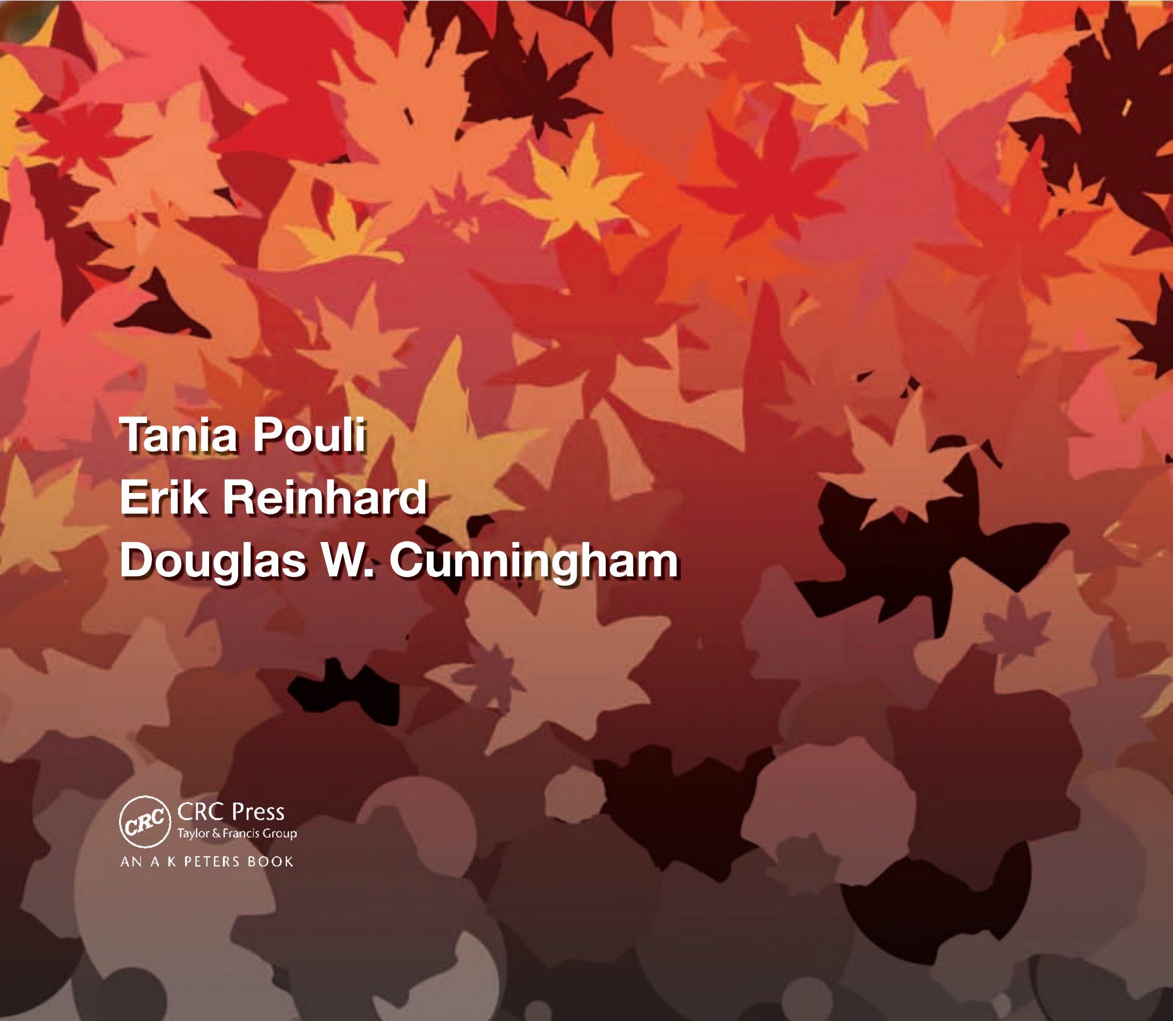
Image Statistics
in Visual Computing



Tania Pouli
Erik Reinhard
Douglas W. Cunningham

CRC Press
Taylor & Francis Group
AN A K PETERS BOOK

Image Statistics

in Visual Computing

Image Statistics
in Visual Computing

Tania Pouli
Erik Reinhard
Douglas W. Cunningham

CRC Press
Taylor & Francis Group
AN A K PETERS BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20130712

International Standard Book Number-13: 978-1-4665-3982-2 (eBook - Kindle)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

ER & TP: To our parents Dora & Panagiotis and Ineke & Franz, for their support, patience, and love despite our always forgetting to call them.

DWC: I would like to thank my wife, Monica, and my son, Samuel, not only for their support but also the useful reminders that there is a world out there beyond book-writing.

Contents

II Image Statistics 69

III Beyond Two Dimensions 225

Foreword

To me, the term *natural image statistics* does not merely refer to a collection of statistics one can collect from images but signifies a paradigm shift in the way we reason about images of natural scenery. Instead of modeling them as a deterministic outcome of a physical process using explicit geometric entities and illumination laws, we started considering natural images as random variables characterized by unique statistical regularities.

In fact, these two modeling approaches, integrated in the Bayesian framework, have proved themselves most effective for a wide range of image restoration tasks. For example, when removing noise or blur from an image, the signal-to-noise level becomes very low across the high-frequency band and therefore a large number of unknowns must be recovered. State-of-the-art denoising and deblurring methods fill in this missing data based on the sparse distribution that derivative and wavelet coefficients exhibit in natural images. Similarly, when filling holes in an image or increasing its resolution, millions of pixels must be determined. Non-parametric patch-based statistical models show a remarkable ability to recover this seemingly missing data by exploiting various self-similarities found within and across natural images.

The statistical investigation of natural images has also proved itself valuable for studying the human visual system, which was exposed and adapted to this stimulus throughout its evolution. Researchers showed, for example, that image transformations based on some optimal statistical criterion, such as compact representation, provide important insights over the response of various sensory neurons.

These practical and theoretical implications make the search for new and more accurate natural image priors a very active line of research in image understanding, processing, and neuroscience.

The book *Image Statistics in Visual Computing* offers a delightful coverage of this fascinating subject. It starts by describing fundamental concepts of human visual perception, as well as important calibration aspects behind the collection of natural image databases. The book thoroughly covers key image transformations and describes their unique statistical behavior. It devotes a chapter to Markov random field models, which constitute the most widely used framework for defining image priors. Finally, the book surveys several important extensions of the statistical study to the behavior of color, motion, and range images.

Raanan Fattal
School of Computer Science and Engineering
The Hebrew University of Jerusalem

Preface

Natural images exhibit statistical regularities that differentiate them from random collections of pixels. Moreover, the human visual system appears to have evolved to exploit such statistical regularities. The field of natural image statistics is traditionally concerned with trying to understand how human vision operates by identifying statistical regularities in natural images that may be used in some way.

There are some amazing examples of how the human visual system is beautifully optimized to observe and interpret natural images. One example that put us on the trail of natural image statistics is described in detail in Chapter 10: Take a large set of pixels from a collection of natural images and transform them to cone response space. If these three color channels are then rotated using principal components analysis, the result is a new color space that is formally decorrelated.

As it happens, the color space created this way strongly resembles a color opponent space. It is known from many sources that the human visual system also employs a color opponent space to encode the visual signal prior to sending it from the eye to the brain. The beauty of this is, therefore, that human vision actively decorrelates its visual input, provided it is observing natural images. In practice it appears that the three decorrelated channels are closer to independence, which is an even stronger statistical claim.

This has subsequently given rise to applications in image processing that treat the three color channels in color opponent space as independent, simplifying an otherwise more difficult three-dimensional problem into three one-dimensional problems. Color transfer algorithms form a good example of a class of algorithm in which this appoach is successfully explored.

The lesson learned from this example—and the premise for this book—is that the field of natural image statistics is not only helpful in the study of human vision but may also create opportunities that can be exploited in visual engineering disciplines such as computer vision, computer graphics, and image processing. Thus, as many areas of visual computing are concerned either with producing imagery for observation by humans or with processing images to obtain information similar to what humans would extract, it would be prudent to understand which statistical regularities occur in nature so they can be emulated by image synthesis and analysis methods. In this book we introduce all aspects of natural image statistics, ranging from data collection to analysis and finally their applications in computer graphics, computational photography, image processing, and art.

We do not aim to provide an exhaustive list of all applications that use natural image statistics in some way; instead, we see this as a book that may give readers those all-important “aha” moments. We think that there are many possibilities for natural image statistics to be used in a variety of applications, of which the currently known ones only scratch the surface. We hope that this book will serve as a trigger for new and fresh ideas, leading to novel insights and applications in all disciplines that relate to visual computing—both in academic and industrial settings.

To serve this purpose, we have tried to keep the book as accessible as possible. Wherever appropriate, we provide the basic mathematics to understand the transforms that emphasize some of the many statistical regularities present in natural images. The statistical regularities and the patterns that can arise once they are transformed are described in some detail. We then give examples of applications that have successfully used these patterns. To keep the book manageable, we definitely do not aim to provide an exhaustive list of examples, nor do we intend to describe each example in full mathematical detail. We do, however, provide many references so that more information about any topic and specific application discussed in this book can be easily found. In that sense, we hope that the book may serve as an entry point into a highly interesting area of research that brings together aspects of statistics, perception, neuroscience, physics, image processing, computer vision, and computer graphics.

Cover Image

One of the most fascinating aspects of natural image statistics is their ability to deconstruct complex images of natural environments into compact descriptors. The seemingly endless variety of nature leads to striking—and repeatedly occurring—regularities. The image used in the cover tries to show exactly this. Starting from a photograph of a natural scene, the maple leaves are slowly transformed to progressively simpler shapes, eventually converging to circles arranged according to the “Dead leaves model,” discussed in Chapter 11.

Acknowledgments

We are grateful for the help offered by many friends and colleagues who have assisted either directly or indirectly with the production of this book. In particular,

we would like to thank A K Peters, the Max Planck Institute for Informatics, and the Brandenburg Technical University for their support. We would also like to thank Ayan Chakrabarti, Mark Fairchild, Bill Geisler, Jan-Mark Geusebroek, and Dale Purves for allowing us to use their images. Further, we would like to thank Karol Myszkowski, Patrick Perez, and Hans-Peter Seidel. We are also very grateful for the great foreward that Ranaan Fattal has written for us.

Many researchers have made their work available online, and this has proved to be a real help, allowing us to reproduce their results on our own images. For instance, for the production of several figures we have adapted Stanford's Wavelab package,[1] which is a suite of MATLAB tools for experimenting with wavelets. The Wavelab team and list of contributors consists of David Donoho, Arian Maleki, Morteza Sharam, Jon Buckheit, Maureen Clerc, Jerome Kalifa, Stephane Mallat, Thomas Yu, Mark Reynold Duncan, Xiaoming Huo, Ofer Levi, Shaobing Chen, Iain Johnstone, Jeffrey Scargle, Rainer von Sachs, Thomas Yu, Jeffrey Scargle, and Eric Kolaczyk. We gratefully acknowledge their efforts.

Stéphane Mallat's book *A Wavelet Tour of Signal Processing*[2] is accompanied on the internet by a large set of numerical experiments that can be used in the classroom.[3] These are designed and maintained by Gabriel Peyré. We thank him for providing us with the answers to these excercises, significantly simplifying our experiments and the production of further figures for our book. The "Dead leaves" figures in Chapter 11 (and indeed part of the front cover) were made using Gabriel's MATLAB "Toolbox Image."

We would also like to thank Ivan Selesnick, Shihua Cai, Keyong Li, Levent Sendur, and A. Farras Abdelnour at the Department of Electrical and Computer Engineering of Brooklyn Poly for making available on the internet a collection of MATLAB routines to experiment with several wavelet transforms.[4] In particular, we have used their dual-tree complex wavelet transform implementation for experimenting with correlations across scale of the magnitude of complex wavelets.

Some of the ICA related figures in Chapter 7 were created by modifying a version of the FastICA code that is kindly made available by Aapo Hyvärinen, Jarmo Hurri, and Patrik Hoyer as part of their book on natural image statistics,[5] a book that served as an inspiration for ours, albeit with a different focus.

The texture synthesis examples were generated with the code kindly made available by Javier Portilla and Eero Simoncelli.[6] This code relies on the multiscale image processing toolbox made available by Eero Simoncelli.[7]

[1] www-stat.standord.edu/~wavelab

[2] www.ceremade.dauphine.fr/~peyre/wavelet-tour/

[3] www.ceremade.dauphine.fr/~peyre/numerical-tour/

[4] eeweb.poly.edu/iselesni/WaveletSoftware/index.html

[5] www.naturalimagestatistics.net

[6] www.cns.nyu.edu/lcv/texture

[7] www.cns.nyu.edu/~eero/software.php

Chapter 9 ("Markov random fields") contains some figures that were generated with code kindly made publicly available by Uwe Schmidt, Qi Gao, and Stefan Roth.[8]

We would like to also thank Christian Winger, a student at the Brandenburg Technical University in Cottbus for allowing us to use his implementation of the inpainting algorithm by Crimini et al. [134] for the figure in Chapter 5. This method was implemented and extended as part of Christian's bachelor's thesis.

Our gratitude also goes to all the great people at A K Peters and CRC Press for their help and guidance throughout the publishing process. Our editors, Sarah Chow and Charlotte Byrnes, have been a pleasure to work with, and their ideas and corrections have helped make the book better. Finally, we'd like to especially thank Alice and Klaus Peters as well for signing us up in the first place; we hope they're enjoying their well-earned retirement.

[8] www.gris.tu-darmstadt.de/research/visinf/software/index.en.htm

Part I

Background

Chapter 1

Introduction

For many reasons, including basic survival, we continuously sense the environment around us. Light is perhaps the premier carrier of information about our environment. The signal it conveys is also the most complex to observe and interpret. It should therefore not come as a surprise that a significant portion of the human brain is dedicated to processing the light reaching our eyes (although there appear to be significant differences between individuals in terms of how much of the brain is devoted to the processing of visual stimuli [17]). Images—both still and moving—contain a wealth of information: a quick look through the nearest window is enough to tell us what the weather is like, whether we are in the countryside or the city, and what time of the day it might be. Movement in a scene might tell us of an approaching predator or that the bus is approaching, while color and texture may inform us about the state of ripeness of fruit or whether our loaf of bread has gone off again.

Considering that human vision takes as input at any point in time an image of the world through each eye—effectively two projections of the world onto the retina—it is remarkable that we are able to infer so much of the world around us. These two streams of flat images are sufficient to allow us to navigate complex environments, meaningfully respond to events, and recognize individuals. Anyone who has tried to implement an algorithm to reconstruct a three-dimensional (3D) environment from a pair of two-dimensional (2D) images knows how impossibly difficult this is. Yet, human vision appears to accomplish this effortlessly.

Thus, human vision has the task of inferring at least some aspects of the 3D configuration of our world, given 2D image inputs. This is known to be an underconstrained problem: in essence, human vision has to invent one dimension which has gone missing as a result of the projection of light on the retina. It has been repeatedly shown that human vision employs many heuristics, some of which are known. As an example, human vision generally expects a single light source [598] that comes from above [407]. Face recognition, for instance, becomes significantly impaired if a human face is illuminated from below.

There is a long history of research on the processes taking place within the human (and animal) visual system that enable us to perceive and interpret images. Vision scientists typically study such processes through carefully constructed psychophysical experiments that measure overt behavior to systematic changes in simple light patterns [498, 554, 138] or by directly measuring electrophysiological responses while the retina is stimulated with a simple light pattern [354].[1] In the latter case, the light pattern that maximally stimulates a given cell is known as that cell's preferred stimulus, and it carries considerable information about the structure and size of the cell's receptive field. Alternatively, one could measure behavior or cell responses while the retina is stimulated with natural images, since that is what the retina usually processes [445, 616, 151, 30, 764].

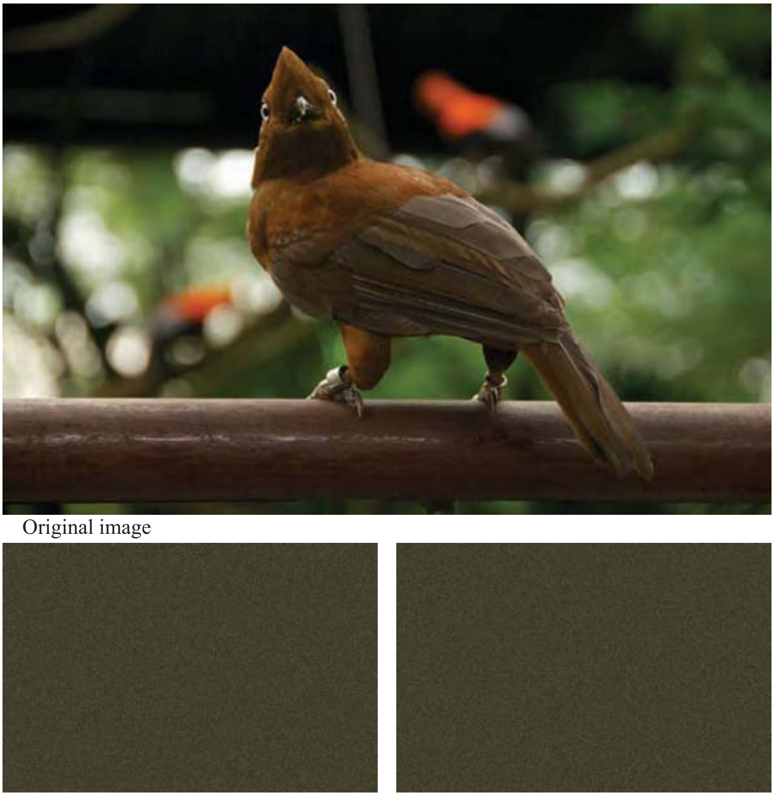
Naturally, the ability of the visual system to analyze information is very much related to the information that is there to be processed, which just happens to be images. The field of natural image statistics has focused on exactly this relationship. Initially the field focused on describing images such as television signals [168, 424], but it has since expanded into explicitly examining the many links between the structure of a wide variety of image types on the one side and the workings of the visual system on the other. In general, natural image statistics research collects a set of natural images, transforms them according to a statistical optimization criterion, and then evaluates whether the result in some sense forms a reasonable description of the response properties of some neurons [645, 234, 26, 546, 691].

The argumentation is roughly as follows. The retina has a large number of discrete photoreceptors, so that the projection of light is spatially discretized. Equivalently, photographs contain a large number of discrete pixels.[2] The values that these elements can take is not fully random. This can, for instance, be seen in Figure 1.1, where at the top a natural image is displayed. Below, we have shuffled the order of the pixels. All pixels in the top image are also present in the bottom images. However, the bottom images do not carry meaningful information. Moreover, human vision is not well equipped to discern differences between images such as these.

As the images at the bottom of Figure 1.1 are effectively recombinations of the pixels of the image at the top, one may ask how many combinations of pixels there exist for every random image [637]. Assuming that we have a small-ish image of 4 Mpix = 4×10^6 pixels, then the number of ways these pixels could be recombined would be $(4 \times 10^6)!$, which for practical purposes approaches infinity. This means that the number of natural images that we might encounter in real life is infinitely smaller than then number of possible combinations in which the photoreceptors in the retina could be stimulated [395, 215, 154, 639].

[1] See hubel.med.harvard.edu.

[2] We are assuming for this discussion, as well as for the remainder of the book, that images are digitally stored and represented, and that they are therefore spatially discretized in a pixel grid.



Original image

A re-combination of pixels

Another re-combination of pixels

Figure 1.1. The pixels of the top image were randomly shuffled twice to form the images at the bottom. These are only two of an almost infinite number of recombinations possible. Note how human vision does not easily detect differences between these two images. (Andean Cock-of-the-rock, Jurong Bird Park, Singapore, 2012)

As a consequence we might argue that natural images are in fact incredibly rare. For every natural image there is an infinite number of stimuli that are random combinations of pixels. Most of these random combinations, however, do not reveal structure that can be interpreted by human vision. The subset of natural images is sometimes called an image manifold [704]. It therefore stands to reason that human vision has evolved to observe, interpret, and make sense of this incredibly rare subset of all images [257].

Despite being a rare subset of all possible images, natural images still contain an immense degree of variation. Within this plethora of differences are statistical regularities that make natural images amenable to human interpretation. What are these features that natural images have in common with each other that differentiate them from random images? Furthermore, in what ways does the human visual system expect "naturalness"? In this book we look at these and similar questions, which are core to this exciting area of research called natural image statistics.

Beyond helping to understand how the human visual system works, the regularities in natural images are very valuable to disciplines that rely on images, such as computer graphics, computer vision, or image processing. Similar in spirit to the processing blocks within the visual system, computational models can take advantage of the regularities in natural images to produce more plausible results or to infer information that may otherwise be lost.

If, for instance, the aim of a particular technique is to reconstruct—or *inpaint*—areas of an image that may be hidden, statistical information from nearby regions can be used to create a patch that looks similar, yet not identical, to its neighbors [568] (see Section 5.8). In another scenario, statistical regularities computed over a collection of images can serve as priors, guiding (optimization) algorithms [211, 453, 667]. The distribution of edge magnitudes in natural images, for example, can aid algorithms aiming to *deblur* images (see Section 5.6). In effect, this produces an expectation of how edges are typically distributed in images, helping the deblurring algorithm in question produce a solution that conforms to that expectation. Likewise, an analysis of art has shown that paintings often have the same statistics as real-world images, and thus a number of existing techniques can be used to analyze, recover, or otherwise alter paintings (for reviews, see [279, 602, 709]).

Although the distinction is not always as clear cut, images can be analyzed in ensembles or individually. In the first case, the statistical analysis can provide insights about regularities and properties of general scene types and situations, which can in turn be applied when manipulating further images of that type. In the latter case, the information acquired by looking at an individual image is only useful for dealing with that particular image. However, this approach has shown to provide powerful statistics internal to the image itself [829]. Statistics are of course also commonly used in a third way; namely, for analyzing the results of studies or experiments often performed for assessing imaging algorithms [138], although this is somewhat outside the scope of this book. The following sections will briefly introduce the possible applications of statistics in the context of imaging disciplines.

1.1 Statistics as Priors

In the same way that human vision has certain expectations or heuristics regarding its visual input, we may employ natural image statistics to guide algorithm design to produce solutions that may be perceived as more natural or realistic. Examples include image deblurring [211] and image restoration or denoising [831] (see

Sections 5.6–5.8). The most direct way in which this may happen is in optimization-type algorithms.

An optimization algorithm requires an objective function—the function that is minimized—which encodes a set of opposing characteristics that the result ought to have. The algorithm typically begins by guessing an initial solution. The quality of the solution is then measured by inserting it into the objective function. Then iteratively a new candidate solution is guessed and compared against the previous best solution. There are several more or less standard choices for optimization processes, defining how the next guess is selected, but the main challenge is in the design of an appropriate objective function.

Objective functions usually consist of multiple terms. The first term is usually a norm that compares the current guess against some desired goal. The closer the image is to the goal, the smaller the number that this term returns. As this term will admit many solutions that are mathematically equally good, extra conditions are normally added to steer the optimization into a direction that is good according to some other, higher-level criteria.

Thus, such additional terms are designed to penalize wrong solutions and promote more appropriate ones respectively. They act as priors in the sense that they encode expectations regarding the mathematical shape of the solution. Often such priors take the form of smoothness terms. When the goal of the optimization algorithm is to filter or generate an image, natural image statistics have proven to be a good source of priors, effectively allowing some statistical form of naturalness to be encoded. An example is the deblurring techniques mentioned earlier: the average distribution of gradients in images can serve as a prior to suppress spurious edges that might arise from the deblurring process [211, 667].

In the particular case of image gradients, they are found to often have small values and only sometimes have large values, as shown in Figure 1.2. In other words, the distribution of gradients is heavy-tailed rather than Gaussian. Note that although here we show the results computed on a single image, this trend holds in general for natural images. Thus, if we were to compute the gradient distribution over a large set of natural images, we would find very similar results, as we will see in Chapter 5. This means that it is possible to use this distribution as a general prior in optimization problems, whenever we wish our resulting image to have a gradient distribution that is in some sense “natural.” Accounting for the non-Gaussian distribution of gradients has helped better solve several problems, including image deblurring [211], image denoising [631, 686], superresolution, and demosaicing [726].

1.2 Statistics as Image Descriptors

More often than not when manipulating images, the most useful source of information is none other than the image itself. If the goal is to add objects to a photo-

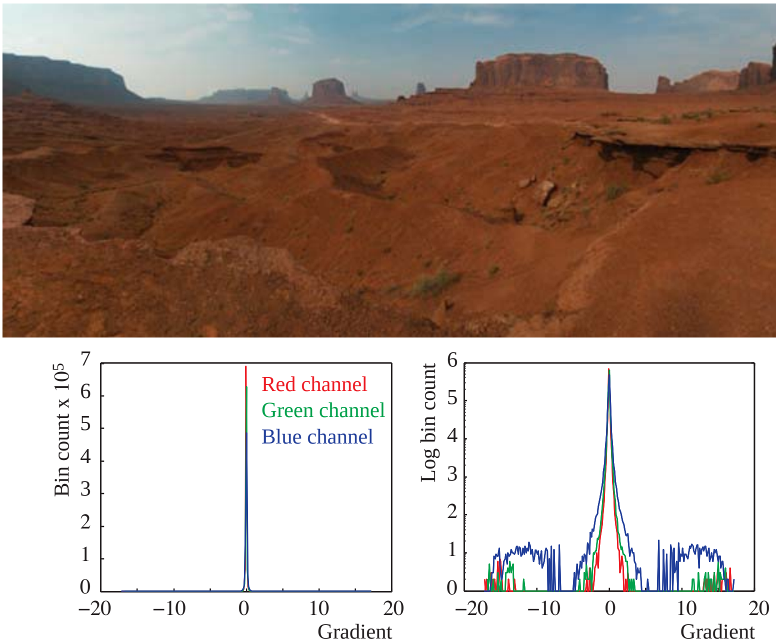


Figure 1.2. Image gradients are heavy-tailed, as shown in this example. The image at the top was first converted to log space, and then the horizontal gradients were computed. They are plotted in a linear plot (bottom left) and for visualization in a log-linear plot (bottom right). Note that most gradients are small, leading to a heavy-tailed distribution. (Monument Valley, USA, 2012)

graph, for instance, clues as to how the new object would look if it were actually placed in the real scene can be obtained by analyzing the image itself. This can, for instance, provide information regarding color and illumination (mis-)matches.

Conversely, if we wish to remove an object from an image, it is possible to analyze regions that neighbor the pixels to be removed. Here, statistics may help to infer what would be the most likely scene that lies behind the removed object. In this case, no general statistical assumptions are made about the input image, i.e., it is not desirable to try and infer a specific background based on ensemble statistics.

Instead, statistical tools are used to analyze the scene and purely based on those, the missing parts can be reconstructed. For instance, in the image shown in Figure 1.3, we may want to remove the camel without making it obvious that the resulting image was manipulated in any way. Areas surrounding the camel contain pixels similar to what would likely be behind it, and as such they can be used to estimate the missing information. Since only a single image is available, an accurate reconstruction would not be possible, but relying on statistical similarities

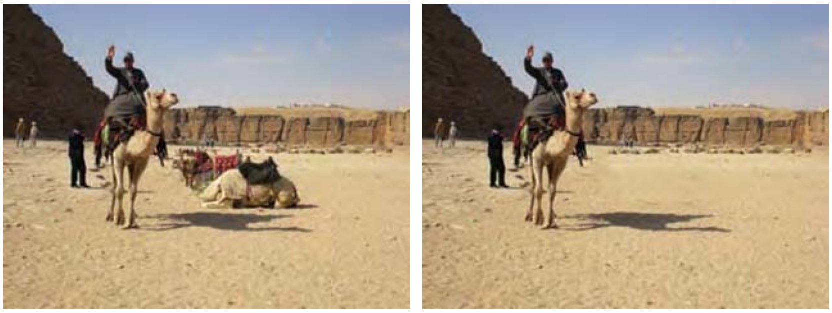


Figure 1.3. Inpainting example. Note that this is only an illustration, created semi-automatically with Photoshop. See Section 5.8 for statistics-based inpainting. (Giza, Egypt, 2007)

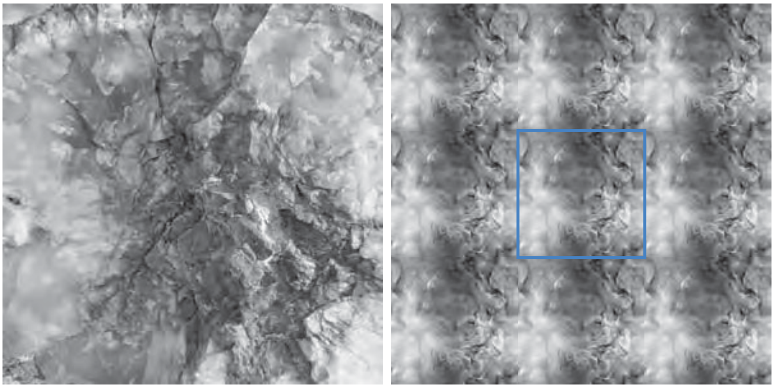


Figure 1.4. An example of a new texture (right) that was synthesized based on the example texture given on the left. The texture on the right was tiled nine times.

could be sufficient to produce a plausible result. Image inpainting is discussed in more detail in Section 5.8.

Another example of where the statistics of an individual image can be meaningfully used is *texture synthesis*. Given an example image with a desired appearance, the goal of texture synthesis algorithms is to generate a texture tile with similar appearance. There are many algorithms based on resampling [787, 195, 157] and some that explicitly take image statistics into account [692, 581, 582]. In the latter case, statistics based on a complex wavelet decomposition in the output texture are forced to match those of an example texture. An example is shown in Figure 1.4 and further discussed in Chapter 8.

There are cases where statistics can serve both as a global prior, capturing general assumptions and expectations about images, and as a tool to analyze specific input. If, for instance, the goal is to white balance an image such that color casts of a non-white light source are removed, one can analyze the color content of that image to that effect [69]. The statistical assumption made in general about scenes is that the average reflectance of all materials is achromatic (the gray-world assumption). This means that if we illuminate a scene with a colored light source, the light reflected off objects (and subsequently captured in an image) would be white only if the color of the illuminant were white. If the average color of an image therefore deviates from gray, that difference is an indication of the color of the light source.

In this case, the expectation that images will average to middle gray serves as a global prior, albeit in a simple sense. On the other hand, the specific statistics of each image, namely, the average of all its pixel values, describes that particular image and how it differs from the expected value. This technique for white balancing, as well as variants thereof, are discussed further in Chapter 10.

1.3 Statistical Pipeline

Whether statistical regularities are computed over image ensembles or single images, a similar process takes place. First, a dataset is constructed (Chapter 3). This may be a collection of images (an ensemble), a collection of patches selected from one or more images, or even single pixels, depending on the application.

The dataset is then transformed from the original image space to a different space, which allows the study of a particular aspect of the image data. This may be a transform to the gradient domain if edges are of interest or to a different color space, for instance. In effect, this transformation takes the form of a function that operates on single or multiple pixels. As an example, consider a color space transform from sRGB to the LMS color space. In this case, the function operating on the image pixels is a matrix transformation applied to each triplet of color values (Chapter 10). If one is studying gradients instead, two or more pixel values are considered to determine the rate of change between them (Chapter 5). The transform is usually designed to extract or emphasize some statistical regularity present in images. Often, the transform is related to some aspect of human vision, in that the transform is matched to the processing thought to occur in the human visual system.

Once the data has been transformed to the appropriate space, the extracted features need to be analyzed. This is achieved through statistical analysis similar to what would be employed to compute trends in any type of data. One may average along one or more dimensions, look at the distribution of values, or construct a model that predicts the measured data.

It is important to note that what this process typically aims to achieve is twofold. First, statistical measures can reduce the dimensionality of the data while retaining information relevant to a particular application. This is one of the key factors making image statistics a very useful tool for analyzing images, which by their nature are rich in data. Second, the statistical analysis can aid in deriving models for the data.

1.4 Natural Images

At this point, it is perhaps worthwhile to briefly discuss what we mean by natural images. In the context of visual computing, an image is typically a 2D array of (colored) pixel values. The content depicted in an image determines whether this image could be considered natural or not. There are many images that are not natural images, including X-rays, images produced with magnetic resonance imaging (MRI), and forward looking infrared (FLIR) imagery. Often, an image is considered to be the realization of a spatial stochastic process [704]. It is common to assume that this stochastic process is stationary, which means that the statistics are the same everywhere in the image.

It would also be prudent to limit oneself to images that have sufficient complexity, ranging from surface reflectance and texture to object shapes, and to require that objects sit on a ground plane with the camera looking more or less forward. This is consistent with how humans normally observe the world, and it may therefore be reasonable to assume that human evolution, as well as personal development, has occurred under similar circumstances.

To what extent human vision is the result of eons of evolution versus the result of personal development may also impact the choice of images that are studied. If it is assumed that the organization of human vision is due to evolution only, then it would make sense to exclude images that depict built-up environments. On the other hand, if individual development plays an important role, then natural image ensembles may well include images of interiors and built-up environments.

In this book, we make no strong assumptions either way. We are interested in natural image statistics and how they may help solve problems in visual computing. As the need arises, for certain applications it would make sense to restrict the type of images analyzed further, or conversely expand the set of images considered. We have therefore deliberately left the title of this book at *Image Statistics* rather than the perhaps more expected *Natural Image Statistics*. Nonetheless, when capturing a dataset of subsequent analysis, certain care must be taken to not inadvertently introduce bias. This topic is discussed further in Chapter 3.

1.5 Discussion

We have found that most transforms that could be applied to pixel data and which have some bearing on human vision all more or less point into the same direction, which is that of sparseness. Statistical distributions of transformed data tend to have high *kurtosis*, which is another way of saying that the data tend to have most values clustered around one value, with relatively few data points far away from this value. Such behavior is apparent when looking at distributions of gradient magnitudes, power spectra, and wavelet coefficients, as well as principal and independent component analysis. This means that many of these transforms lead to probability distributions that are non-Gaussian. Sparse, highly kurtotic distributions appear to be a feature of the 3D environment we live in (Chapter 11), and this persists after projecting 3D environments to 2D images.

Another way to look at this is that the early parts of the human visual system in some sense remove information that is expected, while keeping only data that is informative [741]. It can therefore be argued that the human visual system aims to reduce redundancy as much as possible [34, 342].

The knowledge of various sparse distributions has given many insights into the way human vision operates, and it has led the way for a good number of applications in computer graphics, computer vision, and image processing. There is, however, a trend that is apparent. If we know what statistical distributions correspond to natural images for certain image transforms, then one could measure deviations from naturalness in the statistical sense. It is then possible to design algorithms that enforce such naturalness to individual images or image regions. This is directly exploited, for example, in the image inpainting examples mentioned earlier. Likewise, motion deblurring often includes priors that enforce heavy tails in gradient distributions. Natural image statistics have also found applications in image restoration, superresolution, and texture synthesis algorithms. In each case, an input image is processed such that its improvement is guided using image statistics.

The statistics in a sense form an oracle as to what the most likely improvement to the image might be. Thus, in general, any algorithm that aims to enhance or improve images could in principle be designed to take advantage of knowledge of natural image statistics. Throughout this book, we will discuss many such applications that already demonstrate the usefulness of image statistics, and we will show how statistical findings provide insight into human visual processes that can be directly applied to imaging algorithms.

The book is structured in three parts. The first part provides some useful background in the context of human vision and perception, which we will rely upon throughout later chapters to show how image statistics link with vision. We also discuss the techniques and challenges related to capturing and calibrating images for the purpose of statistical image analysis.

The second part forms the core of the book and discusses several classes of image transforms. Each chapter in that part explores a different space through which images may be transformed (including but not limited to gradients, wavelets, and histograms). We will present the mathematical underpinnings of each transform, discuss its links with human vision, and present statistical regularities that arise. As this book focuses on imaging applications, examples from computer graphics, computer vision, and image processing will be given in each chapter, demonstrating how that particular transform and the statistical structures within it may be used in image-related disciplines.

Finally, the third part of the book focuses on different types of image data. Most of the natural image statistics literature centers around luminance or intensity data, ignoring aspects such as color, depth, or temporal data. While many powerful techniques are possible without these dimensions, they undoubtedly provide a lot of additional information. To that end, the last part of the book focuses on the statistical analysis of these dimensions and the applications that arise from them.

Chapter 2

The Human Visual System

In this chapter, we briefly discuss the human visual system and human visual perception. We begin by defining relevant radiometric and photometric terms. Then, the human visual system is described by following light as it enters the eye and is transduced into an electrical signal by the photoreceptors. We then follow the path this signal takes through several layers of cells in the retina and onwards to the visual cortex. Although much is known about cortical visual processing, there is even more that remains unknown. We therefore describe cortical processing only at a very abstract level and only sofar as is relevant in the context of natural image statistics.

In essence, the nature of human vision is such that natural images can be observed, interpreted, and understood at a level necessary to allow humans to function in a manner appropriate to their environment. Thus, this chapter concludes with a discussion of the implications of visual processing.

2.1 Radiometric and Photometric Terms

To be able to describe light as it is observed by humans or captured and displayed by various technologies, several important terms need to be defined. Radiometric terminology describes light as energy. Radiometry is the science of measuring optical radiation, which occurs in a range of wavelengths between 10 nm and 10^5 nm and includes ultraviolet, visible, and infrared radiation. Radiometric quantities are listed in Table 2.1.

Photometry parallels radiometry, with the important exception that quantities are weighted according to the spectral sensitivity curve of the human visual system (given in Figure 2.1) and therefore only extends between around 400 and 800 nm. Photometric quantities are listed in Table 2.2.

Quantity	Unit	Description
Radiant energy (Q_e)	J (Joule)	Total energy of a beam of radiation
Radiant flux (P_e)	J/s = W (Watt)	Rate of change of radiant energy
Radiant exitance (M_e)	W/m^2	Radiant flux emitted from a delta surface area
Irradiance (E_e)	W/m^2	Radiant flux striking a delta surface area
Radiant intensity (I_e)	W/sr	Flux emitted into a given direction
Radiance (L_e)	W/m^2/sr	Flux per area emitted into a given direction

Table 2.1. Radiometric quantities.

Quantity	Unit	Radiometric equivalent
Luminous energy (Q_v)	lm s	Radiance energy
Luminous flux (P_v)	lm (lumen)	Radiant flux
Luminous exitance (M_v)	lm/m^2 = lx (lux)	Radiant exitance
Illuminance (E_v)	lm/m^2 = lx (lux)	Irradiance
Luminous intensity (I_v)	lm/sr = cd (candela)	Radiant intensity
Luminance (L_v)	lm/m^2/sr = cd/m^2	Radiance

Table 2.2. Photometric quantities.

In human vision one further quantity is often used, which is the Troland (td), a measure of retinal illuminance. The reason for this is that the light that hits the retina has passed through the pupil, which has a diameter dependent on the amount of light incident. The Troland accounts for pupil size by multiplying luminance (in cd/m^2) by the size of the pupil (in mm^2).

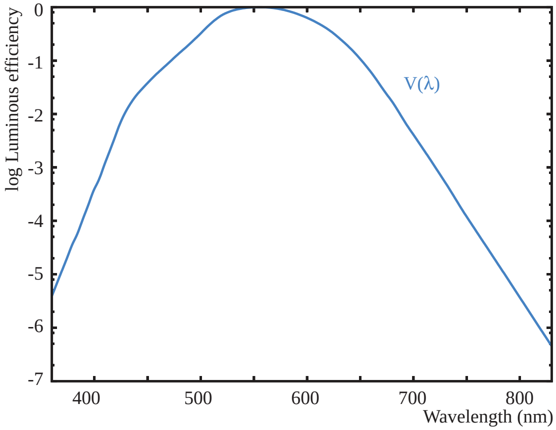


Figure 2.1. The photopic luminous efficiency curve.

2.2 Human Vision

The human visual system forms a significant portion of the brain and is dedicated to processing visual information. Light enters the eye and is eventually transduced into an electrical signal by the photoreceptors. Through several layers of cells with associated processing, a heavily transformed signal is passed to the lateral geniculate nucleus (LGN), a part of the brain that can be thought of as acting as a relay station. After a small amount of further processing, the signal travels from the LGN to the visual cortex where it is processed by a multitude of modules, each responsible for increasingly abstract and high-level tasks [200]. The routes that the signal takes from the eyes to the brain and each of these modules are known as the visual pathways.

Human vision can be studied in many different ways. Visual psychophysics, for instance, presents meticulously designed light patterns (called stimuli) while asking observers to perform certain tasks [258, 554, 402, 138, 699]. Task performance is measured a large number of times for a sufficiently large number of observers using systematically altered versions of the stimuli, which then allows the experimenter to infer something about the inner workings of the human visual system.

In neuroscience, a different approach is taken, one that involves measuring cell responses using electrodes [357]. This cannot normally be done on humans, and is therefore typically performed on animals with visual systems that are thought to be comparable to those found in humans. Many aspects of low-level visual processing have been uncovered in this manner, especially relating to receptive fields of a variety of specific cells. The receptive field of a cell has several critical characteristics, such as the location on the retina or in the environment to which is responds and the specific size, shape, and pattern of light that needs to be shone onto that cell for it to respond maximally. It is reasonable to expect that cells in the retina have simpler receptive fields than those in the visual cortex.

Recent advances in measuring cell responses have made it possible to record and measure groups of cells [542, 394]. Nonetheless, it remains difficult to obtain a view of how the visual system as a whole responds to stimuli. Visual psychophysics is only able to indirectly infer visual processing, while cell recordings often include only single cells. The field of natural image statistics has emerged as another tool for helping to understand human vision. The thought is that the human visual system over a long period has evolved to make sense of natural scenes, and as such is in some sense optimized for this task. It is therefore a good idea to understand the statistical regularities of natural scenes as a further tool to help understand human vision. Hence the emergence of natural image statistics as a field of study.

In our opinion, natural image statistics can also be useful in designing visual algorithms—for instance, in computer graphics, computer vision, and digital image processing. In this chapter, following the visual pathway, we give an overview

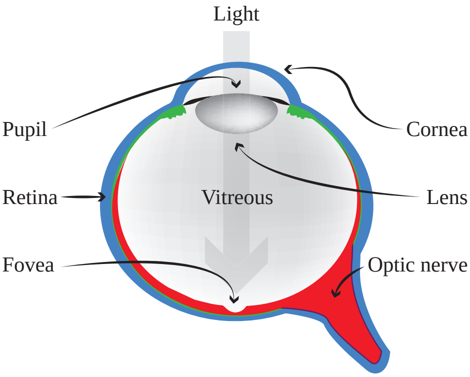


Figure 2.2. A simplified cross section of the human eye. (Adapted from [610].)

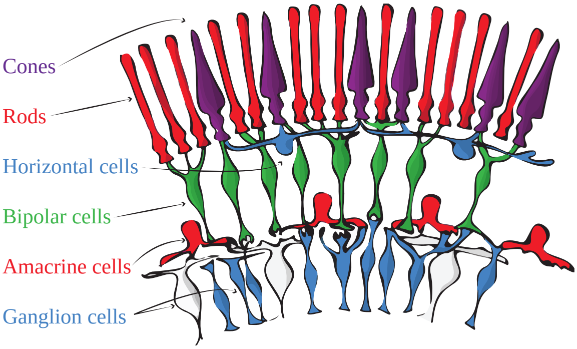


Figure 2.3. The human retina consists of several layers of neurons that mediate vision. (Adapted from [325].)

of the human visual system and its structure, as well as some of its idiosyncrasies. We note that by and large the retina is easier to measure than any of the later stages. As a result, more is known about the retina. Consequently, this chapter focuses mostly on the processing encountered in the human eye. It is intended to serve as background for subsequent chapters.

2.3 The Eyes

The eyes form the first part of a long chain of processing that allows humans to "see" the world [610]. The eyes sit in sockets that allow them to rotate, a feature that is useful for two reasons. First, it allows features of interest to be projected onto the fovea where vision has the highest acuity. Second, it allows humans to follow moving objects without head movement.

Figure 2.2 shows an annotated cross section of the human eye. Light enters the eye through the pupil, travels through the ocular media, and finally hits the retina. The retina is a layer of neural sensors lining the back of the eye. These sensors transduce light into a signal that is transmitted to the brain. Several layers of neurons are present in the retina (shown in Figure 2.3). Only the photoreceptors (i.e., the rods and cones) are sensitive to light.

2.3.1 Optical Media and the Retina

Light enters the eye through the cornea and then passes through the aqueous humor, a cavity filled with clear liquid, before reaching the iris and the lens. The visual field over which light can enter the eye extends to about 60 degrees nasally and 95 degrees temporally. Vertically, the field of view extends 60 degrees above and 75 degrees below the horizontal meridian. Both eyes together therefore achieve a field of view of almost 180 degrees.

