perception of inkblots is recorded and analyzed in terms of everyday concepts.

13.3.2 Comparison and Analysis

Behavioral biometrics measure human actions, which can result from human skills, style, preference, knowledge, motor-skills or strategy. Table 13.4 summarizes what precisely is being measured by different behavioral biometrics, as well as lists some of the most frequently used features for each type of behavior. Indirect HCI-based biometrics are not included as they have no meaning independent of the direct human–computer interaction which causes them.

Motor-skill-based biometrics measure innate, unique and stable muscle actions of users performing a particular task. Table 13.5 outlines which muscle groups are responsible for a particular motor-skill, as well as lists some of the most frequently used features for each muscle control-based biometric approach.

While many behavioral biometrics are still in their infancy, some very promising research has already been done. The results obtained justify feasibility of using behavior for verification of individuals and further research in this direction is likely to improve accuracy of such systems. Table 13.6 summarizes obtained accuracy ranges for the set of direct behavioral biometrics for which such data are available. Table 13.7 presents accuracy rates for biometric methodologies not reviewed in our previous surveys.

13.4 Privacy Concerns

An unintended property of behavioral profiles is that they might contain information which may be of interest to third parties, who have potential to discriminate against individuals based on such information. As a consequence, intentionally revealing or obtaining somebody else's behavioral profile for the purposes other than verification is highly unethical. Examples of private information which might be revealed by some behavioral profiles follow.

Table 13.4 Behavioral biometrics with traits and features [32]

Behavioral Biometric	Measures	Features
Biometric Sketch	Knowledge	Location and relative position of different primitives
Calling Behavior	Preferences	Date and time of the call, duration, called ID, called number, cost of call, number of calls to a local destination, number of calls to mobile destinations, number of calls to international destinations
Car driving style	Skill	Pressure from accelerator pedal and brake pedal, vehicle speed, steering angle
Command Line Lexicon	Technical Vocabulary	Used commands together with corresponding frequency counts, and lists of arguments to the commands
Credit Card Use	Preferences	Account number, transaction type, credit card type, merchant ID, merchant address
Email Behavior	Style	Length of the emails, time of the day the mail is sent, how frequently inbox is emptied, the recipients' addresses
Game Strategy	Strategy/Skill	Count of hands folded, checked, called, raised, check-raised, re-raised, and times player went all-in
Haptic	Style	3D world location of the pen, average speed, mean velocity, mean standard deviation, navigation style, angular turns and rounded turns
Keystroke Dynamics	Skill	Time durations between the keystrokes, inter-key strokes and dwell times, which is the time a key is pressed down, overall typing speed, frequency of errors (use of backspace), use of numpad, order in which user presses shift key to get capital letters
Mouse Dynamics	Style	x and y coordinates of the mouse, horizontal velocity, vertical velocity, tangential velocity, tangential acceleration, tangential jerk and angular velocity
Painting Style	Style	Subtle pen and brush strokes characteristic
Programming Style	Skill, Style, Preferences	Chosen programming language, code formatting style, type of code editor, special macros, comment style, variable names, spelling and grammar, language features, the ratio of global to local variables, temporary coding structures, errors
Soft Behavioral Biometrics	Intelligence, Vocabulary, Skills	Word knowledge, generalization ability, mathematical skill
Text Authorship	Vocabulary	Sentence count, word count, punctuation mark count, noun phrase count, word included in noun phrase count, prepositional phrase count, word included in prepositional phrase count, and keyword count

Table 13.5 Motor-skill biometrics with respective muscles and features [174]

Motor Skill-based Biometric	Muscles Involved	Extracted Features
Blinking	orbicularis oculi, corrugator supercilii, depressor supercilii	time between blinks, how long the eye is held closed at each blink, physical characteristics the eye undergoes while blinking
Dynamic Facial Features	levator labii superioris, levator anguli oris zygomaticus major, zygomaticus minor, depressor labii inferioris, depressor anguli oris, buccinator, orbicularis oris	motion of skin pores on the face, skin folds, wrinkles
Gait/Stride	tibialis anterior, extensor hallucis longus, extensor digitorum longus, peroneus tertius, extensor digitorum brevis, extensor hallucis brevis, gastrocnemius, soleus, plantaris, popliteus, flexor hallucis longus flexor digitorum longus	amount of arm swing, rhythm of the walker, bounce, length of steps, vertical distance between head and foot, distance between head and pelvis, maximum distance between the left and right foot
Handgrip	abductor pollicis brevis, opponens pollicis, flexor pollicis brevis, adductor pollicis, palmaris brevis, abductor minimi digiti, flexor brevis minimi digiti	resistance measurements in multiple points
Haptic	abductor pollicis brevis, opponens pollicis, flexor pollicis brevis, adductor pollicis, palmaris brevis, abductor minimi digiti, flexor brevis minimi digiti, opponens digiti minimi, lumbrical, dorsal interossei, palmar interossei	3D world location of the pen, average speed, mean velocity, mean standard deviation, navigation style, angular turns and rounded turns
Keystroke Dynamics	abductor pollicis brevis, opponens pollicis, flexor pollicis brevis, adductor pollicis, palmaris brevis, abductor minimi digiti, flexor brevis minimi digiti, opponens digiti minimi, lumbrical, dorsal interossei, palmar interossei	time durations between the keystrokes, inter-key strokes and dwell times, which is the time a key is pressed down, overall typing speed, frequency of errors (use of backspace), use of numpad, order in which user presses shift key to get capital letters
Lip Movement	levator palpebrae superioris, levator anguli oris, mentalis, depressor labii inferioris, depressor anguli oris, buccinator, orbicularis oris, risorius	Mouth width, upper/lower lip width, lip opening height/width, distance between horizontal lip line and upper lip
Mouse Dynamics	abductor pollicis brevis, opponens pollicis, flexor pollicis brevis, adductor pollicis, palmaris brevis, abductor minimi digiti, flexor brevis minimi digiti, opponens digiti minimi, lumbrical, dorsal interossei	x and y coordinates of the mouse, horizontal velocity, vertical velocity, tangential velocity, tangential acceleration, tangential jerk and angular velocity

Table 13.5 (Continued)

Motor Skill-based Biometric	Muscles Involved	Extracted Features
Signature/ Handwriting	abductor pollicis brevis, opponens pollicis, flexor pollicis brevis, adductor pollicis, palmaris brevis, abductor minimi digiti, flexor brevis minimi digiti, opponens digiti minimi, lumbrical, dorsal interossei, palmar interossei	coordinates of the signature, pressure at pen tip, acceleration and pen-tilt, signing speed, signature bounding box, Fourier descriptors of the signature's trajectory, number of strokes, and signing flow
Tapping	abductor pollicis brevis, opponens pollicis, flexor pollicis brevis, adductor pollicis, palmaris brevis, abductor minimi digiti, flexor brevis minimi digiti	Pulse height, pulse duration, and the duration of the first inter-pulse interval
Voice/Speech	cricothyroid, posterior ricoarytenoid, lateral cricoarytenoid, arytenoid, thyroarytenoid	logarithm of the Fourier transform of the voice signal in each band along with pitch and tone

- **Calling behavior:** Calling data are a particularly sensitive subject since they might contain highly personal information.
- **Car driving style:** Car insurance companies may be interested to know if a driver frequently speeds or whether he or she is an overall aggressive driver, in order to charge an increased coverage rate or to deny coverage all together.
- **Command line lexicon:** Information about proficiency with the commands might be used by an employer to decide if you are sufficiently qualified for a job involving computer interaction.
- **Credit card usage:** Credit card data reveal information about what items you frequently purchase and in what locations you can be found, violating your expectation of privacy. For example an employer might be interested to know if an employee buys a case of beer every day, indicating a problem of alcoholism.
- **Email behavior:** An employer would be interested to know if employees send out personal emails during office hours.
- **Game strategy:** If information about game strategy is obtained by the player's opponents, they might be analyzed to find weaknesses in player's game and as a result give an unfair advantage to the opponents.
- **Programming style:** Software metrics obtained from analysis of code may indicate a poorly performing coder and as a result jeopardize the person's employment.

Additionally, any of the motor-skill-based biometrics may reveal a physical handicap of a person and so result in potential discrimination. Such biometrics as voice can reveal emotions, and the face images may reveal information about emotions and health [181]. Because behavioral biometric indirectly measures our thoughts

Table 13.6 Recognition and error rates of behavioral biometrics [32]

Behavioral Biometric	Publication	Detection Rate	FAR	FRR	EER
Biometric Sketch	Bromme 2003 [40]				7.2%
Blinking	Westeyn 2004 [44]	82.02%			
Calling Behavior	Fawcett 1997 [47]	92.5%			
Car driving style	Erdogan 2005 [175]	88.25%			4.0%
Command Line Lexicon	Marin 2001 [176]	74.4%		33.5%	
Credit Card Use	Brause 1999 [56]	99.995%		20%	
Email Behavior	de Vel 2001 [64]	90.5%			
Gait/Stride	Kale 2004 [77]	90%			
Game Strategy	Yampolskiy 2007 [83]				7.0%
Handgrip	Veldhuis 2004 [88]				1.8%
Haptic	Orozco 2006 [90]		25%		22.3%
Keystroke Dynamics	Bergadano 2002 [177]		0.01%	4%	
Lip Movement	Mok 2004 [103]				2.17%
Mouse Dynamics	Pusara 2004 [23]		0.43%	1.75%	
Programming Style	Frantzeskou 2004 [178]	73%			
Signature	Jain 2002 [111]		1.6%	2.8%	
Handwriting	Zhu 2000 [121]	95.7%			
Tapping	Henderson 2001 [127]				2.3%
Text Authorship	Halteren 2004 [129]		0.2%	0.0%	
Voice/Speech	Colombi 1996 [179]				0.28%
Singing	Tsai 2006 [180]				29.6%

Table 13.7 Accuracy rates for new biometric modalities previously not reviewed

Behavioral Biometric	Publication	Detection Rate	FAR	FRR	EER
Center of Gravity	[52]	50%			18%
Finger Pressure	[65]	99%			1%
Floor Pressure	[66]	92.3%	6.79%		
Gaze/Eye Tracking	[75, 76]	100%			
Human Shadows	[92, 93]	90%			
Motion of Fingers	[108]	97%			
Short Term Memory	[122, 123]	90%	0.52%	26.14%	
Visual Scan/Search	[122, 123]				3.88%

and personal traits any data collected in the process of generation of a behavioral profile need to be safely stored in an encrypted form.

13.5 Summary

This chapter presented an overview and classification of security approaches based on computer analysis of human behavior. In particular the following broad categories of behavior-based authentication mechanisms were examined: Behavioral Biometrics (Authorship based, Human Computer Interaction Based, Motor Skill, and Purely Behavioral), Behavioral Passwords (syntactic, semantic, one-time methods and visual memory based), Biosignals (Cognitive and semi-controllable biometrics) and Virtual Biometrics (representations of users in virtual worlds).

We have presented only the most popular behavioral biometrics, but any human behavior can be used as a basis for personal profiling and for subsequent verification. Some behavioral biometrics, which are quickly gaining ground, but are not a part of this chapter include profiling of shopping behavior based on market basked analysis [182], web browsing and click-stream profiling [183–185], and even TV preferences [186]. To make it easier to recognize newly proposed approaches as behavioral biometrics, we propose a definition of what properties constitute a behavioral biometric characteristic. We define a behavioral biometric as any quantifiable actions of a person. Such actions may not be unique to the person and may take a different amount of time to be exhibited by different individuals.

Behavioral biometrics are particularly well suited for verification of users, who interact with computers, cell phones, smart cars, or points of sale terminals. As the number of electronic appliances used in homes and offices increases, so does the potential for utilization of this novel and promising technology. Future research should be directed at increasing overall accuracy of such systems, for example by looking into possibility of developing multimodal behavioral biometrics, as people often engage in multiple behaviors at the same time, for example, talking on a cell phone while driving, or using keyboard and mouse at the same time [187–189].