The iris contains muscles to control the diameter of the pupil and thereby helps regulate how much light passes through.[1] The lens is attached to muscles that control its shape and is used to focus on objects located at different distances to the eye. Note, however, that the cornea performs most of the focusing due to its curvature and refractive index ($n = 1.376$), while the lens contributes the remainder [554].

After passing through the lens, light traverses the vitreous humor before reaching the retina, a neurosensory layer able to transduce light into signals and transmit those to the brain. The retina consists of several layers of cells as well as several regions with different properties. Light first passes through all layers of the retina before being transduced by the photoreceptors. The photoreceptors pass their electrical signal to a layer of bipolar cells, which in turn transmit the signal to the ganglion cells. Although the ganglion cells are located in the retina, they synapse in the LGN. Their axons are collectively called the *optic nerve bundle*. In between the photoreceptors and the ganglion cells, there are two layers of lateral connections, formed by horizontal cells and amacrine cells. As such, there is already a substantial amount of visual processing occurring within the retina. This processing is discussed further in the following sections.

As mentioned, the retina consists of a number of regions that perform somewhat different functions and are anatomically different. The central part of the retina contains the fovea, an area of 1.5 mm in diameter. In the center of this region is a smaller disk of 0.35 mm diameter, called the foveola, which is where the density of photoreceptors is highest. As a result, this is where visual acuity is highest as well. The region outside the fovea is the peripheral retina. This region also contains the optic disk, a small area known as the blind spot.

[1] Recent research has shown that photo-responsive ganglion cells regulate this mechanism [318].

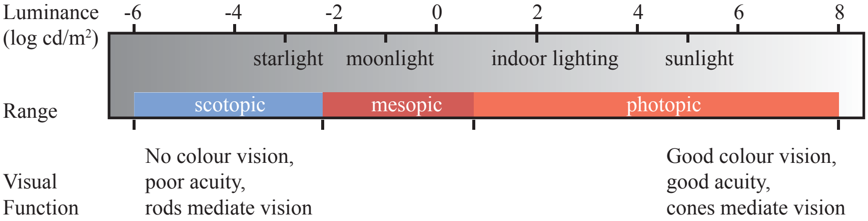


Figure 2.4. The ranges of luminance where the different types of photoreceptors are operational. Rods operate in low light conditions, known as scotopic, while cones cover the brighter range of illumination known as photopic. Some overlap exists between the two photoreceptor types. This is known as the mesopic range. (Adapted from [214].)

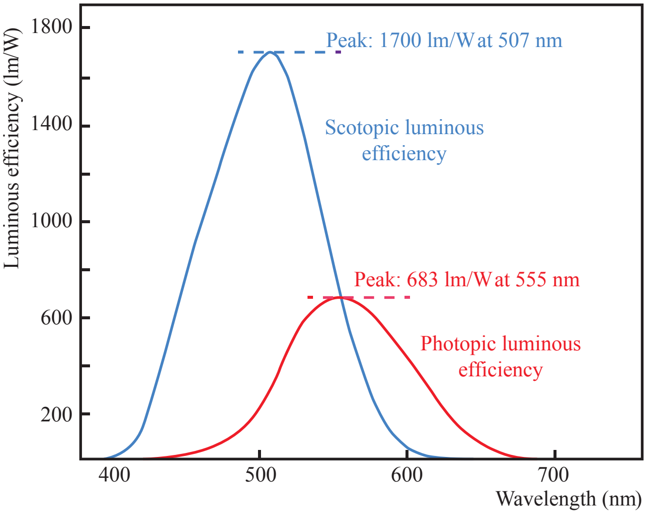


Figure 2.5. Photopic and scotopic luminous efficiency curves.

2.3.2 Photoreceptors

Two types of photoreceptors exist in our eyes, namely, cones and rods. These exist in different densities and fulfill different purposes [236]. Rods are sensitive to low light levels and allow us to detect contrast, brightness, and motion—but not color—in these conditions. The range of illumination that rod photoreceptors operate in is termed *scotopic*. Cones, on the other hand, operate at higher light levels, known as *photopic* conditions. In addition to contrast, motion, and brightness, cones are involved in seeing color (see Figure 2.4). There exists a range of light levels under which both rods and cones are active. This range is called *mesopic*. The responsivities to different wavelengths for the rod and cone systems

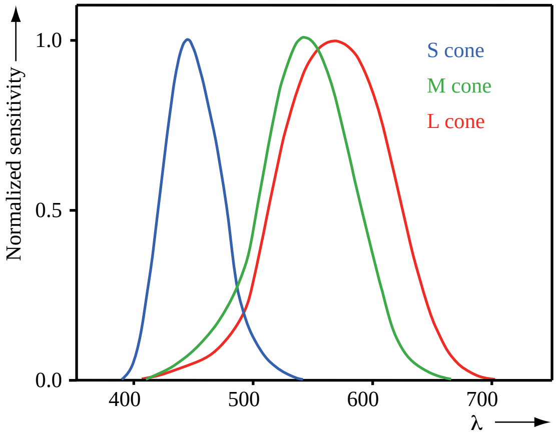


Figure 2.6. The sensitivities of the three cone types for different wavelengths. (Adapted from [68].)

are known as the *scotopic* and *photopic luminous efficiency*, which are plotted in Figure 2.5.

Photoreceptor Types. Three types of cones exist in the human eye and are classified according to the wavelengths that they are most sensitive to, namely, short, medium, and long (referred to as S, M, and L). These translate to peak sensitivities of approximately 440 nm, 545 nm, and 565 nm, respectively, corresponding roughly to blue, green, and red light, although as shown in Figure 2.6 significant overlap exists between the L, M, and S cone sensitivities [68, 707].

As each photoreceptor integrates over a wide range of wavelengths, there are many combinations of colored light that would lead to the same photoreceptor output. For instance, supposing we could change a monochromatic light from one wavelength to another and also change its intensity, it should be possible to produce a new stimulus that leaves the output of a given cone unchanged. If two different spectral distributions lead to the same response of all three photoreceptor types, they are visually equivalent to humans and are referred to as *metamers*. A second consequence of this integration is that a single photoreceptor type is not able to distinguish between different colors. For color vision to occur, the outputs of multiple cone types must be combined [700]. Processing in the retina does indeed achieve this, as the S, M, and L cones provide sufficient information to the remainder of the visual system to mediate color vision. Rods exist in only one type, and as a result color vision is impaired or even absent under scotopic lighting conditions.

Adaptation. A central property of photoreceptors, as well as many if not all cells involved in neural activity, is that of adaptation [783]. Photoreceptors can simultaneously transduce a range of illumination that spans about four orders of magnitude [431]. Our environment has illumination levels that span from starlight to direct sunlight, a range of about ten orders of magnitude (see Figure 2.4). Adaptive processes allow photoreceptors to function over this much greater range. We experience its effect on a daily basis: it is sufficient to walk into a dark room from bright sunlight to realize that our eyes adjust to allow us to see under the new conditions. Effectively, our visual system adapts to the prevailing lighting conditions so that we can still distinguish contrasts in the scene [774].

Adaptive processes give rise to both light and dark adaptation as well as chromatic adaptation. A demonstration of the latter is shown in Figure 2.7, where an image is shown with a yellow and a cyan filter applied. To recover the correct colors of the image, it is possible to adapt the retina to the yellow and cyan color casts. This is achieved by staring at the fixation cross at the top of the image for at least 30 seconds. The image below also has a fixation cross. By focusing on this fixation cross after the adaptation period, the image should regain its normal coloring.

Light Sensitivity. The light sensitivity of rods and cones is mediated by photosensitive pigments that react with light. As light hits the photoreceptors, the photosensitive pigments break down (bleaching). In the case of rods, these photopigments are completely depleted in lighting conditions above the mesopic range, while cones are not fully depleted even in very bright conditions [179]. Photopigment bleaching causes the electrical current present in a photoreceptor to change, which results in a neural response transferred to the brain. Through this transduction process, the retina can translate light energy into a neural response, which effectively is what allows us to see [325].

A final set of mechanisms occurring in the photoreceptors prevents them from saturating while adapting to the enormous range of illumination in nature [180]. Several studies have shown that the photoreceptors do not respond linearly to the full range of illumination [339, 5, 760, 179]. This nonlinearity was first studied by Naka and Rushton [525] and provides a very effective mechanism for the compression of dynamic range. Intensities in the middle of the range lead to approximately logarithmic responses while as the intensity is increased, the response tails off to a maximum. After that maximum is reached, further increases in intensity will not lead to a corresponding increase in response of the photoreceptors, ensuring that saturation will not occur [179].

Figure 2.8 shows the nonlinear response of the rod and cone cells to increasing luminance values. The response for both types of photoreceptors can be modeled with what is known as the Naka-Rushton equation:

$$\frac{V}{V_{\max}} = \frac{I^n}{I^n + \sigma^n} \tag{2.1}$$

where V is the response at an intensity I, $V_{\max}$ is the peak response at saturation, and σ is the intensity necessary for the half-maximum response (also known as

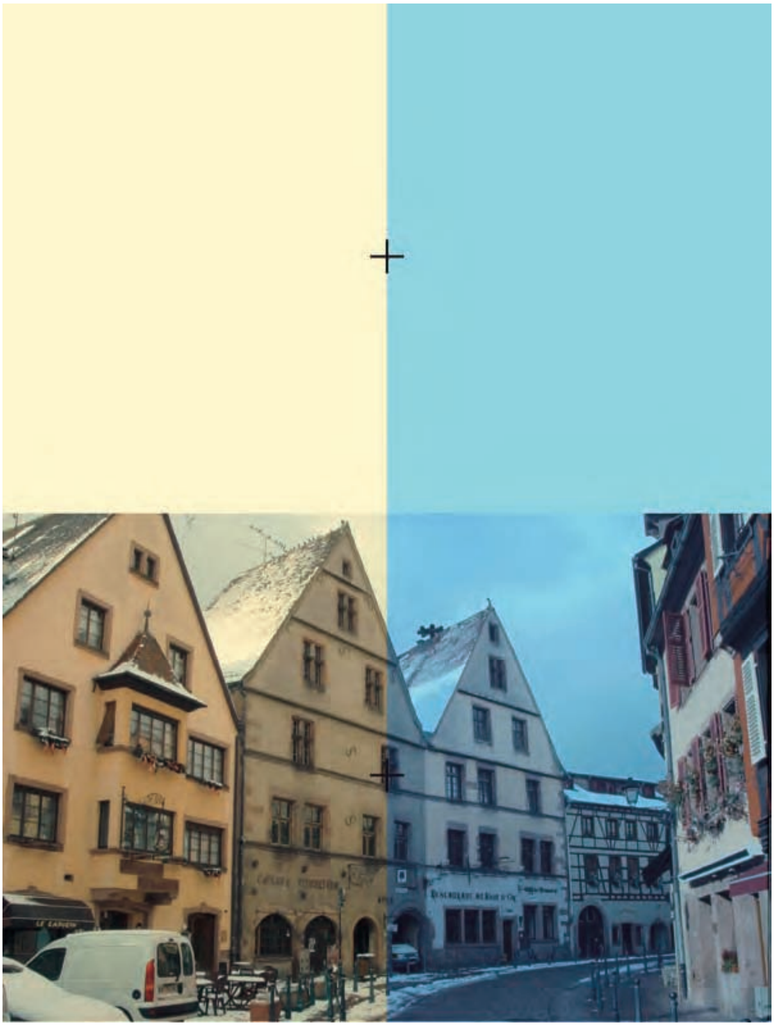


Figure 2.7. Adaptation is local, as can be observed in this demonstration. To see the correct colors of the image below, first fixate on the upper cross for at least 30 seconds to allow the retina to locally adapt. By then focusing on the bottom fixation cross, the image should appear normally colored. (Keysersberg, Alsace, France, 2013)

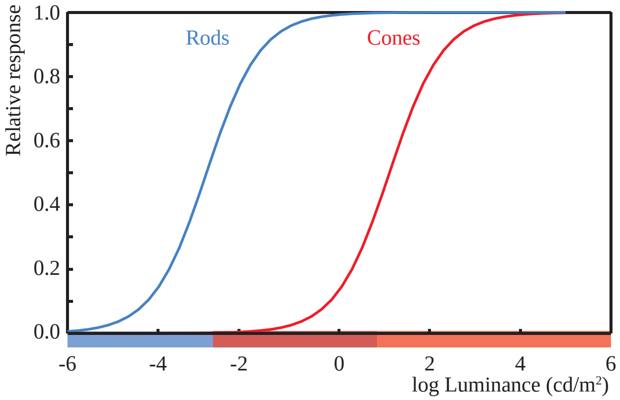


Figure 2.8. Photoreceptors respond nonlinearly to increasing luminance levels. The Naka-Rushton equation models this nonlinearity, producing S-shaped curves known as sigmoids.

the semi-saturation constant). The exponent n controls the slope of the function and is generally reported to be in the range of 0.7 to 1.0 [179]. This functional form is known as a sigmoid because it forms an S-shaped curve when plotted on log-linear axes. In this model adaptation can be accounted for by changing the value of σ. Moreover, this value could be used to model temporal adaptation.

Note that the response function of photoreceptors can be linked to natural image statistics. It has been shown that the joint probability density function of various components that make up images (illumination, reflectance, and textured regions) leads to a lightness scale[2] as function of illuminance that is remarkably similar to the sigmoidal response function of photoreceptors that are shown in Figure 2.8 [615]. More recently, lightness perception has been shown to directly correlate with this sigmoidal response function [11]. Finally, temporal adaptation is implicated in the enhancement of efficient gain control in the context of natural images [694], as well as redundancy reduction [35, 37, 767].

Photoreceptor Mosaic. Photoreceptors are located in the retina with a somewhat unexpected layout [8]. The S cones are sparse in the fovea and absent in the foveola. In the peripheral retina, S cones are packed at regular distances [798]. The L and M cones, on the other hand, are essentially randomly distributed over the retina, although the ratio between the two increases toward the periphery [700].

The ratio between the L and M cones varies significantly between individuals [512, 630], although this has only a minor effect on the perception of color [77, 576]. It has been postulated that this may be the result of plasticity in the

[2] The official definition of lightness is as follows: lightness is the attribute of visual sensation according to which the area in which the visual stimulus is presented appears to emit more or less light in proportion to that emitted by a similarly illuminated area perceived as the "white" stimulus.

human visual system, whereby later processing in essence compensates for the individual's precise photoreceptor mosaic [531]. Similar plasticity could help adjust an individual's visual system to long-term prevailing viewing conditions.

Given that at any given location in the retina there exists at most one cone, color vision is necessarily also a spatial process: different cone types need to be compared, and different cones are necessarily located in different positions. This places a limitation on the visual acuity of color vision.

2.3.3 Horizontal and Bipolar Cells

The retina consists of several layers of neural cells. The photoreceptors make up the first layer. As discussed above, their output is transmitted to a layer of bipolar cells, which subsequently convey the signal to the ganglion cells. These in turn carry the signal through the optic nerve to the LGN.

Bipolar cells take their input primarily from the photoreceptors and form the start of several pathways. Some bipolar cells respond to light against a darker background, forming the start of a so-called ON pathway. Others respond to dark spots against a lighter background, and are part of the OFF pathway. One could argue that the existence of separate ON and OFF pathways helps encode polarity. There are approximately two times more retinal OFF pathways, which encode negative contrasts, than ON pathways, which encode positive contrasts [7]. From the bipolar cells onwards, electrical signals are transported in the form of discrete spike trains. The firing rate of such spike trains encodes the quantity transmitted. One can think of spike trains as encoding only positive numbers. To also encode negative numbers (for instance, those corresponding to dark patches against a light background), a separate pathway needs to be constructed. Both L and M cone types are involved in both ON and OFF pathways, which can be either chromatic or achromatic [780]. The S cones are only involved in separate chromatic ON pathways [421].

In between the photoreceptors and the bipolar cells sits a further layer of cells that provide lateral connectivity. These cells are known as horizontal cells. These cells tend to have wide receptive fields, meaning that they connect to a large number of photoreceptors and bipolar cells over a significant spatial extent.

Due to the connectivity of horizontal cells, they are able to change the input to bipolar cells dependent on signals carried from photoreceptors some distance away. Horizontal cells provide inhibitory (opponent) signals to bipolar cells. This means that if their input from photoreceptors is large, they reduce the input signal available to the bipolar cells, and vice versa. Horizontal cells, therefore, form the mechanism by which bipolar cells are able to respond to light patches on a dark background or dark patches on a light background.

Thus, the opponent processing of the horizontal cells causes the receptive fields of bipolar cells to be of the center-surround variety. A reasonable way to think of such receptive fields is as if both center and surround are obtained by

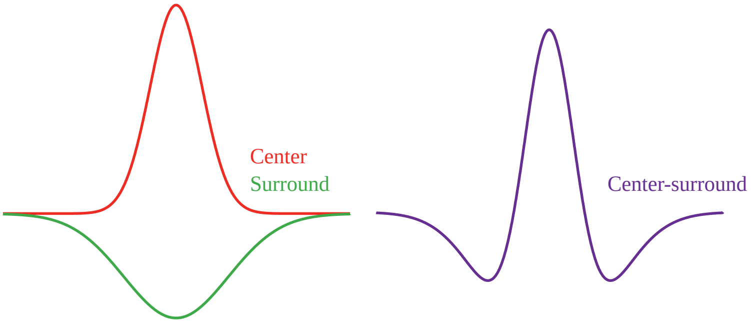


Figure 2.9. Cross-section of a center-surround receptive field. This receptive field was created by subtracting two Gaussian profiles from each other.

maximally responding to a Gaussian-blurred version of the image, although the size of the Gaussian filter for the surround is wider than the center. As shown in Figure 2.9, this leads to a receptive field that resembles the classical Mexican hat shape.

To visualize the responses carried by center-surround bipolar cells, we have blurred a grayscale image twice and subtracted the result, shown in Figure 2.10. The size of the receptive field was arbitrarily chosen and does not reflect actual processing in the visual system. Note that circularly symmetric center-surround receptive fields respond most strongly to circular features matched in size to the filter width. Further, this type of center-surround processing also responds to edges.

2.3.4 Amacrine and Ganglion Cells

Between the bipolar and ganglion cells lies a further layer of cells providing lateral connectivity. This layer of amacrine cells also provides connections between ON and OFF pathways as well as between rod and cone pathways. There exists a large number of different amacrine cells, each with different functionality. They appear to be involved in processes that allow the visual system to adapt to its environment.

As a result of the multitude of amacrine cell types, the receptive fields of ganglion cells are significantly varied [149]. They exist with center-surround receptive fields and may also show color opponency [129]. The latter happens, for instance, when the ganglion cell is stimulated by a pathway originating with an L cone, and it is inhibited by a pathway that started with an M cone. Thus, the ganglion cells effectively encode a very different color space than the cones that approximately correspond to red, green, and blue light. In fact, many ganglion cells in the fovea are color opponent, encoding in their signal the proportion of red against green, or alternatively the amount of yellow against blue.

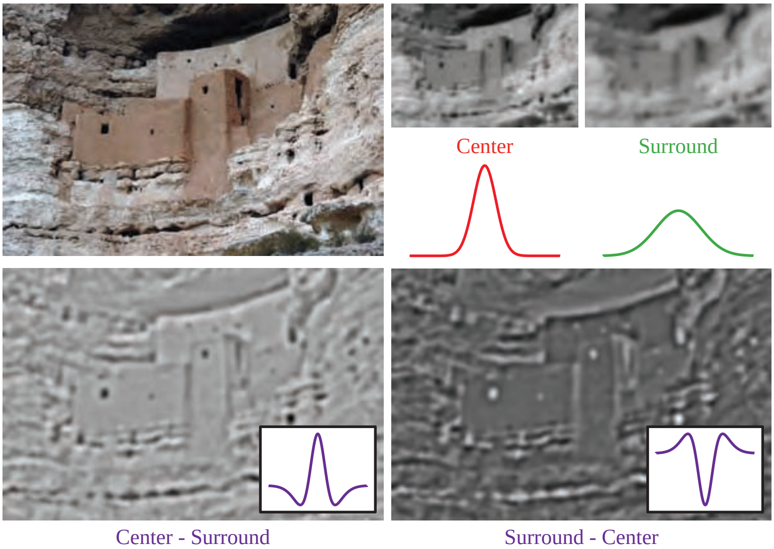


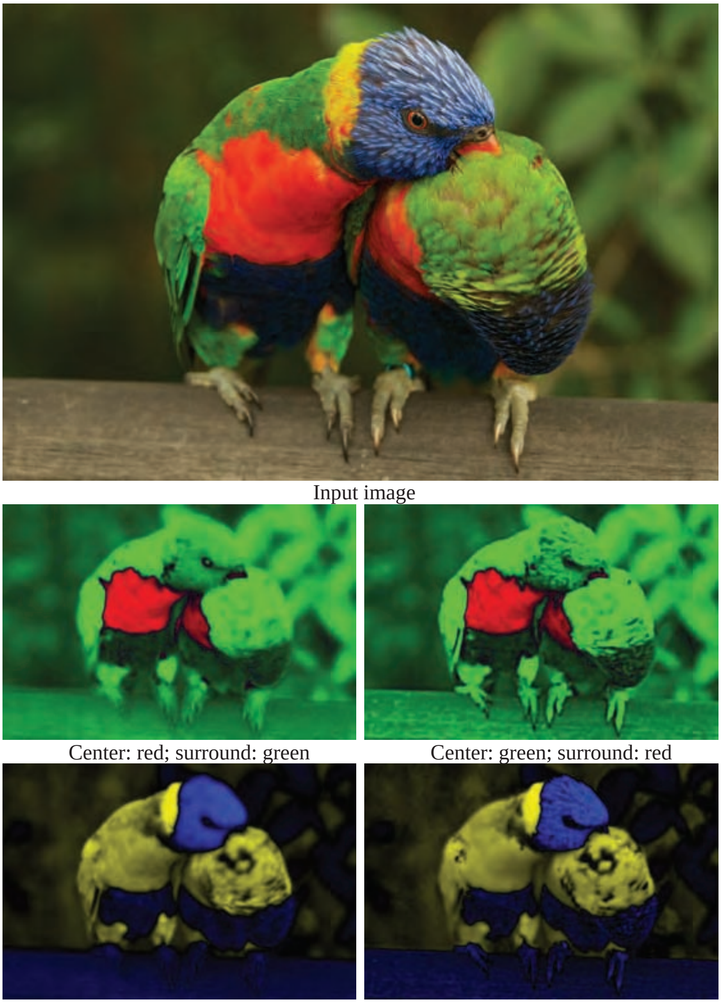
Figure 2.10. The receptive field images at the bottom were processed by subtracting two differently blurred achromatic versions of the input image. (Montezuma's Castle, Arizona, 2012)

Color opponency and spatial center-surround processing can occur simultaneously in a process called *double opponency*. Here, if the center of the receptive field is red, then the surround will be green and vice versa. Likewise, a yellow center will be flanked by a blue surround and vice versa. The result of such opponency is simulated in Figure 2.11.

Although the opponency is normally denoted as red-green and yellow-blue, note that these cardinal color directions do not completely correspond to perceptual experience. Later modules in the visual cortex are thought to perform further processing to arrive at the percepts of hues in terms of red, green, yellow, and blue [159, 160, 161]. Finally, there exist ganglion cells that are selective for direction and motion.

2.4 The Lateral Geniculate Nucleus and Cortical Processing

The lateral geniculate nucleus (LGN) is the dominant recipient of signals transmitted from the retina, and it is thought of as a relay station, effectively broadcasting the signal to modules in the visual cortex. The receptive fields of neurons in this area resemble those seen for ganglion cells, showing spatial and chromatic



Input image

Center: red; surround: green

Center: green; surround: red

Center: yellow; surround: blue

Center: blue; surround: yellow

Figure 2.11. Double opponency demonstrated on an image of a pair of lorikeets. Shown here are red-green, green-red, yellow-blue, and blue-yellow double opponent results. Note that these images are visualizations only; the results were created at an arbitrary spatial scale and they were computed in sRGB color space rather than LMS. The ratios between different cone types were approximately taken into account. (Jurong Bird Park, Singapore, 2012)

opponency. The LGN is organized in layers [173], termed *magnocellular*, *parvocellular*, and *koniocellular*.

The magnocellular layers carry an achromatic center-surround signal. The parvocellular layers mediate red-green color opponency, especially in the fovea [147, 148]. This layer does not appear to receive input from S cones. The koniocellular layers transmit several differentiated types of information, including blue-yellow color opponency [173, 327]. The organization of the layers in the LGN is retinotopic, in that the topology of cells reflects that of the retina. Thus, nearby features in an image projected onto the retina will be processed by nearby cells in the LGN. There is evidence that the ON and OFF pathways are augmented with lagged and non-lagged responses, i.e., some signals are transmitted more rapidly than others [175]. This would allow the calculation of temporal changes and could form the basis for motion detection. There is also some evidence that the spatial properties of some LGN cells change as a function of time (usually, the spatial organization changes in a manner that looks like motion in a single direction [162]).

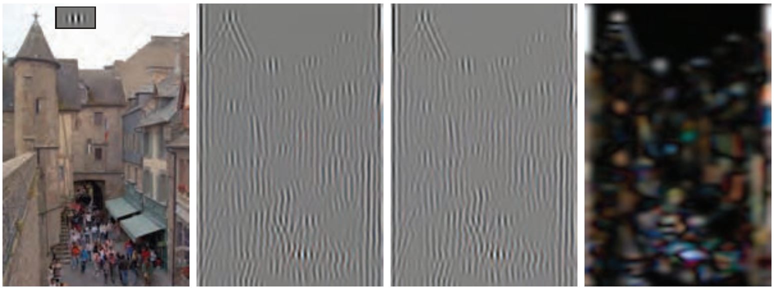
Most of the signal transmitted by the LGN arrives in the visual cortex in a module named V1, although some of the koniocellular layers transmit to the extrastriate cortex and bypass V1 altogether [626, 693, 815]. Nonetheless, V1 is the primary entry point for visual signals in the brain. From there, processing passes through more than 30 visually responsive cortical areas [200].

Area V1 receives signals from the lateral geniculate cortex, and consists of six different layers. Some of these layers are subdivided into sublamina [79, 476, 67]. There are further structures called blobs or patches. It is thought that the structure of V1 relates to its function. Many cells in this area are orientation-sensitive [355], a simulation of which is shown in Figure 2.12. This means that their receptive fields are formed by elongated structures rather than the circularly symmetric receptive fields in the retina and LGN [293, 319, 618]. Cells may also be responsive to motion as well as changes to scale [777] or to binocular signals [356, 477]. The retinotopic layout of the LGN is preserved in area V1 [199]. Area V1, like most, if not all, modules in the visual cortex has both feed-forward as well as feed-back connections with many other processing units [38].

Other cortical areas, such as V2, V3, V4, and MT, are implicated in a variety of different functions. Area MT, for instance, is known to be involved in the perception of motion, while V4 may take part in color perception.

2.5 Implications of Human Visual Processing

The substrate of visual processing is organized in a way that may not be considered obvious. A small selection of features known to exist in the human visual system were discussed in previous sections. For instance, we have presented ON and OFF pathways, color opponency, and center-surround processing, as well



Input image

Even symmetric filter (cosine)

Odd symmetric filter (sine)

Magnitude

Figure 2.12. A simulation of direction-selective receptive fields. An even symmetric Gabor patch (a windowed cosine, see inset, top left) was applied to the left image, yielding the second image. An odd symmetric Gabor filter was applied to the image as well. As an aside, the magnitude of the two receptive fields leads to a response shown on the right. (Mont St. Michel, France, 2007)

as combinations of these. As a result, the signal that the brain receives is not a straightforward image, but one that is already heavily processed. The organization and structure of the human visual system leads to several important characteristics [736], some of which are discussed in the following sections.

2.5.1 Visual Acuity

Visual acuity is defined as the resolving power of the human visual system [387]. The ability to detect and perceive fine detail is affected by various aspects of vision, including diffraction and aberrations in the ocular media, as well as photoreceptor density [696]. Note in particular that photoreceptor density varies across the retina, cones being packed most densely in the fovea. This means that visual acuity is highest in the fovea. Here, a single cone subtends approximately 28 minutes of arc, which means that this is also theoretically the highest resolving power that cones could achieve. This would lead to a resolving power of about 60 cycles per degree [104]. The rod system has lower visual acuity, pooling rod responses to increase light sensitivity.

However, visual acuity may be limited by diffraction in the ocular media due to the edge of the pupil. A point light source would project according to a point spread function (PSF) on the retina. Diffraction would cause this point spread function to consist of concentric rings of light, i.e., an Airy disk pattern. To be able to separate a point source from a second point source some distance away, the point spread functions of both light sources should not overlap. Two point lights are considered sufficiently separated if the center of one PSF lies on the first trough of the second PSF. This is formalized by Raleigh's criterion, which links wavelength of light λ and the diameter of the pupil d:

$$a = 1.22\frac{\lambda}{d} \tag{2.2}$$

Here, a is the angular radius of the first ring of the Airy disk pattern.

Further factors affecting visual acuity include refractive error, which occurs if light refracted through the ocular media is not focused sharply on the retina. Visual acuity is also dependent on the amount of illumination present in the scene [617] and the corresponding state of adaptation of the eye.

Under certain circumstances, visual acuity may be significantly higher than could be expected from the limitations imposed by diffraction and cone spacing [797]. This is generally known as *hyperacuity*. An example of when this occurs is during assessment of the alignment of two line segments—for instance, when reading a caliper. This can be done with an accuracy of about five to ten times higher than the normal resolving power of the human visual system. In essence, a set of photoreceptors are involved in localizing the line segments. Assuming the cones are of the L and M variety, their placement is randomized, which may help in such localization tasks.

In some sense, hyperacuity achieves for human vision what superresolution does in image processing. The randomization of cones helps with anti-aliasing. One implication for the study of natural image statistics may be that image resolution needs to be chosen high enough to anticipate the perception of edges and their relative alignment.

2.5.2 Temporal Resolution

To be able to detect motion, the eye must be able to detect rates of change in the image that falls on the retina. Some amount of time is needed to transduce light and propagate the resulting signal to the brain. This poses a limit on how fast or slow a signal can change and still be perceived as fluent motion [386]. For instance, if the rate of change is too slow, then the sensation of fluent motion is lost, and the stimulus is seen as flicker. At photopic light levels, the eye is most sensitive to flicker at frequencies of around 15 to 20 Hz [307].

To detect two flashes of light separated in time, rods require an integration time of 100 ms, while cones require 10 to 15 ms. Thus, the rod system is slower to help improve light sensitivity.

2.5.3 Contrast

Contrast involves the relative values between points or regions in an image and can be measured in many different ways. For a region in an image, it would, for instance, be possible to compute root mean square contrast, which is the standard deviation of a set of pixels:

$$C_{\mathrm{RMS}} = \sqrt{\frac{\sum_{n \quad N} (L_n - \mu_L)^2}{N}} \tag{2.3}$$

It would also be possible to normalize this measure of contrast by the mean μ_L:

$$C_{\mathrm{norm}} = \frac{C_{\mathrm{RMS}}}{\mu_L} \tag{2.4}$$

However, it can be argued that for complex images a more involved measure would be appropriate—for instance, by computing the contrast on band-pass filtered images [567].

If an image/stimulus is a small feature on a uniform background, then the background luminance L_{b} will be close to the average luminance. In that case, Weber contrast is appropriate. It is defined as:

$$C_{\mathrm{Weber}} = \frac{L - L_{\mathrm{b}}}{L_{\mathrm{b}}} \tag{2.5}$$

For sinusoidal gratings, a different contrast measure is commonly used, a measure known as Michelson contrast. This measures the differences between the peaks and troughs of the grating:

$$C_{\mathrm{Michelson}} = \frac{L_{\mathrm{max}} - L_{\mathrm{min}}}{L_{\mathrm{max}} + L_{\mathrm{min}}} \tag{2.6}$$

Humans are sensitive to contrast in images (whether natural or psychophysical stimuli). This sensitivity depends on the frequency at which the contrast occurs, as well overall light levels. More recently, it has been shown that contrast sensitivity in natural scenes also depends on the distribution of local edges [54]. It would be possible to plot threshold contrast against frequency, which leads to the well-known contrast sensitivity function; this function is plotted for different light levels in Figure 2.13 [106]. Note that for low light levels, the response is low-pass, whereas at high light levels the response is band-pass.

2.5.4 Color Processing

Under photopic lighting conditions, the three types of cones are active, each integrating over the visible spectrum over a slightly different range and with different peak sensitivities. Thus, human vision is trichromatic under these conditions [814, 326]. As a result, in each region of the fovea where there are multiple cone types active, color vision can emerge. However, as noted, different spectra may lead to the same three integrations, which means that the output of the cones will be identical and these different spectra cannot be differentiated. These spectra are said to be *metameric*.

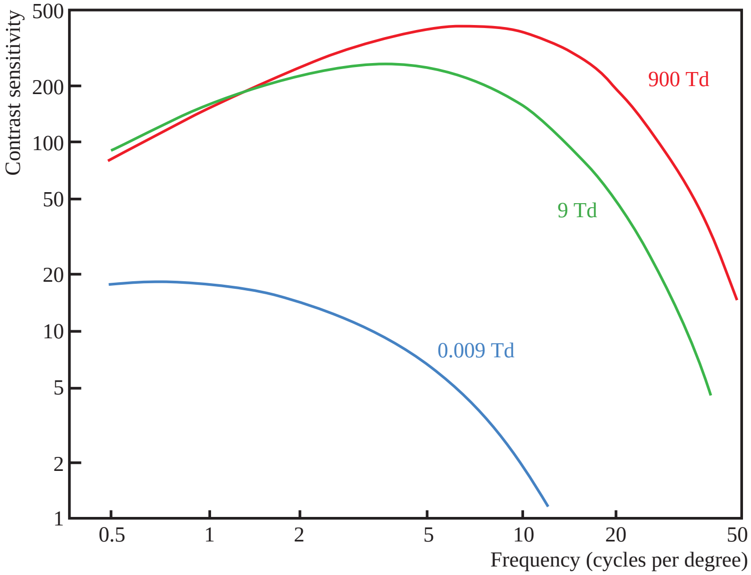


Figure 2.13. Contrast sensitivity functions for different light levels [386, 433].

Trichromacy has a second implication, which is that almost any color can be matched by mixing three other colors. This can be achieved, for instance, by matching colors on a bipartite field, which is in essence a disk partitioned into two halves. One of the halves displays the color to be matched, for instance, by means of projection. The other half is illuminated by three projectors, each projecting a given color by an amount that is user-controllable.

This method can also be used to infer the luminance of a colored patch. Here, the colored patch is held fixed while the other half displays achromatic white in an amount controlled by the user. The amount of light that is necessary to make the border between the left and right half of the bipartite field minimally distinct then represents the luminance of the colored patch [71].

When carrying out such experimentation on a display, significantly higher accuracy can be achieved by replacing the bipartite field with a two-tone image of a human face [400]. The accuracy gain stems from two characteristics of human vision. First, humans appear to make the assumption that objects, including faces, are predominantly illuminated from above (see Figure 2.14). Second, humans are particularly adept at recognizing human faces, but only if illuminated from above.

Color opponent theory predicts that colors are perceived along three different dimensions, namely, light-dark, red-green, and yellow-blue [328]. This is consistent with perceptual experience [362] and after-images, as well as what is known about retinal processing (see Section 2.3). There is also a strong link between color opponent processing in the retina and statistical attributes of natural scenes, as discussed in Section 10.3.

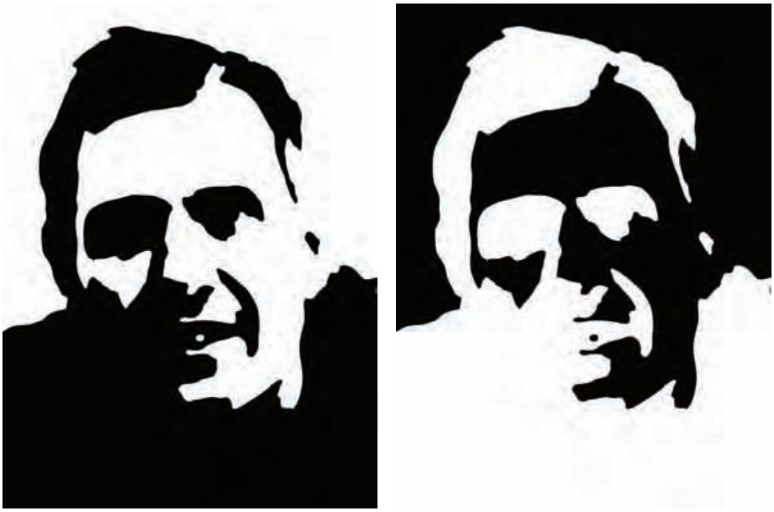


Figure 2.14. Two-tone images of human faces are easily recognized if lit from above (left). The inverted version of this image (right) is not easily recognized. Images such as these are therefore good candidates for accurate color matching experiments [400].

2.5.5 Visual Illusions

Human vision solves a complicated problem, which is to makes sense of the world around us. One of complications stems from the fact that scene understanding and navigation requires a mental image of a three-dimensional world that must be derived from a pair of two-dimensional retinal images [413]. That this is a complicated problem is well known and is, for instance, exemplified by the difficulty of figuring out shape from shading in computer vision–related applications [819, 589].

It appears that human vision makes many assumptions that help in image understanding. Figure 2.14 already showed an example, namely, that face recognition is directly facilitated by ensuring that light comes from above. When this assumption is broken, face recognition becomes significantly more difficult. Similarly, a human face is more easily recognized when it is positioned right-side up.

The human visual system appears to assume that the scene being observed does not contain anything special, i.e., there are no accidental viewpoints, light comes from above and, possibly above all, the scene is in some sense natural. When these expectations are broken, then the scene may be viewed as having artefacts or, alternatively, a visual illusion may arise [283, 285].

In computer graphics, for example, achieving visual realism has been difficult for many years [608]. Only recently have modeling and rendering systems

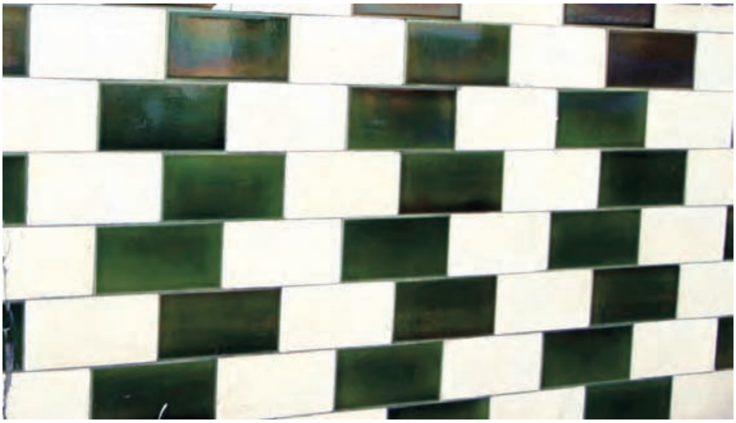


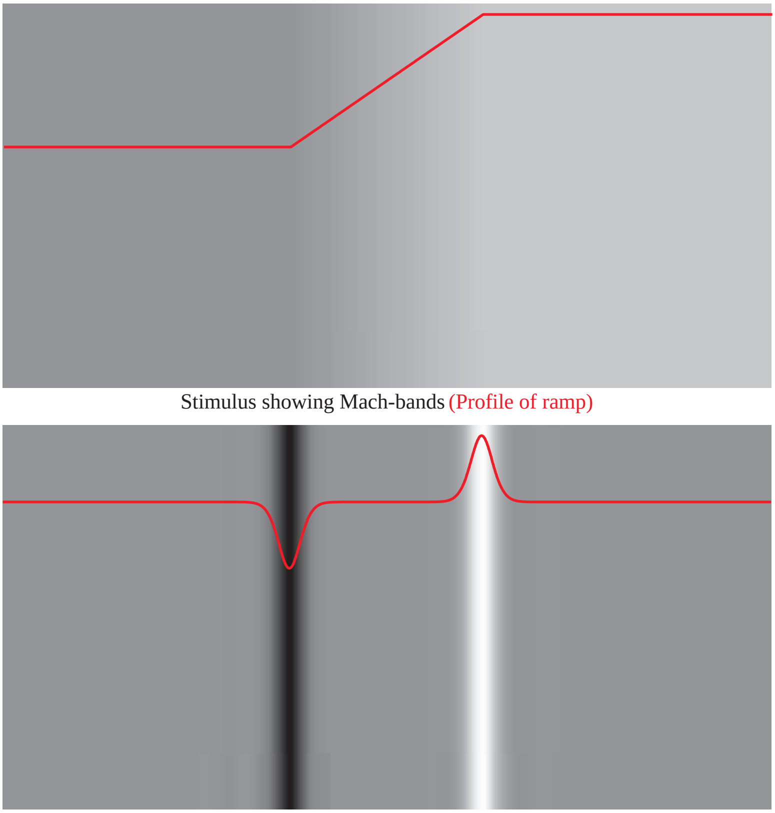
Figure 2.15. Cafe wall illusion. (Bristol, UK, 2006)

become available that add enough dirt, grime, and general detail that rendered images have become difficult to differentiate from photographs. Leaving out such detail, or omitting certain light paths because they are computationally costly to render, may lead to images that are generally scene as plastic or fake.

Visual illusions occur when expectations are broken in specific ways. It can be argued that human vision is betting on what is likely to be true about the world given the evidence available [285]. Thus, the study of visual illusions may lead to insights into human visual processing [284]. There are too many visual illusions to list in this book; moreover, many of them involve very simple line drawings. Illusions that can be reproduced in photographs or can be demonstrated as shaded surfaces are less numerous.

The most common illusion that can be seen in a photograph is perhaps the cafe wall illusion, discovered in Bristol by Richard Gregory [286] (see Figure 2.15). Here, a set of rectangular tiles are positioned such that the horizontal grout lines appear to be oriented differently. Due to the camera angle, however, these lines should (and do!) all converge at the same point. In three dimensions, these lines are parallel.

The center-surround organization found in the retina is thought to give rise to several interesting phenomena. One consequence of center-surround processing is the occurrence of Mach-bands [119, 483]. In essence, wherever a uniform luminance gradient abruptly changes there may be the perception of either under- or over-shoot. An example is shown in Figure 2.16 (top) where a linear ramp abuts two uniform areas. At the transition point there is C^0 continuity, but a discontinuity in gradient. Processing such an image with a center-surround filter shows the cause for the perceived under- and over-shoot as seen in Figure 2.16 (bottom).



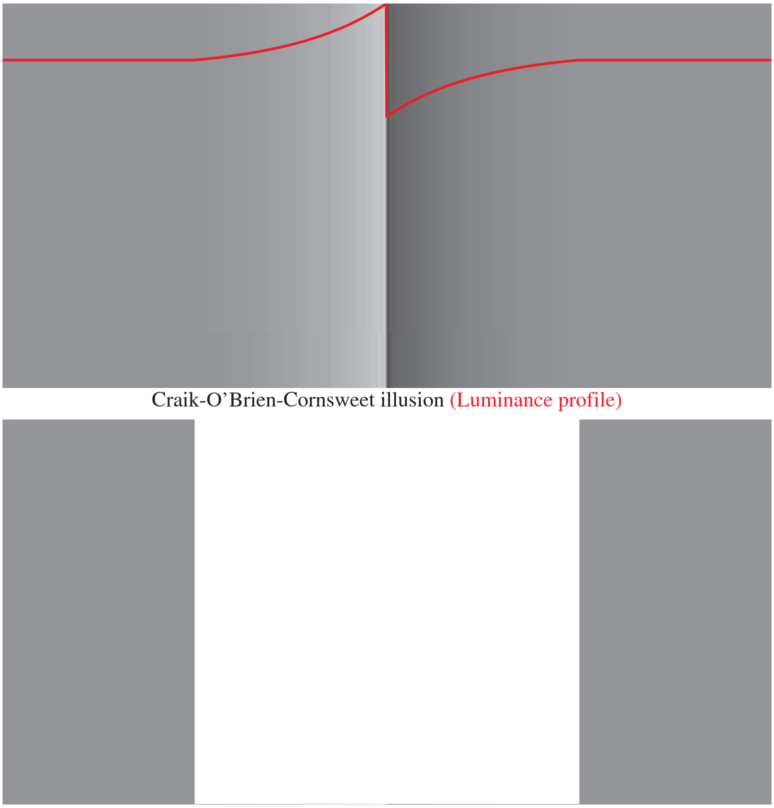
Stimulus showing Mach-bands (Profile of ramp)

Exaggerated response to ramp (Profile of response)

Figure 2.16. C^0 continuity may lead to Mach-bands, seen as under- or over-shoot where gradients are discontinuous.

A further consequence of center-surround processing is the Craik-O'Brien-Cornsweet illusion, shown in Figure 2.17. This figure contains a step edge flanked by ramps that smoothly vary their gradients [541, 674, 133, 401]. As a result, a step edge is seen, suggesting that the left part of the image is lighter than the right part. The shape of the ramps is chosen such that they generate little to no response from center-surround mechanisms in the retina (see Figure 2.18). As the step edge in the center does evoke a response from these mechanisms, the resulting percept is similar as that which would be generated by an actual step edge flanked by uniform regions.

A consequence of this type of retinal processing is that the visual cortex receives a signal that highly emphasizes edges. Perceiving the structure of a scene in between edges may be based on a process called filling-in [197, 291, 423, 771],



Craik-O'Brien-Cornsweet illusion (Luminance profile)

Masking the middle part reveals that the left and right parts have the same luminance

Figure 2.17. The Craik-O'Brien-Cornsweet illusion. In the top panel, the left and right quarter of the image will appear to have a different gray level. Masking the center half of the image reveals that the actual luminance levels are identical.

suggesting that the brain completes its perception of a scene predominantly on the basis of edges. One may therefore expect edges and their statistics to be of crucial importance in human vision. Edge statistics are discussed further in Section 5.3.3.

An alternative suggestion is that filling-in is aided by other statistical regularities in natural images, specifically the nature of the power spectrum [150], which is known to give a relation between power and frequency of $1/f^2$. Chapter 6 is devoted to this particular natural image statistic, which has proven to be one of the most ubiquitous.

That the Craik-O'Brien-Cornsweet illusion may have yet a different and higher level explanation can be seen in Figure 2.19 [593]. Here, the edge of the image is changed to suggest curvature in depth, and this appears to enhance

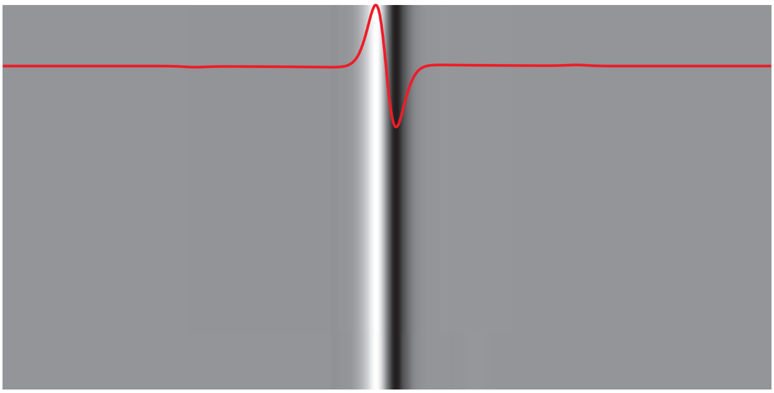


Figure 2.18. The Craik-O'Brien-Cornsweet illusion passed through a center-surround filter. Note that there is virtually no response for most of the image, except near the edge.

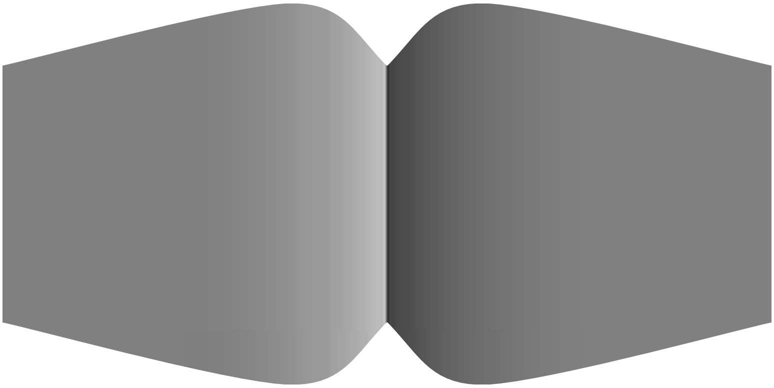
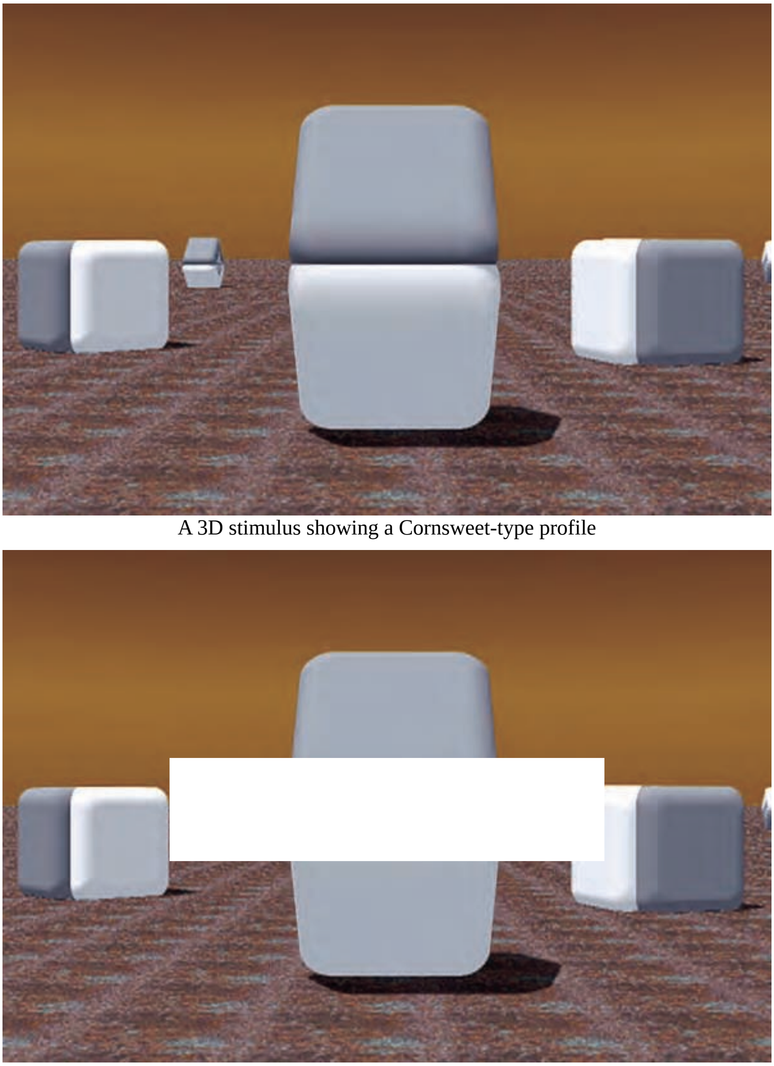


Figure 2.19. The Craik-O'Brien-Cornsweet illusion becomes stronger if it is placed in context. Here, for instance, the shape of the image suggests a 3D surface, which enhances the effect [593].

the effect. It may be that this simple change to the figure increases the probability that the luminance profile is due to illumination, and that the human visual system is able to detect this. This effect can be enhanced further by embedding it into a scene that affords further 3D cues, as demonstrated in Figure 2.20. Incidentally, the strength of this illusion also relates to the assumption that light comes from above. Thus, while center-surround processing may be involved, it is by no means the only component of human visual processing that contributes to this variant of the illusion.



A 3D stimulus showing a Cornsweet-type profile

Masking the center part shows that top and bottom parts have identical luminance

Figure 2.20. The Craik-O'Brien-Cornsweet illusion becomes stronger if it is placed in context. Here, for instance, the shape of the image suggests a 3D surface, which enhances the effect [593]. (Image used by permission from Dale Purves.)

Chapter 3

Image Collection and Calibration

To analyze the world around us, we first need to capture a representation of it. Although information in the real world spans many different modalities, current technologies can only capture a subset of that wealth of information. A photograph of a natural scene will only show a small portion of that scene, compressed to two dimensions and capturing only a subset of the spectral information available. Image statistics, consequently, capture not only regularities in nature but also inherently those of the image acquisition process. To that end, before proceeding to analyze images using the statistical tools that we will present in the remainder of this book, we will explore different capture possibilities and understand their respective limitations.

The statistical analysis of images typically involves the collection of large ensembles. Various image data sets are already available and some existing sets will be discussed in Section 3.3, but in order to create new ensembles, several points need to be considered. One might need to create new ensembles if the intended application involves a specific class of images, such as high dynamic range imagery, or if the study is to understand human (or animal) vision under specific viewing conditions.

Advances in camera technology have made the capture of higher resolution, higher dynamic range, or even hyperspectral imagery possible. Each such type of image presents its own advantages in terms of information gain but also its own specific challenges regarding capture, calibration, and analysis.

Issues relating to the collection of accurate image ensembles of different types of imagery will be discussed in Section 3.1. Before analyzing the collected images, it is prudent to remove artifacts and irregularities that are due to the camera, lens, or the settings used to capture the scenes. These can include noise, distortions due to imperfections of the lens, or variations from changes in illumination, or even weather conditions. The calibration process and preprocessing steps necessary to account for such irregularities will be covered in Section 3.2. Finally, existing image collections suitable for image statistics applications are discussed in Section 3.3.

3.1 Image Capture

In most digital imaging applications, image data is typically captured, processed, and displayed as a 2D pixel grid, where each pixel is encoded as a triplet of 8-bit numbers. This representation is the most common form of digital imagery and the one we will use for the remainder of this book, unless otherwise specified.

Although adequate for most applications, traditional images have a number of limitations in terms of resolution, color gamut, spectral information, and dynamic range. Additionally, the type of intended application will pose further restrictions on the required number, resolution, and type of images. Finally, the choice of capture device and image modality (e.g., multi-spectral, depth, RGB) will profoundly affect the regularities and effects that can be measured within an ensemble. As such, a number of issues need to be considered before capturing images for an ensemble. The following largely depend on the aims of the study:

- **Type of imagery.** For most purposes, traditional imagery as discussed above is typically sufficient. To fully capture the visible information in real environments, however, additional imaging modalities are necessary. For instance, high dynamic range (HDR) images can capture a wider range of luminance values, making them more appropriate for scenes of more extreme lighting (such as indoor scenes or night scenes where artificial light sources are much brighter than the rest of the scene), while panoramas or wider angle images can capture a larger part of the field of view—or even the full scene in the case of spherical panoramas. Some examples will be discussed in more detail in the following sections, although a comprehensive account of all image modalities is beyond the scope of this book.

- **Number of images.** Generally, the accuracy and confidence of statistical findings increases with larger sample sizes [138]. However, in the case of image statistics, the trade-off between number of images and data collection time often needs to be resolved. The number of images appropriate for a particular study largely depends on its aims, with studies varying from a few hundred [825, 640, 654] to a few thousand images [316, 352].

- **Camera.** The selection of capture device will of course largely depend on the image modality of interest. For standard digital imaging, although most cameras currently available on the market offer high-resolution acquisition, a several trade-offs may need to be considered, including noise, the quality of lens optics, and controllability of the camera settings. These issues will be discussed further in Section 3.2.

3.1.1 Photographer and Camera Bias

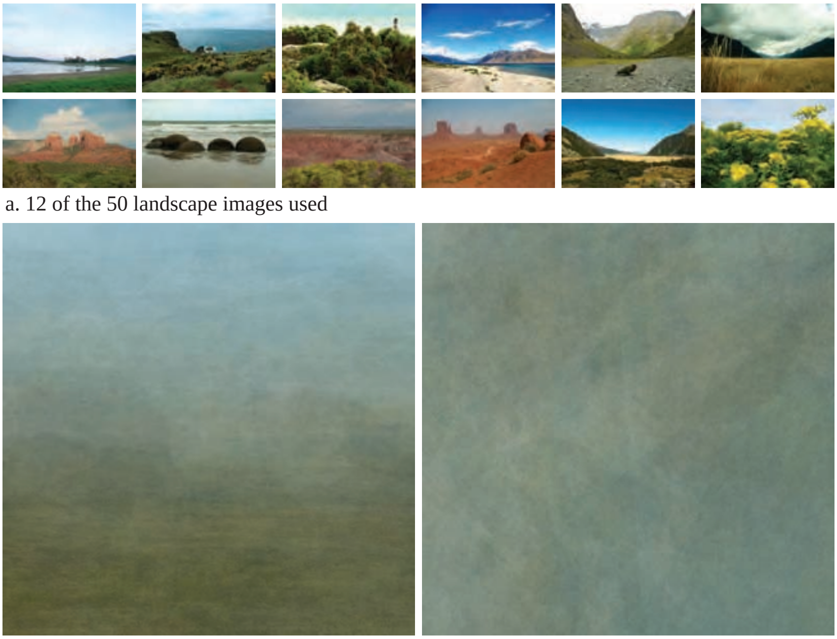
When collecting images for statistical analysis, the camera settings, lens characteristics, and composition of the images can all introduce bias. Depending on the application, some forms of bias may be desirable or even necessary. For instance, in studies linking statistical regularities of natural scenes to properties of the visual system, the field of view and orientation of images may be chosen such that they approximate the image formed in the retina [751]. In such a scenario, it is likely that most images will be oriented such that the horizon is in fact horizontal and will contain both land and sky. The inclusion or exclusion of such forms of bias is likely to influence the statistical regularities that arise from an image collection.

Figure 3.1 shows the effect of orientation bias for a simple scenario. For this example, 50 landscape images were considered. A subset of these images is shown in Figure 3.1a. A 400×400 square was cropped from the center of each image and averaged to produce the result in Figure 3.1b. As these images were captured with aesthetic composition in mind, the horizon was aligned with the horizontal axis with the image, resulting in the horizontal structure still visible in Figure 3.1b. In contrast, Figure 3.1c was produced by applying a random rotation to each image, thereby removing the orientation bias of the photographer and the resulting structure in the average image.

Although Figure 3.1c could be seen as a less biased representation of the "average landscape," either version can be useful depending on the application. Humans tend to see a horizontally aligned world, and therefore Figure 3.1b is a closer representation of the average natural scene according to the human visual system. Other forms of bias, however, such as lens distortions, fixed pattern noise, or a limited type of images (e.g., only including images of one location when analyzing natural images in general) are not inherent to the application of subject of the study but are introduced by the researcher. Much like in experimental design [138], such forms of bias should be avoided or corrected for where possible. These are further discussed in the context of image correction and calibration in Section 3.2.

3.1.2 High Dynamic Range Imaging

As we saw in the previous chapter, the human visual system employs a number of processes in order to adapt to the wide range of illumination present in nature. Standard imaging techniques can only capture and process a subset of that range: whereas the visual system can simultaneously adapt to around four orders of magnitude [431], 8-bit imagery can only represent 256 distinct levels per channel.



a. 12 of the 50 landscape images used

b. Average of 50 landscape images horizontally oriented

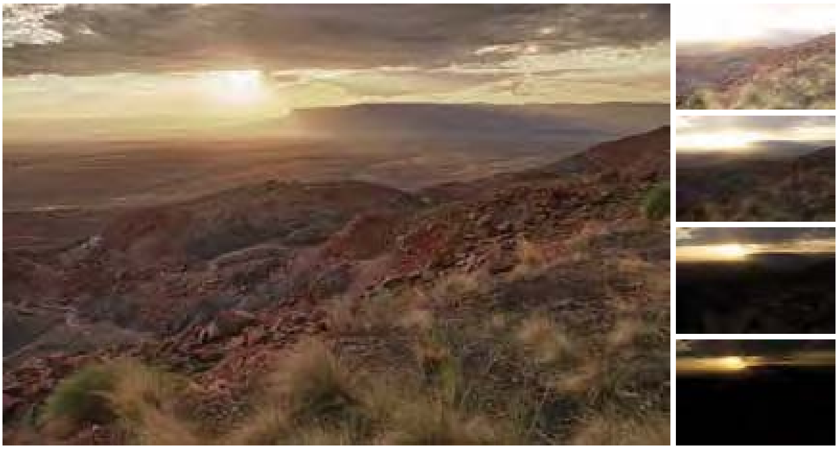
c. Average of 50 landscape images randomly oriented

Figure 3.1. A set of 50 landscape photographs (a) were analyzed showing the effect of orientation bias. A 400 × 400 pixel square was cropped from the center of each image and averaged. In (b) the images were horizontally aligned, while in (c) a random rotation was first applied to each image, removing orientation bias.

To counter these limitations, hardware and software solutions have been developed that allow the capture, processing, and display of images with a wider dynamic range of intensities. These solutions are known as high dynamic range (HDR) imaging [609]. In contrast to standard imagery, HDR data is (at least conceptually) represented using floating point numbers, allowing for a virtually unlimited range to be represented, with practically no quantization.

Capture. HDR scenes can be captured in a number of ways. The most commonly used approach involves capturing a series of differently exposed versions of the scene, known as exposure bracketing, that are then merged into the HDR image [494, 164, 511]. An example is shown in Figure 3.2. Several modern digital SLR cameras offer automated settings for bracketing (*"Auto-exposure Bracketing"*), greatly simplifying the process. The number of brackets offered typically ranges from two or three to a maximum of nine in higher range models of each brand. These are then merged into a single HDR image.

Recently, cameras capable of directly capturing HDR still content have become available. Specialized cameras exist that are capable of capturing 360°



a. HDR image (manually tonemapped for print) b. Exposures

Figure 3.2. A series of differently exposed versions of the same scene are merged into a single high dynamic range image. To display or print an HDR image, the wide dynamic range is compressed to that of the display or printed medium through tonemapping. The result preserves more details than any single exposure can. (Near Page, Arizona, USA, 2012)

spherical images. Some examples are the SpheroCamHDR by Spheron VR, which uses a single-line CCD and offers up to 26 f-stops of dynamic range, and the Civetta by Weiss AG, which uses a full frame sensor and a fish-eye lens to capture up to 30 f-stops of dynamic range. For capturing HDR video content, a camera was recently developed, which simultaneously captures differently exposed versions of the same scene [742]. Please refer to Christian Bloch's book [66] for an up-to-date account of HDR camera developments.

Merging. The aim of the merging process is to reconstruct a linear, floating point representation of the illumination in the scene from the differently exposed images, which poses a number of issues. If images are not captured in RAW format, the camera is likely to apply a nonlinearity (usually close to the sRGB curve), which needs to be corrected. Additionally, each exposure is likely to contain some over- and under-exposed pixels, and movement of elements in the scene or of the camera will lead to mismatches between the exposures, which in turn can result in ghosting artifacts.

A sequence of differently exposed images can be used directly to recover the response curve. A commonly used method is proposed by Debevec and Malik [164]. By exploiting a property of both physical film and digital sensors known as *reciprocity* (meaning that halving the irradiance hitting the sensor E and simultaneously doubling the exposure time Δt will result in the same pixel values p), a linear system is derived. The solution to that yields the inverse function g^{-1} to the camera response curve. Additional methods for recovering the response curve of a camera and for converting image data to accurate radiometric quantities are discussed in Section 3.2.

After the response curve g is obtained, the image irradiance E_n for a given pixel of an exposure n can be computed given the exposure time Δt_n and the value of that pixel p_n as follows:

$$E_n = \frac{g^{-1}(p_n)}{\Delta t_n} \tag{3.1}$$

If a sufficient number of exposures is collected, each part of the scene should be correctly exposed in at least one of the images. The irradiance value for a given pixel p in the HDR image can then be computed as a weighted average of the corresponding pixels of each exposure.

Several weighting functions have been proposed in the literature. Mann and Picard [494] merge the exposures using *certainty functions* computed from the derivatives of the response curves for each differently exposed image. A simpler approach is used in [164] where a hat function suppressing values that are likely to be under- or over-exposed is applied. Finally, inspired by signal theory, Mitsunaga and Nayar [511] propose a weighting function that takes into account the signal-to-noise ratio. Despite their differences, any of these approaches will yield satisfactory results. More recently, a solution for merging image at larger exposure intervals was developed, allowing for scenes with extreme dynamic ranges to be captured with relatively few exposures [742].

Many software packages are capable of merging exposures into an HDR image [66]. The Adobe Photoshop package offers a "Merge to HDR" function that allows for manual control over the image alignment. Other packages include HDRSoft's Photomatix [320] and HDRShop [163], both of which offer batch processing functionality. Finally, a free option that is only available for Apple computers is Greg Ward's Photosphere [20] (see anyhere.com), which is also an excellent viewer for HDR image collections. For more information regarding available HDR software and creative solutions, refer to [66] or to http:\hdrlabs.com.

Although many algorithms exist currently for correcting ghosting artifacts due to movement in the scene (many included in the software discussed above), it cannot be guaranteed that they will not affect the statistical properties of the resulting images [609]. As such, unless the effect of such algorithms is of interest, we recommend using images with no such artifacts for the purposes of statistical analysis. Thus, to minimize ghosting and alignment artifacts, scenes with minimal movement should be chosen and the camera should ideally be placed on a tripod.

3.1.3 Field of View

Different lenses and camera setups can capture different portions of the scene. Any given scene can be thought of as projected onto a sphere with the observer

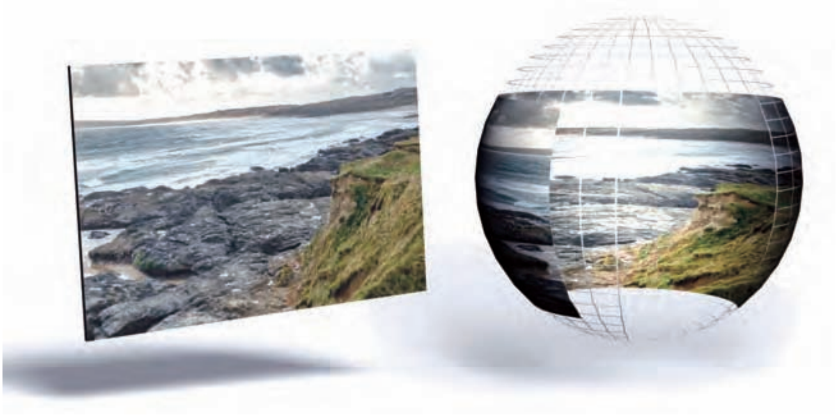


Figure 3.3. A given scene can be represented by a (hemi)sphere, resulting in a 360-degree field of view. Any image from that scene forms a slice on the sphere. To capture the complete scene, several overlapping slices are necessary.

or imaging device placed at its center, while a given image can be seen as a slice of that sphere (Figure 3.3). The angle subtended by the image is the *field of view* (FoV) of the lens. A fish-eye lens can capture almost 180 degrees of the hemisphere and therefore has a wide FoV, while a zoom lens will only be able to capture a smaller slice and thus has a narrower FoV.

To study entire scenes or to remove bias due to orientation or composition, the complete sphere representing a scene can be captured in the form of a *panorama*. Panoramic images are increasingly finding applications within computer graphics applications. Such images are often combined with the HDR techniques described earlier to capture accurate scene illumination for relighting or compositing [609, 66], or they are used for creating immersive virtual reality environments [118, 720].

With the exception of a few specialized (and considerably expensive) camera systems such as the SpheroCamHDR by SpheronVR[1] or the Ladybug series by PointGray,[2] commercially available lenses can capture at most 180 degrees of visual angle. To capture the complete sphere for a scene, an FoV of 360 degrees horizontally and 180 degrees vertically is necessary. This is typically achieved by capturing and combining (stitching) a series of images, with each image capturing a partially overlapping slice of the hemisphere. Figure 3.4 shows a 2D illustration of this process.

As with HDR capture, to be able to combine a series of images into a panorama, the overlapping parts of consecutive images need to align correctly. In addition, the exposure settings of the camera need to remain the same for the whole scene.

[1] http://www.spheron.com/en/spheron-cgi/products/spherocam-hdr.html

[2] http://www.ptgrey.com/products/ladybug2/ladybug2_360_video_camera.asp

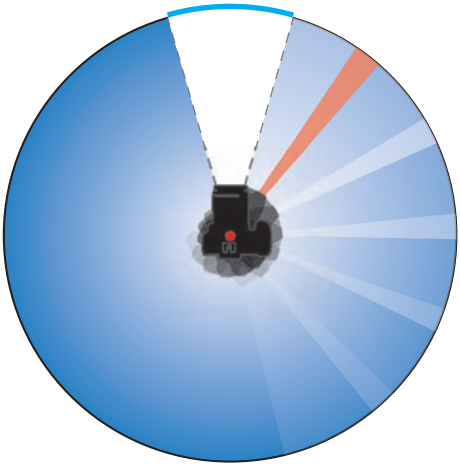


Figure 3.4. To capture a panorama, a sequence of overlapping images is taken by rotating the camera on the focal point (shown as a red dot). To combine the images into the panorama, stitching algorithms find matching features between images and as such require some amount of overlap between subsequent shots (the red slice in the image).

Note that especially in outdoor photography, it is likely that a single exposure setting will not be enough to capture the full dynamic range of the scene both towards and against the sun. To capture such scenes, techniques from the previous section can be combined with the ones discussed here. Namely, for each panoramic slice, multiple exposures can be shot and combined [66].

Although with some luck and a very stable hand, it is possible to shoot artifact-free panoramas handheld; to ensure that images can be correctly aligned and stitched, using a tripod is recommended. Ideally, the nodal point of the tripod head should coincide with the focal plane of the camera, usually indicated with a symbol similar to ⊖. Such tripod heads are known as panorama heads and can be either manually constructed[3] or commercially obtained.[4]

Once an image sequence covering the full hemisphere has been captured, stitching algorithms can be employed to combine them into a single panoramic image [718, 80]. Interestingly, some of these algorithms rely on statistical regularities themselves [456]; one example will be discussed in Chapter 5. Stitching algorithms first detect features on each of the images and, using feature matching techniques, align correspondences. Once the images are aligned, distortions and variations in exposure or color may be removed and finally the images are blended

[3] http://www.stockholmviews.com/diyphotogear/pano_head.html

[4] http://wiki.panotools.org/Heads

into a single image. Currently, stitching functionality is offered both in general image editing packages such as Adobe Photoshop and dedicated packages such as the open-source Hugin.[5]

Statistical Analysis of Panoramas. Despite the obvious benefit of depicting the whole scene rather than a portion of it, panoramic images pose additional challenges in terms of analysis, especially in the context of image statistics. Inevitably, to store, analyze, and view panoramas, a mapping between the spherical scene and a 2D representation is necessary, a problem similar to that faced by map designers. Several different projections have been employed, both in mapping and in panoramic imaging, to achieve such a mapping [111, 643].

To analyze panoramic images, the 2D rectangular mapping of the 3D scene can be directly considered. However, distortions due to the projection chosen may affect the statistical regularities that arise. For instance, straight lines may no longer appear straight. Alternatively, techniques that operate on spherical surfaces may be employed, such as spherical harmonics [335, 484] or spherical wavelets [657]. For a more detailed discussion on the statistical analysis of spherical panoramas, please refer to the work of Dror et al. [182, 185].

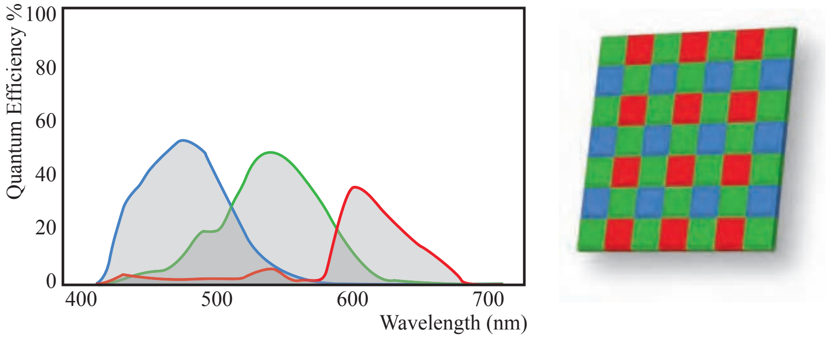
3.1.4 Multispectral and Hyperspectral Imaging

Image modalities discussed so far are limited in their representation of the visible spectrum. HDR technologies allow for a wider range of intensities to be captured, but color is typically restricted to a three-channel encoding. Light in nature is defined along a continuous spectrum of wavelengths and the human visual system is capable of perceiving light between 400 and 700 nm [774] (see Figure 3.6). As we have briefly discussed in the previous chapter, the human visual system can obtain a sufficiently accurate representation of the visible spectrum from three types of photoreceptors, each with a wide response centered at different wavelengths. Another way to think about it would be that the visual system "samples" the spectrum using three overlapping sample ranges.

Similarly, typical camera sensors filter light using three filters with different sensitivities. The filters are placed over the sensor pixels in a specific pattern [526]. A commonly used pattern, known as *Bayer* pattern [40], and the sensitivities of an example camera are shown in Figure 3.5. Although the sensitivities of cameras are different to those of the photoreceptors in the visual system, their purpose is similar [683]. They aim to capture a representation of visible light with the least amount of data possible.

This representation is sufficient for most imaging scenarios. However, some applications, such as studies of animal vision [500, 143], might require either a denser representation of the visible spectrum or information outside the visible (by humans) spectrum. Similarly, to accurately study the responses of human photoreceptors, the full spectrum may need to be captured [640].

[5] http://hugin.sourceforge.net/



a. Spectral sensitivities of Nikon D300 camera

b. Typical Bayern pattern

Figure 3.5. The spectral sensitivities of a Nikon D300 camera are shown on the left (adapted from [683]). Although very different quantitatively to the sensitivities of the three types of photoreceptors in the visual system, conceptually they serve a similar purpose: they sample the light spectrum using three overlapping responses. On the right, a typical filter array is shown. Such an array is placed on the camera sensor to filter incoming light.

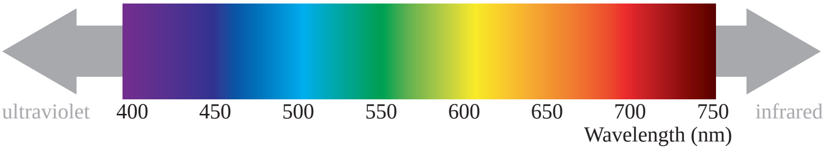


Figure 3.6. An illustration of the visible spectrum.

Three main solutions are available for capturing spectral data. Tunable[6] or fixed narrow-band filters are often used to acquire spectral data in the context of image statistics, each filter capturing a narrow slice of the spectrum [78, 302, 303, 116]. Although this is a simple solution and can be used in conjunction with non-specialist cameras, it poses similar issues to HDR capture as it requires the same scene to be imaged multiple times. If objects in the scene move, occluded parts of the scene will be missing in some spectral bands.

Instead of sampling the spectral dimension, a prism or diffraction grating may be used to capture the full spectral distribution. In this case, however, only limited spatial resolution is possible. At the extreme, a single point is sampled (e.g., through a spectrometer) or a scanline (e.g., pushbroom scanning) [635]. More recently, an occlusion mask was employed to increase the spatial resolution of the prism-based approach, albeit at the cost of increased signal to noise ratio [186].

The third class of spectral imaging is known as Fourier Transform spectroscopy [45, 42]. Using an interferometer and a set of movable mirrors, an

[6] For example, http://www.spectralcameras.com/varispec.

interferrence pattern can be captured and subsequently analyzed to obtain the spectral distribution of the incoming light. This form of spectroscopy is used mostly in chemistry research and is typically limited to a single point of light.

The filter-based approach has often been used in the context of image statistics as it offers the most flexibility in terms of capture. Since each image depicts the same 2D scene at a different band of wavelengths, existing image processing techniques may be directly applied to the data, simplifying the analysis of the resulting data. Several hyper- and multi-spectral databases of natural and urban imagery collected using such an imaging system are discussed in Section 3.3.

3.1.5 Depth and Range Capture

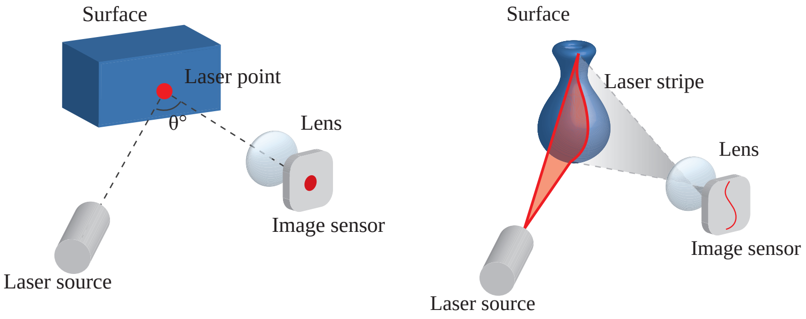
All image modalities discussed so far map a 3D scene to a 2D representation, resulting in the loss of depth information. Within real environments, our visual system makes use of several cues to determine depth in the scene [554]. Cues such as perspective or atmospheric haze are still available in 2D images, while motion parallax and binocular cues cannot be reproduced without stereo or motion information.

Broadly, a 3D representation of a scene can be obtained through reconstruction from one or more two-dimensional images of the same scene [719, 660] or through 3D scanning [142, 27, 50]. The former class of approaches offers interesting applications in the context of computer vision and graphics [50]. Inherent limitations of the registration and reconstruction processes, however, mean that 3D data generated through such methods is unlikely to be an accurate representation of the real scene geometry.

The latter class of 3D acquisition processes, namely 3D scanning, takes many forms, each with its respective benefits, limitations, and suitable applications. In contrast to the 2D reconstruction solutions discussed earlier, where normal 2D cameras can be used, 3D scanning is achieved through specialized devices and optical systems. The remainder of this section will give an overview of the main 3D scanning techniques, focusing in particular on the ones best suited for capturing the geometry of natural scenes and environements and will briefly discuss issues related to processing 3D scanning data.

Contact Scanning. The most accurate way to obtain 3D geometry information is through contact scanning systems, where a mechanical arm or probe is physically moved over the surface to be scanned. Such systems are found mostly in manufacturing and offer a high degree of accuracy but require contact with the scanned objects.

Structured Light. With this approach, a pattern of light with known characteristics is projected onto the object. Deformations in the pattern are then used to recover the geometry of the object. This method allows for an entire two-dimensional field of view to be captured at the same time and, as such, is one of the most efficient means of capturing geometry. Additionally, depending on



a. Triangulation scanning—point b. Triangulation scanning—stripe

Figure 3.7. An illustration of triangulation-based laser scanning using a point laser source and a laser stripe.

the pattern used and the precision of the calibration process, this method can achieve very accurate reconstruction [644, 551, 623, 513]. Recently, structured light–based solutions employing infrared light have been used to great success in Microsoft's Kinect.

As this class of 3D scanning techniques relies on light being reflected from the surface that is scanned, transparent and translucent objects as well as geometric configurations that lead to inter-reflections are likely to distort the recovered geometry [432, 292]. Additionally, structured light methods offer a limited spatial range and thus are better suited for scanning small-scale scenes such as objects or indoor environments.

Triangulation-based Scanning. This form of 3D scanning illuminates the object using a laser point light source and relies on trigonometric triangulation to determine the distance between the scanned surface and the scanner. To achieve that, a sensor is placed at a known distance and angle from the laser source. As the light from the laser source reflects off the object surface towards the sensor, the distance can be calculated from the position of the reflected light on the 2D sensor [123, 142]. This form of scanning is illustrated in Figure 3.7.

As the laser source is approximately a point light source, only a single point of the surface is captured each time. To reconstruct the whole surface, the laser ray may be fanned into a stripe, effectively capturing a scanline of the 3D surface. Using mirrors, the laser stripe can be swept over the object, capturing the full surface, as illustrated in Figure 3.7b [48, 155].

Although triangulation 3D scanners can achieve a high degree of accuracy, they are restricted to a range of a few meters and are therefore suitable for small to medium scale scenes, such as cultural heritage applications [459]. Additionally, the laser source needs to be at a certain distance from the imaging sensor (baseline) to enable the triangulation calculation.

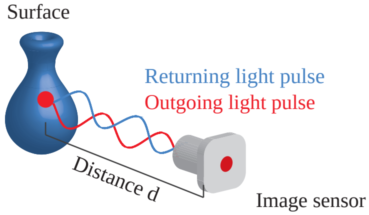


Figure 3.8. An illustration of the time of flight principle. A pulse of light reflects off the object surface and is captured by a sensor element in the camera. As the speed of light is known, the distance between the camera and the object can be computed.

Commercial examples of this technology include the DLT Laser Range Scanner[7] and the Konica Minolta series of laser scanners.[8]

Time of Flight Cameras. An alternative 3D scanning solution is implemented in *time of flight* (ToF) cameras. The principle behind this technology relies on the known speed of light c. A pulse of light is sent to the object and the time t it takes for the pulse to return to the camera is measured. From this, the distance to the surface can be computed as $d = \frac{tc}{2}$ [301]. Figure 3.8 shows an illustration of the ToF principle.

ToF cameras typically consist of an illumination element, such as infrared light [436] or a laser source [373], a lens, and an image sensor, such as a CMOS sensor.[9] Unlike triangulation systems, the illumination element can be placed very near the sensor, allowing for very compact designs. Since the travel time of the light pulse is extremely short, the accuracy of such a system depends highly on the precision of the timing measurements and the type of light source selected. Additionally, since the sensing element of a ToF camera needs to detect light, it is inherently sensitive to ambient light and scattering. Many commercial solutions employ the time of flight technology for 3D scanning. Some examples include the Leica[10] and Riegl[11] range scanners.

3D scanned data typically takes the form of a 3D point cloud. Depending on the scene, discontinuities and holes may be present in the point cloud. In computer graphics applications, such holes may be filled and the resulting mesh may be smoothed for further processing (e.g., model acquisition) [778, 156]. Although such processing is appropriate in applications where visual artefacts are undesirable, depth information may be distorted and therefore not suitable in the

[7] http://www.dlr.de/rm/en/desktopdefault.aspx/tabid-3932/6095_read-8882/

[8] http://sensing.konicaminolta.us/technologies/laser-triangulation/

[9] Fotonic range cameras: http://www.fotonic.com/content/Products/Default.aspx.

[10] http://www.leica-geosystems.us/forensic/

[11] http://www.riegl.com/

context of scene statistics. Statistical applications that make use of range data are discussed in detail in Chapter 11.

3.2 Post-processing and Calibration

When measuring the statistical regularities of an image ensemble, some care is necessary to ensure that irregularities, artifacts, and nonlinearities due to the camera, lens, or shooting conditions do not affect the results. To avoid such effects, the properties of the equipment used need to be accounted for. Moreover, to ensure that the images are calibrated and, therefore, pixel values between them are directly comparable, properties of the scene need to be measured.

3.2.1 Radiometric Calibration

Images captured with a digital camera are typically encoded in a device-dependent space, which represents the response of the particular sensor. Although SLR cameras often offer settings for sRGB and Adobe RGB encodings, due to manufacturing restrictions and tolerances, actual sensor responsivities may deviate from the ideal sRGB curve. In order to map pixel values to real-world radiometric quantities, the response function of the camera, which is the mapping between sensor illuminances to pixel values, needs to be known first. Although this information is not generally available, the nonlinearities for a particular camera or sensor can be recovered through the use of calibrated input [719].

The relationship between scene radiance and pixel values can be recovered in two ways, and the process is known as *camera characterization*. Specialized equipment can be used, such as a radiance meter, which can measure the response of the camera to particular wavelengths [610]. More commonly, a specific color target (such as the GretagMacbeth ColorChecker Chart) can be used in conjunction with a spectrophotometer to match measured XYZ values to the RGB values captured by the camera [503, 562].

Once a number of matches between patches (which are either measured or known through a color target) and corresponding pixels in the image are made, a mapping needs to be determined for all possible RGB values. Lookup tables store the data of such measured/captured pairs [361], which can be used for future reference. If a large enough number of samples is collected so that the camera's gamut is densely populated, then XYZ values corresponding to RGB samples not in the lookup table can be estimated using three-dimensional interpolation. Alternatively, if the number of samples available is not sufficient for interpolation, a function can be fitted to the available data, at the cost of increased computational complexity [338]. More recently, a comprehensive database of the space of possible and plausible camera responses was created, allowing for accurate modeling of camera responses with fewer parameters [290].

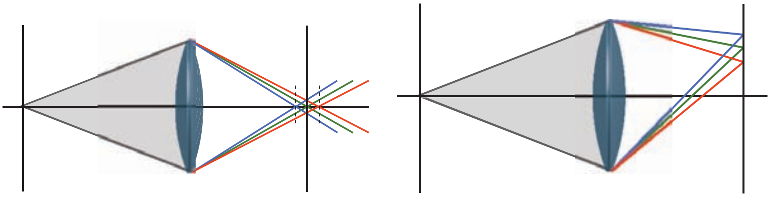


Figure 3.9. Longitudinal (left) and lateral (right) chromatic aberrations are shown (adapted with permission from [610]).

3.2.2 Lens Aberrations

The lens is responsible for focusing the light coming from the scene to an image plane (which may be a sensor in digital cameras or the film in analog cameras). For many applications, it is sufficient to model the camera as a simple pinhole whereby no lens is present and the aperture through which the light enters is a point. To model the various effects of an optical system to the image, however, more complex lenses need to be considered.

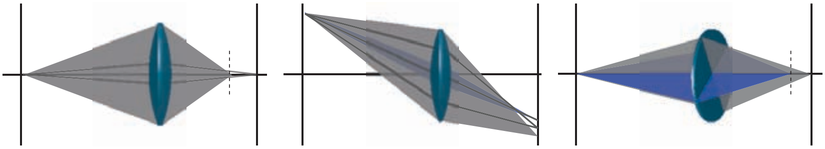
Lenses can consist of a single element, in which case they are referred to as *simple lenses* or multiple elements, called *compound lenses*. For a detailed description and models of lenses, the reader is referred to [799, 677, 610]. Assuming Gaussian optics (lenses are arranged along a single reference axis, and all light rays make only small angles ϕ with this axis, so that $\sin(\phi) \approx \phi$ and $\cos(\phi) \approx 1$), the relationship between the focal length of a lens f, the distance to an object s_0, and the distance behind the lens where the image is formed (s_i) is given by:

$$\frac{1}{f} = \frac{1}{s_0} + \frac{1}{s_i} \tag{3.2}$$

Imperfections in the design or construction of such lenses can cause them to violate the assumption of Gaussian optics. These deviations are known as *aberrations* and can be broadly categorized as chromatic, where the optical system behaves differently for different wavelengths, and monochromatic, where the distortions in the image are apparent in all color channels.

Chromatic Aberrations. Lens elements can have wavelength-dependent indices of refraction, resulting in different colors focusing in slightly different places. This causes *longitudinal* or *axial* chromatic aberration, which appears as chromatic image blur [75] (see Figure 3.9). Typically this form of aberration is managed optically, using a pair of thin lenses with different refractive indices, known as *achromatic doublets*. Such systems can reduce longitudinal aberration effects but not always completely remove them.

Another type of chromatic aberration is *lateral* chromatic aberration, which is the result of different wavelengths hitting the image plane at slightly different



a. Spherical aberration b. Coma c. Astigmatism

Figure 3.10. Three of the four blurring aberrations.

positions. This is more commonly known as *color fringing* and can be reduced digitally. This can typically be achieved by realigning the color channels of the image [491], which can compensate for the effect of the color shifting to some extent. Alternatively, for a more accurate reconstruction, a more complete characterization of the optical system can be utilized [389].

One particular case of chromatic aberration specific to digital cameras, especially ones using a charge coupled device sensor (CCD), is known as *purple fringing*. This effect is caused by a combination of factors that operate in addition to the lens aberrations described earlier. Highlights in the scene that are overly bright can cause some of the CCD quantum-wells to flood, creating the visual effect of blooming. Additionally, a false color effect can be created by the demosaicing process [398].

A number of digital solutions are available for correcting the effects of chromatic aberrations in images with various levels of accuracy. DxO Optics handles chromatic aberrations by taking into account the characteristics of the lens where possible [188]. A similar approach is also found in PTLens [198]. Adobe Photoshop (version CS4 tested) provides simple sliders to account for red/cyan and blue/yellow shifting.

Monochromatic Aberrations. A wider range of artifacts in images appears as the result of monochromatic aberrations. These include blurring aberrations and geometric distortions. Blurring aberrations can be further divided into *spherical aberration*, *coma*, *astigmatism*, and *Petzval field curvature*.

Spherical lenses are commonly used in optical systems as they are relatively easy to manufacture, but their shape is not ideal for the formation of a sharp image. Spherical aberrations are the consequence of light rays farther from the lens axis (marginal) focusing at a different position than rays passing through the middle of the lens. When marginal rays focus nearer the lens than the principal rays, the effect is known as *positive* or *undercorrected* spherical aberration, while a marginal focus farther away from the lens than the paraxial focus causes *negative* or *overcorrected* spherical aberration [772]. (See Figure 3.10a.)

Comatic aberrations cause objects farther from the optical axis to appear distorted. *Positive coma* occurs when marginal rays focus farther away from the optical axis than principal rays, while *negative coma* is the oposite effect, namely, the marginal rays focusing closer to the optical axis. (See Figure 3.10b.)