Fields as diverse as marketing, game theory, security and law enforcement all can greatly benefit from accurate modeling of human behavior. One of the aims of this chapter was to show that the problem at hand is not unique to any given field and that a solution found once might benefit many industries without a need for re-discovering it for each subfield.

Because many of the presented technologies represent behavioral biometrics which are not strong enough to serve as a backbone of a complete security system on their own, we suggest that a lot of research in behavioral biometrics be geared toward multimodal behavioral biometrics. Successful research in this area would allow for development of systems with accuracy levels sufficient not just for identity verification, but also for person identification obtained as a result of combining different behaviors. Breakthroughs in purely behavioral biometrics research will also undoubtedly lead to improvements in associated applications such as product customization, development of tailored opponents in games as well as multitude of competency assessment tools.

Future of behavioral research looks very bright. The next decade will bring us technologies providing unprecedented level of security, product customization, social compatibility and work efficiency. Ideas presented in the section on novel behavioral biometrics provide a wealth of opportunities for interesting research and

development. A great side effect of such research would be general greater understanding of human behavior, personality and perhaps human mind itself.

13.6 Questions

(1) Describe different authentication mechanism categories presented in the chapter.
(2) Which behaviors tend to have the highest degree of uniqueness leading to better authentication accuracies?
(3) List behavioral biometrics classified in multiple categories and explain underlining reasons for that.
(4) What are the issues of concern with use of biometrics based on analysis of human behavior?
(5) If you were designing a behavioral biometric-based security system, which behavior would you select and why?

13.7 Glossary

- *Acceptability*: Willingness of people to utilize a biometric modality.
- *Authentication*: The act of confirming identity.
- *Biometric*: Intrinsic physical or behavioral characteristic.
- *Behavioral Biometric*: Biometric based on the behavior of a person.
- *Circumvention*: A way to bypass biometric authentication.
- *Collectability*: Easy acquisition of biometric data.
- *FAR*: False Accept Rate, the likelihood that the biometric system will incorrectly match an individual to the wrong template in the database
- *FRR*: False Reject Rate, the likelihood that the biometric system will fail to detect a match between a person and the correct template in the database.
- *Feature*: A distinguishing characteristic of a pattern.
- *Performance*: Accuracy, speed, and robustness of the biometric algorithm.
- *Permanence*: Invariance of a biometric trait with respect to time.
- *Recognition*: Identification of an individual from a list of known users.
- *Uniqueness*: Discriminative ability of a biometric modality.
- *Verification*: Confirmation used to verify that the individual is who he claims to be.
- *Universality*: The need for universal availability of a biometric characteristic in all individuals.

References

1. Angle, S., Bhagtani, R., Chheda, H.: Biometrics: a further echelon of security. In: First UAE International Conference on Biological and Medical Physics (2005)

2. Dugelay, J.-L., et al.: Recent advances in biometric person authentication. In: IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), Special Session on Biometrics, Orlando, Florida (2002)
3. Lee, K., Park, H.: A new similarity measure based on intraclass statistics for biometric systems. ETRI J. **25**(5), 401–406 (2003)
4. Cappelli, R., et al.: Performance evaluation of fingerprint verification systems. IEEE Trans. Pattern Anal. Mach. Intell. **28**(1), 3–18 (2006)
5. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. In: IEEE Trans. Circuits Syst. Video Technol. (2004)
6. Bolle, R., et al.: Guide to Biometrics. Springer, Berlin (2003)
7. Jain, A.K., et al.: Biometrics: a grand challenge. In: International Conference on Pattern Recognition, Cambridge, UK (2004)
8. Uludag, U., et al.: Biometric cryptosystems: issues and challenges. Proc. IEEE **92**(6) (2004)
9. Delac, K., Grgic, M.: A survey of biometric recognition methods. In: 46th International Symposium Electronics in Marine, ELMAR-2004, Zadar, Croatia (2004)
10. Ruggles, T.: Comparison of biometric techniques (2007). Available at: http://www.bio-tech-inc.com/bio.htm
11. Solayappan, N., Latifi, S.: A survey of unimodal biometric methods. In: Security and Management, Las Vegas, Nevada, USA (2006)
12. Bioprivacy.org: FAQ. BioPrivacy Initiative (2005). July 22, 2005. Available from: http://www.bioprivacy.org/faqmain.htm
13. Bromme, A.: A classification of biometric signatures. In: International Conference on Multimedia and Expo (ICME '03) (2003)
14. Yampolskiy, R.V.: Human computer interaction based intrusion detection. In: 4th International Conference on Information Technology: New Generations (ITNG 2007), Las Vegas, Nevada, USA (2007)
15. Bioprivacy.org. FAQ's and Definitions. International Biometric Group, LLC (2005). October 2, 2005. Available from: http://www.bioprivacy.org/bioprivacy_text.htm
16. Yampolskiy, R.V.: Indirect human–computer interaction-based biometrics for intrusion detection systems. In: The 41st Annual IEEE International Carnahan Conference on Security Technology (ICCST 2007), Ottawa, Canada (2007)
17. Denning, D.E.: An intrusion-detection model. In: IEEE Transactions on Software Engineering (1987)
18. Ilgun, K., Kemmerer, R.A., Porras, P.A.: State transition analysis: A rule-based intrusion detection approach. In: Software Engineering (1995)
19. Ghosh, A.K., Schwartzbard, A., Schatz, M.: Learning program behavior proles for intrusion detection. In: First USENIX Workshop on Intrusion Detection and Network Monitoring (1999)
20. Apap, F., et al.: Detecting malicious software by monitoring anomalous windows registry accesses. In: Fifth International Symposium on Recent Advances in Intrusion Detection, pp. 16–18 (2002)
21. Pennington, A.G., et al.: Storage-based intrusion detection: Watching storage activity for suspicious behavior. Carnegie Mellon University (2002)
22. Feng, H.H., et al.: Anomaly detection using call stack information. In: Proceedings of IEEE Symposium on Security and Privacy (2003)
23. Pusara, M., Brodley, C.E.: User re-authentication via mouse movements. In: VizSEC/DMSEC '04: Proceedings of the ACM Workshop on Visualization and Data Mining for Computer Security. ACM, Washington (2004)
24. Garg, A., et al.: Profiling users in GUI based systems for masquerade detection. In: The 7th IEEE Information Assurance Workshop (IAWorkshop 2006), West Point, New York, USA (2006)
25. Yampolskiy, R.V.: Motor-skill based biometrics. In: Dhillon, G. (ed.) Assuring Business Processes, Proceedings of the 6th Annual Security Conference. Global Publishing, Las Vegas (2007)

26. Caslon.com.au: Caslon-Analytics. October 2, 2005. Available from: http://www.caslon.com.au/biometricsnote8.htm
27. Jain, A.K., Bolle, R., Pankanti, S.: BIOMETRICS: Personal Identification in Networked Society. Kluwer Academic, Dordrecht (1999)
28. Adler, A., Youmaran, R., Loyka, S.: Towards a measure of biometric information (2006). Available at: http://www.sce.carleton.ca/faculty/adler/publications/2006/youmaran-ccece2006-biometric-entropy.pdf
29. Koychev, I., Schwab, I.: Adaptation to drifting user's interests. In: Proceedings of ECML2000 Workshop: Machine Learning in New Information Age, Barcelona, Spain (2000)
30. Tsymbal, A.: The problem of concept drift: definitions and related work. Technical Report TCD-CS-2004-15, Computer Science Department, Trinity College, Dublin, Ireland (2004)
31. Schuckers, S.A.C.: Spoofing and anti-spoofing measures. Information Security Technical Report (2002)
32. Yampolskiy, R.V., Govindaraju, V.: Behavioral biometrics: a survey and classification. Int. J. Biom. **1**(1), 81–113 (2008)
33. Oursler, J.N., Price, M., Yampolskiy, R.V.: Parameterized generation of Avatar face dataset. In: 14th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games, Louisville, KY (2009)
34. Yampolskiy, R., Gavrilova, M.: Applying biometric principles to avatar recognition. In: International Conference on Cyberworlds (CW2010), Singapore, October 20–22 (2010)
35. Ajinal, S., Yampolskiy, R.V., Amara, N.E.B.: Authentification de Visages D'Avatar. In: Confere 2010 Symposium, Sousse, Tunisia, July 1–2 (2010)
36. Yampolskiy, R.V., Govindaraju, V.: Behavioral biometrics for verification and recognition of malicious software agents. In: Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense VII. SPIE Defense and Security Symposium, Orlando, Florida, March 16–20 (2008)
37. D'Souza, D., Yampolskiy, R.V.: Avatar face detection analysis using an extended set of Haar-like features. In: Kentucky Academy of Science, Annual Meeting, Bowling Green, Kentucky, November 12–13 (2010)
38. Lyons, M., et al.: Avatar creation using automatic face recognition. In: ACM Multimedia 98, Bristol, England, Sept. 1998, pp. 427–434 (1998)
39. Yanushkevich, S., et al.: Image Pattern Recognition: Synthesis and Analysis in Biometrics. Machine Perception and Artificial Intelligence, vol. 67. World Scientific, Singapore (2007)
40. Brömme, A., Al-Zubi, S.: Multifactor biometric sketch authentication. In: BIOSIG, Darmstadt, Germany (2003)
41. Al-Zubi, S., Brömme, A., Tönnies, K.: Using an active shape structural model for biometric sketch recognition. In: DAGM, Magdeburg, Germany (2003)
42. Renaud, K.: Quantifying the quality of web authentication mechanisms. A usability perspective. J. Web Eng. (2003). Available at: http://www.dcs.gla.ac.uk/~karen/Papers/j.pdf
43. Westeyn, T., et al.: Biometric identification using song-based eye blink patterns. In: Human Computer Interaction International (HCII), Las Vegas, NV (2005)
44. Westeyn, T., Starner, T.: Recognizing song-based blink patterns: applications for restricted and universal access. In: Sixth IEEE International Conference on Automatic Face and Gesture Recognition (2004)
45. Hilas, C., Sahalos, J.: User profiling for fraud detection in telecommunication networks. In: 5th International Conference on Technology and Automation (ICTA 2005), Thessaloniki, Greece (2005)
46. Grosser, H., Britos, H., García-Martínez, R.: Detecting Fraud in Mobile Telephony Using Neural Networks. Lecture Notes in Artificial Intelligence. Springer, Berlin (2005)
47. Fawcett, T., Provost, F.: Adaptive fraud detection. In: Data Mining and Knowledge Discovery. Kluwer Academic, Dordrecht (1997)
48. Erzin, E., et al.: Multimodal person recognition for human-vehicle interaction. In: IEEE MultiMedia (April 2006)