A third type of blurring aberration is known as *astigmatism* and is the result of off-axis rays of different orientations focusing at slightly different positions. The *meridional plane* for an off-axis object point is defined by that point and the optical axis. The plane orthogonal to the meridional plane is known as the *saggital plane*. Rays along these two planes focus at different positions on the optical axis. As such, if the lens is focused such that the meridional image is sharp, blur will occur in the saggital direction and vice versa. (See Figure 3.10c.)

A final aberration considered here is a consequence of the *field curvature*. The image and object planes are generally considered to be planar. However, in the case of spherical lenses they are in fact spherical, which is known as *Petzval field curvature*. Regions of the image farther from the optical axis increasingly deviate from the planar assumption. This is more evident when a planar image plane is used (such as a sensor), in which case the image is less sharp towards the periphery [610].

The effect of blurring aberrations can be reduced in post-processing by sharpening the image or during the image capture by using smaller aperture settings and thus only allowing rays closer to the optical axis to enter. Optical solutions do exist and generally involve the use of lens doublets or carefully designed lenses. These solutions, however, complicate the design of the optical system and are, as a result, found in more expensive lenses.

A different class of monochromatic aberrations consists of radial distortions. These cause straight lines in the scene to appear curved in the resulting image and can be corrected digitally either manually or using information about the characteristics of the lens that was used. Radial distortions are the result of deviations from a linear projection model and are more pronounced for wider angle lenses.

Although interactions between the various lens elements of a complex optical system can produce some less well-defined distortions, the two main types considered here are *barrel* and *pincushion* distortions, shown in Figure 3.11. Barrel distortion is the result of decreasing lateral magnification and it causes straight lines to curve away from the image center. Pincushion distortion occurs in lenses with increasing lateral magnification and causes straight lines to curve towards the image center [794].

Radial distortions can be adequately estimated as a fourth-degree polynomial:

$$x = x(1+\kappa_1 r^2 + \kappa_2 r^4) \tag{3.3}$$

$$y = y(1+\kappa_1 r^2 + \kappa_2 r^4) \tag{3.4}$$

where $r^2 = x^2 + y^2$ and $\kappa_{1,2}$ are the *radial distortion parameters* [719]. Several methods have been proposed in the literature for estimating the distortion parameters of Equation (3.4). In the simplest case, by photographing a test scene containing straight lines, the distortion parameters can be adjusted until the lines in the resulting image are also straight [196, 75]. This process can, of course, be simplified even further if the lens characteristics are provided by the manufacturer.

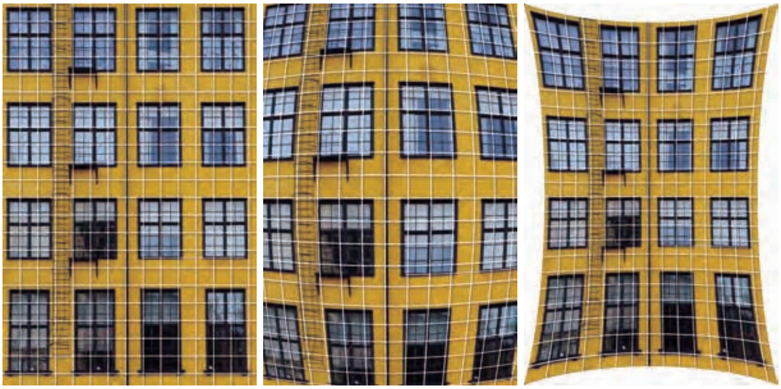


Figure 3.11. This figure demonstrates the effect of barrel (middle) and pincushion (right) distortions on an image (left). In this case, the distortions are exaggerated and overlaid with a grid for demonstration purposes.

More complex solutions have also been proposed, combining image alignment with distortion parameter estimation [648, 706] or by recovering the radial distortion characteristics together with intrinsic camera parameters [124, 712, 308].

Barrel and pincushion distortions are not the only possible geometric distortions. The interactions of the various lens elements and their resulting distortions can result in effects that require more complex solutions. In these cases, approaches using calibrated targets can still be of use. Various software packages offer solutions that reduce or remove the effect of radial distortions. Options such as DxO Optics or PTLens use lens-specific information while other solutions exist that require more manual input, such as Hugin or the Lens Correction module in Adobe Photoshop.

Vignetting. Vignetting is a common artifact appearing in photographs, which causes areas of the image further from the center to appear darker. Although this effect often goes unnoticed and sometimes can even be added to images for artistic purposes, the magnitude differences it causes can be large. It has been found that for certain optical systems, the amount of light reaching the corners of the image can be several times less than the light going through the center [821]. Such differences may affect the results in statistical or computer vision applications.

Several factors contribute to vignetting [75]. *Mechanical vignetting* can be caused by lens extensions or accessories that physically block parts of the lens. *Optical vignetting* arises when rays arrive from an oblique angle. Parts of the lens may be obstructed by the rim of the barrel, reducing the effective aperture and thus allowing less light to enter the lens for these angles. Finally, a third cause of vignetting is known as *natural vignetting* and is the effect of natural light falloff. The severity of this effect depends on the angle θ that light rays exit the rear element and hit the image sensor and, for simple lenses, can be modeled as $\cos^4(\theta)$ falloff [269].

Both optical and natural vignetting can be reduced digitally after the image is captured. The simplest approaches for estimating the vignetting for a given optical system involve capturing a photograph of a uniformly colored and illuminated surface [75, 719]. Using such an image, the vignetting function can be computed and its effect can then be removed from subsequent images taken using the same system.

A limitation of this approach is that it requires such a calibration image to be captured with the same camera and lens as the image(s) to be corrected [22, 390]. This may not always be possible. To that end, information from a single image can be used to reduce the effects of vignetting. This can be achieved by locating slow intensity variations in the radial direction [821]. Alternatively, in cases where multiple overlapping images are available (for instance, when capturing a panorama), corresponding pixels that appear in more than one image can help recover the vignetting function [467, 269].

3.2.3 Noise

Noise can be a major cause of image degradation and can be defined as any random variation in brightness or color in the image caused by the sensor, circuitry, or other parts of the camera. Noise can have various sources, each requiring different treatment. Some common types of image noise will now be discussed.

Fixed Pattern Noise. Fabrication errors can cause different pixels of a sensor array to have different responses to constant illumination. This is known as *fixed pattern noise* and appears in two forms. *Dark signal non-uniformity (DSNU)* (DSNU) is the fixed offset from the average across each pixel of the image sensor, occurring even when no illumination is present. *Photo response non-uniformity* (PRNU) is the variation in the responsivity of pixels under constant illumination.

Fixed pattern noise can depend on camera settings and conditions such as exposure time and temperature but can be removed, either in the camera hardware or in post-processing. In the latter case, DSNU can be removed simply by capturing and subtracting an image with no illumination (for instance, by keeping the lens cap on) in otherwise the same conditions as the images to be calibrated. To additionally remove noise due to PRNU under various illumination levels, an image of a flat, uniformly lit surface can be captured with the illumination set such that it nearly saturates the sensor pixels. Then using the DSNU calibration image described previously and the new image, the response of each pixel can be linearly interpolated for a particular set of conditions (temperature, camera settings, etc.) [765].

Dark Current. Even if no light reaches the sensor, electrons can still reach the quantum-wells, generating *dark current shot noise*. The severity of this type of noise is affected by temperature and consequently can be reduced by cooling the sensor [610].

Photon Shot Noise. The number of photons collected by each individual quantum-well of an image sensor is not constant for a given intensity. Due to this randomness, a given pixel can receive more or fewer photons than the average. The distribution of photons arriving at the sensor can be modeled by Poisson statistics (mean μ equals variance σ^2). As such, if the intensity of the source increases, so will the variance in the signal. Since the signal-to-noise ratio can be defined as $SNR = \mu/\sigma = N/\ \ \overline{N} = \ \ \overline{N}$, where N is the mean number of photons hitting the image sensor, large values of N cause the SNR to be large too, thus reducing the effect of this type of noise.

Most of the noise sources discussed do not depend on illumination levels in the scenes. As such, in darker scenes, the signal-to-noise ratio becomes smaller, effectively amplifying the effect of noise in the image. Although dark current and fixed pattern noise can be minimized though the use of appropriate calibration, other noise sources (including photon shot noise) cause less predictable patterns. The reader is referred to the work by Healey and Kondepudy [321] for a detailed account on noise estimation and camera calibration.

3.3 Image Databases

The previous sections dealt with issues related to constructing an image collection from scratch. Although collecting a novel set of images allows for full control over the capture and calibration procedures as well as the scene selection, many existing databases are available and suitable for statistical analysis and related applications. This section will describe some of these databases. The aim of this section is not to provide a comprehensive list of all image databases available but rather to highlight collections that have been constructed with a high degree of accuracy in mind.

3.3.1 Van Hateren's Natural Image Database

This dataset is one of the largest and earliest collections of natural images used in natural image statistics literature [316, 352]. Although the images encode only intensities and are therefore monochromatic, they are calibrated and represent a wide variety of natural scenes.

- **Number of images:** 4212

- **Image resolution:** 1536 × 1024 pixels

- **Channels:** 1 (Intensity only)

- **URL:** http://bethgelab.org/datasets/vanhateren/

3.3.2 University of Texas at Austin Databases

This collection contains over a thousand high-resolution RGB images of natural scenes [252] and a smaller set of images (90) taken around the University of Texas at Austin campus, containing manmade objects, buildings, and people [90]. Additionally, data for hand-segmented contours are provided for 20 images [253]. The images are provided as PPM files, directly converted from the RAW sensor data, and the camera sensitivity functions are given on the database webpage to allow for conversions to other color spaces. Figure 3.12 shows a few sample images from the natural image set.

- **Number of images:** 1204 (natural), 90 (campus), 20 (hand segmented)

- **Image resolution:** 4284 × 2844 pixels

- **Channels:** 3 (RGB)

- **URL:** http://www.cps.utexas.edu/natural_scenes/db.shtml

3.3.3 UPenn Natural Image Database

Similar to the Van Hateren database (Section 3.3.1), this collection focuses on natural scenes, specifically scenes depicting baboon habitat in Botswana. It contains around four thousand images, which are calibrated and available in different formats. The database can be accessed and browsed through an online gallery page, which allows for a custom selection of images to be downloaded easily. The aim of this database is to represent the type of environment where the human visual system initially evolved [740].

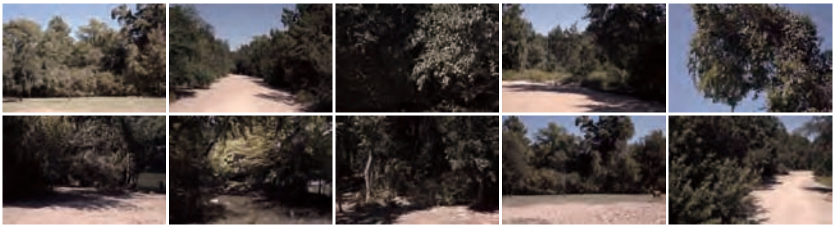


Figure 3.12. A small set of natural images from the University of Texas at Austin database [252]. (The images are shown with kind permission from the author, who retains the copyright.)

- **Number of images:** approximately 4000
- **Image resolution:** 3040 × 2014 pixels
- **Channels:** 3 (sRGB)
- **URL:** http://tofu.psych.upenn.edu/~upennidb/

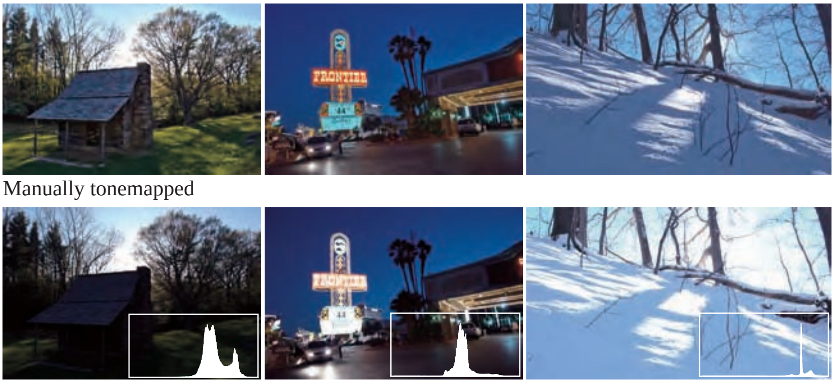
3.3.4 The Barcelona Calibrated Images Database

Both natural and urban scenes were included in this database. In contrast to most other collections, the natural scenes also include snow and sea landscapes in addition to foliage. An approximately diffuse gray sphere was placed in front of the camera in all images, to allow for recovery of the illumination. The database is calibrated and provided in three device-independent forms, namely, CIE XYZ and two LMS cone response spaces [697, 708]. Particular effort has gone into the accurate calibration and camera response recovery for this database [761, 561, 559].

- **Number of images:** 448 in 7 categories
- **Image resolution:** 567 × 378 pixels
- **Channels:** 3 (XYZ or LMS)
- **URL:** http://www.cvc.uab.es/color_calibration/

3.3.5 HDR Photographic Survey

This is one of the very few high dynamic range image databases and it has been constructed with significant care going towards accurate color and luminance reproduction [203]. It contains more than a hundred HDR images of natural and urban scenes, all of which are given as linear, calibrated data. To obtain radiometrically accurate quantities, the images need to be multiplied with a luminance factor, which is provided separately. Additionally, for many of the images, colorimetric measurements and corresponding color appearance data is given for selected patches. Although the database contains a variety of scenes, we note that their choice may be biased towards more extreme examples of dynamic range and as such, they may not be an accurate representation of luminance distributions overall. Figure 3.13 shows a few example images from this database as well as their luminance histograms.



Manually tonemapped

Linearly compressed

Figure 3.13. Some example images from the HDR Photographic Survey [203]. The top row has been manually tonemapped, while the images in the bottom row have been linearly compressed. The histograms in the insets show the luminance distributions of the images in logarithmic scale. (The images are shown with kind permission from the author, who retains the copyright.)

- **Number of images:** 106
- **Image resolution:** varied, around 10–12 MP
- **Channels:** 3 (RGB)
- **URL:** http://www.cis.rit.edu/fairchild/HDR.html

3.3.6 IPL Calibrated Color Image Database

This calibrated database was recently constructed for the purpose of studying chromatic adaptation using color statistics. It contains small scale scenes, such as plants and office objects, as well as ten pages of the Munsell Book of Color. All images were captured under two controlled illuminations, namely, CIE D65 and A [440]. Although many databases include varied illumination, the advantage of this collection comes from the controlled setup used, where all scenes are captured with the same set of illuminations.

- **Number of images:** 130 plus ten pages of the Munsell Book of Color
- **Image resolution:** 1000×1280 pixels

- **Channels:** 3 (CIE XYZ)
- **URL:** http://isp.uv.es/data_color.htm

3.3.7 McGill Calibrated Color Image Database

This collection contains images in nine different image classes, including animals, textures, and shadows. Calibration data is provided on the database website to allow conversions to linear RGB data or to LMS cone responses [544]. Although many of the image collections discussed include natural images, this database contains a number of more unusual image classes.

- **Number of images:** 850
- **Image resolution:** 786 × 576 pixels (scaled down from 1920 × 2560 pixels, which are available by contacting the database owners)
- **Channels:** 3 (RGB)
- **URL:** http://tabby.vision.mcgill.ca/

3.3.8 Hyperspectral Images of Natural Scenes

These two collections contain both natural and urban scenes captured in Portugal, which were originally obtained for studying phenomena relating to color vision [527, 240, 239]. The scenes were illuminated by direct sunlight and no artifical light sources. Calibration data is provided to convert scene reflectances to radiances. Note that only eight images from each of the sets are available online.

2002 Collection

- **Number of images:** 30
- **Image resolution:** cropped from 1024 × 1024 pixels, maximum 820 × 820 pixels
- **Channels:** 33 spectral measurements
- **URL:** http://personalpages.manchester.ac.uk/staff/david.foster/Hyperspectral_images_of_natural_scenes_02.html

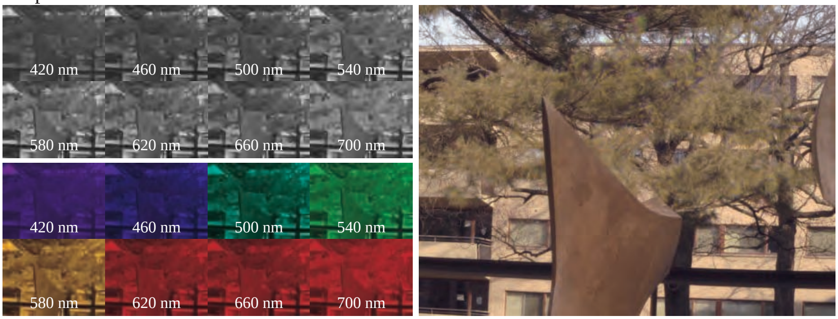
2004 Collection

- **Number of images:** 25
- **Image resolution:** cropped from 1024 × 1344 pixels
- **Channels:** 33 spectral measurements
- **URL:** http://personalpages.manchester.ac.uk/staff/david.foster/Hyperspectral_images_of_natural_scenes_04.html

3.3.9 Real-World Hyperspectral Images Database

Recently, a collection of 50 hyperspectral images was compiled, including both natural and artificial scenes [116]. Each scene consists of 31 narrow-band spectral measurements of approximately 10 nm in bandwidth, spanning the range between 420 nm and 720 nm, which approximately coincides with the visible spectrum according to the human visual system. An example scene with a subset of the spectral measurements is shown in Figure 3.14.

- **Number of images:** 50
- **Image resolution:** 1392 × 1040 pixels
- **Channels:** 31 spectral measurements
- **URL:** http://vision.seas.harvard.edu/hyperspec/



a. Spectral bands

b. Spectral bands (mapped to RGB)

c. RGB scene

Figure 3.14. A sample image from the Hyperspectral Database presented in [116]. (The images are shown with kind permission from the author, who retains the copyright.)

3.3.10 Bristol Hyperspectral Images Database

Although this is a smaller set than other hyperspectral datasets discussed, it is one of the oldest such collections [560]. The database contains natural scenes, captured with a custom-made hyperspectral camera [78].

- **Number of images:** 29
- **Image resolution:** 256 × 256 pixels
- **Channels:** 31 spectral measurements
- **URL:** http://www.cvc.uab.es/color_calibration/Bristol_Hyper/

3.3.11 Amsterdam Library of Object Images (ALOI)

Unlike the other collections discussed, this database focuses on small objects, depicted in isolation, under controlled illuminations [259]. The database contains 1000 small objects, each photographed using 8 different illumination directions, 3 camera positions, 12 illumination colors and 72 object viewpoints. Additionally, wide baseline stereo images are provided for 750 of the objects.

Several other object databases exist [532, 529, 420], but ALOI is the largest and most comprehensive to our knowledge both in terms of individual objects included and viewpoint and illumination variations. Although this type of imagery is likely to mostly find applications in evaluating computer vision and image processing algorithms, we include it here due to the wealth of data it provides. Figure 3.15 shows a series of illumination directions and camera positions for a single object.

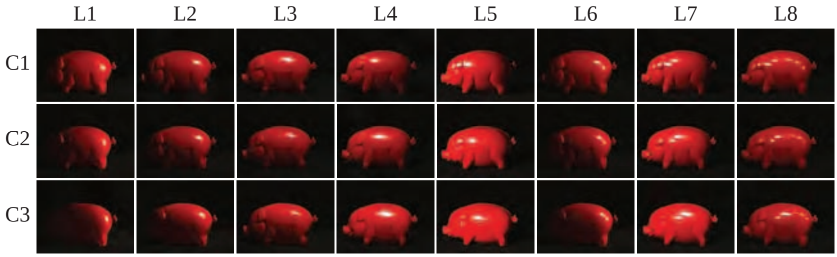


Figure 3.15. A series of illumination directions (L1 to L8) and camera positions (C1 to C3) for one object from the ALOI database. Please refer to [259] for details. (The images are shown with kind permission from the author, who retains the copyright.)

- **Number of images:** 110,250
- **Image resolution:** 768 × 576 pixels
- **Channels:** 3
- **URL:** http://staff.science.uva.nl/~aloi/

3.3.12 Caltech-256 Object Category Dataset

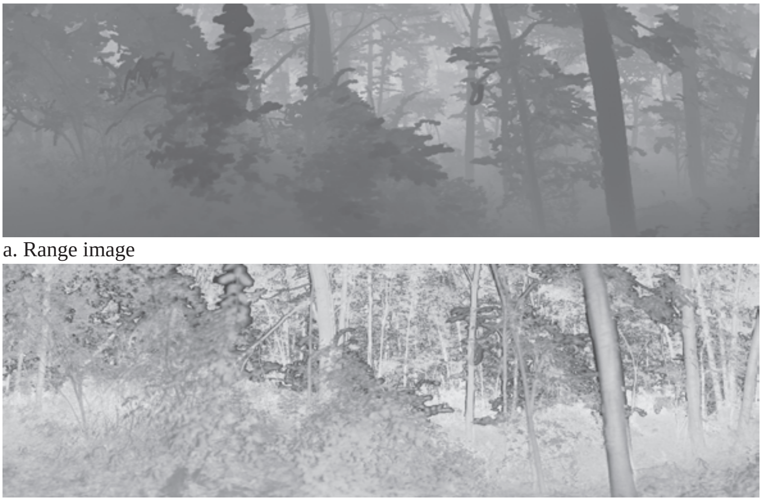
This collection groups over 30,000 images in 256 distinct categories. Each category contains a minimum of 80 images, collected from internet sources (Google Image search). Although this database by its nature is uncalibrated, we include it here because of the large number of categories included, which could be used in studies focusing on the analysis of images of different categories. Details on the categories used and the collection procedure are described in [288]. This collection supersedes the previous Caltech-101 database [208].

- **Number of images:** 30,607 (in 256 categories)
- **Image resolution:** varied
- **Channels:** 3 (RGB)
- **URL:** http://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-001

3.3.13 Brown Range Image Database

This database consists of 197 range images of both natural and urban scenes. The images have a vertical field of view of 80 degrees and horizontal of 259 degrees, and were captured using a laser range-finder with a rotating mirror. An example range image and the corresponding intensities are shown in Figure 3.16. The database is described in more detail in [350, 447].

- **Number of images:** 197
- **Image resolution:** 444 × 1440 pixels
- **Channels:** 4 (depth, intensity, bearing, and inclination)
- **URL:** http://www.dam.brown.edu/ptg/brid/range/index.html



a. Range image

b. Corresponding intensities

Figure 3.16. An example image from the Brown Range Image Database. The range data is visualized by normalizing and compressing the depth values with an exponent of 0.3. (The images are shown with kind permission form the author, who retains the copyright.)

Part II

Image Statistics

Chapter 4

First-Order Statistics

Digital images consist of an array of pixels, each representing the average incoming illumination from a small part of the scene. Effectively, each pixel of a digital camera sensor integrates light over a small area, storing it as a single number. If we were to use a very low resolution sensor consisting of a single pixel, our digital image would simply be a single number representing the average light in the scene—this is, after all, what a photometer does. This is illustrated in Figure 4.1.

Our visual system performs a similar task as photoreceptors in the retina, spatially sample light (see Chapter 1). Some creatures such as mollusks, however,

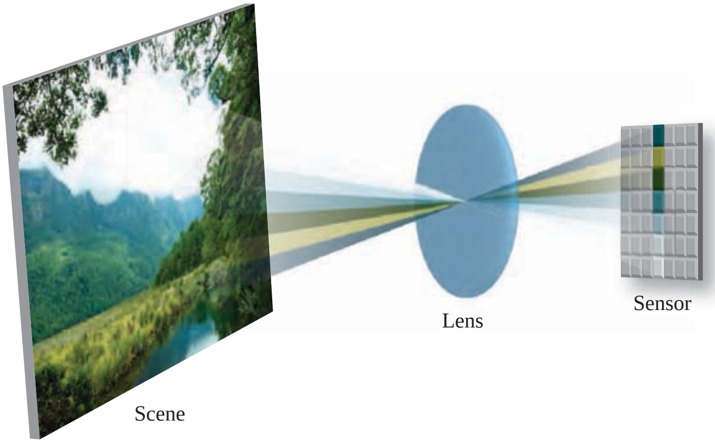


Figure 4.1. Pixels on the camera sensor integrate light from the scene, effectively capturing the average incoming illumination over a small area of the scene.

sense light in a way more akin to our single-pixel camera example, suggesting that useful information may be found in images even at the single pixel level, without any spatial information [555].

4.1 Histograms and Moments

The distribution of pixel values in images can be represented with *histograms*. For a given intensity vector I, we define its histogram H with B bins of width V as follows:

$$\begin{aligned}
H &= (h(1), v(1)), \ldots, (h(B), v(B)) && (4.1)\\
B &= \left\lceil \frac{\max(I) - \min(I)}{V} \right\rceil && (4.2)\\
h(i) &= \sum_{p=1}^{N} P(I(p), i), \quad i \quad [1, B] && (4.3)\\
v(i) &= \min(I) + (i-1)V && (4.4)\\
P(I(p), i) &= \begin{cases} & i = \left\lfloor \dfrac{I(p) - \min(I)}{V} + 1 \right\rfloor \\ 0 & \text{otherwise} \end{cases} && (4.5)
\end{aligned}$$

where H is the set of all pairs $(h(i), v(i))$ for all $i \quad [1, B]$ corresponding to the number of elements and value of the i^{th} bin of the histogram. $I(p)$ is the value of the p^{th} pixel of vector I, which contains a total of N pixels, and $P(I(p), i)$ represents the probability of a pixel $I(p)$ belonging to a bin i.

Figure 4.2 shows the intensity histograms of some example images. By visual inspection, we can already surmize that the shape of a histogram depends on the content of the image: scenes with uniform illumination (top left) will typically have an approximately Gaussian shape (top right). If, on the other hand, there are shadows and highlights present (bottom left), the histogram will likely exhibit a bimodal distribution (two peaks corresponding to the different parts of the bottom-right image in Figure 4.2).

Although such high-level characterization and understanding of histograms might be useful in photography, to derive an analytical description of the shape of an image histogram, properties of the intensity distribution may be computed, such as its mean and variance. *Statistical moments* are commonly employed to quantitatively describe the shape of a distribution. The k^{th} moment of a distribution can be computed as follows:

$$m_k = \sum_{p=1}^{N} \frac{(I(p) - c)^k}{N} \tag{4.6}$$

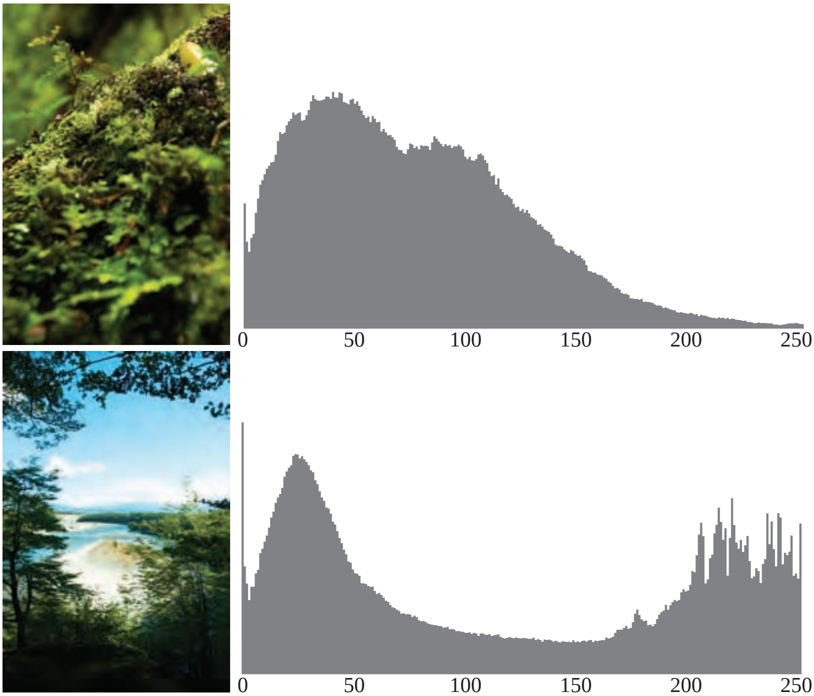


Figure 4.2. The top image is uniformly illuminated, leading to an approximately Gaussian distribution. In the bottom image, parts of the scene are in the shade while other parts are brightly illuminated, leading to a bimodal distribution.

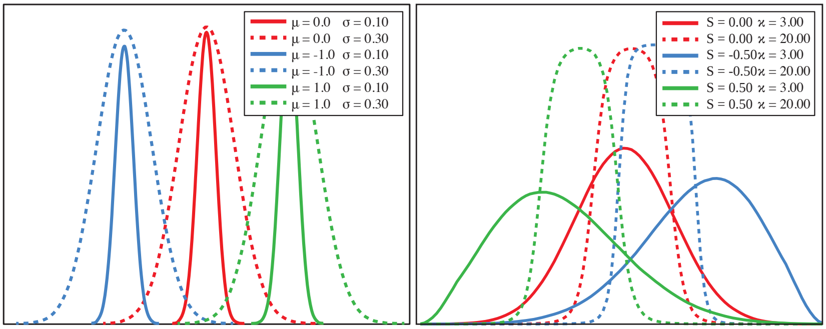
where c can be any constant. Generally, if $c = 0$, then the above equation computes the raw moments of the distribution, while setting $c = \mu$ gives us the central moments (i.e., centered at the mean). The first moment corresponds to the mean μ of the distribution and the second is the variance σ^2 (which is by itself the square of the standard deviation σ).

The meaning of further moments is less straightforward but the *skewness* S and *kurtosis* κ of a distribution relate to the third and fourth moments, respectively. More specifically, the skewness and the kurtosis are defined as:

$$S = \frac{m_3}{\sigma^3} \tag{4.7}$$

$$\kappa = \frac{m_4}{\sigma^4} \tag{4.8}$$

respectively, where m_3 and m_4 are the third and fourth central moments, respectively. Skewness encodes asymmetry in the distribution while kurtosis closely relates to sparseness. Figure 4.3 shows some example distributions with specific values for mean, standard deviation, skewness, and kurtosis.



a. Different mean and standard deviation
S = 0.00, ϰ = 3.00

b. Different skewness and kurtosis
μ = 0.0, σ = 0.1

Figure 4.3. Probability distributions with specific mean, standard deviation, skewness, and kurtosis values from sets of random numbers drawn from the Pearson system [376].

4.1.1 Image Moments and Moment Invariants

In our analysis so far, we have treated images as one-dimensional vectors of intensity values, whereby the intensity of a pixel p is given by $I(p)$. We can expand this definition to consider the horizontal and vertical position of a pixel and define its intensity value at position x,y as $I(x,y)$, where $x \quad M$ and $y \quad N$ and $M,N \quad \mathbb{R}$ are the dimensions of the image. The moments of a 2D image $I(x,y)$ can now be defined as:

$$m_{j,k} = \sum_{x=1}^{M}\sum_{y=1}^{N}(x^j - c_x)(y^k - c_y)I(x,y) \tag{4.9}$$

where j,k define the order of the moment in the horizontal and vertical direction, respectively. Setting c_x, c_y to 0 leads to the raw moments of the image, while $c_x = m_{1,0}/m0,0$ and $c_x = m_{0,1}/m0,0$ lead to central moments $m_{i,j}$.

Similar to (4.6), the above definition can be normalized by the number of pixels or *area* of the image. Using the formulation in (4.9), the area $m_{0,0}$ can computed as:

$$m_{0,0} = \sum_{x=1}^{M}\sum_{y=1}^{N} I(x,y) \tag{4.10}$$

The first-order moments $m_{1,0}$ and $m_{0,1}$ correspond to the *centroids* of the image, while second-order central moments can be used to determine the orientation of an image.

Image moments offer a very powerful basis for representing and describing images. An image can be uniquely described and reconstructed by its (raw) moments—a property known as the *uniqueness theorem* [233] (although this is not a case for a general distribution). By manipulating or combining moments, image descriptors can be constructed that can characterize images uniquely up to certain transformations. That is they are *invariant* to specific transformations.

In the simplest case, we can consider the central image moments. As they effectively align the image centroid with the origin of the coordinate space, they are invariant to translations—they remain constant, irrespective of the position of the image (or segment of interest) within the coordinate space. In addition to translational invariance, moments can be combined and manipulated to define bases that are scale, rotation, and other property invariant [233]. Such bases are known as *moment invariants* and were first introduced in the pioneering work of Hu et al. [349], where a set of moments $\phi_{1,\cdots,7}$ was introduced to obtain rotational invariance. These are the following:

$$\phi_1 = m_{2,0} + m_{0,2} \tag{4.11}$$

$$\phi_2 = (m_{2,0} - m_{0,2})^2 + 4m_{1,1}^2 \tag{4.12}$$

$$\phi_3 = (m_{3,0} - 3m_{1,2})^2 + (3m_{2,1} - m_{0,3})^2 \tag{4.13}$$

$$\phi_4 = (m_{3,0} + m_{1,2})^2 + (m_{2,1} + m_{0,3})^2 \tag{4.14}$$

$$\begin{aligned}\phi_5 = {} & (m_{3,0} - 3m_{1,2})(m_{3,0} + m_{1,2})((m_{3,0} + m_{1,2})^2 \\ & - 3(m_{2,1} + m_{0,3})^2) + (3m_{2,1} - m_{0,3})(m_{2,1} + m_{0,3}) \\ & \times (3(m_{3,0} + m_{1,2})^2) - (m_{2,1} + m_{0,3})^2)\end{aligned} \tag{4.15}$$

$$\begin{aligned}\phi_6 = {} & (m_{2,0} - m_{0,2})((m_{3,0} + m_{1,2})^2 - (m_{2,1} + m_{0,3})^2) \\ & + 4m_{1,1}(m_{3,0} + m_{1,2})(m_{2,1} + m_{0,3})\end{aligned} \tag{4.16}$$

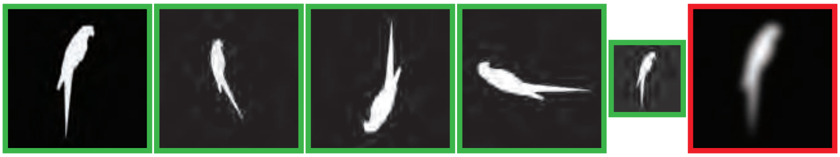
$$\begin{aligned}\phi_7 = {} & (3m_{2,1} - m_{0,3})(m_{3,0} + m_{1,2})((m_{3,0} + m_{1,2})^2 \\ & - 3(m_{2,1} + m_{0,3})^2) - (m_{3,0} - 3m_{1,2})(m_{2,1} + m_{0,3}) \\ & \times (3(m_{3,0} + m_{1,2})^2 - (m_{2,1} + m_{0,3})^2)\end{aligned} \tag{4.17}$$

Figure 4.4 shows the Hu moments for an image with different transformations. Scale and rotation does not affect the resulting moments, while blur does.

Since then, several different moment invariants have been proposed, initially for improved rotational invariance and later for achieving invariance to affine transforms [229, 230, 715], convolution [232, 714], and even elastic transformations [228]. Because of the descriptive power of such bases, in conjunction with their ability to ignore image transforms or degradations [231], moment invariants have found applications in fields as varied as character recognition [230, 802], forgery detection [485], and mesh simplification [725].

4.1.2 Histogram Adjustments

Despite their simplicity, first-order statistics have now found several applications in image processing. Studies have shown correlations between first-order statistical regularities in images and properties of the illuminant [610, 4], which has



ϕ_1	0.5621	0.5621	0.5621	0.5621	0.5621	0.5945
ϕ_2	0.2618	0.2618	0.2618	0.2618	0.2618	0.2616
ϕ_3	0.0791	0.0791	0.0791	0.0791	0.0791	0.0787
ϕ_4	0.0540	0.0540	0.0540	0.0540	0.0540	0.0537
ϕ_5	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035
ϕ_6	0.0243	0.0243	0.0243	0.0243	0.0243	0.0242
ϕ_7	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003

Figure 4.4. Rotation and scaling transformations lead to the same Hu moments [349] while blurring the image does not. More recent moment invariants have been developed to handle blurring as well as other transforms [233].

proven useful in areas such as white balancing. Moreover, transferring statistical moments between images in appropriate color spaces has been demonstrated in what is now known as color transfer [607]. These color-related applications are discussed in detail in Chapter 10.

First-order statistics can also be computed on a single image basis. By manipulating the distribution of values within a single image, a variety of effects can be achieved. For instance, the contrast of an image that only covers a small portion of the available range of intensities can be increased by adjusting the pixel values such that the full range of intensities is more equally represented. This process is known as *histogram equalization* and an example can be seen in Figure 4.5.

A more general version of histogram equalization is *histogram matching*, where the histogram of a source image (I_s) is matched to that of a given target (I_t). First, the cumulative histograms of the source and target are computed:

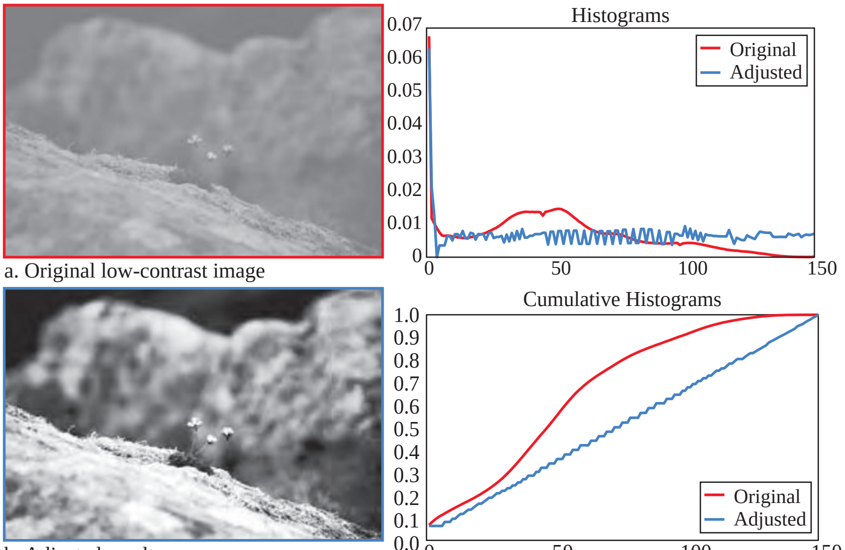
$$C_s(i) = \sum_{i=1}^{B} h_s(i) \tag{4.18}$$

$$C_t(i) = \sum_{i=1}^{B} h_t(i) \tag{4.19}$$

after which an image is matched to another according to these two cumulative histograms:

$$I_0(p) = v_t\left(C_t^{-1}\left(C_s\left(\frac{I(p) - \min(I) + 1}{V}\right)\right)\right) \tag{4.20}$$

Here, a cumulative histogram C is defined as a function mapping a bin index to a cumulative count. The inverse function C^{-1} acts as a reverse lookup on the histogram, returning the bin index (and therefore the intensity value) corresponding



a. Original low-contrast image

b. Adjusted result

Figure 4.5. Histogram equalization can increase the contrast of a low contrast image (top left) by reshaping the intensity distribution more equally. The equalized resulting image is shown at the bottom left.

to a given count. An example of this technique applied on a source-target pair of intensity images is shown in Figure 4.6.

4.2 Moment Statistics and Average Distributions

Histogram moments have been used in the context of natural image statistics as a means to characterize images of specific classes. Depending on the content of images, different values will arise for the various moments as we will see in this section.

In a large-scale image statistics study, Huang and Mumford analyzed more than 4000 grayscale images of natural scenes (taken from the database created by J. H. van Hateren [316]; see Section 3.3.1) by computing central moments of logarithmic histograms [352]. The values found for the mean, standard deviation, skewness, and kurtosis were, respectively, $\mu = 0$, $\sigma = 0.79$, $S = 0.22$, and $\kappa = 4.56$. The value of the skewness shows that the distribution is not symmetric, which can be attributed, at least partly, to the presence of sky in many of the images, resulting in a bias towards higher intensities. In addition to that, the values of both the skewness and the kurtosis show that the distribution is non-Gaussian.

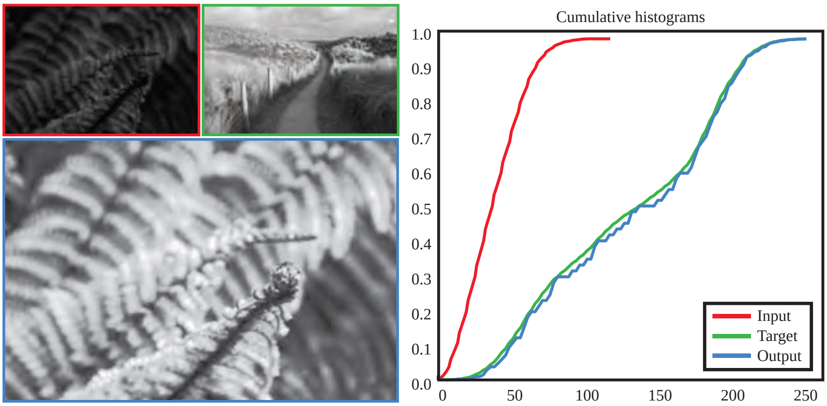


Figure 4.6. The source image (top left, red) is matched to the target (top right, green) using histogram matching to produce the image at the bottom (blue). The corresponding cumulative histograms are shown.

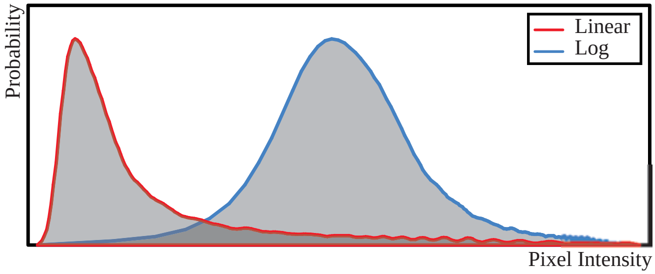


Figure 4.7. The linear and log intensity histograms found by Brady and Field from their analysis of 46 natural images. (Adapted from [74].)

A less skewed distribution was found by Brady and Field [74] in their analysis of 46 logarithmically transformed natural images, while the linear images resulted in a distribution skewed towards darker values. Although no exact values were provided for the moments of the distributions, Figure 4.7 shows the resulting histograms for both cases. As can be seen from these examples, although in both cases natural images were used and analyzed, the results can vary. Generally, the distribution of log intensities for natural images does not deviate far from Gaussian.

Results, however, do depend on the choice of images. In all the examples given, the images used to compute the distributions were captured using traditional imaging techniques and therefore were restricted in terms of the dynamic range they could represent (e.g., 8 bits in [585] and 12 bits in [352, 74]). As discussed in Section 3.1.2, high dynamic range (HDR) imaging allows for scenes with more extreme illumination to be captured. HDR images have been analyzed

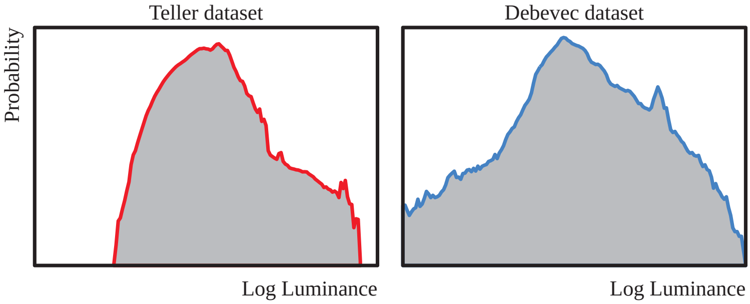


Figure 4.8. The log luminance histograms for the environment map collections analyzed in [182].

and compared to traditional imagery in a small set of studies, indicating marked differences between high dynamic range (HDR) and low dynamic range (LDR) ensembles [182, 185, 585].

Dror et al. analyzed HDR environment maps to determine statistical regularities in real-world illumination [182, 185]. Environment maps capture the full field of view from a particular viewpoint and represent it as a hemisphere (or sphere—see Section 3.1.3). The illumination arriving at the centre of the hemisphere is thus fully represented in such an image. Figure 4.8 shows the histograms for the two different datasets used in this study.

In a more recent study, Pouli et al. collected and analyzed images of four different scene categories for the purpose of determining the statistical differences between HDR and LDR imagery [585]. For each scene, a series of nine differently exposed images was captured to form the HDR image. The best-exposed image of the series was then selected to form the corresponding LDR ensemble, effectively constructing two sets depicting the same scenes but using different capture techniques (example images from these datasets can be seen in Figure 6.14).

Figure 4.9 shows the average logarithmic histograms for the HDR and LDR sets discussed as well as the images from the HDR Photographic Survey [203]. Visual analysis of the histograms quickly shows considerable differences both between LDR and HDR as well as between scene classes. Unsurprisingly, given the 32-bit encoding of the HDR images, the quantization seen toward the left of the LDR ensembles is absent in the HDR histograms. On the other hand, long tails can be observed in the HDR distributions towards higher luminance values, which can be attributed to the presence of light sources or highlights that cannot be adequately represented with LDR imagery.