49. Liu, A., Salvucci, D.: Modeling and prediction of human driver behavior. In: 9th HCI International Conference, New Orleans, LA (2001)
50. Oliver, N., Pentland, A.P.: Graphical models for driver behavior recognition in a SmartCar. In: Proceedings of the IEEE Intelligent Vehicles Symposium (2000)
51. Kuge, N., Yamamura, T., Shimoyama, O.: A driver behavior recognition method based on driver model framework. In: Society of Automotive Engineers Publication (1998)
52. Porwik, P., et al.: Biometric recognition system based on the motion of the human body gravity centre analysis. J. Med. Inform. Technol. **15** (2010)
53. Schonlau, M., et al.: Computer intrusion: detecting masquerades. Stat. Sci. **16**(1), 1–17 (2001)
54. Maxion, R.A., Townsend, T.N.: Masquerade detection using truncated command lines. In: International Conference on Dependable Systems and Networks (DNS-02). IEEE Comput. Soc., Los Alamitos (2002)
55. Dao, V., Vemuri, V.: Profiling users in the UNIX OS environment. In: International ICSC Conference on Intelligent Systems and Applications, University of Wollongong, Australia (2000)
56. Brause, R., Langsdorf, T., Hepp, M.: Neural data mining for credit card fraud detection. In: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence (1999)
57. Pamudurthy, S., et al.: Dynamic approach for face recognition using digital image skin correlation. In: Audio- and Video-based Biometric Person Authentication (AVBPA), New York (2005)
58. Mainguet, J.-F.: Biometrics (2006). Available at: http://perso.orange.fr/fingerchip/biometrics/biometrics.htm
59. Ito, A., et al.: Smile and laughter recognition using speech processing and face recognition from conversation video. In: Proceedings of the International Conference on Cyberworlds (2005)
60. Tsai, P., Hintz, T., Jan, T.: Facial behavior as behavior biometric? An empirical study. In: IEEE International Conference on Systems, Man and Cybernetics, Montreal, Quebec, October 7–10, pp. 3917–3922 (2007)
61. Benedikt, L., et al.: Assessing the uniqueness and permanence of facial actions for use in biometric applications. IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. **40**(3), 449–460 (2010)
62. Stolfo, S.J., et al.: A behavior-based approach to securing email systems. In: Mathematical Methods, Models and Architectures for Computer Networks Security. LNCS, vol. 2776, pp. 57–81 (2003)
63. Stolfo, S.J., et al.: Combining behavior models to secure email systems. CU Tech Report (2003). Available at: www1.cs.columbia.edu/ids/publications/EMT-weijen.pdf
64. Vel, O.D., et al.: Mining email content for author identification forensics. SIGMOD Rec. **30**(4), 55–64 (2001). Special Section on Data Mining for Intrusion Detection and Threat Analysis
65. Saevanee, H., Bhattarakosol, P.: Authenticating user using keystroke dynamics and finger pressure. In: 6th IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, January 10–13, pp. 1–2 (2009)
66. Qian, G., Zhang, J., Kidane, A.: People identification using gait via floor pressure analysis IEEE Sens. J. **10**(9), 1447–1460 (2010)
67. Addlesee, M., et al.: The ORL active floor. IEEE Pers. Commun. 35–41 (1997)
68. Jung, J., et al.: Dynamic-footprint based person identification using mat-type pressure sensor. In: International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2937–2940 (2003)
69. Pirttikangas, S., et al.: Footstep identification from pressure signals using hidden Markov models. In: Finnish Signal Processing Symposium, pp. 124–128 (2003)
70. Middleton, L., et al.: A floor sensor system for gait recognition. In: IEEE Workshop on Automatic Identification Advanced Technologies, pp. 171–176 (2005)

71. Yun, J., et al.: The user identification system using walking pattern over the ubiFloor. In: International Conference on Control, Automation, and Systems, pp. 1046–1050 (2003)
72. Orr, R.J., Abowd, G.D.: The smart floor: a mechanism for natural user identification and tracking. In: Conference on Human Factors in Computing Systems, pp. 275–276 (2000)
73. Yoon, J., Ryu, J., Woo, W.: User identification using user's walking pattern over the ubiFloorII. In: International Conference on Computational Intelligence and Security, pp. 949–956 (2005)
74. Suutala, J., Röning, J.: Methods for person identification on a pressure-sensitive floor: Experiments with multiple classifiers and reject option. Inf. Fusion **9**(1), 21–40 (2008)
75. Maeder, A.J., Fookes, C.B.: A visual attention approach to personal identification. In: Eighth Australian and New Zealand Intelligent Information Systems Conference, December 10–12 (2003)
76. Maeder, A.J., Fookes, C.B., Sridharan, S.: Gaze based user authentication for personal computer applications. In: International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, China, October 20–22 (2004)
77. Kale, A., et al.: Identification of humans using gait. IEEE Trans. Image Proc. **13**(9) (2004)
78. BenAbdelkader, C., Cutler, R., Davis, L.: Person identification using automatic height and stride estimation. In: IEEE International Conference on Pattern Recognition (2002)
79. Nixon, M.S., Carter, J.N.: On gait as a biometric: progress and prospects. In: EUSIPCO, Vienna (2004)
80. Kalyanaraman, S.: Biometric authentication systems. A report. 2006. Available at: http://netlab.cs.iitm.ernet.in/cs650/2006/TermPapers/sriramk.pdf
81. Yampolskiy, R.V.: Behavior based identification of network intruders. In: 19th Annual CSE Graduate Conference (Grad-Conf2006), Buffalo, NY (2006)
82. Yampolskiy, R.V., Govindaraju, V.: Use of behavioral biometrics in intrusion detection and online gaming. In: Biometric Technology for Human Identification III. SPIE Defense and Security Symposium, Orlando, Florida (2006)
83. Yampolskiy, R.V., Govindaraju, V.: Dissimilarity functions for behavior-based biometrics. In: Biometric Technology for Human Identification IV. SPIE Defense and Security Symposium, Orlando, Florida (2007)
84. poker-edge.com: Stats and analysis (2006). June 7, 2006. Available from: http://www.poker-edge.com/stats.php
85. Ramon, J., Jacobs, N.: Opponent modeling by analysing play. In: Proceedings of the Computers and Games workshop on Agents in Computer Games, Edmonton, Alberta, Canada (2002)
86. Jansen, A.R., Dowe, D.L., Farr, G.E.: Inductive inference of chess player strategy. In: Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence (PRICAI'2000) (2000)
87. Kauffman, J.A., et al.: Grip-pattern recognition for smart guns. In: 14th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC), Veldhoven, The Netherlands (2003)
88. Veldhuis, R.N.J., et al.: Biometric verification based on grip-pattern recognition. In: Security, Steganography, and Watermarking of Multimedia Contents (2004)
89. Orozco, M., et al.: Automatic identification of participants in haptic systems. In: IEEE Instrumentation and Measurement Technology Conference, Ottawa, Canada (2005)
90. Orozco, M., et al.: Haptic-based biometrics: a feasibility study. In: IEEE Virtual Reality Conference, Alexandria, Virginia, USA (2006)
91. Trujillo, M.O., Shakra, I., Saddik, A.E.: Haptic: the new biometrics-embedded media to recognizing and quantifying human patterns. In: MULTIMEDIA '05: Proceedings of the 13th Annual ACM International Conference on Multimedia, Hilton, Singapore. ACM, New York (2005)
92. Stoica, A.: Towards recognition of humans and their behaviors from space and airborne platforms: extracting the information in the dynamics of human shadows. In: Symposium on Bio-inspired Learning and Intelligent Systems for Security (BLISS '08), Edinburgh, August 4–6, pp. 125–128 (2008)

93. Iwashita, Y., Stoica, A., Kurazume, R.: Person identification using shadow analysis. In: British Machine Vision Conference, September, pp. 35.1–35.10 (2010)
94. Ilonen, J.: Keystroke dynamics (2006). Available at: www.it.lut.fi/kurssit/03-04/010970000/seminars/Ilonen.pdf
95. Bella, S.D., Palmer, C.: Personal identifiers in musicians' finger movement dynamics. J. Cogn. Neurosci. **18** (2006)
96. Gamboa, H., Fred, A.L.N., Jain, A.K.: Webbiometrics: User verification via web interaction. In: Biometrics Symposium, Baltimore, MD, September 11–13, pp. 1–6 (2007).
97. Shipilova, O.: Person recognition based on lip movements (2006). Available at: http://www.it.lut.fi/kurssit/03-04/010970000/seminars/Shipilova.pdf
98. Broun, C.C., et al.: Automatic speechreading with applications to speaker verification. In: Eurasip Journal on Applied Signal Processing, Special Issue on Joint Audio-Visual Speech Processing (2002)
99. Luettin, J., Thacker, N.A., Beet, S.W.: Speaker identification by lipreading. In: Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96) (1996)
100. Wark, T., Thambiratnam, D., Sridharan, S.: Person authentication using lip information. In: Proceedings of IEEE 10th Annual Conference. Speech and Image Technologies for Computing and Telecommunications (1997)
101. Mason, J.S.D., et al.: Lip signatures for automatic person recognition. In: IEEE Workshop, MMSP (1999)
102. Jourlin, P., et al.: Acoustic-labial speaker verification. In: Pattern Recognition Letters (1997)
103. Mok, L., et al.: Person authentication using ASM based lip shape and intensity information. In: International Conference on Image Processing (2004)
104. Ahmed, A.A.E., Traore, I.: Detecting computer intrusions using behavioral biometrics. In: Third Annual Conference on Privacy, Security and Trust, St. Andrews, New Brunswick, Canada (2005)
105. Ahmed, A.A.E., Traore, I.: Anomaly intrusion detection based on biometrics. In: Workshop on Information Assurance, United States Military Academy, West Point, NY (2005)
106. Gamboa, H., Fred, V.-A.: A behavioral biometric system based on human–computer interaction. In: Proceedings of SPIE (2004)
107. Gamboa, H., Fred, A.: An identity authentication system based on human–computer interaction behaviour. In: Proc. of the 3rd Intl. Workshop on Pattern Recognition in Information Systems (2003)
108. Nishiuchi, N., Komatsu, S., Yamanaka, K.: A biometric identification using the motion of fingers. In: International Conference on Biometrics and Kansei Engineering, Cieszyn, Poland, June 25–28, pp. 22–27 (2009)
109. Lyu, S., Rockmore, D., Farid, H.: A digital technique for art authentication. In: Proceedings of the National Academy of Sciences (2004)
110. Spafford, E.H., Weeber, S.A.: Software forensics: can we track code to its authors? In: 15th National Computer Security Conference (1992)
111. Jain, A., Griess, F., Connell, S.: On-line signature verification. Pattern Recognit. **35**, 2963–2972 (2002)
112. Nalwa, V.S.: Automatic on-line signature verification. Proc. IEEE **85**, 215–239 (1997)
113. Herbst, B., Coetzer, H.: On an offline signature verification system. In: Proceedings of the 9th Annual South African Workshop on Pattern Recognition (1998)
114. Lei, H., Palla, S., Govindaraju, V.: ER2: an intuitive similarity measure for on-line signature verification. In: IWFHR '04: Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR'04). IEEE Comput. Soc., Los Alamitos (2004)
115. Riha, Z., Matyas, V.: Biometric authentication systems. In: FI MU Report Series (2000)
116. Muralidharan, N., Wunnava, S.: Signature verification: a popular biometric technology. In: Second LACCEI International Latin American and Caribbean Conference for Engineering and Technology (LACCEI'2004), Miami, Florida, USA (2004)
117. Plamondon, R., Lorette, G.: Automatic signature verification and writer identification: the state of the art. Pattern Recognit. **22**(2), 107–131 (1989)

118. Ballard, L., Monrose, F., Lopresti, D.P.: Biometric authentication revisited: understanding the impact of wolves in sheep's clothing. In: Fifteenth USENIX Security Symposium, Vancouver, BC, Canada (2006)
119. Ramann, F., Vielhauer, C., Steinmetz, R.: Biometric applications based on handwriting. In: IEEE International Conference on Multimedia and Expo (ICME '02) (2002)
120. Ballard, L., Lopresti, D., Monrose, F.: Evaluating the security of handwriting biometrics. In: The 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR06), La Baule, France (2006)
121. Zhu, Y., Tan, T., Wang, Y.: Biometric personal identification based on handwriting. In: 15th International Conference on Pattern Recognition (ICPR'00) (2000)
122. Hamdy, O., Traoré, I.: Cognitive-based biometrics system for static user authentication. In: Fourth International Conference on Internet Monitoring and Protection, Venice/Mestre, Italy, May 24–28, pp. 90–97 (2009)
123. Hamdy, O., Traoré, I.: New physiological biometrics based on human cognitive factors. In: International Conference on Complex, Intelligent and Software Intensive Systems, Fukuoka, Japan, March 16–19, pp. 910–917 (2009)
124. Jain, A.K., Dass, S.C., Nandakumar, K.: Can soft biometric traits assist user recognition. In: SPIE Defense and Security Symposium, Orlando, FL (2004)
125. Jain, A.K., Dass, S.C., Nandakumar, K.: Soft biometric traits for personal recognition systems. In: International Conference on Biometric Authentication (ICBA), Hong Kong (2004)
126. Jacob, B.A., Levitt, S.D.: To catch a cheat. In: Education Next. Available at: www.educationnext.org (2004)
127. Henderson, N.Y., et al.: Polymer thick-film sensors: possibilities for smartcard biometrics. In: Proceedings of Sensors and Their Applications XI (2001)
128. Henderson, N.J., et al.: Sensing pressure for authentication. In: 3rd IEEE Benelux Signal Processing Symp. (SPS), Leuven, Belgium (2002)
129. Halteren, H.V.: Linguistic profiling for author recognition and verification. In: Proceedings of ACL-2004 (2004)
130. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic authorship attribution. In: Ninth Conf. European Chap. Assoc. Computational Linguistics, Bergen, Norway (1999)
131. Juola, P., Sofko, J.: Proving and improving authorship attribution. In: Proceedings of CaSTA-04. The Face of Text (2004)
132. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: 21st International Conference on Machine Learning, Banff, Canada (2004)
133. Koppel, M., Schler, J., Mughaz, D.: Text categorization for authorship verification. In: Eighth International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, Florida (2004)
134. Ciota, Z.: Speaker verification for multimedia application. In: IEEE International Conference on Systems, Man and Cybernetics (2004)
135. Sanderson, C., Paliwal, K.K.: Information fusion for robust speaker verification. In: Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH'01), Aalborg (2001)
136. Campbell, J.P.: Speaker recognition: a tutorial. Proc. IEEE **85**(9), 1437–1462 (1997)
137. Ratha, N.K., Senior, A., Bolle, R.M.: Automated biometrics. In: International Conference on Advances in Pattern Recognition, Rio de Janeiro, Brazil (2001)
138. Deshpande, S., Chikkerur, S., Govindaraju, V.: Accent classification in speech. In: Fourth IEEE Workshop on Automatic Identification Advanced Technologies (2005)
139. Lin, X., Simske, S.: Phoneme-less hierarchical accent classification. In: Thirty-Eighth Asilomar Conference on Signals, Systems and Computers (2004)
140. Tsai, W.-H., Wang, H.-M.: Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. IEEE Trans. Audio Speech Lang. Process. **14**(1), 330–341 (2006)
141. Revett, K.: Behavioral Biometrics: A Remote Access Approach. Wiley, Chichester (2008)
142. Marcel, S., Millan, J.: Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 743–752 (2007)

143. Thorpe, J., Oorschot, P.C.V., Somayaji, A.: Pass-thoughts: authenticating with our minds. In: Workshop on New Security Paradigms, Lake Arrowhead, California (2011)
144. Lawson, W.: The new wave ("Biometric access & neural control") (2002). November 24, 2008. Available from: http://www.icdri.org/biometrics/new_wave.htm
145. Mohammadi, G., et al.: Person identification by using AR model for EEG signals. In: World Academy of Science, Engineering and Technology (2006)
146. Gahi, Y., et al.: In: New Technologies, Mobility and Security (NTMS'08), Tangier, Virginia, November 5–7, pp. 1–5 (2008)
147. Shye, A., et al.: Power to the people: leveraging human physiological traits to control microprocessor frequency. In: 41st IEEE/ACM International Symposium on Microarchitecture, Como, Italy, November 8–12 (2008)
148. Korotkaya, Z.: Biometrics person authentication: odor (2003). October 12, 2008. Available from: http://www.it.lut.fi/kurssit/03-04/010970000/seminars/Korotkaya.pdf
149. Beritelli, F., Serrano, S.: Biometric identification based on frequency analysis of cardiac sounds. IEEE Trans. Inf. Forensics Secur. **2**(3), 596–604 (2007)
150. Phua, K., et al.: Human identification using heart sound. In: Second International Workshop on Multimodal User Authentication, Toulouse, France (2006)
151. Preez, J., Soms, S.H.: Person identification and authentication by using "the way the heart beats". In: ISSA 2005 New Knowledge Today Conference, Sandton, South Africa (2005)
152. Scotti, S., et al.: Quantitative evaluation of distant student psychophysical responses during the e-learning processes. In: 27th IEEE Annual Conference on Engineering in Medicine and Biology, Shanghai, China, September 1–4 (2005)
153. Yampolskiy, R.V.: Action based user authentication. Int. J. Electron. Secur. Digit. Forensics **1**(3), 281–300 (2008)
154. Bekkering, E., Warkentin, M., Davis, K.: A longitudinal comparison of four password procedures. In: Proceedings of the Hawaii International Conference on Business, Honolulu, HI, June (2003)
155. Podd, J., Bunnell, J., Henderson, R.: Cost-effective computer security: cognitive and associative passwords. In: Sixth Australian Conference on Computer-Human Interaction, Hamilton, New Zealand, November 24–27, pp. 304–305 (1996)
156. Brostoff, A.: Improving password system effectivness. PhD Dissertation, Department of Computer Science University College London, September 30, 2004
157. Brostoff, A.: The science behind passfaces. In: Real User Corporation, June 2004. Available at: http://www.realuser.com/
158. Dhamija, R., Perrig, A.: Deja vu: a user study. Using images for authentication. In: Proceedings of the 9th USENIX Security Symposium, Denver, Colorado, August (2000)
159. Angeli, A.D., et al.: Usability and user authentication: Pictorial passwords vs. PIN. In: Contemporary Ergonomics, pp. 253–258. Taylor & Francis, London (2003)
160. Jansen, W., et al.: Picture password: a visual login technique for mobile devices. Retrieved October 24, 2005. Available at: http://csrc.nist.gov/publications/nistir/nistir-7030.pdf
161. Pointsec. PicturePINs. November, 2002. Available at: http://www.pointsec.com/news/download/Pointsec_PPC_2.0_POP_PA1.pdf
162. Gibson, M., et al.: Musipass: authenticating me softly with my song. In: New Security Paradigms Workshop (NSPW'09), Oxford, UK, September 8–11 (2009)
163. Wiedenbeck, S., et al.: Authentication using graphical passwords: basic results. Retrieved October 23, 2005. Available at: http://clam.rutgers.edu/~birget/grPssw/susan3.pdf
164. Wiedenbeck, S., et al.: PassPoints: design and longitudinal evaluation of a graphical password system. Int. J. Human-Comput. Stud. **63**(1–2) (2005)
165. Blonder, G.E.: Graphical passwords. United States Patent 5559961 (1996)
166. Varenhorst, C.: Passdoodles; a lightweight authentication method. July 27, 2004. Available at: http://people.csail.mit.edu/emax/papers/varenhorst.pdf
167. Jermyn, I., et al.: The design and analysis of graphical passwords. In: Proceedings of the 8th USENIX Security Symposium, Washington, D.C., August 23–36 (1999)

168. Thorpe, J., v. Oorschot, P.: Towards secure design choices for implementing graphical passwords. In: 20th Annual Computer Security Applications Conference, Tucson, Arizona, December 6–10 (2004)
169. Ross, S.: Is it just my imagination? Retrieved November 4, 2005. Available at: http://research.microsoft.com/displayArticle.aspx?id=417
170. Renaud, K., McBryan, T.: How viable are Stubblefield and Simon's inkblots as password cues? In: PUMP 2010, University of Abertay, Dundee, 6 September (2010)
171. Renaud, K., McBryan, T., Siebert, P.: Password cueing with cue(ink)blots. In: IADIS Computer Graphics and Visualization 2008 (CGV 2008), Amsterdam, The Netherlands (2008)
172. Stubblefield, A., Simon, D.: Inkblot authentication. Microsoft TechReport# MSR-TR-2004-85 (August 2004). Available at: http://research.microsoft.com/pubs/70086/tr-2004-85.pdf
173. Porter, S.: Stronger passwords through visual authentication: handwing. University of Glasgow. Retrieved November 4, 2005. Available at: http://www.dcs.gla.ac.uk/~porters/thesis.pdf
174. Standring, S.: Gray's Anatomy: The Anatomical Basis of Medicine and Surgery. Churchill Livingstone, Oxford (2004)
175. Erdogan, H., et al.: Multi-modal person recognition for vehicular applications. Lect. Notes Comput. Sci. **3541**, 366–375 (2005)
176. Marin, J., Ragsdale, D., Surdu, J.: A hybrid approach to the profile creation and intrusion detection. In: DARPA Information Survivability Conference and Exposition (DISCEX II'01) (2001)
177. Bergadano, F., Gunetti, D., Picardi, C.: User authentication through keystroke dynamics. ACM Trans. Inf. Syst. Secur. **5**(4), 367–397 (2002)
178. Frantzeskou, G., Gritzalis, S., MacDonell, S.: Source code authorship analysis for supporting the cybercrime investigation process. In: 1st International Conference on eBusiness and Telecommunication Networks—Security and Reliability in Information Systems and Networks Track, Setubal, Portugal. Kluwer Academic, Dordrecht (2004)
179. Colombi, J., et al.: Cohort selection and word grammer effects for speaker recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA (1996)
180. Tsai, W.-H., Wang, H.-M.: Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. In: IEEE Transactions on Audio, Speech and Language Processing, January 2006
181. Crompton, M.: Biometrics and privacy: the end of the world as we know it or the white knight of privacy? In: 1st Biometrics Institute Conference (2003)
182. Prassas, G., Pramataris, K.C., Papaemmanouil, O.: Dynamic recommendations in internet retailing. In: 9th European Conference on Information Systems (ECIS 2001) (2001)
183. Liang, T.P., Lai, H.-J.: Discovering user interests from web browsing behavior. In: Proceedings of the Hawaii International Conference on Systems Sciences, Hawaii, USA (2002)
184. Fu, Y., Shih, M.: A framework for personal web usage mining. In: International Conference on Internet Computing (IC'2002), Las Vegas, NV (2002)
185. Goecks, J., Shavlik, J.: Learning users' interests by unobtrusively observing their normal behavior. In: Proceedings of the International Conference on Intelligent User Interfaces, New Orleans, LA (2000)
186. Democraticmedia.org: TV that watches you: the prying eyes of interactive television. A report by the center for digital democracy, June 2001. Available from: www.democraticmedia.org/privacyreport.pdf
187. Jain, K., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. In: Pattern Recognition (2005)
188. Dahel, S.K., Xiao, Q.: Accuracy performance analysis of multimodal biometrics. In: IEEE Information Assurance Workshop on Systems, Man and Cybernetics Society (2003)
189. Humm, A., Hennebert, J., Ingold, R.: Scenario and survey of combined handwriting and speech modalities for user authentication. In: 6th International Conference on Recent Advances in Soft Computing (RASC'06), Canterbury, UK (2006)

Chapter 14
Human Behavior Analysis in Ambient Gaming and Playful Interaction

Ben A.M. Schouten, Rob Tieben, Antoine van de Ven, and David W. Schouten

14.1 Introduction

Game developers are not primarily driven by technology. The main driver for game developments is the *gameplay* itself. Gameplay refers to the overall game experience or the essence of the game itself. There is some confusion as to the difference between game *mechanics* and gameplay. Game mechanics is a construct of rules (not necessarily computable rules), introduced to produce an enjoyable game. For some, gameplay is nothing more than the set of game mechanics. For others, gameplay determines the overall characteristics of the game itself, which is partly in the perception of the game player.

Before we begin our survey, it is important to also underline the importance of game design, an issue which is beyond the scope of this chapter. To illustrate, take a simple puzzle game, where one does not need advanced input possibilities, or realistic feedback. In a realistic and natural golf simulator, however, one cannot truly experience a swing without advanced multi-modal input to measure the result, or the feedback of force and wind during the swing.

B.A.M. Schouten (✉) · R. Tieben
Department of Industrial Design, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: b.a.m.schouten@tue.nl

R. Tieben
e-mail: r.tieben@tue.nl

A. van de Ven
Fontys University of Applied Sciences, Postbus 347, 5600 AH Eindhoven, The Netherlands
e-mail: antoine.vandeven@fontys.nl

D.W. Schouten
S. Nicolas Highschool, Prinses Irenestraat 21, 1077 WT Amsterdam, The Netherlands

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_14,

Fig. 14.1 From Pac-Man (1980) to Call-of-Duty (Black ops, 2010); from limited gameplay and 2D visualization to realistic gameplay and output

If we compare the historic game *Pac-Man* [44] with a modern game like *Call-of-Duty* [12] (for PC, see Fig. 14.1), we can see obvious changes in visuals, gameplay, level design, and so on. However, the interaction (input and output) is basically still delivered in the same way, through a (physical) controller and a (video)screen.