Looking at specific differences across the different image classes, a few more observations arise. In the LDR case, the average histograms of all datasets look very similar. When captured in HDR though, the same scenes lead to significant differences between categories.

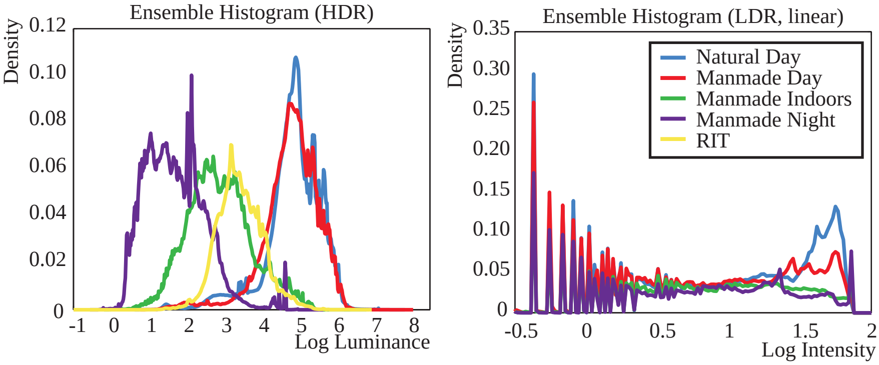


Figure 4.9. Log histograms for the HDR and LDR datasets from [585]. Different scene types lead to much larger differences when captured in HDR.

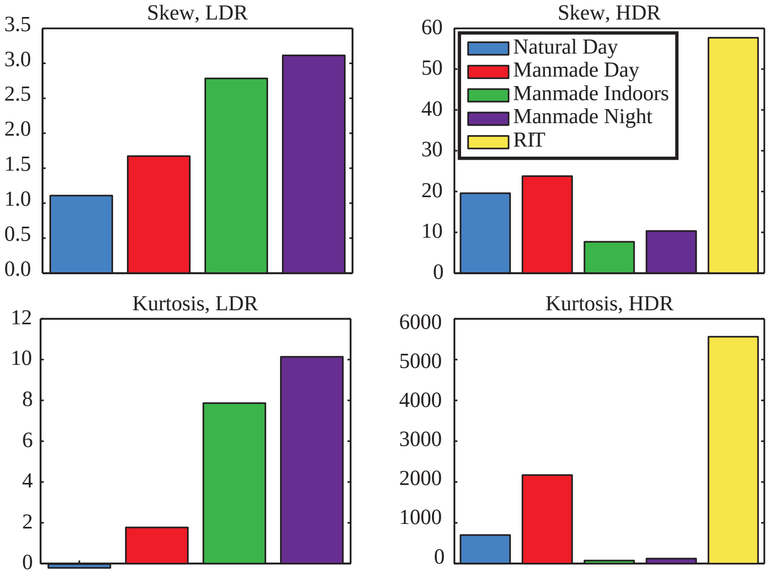


Figure 4.10. Skewness and kurtosis values for the histograms in Figure 4.9.

Although the two capture methods lead to measurable differences, it is important to note that only qualitative conclusions can be drawn from this data. As the HDR images are linear but not radiometrically calibrated, their pixel values are correct up to a scaling factor, which may be different for each image. Consequently, to draw quantitative conclusions regarding distributions of radiances in real scenes, a fully calibrated HDR set would be necessary. To some extent, this is fulfilled by the HDR Photographic Survey [203]. However, the scenes in this collection were selected with the aim of depicting both deep shadows and bright highlights within the same image and thus their distributions may be biased.

4.3 Material Properties

Histograms of image categories, as we have seen, already hint at differences between scene types. In the general scenarios we have discussed so far, however, they do not offer sufficient discriminative power. For instance, although the average distributions for natural and manmade scenes may have measurable differences (e.g., Figure 4.9), it would not be possible to accurately categorize the content of any given image as one or the other based solely on its histogram. If we shift our focus to more specific image classes or properties, however, histograms and their moments have been found to be more descriptive.

One notable example is the analysis of materials [225, 226, 227, 516]. Human observers are very capable of making accurate judgements about material properties of objects and surfaces around them. Although the appearance of objects depends on complex interactions between light, material properties, and geometry, psychophysical evidence suggests that simple cues may be sufficient for assessing whether a surface is matte or glossy [41]. It is reasonable to expect that such cues may be directly or indirectly linked to statistical properties of images. For instance, Figure 4.11 shows renderings of the same object but with different material properties and their corresponding histograms. The shape of the histogram changes significantly with material changes even though both the lighting and geometry remain constant in the three images.

To determine whether a link between the perception of materials and statistical proprties exists, Motoyoshi et al. [517] studied the perceived surface qualities of various materials and analyzed their relation to the associated histograms and specifically their skewness. For patches of simple materials with mesostructure, such as stucco or crumpled paper, they found a correlation between histogram skewness and specular intensity, while an inverse relation was observed between skewness and the diffuse reflectance of the material. In addition to the physical properties of the materials, perceptual attributes of lightness and glossiness were also measured. Figure 4.12 shows the perceptual attributes and physical material properties in relation to the image skewness for a stucco material, while Figure 4.13 gives the skewness for the images in Figure 4.11.

Although skewness measurements correlate with material glossiness both physically and perceptually, it is important to note that such correlation only holds when certain assumptions are present [14]. Crucially, images are expected to depict surfaces of nearly constant *albedo* (the proportion of incident light reflected by a surface) and illumination. If these assumptions are violated, skewness

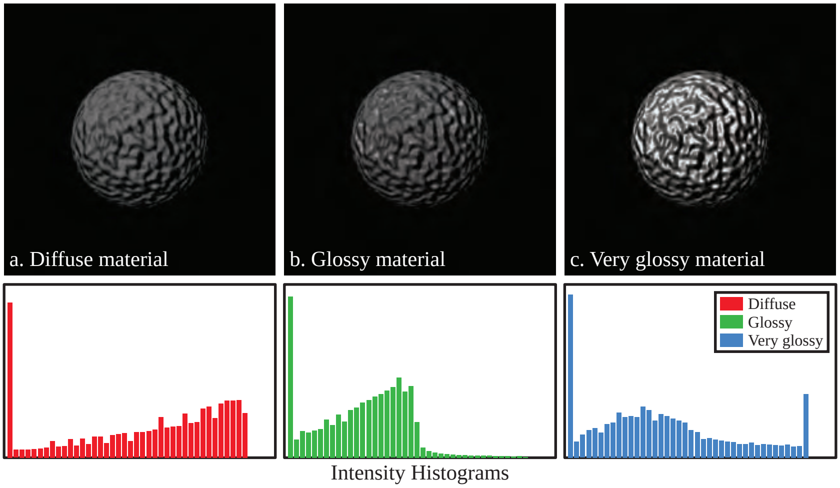


Figure 4.11. The three images were generated in a 3D modeling program using different properties of the Phong material so that they look progressively glossier. Their log intensity histograms are shown in the bottom row. The histograms were computed only on the pixels within the sphere.

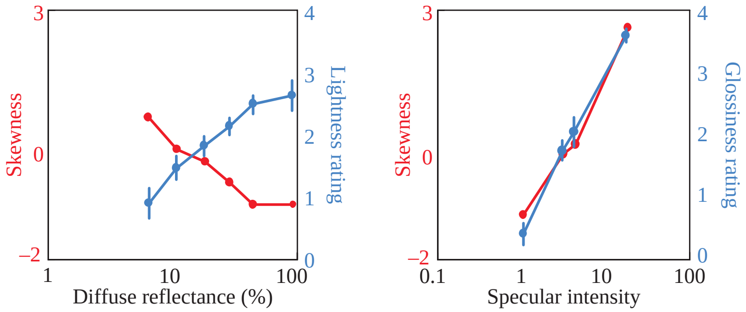


Figure 4.12. As the diffuse reflectance of the surface increases, the lightness rating as perceived by the observers also increases while the corresponding histogram skewness decreases (left). An increase in specular intensity also results in an increased rating of glossiness as well as higher skewness value (right). (Adapted from [517].)

measurements will likely not be indicative of specific material properties. This is demonstrated in Figure 4.14, where the image of the diffuse object (Figure 4.11a) is inverted, therefore changing the skewness of its histogram to a value close to that of Figure 4.11b ($s = 0.3273$). Despite the increased skewness, the object does not appear glossier, indicating that other aspects need to be considered, such as additional histogram moments or percentile statistics [670].

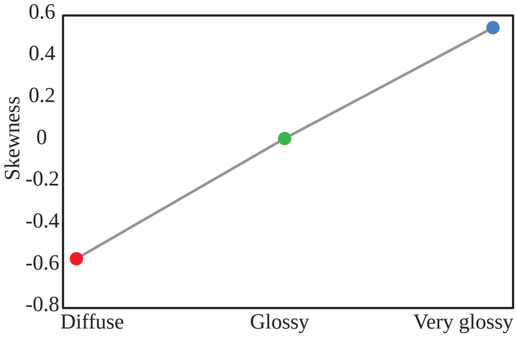


Figure 4.13. Skewness measurements from the images shown in Figure 4.11. Similar to the findings of Motoyoshi et al. [517], skewness increases proportionally to the glossiness of the material.

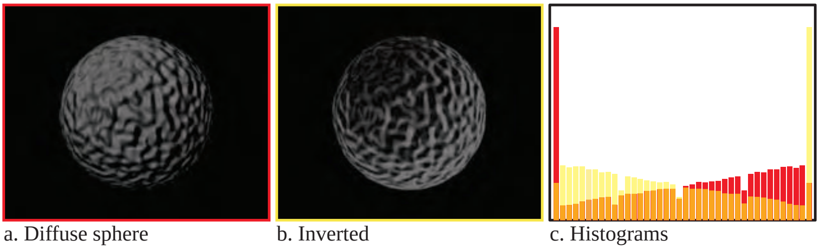


Figure 4.14. The image of the diffuse object (a) is inverted to produce (b). Although this process changes the skewness of the distribution as the corresponding histogram is mirrored (c), it does not change the perceived glossiness of the object. Instead, the object appears as if it is illuminated from the opposite direction, indicating that skewness alone may not be a sufficient indicator for material glossiness.

4.4 Nonlinear Compression in Art

First-order statistics have also seen very widespread use in the study of paintings. Paintings form a very interesting category of images as they are generally abstracted or manipulated representations of real scenes, purposefully created for human viewers. Different creative styles can represent the same scene in totally different ways—from almost realistic to completely abstract—eliciting different reactions and emotions. Yet, despite art often representing a reduced, abstracted, or modified version of reality, we are not only capable of visually processing it but also find it aesthetically pleasing and engaging.

In an effort to understand the regularities in art and how the compare with real scenes, first-order (as well as higher-order) statistics have been extensively used to examine different properties of art [273, 274, 275, 277, 276, 278]. This included various epochs of art, including properties of the illuminants, differences between epochs or cultures, and the relationship between image statistics and either the human visual system, personal preference, or real-world scenes.

One particularly interesting aspect of art studied in this work is the nonlinearity of intensities found in paintings compared to real-world illumination and the nonlinearities in the visual system [277, 275, 273]. As we have seen in Section 3.1.2, the range of illumination in natural scenes far exceeds the capabilities of modern cameras and displays, requiring nonlinear compression schemes to preserve detail in the scene while reducing its dynamic range to fit within more restricted devices. The human visual system faces a similar problem, as discussed in Section 2.3.2, employing a sigmoidal compression in the photoreceptors [525].

Unsurprisingly, artists are also confronted with this issue, especially given the limited range that can be represented with paints on a canvas. Graham and colleagues compared a set of paintings to calibrated photographs of natural scenes to determine how the nonlinearity employed in paintings relates to natural scene illumination and the responses of photoreceptors [273]. Images in both sets were processed with a difference-of-Gaussians (DoG) filter, simulating ganglion responses, and with Gabor filters, simulating cortical responses. Further, images were analyzed both in linear and logarithmic (simulating the photoreceptor responses) domain. Skewness and kurtosis statistics were collected for each transform and image class (art or natural).

Interestingly, paintings tend to have a significantly lower skewness and sparseness than natural scenes, both when pixels are analyzed directly and after filtering with a DoG in the linear case, but this relation reverses in the log domain, as shown in Figure 4.15. As the paintings already represent a compressed version of real-world intensities, they are less affected by the effect of the logarithmic compression [273]. Although nonlinearities in the way light is represented in art are common, no single function was found capable of modeling this compression across all art [274], suggesting that much like tonemapping photographic images of high dynamic range, no single solution will work for all scenes [99].

In addition to learning about artists' nonlinear compression, art has also been analyzed using statistics to estimate elements involved in the process of painting, including the position of light sources in Caravaggio-style paintings, whether optical projections were used, the position and properties of a virtual camera that would have taken the picture if it were a photograph, and a 3D version of the scene [135, 382, 710, 826]. Finally, a number of researchers in addition to Graham have examined distinguishing art from different periods from either each other or from real scenes [770, 146, 451, 496] or to digitally restore art [556] using first-order statistics combined with higher-order statistics. For a more detailed review of this area, we refer the reader to [233].

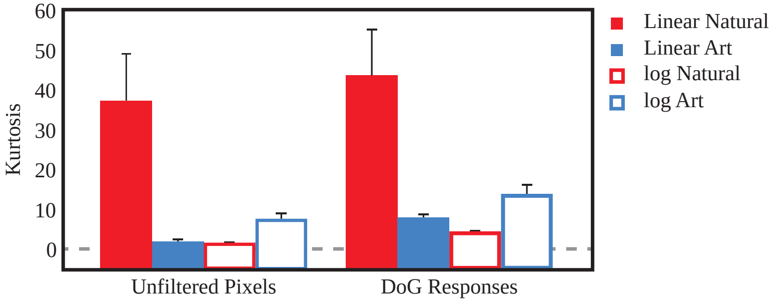


Figure 4.15. Natural scenes exhibit much higher sparseness (kurtosis) in the linear case. However if intensities are logarithmically compressed, artistic images have sparser distributions than natural scenes. (Adapted from [273].)

4.5 Dark-Is-Deep Paradigm

Photographs can capture in a two-dimensional array of pixels the appearance of a complex scene. Looking at an image, we can easily form an understanding about the shape of the depicted objects, the illumination, or the materials within the image. In the previous section we have seen how simple histogram statistics may be linked with material appearance in images. Substantial effort has been devoted to algorithms that can similarly recover information about the geometry and the estimation of the light sources present in the scene. Depth and shape, however, have been linked both perceptually and in terms of image reconstruction to the intensities of image pixels [743, 599, 438].

The Shape from Shading (SFS) problem was first formulated by Horn in the context of computer vision and can be understood as the extraction or recovery of 3D geometry from a 2D image using shading information [340]—essentially relating luminance with depth. The intuitive simplicity of this idea has led to a remarkable amount of follow-up research, spanning several decades [819]. Although many of the SFS methods perform well in restricted situations, where the materials and illumination are well controlled, estimations tend to be less accurate in more general scenarios.

Human vision also relies on certain assumptions about the scene to interpret the shape of the objects. Mathematically, without prior information about the illumination in a scene, it is impossible to determine the geometry of an object. Consider, for instance, the spheres in Figure 4.16; without knowledge of the location of the illumination we cannot accurately determine whether the shapes are convex or concave. Despite this ambiguity, the third shape is perceived as concave while the others appear convex. On the other hand, if all shapes are rotated by 90 degrees as shown in Figure 4.17 such a distinction cannot be made.

One of the main priors that the human visual system uses to determine shape is the “light from above” assumption [599] (although evidence suggests that a

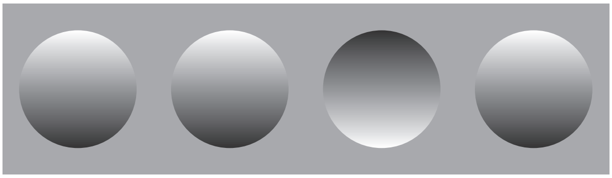


Figure 4.16. Although it is impossible to accurately determine whether these objects are concave or convex from this viewpoint, the third sphere is perceived as concave while the remaining objects are seen as convex. Our visual system employs priors to make such assessment relating both to the direction of the illumination and global convexity.

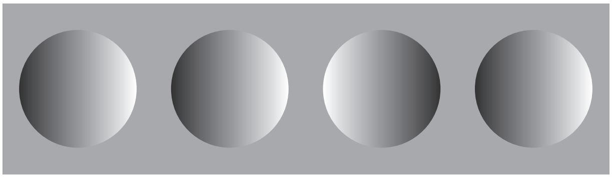


Figure 4.17. Unlike the objects in Figure 4.16, it is harder to assess the convexity of these objects as illumination appears to come from either side.

left bias also exists [492, 506]). This is not surprising, as our visual system has evolved both genetically and personally in environments where illumination typically comes from above. Further, Langer et al. have found evidence that a global convexity assumption exists when assessing shape [439].

Although in such simple examples image intensity appears to be related to depth and geometry—at least to the extent that it conforms to the assumptions discussed—natural scenes consist of much more complex configurations (varying material properties, structure at different scales, translucency, and so on). To derive links between 2D image content and 3D depth, more complex statistics need to be considered. A detailed discussion of depth and range statistics and their connection to 2D statistical regularities is given in Chapter 11.

4.6 Summary

Statistical analysis and manipulation of image histograms can be used to derive powerful descriptors. Using well-chosen assumptions, first-order statistics have been linked to depth in scenes as well as material properties. However, these statistics do not capture spatial relations between pixels as the position of pixels is not considered. Figure 4.18 shows an example of two very different images that would result in identical first-order statistics when spatial information is ignored. The right image is constructed by randomly permuting the pixels of the left image, yet it appears unnatural—it carries no recognizable information.

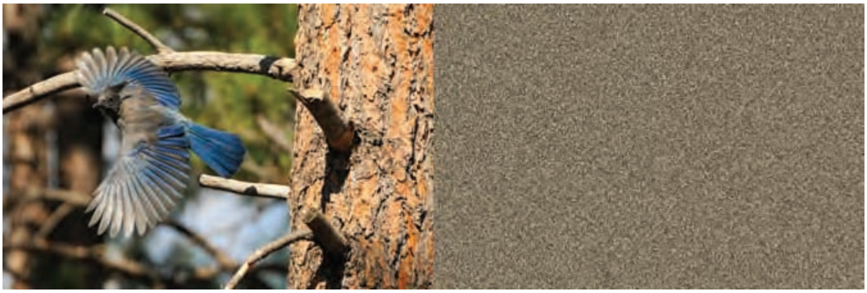


Figure 4.18. The right image was created by randomly permuting the pixels of the image on the left, resulting in identical first-order statistics. (Bryce Canyon, USA, 2009)

As they only assess single-pixel properties, the histogram-based statistics discussed in this chapter are the simplest to compute and interpret. To further explore the spatial relations between pixels and their associations with aspects of human vision, more complex transforms are necessary. These will be discussed in the following chapters.

Chapter 5

Gradients, Edges, and Contrast

The information contained within a single pixel is essentially just a color, which is normally represented with three values. As discussed in Chapters 4 and 10, we can derive useful information from the analysis of individual pixel values. We can even perform some very elegant image manipulations with the information contained in single pixels. Nonetheless, the information obtained by analyzing single pixels is still very limited. In particular, a single pixel cannot, in principle, tell us anything about the structures depicted in the image. Does the image, say the one in Figure 4.18, depict a bird? Is there a tree present, and if so, where does it start or stop? Is the tree closer to us than the bird or farther away? The information present in a single pixel is not enough to answer such questions.

As Figures 1.1 and 4.18 demonstrated, simply rearranging the location of the pixels completely eliminates the visible structure. While this demonstration again emphasizes that single pixels by themselves do not tell us anything about structure, it also suggests a way to get at structural information. Pixels do in fact have one more critical piece of information: their relative location. The change in color or luminance between pixels placed at specific locations is the key to discovering more about the content of an image.

5.1 Real-World Considerations

Since we usually try to infer something about the real-world scenes depicted in images, it will be worthwhile to spend a few minutes thinking about the real world. From a physical point of view, every point of humanity's natural environment is filled with some form of matter. Some of this matter is in solid form (e.g., rocks, trees, glass), some in liquid form (e.g., ink, milk, water), and sometimes in gaseous form (e.g., fog, clouds, air). We tend to call spatially localized regions of similar material "objects," with gaseous media such as air being a special case of object. In other words, the real world is completely filled with objects. Thus, the surface of one object *always* abuts against the surface of another object, which means that a surface is really just a transition region between two kinds of matter.

5.1.1 Perceptual Consequences

Light is structured by its interaction with objects. In most cases, the surface of an object (and occasionally, such as in subsurface scattering, a very short distance under the surface) is solely responsible for the altering of the properties of light rays [610]. Since objects are spatially localized regions of similar material, neighboring points on an object's surface will tend to affect light in a similar way. Likewise, different objects are usually made of different materials and thus will tend to affect light in different ways.

Perceptually, then, it can be said that the visual world consists solely of surfaces, the properties of which we infer from how they structure light [263]. As a simple example, we can reverse the observations given earlier: if two neighboring points affect light in the same way, it is very likely that they come from the same object. If they affect light (significantly) differently, then they probably belong to different objects.

5.1.2 Image Space Consequences

Obviously, images are constructed from light rays that have previously interacted with real-world objects. It is not surprising, then, that some of the same light structures used by the visual system to infer properties of the real world are also present in images. For an image, therefore, we have the observation that two-dimensional images consist entirely of surfaces, and the place where two surfaces meet is called an edge. If two neighboring points in an image affect light in the same way, it is very likely that they come from the same surface. If they affect light (significantly) differently, then they probably belong to different surfaces. Thus, by looking for strong changes in luminance or color between neighboring points, we can find edges in an image.

The next part of this chapter examines methods for looking at luminance changes. Following this, methods for looking specifically at edges will be examined.

5.2 Gradients

There are two common methods for determining the magnitude of the change in the luminance function: using the first derivative and using the second derivative. The simplest approach to calculating the first derivative is to compare a given pixel with some subset of its neighbors, which provides a description of how the luminance in the image changes as a function of spatial direction. There are four primary methods for calculating a discrete approximation of the gradient: forward difference, backward difference, central difference, and the Söbel operator. The first three can be derived from a Taylor polynomial around the pixel of interest x.

5.2.1 The Forward Difference Method

To derive an estimate of the gradient at a given pixel location, we can start with a Taylor polynomial, where we try to calculate the value at the next point in a function based on the current value and the derivatives of the function at the current point. More specifically, we can employ a weighted sum of the derivatives, which in one spatial dimension leads to:

$$I(x+h) = \sum_{n=0}^{\infty} \frac{I^{(n)}(x)}{n!} h^n \tag{5.1a}$$

$$= I(x) + hI\ (x) + \frac{h^2}{2!} I\ \ (x) + \ldots \tag{5.1b}$$

where $I(x)$ is the value at pixel x, $I(x+h)$ is the value at $x+h$ with h being some constant, and $I^{(n)}(x)$ is the n^{th}-order derivative of I evaluated at point x. Rewriting Equation (5.1b) to isolate the first derivative $I\ $ yields:

$$I\ (x) = \frac{I(x+h) - I(x) - \sum_{n=2}^{\infty} \frac{h^n}{n!} I^{(n)}(x)}{h} \tag{5.2}$$

Given that gradients are effectively measures of the derivative of a function, we could ignore the higher-order terms, and rather than compute derivatives, we can compute gradients. If we additionally assume that the step between pixels h is 1, then we obtain the forward difference method for calculating the gradient $\mathbf{D} = (D_x, D_y)$ at a pixel (see Figure 5.1). Since we are dealing with a two-dimensional image, horizontal and vertical differences are calculated separately:

$$D_x(i,j) = I(i+1,j) - I(i,j) \tag{5.3a}$$

$$D_y(i,j) = I(i,j+1) - I(i,j) \tag{5.3b}$$

where $I(i,j)$ is the luminance for pixel (i,j), D_x is the horizontal gradient and D_y is the vertical gradient. Commonly, gradients are computed in log space, which has the advantage that they represent contrasts, as pixel ratios become pixel differences:

$$D_x(i,j) = \log I(i+1,j) - \log I(i,j) \tag{5.4a}$$

$$D_y(i,j) = \log I(i,j+1) - \log I(i,j) \tag{5.4b}$$

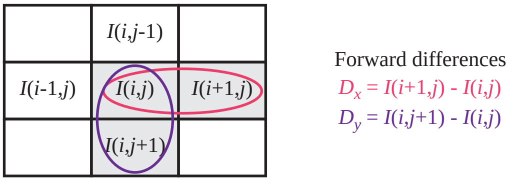


Figure 5.1. Forward difference method for calculating gradients.

We can construct a pair of convolution kernels to represent the forward difference operator:

$$\frac{\partial}{\partial x} = \begin{bmatrix} -1 & 1 \end{bmatrix} \tag{5.5a}$$

$$\frac{\partial}{\partial y} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{5.5b}$$

A vector valued gradient image (dI_x, dI_y) can then be computed by applying these convolution kernels to the input image I in log space:

$$dI_x = \log I * \frac{\partial}{\partial x} \tag{5.6a}$$

$$dI_y = \log I * \frac{\partial}{\partial y} \tag{5.6b}$$

It is important to note that the high-order terms are omitted. This means that this method yields only an approximation of the first derivative—albeit a rather accurate one for small values of h. The error can be estimated from the truncated terms and has the order $O(h)$.

5.2.2 The Backward Difference Method

The backward difference method is essentially the same, with the exception that the left pixel $I(i-1, j)$ is compared to the pixel of inters $I(i, j)$ rather than $I(i+1, j)$ (see Figure 5.2). Since the step h between pixels is now negative, every odd-powered derivative in the series also becomes negative:

$$I(x-h) = \sum_{n=0}^{\infty} \frac{I^{(n)}(x)}{n!}(-h)^n \tag{5.7a}$$

$$= I(x) - hI\ (x) + \frac{h^2}{2!}I\ (x) - \frac{h^3}{2!}I\ (x) + \ldots \tag{5.7b}$$

Rewriting Equation (5.7b) to isolate the first derivative $I\ $ yields:

$$I\ (x) = \frac{I(x) - I(x-h) - \sum_{n=2}^{\infty} \frac{(-h)^n}{b!} I^{(n)}(x)}{h} \tag{5.8}$$

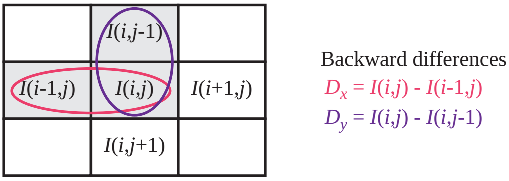


Figure 5.2. Backward difference method for calculating gradients.

The gradients are therefore computed in horizontal and vertical directions using:

$$D_x = \log I(i,j) - \log I(i-1,j) \tag{5.9a}$$

$$D_y = \log I(i,j) - \log I(i,j-1) \tag{5.9b}$$

where, similar to the previous section, we have also added the log space computation.

The convolution kernel for the backward difference operator is the same as for the forward difference. The two operators differ merely in which of the cells is the pixel of interest. Just as with the forward difference, the high-order terms are truncated. The error, then, still has the order $O(h)$.

5.2.3 The Central Difference Method

Combining the forward and the backward difference methods yields a more accurate and robust technique called the central difference method, which is illustrated in Figure 5.3:

$$\begin{aligned} I(x+h) - I(x-h) &= \left(I(x) + hI\ \,(x) + \frac{h^2}{2!}I\ \ (x) + \frac{h^3}{3!}I\ \ \ (x) + \ldots \right) && (5.10a) \\ &\quad - \left(I(x) - hI\ \,(x) + \frac{h^2}{2!}I\ \ (x) - \frac{h^3}{3!}I\ \ \ (x) + \ldots \right) \\ &= 2\left(hI\ \,(x) + \frac{h^3}{3!}I\ \ \ (x) + \ldots \right) && (5.10b) \end{aligned}$$

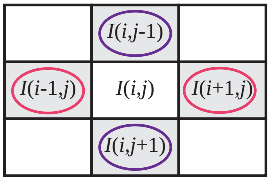
Rewriting Equation (5.10b) to isolate the first derivative $I\ $ yields:

$$\frac{I(x+h) - I(x-h)}{2} = hI\ \,(x) + \frac{h^3}{3!}I\ \ \ (x) + \ldots \tag{5.11}$$

Ignoring the higher-order terms, assuming h is 1, moving into log space, and computing the horizontal and vertical gradients separately yields:

$$D_x(i,j) = (\log I(i+1,j) - \log I(i-1,j))/2 \tag{5.12a}$$

$$D_y(i,j) = (\log I(i,j+1) - \log I(i,j-1))/2 \tag{5.12b}$$

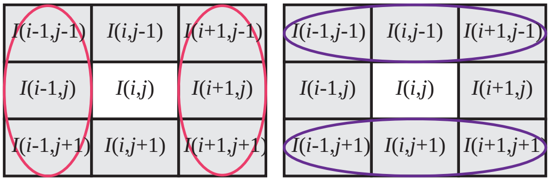


Central differences

$D_x = (I(i+1,j) - I(i-1,j))/2$

$D_y = (I(i,j+1) - I(i,j-1))/2$

Figure 5.3. Central difference method for calculating gradients.


$I(i\text{-}1,j\text{-}1)$
$I(i,j\text{-}1)$
$I(i\text{+}1,j\text{-}1)$
$I(i\text{-}1,j)$
$I(i,j)$
$I(i\text{+}1,j)$
$I(i\text{-}1,j\text{+}1)$
$I(i,j\text{+}1)$
$I(i\text{+}1,j\text{+}1)$
$I(i\text{-}1,j\text{-}1)$
$I(i,j\text{-}1)$
$I(i\text{+}1,j\text{-}1)$
$I(i\text{-}1,j)$
$I(i,j)$
$I(i\text{+}1,j)$
$I(i\text{-}1,j\text{+}1)$
$I(i,j\text{+}1)$
$I(i\text{+}1,j\text{+}1)$

Söbel operator

$D_x = I(i+1,j-1) + 2\,I(i+1,j) + I(i+1,j+1)$
$\quad - I(i-1,j-1) - 2\,I(i-1,j) - I(i-1,j+1)$

$D_y = I(i-1,j+1) + 2\,I(i,j+1) + I(i+1,j+1)$
$\quad - I(i-1,j-1) - 2\,I(i,j-1) - I(i+1,j-1)$

Figure 5.4. Söbel operator for calculating gradients.

We likewise construct a pair of convolution kernels for the central difference operator:

$$\frac{\partial}{\partial x} = \begin{bmatrix} -1/2 & 0 & 1/2 \end{bmatrix} \tag{5.13a}$$

$$\frac{\partial}{\partial y} = \begin{bmatrix} -1/2 \\ 0 \\ 1/2 \end{bmatrix} \tag{5.13b}$$

These convolution kernels can then be applied according to Equations (5.6a) and (5.6b) to compute the gradient image.

Since the second derivative term in the forward difference method cancels out the second derivative term in the backward difference method, the error is of order $O(h^2)$. Although the central difference method is more accurate, it has the disadvantage that oscillating functions can yield a zero derivative.

5.2.4 The Söbel Operator

The forward and backward difference methods both incorporated a single neighboring pixel into the computation. The central difference examines two neighbors and is more accurate. The Söbel operator additionally includes diagonal neighbors (Figure 5.4) and is the most accurate of the basic methods. To calculate a gradient using the Söbel operator, convolution kernels with the following weights

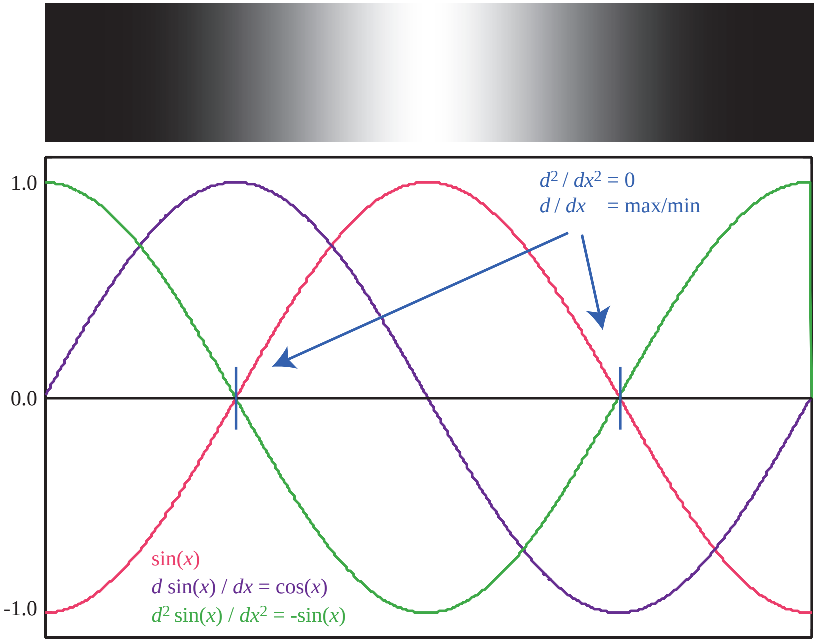


Figure 5.5. A sinusoidal texture gradient. The luminance function and its first and second derivatives (both normalized to [–1, 1]) are plotted. Note that where the second derivative is 0 the luminance function has maximum variation, i.e., the first derivative is either at its maximum or minimum.

are applied to the image:

$$\frac{\partial}{\partial x} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{5.14a}$$

$$\frac{\partial}{\partial y} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{5.14b}$$

Once again, these kernels are convolved with the logged input image to produce the vector-valued gradient image as per Equations (5.6a) and (5.6b).

5.2.5 Second Derivative Methods

Intuitively, the place of greatest change in luminance will be important. The places of greatest change in the luminance function correspond to the maxima in the first derivative. Since it is not always easy in practice to uniquely identify maxima, many applications take advantage of the fact that maxima in the first derivative will show up as zero crossings in the second derivative. Figure 5.5 visualizes the relationship between the luminance function and its derivatives. The top of the figure shows a simple horizontal sinusoid. The lower part of the graph shows the luminance function, as well as its first and second derivatives. The vertical bars represent the location of maximal luminance changes.

There are several common ways of estimating the second derivative. The most common is the Laplacian, which is the divergence of the gradient operator. The Laplacian ∇^2 can be calculated as $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$, or as a single convolution kernel, applied to the image:

$$\nabla^2 I = I * \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{5.15}$$

This is a 2D convolution, which is separable, meaning that it can also be written as the sum of two 1D convolutions. In that case, a straightforward pair of convolution kernels would be:

$$\frac{\partial^2}{\partial x^2} = \begin{bmatrix} 1 & -2 & 1 \end{bmatrix} \tag{5.16a}$$

$$\frac{\partial^2}{\partial y^2} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \tag{5.16b}$$

The Laplacian of an image is then given by:

$$\nabla^2 I = I * \frac{\partial^2}{\partial x^2} + I * \frac{\partial^2}{\partial y^2} \tag{5.17}$$

Wherever the Laplacian of an image becomes small, the gradient of the image will be either at a maximum or a minimum.

5.2.6 Gradient Magnitude

In all cases, it is common to ignore the direction of the gradient, and instead calculate the mean gradient magnitude at a given location:

$$D(i,j) = \sqrt{D_x(i,j)^2 + D_y(i,j)^2} \tag{5.18}$$

In some cases, however, one may choose to keep the vertical and horizontal gradient magnitudes separate (as, for instance, in Levin's motion deblurring technique [453]):

$$D(i,j) = \begin{bmatrix} D_x(i,j) \\ D_y(i,j) \end{bmatrix} \tag{5.19}$$

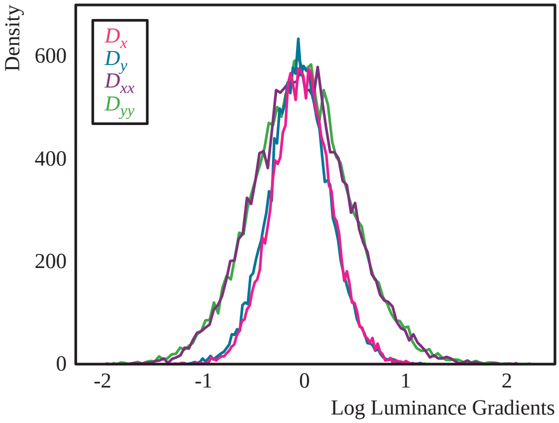


Figure 5.6. Gradient distributions for a collection of natural images: D_x and D_y are the horizontal and vertical gradients (first derivative of the luminance function), respectively, and D_{xx} and D_{yy} are the horizontal and vertical second derivatives [586].

5.2.7 Gradient Statistics

Although gradients can carry considerable information, especially about the texture of objects and surfaces, little is known about the statistics of gradients in natural images. It is known, however, that the gradient histogram has a very sharp peak at zero (see Figure 5.6) and then falls off very rapidly (with long tails at the higher gradients) [639, 637, 352, 350, 182, 185]). The distribution can be modeled as:

$$\exp(-x^{\alpha}) \tag{5.20}$$

with $\alpha < 1$ [457, 489, 684]. The reason for the specific shape of the gradient distribution seems to be precisely the observations outlined earlier: images contain many large surfaces that tend to be smooth and somewhat homogeneous. Such surfaces will have gradient magnitudes near zero. There will, of course, be a few high-contrast edges, which will yield very high gradients [32]. Interestingly, similar objects tend to cluster together, which means that the transition from one object to another will be similar in any direction. This is reflected in the symmetry of the gradient distribution around the central peak. It has also been shown that the gradient histogram is roughly invariant to changes in scale [215, 350, 639].

5.2.8 Single-Image Gradient Statistics

Instead of assessing average statistics over ensembles, it is possible to compute gradient statistics on a single image. Specifically, it is interesting to look at small patches and assess the probability of such patches recurring elsewhere in the image [829]. To characterize a 5 × 5 patch **p** centered at pixel location (x,y), first its mean gradient magnitude $\bar{D}_{\mathbf{p}}$ is computed:

$$\bar{D}_{\mathbf{p}}(x,y) = \sum_{i=-2}^{2} \sum_{j=-2}^{2} D(i+x, j+y) \tag{5.21}$$

Then, an empirical density d is computed for each patch within a neighborhood of radius r around the patch. This neighborhood has an area $A = \pi r^2$:

$$d(\mathbf{p};r) = \sum_{\mathbf{p}_j \quad A} \frac{G(\quad \mathbf{p}-\mathbf{p}_j \quad {}^2_2)}{A} \tag{5.22}$$

where G is a Gaussian kernel. We can then calculate an average density for each mean gradient magnitude:

$$\bar{d}(r, D_{\mathbf{p}}) = \frac{\sum_{D_{\mathbf{p}_j}=D_{\mathbf{p}}} d(\mathbf{p}_j;r)}{n} \tag{5.23}$$

where n is the number of elements over which we sum. The average number of matching nearest neighbors N with a distance r can then be computed as:

$$N(r, D_{\mathbf{p}}) = A\, \bar{d}(r, D_{\mathbf{p}}) \tag{5.24}$$

Calculations such as these, if carried out on a large set of images, reveal that within each image, smooth patches occur frequently, whereas structured patches, i.e., those with sharp gradients, occur less often. Moreover, patches tend to occur in clusters, meaning that a given patch is more likely to occur nearby than farther away. However, the distance over which a patch has a matching counterpart depends on its gradient content. In particular, the higher the mean gradient magnitude, the larger the search space r should be. This notion has been empirically quantified with the following expression, relating search radius to mean gradient magnitude:

$$r(D_{\mathbf{p}}) = a + b \exp\left(D_{\mathbf{p}}/10\right) \tag{5.25}$$

It was found that the variables a and b depend on the average number of matching neighbors N as follows:

$$a = 5 \cdot 10^{-3} N + 0.09 N^{0.5} - 0.044 \tag{5.26}$$

$$b = 7.3 \cdot 10^{-4} + 0.3234 N^{0.5} - 0.35 \tag{5.27}$$

The interpretation of these equations is that a patch has a certain amount of contrast, quantified by its mean gradient magnitude $D_{\mathbf{p}}$. The search radius to find a given number N of matching patches for this patch is then given by Equation (5.25).

Conversely, for a given patch and search radius, it is possible to calculate the number of matching neighbors N that are likely to be found:

$$N(r, D_{\mathbf{p}}) = \left(\frac{-b + \ \overline{b^2 - 4ac}}{2a} \right)^2 \tag{5.28}$$

where

$$a = 0.001(5 + 0.73 \exp(D_{\mathbf{p}}/10)) \tag{5.29}$$

$$b = 0.1(0.9 + 3.24 \exp(D_{\mathbf{p}}/10)) \tag{5.30}$$

$$c = -0.1(0.44 + 3.5 \exp(D_{\mathbf{p}}/10) + r) \tag{5.31}$$

Applications which might benefit from knowing the search radius to find similar patches include, for example, image denoising [829]. The Non-Local Means denoising algorithm replaces the center pixel in a patch with the mean value of center pixels found in similar patches elsewhere in the image [84]. The standard algorithm limits the search space to a relatively small and fixed neighborhood. By making the search space adaptive as described earlier, however, it is possible to find better candidate patches and therefore obtain better results. More recently, it was shown that denoising results may benefit from a combination of single-image statistics and ensemble statistics [515].

It was also discovered that matching patches can be found at different spatial scales. In other words, an image can be repeatedly downsampled by some factor, leading to a stack of images (an image pyramid). Searching matching patches between scales will find strong candidates in nearby scales at nearby locations. In essence, this means that patches are self-similar across scales. This is a concept that will return several times throughout this book, in particular with the discussion of wavelets in Chapter 8. Local self-similarity, detected within a single image, has been used in an image and video upscaling application [245].

Finally, we note that the likelihood of similar patches occurring drops off with distance. This is also seen with single pixel values, which can be computed with the autocorrelation function, or equivalently with the Fourier transform. Such correlations are the topic of Chapter 6.

5.3 Edges

As stated in Section 5.1.2, an edge is the meeting of two surfaces. It follows, then, that an edge contains information about both surfaces and as such provides a highly localized, very dense set of insights into the surfaces—or objects—on either side.

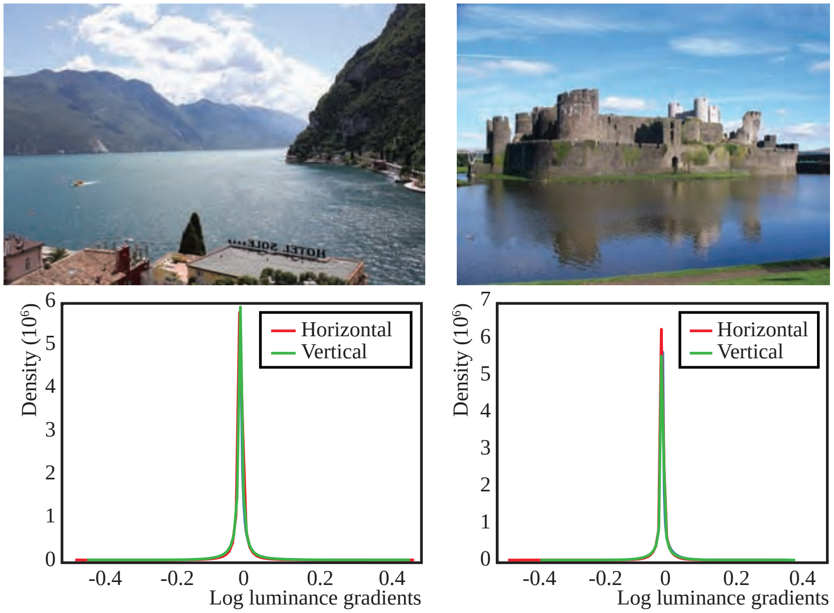


Figure 5.7. Two color images and the gradient histograms of grayscale versions of the images, calculated using the central difference method. (Lago di Garda, Italy, 2012; Caerphilly Castle, UK, 2010)

5.3.1 Definition of an Edge

The term *edge* has many definitions, which can differ considerably for different application domains. Given the observations already mentioned, though, it is clear that an edge is essentially a region of rapid or even maximal change in some underlying function. Quite often, edges are regions where there is a discontinuity. For the real world, the function can be seen as the type of matter across spatial location, and the edges of an object—its surface—are discontinuities in type of matter. For images, the underlying function is image luminance or color across spatial location, and edges are large, sudden changes in luminance. For light field photography, the underlying function is either radiance or intensity across light rays [256]. For 3D volume recordings, the function is often material density across spatial location. Note that it is even possible to extend this definition of edges to concepts, with edges in conceptual space. For our present purposes, however, we will restrict ourselves to color or luminance magnitude functions (Figure 5.7).