Call of Duty was introduced on the PC, and later expanded to other consoles in order to enhance the game experience and allow for a better and more natural interaction to game action. These consoles allow advanced input (and limited output) by controllers like gamepads, joysticks, steering wheels, trackballs, motion sensing etc. Sometimes these controllers are equipped with LED lights or haptic or auditory feedback or a rumble pak (to enable force feedback).

In the last decades, game developers have focused on creating more natural and realistic gameplay, enabled by fast technological progress. This chapter focuses on the technology; the design and development of games as enabled by this technology is a different topic. In Sect. 14.2, we will present a brief history of games in relation to computer analysis of human behavior. Section 14.3 will cover the input modalities, the different ways in which players interact with the gaming systems. In specific, we focus on the role of technology and computer analysis of behavior. In Sect. 14.4, we cover the game experience (sensory output as well as perception) of modern games, and the way in which human behavior analysis and technology influence this experience. In addition, we show a trend toward games that include principles from ambient technology, defined as *ambient gaming*.

We conclude this chapter with challenges and opportunities for human behavior analysis in the near future, in relation to game development (Sect. 14.5).

14.2 History of Games

The predecessor of all console game genres is considered to be the ball-and-paddle game, called *Pong* [50]. In 1973, after the success of the original PONG coin-op, an Atari engineer by the name of Harold Lee came up with the idea of a home PONG unit. Pong could be played on your home television set. Many of the concepts from

arcade video games were ported by Atari to different consoles, creating a mass market. The Atari 2600 [7], released in 1977, is the first successful video game console to use plug-in cartridges instead of having one or more games built in.

Almost all the earliest video games were action games. *Space Invaders* [58] from 1978, *Asteroids* [6] from 1979, and *Pac-Man* [44] from 1980 are some of the earliest video games, and have since become iconic examples from the action genre.

Donkey Kong [16], an arcade game created by Nintendo, released in July 1981, was the first game that allowed players to jump over obstacles and across gaps, making it the first true platformer.[1] This game also introduced *Mario* [37], an icon of the genre. Donkey Kong was ported to many consoles and computers at the time, and the title helped to cement Nintendo's position as an important name in the video game industry internationally.

Mario also paved the way to more advanced forms of interaction and ludic activity. Role-playing video games (RPG) draw their gameplay from traditional role-playing games like *Dungeons & Dragons* [18]. Most of these games cast the player in the role of one or more 'adventurers' who specialize in specific skill sets (such as melee combat or casting magic spells) while progressing through a pre-determined storyline. Massively multiplayer online role-playing games, or MMORPGs, emerged in the mid to late 1990s as a commercial, graphical variant of text-based MUDs (multiplayer real-time virtual world described primarily in text) which had existed since 1978. By and large, MMORPGs feature the usual RPG objectives of completing quests and strengthening one's player character, but involve up to hundreds of players interacting with each other on the same persistent world in real-time. The massively multiplayer concept was quickly combined with other genres. Fantasy MMORPGs like *The Lord of the Rings Online: Shadows of Angmar* [35], remain the most popular type, with the most popular 'pay-to-play' game being *World of Warcraft* [63] (by Blizzard) which holds over 60% of the MMORPG market, and the most popular free game, *RuneScape* [52], by JaGex Studios, yet other types of MMORPG are appearing. Other massively multiplayer online games which do not have a conventional RPG setting such as *Second Life* [55] may still sometimes be classed as RPGs.

To support these trends in contemporary gaming, recently we see a shift from advanced computer graphics to better interaction based on sensory input, the integration of different modalities, *tangible computing* and the analysis of human behavior.

Tangible computing [17] is an area of Human–Computer Interaction (HCI) research in which people are exploring how we can move the interface 'off the screen' and into the real world. The objective is to interact with physical objects, which have become augmented with computational abilities. This lets designers offer new sorts of interactions, or take advantage of our physical skills (like being able to use two hands, or to rearrange space to suit our needs), or even to directly observe and respond to our physical activities in the world (perhaps by knowing where we are and

[1] The platform game (or platformer) is a video game genre characterized by requiring the player to jump to and from suspended platforms or over obstacles (jumping puzzles).

who we are with, and responding appropriately). In the next section we will see some examples.

Despite all these (conceptual) trends, it is important to say that human behavior analysis for gaming, as a technology, is still in its early years. Most of the applications limit themselves to simple (biometrical) recognition, enabling the user to shift away from the traditional input devices and allowing to be tracked and traced. More advanced features as emotion or activity recognition are still in the research domain. We will discuss some of these challenges at the end of this chapter.

14.3 The Gamer Put into Action

A game controller is a device used in games or entertainment systems to control a playable character or object, or otherwise interact in a computer game. A controller is typically connected to a game console or computer by means of a wire, cord or nowadays, by means of wireless connection [59].

Controllers vary from keyboards and joysticks to light guns and physical objects. The input to a game console can vary from simply (pushing) a button, to rich multimodal interaction from distributed intelligent environments equipped with sensors. We like to distinguish between several categories of input:

1. Direct Input. Controllers to activate commands and other in-game actions.
2. Audio-visual based input. Cameras and microphones to detect & recognize actions.
3. Input provided by other (physiological) sensors and *wearables*.

In most of the games we play, input is provided to a device (controller) that is connected directly to a game console; the player activates a signal through a controller or other instrument (e.g. mouse & keyboard), and this is metaphorically mapped on a specific input for game action. The most well known solution to this problem (metaphor) is of course the left–right (or a–d) buttons or up–down (or w–s) buttons, which are used for in-game navigation. Adjacent buttons (like q and e) are used for special actions such as to jump or crouch. Moreover, consoles can have joysticks to navigate, d-pads and other (action) buttons for shooting etc.

To enable natural interaction, it is important to create a natural mapping from input device into action. As an example a steering wheel (Fig. 14.2) is better used to replace a button input in a racing game, or a real bike in order to achieve the need for speed to climb a virtual mountain hill.

14.3.1 More Advanced Interaction: Audiovisual Based Input

To allow some freedom in interaction, not limited to display and keyboard, artists and designers in the mid-1990s created interactive play environments, based on the

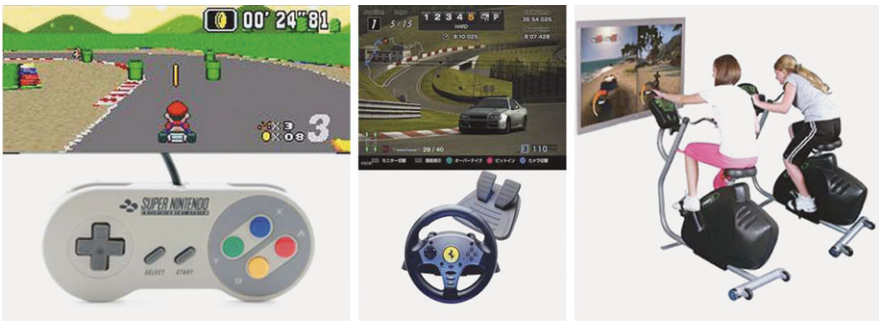


Fig. 14.2 Different Input Devices: SNES controller [56] with d-pad and buttons; GT Steering Wheel [28] with pedals and wheel; Exerbike [20] with cycling on a bike

Fig. 14.3 Three types of visual based input: Daisies (2005) [15], an interactive installation where daisies are projected and die if occluded; EyeToy Play 3 (2005) [21], where player movements result in game character actions; and Xbox Kinect (2010) [39] where full motion recognition is used in a variety of games

projections of video images and interactive sounds. The human–computer interaction was based on simple computer vision algorithms, like in *Daisies*, by Theodore Watson [15], see Fig. 14.3. In this interactive installation, daisies are projected on a floor; cheerful music can be heard. If the projection on the floor is blocked by human appearance, for instance by somebody dancing on the music, daisies will disappear (as if they die) around the body of the user and new daisies will grow, when the projection is restored. A good and simple example of experience design; children loved it and were excited as if they were dancing through a 'real' flowerbed.

In modern game design, due to the progress in scientific research (computer vision) as well as the lowering prices of capturing devices and sensors, direct input can be enriched with audiovisual modalities. These mainly audiovisual signals are captured and analyzed to detect humans and recognize activities and objects. Common technologies vary from relatively simple edge detection and color tracking in Sony's EyeToy [21], to gesture recognition, facial recognition, head tracking, voice and speech recognition in the Xbox Kinect [39], see Fig. 14.3. In more recent consoles, advanced technologies are used such as fingerprint recognition in the Microsoft Surface Tabletop System [38], which allow multi-user tangible interaction.

In the new Kinect [39] for Xbox 360 games, see Fig. 14.3, objects can be scanned and put into virtual action. Microsoft's Kinect (earlier named as project Natal) is based on software technology developed internally by *Rare*, a subsidiary of Mi-

crosoft Game Studios and range camera technology by Israeli developer *PrimeSense*, which interprets 3D scene information from a continuously projected infrared structured light. The Kinect sensor is a horizontal bar connected to a small base with a motorized pivot and is designed to be positioned lengthwise above or below the video display. The device features a 'RGB camera, depth sensor and multi-array microphone running proprietary software' [61] which provide full-body 3D motion capture, facial recognition and voice recognition capabilities. The depth sensor has a fixture that emits structured infrared light and by analyzing the distortions on the sensed patterns, a 3D image of the user and his environment is constructed (*depth map*). According to information supplied to retailers, the Kinect is capable of simultaneously tracking up to six people, including two active players for motion analysis with a feature extraction of 20 joints per player. The sensing range of the depth sensor is limited but adjustable, with the Kinect software capable of automatically calibrating the sensor based on gameplay and the player's physical environment, such as the presence of furniture.

The software technology enables advanced 3D view independent gesture recognition based on a patented algorithm from a company called *Canesta* [26], which is acquired by Microsoft. Three-dimensional position information is used to identify the gesture created by a body part of interest. At one or more instances of an interval, the posture of a body part is recognized, based on the shape of the body part and its position and orientation. The posture of the body part over each of the one or more instances in the interval are recognized as a combined gesture. The gesture is then classified for determining an input into a related electronic device.

Face tracking and facial expression recognition are based on 3D deformable face models and a support vector machine classifier [26]. Voice recognition is supported only in a few countries like the US, UK, Mexico and Japan. The Kinect sensor's microphone array enables the Xbox 360 to conduct acoustic source localization and ambient noise suppression, allowing for things such as headset-free party chat over Xbox Live. The first official games that are supported by the Kinect are *Fable III* [23] and *Ghost Recon: Future Soldier* [25].

More novel are the developments in entertainment robots. *NAO* [27] is an autonomous and interactive humanoid robot developed by *Aldebaran Robotics* that is completely programmable. NAO replaced the robot dog Aibo by *Sony* as the robot used in the *Robocup* ('Robot Soccer World Cup') Standard Platform League (SPL), an international robotics competition. It is currently the most-sold humanoid research and educational robot in the world.

NAO's vision is provided by two CMOS 640×480 cameras, which can capture up to 30 images per second. Algorithms in its on-board computer can detect and track faces and shapes to be able to recognize and follow the person talking to it, to find a ball or more complex objects. NAO's SDK makes it possible to program and apply many different possible behaviors and computer vision algorithms which can run on a remote computer, by interfacing with OpenCV (the Open Source Computer Vision library initially developed by Intel) for computer vision.

NAO uses the Haar Feature-based Cascade Classifier for Object Detection [34], eigenfaces for face recognition [62], and the Continuously Adaptive Mean-Shift

Fig. 14.4 Different types of audio based input: Rockband 3 (2010) [51] measures changes in pitch and length of silences; Nintendogs (2005) [43] can be trained to recognize certain words; and Aibo (1999) [2] responds to voice commands and can learn to recognize it's own name as well as its owner's name

(Camshift) algorithm [9] for face tracking, as well as other methods [9]. Through different software platforms, one is able to implement navigation algorithms like Visual SLAM (Visual Simultaneous Localization and Mapping) [33] and to use speech recognition based on Hidden Markov Models (HMM) [32]. The robot can be programmed to retrieve and express emotions in social games [36]. It has 25 degrees of freedom, including functional hands that can pick up and grasp objects, an inertial sensor, two speakers, four microphones, sonars to detect obstacles and touch-sensors to detect touch. It can express itself by movements, gestures and multicolor LEDs in its eyes and on its body. Other platforms with audio input are exemplified in Fig. 14.4.

14.3.2 Other (Physiological) Sensors and Wearables

Games in this category measure physiological behavior and other characteristics of the human body. Several gaming applications analyze brain activity, heart rate (ECG, EEG, EMG, HEG), respiration (GSR), temperature, iris activity, or glucose blood levels (see Fig. 14.5).

For example, *The Journey to Wild Divine* [31] measures skin conductance level and heart rate variability, translating this to stress and pathologic conditions used in an adventure game. Its controller is a USB-based biofeedback device, which can be used with other biofeedback programs. *Brainball* [10] uses EEG sensors to measure brain activity, and translates this into a competition between two players: the higher the brain activity, the further the ball is pushed away.

Emotiv [19] provides a head set with a series of sensors and integrated algorithms, resulting in an API with three types of measurements. First of all, there is facial expression recognition, by mapping muscle EEG measurements to a human face model. Second, emotional state is detected by recognizing active EEG brain activity clusters. Last but not least, EEG is used to train and recognize thought patterns, which can be mapped to game actions.

In addition to physiological input, wearable sensors are often equipped to the player: either attached to the user's body or clothing, or carried in a device such as a

Fig. 14.5 Different physiological input devices: the Wild Divine (2001) [31] USB biofeedback hardware; the Brainball (2000) [10] EEG installation; and the Emotiv (2010) [19] wireless headset with advanced EEG measurements

mobile telephone. Sensors commonly used for this sort of measurements are inertia sensors (accelerometers, gyroscopes, magnetometers), location sensors (GPS, proximity sensors), mini-cameras and muscle tension detectors. The Wii Remote [42] allows the user to interact with and manipulate items on screen via gesture recognition and pointing through the use of accelerometer and optical sensor technology. The movements of the controller result in similar movements in the game; e.g. swinging the controller results in a swing of a golf club.

The Pokéwalker [49] is a device that connects to the Pokémon [48] games, a pedometer (stepping counter) that measures the player's physical activity. For every step, the Pokémon in the game gains experience points and the player earns 'watts', which can be exchanged for in-game items.

The widespread availability of accelerometer sensors and gyroscopes in mobile phones has also introduced new categories of gaming. The iPhone and iPad for instance are equipped with proximity, motion and acceleration sensors, as well as ambient light sensors which automatically adjust the brightness of the screen in order to conserve battery life [4]. The iPhone 4 adds another sensor: a three-axis gyroscope. When combining the gyroscope with the accelerometer, this gives the iPhone 4 six axes on which it can operate. This is designed to make the iPhone 4 more sensitive and responsive [5]. *Brothers In Arms 2: Global Front* [11], is a first person shooter game which is situated in a Second World War setting and allows gyroscopic 3D control. One of the most eye catching games is *IThrown* [30]. It uses the iPhone's built in accelerometer to measure your virtual throw and how far the phone would have flown if you actually have let it go (see Fig. 14.6).

At the end of this section, based on the analysis described above, we would like to provide the reader with an overview of games used in this chapter including the type of sensors used as input, as well as the enabled game-actions and the corresponding measured human behavior in Fig. 14.7.

14.4 Game Experience and Human Behavior Analysis

In the previous sections we mainly focused on how a game can be put into action through the input of the user. In this section we want to elaborate on the game expe-

Fig. 14.6 Different types of wearables: Wii Remote (2006) [42], which measures movements using acceleration sensors; Pokéwalker (2010) [49], which measures physical steps using a pedometer; and IThrown (2008) [30], which uses the iPhone's accelerometer to measure your virtual throwing distance

rience, which relies in modern games mainly on high definition graphics and other audio-visual output. Sound (-tracks) are often used to enhance the gameplay, sometimes supported by force-feedback (vibrating controller to imitate tactile feedback when e.g. shooting a gun).

However, in recent years, new forms of gaming (experience) and playful interaction have emerged. The interaction can be in the real world (e.g. through play-objects) but also in hybrid environments. With the availability of cheap sensors and system architectures, like *Arduino*, which allows for human interaction with physical objects, games tend to shift away from the computer. The movie *Minority Report* [41] is an example for the vision of how human–computer interaction will eventually be more natural (using less devices, interaction through natural objects and actions).

As an example, we like to mention the *ColourFlare* [8], which is an object that can be carried in one hand, which changes color when rolled, and which starts blinking when shaken (see Fig. 14.8). When it blinks, it can send its color to other objects in the neighborhood using infra-red technology. The ColourFlare allows children to use their creativity to make their own games (*open-ended play*) in which they allocate meaning to the behavior of the object when shaken and rolled. Children will have to discuss ideas for game goals and rules, and thus also practice their social interaction and negotiating skills.

Mark Eyles [22] mentions the class of games labeled *pervasive/ambient games* allowing the player to act freely in everyday locations while playing. In addition and according to the properties of Ambient Intelligence [1], some new qualities for an enriched game experience can be derived:

1. *context-aware*: (game) devices can recognize you and your situational context
2. *personalized*: the functionality is tailored to your needs and preferences (short timescale, e.g. installing personal settings)
3. *adaptive*: the system can change/adapt in response to you and your environment (adjustments resulting from longer monitoring)

			In chapter examples		Remarks	Other Examples
	Sensor/data	**Example Analysis**	**Example Game Action**	**Example game**		
Direct Input	Button presses		Platforming	**Mario**	Both controller input as well as more natural, metaphorical input	Flight simulator Dance Dance Revolution Motivatrix exergames Golf simulators
	Direct metaphor: Steering wheel Cycling on bike		Guided by the metaphor	**Gran Turismo** **Exerbiking**		
Audio Visual Input	Camera Infrared Camera	Edge detection Color tracking Motion detection Facial recognition Emotion analysis	Relative spatial movement	**EyeToy Play 3** **Daisies**	Simple-to-advanced camera tracking	Playstation Move
			Character facial actions	**Kinect Avatar**		
			Full-body character actions	**Kinect Adventures, Kinect Sports**		
	Simple microphone	Voice – Sound detection Speech – Speaker recognition	Command Control Recognin	**RockStar**	Sound/voice comparison to known data	iPhone Ocarina Talking Pets Crowd-noise competitions
	Multi-array microphone	Speaker - Speech recognition	Input	**Kinect menu**	Speech analysis	Google Voice
Other Sensory Input	Accelometer	Gesture recognition	Virtual actions	**Wii Sports** **iThrown**	Inertia measurements	Powerglove DirectLife Nike+iPod VR gloves
	Pedometer	Counting steps	Character walking progress	**Pokewalker**		
	GPS	Location	Virtual location	**Geocaching**	Location measurements	Foursquare Latitude Urban games
	EEG	Brain activity Thought pattern recognition	Brain activity based competition and actions	**Brainball** **Emotiv**	Physiological measurements	Mindball Mindflex Glucoboy
	Skin conductance		Relaxation goals	**Journey to Wild Divine**		

Fig. 14.7 Overview of games, extracted from this chapter to illustrate sensory input, in game actions as well as the corresponding human behavior

Fig. 14.8 ColourFlares [8], interactive objects that elicit open-ended playful interaction by changing colors when they are rolled and shaken

Fig. 14.9 High-definition output: AmBX (2005) [3] enriches games with visual effects and tactile feedback; Pandadroom (2002) [45] allows visitors to experience a 4D experience with 3D glasses, vibrating chairs, and water spraying

As an example, in the *AmBX* [3] system from Philips (see Fig. 14.9), visual effects and tactile feedback (vibration and wind effects) are added to the gaming experience, by responding to certain game events. AmBX code acts as a conversion middleware (sitting between source and output device) that takes generic or specifically scripted (via AmBX SDK) input signals from video, audio, PC or media content, then outputs it to suitable hardware such as LED lights, rumble boxes or similar devices via cable or wireless, subject to hardware. In the theme-park 4D theatre *Pandadroom* [45], 3D effects, force-feedback chairs, and water spraying make the experience multi-modal and more intense.

One example in which the console is partly context-aware and adaptive, is the *CAVE Automatic Virtual Environment* [13] (see Fig. 14.10). In this application, the environment is projected on all walls and the ceiling of a room, creating a three-dimensional effect—for example from the Unreal Tournament world. Using the headset, the position and the direction of the user's head are detected, and the output is adapted to the user's perspective. The output is thus, among others, dependent on the height of the user.

A more recent example of context-awareness and personalization is the *Kinect Avatar* [40]: the Kinect system recognizes a player, loads a profile with settings, and creates a matching avatar. Technologies such as face recognition, expression analysis, speech recognition and other motion recognition translate the player's movement into a personal Avatar (see Fig. 14.10). In addition, the Kinect uses a com-

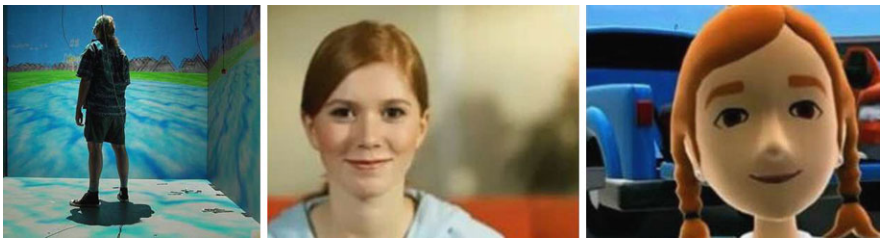


Fig. 14.10 CAVE (1992) [13] adapts its output to the perspective of the player; Kinect Avatar (2011) [40] recognizes the player, and personalizes the gaming experience

Fig. 14.11 Pervasive and locative games: Geocaching (2000) [24], finding hidden caches throughout the world using GPS; Parallel Kingdoms (2010) [46], conquering areas depending on your physical location; and Head Up Games (2010) [57], playing games depending on your proximity to other players and game objects

bination of context-awareness and adaptation to setup the sound output: a special learning algorithm adapts the sound output to the physical characteristics of the room, including the position of the players and objects.

Pervasive and locative games are another example of games that use aspects from ambient intelligence. These games blend the virtual and real world and are interacted through multiple ubiquitous devices. A *location-based game* (or location-enabled game) is one in which the gameplay evolves and progresses through a player's location. Thus, location-based games almost always support some kind of localization technology, for example by using satellite positioning (GPS). Current research trends use other embedded mobile protocols like Near Field Communication and Ultra Wide Band Wireless (UWB).

Urban gaming or *Street Games* are typically multi-player, location-based games. The playground is the city itself. An example of such a pervasive game is *Geocaching* [24], treasure hunting with the help of GPS, a popular activity in which players search hidden caches around the world (see Fig. 14.11). The caches and puzzles have been created by other players. In *Parallel Kingdom* [46], players use their location-aware telephone to conquer different areas of the map. The playground is the current real-world location of the players, moreover its location is constantly changing by players that travel around in the real world. In a recent publication, Soute and Markopoulos used the notion of *Head Up Games* [57], because children can play these games without having to focus on a screen or other device, using

wearable sensors and actuators. The technology is used to support the playful interaction. Gameplay is more open-ended, rules originate from the players themselves.