5.3.2 Edge Detection Processes

Since both the first derivative and second derivative methods for calculating gradients examine neighboring pixels to determine how much the luminance function

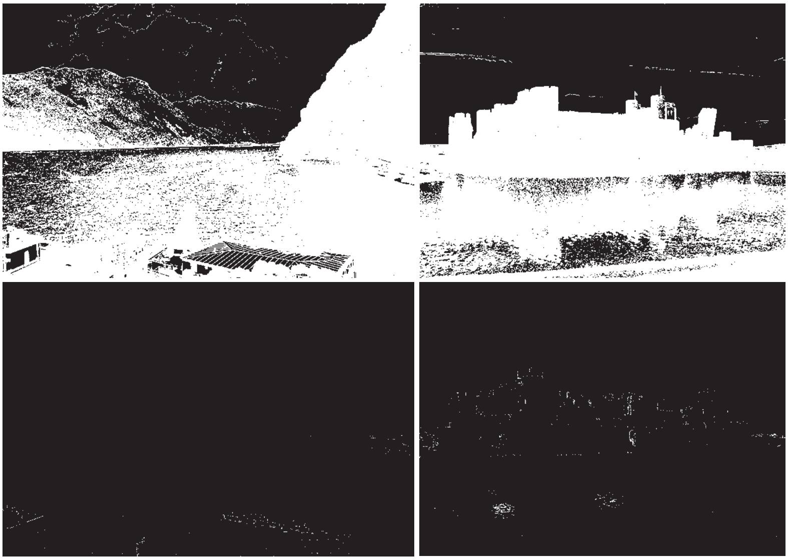


Figure 5.8. The gradient magnitudes that are above a low (top) or high (bottom) threshold are shown in white.

changes (making them very local), they would seem to be natural choices for discovering edges. All we need to do is to decide precisely what we mean by a *large* change in luminance. At first blush, this would seem to be simple: we simply look for zero crossings if we are using a second derivative or really big gradients if we are using a first derivative. In practice, it is not so easy. For example, we could take the two images from Figure 5.7. Both images have large, relatively homogenous regions such as lakes, the sky, and buildings. There are also a few high-contrast edges, such as the edges of the mountains or the silhouette of the buildings. The gradient histograms for these pictures are shown in the second row of Figure 5.7 with a focus on the gradients with strengths between -0.5 and 0.5. As expected, there is a sharp peak at zero, confirming that there are many pixels in relatively homogeneous regions. The plots also exhibit heavy tails, showing that there are some strong edges.

Figure 5.8 shows the effects of two different threshold values (a low threshold in the top row and a higher threshold in the bottom row. Neither result is really satisfactory. The top row has too many edges, especially in the water, mountains, and buildings. The bottom row, on the other hand, is missing large, important edges (such as part of the castle or mountains) and yet still has some edges in the water. Although there are many reasons for these and similar difficulties, one central issue is that no surface is perfectly homogeneous. Thus, small local inhomogeneities will lead to small local fluctuations in luminance and therefore in the first derivative. This is particularly true for textured objects. Likewise, the presence of noise will lead to spurious, large changes in the gradient.

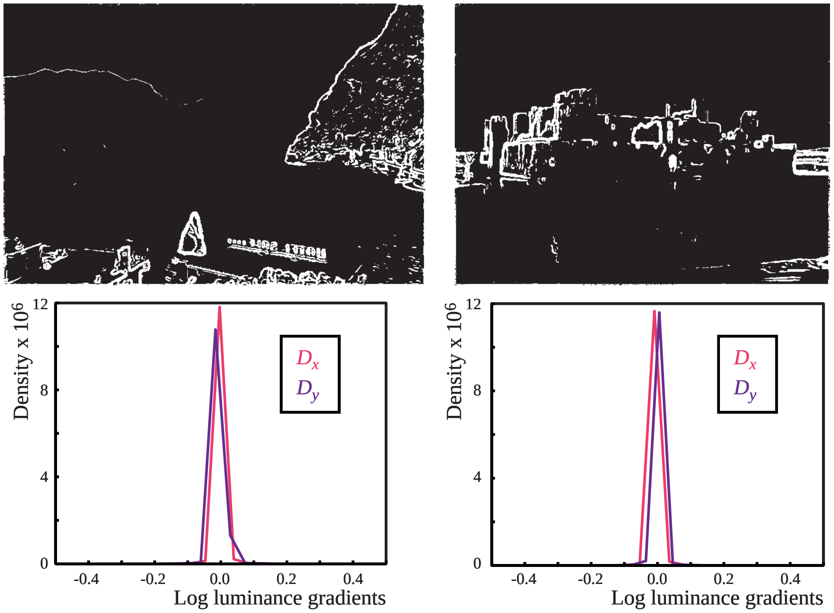


Figure 5.9. The effect of filtering. The gradients (top) that are above the same low threshold used in Figure 5.8 and the gradient histograms (bottom) are shown for the two images after they have been smoothed with a Gaussian filter.

To help reduce the effect of noise and highly local edges, many edge detection procedures add a number of steps and constraints. The first step that is often taken is to reduce the local fluctuations by preprocessing the image with a local smoothing operator like a Gaussian filter. The top row in Figure 5.9 shows the gradient histograms of the two images from Figure 5.7 after they have been smoothed with a Gaussian filter. The bottom row shows the gradients in the image that are above the same low threshold used in the top row of Figure 5.8. As expected, the histograms have become even more spiked around zero, leading to higher kurtosis.

Finally, if we remember that edges are merely the meeting places of two surfaces or objects and that an object is a spatially localized region of similar matter, it becomes clear that edges have another useful property: they are spatially extended. In other words, edges will cover many neighboring pixels, especially if they are important edges. Thus, one can use the presence or absence of an edge in a neighboring pixel to strengthen or weaken the response of an edge detector at a given pixel. This is part of the reason that local smoothing operators help to find “important” edges (anything that is extended over a larger space will be partly spared by a low-band pass filter). For a more detailed discussion of edge detection methods, we recommend several standard computer vision books, including Shapiro and Stockman’s [669] and Forsyth and Ponce’s [237].

5.3.3 Edge Statistics

Edges play a central role in most theories of human visual perception. In fact, many theories focus solely on edges, ignoring any surface or texture qualities. This can be seen, for example, in the first computational theory of visual perception [498]. In Marr's theory, the human visual system first extracts important primitives in an image, which should reflect geometric structure as well as illumination effects, such as highlights. In practice, these features are extracted using the zero crossings of a Laplacian of a 2D Gaussian, which yields edges.

Although the features are sometimes modified to make bars or blobs, nearly all features in the implementation of Marr's theory can be considered to be "edges with some width or length." These (edge-like) features jointly form the *primal sketch*. The edges are then grouped into coherent units, and some depth information for the newly-formed units is extracted to make the 2.5D sketch. Naturally, further processing occurs at subsequent stages of the model. For the present purposes, though, we can see that essentially everything in the image is discarded with the exception of edges. Note that Marr's primal sketch can be formalized into a scale-invariant representation of edges [564], leading to a sparse representation that models the probability distribution of edges in natural images.

Arguably, nearly all modern computational theories of human visual perception follow Marr's basic model to a surprising degree, especially with regard to the focus on edges. This emphasis is to some degree justifiable, since most physiological studies on the neural basis of early visual processing show that one of the first steps in the visual cortex is to extract edges [355, 735]. Of course, subsequent stages extract color, motion, depth, and many other useful features, but edges do seem to be a critical first stage.

Given this central role that edges play, it is very surprising that there are very few empirical studies on the statistics of edges in natural images. There are, of course, many theoretical discussions of what edges in an image *should* do, going back as far as the Gestalt psychologists, who had a lot to say about edges and their role in perceptual organization (for more on the Gestalt laws, please see [414]). Very little work, however, has been done what what edges *actually* do.

While it is true that a considerable amount of work has been done in the machine learning and computer vision fields on learning to detect edges and segment images using a wide variety of filters and that many of these methods rely on the (learned) statistics in image corpuses, most of the statistical descriptions of edge variations in an image corpus are hidden in various classifiers (see, for example, Konishi et al. [417]). Thus, these studies do not provide any *explicit* information about the statistics of gradients or edges.

A few studies have examined the statistics of edges as well as specific edge properties such as collinearity (e.g., [253, 427]). Much of this work has focused on finding empirical evidence showing that the various Gestalt laws of perceptual organization can be derived from image statistics. The most intense focus has been on the law of similarity and the law of good continuation. The first law essentially states that similar elements will be grouped together into a coherent whole.

The second law has two major components. First, in a refinement of the law of similarity, it states that elements with similar orientations will be grouped together. Second, and more critically, it asserts that when several elements intersect, elements that have similar orientations will be joined into one continuous unit. In other words, the human visual system will try wherever possible to join neighboring edges into long, smooth contours.

Empirical evidence for the these laws comes from Geisler and colleagues [253], who extracted edges (using a multistage procedure based on Gabor functions; see Chapter 8) from 20 images representing a wide range of natural (i.e., not manmade) scenes. After extracting the edges, they examined the geometrical relationship between all pairs of edges, focusing on the distance between edge centers, the orientation difference, and the direction of the second edge relative to the orientation of the first edge. Among other things, they found that regardless of the separation of two edges or their orientation, the two edges will most likely have the same orientation. This is even more the case when the two edge segments are near to each other and when they are collinear. They also found that for similarly oriented edges, regardless of how far apart they were, there is a high probability that the second edge is nearly co-circular (i.e., tangent to the same circle). This is reflected in the Gestalt law of good continuation. Using a very different methodology, Krüger and Wörtgötter [427] found very similar results.

Gradients and edges detection processes have also been applied to the study of art. Several researchers have examined the statistics of various categories of art and compared them to each other as well as to real scenes (including comparing a painting of a specific scene to a photograph of that exact scene), in part to learn about how humans represent or encode visual information [277, 278, 603]. They have found that, in general, the statistics of most art are very similar to those of real-world scenes.

5.4 Linear Scale Space

As was mentioned in Section 5.3, one of the first steps in edge detection is often to smooth the image with a Gaussian filter in order to remove the very local, very high-frequency edges. How much the image is blurred will determine which edges get removed. Some edges like the leaves of a tree are very small, with lots of rapid changes in direction and thus only visible in the higher-frequency range. These disappear quickly with little blurring. Other edges, like the trunk of a tree, are larger, with fewer rapid changes and as such, they are visible in the high, middle, and possibly even lower frequencies. Thus, if we want to capture the leaves or fine texture details, we should blur very little. Of course, since both the coarse edges and the fine edges are visible with little blur, it will be hard to decide with little blur if an edge is a fine detail or coarse without additional processing. If, on the other hand, we want to focus on the larger, more spatially extended, coarser structure in an image, then blurring a lot will make these easier to see since they will remove the finer edges.

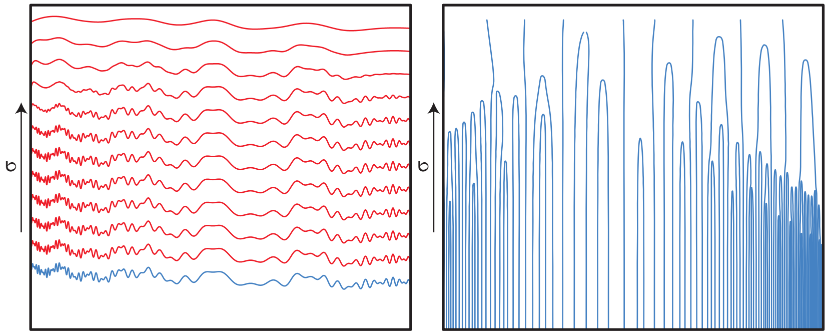
So, what do we do if we want to clearly see the coarse structure as well as the fine details at the same time? Or to remove the coarse and only focus on the fine detail? One way to do this is to look at the image at a number of different blur levels simultaneously. This is the core of *scale-space theory*, in which an image is represented as a family of one-parameter, smoothed images. The parameter is the size of the smoothing function used. Scale spaces have been applied to an incredible variety of fields in computer vision and computer graphics, including image segmentation, image denoising, medical imaging, optical flow,[1] texture analysis, shape-from-texture, object recognition, image matching, even temporal changes (for a recent overview, see [732]).

In 1959, Taizo Iijima axiomatically derived scale space (as cited in [788]). Since the original article was in Japanese, it was not widely read. A little over twenty years later, the idea was independently (re-)introduced by Witkin [800], who focused on one-dimensional signals. The idea was immediately seized upon by the psychophysicist Jan Koenderink [411], who applied it to image analysis and used it to explain many aspects of human visual perception.

Tony Lindeberg was central in further developing the mathematical and theoretical basis of scale space, as well as determining a number of its properties. A very thorough overview of the various methods for deriving linear scale space—including discrete versions of the operators—can be found in Lindeberg's book [464]. In brief, by assuming a few very general qualities, it can be shown that the Gaussian kernel is the only possible choice for the smoothing kernel. Perhaps the most critical property of the smoothing kernel is that the smoothed versions should be simplifications of the original signal. This criterion (often called *causality*) means that no new properties should be introduced by the smoothing (see Figure 5.10). Other critical properties include that the operator not require any special knowledge of the structures in the image (e.g., it should be linear, scale invariant, isotropic, and shift invariant).

Figure 5.10 shows on the left side a simple one-dimensional signal, at several different scales (the original is at the bottom). Notice that the curve becomes smoother the more it is blurred. That is, higher-frequency information is being increasingly suppressed. On the right, the zero crossings in the second derivative are shown as a function of increasing blur. Several critical things can be seen here. First, no new zero crossings are introduced as the width of the blur kernel is increased. Second, and perhaps more critically, the zero crossings form paths

[1] In the human vision literature, this concept is often referred to as *optic flow*. Here we follow the computer vision literature and refer to this phenomenon as *optical flow*.



a. 1D signal smoothed at different scales

b. Zero-crossings at different scales

Figure 5.10. Scale-space example for a one-dimensional signal. (a) A one-dimensional signal (bottom) is convolved with a Gaussian kernel of increasing width. (b) The zero crossings in the second-order derivative are shown as a function of increasing kernel size σ. Notice that no new zero crossings are introduced by the blur and that the zero crossings form paths across the blur levels. The change of features as a function of blur level is referred to as *deep structure*.

through scale space [800]. The nature of features across scales is referred as *deep structure* and can be used to infer rather interesting properties of edges in images.

One issue with scale-space approaches that can be seen from this figure is that dependent on the scale at which a signal is analyzed, zero crossings are detected at different spatial locations. In other words, edges drift across scale space. This is generally an undesirable property of scale-space approaches. This can, however, be overcome by processing the signal with edge-preserving or edge avoiding filters such as anisotropic diffusion [570] or the bilateral filter [747, 557]. Edge-preserving scale-space decompositions can also be obtained by carefully designing filters that can be steered to smooth edges of specific strengths while avoiding others [205, 207, 558]. Such methods have found many applications in visual computing disciplines as they allow images to be processed at different scales, while minimizing the artefacts caused by traditional scale-space approaches [753].

For a two-dimensional image $I(x,y)$, the family of one-parameter, smoothed images can be represented as (see the leftmost column of Figure 5.11):

$$I(x,y;t) = g(x,y;t) * I(x,y;0) \tag{5.32}$$

for $t > 0$ where t is the scale parameter and g is the Gaussian kernel:

$$g(x,y;t) = \frac{1}{(2\pi t)} e^{\frac{x^2+y^2}{2t}} \tag{5.33}$$

The semicolon implies that the convolution is only applied over the variables before the semicolon. The value t defines the scale. Koenderink proved that

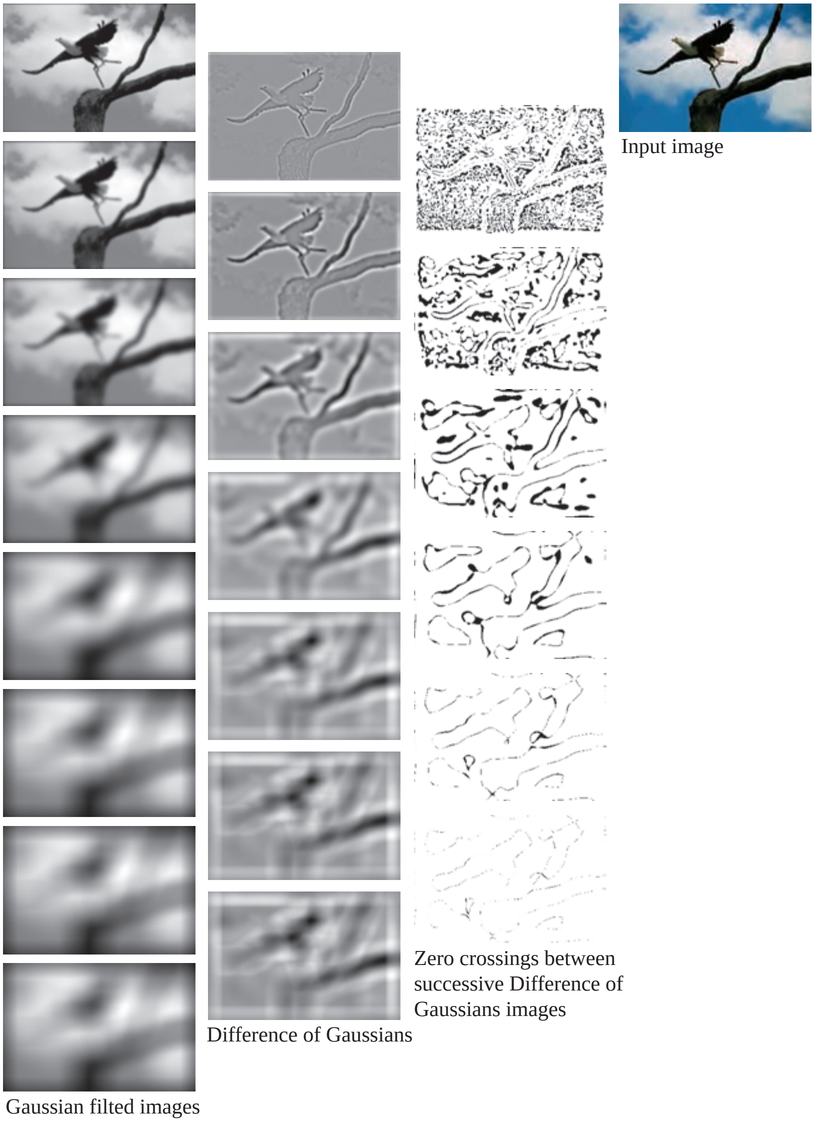

Input image
Zero crossings between successive Difference of Gaussians images
Difference of Gaussians
Gaussian filted images

Figure 5.11. An input image was filtered several times with increasing blur kernels (first column). Pairs of successively filtered images where then subtracted to create a difference-of-Gaussians stack (second column). Then, pairs of difference of Gaussian images were tested to detect zero crossings for each pixel (third column), effectively forming the 2D equivalent of Figure 5.10. (Falknerei Potzberg, Germany, 2013)

convolution of the image with a Gaussian filter is the general solution to the heat diffusion equation (for a homogeneous medium with uniform conductivity [411]):

$$\partial_t I = \nabla^2 I \tag{5.34}$$

In other words, the different scales can be thought of as how heat (or in this case, luminance) would spread over time with the scale parameter t representing time. As the second and third columns of Figure 5.11 show, detecting zero crossings of the second derivative in image space leads at appropriately chosen scales to an edge detector. It has also been shown that, given specific assumptions, scale space can be seen as a special case of wavelets. For more on wavelets, see Chapter 8.

5.4.1 The N-Jet and Feature Detectors

Linear scale spaces become really useful for image statistics once one realizes that differentiation commutes with convolution:

$$\frac{\partial}{\partial x}(I * g) = I * \frac{\partial g}{\partial x} \tag{5.35}$$

Thus, rather than convolving the image with a Gaussian and then calculating the derivative, we can convolve the original image with a Gaussian derivative operator. This naturally leads to a multiscale consideration of gradients as a form of feature detection.

As we saw in Equation (5.1b), the Taylor expansion shows how the luminance for a given pixel is related to its neighbors based on a weighted combination of derivatives of increasing order. Similarly, scale space uses a set of derivatives up to a given order. More specifically, the set of derivatives up to the order N (at a given scale t) is referred to as the *local N-Jet*. The set of derivatives up to a given order N for all scales is the *multiscale local N-Jet*.

By combining Gaussian derivatives, a number of feature detectors can be readily constructed. In order to ensure that the detectors have certain properties, however, it is first necessary to construct a local coordinate system. For an edge detector, for example, it would be useful if the local coordinate system were aligned with the edge, thus ensuring rotational invariance of the detector. Since—as was shown earlier (see Figure 5.5)—the first derivative is at a maximum for edges, it is possible to define the axes in terms of the first derivative and thereby ensuring that the edge detector is aligned with the edge. We can do this by defining one axis v of the local coordinate system as the direction of the gradient (and thus the perpendicular to the edge) and the other axis u as perpendicular to the direction of the gradient.

Now that we have an edge-aligned coordinate system, edges can be located by looking for maxima in the first-order derivative. This is often done using the second-order derivative, which crosses zero at maxima and minima in the first-order derivative. To determine if a given zero crossing is a maxima or a minima, we can look at the third derivative, which will be less than zero for maxima. Thus, the 3-Jet carries information about edges. Since the first derivative perpendicular to the gradient (that is, along u) is by definition zero, the equations simplify considerably within the local coordinate system. An edge can be defined as those points where the second-order derivative along the axis v is zero and the third is less than zero:

$$I_v = 0 \tag{5.36a}$$

$$I_v < 0 \tag{5.36b}$$

As Lindeberg [464] pointed out, due to the discrete nature of an image, the actual zero crossing in the second-order derivative might be between pixels. Interpolating in the second-order derivative to find zero crossings using the constraint that the third-order derivative needs to be negative is an easy way to obtain subpixel edge detection.

Likewise, ridge detectors can be constructed by looking for zero crossings in I_{uv} that satisfy the condition that ${I_u}^2 - {I_v}^2 < 0$. Blob detectors are constructed from maxima or minima in the Laplacian. Junction detectors use a measure of curvature that can be constructed by combinations of Gaussian derivative operators. By selecting the scale(s) at which the feature detector's response is strongest, it is possible to automatically detect the proper scale(s) for examining a given feature[464].

5.4.2 Implications for Human Perception

A considerable amount of work has been done connecting different properties of scale spaces to aspects of the human visual system. For example, Koenderink has suggested that the early stages of the human visual system use multiscale local N-Jets to represent the input [412]. It has also been shown that the structure of receptive fields in the retina can be modeled using a Gaussian scale space [465]. Likewise, the "center-surround" nature of receptive fields in early visual areas can also be modeled using the Laplacian of the Gaussian. More intriguingly, the layout of receptive field sizes on the retina (a linear increase in receptive field size as a function of distance from the retina) is precisely the layout that is predicted if the retina were trying to create a scale-space representation [465]. Finally, the fact that there are massive downward connections from higher visual areas to lower visual areas is consistent with scale-space theory [732], and it suggests that the lower area receptive fields may be modified (something for which there is some physiological evidence; see, for example, [18]). Once nonlinear scale spaces are allowed, even more aspects of human vision (in particular, higher visual processing) can be explained. For more on the relationship between scale space and human vision, we recommend [732].

5.4.3 Scale-Space Statistics

Several researchers have created created a wide variety of feature detectors using scale-space theory and tested them on natural images (e.g., [250, 378, 411, 466, 487]). Interestingly, a large amount of this work is directly interested in deriving 3D shape information from the 2D image information, usually by focusing on the distortions that shape information should undergo during perspective projection. Although these works present statistics on the recovered depth, edge, and orientation values, they focus on the accuracy of the detectors and less on the statistical analysis of natural images.

Salden and colleagues have provided extensive discussions of how one can use scale space to construct complete feature representation systems that are appropriate for natural images (see, for example, [642]). Altough the statistical regularities of those features in natural images remains an interesting prospect for future research, there is some research on the statistical properties of the feature detector. Lindeberg [464], for example, extensively examined how one might extract specific structures from images, including testing these detectors on sample images. In this work he also derived the statistical properties that one should expect from the detectors, then he examined synthetic images to see if the feature detectors functioned as expected. For example, he determined how the number of extrema should decrease as a function of scale (the density of extrema is essentially inversely proportional to the scale parameter). The expected pattern was more or less obtained. Since noise data was used, however, this study is more focused on the statistics of the detector than of images.

Edge statistics were assessed in scale space by Pedersen and colleagues [565]. They examined Van Hateren's database (see Section 3.3.1)) using a 3-Jet to determine edge statistics. In particular, they point out that when constructing an edge detector, at least three dimensions need to be described. Specifically, any given edge will have an orientation, a position within the receptive field, and a blur or scale level. They show that ideal step edges form a two-dimensional manifold in this space that depends solely on the orientation and the ratio of the position to the scale. They then construct a nine-dimensional 3-Jet space, consisting of the two first-order derivatives (I_x and I_y), the three second-order derivatives(I_{xx}, I_{yy}, and I_{yx}) and the four third-order derivatives and project the edge manifold into this space. In this theoretical part, they show that the ideal step edges still trace a 2D manifold.

The data is then whitened and contrast normalized. Whitening is essentially the removal of first- and second-order correlations. As discussed in Section 7.1.1, this can be achieved with the aid of principal components analysis. As a result, the covariance matrix of each nine-dimensional data point becomes the identity matrix. Contrast normalization subsequently takes these data points and divides each point by its norm.

After whitening (see Section 7.1.1) and contrast normalization of the data (to eliminate the effects of lighting and thus be able to focus on geometry), they show that all 3-Jets of ideal step edges will lie on a eight-dimensional sphere in the nine-dimensional space, and thus the distance between and two 3-Jets can be measured by their angular separation. Finally, in the empirical part, they measure the angular distance between the ideal edge manifold and the measured 3-Jets, obtaining the full probability curve for all points in the image. They found that most of the points in an image are very close to the edge manifold, regardless of the scale! In other words, the 3-Jets of the points in the images are clustered near the ideal edge manifold. Furthermore, the edge statistics are roughly scale invariant. They conclude that natural images are extremely sparse, with most of the points in an image clustering around low-dimensional structures like edges, blobs, ridges, and junctions.

5.5 Contrast in Images

An alternative method of assessing relations between pixels is through the computation of contrast. Simple measures of contrast are discussed in Section 2.5.3. Recall in particular Weber's contrast:

$$C_{\text{Weber}} = \frac{L - L_{\text{b}}}{L_{\text{b}}} \tag{5.37}$$

This measure is intended for uniform patches on uniform backgrounds. It can be extended to measure contrast in images in the form of center-surround processing (see Section 2.3.4) and is typically computed as a difference of Gaussians. First, we apply to the image a Gaussian filter twice, but with different spatial extents, leading to L^{center} and L^{surround}. These are then subtracted and optionally normalized to yield a contrast measure for each pixel [723]:

$$C_{\text{DoG}} = \frac{L^{\text{center}} - L^{\text{surround}}}{L^{\text{surround}}} \tag{5.38}$$

This computation gives rise to the filter profile shown in Figure 2.9. If we apply such processing to an image, then we obtain results such as shown in Figure 5.12.

In essence, such center-surround processing can be seen as a method to extract edges from an image; it serves as a rudimentary edge detector. As discussed in Section 2.3.4, such processing also occurs in the human visual system. It removes spatial correlations [36, 703, 25] and effectively returns a zero response in regions where there are no contrasts. This can also be seen as removing image components that are predictable and therefore uninformative. Conversely, the information that is left is unpredictable and therefore informative [741]. Alternatively, it can be understood by saying that surfaces are consistent [289]; edges

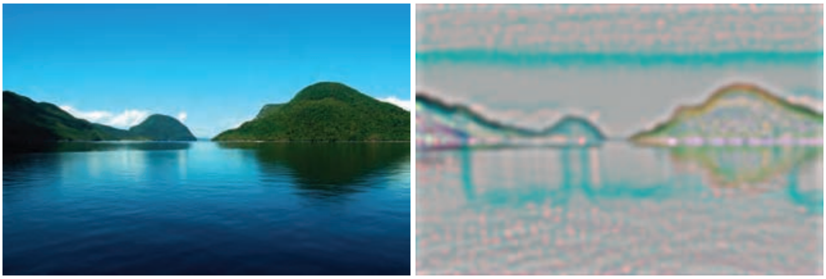


Figure 5.12. Example of center-surround processing. Here, we applied separate processing to each of the three RGB color channels. The filter parameters were $\sigma^{\text{center}} = 6.25$ and $\sigma^{\text{surround}} = 12.5$ pixels. (Doubtful Sound, New Zealand, 2012)

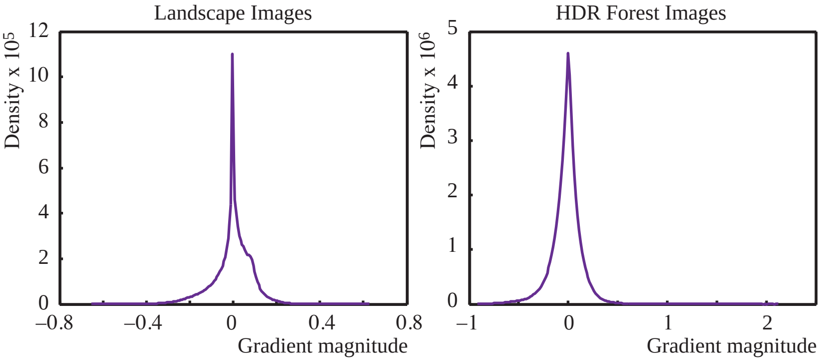


Figure 5.13. Magnitude distribution of difference-of-Gaussian processing, performed on the linearized landscape images of Figure 5.14 (left) and the high dynamic range forest images of Figure 7.8 (right).

form discontinuities that are transmitted to the brain. Anything in between edges is thought to be inferred by the human visual system from the edges alone, a process known as *filling-in* [771, 423, 291, 197, 533, 150].

As gradients compute differences between neighboring pixels and center-surround processing computes differences between pixel areas, one would expect that histograms of the magnitude of difference of Gaussian responses follow the same behavior as gradient histograms. That this is indeed the case is shown in Figure 5.13, where a center-surround computation was carried out on two image ensembles. The 15 images of the first ensemble are shown in Figure 5.14, whereas the second ensemble is shown in Figure 7.8. The first ensemble consists of 15 landscape images, whereas the second ensemble consists of 15 high dynamic range forest scenes, which do not contain significant amounts of sky.

The forest image ensemble produces a longer tail for positive gradients, which is likely due to the fact that these images were captured in high dynamic range.

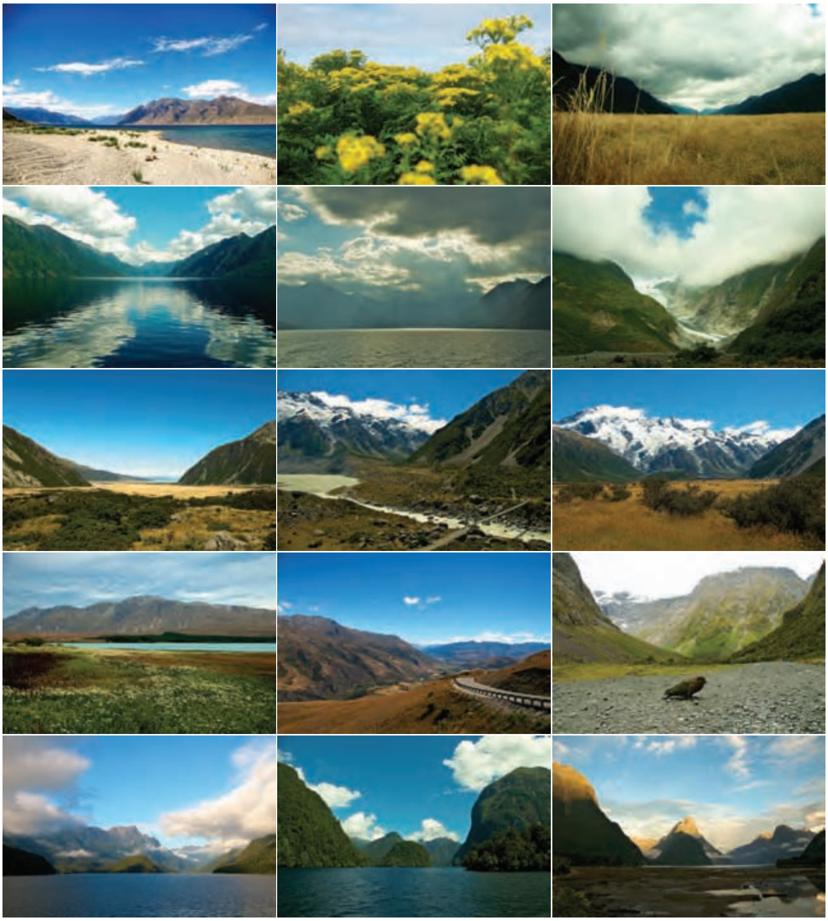


Figure 5.14. An ensemble of 15 images containing landscapes. (Various, South Island, New Zealand, 2012)

The landscape image ensemble was linearized prior to the calculations. The shape of both plots suggests that a highly kurtotic distribution is present, consistent with the idea that center-surround processing removes unimportant information. High dynamic range imaging, which is a closer representation of light available in scenes than afforded by conventional images, appear to exhibit higher kurtosis, an issue also seen in moment statistics, which were discussed in Section 4.2.

It has been suggested that negative center-surround responses to natural images are more numerous than positive responses [31]. The difference between the number of negative versus positive responses was measured to be as high as 50%. This could be an explanation for the fact that there are twice as many OFF pathways in the human visual system as there are ON pathways (see Section 2.3.3).

Unfortunately, we were not able to reproduce this result. For the HDR forest ensemble the ratio between positive and negative responses is 1:1.15. In other words, the negative responses are 15% more numerous. Given that landscape images tend to have bright sky near the top of the image, with darker ground and foreground areas, we carried out the same experiment on the 15 landscape images of Figure 5.14. Here, the ratio is 1:1.025, albeit that in this case the positive responses were slightly more numerous (by 2.5%).

So far, we have relatively arbitrarily fixed the size of the center Gaussian to $\sigma^{\text{center}} = 6.25$ pixels and the surround to $\sigma^{\text{surround}} = 1.5$ pixels. There is of course no reason to believe that these values are in some sense optimal. Further, human vision is sensitive to contrasts at different scales as modeled by the contrast sensitivity function (see Section 2.5.3). This has led to the development of contrast measurements that take different spatial scales into account, which can be obtained by first band-pass filtering the image [567]. Typically, center-surround processing at different spatial scales begins by creating a stack of Gaussians, as outlined in Section 5.4. To create differences of Gaussians at different scales, pairs of neighboring scales are simply subtracted.

Finally, note that contrast as well as luminance values vary greatly across an image. Moreoever, contrast and luminance values tend to vary independently. Interestingly, human vision appears to process contrast and luminance separately, showing yet another form of adjustment to statistical regularities in natural scenes [495].

This also affects adaptation in the human visual system, as saccadic eye movements will continuously cause the retina to focus on different patches with correspondingly different local contrasts and luminances [244]. This means that human vision does not tend to approach a state of full adaptation, as the retina is presented with different luminance and contrast statistics every few hundred milliseconds or so. This is important to maintain vision. Consider a situation whereby every photoreceptor is fully adapted to its input. In that case, the semi-saturation constant σ of Equation (2.1) would equal its input I, and as a result, the response of every photoreceptor would be the same, namely, half-maximal. If all photoreceptors emit the same signal, vision would essentially be blind! Non-stationary statistics combined with saccadic eye movements prevent this from happening.

Of course, motion of objects in scenes will also cause significant deviations from the zero response and are therefore particularly salient. The temporal element of image statistics is further discussed in Chapter 12.

5.6 Image Deblurring

When taking a photograph of a scene, it is not uncommon that either the camera or an object in the scene moves. The longer the aperture is open, the more likely this is to be the case. As a result, all or part of the image will be blurred. A number of approaches for sharpening an image have been proposed. One type of approach, *blind motion deconvolution*, essentially treats the photograph as the result of a convolution between an unknown sharp image and an equally unknown blurring kernel. The goal is to estimate the blur kernel so that it can be deconvolved with the blurred image to yield the sharp image. Naturally, this is an underspecified problem, so additional constraints are needed; recently, a number of researchers have employed natural image statistics to provide them.

For example, gradient distributions derived from natural images can be used to estimate the blur kernel [211]. They can be effectively modeled with a hyper-Laplacian distribution [425]:

$$p(x) \propto \exp\left(-\alpha\, x^{\beta}\right) \tag{5.39}$$

The heavy-tailed distribution of gradient histograms was shown to be best modeled if parameter β was set to approximately $2/3$ [425]. Hyper-Laplacian distributions have also been successfully applied to applications such as transparency separation [455], image segmentation [324] and superresolution [726] (see Section 5.7).

Of course, other image statistics approaches are possible, for instance, by utilizing the typical $1/f^{\beta}$ distribution of power spectra (Section 6.4.2), which would hold for sharp images but not necessarily for blurred ones [112]. An interaction between power spectra and wavelets can also be used [371, 530].

All of these approaches assume camera motion—that is, that there is a single blur kernel for the entire image. In an alternate approach, the gradient structure can be used to find those pixels that are blurred and segment them from the rest of the image, as demonstrated by Levin [453]. Specifically, a correspondence is found between the gradient distribution and the blur present in the image, allowing the discovery of the blur kernel for a given image.

One primary feature of motion blurring is the attenuation of higher frequencies. This shows in the slope of the power spectra (by increasing β) as well as in gradient histograms (in particular by removing the longer tails at the higher gradients). Levin attempts to recover the blur kernel by applying different blurs to the image to find the one that produces a gradient histogram that matches the blurred one. This requires a non-blurred image or image region. Using a completely different image tends not to produce reliable results (due to differences in the underlying gradient histogram). Since much terrestrial motion tends to be horizontal, the majority of the effects of motion blurring are also horizontal. Thus, the vertical gradients can, under some circumstances, be used to calculate the blur of the horizontal components. In a further refinement, Shan et al. [667] employ an iterative alternation between optimizing the blur kernel and the sharp image using priors from natural image statistics.

5.7 Superresolution

There is a growing field of work that aims to infer information from an image that is not actually present in the image. Image deblurring is one example, where sharpness is recreated. One core application in this area is superresolution, which essentially tries to increase the resolution of an image by intelligently guessing what the missing higher-frequency information is. For example, Fattal notes that there is a difference between shadow edges and texture edges [206]. Not only are humans sensitive to these differences, but information about what type of edge is present at a pixel can be found in the gradient structure. The unique intensity falloff found at the edge of a shadow—the penumbra—seems to be used by human vision to find shadows and separate illumination changes from reflectance changes [270].

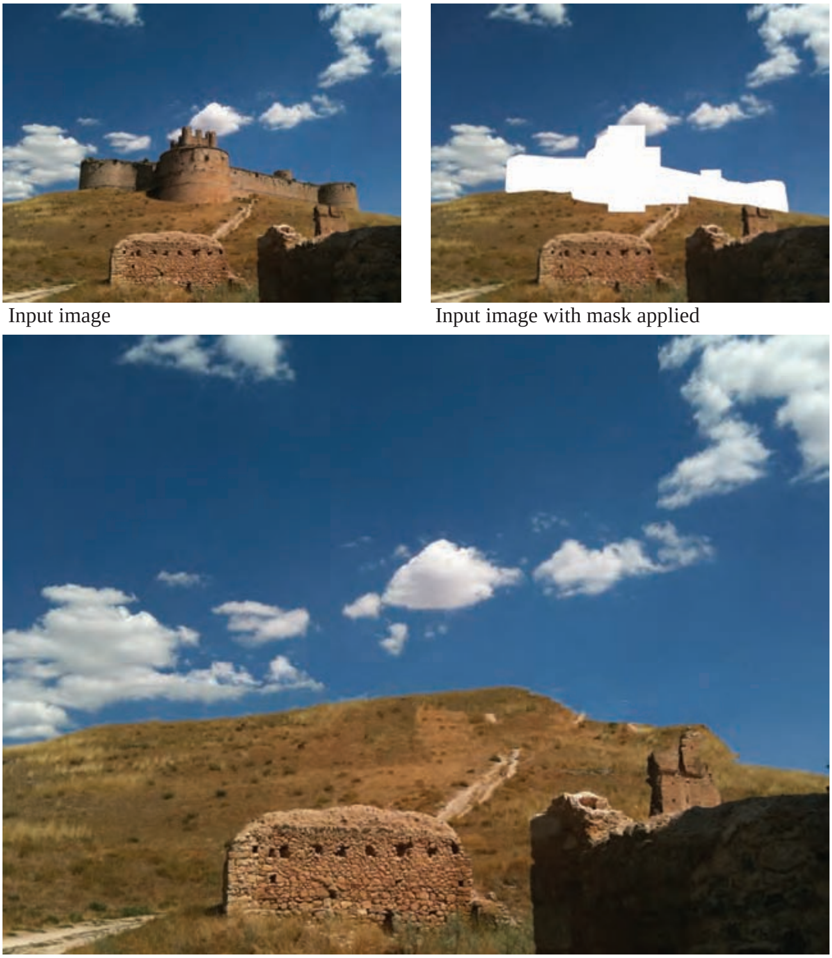
Following this insight, Fattal et al. examined 15 indoor images, looking at the statistics of pixels in a small area to see if the sharpness of an edge in the high-resolution image could be estimated from the low-resolution version of the image. To do this, they calculated a gradient field (using central differences) and then searched for the nearest edge (along the gradient) for each pixel. Finally, they determined how much the luminance changes along that edge, how far the edge is from the pixel, and how sharp the edge is. They found a clear dependency between the three measures and the continuity profiles exhibited by the image intensities, and they were able to use this information to improve the quality of a superresolution technique.

In a different approach, Tappen et al. [726] use the gradient distribution to fill in the holes produced when a high-resolution image is produced from a low-resolution image. They also apply the same procedure to yield a full-color image from an achromatic image.

5.8 Inpainting

No imaging system is perfect. All imaging systems contain intrinsic noise and occasionally yield results with holes (sometimes due simply to self-occlusion). This is also true of the human visual system. The “blind spot,” which is the section of the retina where the optic nerve leaves the eye, has no photoreceptors. Yet we do not see a blank spot there. Aside from that, there are many situations whereby it is desirable to remove an object from an image—for instance, a photograph shot through a wire fence may benefit from the removal of the fence. Distracting objects may also need to be removed, requiring algorithms to guess what may have been behind those objects.

There is a range of computer graphics algorithms for “hole filling” or image completion, which can be captured under the title *inpainting* [51, 134]. Inpainting techniques can be divided into several classes based on the central algorithm. The



Input image

Input image with mask applied

Inpainted result

Figure 5.15. An example of inpainting using Crimini's [134] algorithm. After regions from an image (top left) are deleted (top right), the hole is filled (bottom) using this and similar images. (Castilla y Leon, Spain, 2010)

three typical categories are texture synthesis, methods relying on partial differential equations, and exemplar-based techniques.

The first category, texture synthesis, is generally considered to be the oldest form of inpainting, likely originating with Fournier et al. [241]. These methods have as their goal the synthesis of a new image patch that is perceptually similar to a user-defined patch. For example, the technique developed by Heeger and Bergen [323] generates stochastic textures using a straightforward model of the human visual system. Specifically, their method takes a white noise patch—of arbitrary size—and alters it using, in part, a pyramid-based histogram matching algorithm so that the first-order statistics of a series of oriented spatial-frequency filters are similar to those of the patch. For more on such filters, see Chapter 6. A wavelet-based approach to texture synthesis is discussed in Section 8.14.

The second class of approaches is based on the theory of variational methods and partial differential equations (PDE). One of the first of these was proposed by Bertalmio et al. [51], who used an iterative procedure, where edges (detected using gradients) at the border of the missing region are extended into the missing region. Chan and Shen [117] extended this model using the theory of total variation, as well as an anistropic diffusion based on the strength of the gradients. Bertalmio et al. [52] also proposed a hybrid approach, which uses PDE-like methods to fill in the structure of an image and then applies texture synthesis methods to finish by filling in the texture.

Finally, there are the examplar approaches, which use image statistics to find optimal replacements for the missing texture, usually by assuming that a patch similar to the missing bit can be found somewhere in the remaining image. For example, Hirani and Totsuka [332] use a combination of spectral and spatial information to find the replacement patch. The seminal work of Efros et al. [195, 194] has done much to define the field of exemplar-based inpainting. Levin [458] suggested that the missing section can be filled in based on the gradients at the boundary region as well as some term that maximizes the match to the global gradient histogram.

Likewise, Criminisi et al. [134] included information from gradients in the algorithm that decides which regions should be completed first. They then select a small patch of pixels around the region to be completed and then find similar patches elsewhere in the image. Image editing can also proceed entirely in the gradient domain. In that case, an image can be reconstructed by solving the Poisson equation. This allows gradients to be directly interpolated into the missing region [568, 673]. Alternatively, the patch-copying procedure can be treated as a global optimization of a Markov random field [416] (see Chapter 9).