14.5 Summary

In this chapter we showed some new developments in game design and technology. Inspired by ambient intelligence [53, 60], ambient gaming will become context-aware, adaptive, personalized and anticipatory. Games will be developed that allow us to interact freely, not depending on a central computer but supported by sensors embedded in play objects and toys. Also gaming will be more playful, open ended such that rules can easily be altered and be supportive to other activities. Hybrid graphical environments and other actuators will enrich the game experience.

In serious game design, another aspect can be added to the notion of ambient gaming. Schouten [54] envisions gaming in a context in which they are a part of everyday activities; a playful approach in which games are not always 'present', but can be called upon when necessary as part of existing applications in learning, social networks, health care etc. Besides real-time analysis of behaviors, this requires a social intelligence in game design and will lead to games that are embedded in systems of social meaning; fluid and negotiated between us and the other people around us (an early example is *Cityville* [14] in Facebook). In this way game design focuses on interactive products as creators, facilitators and mediators of experiences. Experience comprises of perception, action, motivation and cognition [29].

In general we can say that computer enabled human behavior analysis can play an important role in two main areas.

1. Physiological behavior, activities and human events. Human behavior analysis can play an important role in gaming experiences on a physiological level. For instance, games can be used for rehabilitation of injured medical patients or the disabled. But also in learning or training activities for sport and similar other activities, feedback through games could improve results. For the elderly, an activity program based on their personal capabilities could improve the quality of life.
2. Psychological and social behavior. If computers can measure the emotions and expressions of the player(s), then games can adapt to playing styles and maximize the gaming experience. Imagine a gaming character interacting in a specific way to a calm couch-hanging player, or to a group of excited friends. On a personal level, if the gaming experience can adapt itself to the emotional state of the player, e.g. to the arousal level, then the immersion and in-the-flow level can be optimized. Furthermore, one can imagine focus recognition, as is suggested by Peters and Itti [47], to respond to the point of the player's attention; for instance creating an enemy at the spot where the user is not paying attention to. If the player always acts in a certain way, the game can predict or alter this.

In short game design and playful interaction are among the first application areas which will benefit from new developments in computer analysis of human behavior.

Other application areas will follow like health care or education, where results are more critical. Such technologies will play an important role to make games even more exciting.

14.6 Glossary

- *Adaptive systems*: Systems that can change/adapt in response to the user and his/her environment (adjustments resulting from longer monitoring)
- *Ambient gaming*: Games that are context-aware, adaptive, personalized and anticipatory
- *Ambient intelligence*: Vision on technological development, in which systems are distributed, context-aware, adaptive, personalized and anticipatory
- *Arduino*: Easy-to-use open-source microcontroller system
- *Context-aware*: Characteristic of a system that can recognize you and your situational context
- *Depth map*: An image of the user and his/her environment that contains information related to the distance of the surfaces from a viewpoint
- *Distributed intelligent environment*: Environment with embedded sensors and actuators, that responds to user actions in an adaptive, anticipatory and personalised way
- *Game controller*: A device used in games or entertainment systems to control a playable character or object, or otherwise interact in a computer game
- *Game mechanics*: A construct of rules (not necessarily computable rules), introduced to produce an enjoyable game
- *Gameplay*: The overall game experience or essence of the game itself
- *Humanoid robot*: Robot with human-like attributes or characteristics
- *Locative games*: Games that take place between physical locations, or that can be played in any location
- *Multi-player location-based game*: Game in which groups of users play through physically moving to different locations
- *Open-ended play*: Play that is not fully determined by rules, but allows the adaption of existing, and creation of new rules
- *Personalized systems*: Systems that tailor the functionality to your needs and preferences (short timescale, e.g. installing personal settings)
- *Pervasive games*: Games that are integrated in the physical space
- *Platformer*: A video game genre characterized by requiring the player to jump to and from suspended platforms or over obstacles
- *Role-playing video games*: Video games in which the player plays the role of a certain character, and develops this character by making game decisions
- *Social games*: Games that are based on social interaction, and utilize social characteristics, often found on profile sites such as Facebook
- *Tactile*: Designed to be perceived by touch; feedback that one can feel
- *Tangible computing*: An area of Human–Computer Interaction research in which people are exploring how we can move the interface ‘off the screen’ and into the

real world. The objective is to interact with physical objects which have become augmented with computational abilities
- *Urban gaming*: Gaming that takes place in an urban environment, e.g. throughout a city

References

1. Aarts, E.H.L., Marzano, S.: The New Everyday: Views on Ambient Intelligence. 010 Publishers, Rotterdam (2003). 9064505020
2. Aibo (1999). http://en.wikipedia.org/wiki/Aibo. Retrieved on 2011-02-12
3. amBX Technology. http://ambx.com. Retrieved on 2011-01-12
4. Apple Battery Information (2010). http://www.apple.com/batteries/iphone.html. Retrieved on 2012-02-12
5. Apple iPhone 4 (2011). http://en.wikipedia.org/wiki/IPhone_4. Retrieved on 2012-02-12
6. Asteroids (1979). http://en.wikipedia.org/wiki/Asteroids_(video_game). Retrieved on 2011-02-05
7. Atari 2600 (1977). http://en.wikipedia.org/wiki/Atari_2600. Retrieved on 2011-02-07
8. Bekker, T., Hummels, C., Nemeth, S., Mendels, P.: Redefining toys, games and entertainment products by teaching about playful interactions. Int. J. Arts Technol. **3**(1), 17–35 (2010)
9. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, Sebastopol (2008)
10. Brainball (2000). http://www.tii.se/touchingtheinvisible/brainball.html. Retrieved on 2010-12-05
11. Brother in Arms 2: Global Front (2010). http://uk.wireless.ign.com/articles/107/1070914p1.html. Retrieved on 2012-02-12
12. Call of Duty 4: Black Ops (2010). http://en.wikipedia.org/wiki/Call_of-Duty. Retrieved on 2010-11-28
13. CAVE Automatic Virtual Environment (1992). http://en.wikipedia.org/wiki/Cave_Automatic_Virtual_Environment. Retrieved on 2010-12-10
14. CityVille (2010). http://en.wikipedia.org/wiki/CityVille. Retrieved on 2011-02-12
15. Daisies (2007). http://www.theowatson.com/site_docs/work.php?id=18. Retrieved on 2011-01-06
16. Donkey Kong (1981). http://en.wikipedia.org/wiki/Donkey_kong. Retrieved on 2011-02-05
17. Dourish, P.: Where the Action Is: the Foundations of Embodied Interaction. MIT Press, Cambridge (2004)
18. Dungeons and Dragons (1974). http://en.wikipedia.org/wiki/Dungeons_and_dragons. Retrieved on 2011-02-05
19. Emotiv Systems (2010). http://en.wikipedia.org/wiki/Emotiv_Systems. Retrieved on 2011-01-05
20. Exerbike Xg (2009). http://www.exerbikeusa.com. Retrieved on 2010-11-20
21. EyeToy Play 3 (2005). http://en.wikipedia.org/wiki/EyeToy. Retrieved on 2010-11-20
22. Eyles, M., Eglin, R.: Ambient games, revealing a route to a world where work is play? Int. J. Comput. Games Technol. **2008**, 1–7 (2008)
23. Fable 3 (2010). http://en.wikipedia.org/wiki/Fable_III. Retrieved on 2011-01-05
24. GeoCaching (2000). http://www.geocaching.com. Retrieved on 2010-12-10
25. Ghost Recon: Future Soldier (2011). http://en.wikipedia.org/wiki/Ghost_Recon:_Future_Soldier. Retrieved on 2011-02-12
26. Gokturk, S.B. et al.: Gesture recognition system using depth perceptive sensors (2008). http://www.google.nl/patents/about?id=8JKpAAAAEBAJ. Retrieved on 2011-02-05
27. Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., Maisonnier, B.: The NAO humanoid: a combination of performance and affordability. arXiv:0807.3223 (2008)

28. GT Steering Wheel (2004). http://www.logitech.com/en-us/gaming/wheels/devices/4172. Retrieved on 2010-11-20
29. Hassenzahl, M.: Encyclopedia entry on user experience and experience design (2011). http://www.interaction-design.org/encyclopedia/user_experience_and_experience_design.htm. Retrieved on 2011-03-14
30. IThrown (2011). http://www.freshapps.com/ithrown/. Retrieved on 2011-02-12
31. Journey to Wild Divine (2001). http://en.wikipedia.org/wiki/Journey_to_Wild_Divine. Retrieved on 2010-12-05
32. Juang, B.H., Rabiner, L.R.: Hidden Markov models for speech recognition. Technometrics **33**(3), 251–272 (1991)
33. Karlsson, N., Di Bernardo, E., Ostrowski, J., Goncalves, L., Pirjanian, P., Munich, M.E.: The vSLAM algorithm for robust localization and mapping. In: Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on. IEEE Press, New York (2005)
34. Lienhart, R., Maydt, J.: An extended set of Haar-like features for rapid object detection. In: Proceeding of the International Conference on Image Processing, 2002, vol. 1, p. 900. IEEE Press, New York (2002).
35. Lord of the Rings Online: Shadows of Angmar (2007). http://en.wikipedia.org/wiki/The_Lord_of_the_Rings_Online:_Shadows_of_Angmar. Retrieved on 2011-02-05
36. Lourens, T., Barakova, E.: Humanoid robots are retrieving emotion from motion analysis. (2011, submitted)
37. Mario (1981). http://en.wikipedia.org/wiki/Mario. Retrieved on 2011-02-05
38. Microsoft Surface (2011). http://blogs.msdn.com/b/surface/. Retrieved on 2011-01-06
39. Microsoft Xbox Kinect (2010). http://en.wikipedia.org/wiki/Kinect. Retrieved on 2011-01-02
40. Microsoft Xbox Kinect Avatar (2011)
41. Minority Report (2002). http://en.wikipedia.org/wiki/Minority_Report_(film). Retrieved on 2011-02-12
42. Nintendo Wii Remote (2006). http://en.wikipedia.org/wiki/Wii_Remote. Retrieved on 2010-11-05
43. Nintendogs (2005). http://en.wikipedia.org/wiki/Nintendogs. Retrieved on 2010-12-05
44. Pac-Man (1980). http://en.wikipedia.org/wiki/Pac-Man. Retrieved on 2010-11-20
45. Pandadroom (2002). http://www.efteling.com/NL/Park/Attracties/PandaDroom.html. Retrieved on 2010-12-10
46. Parallel Kingdom (2010). http://www.parallelkingdom.com. Retrieved on 2010-12-10
47. Peters, R.J., Itti, L.: Applying computational tools to predict gaze direction in interactive visual environments. ACM Trans. Appl. Percept. **5**(2), 9 (2008)
48. Pokemon (1999). http://en.wikipedia.org/wiki/Pokemon_(video_game_series). Retrieved on 2010-12-05
49. Pokewalker (2010). http://en.wikipedia.org/wiki/Nintendo_DS_accessories. Retrieved on 2010-12-05
50. Pong (1973)
51. Rockband 3 (2010). http://en.wikipedia.org/wiki/Rock_Band_3. Retrieved on 2010-12-10
52. RuneScape (2001). http://en.wikipedia.org/wiki/Runescape. Retrieved on 2011-02-12
53. Salah, A.A., Morros, R., Luque, J., Segura, C., Hernando, J., Ambekar, O., Schouten, B., Pauwels, E.: Multimodal identification and localization of users in a smart environment. J. Multimodal User Interfaces **2**(2), 75–91 (2008)
54. Schouten, B.: Play as source for ambient culture. Inaugural Speech Professor of Serious Gaming, Fontys University of Applied Science (2008). http://www.fontys.nl/generiek/bronnenbank/sendfile.aspx?id=189061
55. Second Life (2003). http://en.wikipedia.org/wiki/Second_life. Retrieved on 2011-02-12
56. SNES Controller (1992). http://en.wikipedia.org/wiki/Super_Nintendo_Entertainment_System. Retrieved on 2010-11-20
57. Soute, I., Kaptein, M., Markopoulos, P.: Evaluating outdoor play for children: virtual vs. tangible game objects in pervasive games. In: Proceedings of the 8th International Conference on Interaction Design and Children, pp. 250–253. ACM, New York (2009).