Chapter 6

Fourier Analysis

As mentioned in the previous chapter, the interaction of light rays with an object systematically alters light according to the surface properties of the object. This means that the pattern of change across a set of light rays provides direct, reliable information about the object. This can also be seen by permuting the pixels of an image (Figure 6.1). Although the exact same pixels are available in both images, the permutation has resulted in the loss of meaningful information. Thus, a statistical assessment of an image based on individual pixels alone will give only very partial insight into the structure of an image.

Of course, the structure in images is of central interest to the human visual system. It should not be surprising, then, that considerable attention has been directed toward investigating second- and even higher-order statistics. It has been shown, for example, that people can discriminate almost without effort between images that differ in their second-order statistics [381, 100]. Second-order statistics are those that typically involve covariances or correlations between pairs of pixels.

In Chapter 5, we showed that looking at the difference between pairs of neighboring pixels provided a large amount of information about structure. Gradient analysis is, however, very limited in that it requires the two pixels to be neighbors. In this chapter we address ways to analyze relationships between pixels that are not necessarily adjacent in images. In a given image, the probability that neighboring pixels have similar values tends to be high, whereas pixels some distance away have less ability to predict a pixel's color. This phenomenon is described by the autocorrelation function, which gives the correlation between pixels at different points in image space. It is explained in more detail in Section 6.1.

Although the autocorrelation function could provide useful insights into image structure, it tends to be computationally inefficient to compute in image space. An equivalent analysis can be performed in frequency space by examining the power spectrum. This requires the use of Fourier transforms, which are introduced in Section 6.2. The equivalence of the autocorrelation function and the

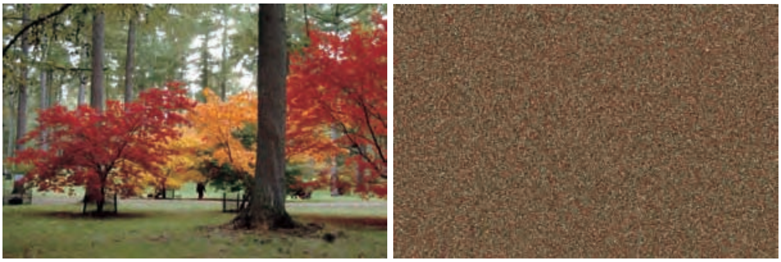


Figure 6.1. The pixels of the image on the left were permuted to create the image on the right, in the process destroying higher-order structure that prevents the image to be meaningfully interpreted. (Westonbirt Arboretum, UK, 2011)

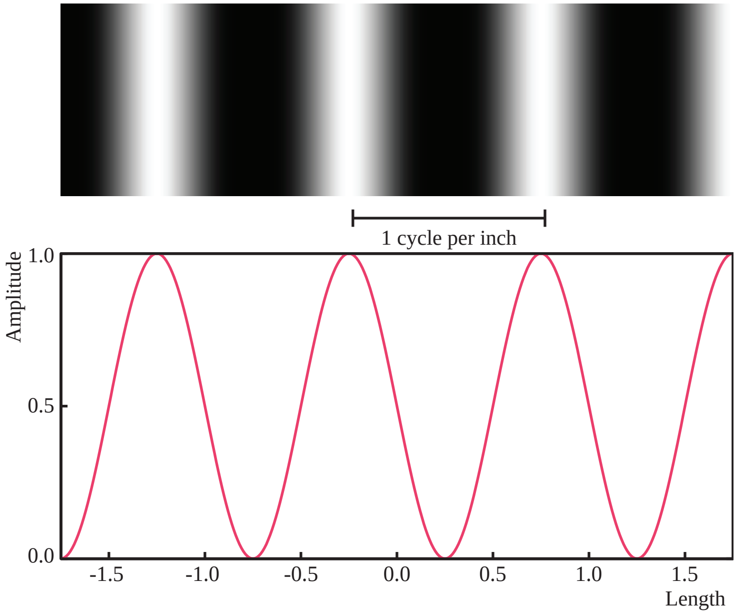


Figure 6.2. A sine function encodes a single frequency, in this case at one cycle per inch.

power spectrum is known as the Wiener-Khintchine theorem, which states that these form a Fourier transform pair. This is explained in Section 6.3.

The power spectrum encodes how much of each spatial frequency is available in the image. Spatial frequencies are measures of the number of oscillations (cycles) from dark to light per unit of length of the image (Figure 6.2). In vision research it is common to express this as the number of cycles per degree (cpd)—effectively a measure of oscillations per unit of visual angle (α in Figure 6.3). The two ways to express spatial frequencies convey the same information,

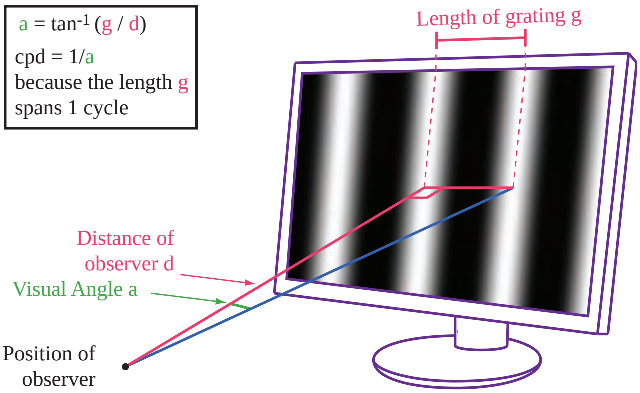


Figure 6.3. The angular frequency (in cycles per degree, cpd) depends on the size *g* at which a sine wave is displayed as well as the distance *d* between the observer and the display.

although the latter assumes that the distance between the observer and the display is known. The importance of power spectral analysis is discussed in Section 6.4.

Finally, Fourier decompositions return complex valued elements. The magnitude of the complex numbers gives information of the frequency content, whereas the angle provides information regarding the position within the image, also known as *phase*. There have been attempts to discover statistical regularities in the phase content of images, but although most information that characterizes an image is known to be encoded into the phase spectrum, analyses of the phase structure of images have not been as overwhelmingly successful as frequency-domain analysis. Fourier-based phase structure is discussed further in Section 6.5.

6.1 Autocorrelation

The autocorrelation function is a special case of cross-correlation. Cross-correlation is a measure of how similar an image is to a mask. Assume we have an image I and a mask M, with the mask having a size of $m_x \times m_y$ pixels, which is smaller than the size of the image $(n_x \times n_y)$. Image and mask will have means μ_I and μ_M and standard deviations σ_I and σ_M, respectively.

Each pixel in the image now forms the center of a region the size of the mask. The similarity of the region with the mask can then be computed using the normalized cross-correlation $r(x,y)$:

$$r(x,y) = \sum_{x_s=-m_x/2}^{m_x/2-1} \sum_{y_s=-m_y/2}^{m_y/2-1} \frac{I(x+x_s, y+y_s)\, M(x_s+m_x/2, y_s+m_y/2)}{\sigma_I\, \sigma_M} \qquad (6.1)$$

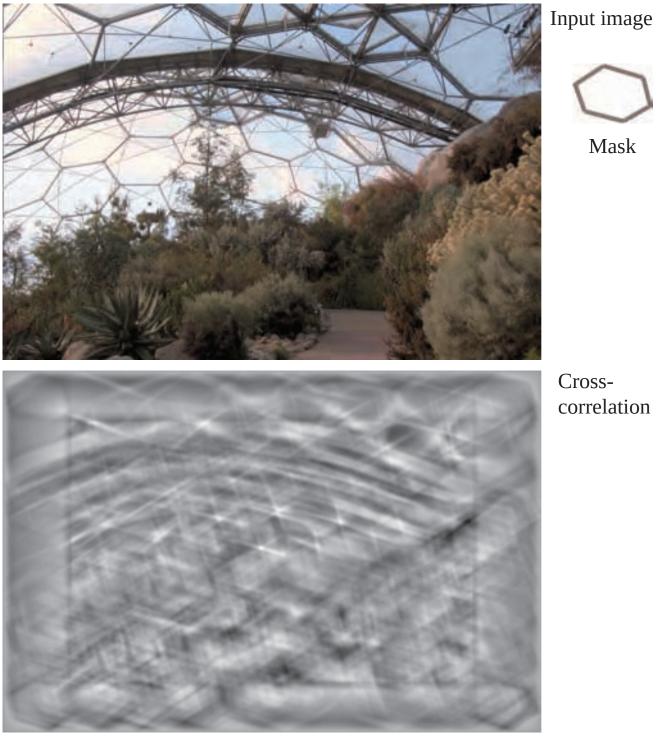


Figure 6.4. Cross-correlation between an image and a template yields values between –1 and 1 for each pixel. (Eden Project, Cornwall, UK, 2011)

The normalized cross-correlation $r(x,y)$ produces values between -1 and 1. It will be close to 1 if the region around (x,y) closely resembles the pattern encoded by the mask. Such a measure is thus useful to detect specific patterns in images. An example is shown in Figure 6.4. Here, the spikes show where the hexagonal template was matched best against the input image.

When such a cross-correlation is computed between the image and a displaced version of the same image, we measure autocorrelation as a function of displacement. If for a given displacement a high autocorrelation is found, then this could be interpreted as an indication that the image exhibits periodicity with a frequency commensurate with this displacement. The autocorrelation function r_a as a function of displacement (u,v) is given by:

$$r_a(u,v) = \sum_{x_s=0}^{n_x-1} \sum_{y_s=0}^{n_y-1} \frac{I(u+x_s, v+y_s)\; I(x_s,y_s)}{\sigma_I^2} \tag{6.2}$$

This function is essentially the cross-correlation of an image with itself, and is therefore a direct consequence of the formulation of the cross-correlation in

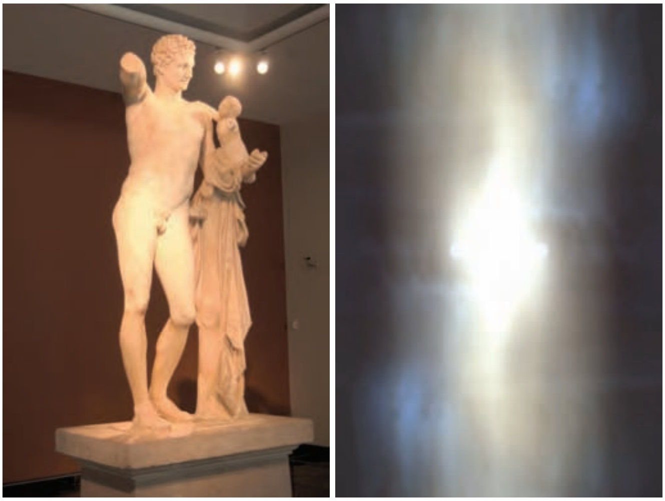


Figure 6.5. The autocorrelation for the left image is shown on the right. The autocorrelation was computed for each of the red, green, and blue channels separately, leading to slight color variations in this image. (Olympia, Greece, 2010)

Equation (6.1). An example is shown in Figure 6.5, where the autocorrelation was computed for each of the red, green, and blue channels separately.

The autocorrelation for the green channel is shown as a surface in Figure 6.6. The zero displacement is shown in the middle of the surface as a large spike. For natural images, larger displacements tend to lead to smaller autocorrelation values. Note also that this image has produced some direction sensitivity, with the autocorrelation remaining at higher values in the vertical direction compared with similar horizontal displacements.

The computation of the autocorrelation function as in Equation (6.2) is expensive to compute. In effect it has a time complexity of $O(n_x^2\, n_y^2)$, which is not practical for decent-sized images. However, it can be shown that the autocorrelation function is equivalent to the power spectrum, which is computed in Fourier space. In fact, when applied to natural image ensembles, the power spectrum shows striking statistical regularities. To discuss these findings, we first briefly explain Fourier transforms (Section 6.2) and show that the autocorrelation and power spectrum constitute a Fourier transform pair (Section 6.3).

6.2 The Fourier Transform

The Fourier transform is exceptionally important in digital image processing, computer vision, and other fields, but it is also important in the statistical analysis of natural images. We therefore briefly review its construction and its properties.

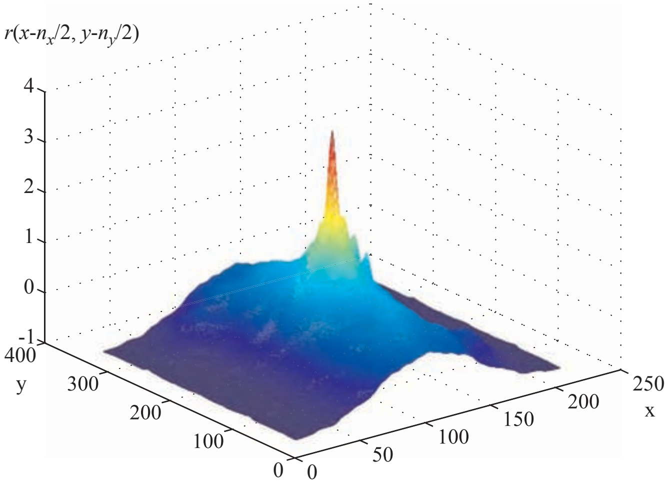


Figure 6.6. The autocorrelation function for the green channel of the left image of Figure 6.5 is shown here. The spike in the middle of the plot coincides with a zero displacement.

We begin by noting that sinusoids—sines and cosines—are periodic functions for all k $\mathbb{Z}$:

$$\sin(x+2\pi k) = \sin(x) \tag{6.3}$$

$$\cos(x+2\pi k) = \cos(x) \tag{6.4}$$

A sinusoidal function $g(x) = A\ \sin(2\pi f x + \theta)$ is periodic with frequency f, amplitude A, and phase θ. The period T_0 of this function is $1/f$. We can add sinusoids with different periods and if these periods are integer multiples of a base period (i.e., they are commensurable), then the resulting sum is also periodic. Figure 6.7 shows some examples.

We can add a large number of sinusoids together using this same progression, i.e., $g(x) = \sum_n A\ \sin(2\pi f x/2^n + \theta)$, as shown in Figure 6.8. As each frequency is a harmonic, i.e., an integer multiple of a base frequency, a step edge is beginning to form, although there is still some undershoot and overshoot of the signal near the edge (known as the Gibbs phenomenon). In general, arbitrary periodic waveforms can be built by summing harmonically related sines and cosines.

Conversely, given an arbitrary periodic waveform, it is possible to find its trigonometric series representation, known as the trigonometric Fourier series of a signal. A trigonometric Fourier series is generally written as:

$$g(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos(2\pi f n x) + \sum_{n=1}^{\infty} b_n \sin(2\pi f n x) \tag{6.5}$$

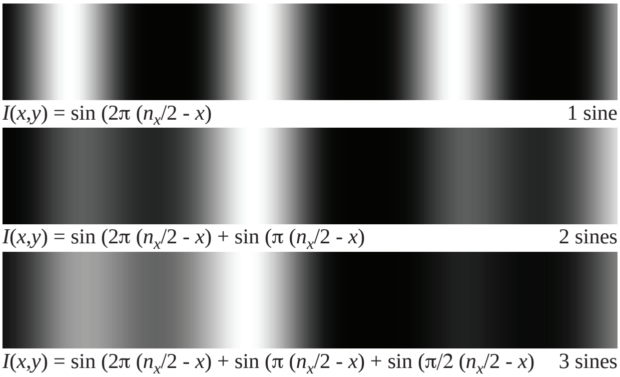

I(x,y) = sin (2π (n_x/2 - x)
1 sine
I(x,y) = sin (2π (n_x/2 - x) + sin (π (n_x/2 - x)
2 sines
I(x,y) = sin (2π (n_x/2 - x) + sin (π (n_x/2 - x) + sin (π/2 (n_x/2 - x)
3 sines

Figure 6.7. Sines with commensurable periods are summed, leading to a new signal which is also periodic.

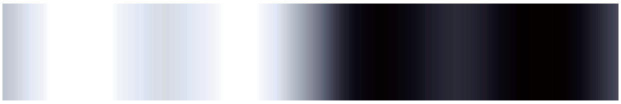


Figure 6.8. Summing a large number of sinusoids (500 in this case) approximates a step edge.

with $-\infty < x < \infty$. Thus, a specific signal $g(x)$ can be represented by the coefficients a_0, a_n, and b_n, which can be computed as follows [827]:

$$a_0 = f \int_{1/f} g(x)\, dx \tag{6.6}$$

$$a_n = 2f \int_{1/f} g(x) \cos(2\pi f n x) dx \qquad n = 0 \tag{6.7}$$

$$b_n = 2f \int_{1/f} g(x) \sin(2\pi f n x) dx \tag{6.8}$$

A Fourier series will converge to $g(x)$ if this function has continuous first and second derivaties over a period $T_0 = 1/f$, except possibly at a finite number of points where it may have finite jump discontinuities. A Fourier series is defined over a period $1/f$. Outside this interval, the Fourier series converges to a periodic extension of $g(x)$.

With the exception of certain special cases, the trigonometric Fourier series may alternatively be expressed in terms of cosines alone by noting that:

$$a_n \cos(2\pi f n x) + b_n \sin(2\pi f n x) = A_n \cos(2\pi f n x + \theta_n) \tag{6.9}$$

As a result, the trigonometric Fourier series is then:

$$g(x) = A_0 + \sum_{n=1}^{\infty} A_n \cos(2\pi f n x + \theta_n) \tag{6.10}$$

The Fourier coefficients are now given by the amplitudes A_n and phases θ_n, which go with frequencies fn. The coefficients A_n plotted against frequencies fn (where $n = 0, 1, 2, \ldots, \infty$) form the amplitude spectrum of the signal, whereas the coefficients θ_n plotted against frequencies fn (where $n = 1, 2, 3, \ldots, \infty$) give the phase spectrum.

A third form of the Fourier series is exponential in nature. We note that using Euler's formula, the cosine in (6.10) can be written as:

$$cos(2\pi fnx + \theta_n) = \frac{1}{2}e^{i(2\pi fnx+\theta_n)} + \frac{1}{2}e^{-i(2\pi fnx+\theta_n)} \tag{6.11}$$

Thus, it is possible to represent a signal $g(x)$ as a sum of complex exponential terms:

$$g(x) = \sum_{n=-\infty}^{\infty} X_n e^{2\pi i fnx} \tag{6.12}$$

The complex coefficients X_n can be computed as follows:

$$X_n = \frac{1}{T_0}\int_{T_0} g(x)\, e^{-2\pi i fnx}\, dx \tag{6.13}$$

which defines the exponential Fourier series. Note that these coefficients directly relate to those of the trigonometric Fourier series, as follows:

$$A_n = 2\ X_n \tag{6.14}$$

$$\theta_n = \tan^{-1}\left(\frac{\text{Im}\ X_n}{\text{Re}\ X_n}\right) \tag{6.15}$$

This makes it convenient to calculate the exponential Fourier series and then derive amplitude and phase spectra according to these equations.

So far we have considered only one-dimensional signals $g(x)$. For images, the extention to two dimensions is as follows. The Fourier transform $F(u, v)$ as function of an image I(x,y) is defined as:

$$F(u,v) = \sum_x \sum_y I(x,y)\ \exp\left(-2\pi i(xu + yv)\right) \tag{6.16}$$

Thus, an image can be transformed into an equivalent representation which is termed Fourier space or frequency space. The reasons why such a conversion represents frequencies will become apparent later in this section. The inverse Fourier transform takes the frequency representation back to image space:

$$I(x,y) = \sum_u \sum_v\ F(u,v)\ \exp\left(2\pi i(xu + yv)\right) \tag{6.17}$$

With this mechanism, it is possible to convert an image to Fourier space, carry out calculations, and then transform the result back to image space. The reason that one might want to go to Fourier space is that some calculations become much more computationally efficient. Noting that for an image with n pixels the time complexity of the forward and backward transforms is $O(n \log n)$ for efficient Fast Fourier Transforms (FFTs), any calculation that costs $O(n \log n)$ in Fourier space, but more in image space, becomes amenable to treatment in this space.

An example is convolution, where in image space a sliding window is moved across the image. For every pixel, a set of neigboring pixels is weighted and summed according to a specified filter kernel δ:

$$I^{\text{conv}}(x,y) = \sum_i \sum_j I(i,j)\ \delta(x-i, y-j) = I \star \delta \tag{6.18}$$

where the "$\star$" is the convolution operator. Convolutions are very useful, for instance to blur an image. In this case, the filter kernel is often chosen to be Gaussian. Although the filter kernel could be capped, summing over only a small pixel neighborhood around each pixel, this would make the convolution approximate. In the formulation above, each pixel contributes to every other pixel, making the time complexity of this operation $O(n^2)$.

The convolution theorem states that a convolution could be equivalently expressed in the Fourier domain as a multiplication of the Fourier transform of the image and the Fourier transform of the filter kernel. Denoting the Fourier transform as $\mathscr{F}[]$, the convolution theorem is given by:

$$\mathscr{F}[I \star \delta] = c\ \mathscr{F}[I]\ \mathscr{F}[\delta] \tag{6.19}$$

where c is a normalizing constant. We can therefore transform both image and kernel, incurring a cost of $O(n \log n)$, followed by a multiplication that costs $O(n)$. Including the inverse Fourier transform results in a total cost of $O(n \log n)$, which is especially for large images significantly cheaper than $O(n^2)$.

At this point, it is interesting to note the mathematical similarity between convolution and cross-correlation. In 1D, the correlation between two functions f and g is given by:

$$r(x) = \int_{-\infty}^{\infty} f(\theta)\ g(x+\theta)\, d\theta \tag{6.20}$$

This is the one-dimensional case of the cross-correlation function given earlier in (6.1). Convolution in 1D is defined as:

$$(f \star g)(x) = \int_{-\infty}^{\infty} f(\theta)\ g(x-\theta)\, d\theta \tag{6.21}$$

Despite mathematical similarities, their interpretation and use is different. In signal processing for instance, convolutions are used to compute the output of a linear system given a specific input. Their Fourier space equivalent is a multiplication.

Correlation is used to compare the similarity of two signals, reaching its maximum for the displacement at which the two signals best match. As mentioned in the previous section, the autocorrelation function is a special case of cross-correlation, whereby a signal is compared to itself for different displacements. This function has the same time complexity as convolution and is therefore also expensive to compute in image space. However, it turns out that the autocorrelation function can also be expressed in Fourier space, an important finding discussed next.

6.3 The Wiener-Khintchine Theorem

The reason to discuss the Fourier transform in some detail is that there is a direct link between computations in Fourier space and those in image space. While this is generally the case, here we refer specifically to the relation between the autocorrelation function (Section 6.1) and power spectra [538]. In other words, Fourier space gives us computationally convenient access to the Fourier equivalent of the autocorrelation function. Recall that the autocorrelation function $r_a(u,v)$ of an image $I(x,y)$ is the cross-correlation between the image and a displaced version of the same image as function of the displacement (u,v).

The Wiener-Khintchine theorem [796, 397] states that the autocorrelation function r_a can be expressed as:

$$r_a = \mathscr{F}^{-1}\left[\, F^{\,2} \right] \tag{6.22}$$

where F is the Fourier transform of the image, and $\mathscr{F}^{-1}$ denotes the inverse Fourier transform, as defined in Equation (6.17). Given Equation (6.14), we see that the amplitudes squared are the Fourier equivalent of the autocorrelation function. It is therefore possible to gain information about the autocorrelation function by analyzing the square of the amplitude spectrum, which is also known as the power spectrum. As discussed in the following section, natural images exhibit striking statistical regularities in their power spectra.

6.4 Power Spectra

Images contain information at different spatial scales, from the large (e.g., the mountains in the distance of Figure 6.9) to the very small (e.g., the grass on the left as well as the fine texture on the mountains due to vegetation). As is well known from Fourier analysis, sharp edges, such as at the silhouette of the mountains against the sky, can be described by a weighted sum of sinusoids, with higher frequencies being weighted less (see Section 6.2 and in particular Figure 6.8). An examination of the relative power of the different spatial frequencies reveals several interesting trends, which are so prominent that many works in image statistics provide an analysis of the power spectrum.

Figure 6.9. A typical natural scene, containing details at various scales. (Milford Sound, New Zealand, 2012)

6.4.1 Slope Computation

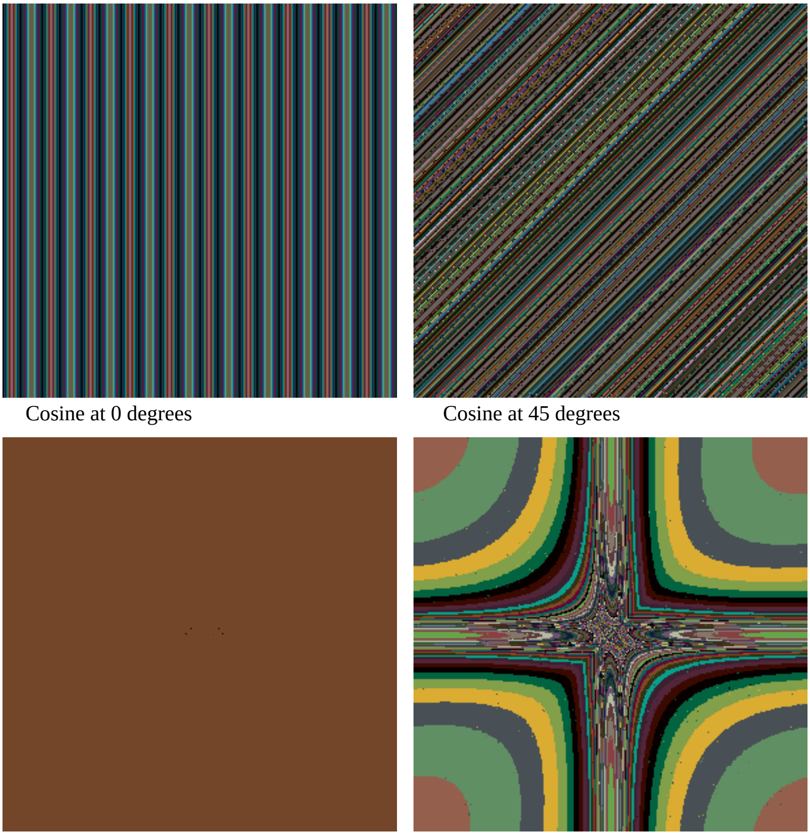
To compute the spectral slope of an ensemble of images, the following procedure can be used [653]. We begin by assuming that images are square. Then, the weighted mean intensity μ is subtracted to avoid leakage from the DC-component of the image, with μ defined as:

$$\mu = \frac{\sum_{(x,y)} L(x,y)w(x,y)}{\sum_{(x,y)} w(x,y)} \tag{6.23}$$

Here, $w(x,y)$ is a weight factor that is defined below.

Next, the images should be prefiltered to avoid boundary effects that are related to the Fourier transform. As the Fourier transform requires the signal to be periodic to give a correct answer, it should be used with care when applied to images. To see this, consider a cosine grating as shown in the top left of Figure 6.10. This grating is aligned with the image and contains an exact number of periods between the left and right edges. The amplitude spectrum in Fourier space will show exactly two spikes, left and right of the center of the image (we have shifted this amplitude spectrum such that the zero frequency is in the middle of the image). These spikes represent the frequency at which this cosine occurs. Edge artifacts are carefully avoided in this image by making the signal truly periodic.

However, this can not be maintained in general, and even rotating the cosine by 45 degrees breaks the assumption of the image being periodic. As the left and



Cosine at 0 degrees

Cosine at 45 degrees

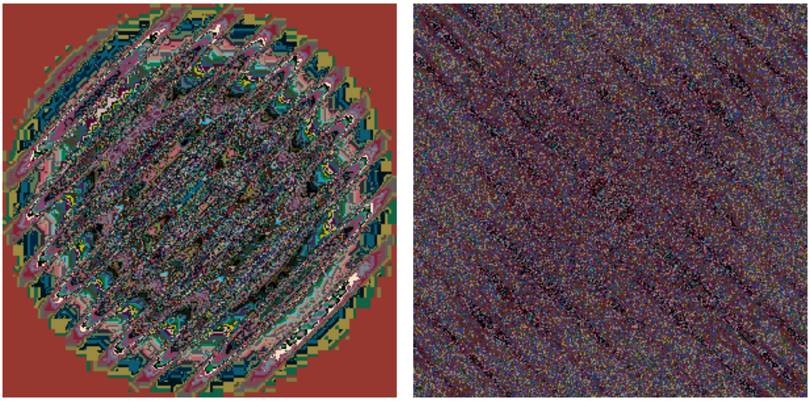
FFT of cosine at 0 degrees: 2 spikes

FFT of cosine at 45 degrees:
2 spikes plus edge effects

Figure 6.10. The top row shows two oriented cosine gratings. Underneath are shown their amplitude spectra.

right edges of the image no longer line up (the same is true for top and bottom images), the amplitude spectrum is no longer a pair of spikes. In fact, it is a pair of spikes augmented with various other structures that exist because opposite edges of the image are not lining up.

In general, images are not periodic and this means that measures ought to be taken to minimize edge artifacts when analysing ensembles in Fourier space. This is normally accomplished by applying a window to the image first, which tapers pixel values to a constant value near the edges. This makes edges on opposite sides of the image have the same value and therefore simulates a periodic signal. The center of the image is largely unaffected to ensure that meaningful statistics can be computed. For the cosine example, the result of windowing is shown in Figure 6.11.



Windowed cosine at 45 degrees

FFT of windowed cosine at 45 degrees:
2 spikes plus reduced edge effects

Figure 6.11. After windowing with a Kaiser-Bessel window (left), the amplitude spectrum shows much fewer edge artifacts (right).

There are many choices of window [305], each with some advantages and disadvantages. A good trade-off between computability, side-lobe level, and main-lobe width is afforded by the circular Kaiser-Bessel window with parameter $\alpha = 2$ [305]. It is computed as follows:

$$w(x,y) = \frac{I_0\left(\pi\alpha\sqrt{1.0 - \left(\frac{x^2+y^2}{(N/2)^2}\right)}\right)}{I_0(\pi\alpha)} \quad : \quad 0 \leq \sqrt{x^2+y^2} \leq \frac{N}{2} \tag{6.24}$$

Here, N is the window size. This weight function is then normalized by letting:

$$\sum_{(x,y)} w(x,y)^2 = 1 \tag{6.25}$$

Further, I_0 is the modified zero-order Bessel function of the first kind, computed as:

$$I_0(x) = \sum_{m=0}^{\infty} \frac{1}{m!\,\Gamma(m+1)} \left(\frac{x}{2}\right)^{2m} \tag{6.26}$$

and Γ is the gamma function, which is an extension of the factorial function:

$$\Gamma(z) = \int_0^{\infty} e^{-t}\, t^{z-1}\, dt \tag{6.27}$$

The windowed images are then Fourier transformed, and the power spectrum $P(u,v)$ can be computed as the square of the magnitudes:

$$P(u,v) = \frac{F(u,v)\ ^{2}}{N^2} \tag{6.28}$$

where $F = \mathscr{F}\left[I\right]$ is the Fourier transform of the image, and (u, v) are pixel indices in Fourier space. Two-dimensional frequencies can also be represented with polar coordinates (f, θ), where f is the spatial frequency and θ is the spatial orientation. They are computed as:

$$f = \sqrt{u^2 + v^2} \tag{6.29}$$

$$\theta = \tan^{-1}\left(v/u\right) \tag{6.30}$$

Conversion back to Cartesian coordinates is given by:

$$u = f\cos(\theta) \tag{6.31}$$

$$v = f\sin(\theta) \tag{6.32}$$

Although frequencies of up to $N/2$ cycles per image are computed, it would be better to use only half as many of the lowest frequencies. Higher frequencies may suffer from aliasing, noise, and low modulation transfer [653].

The estimation of the spectral slope is then performed by fitting a straight line through the logarithm of these data points as a function of the logarithm of $1/f$.

6.4.2 Spectral Slope Analysis

The spectra of individual natural images are likely to vary, as shown in Figure 6.12. This may play a role in the rapid detection of certain scenes or objects [59]). When averaging over a sufficiently large number of images (and across orientation θ), however, a clear pattern arises: the lower frequencies contain the most power, with power decreasing as a function of frequency. In fact, on a log-log scale, amplitude as function of frequency lies approximately on a straight line. That is, the averaged spectrum tends to follow a power law, which can be well modeled with:

$$P(f) = \frac{1}{f^\beta} \tag{6.33}$$

where P is the power as a function of frequency f, and β is the spectral slope.

Figure 6.13 shows the results of analysing three image ensembles, namely, a set of 133 natural images, drawn from the Van Hateren database [316] as well as two sets of synthetic images, whereby the first set of 30 images is of subjectively higher quality than the second set, which contains 18 images [612]. The natural image ensemble yields a slope of 1.88 with a standard deviation of 0.42, which, as discussed next, turns out to be typical for natural scenes. The high- and low-quality image ensembles produced somewhat steeper slopes, with especially the low-quality ensemble showing significantly less energy at high frequencies. This is thought to be because the subjectively low quality synthetic images have relatively little fine detail.

Several studies have reported values for the spectral slope for different image ensembles. While the particulars of the image ensembles vary considerably,

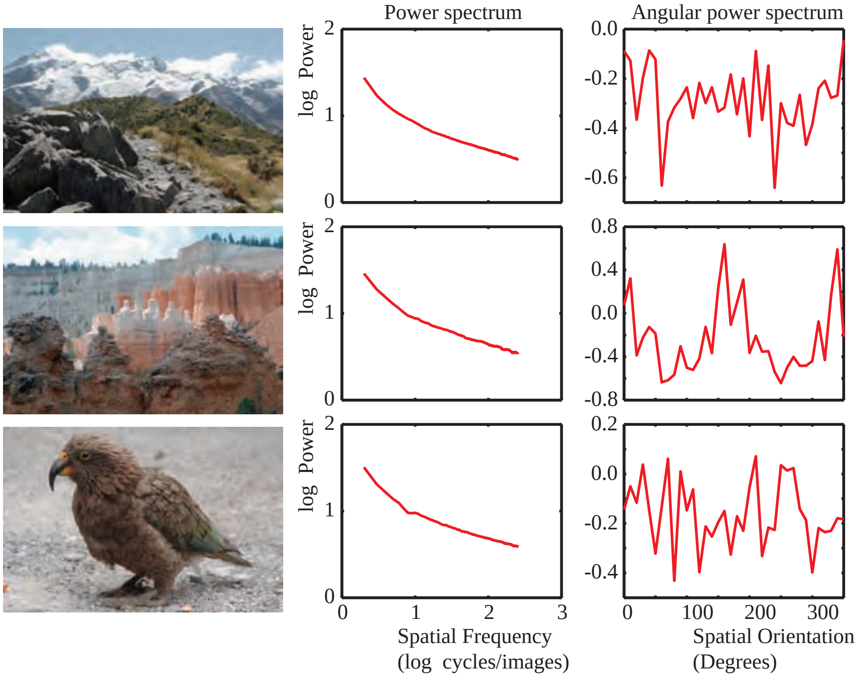


Figure 6.12. The individual power spectra are shown here for three different images (middle column). The angular power spectra are shown on the right. (Top to bottom: Mt. Cook, New Zealand, 2012; Bryce Canyon, USA, 2009; Kea at Homer Tunnel, New Zealand, 2012)

they tend to focus on natural scenes, which usually means simply eliminating carpentered environments. The average power spectral slope varies from $\beta = 1.8$ to $\beta = 2.4$, with most values clustering around 2.0 [97, 174, 185, 215, 216, 218, 312, 350, 560, 607, 612, 639, 637, 654, 739, 746, 751, 785]. A summary of findings is listed in Table 6.1. Note that if the power spectrum has a slope of around $\beta \approx 2.0$, then the amplitude spectrum would have a slope of around $\alpha \approx 1.0$:

$$A(f) = \frac{1}{f^{\alpha}} \tag{6.34}$$

As α hovers around 1.0 in many studies, the amplitude spectrum is therefore often well characterized by $A(f) \approx 1/f$, which is why this particular statistical regularity, found in so many studies, is variously described as the $1/f$ property or $1/f$ statistic.

If the power as function of frequency is given by $P(f) \approx 1/f^2$, this leads to an interesting feature. By *Parceval s theorem* we have that the volume under the

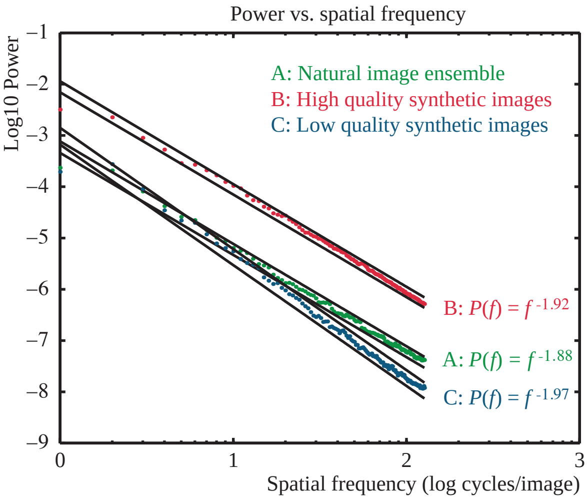


Figure 6.13. The average power spectra for 133 natural images, 30 high-quality rendered images, and 18 lower-quality rendered images.

power spectrum is proportional to the variance of an image. This by itself relates to the available amount of contrast. We can therefore study contrast in terms of the power spectrum.

If we are moving through a natural environment, we do not expect the amount of contrast of the images formed on our retinas to vary considerably. Likewise, if we zoom in to a portion of a digital image, the contrast as measured by the variance in the image is not likely to change much. We may therefore expect the energy in different frequency bands to be roughly equal [216], which points at a source of scale invariance as explained next.

A 2D image will have an energy E summed over all orientations as a function of frequency f given by:

$$E(f) = 2\pi f\, P(f) \tag{6.35}$$

If we consider a range of frequencies between f_0 and nf_0, zooming in to a portion of the image will have the effect that this band of frequencies will shift to be between af_0 and anf_0, where a is a measure of how much zoom has been applied. The energy between f_0 and nf_0 will also shift accordingly. To have an equal amount of energy at all zoom levels, and thereby an equal amount of contrast irrespective of how much an image is zoomed in to, we require that [216]:

$$\int_{f_0}^{nf_0} 2\pi f\; P(f)\; df = K \tag{6.36}$$

where K is a constant. This is indeed the case if $P(f)$ is chosen to be proportional

Study	Ensemble Size	$\beta \pm \sigma$
[97]	19	2.10±0.24
[174]	320	2.30
[185]*	95	2.29
[215]	6	2.00
[216]	85	2.20
[218]	20	2.20±0.28
[312]	117	2.13±0.36
[351]	216	1.96
[560]	29	2.22±0.26
[585]	95	2.22±0.24
[585]*	95	2.24±0.14
[607]	133	1.88±0.42
[639, 637]	45	1.81
[654]	276	1.88±0.42
[739]	82	2.38
[746]	135	2.40±0.26
[751]	12,000	2.08
[785]	48	2.26

Table 6.1. Spectral slopes for natural image ensembles (Starred studies were carried out on high dynamic range data).

to $1/f^2$, as in that case we have:

$$\int_{af_0}^{anf_0} 2\pi c \, \frac{1}{f} \, df = K \tag{6.37}$$

where the constant c is introduced to reflect the fact that the power spectrum is proportional to $1/f^2$. Equation (6.37) evaluates to:

$$2\pi c \left| \log(f) \right|_{af_0}^{anf_0} = K \tag{6.38}$$

$$2\pi c \Big(\log(anf_0) - \log(af_0) \Big) = K \tag{6.39}$$

$$2\pi c \log\left(\frac{anf_0}{af_0} \right) = K \tag{6.40}$$

$$2\pi c \log(n) = K \tag{6.41}$$

The zoom level, given by parameter a, has been factored out of this equation, which means that the amount of energy remains constant irrespective of zoom level. This is of course because the power spectrum behaves according to $1/f^2$. This implies that an image with such a power spectral composition has constant variance at all scales. In other words, it is *scale invariant* [215].

Thus, images tend to have features that appear at multiple scales, which is also a hallmark of fractal images (note that a simple linear transform of the spectral slope yields the image's fractal dimension [144]). In practice, this means that we should be able to zoom in and out of an image, or travel through a natural scene, and expect that the statistics will remain roughly constant. Thus, a scene imaged on one's retina or on the sensor of a camera $I(x,y)$ could be viewed from nearer or farther away, producing approximately $I(kx,ky)$, where k is a constant determining the amount of scaling [518]. Although this image is smaller or larger than $I(x,y)$, dependent on the value of k, it is a depiction of the same scene, and we would therefore expect it to exhibit the same statistics. This is borne out in the power spectrum of natural images, and it can also be seen in the analysis of wavelet coefficients, as discussed in Section 8.8.

As argued in Section 6.3, the power spectrum and the autocorrelation function form a Fourier transform pair. This means that the spectral slope of image ensembles can be interpreted as describing relations between pairs of pixels. Intuitively, this means that since the surface of a given object tends to be rather homogeneous, it is expected that neighboring pixels will be similar and that the farther apart two pixels are, the less similar they will be.

The $1/f$ property of images seems to arise from several sources. Edges, for example, show $1/f$ spectra [717]. Likewise, foliage and landscapes tend to exhibit fractal properties [493]. Further, the clustering of independent objects is such that the distribution of sizes in many scenes also tends to follow a power law [218, 638].

There is also some emerging evidence that the slope of the power spectrum is distinct for different scene types or objects [352, 751, 785, 585]. For example, Huang and Mumford [351] examined 216 images, which had been painstakingly segmented into pixels representing 11 different categories, and found that although the image ensemble had an average slope of 1.96, there were systematic differences in the slopes across categories. Specifically, the slopes were 2.3, 1.8, 1.4, and 1.0 for manmade, vegetation, road, and sky pixels, respectively. Likewise, Webster and Miyahara [785] analyzed the power spectra and RMS-contrast of 48 natural scenes. They found significant differences in both spectral slope and contrast across the three scene types (2.15, 2.23, and 2.4 for the forest, close-up, and distant meadow scenes, respectively).

Finally, Pouli et al. compared three different manmade categories (indoors, night, and day scenes) against natural daytime scenes [585]. Example images from their ensembles are shown in Figure 6.14. Prior to analysis, a window of 1024 × 1024 pixels was cropped from the middle of each image. This helps avoid photographer bias, as shown in the examples of Figure 6.14. The number of images in each ensemble as well as slopes β and their variances are listed in Table 6.2. In this study, it was found that carpentered environments have on average steeper slopes than natural scenes: a two-tailed, independent measures t-test revealed that the difference between the manmade categories and the natural

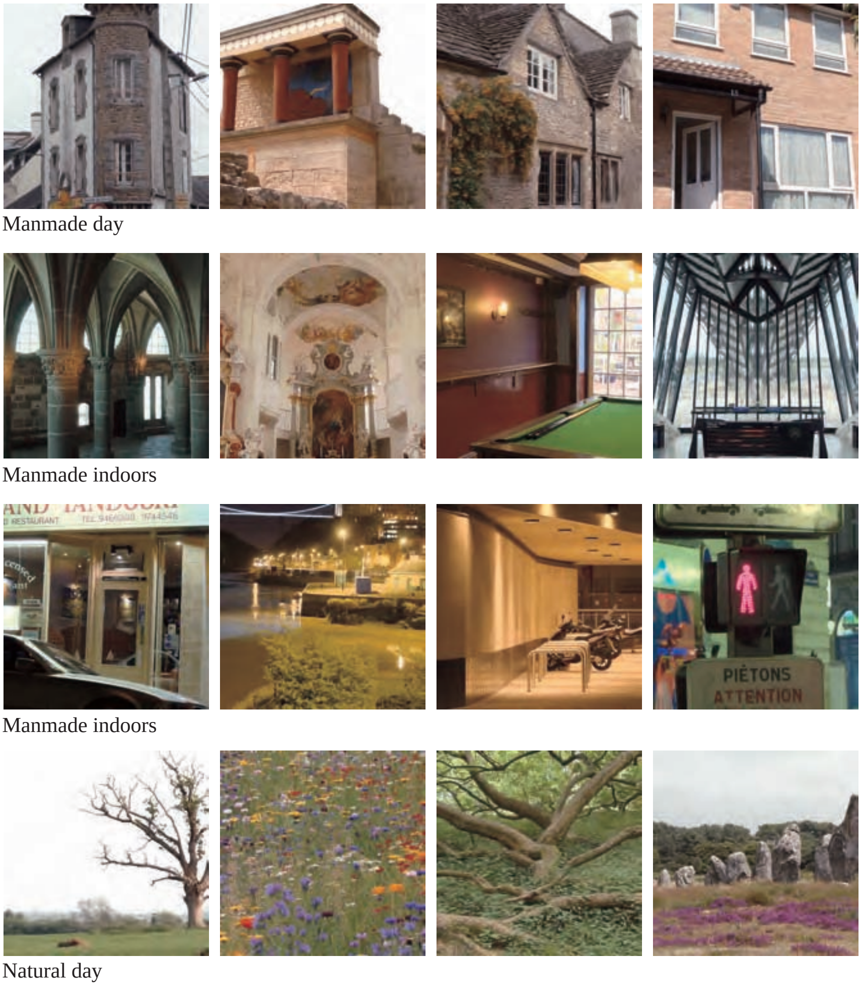


Figure 6.14. Examples from four different classes of images, used to determine to what extent natural image statistics depend on scene type [585].

one is statistically significant ($t(668) = 2.64$, $p < 0.008$), with an average slope of $\beta = 2.32$ for the manmade image classes and $\beta = 2.22$ for the natural set.