58. Space Invaders (1978). http://en.wikipedia.org/wiki/Space_invaders. Retrieved on 2011-02-12
59. Taxonomy of Game Controllers. http://en.wikipedia.org/wiki/Game_controller. Retrieved on 2011-01-06
60. Tistarelli, M., Schouten, B.: Biometrics in ambient intelligence. J. Ambient Intell. Humaniz. Comput. **2**(2), 113–126 (2010)
61. Totilo, S.: Natal recognizes 31 body parts, uses tenth of Xbox 360 computing resources. http://kotaku.com/#!5442775/natal-recognizes-31-body-parts-uses-tenth-of-xbox-360-computing-resources. Retrieved on 2011-02-12
62. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. **3**(1), 71–86 (1991)
63. World of Warcraft (2004). http://en.wikipedia.org/wiki/World_of_warcraft. Retrieved on 2011-02-12

Index

A
Action, 155–157, 159, 168, 174, 178, 326
Activities of daily living, ADL, 326
Activity, 154–157, 159, 160, 169, 170, 173, 174, 178, 326
Affinity matrix, 48
Ambient intelligence, 395
Applications
  affective computing, 269
  Arduino, 395
  iBracelet, 329
  PlayBot, 97
  sensitive artificial listener, SAL, 270
  SignTutor, 142
  SnowGlobe, 338
  STARS, 142, 144, 145
Attention, 69
  attention systems, 97
    AIM, 98
    evaluation, 94
    iNVT, 77, 98
    SaliencyToolbox, 98
    VOCUS, 77, 98
  bottom-up, 74, 78, 87, 94
  computational visual, 69
  covert, 74
  focus of attention, FOA, 77, 87, 303
  guided search model, 76
  human visual attention, 71
  overt, 74
  selective, 69
  top-down, 74, 88, 94
  visual, 69, 302
  visual dominance ratio, VDR, 303
Attentional capture, 74

B
Background subtraction, 56, 57, 65
Bag of words, BoW, 153, 161, 162, 164, 167–169, 241, 308
Baum–Welch algorithm, 143
Bayes' rule, 3, 4, 111, 136
Behavioral biometrics, 347
Belief propagation, 55

C
Camera calibration, 131
Center-surround filters, 78, 80, 100
Chunking, 230
Circumplex model of affect, 257
Classification, 228, 238
Classifiers
  artificial neural networks, ANN, 239
  bi-directional long short-term memory neural networks, BLSTM-NN, 274
  boosting, 54, 309
  combination of, 240
  decision trees, 239
  elastic graph matching, 134
  hidden Markov models, HMM, 27
  k-nearest neighbor, kNN, 239
  random forests, RF, 239
  rank level fusion, 307
  score level fusion, 306
  support vector machines, SVM, 54, 171, 243, 268, 297, 309, 316, 392
Clustering, 9, 10, 28
Conjunction search, 74
Context, 132, 152, 153, 158, 172–175, 178, 296
Context-awareness, 395

D
Depth map, 392
Difference of Gaussians, DoG, 53, 71, 72, 78, 100

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9, © Springer-Verlag London Limited 2011

Direct linear transform, DLT, 108, 120
Dirichlet process, DP, 9, 19
Distance measures
  Bhattacharya distance, 63
  Euclidean distance, 82
  Mahalanobis distance, 56, 169
Distance transform, 44
Distributed intelligent environment, 390

E
Emotions, 256
  models of, 257
Entity, 156–159, 161, 175, 178
Expectation-Maximization algorithm, EM, 64, 309, 310
Eye movements
  evaluation with, 95
  eye movement data online, 95, 98

F
Face detection, 299
Fall detection, 332
Fast Fourier transform, FFT, 233
Feature extraction, 131, 228, 299
  Gabor wavelets, 57, 83, 84, 134
  histograms of oriented gradients, HOG, 50, 54, 153, 161, 164, 165, 168
  Hu moments, 134, 146
  independent components analysis, ICA, 238
  linear discriminant analysis, LDA, 238
  mel-frequency cepstral coefficients, MFCC, 28, 233, 273
  most discriminative features, MDF, 134
  motion energy image, 134
  motion history image, 134, 302
  non-negative matrix factorization, NMF, 238
  principal components analysis, PCA, 57, 64, 134, 238
  relational features, 299
  scale invariant feature transform, SIFT, 50, 53, 161–164
  space time interest points, STIP, 52
  Zernike moments, 134
Feature integration theory, FIT, 75
Feature search, 74
Feature selection, 229, 235
  acoustic features, 235
Functionals, 234
Fuzzy metric temporal Horn logic, FMTL, 155, 176, 178

G
Gameplay, 387
Games
  ambient gaming, 399
  game controller, 390
  game mechanics, 387
  locative games, 398
  multi-player location-based games, 398
  online games
    massively multi user dungeon, MUD, 389
    massively multiplayer online role-playing games, MMORPG, 389
    role-playing video games, RPG, 389
  pervasive games, 398
  platformer, 389
  social games, 393
  urban gaming, 398
Gaussian curvature, 48
Gaussian process, GP, 11, 13, 19
Gaussian pyramid, 52
Gestures
  act gestures, 127
  beats, 126
  deictic gestures, 126, 127, 129
  gesture spotting, 132, 135, 145, 148
  iconic gestures, 126
  metaphoric gestures, 126
  symbol gestures, 126, 127
Gibbs sampling, 308
Global interaction model, 170, 171
Graphical models, 22, 136, 140
  belief propagation, 34
  conditional random field, CRF, 22, 29, 30, 137, 141, 268, 336
  coupled hidden Markov model, CHMM, 311
  Dirichlet process mixture model, DPMM, 11
  factorization of, 23, 26, 30, 36
  hidden conditional random field, HCRF, 32, 33, 138, 140
  hidden Markov model, HMM, 11, 19, 22, 27, 136, 143, 240, 297, 310, 316, 336, 393
  influence model, 310, 311, 316
  input output hidden Markov model, IOHMM, 138–140, 142, 146, 303
  latent Dirichlet allocation, LDA, 307, 308, 318
  latent dynamic conditional random field, LDCRF, 32, 33, 138–140
  linear chain conditional random field, LCCRF, 31
  Markov random field, MRF, 136

H
Harris corner detector, 51, 52, 100
Hessian matrix, 53
Histogram, 50
Histogram intersection kernel, 171
Honest signals, 186
Human event understanding, HEU, 152, 158–160, 172, 176–178
Human visual system, 71
Human–robot interaction, 70
Human–computer interaction, HCI, 125, 152, 350, 389
  gesture based, 129
Hungarian search, 60

I
Inhibition of return, IOR, 88
Integral image, 80, 81
iNVT, 77
Isomap algorithm, 115, 335

K
K-means clustering, 162
Kernel density estimation, KDE, 55, 56
Kernel functions, 14, 15, 239, 240
Kernel trick, 14
Kullback–Leibler divergence, 8

L
Level sets, 44, 59, 62
Local interaction model, 169, 171
Low-level descriptors, 231

M
Markov chain Monte Carlo, MCMC, 7
Maximum a posteriori, MAP, 5, 241, 297
Maximum likelihood estimation, MLE, 5, 55, 241, 268, 297
  overfitting problem, 6
Mean curvature, 48
Medial axis transform, 45, 57
Mixture of Gaussians, MoG, 56, 146
Motion capture, 106, 260, 261
  cardboard models, 110
  markerless, 109, 116
  model based, 110
Movement, 155–157, 178
Multimodal fusion, 265, 268, 274, 276, 277, 313
Multiscale representations
  Gaussian pyramid, 78, 80, 83, 84
  Laplacian pyramid, 83, 84
  Oriented pyramid, 83
  Pyramid of histograms of oriented gradients, PHOG, 168, 169

O
Open-ended play, 395
Optical flow, 47, 58, 61, 83, 110, 153

P
PAD emotion space, 257
Pearson's correlation coefficient, 272, 314
Personality, 295
Petri nets, 155
Pitch, 233
Playful interaction, 399
Primary visual cortex, V1, 72
Privacy, 296
Probability distributions
  Dirichlet, 9, 10, 308
  multinomial, 10, 308
  multivariate Gaussian, 13

R
Recognition
  actions, 46, 52, 153, 172, 331
  activities, 18, 116, 152, 154, 160, 175, 178, 299, 302, 327, 329, 331, 335, 337
  emotions, 247, 258, 393
  events, 154, 157
  facial expressions, 393
  gait, 105, 334
  gestures, 22
  group interactions, 296
  hand gestures, 129, 135, 140
  physiological signals, 261
  postures, 51
  sign language, 142
  social roles, 32, 317
  speech, 242
  vocal behavior, 246, 316
Recursive coarse-to-fine localization, RCFL, 165
Regression, 238
  Linear, 13
Robots
  attentive robots, 96
  humanoid robots, 392
  localization, 96
  PlayBot, 97
  visual SLAM, 97

S
Saliency, 74, 78, 87, 100, 242, 331
  conspicuity map, 85, 86
  in speech, 247

Saliency (*cont.*)
interest points, 51, 65
region of interest, ROI, 89, 331
saliency map, 73, 77, 87
Seeded region growing, 87
Sensors, 304
body worn, 328
cameras, 261, 331
3D, 109
stereo, 332
time-of-flight, TOF, 331, 341
electromyography, 119, 261
force platforms, 119
inertial, 116
accelerometers, 106, 116, 329, 332
gyroscopes, 117, 333
magnetometers, 118
pressure sensors, 117
microphone arrays, 314
microphones, 261, 330
mobile devices, 304
motion, 337
RFID, 328, 339
sensor networks, 329
Shape descriptors
contour, 62
curvature, 48, 58
graph, 48
shape histogram, 51
template, 49
Shape representations, 42
spatial representations, 42
contour, 42–44
contours, 112
silhouette, 43, 45, 50, 56–58, 65, 113
skeleton, 42, 45, 49, 57, 115
spatiotemporal representations, 45
bag of features, 46
space-time cube, 45
trajectory representation, 46
volumetric representation, 46
Situation Graph Tree, SGT, 152, 155, 159, 176–178
Social affiliation networks, SAN, 300
Social attitudes, 202
persuasion, 202, 203
Social emotions, 211
blush, 212
enthusiasm, 215
pride, 212
shame, 211
Social relationships, 207
dominance, 208, 209, 294, 297, 301, 305, 309–311, 315
leadership, 295, 302, 303, 314
power, 294
social roles, 297, 315
social verticality, 294, 310
status, 295
Social signal processing, 21
Social signals, 186
backchannel signals, 201, 247, 317
communicative signals, 190
conflict detection, 193
definition of, 189
extraversion, 186
informative signals, 189
modalities
audio signals, 260
cued speech, 128
facial expressions, 260, 304
fingerspelling, 128
gaze, 196, 197, 295, 302
gestures, 193, 303
head movements, 195, 304
language, 192
posture, 199
prosody, 193, 231, 273, 301
speaker turn, 300
speech, 296
touch, 199
turn-taking behavior, 193
sign language, 128
social signal processing, 187
Speech
speaker diarization, 28, 298
speaker states, 228
speaker traits, 228
Squashing function, 18, 275
Stereophotogrammetry, 108
Summed area table, 80

T

Tactile, 395
Tangible computing, 389
Thin slices of behavior, 186, 293, 299, 308, 316, 317, 319, 320
Tracking, 47, 59, 110, 111, 175, 331
Kalman filter, KF, 61, 112, 119, 131, 132, 145
kernel tracking, 63–65
particle filter, PF, 8, 61, 114, 131, 274
tracking by association, 60, 65

V

Valence-arousal, 244
Visual attention, 69
Visual search, 74, 77, 88, 100
conjunction search, 74

Visual search (*cont.*)
feature search, 74
search asymmetries, 75
Visual SLAM, 96, 97, 393
Viterbi algorithm, 55, 141
VOCUS, 77
Voice quality, 233

W
Winner-Take-All network, WTA, 87
Wizard-of-Oz, WoZ, 246