6.4.3 Dynamic Range

In addition to examining different scene types, as noted above, Pouli et al. also investigated whether the $1/f$ statistic is dependendent on the way the data is captured [585]. In particular, one might expect that high dynamic range capture tech-

Data Set	N	β	σ^2
Natural Day	95	2.22	0.242
Manmade Day	240	2.29	0.153
Indoors	125	2.44	0.121
Night	52	2.47	0.249

Table 6.2. For each of the LDR image ensembles from the study by Pouli et al. [585], the number of images N is shown, as well as spectral slope β and its variance σ^2.

		log HDR	**LDR**		
Data Set	N	β (σ^2)	β (σ^2)	**t-test**	p
Natural Day	95	2.24 (0.144)	2.22 (0.242)	t(188) = 0.390	> 0.69
Manmade Day	240	2.34 (0.099)	2.29 (0.153)	t(478) = 1.683	> 0.09
Indoors	125	2.61 (0.095)	2.44 (0.121)	t(248) = 3.977	< 0.00
Night	52	2.68 (0.152)	2.47 (0.249)	t(102) = 2.362	< 0.02

Table 6.3. Comparison of HDR and LDR image ensembles from the study by Pouli et al. [585]. The number of images N is listed, as well as spectral slope β and its variance σ^2. The t-tests show that the difference between LDR and HDR ensembles is significant for the indoors and night ensembles.

niques would allow the statistical assessment of scenes that cannot be captured with conventional imaging techniques. Many scenes have a range between light and dark that exceed the range captured by conventional cameras. This means that scenes would have to be well exposed before they can be captured, which may lead to bias in the image ensembles and therefore affect the statistical regularities computed on those images. With high dynamic range imaging techniques such limitations are removed (although in many cases there are other limitations, for instance, regarding how much movement can be present in a scene at capture time).

Pouli et al. assessed the $1/f$ behaviour of the four aforementioned scene types, namely natural images, manmade daytime images, nighttime images, and indoors images [585]. The comparison of the spectral slopes found for each category, as captured with both conventional and high dynamic range imaging techniques, is shown in Table 6.3.

Of note in this table is that there is a significant difference between the low dynamic range and high dynamic range capture techniques for the indoors and night scenes only. This means that both the carpentered and natural daytime scenes are equally well represented by both high dynamic range and low dynamic range images, suggesting that previous studies on natural image statistics remain valid, even if the ensembles were captured with conventional image capture techniques.

On the other hand, the same is not true for nighttime and indoors images. Here, the high dynamic range images especially produce much steeper slopes than for the natural image ensemble. This suggests that high dynamic range imaging is a particularly important tool for such image categories.

As spectral slopes obtained for natural images do not translate to other image categories, it appears that any computer graphics applications that would make use of $1/f$ statistics would benefit from considering image type.

6.4.4 Dependence on Image Representation

As many studies have found similar power spectra, it is likely that this statistic is relatively robust against image distortions, for instance due to the choice of camera or lens system. In particular, several post-processing distortions were evaluated, leading to the following observations [612]:

File Formats. The choice of file format does not have an appreciable effect on the $1/f$ statistic. In particular, conversion from a lossless format to a lossy format such as GIF or JPEG does not significantly alter the spectral slope, with the exception of the smoothing parameter available in the JPEG format, which can destroy high frequency content.

Gamma Correction. To correct for display nonlinearities in cathode ray tube monitors, and for reason of backward compatibility also in liquid crystal displays, images should be gamma corrected. This involves applying a power function $I_d = I^{1/\gamma}$, where I_d denotes the pixel values that are sent to the display. The gamma value of typical displays hovers around 2.2–2.4. A similar nonlinearity is encoded in many color spaces, including the often-used sRGB space. This manipulation does not appear to affect the spectral slope significantly. Figure 6.15 shows the relationship between gamma correction value and the resulting spectral slope for the rendering shown in the same figure.

Aliasing. In image synthesis, geometry is normally sampled with point samples, which are then taken to represent small areas. If a single sample is taken per pixel, then this can lead to visible artifacts such as jagged edges and Moiré patterns, collectively known as *aliasing*. Multiple samples per pixel will ameliorate the effect, as more samples will provide a better estimate of the area represented by each pixel. Aliasing does have an effect on the spectral slope, and it has been found that at least 16 samples per pixel would be required to ensure that aliasing is not a factor affecting the $1/f$ statistic.

Rendering Parameters. When synthesizing an image, the choice of rendering parameters may affect the appearance of an image. For instance, rendering soft shadows instead of hard shadows (i.e., using area lights rather than point lights) could be thought to affect the power spectrum. However, it was found that this does not have a significant effect. Likewise, computationally expensive rendering techniques such as adding diffuse interreflection do not affect the power spectrum.

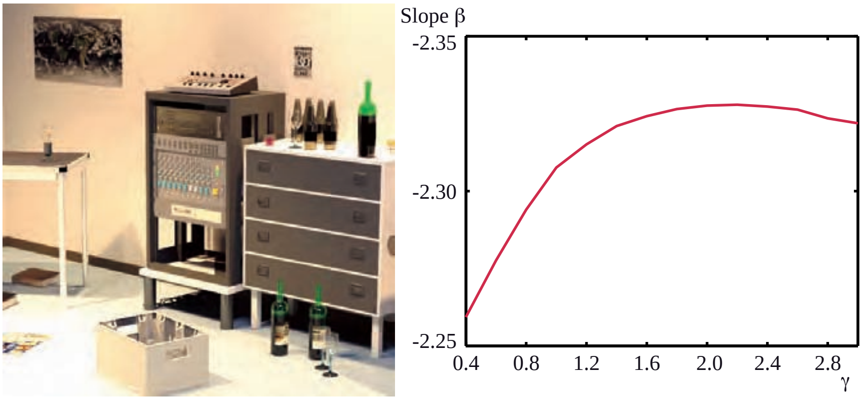


Figure 6.15. The rendering on the left was subjected to gamma correction for a range of gamma values. The resulting spectral slope is plotted on the right.

In all, this suggests that rendering has less of an effect on the power spectrum than modeling, corroborating the idea that the $1/f$ behavior relates to the geometric composition of a scene more than the way it is lit.

6.4.5 Angular Dependence

When power spectra are not averaged over all orientations, it is clear that there is some variation as a function of angle: natural images tend to concentrate their power in horizontal and vertical angles [30, 543, 638, 717, 654]. Figure 6.16 shows the angular spectra for the aforementioned natural and synthetic image ensembles. Manmade structures especially tend to show strong vertical and horizontal lines, leading to similar angular dependence, even in single images (see Figure 6.17).

Further, in examining over 6000 manmade and 6000 natural scenes, Torralba found that the slope varied as a function of both scene type and orientation (with slopes of 1.98, 2.02, and 2.22 for horizontal, oblique, and vertical angles in natural scenes and 1.83, 2.37, and 2.07 for manmade scenes) [751]. Thus, the spectral slope may be useful in object or scene discrimination [59]. It is critical to mention that if two images have similar spectral slopes, then swapping their power spectra will not affect recognition as long as phase information is preserved [746, 721].

6.4.6 Temporal Dependence

It has also been shown that the pattern of change over time follows a power law. That is, if the contrast modulation over time for a given pixel is examined, the power spectra can also be modeled with $P = 1/f^{\alpha}$, where P is the power as function of frequency f, and α is the *temporal spectral slope* [60, 174, 192].

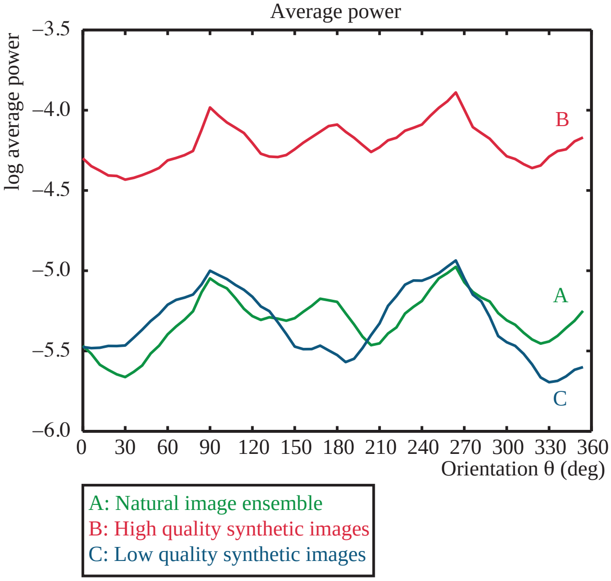


Figure 6.16. Power as a function of spatial orientation for a natural image ensemble, and high- and low-quality rendered image ensembles.

Temporal spectral slopes between 1.2 and 2.0 have been reported for natural image sequences. The temporal spectral slope relates, perceptually, to the apparent speed and jitter of the scene and to some degree with the apparent “purposefulness” of the motion (the higher the slope is, the more persistent the motion will be). Temporal frequency and other temporal phenomena are discussed in more detail in Chapter 12.

6.4.7 1/*f* Failures

Although the $1/f$ statistical regularity found in so many studies is striking, this does not mean that it holds for all imagery. It is important to note that reliable image statistics emerge when a large number of images are analyzed together. Specifically, individual images may deviate from any statistics observed in ensembles. This implies that these results should be used with care if they are somehow to be applied to individual images. In particular, the $1/f$ statistics are prone to failure for individual images, and then specifically at low frequencies, which correspond to large image scales [437].

Further, some deviations from the predicted scale invariance may be observed at high frequencies. This can be attributed to noise in images, which manifests itself in high frequencies [830].

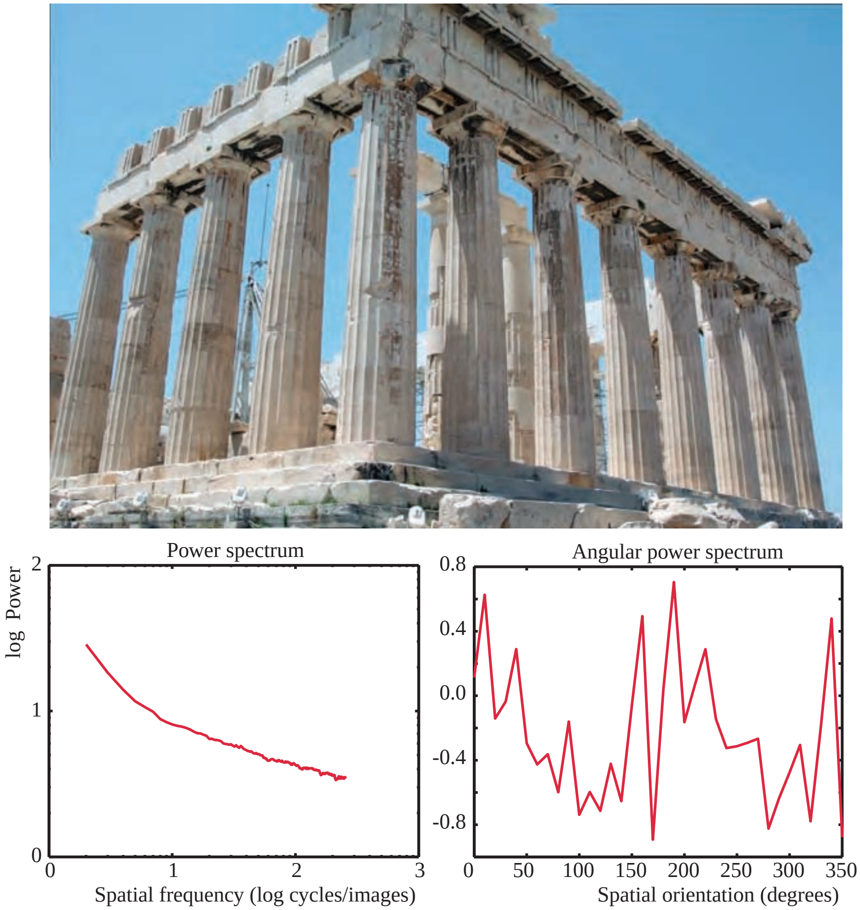
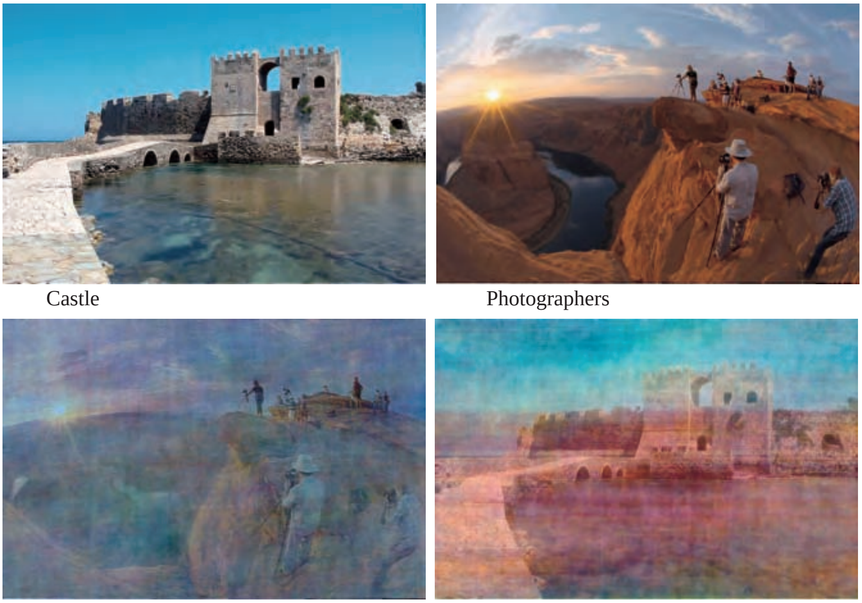


Figure 6.17. The distribution of power over spatial orientation tends to produce peaks at the horizontal and vertical direction for image ensembles. Here, a similar effect is shown for a single image of a manmade structure. (Parthenon, Athens, Greece, 2010)

6.5 Phase Spectra

It can be argued that although statistical regularities are present in power spectra of natural images, much of the perceptually relevant information is encoded in phase spectra [737, 504]. As an example, we have swapped the phase spectrum of two images, shown in Figure 6.18. As can be seen, much of the image structure has swapped. Thus, the image structure is largely encoded into the phase spectrum. Second-order statistics such as the autocorrelation function (and therefore



Castle

Photographers

Castle with photographers phase

Photographers with castle phase

Figure 6.18. In this demonstration, the phases of the top two images are swapped to produce the bottom to images. The amplitude spectra are retained. Note that much of the image structure is located in the phase spectrum and has therefore been swapped. (Left: Methoni Castle, Greece, 2010; right: Horse Shoe Bend, Arizona, 2012)

power spectra in Fourier space) as well as variance are insensitive to signal phase. They therefore do not adequately measure image structure.

To gain access to phase information without polluting the results with first- and second-order information (such as means, variances, and covariances), we can whiten the images first [737, 738]. This amounts to adjusting the spectral slope to become flat. The autocorrelation function will therefore be zero everywhere, except at the origin. Alternatively, Principal Component Analysis (PCA) can be applied to whiten a signal; see Section 7.1. By removing the second moment from consideration, it is now possible to compute skewness and kurtosis on the whitened signal. The whitened skew S_w and whitened kurtosis κ_w are thus a measure of variations in the phase spectra.

The result of applying this procedure to a set of natural images leads to the conclusion that the whitened images are almost always positively skewed and are always positively kurtosed. In contrast, if the phase spectrum is randomized on the same images, the whitened skewness and kurtosis are close to zero.

While positive skewness and kurtosis of phase spectra points in the direction of the presence of statistical regularities in the phase spectrum, these results are relatively weak and do not easily explain aspects of human vision. Furthermore, they do not appear to have found employ in any graphics-related applications that we are aware of (although it may lead to higher-order constraints in texture synthesis algorithms). In the following section, however, we will discuss how extending the present analysis to be localized in space leads to further and stronger insights. Moreover, in Section 8.11 we will discuss a wavelet-based technique to help understand phase structure.

6.6 Human Perception

There is an extremely large body of work examining the relationship between natural image statistics and human perception. At the simplest level, our ability to discriminate two random phase textures based solely on changes in the spectral slope has been examined [58, 59, 408, 722]. Humans are most sensitive to slopes around 2.8 to 3.2, which would represent an image with much less high spatial frequency content than natural images. There is some evidence (albeit controversial) for a second minimum near 1.2. Rainville and Kingdom examined the ability to detect symmetry for white noise images with different spectral slopes and found that one participant was best for slopes near 2.8, consistent with the image discrimination data [595]. The other participant was best for slopes between 1 and 2, consistent with the potential second minima.

Regardless, it is clear that humans are not maximally sensitive to changes in spectral slopes representing natural images. The reasons for this are still unclear, although several hypotheses have been forwarded including that the shift from 2 to 2.8 reflects blur perception [57].

Instead of attempting to determine the tuning of the visual system by measuring discrimination, one can approach the problem more indirectly: one can estimate the sensitivity of the spatial perception mechanisms from an autocorrelation analysis of contrast sensitivity function (which has been referred to as a modulation transfer function of the human visual system).

It is generally accepted that human spatial perception is mediated by several, partially overlapping spatial frequency channels (at least seven at each orientation: see, e.g., [57]). Since similar frequencies are likely to be processed by the same channel, the sensitivity to similar frequencies should be similar. The less similar the frequencies are, the less correlated their sensitivity thresholds should be [61, 550, 571, 661].

Billock [57] examined the correlation between contrast sensitivity thresholds as a function of the spatial frequency separation, and found that (for up to five octaves) the correlation functions were power laws with slopes ranging from 2.1 to 2.4. This held not only for static stimuli but also for slowly flickering stimuli (up to 1 Hz). These slopes are much more in line with the slopes found in natural images, suggesting that human spatial frequency mechanisms may be optimized for natural images. Interestingly, more rapid flicker yielded higher slopes (around 2.6). As mentioned above, higher slopes reflects an attenuation of the higher spatial frequencies. Billock suggested that the higher slopes for rapidly flickering images may represent motion deblurring.

In contrast, the discrimination of temporal spectral slopes appears to be more straightforward. Humans are most sensitive to differences in temporal spectral slope for slopes between 1.8 and 2.0, which is very similar to the range of slopes in natural image sequences.

The existence of spatial frequency channels in the human visual system are also implicated in lightness perception. Dakin and Bex have shown that if the amplitude of the response of these channels to natural stimuli is weighted according to their scale, i.e., with weights w_s proportional to $1/f^{-s}$, their combined response correlates well with the perception of lightness [150]. In particular, it can explain the Craik-O'Brien-Cornsweet [133] and White's illusions [795].

6.7 Fractal Forgeries

Thanks in large part to the seminal work of Mandelbrot [493], many areas of computer graphics use fractals to synthesize textures, surfaces, objects, or even whole scenes (see, e.g., [169, 170, 566]). A subset of this work focuses on fractal Brownian motion in general and fractal Brownian textures in specific, which bear striking resemblance to real surfaces and textures. Since the fractal dimension of such a fractal texture is a linear transform of the spectral slope, these works are essentially relying on the regularities in power spectra. Many of these techniques either explicitly or implicitly choose parameters so that the spectral slope will be similar to natural images. Perhaps the most famous of these synthetic scenes are the eerily familiar landscapes produced in Voss's "fractal forgeries" [766].

6.8 Image Processing and Categorization

As mentioned in Section 6.4.7, despite the fact that a power law description clearly captures the regularities in large collections of images, individual images tend not to be $1/f$. It has been suggested that differences in the spectral slope between parts of an image allow people to rapidly make some simple discriminations (e.g., the "pop-out" effect, see [59, 100, 381]). Others have speculated on the evolutionary advantage of being able to detect spectral slope [57, 59, 105, 298, 627]. Just as knowing about the statistics of natural images in general can inform us about how the human visual system works and how we might build more efficient computer vision and computer graphics algorithms, so too will an understanding of the cause of *variations* in the statistics provide insights.

A number of potential sources for $1/f$ patterns and their variations have been identified [493, 218, 352, 638, 717, 751, 785]. For example, [352] and [785] found different average spectral slopes for different scene categories (both within as well as between images); at the very least, there seem to be differences between manmade, general vegetation, forest, meadow, road, and sky elements. It has also been shown that underwater scenes have different spectral slopes [33]. To help further distinguish between object or scene categories, one can look at the interaction between power spectra and other characteristics [30, 543, 638, 717, 654]. For example, a "spatial envelope" of a scene can be constructed from the interaction between power spectra and orientation combined with some information from Principal Components Analysis (PCA; for more on the PCA see Chapter 7) [543]. This envelope yields perceptual dimensions such as naturalness, roughness, and expansion. Similar categories tend to cluster together in this scene space. This approach was later extended to estimate absolute depth using the relationship between power spectra and orientation and some information from wavelets (for more on wavelets, see Chapter 8) [751].

In a related line of work, Dror and colleagues used a variety of natural image statistics to estimate the reflectance of an object (e.g., metal versus plastic) under conditions where the illumination is unknown [181]. They employed a wide range of image statistics from simple intensity distributions through oriented power spectra to wavelet coefficient distributions.

As noted in Section 6.6, it has been suggested that the differences between the average spectral slope of 2.0 and the peak of human sensitivity to changes in slope (at 2.8) is due to deblurring. Furthermore, it has been suggested that the higher slopes for an autocorrelation analysis of human contrast sensitivity are due to motion deblurring. In an indirect examination of this claim, Dror and colleagues examined the properties of Reichardt correlator motion detections [183, 184]. While there is considerable evidence that the human visual system uses such correlators for the low-level detection of motion, it has also been shown using typical moving gratings that they signal temporal frequency and not velocity. Dror demonstrated that when stimuli that have natural image statistics are used, the response properties of Reichardt detectors are better correlated with velocity and suggest that they make a much better basis for the synthetic processing of motion than previously assumed (for more on motion and motion detection, see Chapter 12).

6.9 Texture Descriptors

Increasingly, the spectral slope is being used as a low-dimensional descriptor of texture [58, 144, 428, 627, 729]. Perceptually, the primary effect of increasing the spectral slope is to increase the coarseness of the texture [58] (as shown in

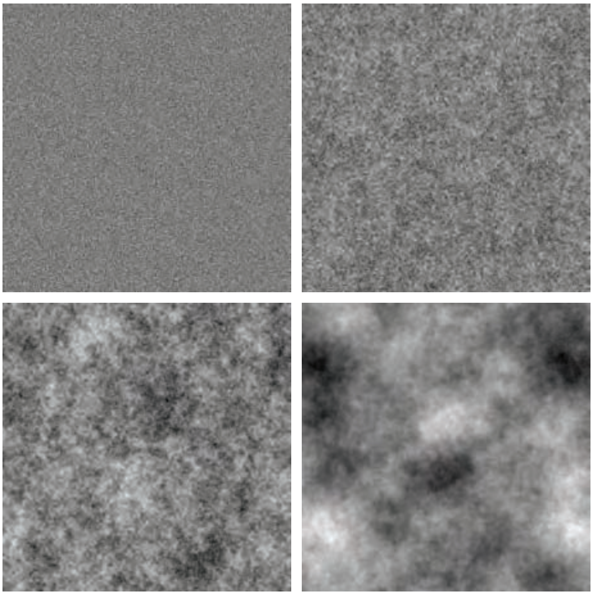


Figure 6.19. Static, random phase patches produced by $1/f^{\beta}$ spatial-frequency filtering of random white noise. The values of the spectral slope are 0.8, 1.6, 2.4, and 3.2 for the top left, top right, bottom left, and bottom right, respectively.

Figure 6.19). Indeed, Billock and colleagues have suggested that a decent model of dynamic textures can be given with the equation $A(f) = K f_s^{-\beta} f_t{}^{-\alpha}$, where K is a constant, and f_s and f_t are the spatial and temporal frequencies, respectively (for more on the separability of temporal frequency, see Chapter 12).

Likewise, several researchers have suggested that the spectral slope might provide a good estimate of the blur in an image, either by using the slope directly [76, 745] or by looking at the relative power at higher frequencies [499, 502]. The spectral slope of an image is in fact often altered to synthetically blur an image [784, 762]. Murray and colleagues, however, have shown that the perception of image blur is not well predicted by the relative energy at higher frequencies [524]. Changes in the relative phases must also be taken into account.

6.10 Terrain Synthesis

The omnipresence of the $1/f$ statistic may be leveraged in applications that generate images involving parameter tuning. For example, in procedural plant modeling [591], the well-known fact that plants tend to exhibit a strong degree of self-similarity is exploited. Many descriptive systems have been developed to take advantage of this characteristic, the most prominent of which are L-systems (see, e.g., [170] for a review). Other areas where the $1/f$ nature of natural images is employed is in displacement mapping [130, 698] or solid texture generation [569].

Here, we show the use of $1/f$ statistics in fractal terrain modeling. The simplest approach is to filter a white noise field with the desired spectral slope to produce the terrain height map [566]. In a more complex approach, called the *midpoint subdivision algorithm* [241], a patch of terrain is iteratively subdivided while displacing the subdivided patches by a random amount. As the size of the subdivided patches decreases, the amplitude of the displacement is reduced. In particular, halving the size of the patch corresponds to a reduction in maximum displacement by a factor of k. Here, k is a user parameter that determines the roughness of the terrain.

Figure 6.20 shows a set of example terrains, created by varying k between 1.5 and 2.6. In each case the number of iterations was ten, resulting in terrains consisting of 524,288 triangles. The spectral slope of these images relates to parameter k and the number of iterations (up to ten) according to the plot in Figure 6.21.

An interesting observation is afforded by showing these images to participants in an informal perceptual experiment. Simply asking observers to indicate the image which looks most natural resulted in the distribution shown in Figures 6.22 and 6.23. As can be seen, there is good correspondence between what observers consider natural, and a $1/f$ slope of around $\beta = 1.86$. Note that this very closely corresponds to the average spectral slope that was found for natural images, as discussed in Section 6.4.2.

6.11 Art Statistics

A considerable amount of research has been conducted on the frequency statistics of art. Usually the goal is to examine the relationship between the real world and depictions of it. The knowledge gained is used to either support theories of how human vision works or for the (semi-) automatic analysis of art. For example, Graham and collegues [273, 275, 277, 276, 278] examined the first- and second-order statistics of over 900 works from many different art epochs from both western and eastern cultures.

In one of their first works [273], they compared the statistics of 124 paintings to a similar number of real-world images from Van Hateren's database.[1] The art images were uncompressed TIFF photographs (taken under controlled conditions by the museum photographer) of art works from the Herbert F. Johnson Museum of Art, Cornell University (for more on the images, see [273]). A patch of 818 × 818 pixels for each image was extracted randomly and examined for a variety of single pixel and frequency-based statistics. They found that the art works had a power spectral slope of 2.46 and the world images had a slope of 2.74. Although

[1] They excluded photos with an undefined amount of blur.

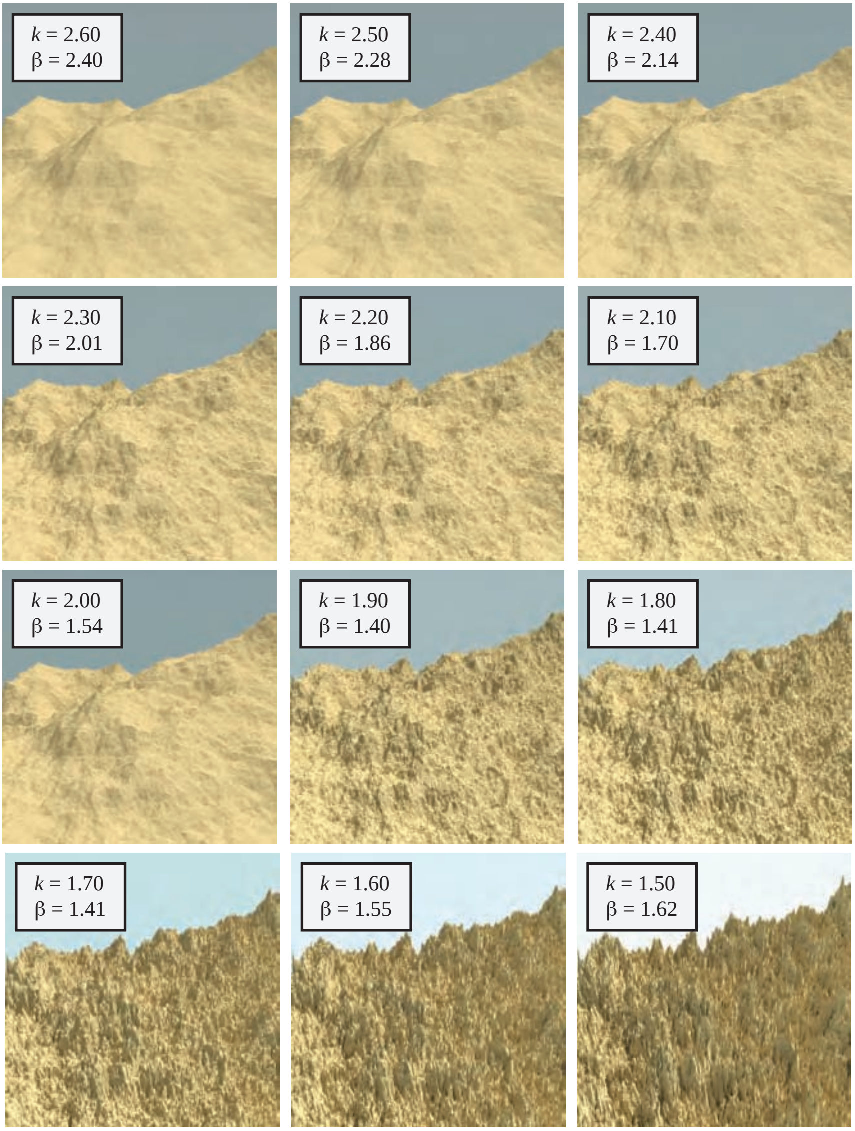


Figure 6.20. A set of 12 fractal terrains, generated with parameter *k*, resulting in a spectral slope of β.

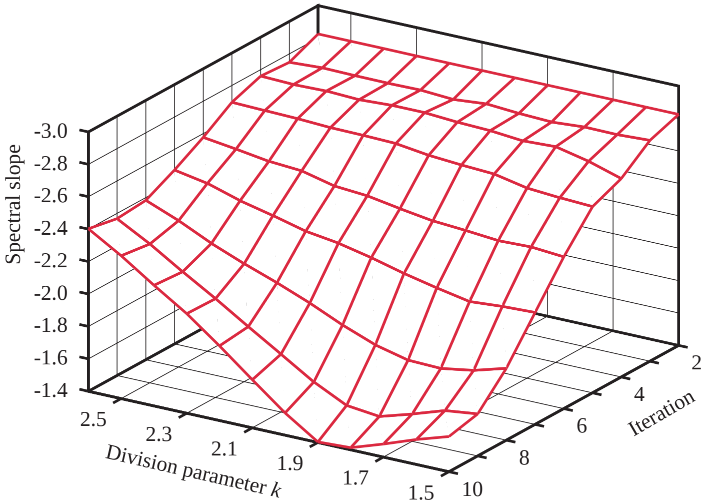


Figure 6.21. The relation between terrain roughness parameter k, the number of iterations, and the spectral slope.

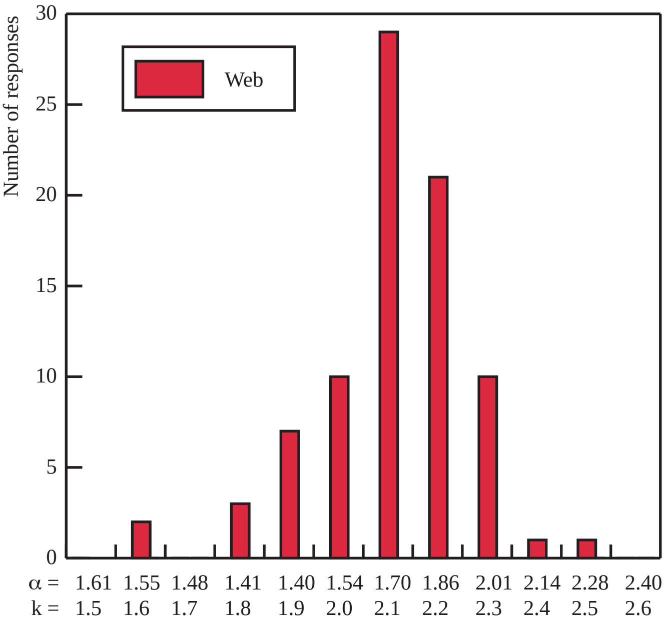


Figure 6.22. Naturalness ratings for the images of Figure 6.21. This experiment was carried out by means of a webpage, which showed images and asked viewers to send an e-mail to the author, stating the number of the image which appeared most natural.

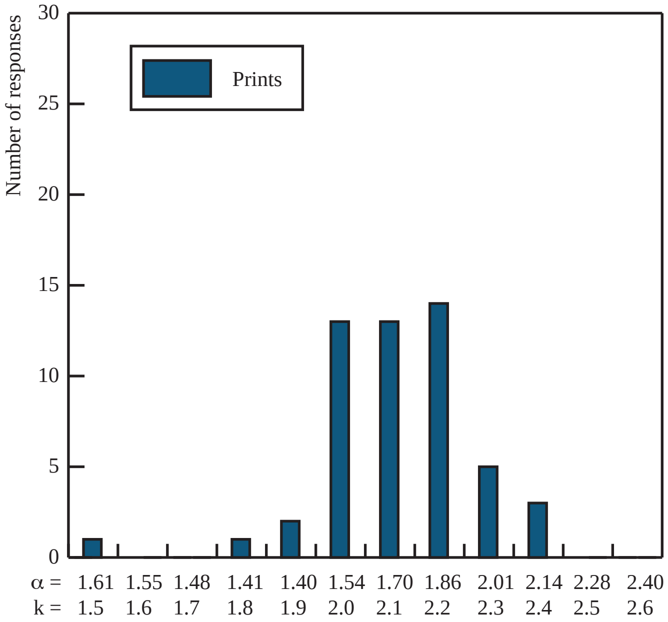


Figure 6.23. Naturalness ratings for the images of Figure 6.21. This experiment was carried out using high-quality prints, whereby participants were asked to select the print that appeared most natural.

this value is above the average found by other labs, it is not that much higher and the difference can be explained by the exclusion of blurry images. The smaller slope value for art implies a relatively smaller amount of higher-frequency information (or the absense of blur in the real-world images).

A subsequent more detailed analysis revealed that beyond blur, the difference between art and photographs seems to be driven by two factors [275]. First, eastern artworks have shallower slopes than western artworks. Second, abstract art had shallower slopes (ca. 2.23) than either landscape or portrait paintings (both roughly 2.5). Furthermore, in [274, 277], Graham and colleagues examine the luminance compression between real-world (which has a high dynamic range) and the very limited dynamic range (usually around 30:1) of paintings and found that painters use a compressive nonlinearity. That is, they use a tone mapping algorithm. They also found that once a similar nonlinearity is applied to real-world scenes, both art and real-world scenes show similar sparsity.

In a similar line of research, Redies and colleagues examined monochrome versions of 200 western artworks from a variety of epochs excluding modern art. The exclusion of modern art allowed them to focus on representational art [605, 602, 410]. The statistics of the artworks were compared to those of pictures of household objects (179 images), plant parts and natural scenes (i.e., 408 natural images from the Van Hateren database), and scientific illustrations (209 images). They found a power spectral slope of 2.1 for the paintings, 2.1 for natural scenes, 2.9 for plant parts and objects (close-up views), and 1.6 for scientific illustrations. In short, the frequency statistics of representational western art seem similar to real-world scenes, but not to close-up views of real-world scenes.

There was an astonishingly consistent slope for the artworks: the spectral slopes did not vary significantly with changes in country of origin, century, painting technique, or subject matter. Interestingly, they also found that representations of faces showed $1/f$ spectra even though photographs of faces are not $1/f$ [604]. These and other results can be related to information from neuroscience for a theory on aesthetics [602]. It has been shown that political cartoons, comics, and Japenese mangas also have roughly similar power spectra (roughly 1.99, 2.04, and 2.08, respectively) [410]. Perhaps most interesting is that the spectral slope of artworks does not vary much as a function of orientation, while it does for natural images (which can be the basis of a simple scene discriminator in photographs).

A large body of work has focused exclusively on Jackson Pollock's drip paintings (see, for example, [728, 13, 126, 368, 9, 521, 522, 523]). While the bulk of the work has focused on descriptive statistics of Pollock's work, some research addressed the issue of authenticating Pollock paintings [126, 727, 9]. Overall, it has been extensively shown that Pollock's artwork can be described well with a $1/f$ like power spectra. It has also been shown that the slope changed systematically over the course of Pollock's career.

Chapter 7

Dimensionality Reduction

Photographs of scenes typically contain a very large number of pixels. It is not uncommon for modern cameras to have in excess of 15 megapixels, and high-end cameras can have 80 megapixels or more. The pixels in an image can be seen as observations of a set of processes that gave rise to this image. Light sources emit light, all objects in the scene reflect, absorp or transmit light, and the medium through which light travels may participate by scattering light. Some light will eventually pass through the optical system of the camera and will be recorded as pixels.

The number of observations made in a single image or a set of images is usually much larger than the number of processes (think of objects in the scene). It would be possible to see the pixels in an image, therefore, as random variables that sample the same random process. The question then is whether it would be possible to learn something about the underlying process that gave rise to the samples/pixels.

If the dimensionality of the underlying process is lower than the number of observations, then the data should form clusters in N-dimensional space. An example is shown in Figure 7.1, where an image of a snowy scene is analyzed by plotting 500 randomly chosen pixels in linear sRGB space. As the process that gave rise to the scene is predominantly snowy weather, it is no surprise that the pixels lie more or less on a straight line in 3D sRGB color space. An example where the pixels stem from multiple processes is shown in Figure 7.2. Here, the butterfly and its background lead to different clusterings of pixels in 3D space.

If clusters lie in lower-dimensional planes as in this example, then there exist various types of analysis that will be able to detect what the dimensionality of the underlying process was [91]. Principal Component Analysis (PCA) is perhaps the most well-known algorithm that accomplishes this [563, 801, 377]. There are many more techniques that can be used for dimensionality reduction, including

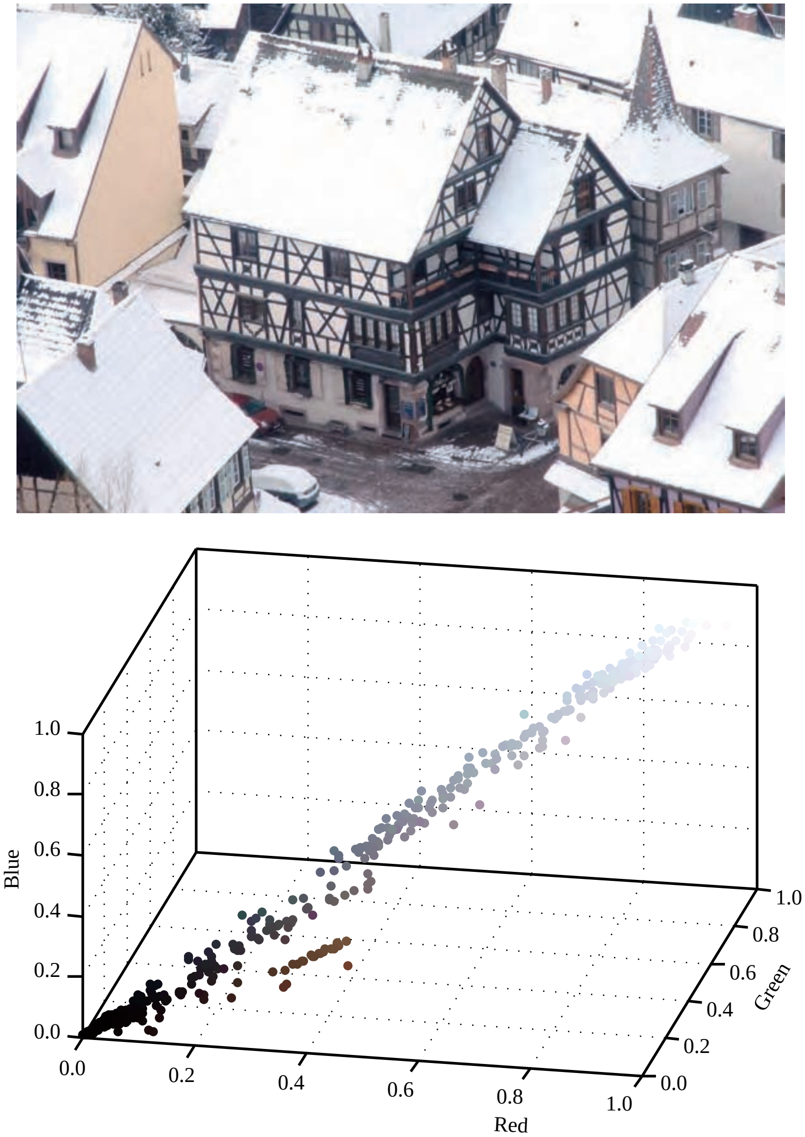


Figure 7.1. We have randomly picked 500 pixels of the image shown at the top and plotted them in linear sRGB space. As a result of the snowy weather, most pixels ended up lying on a straight line in 3D space. (Keysersberg, Alsace, France, 2013)

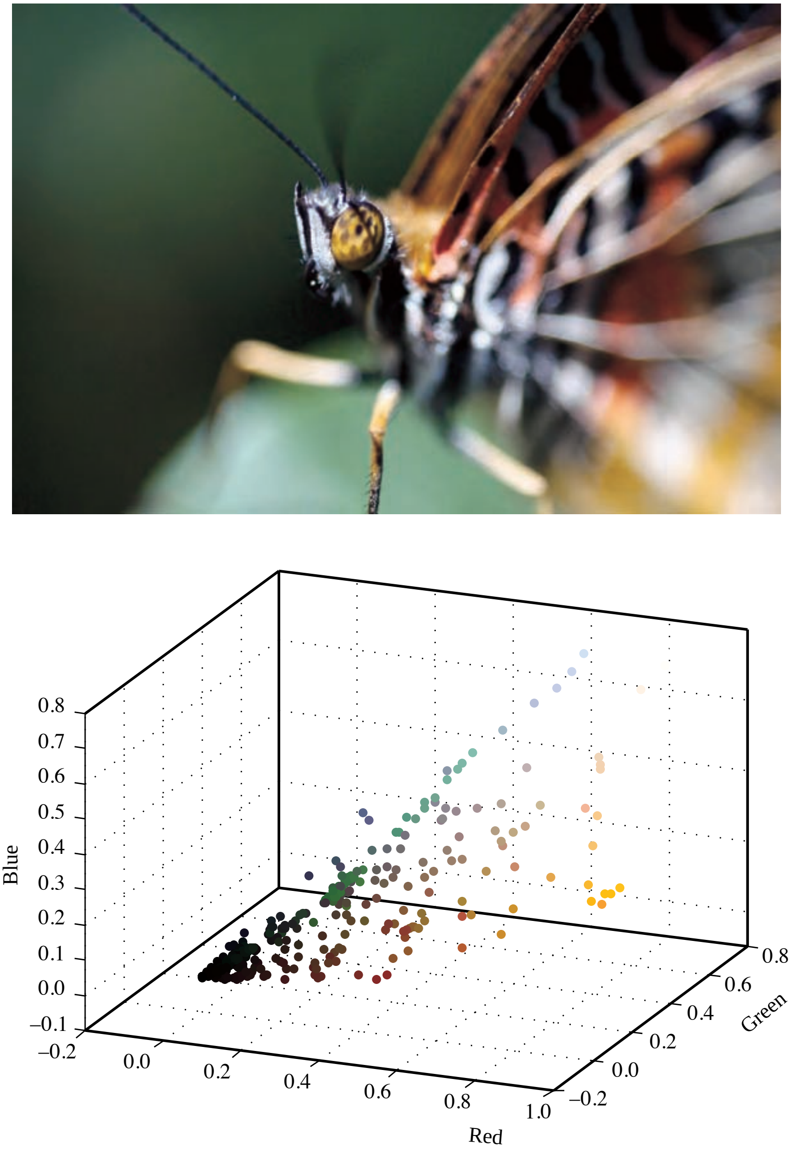


Figure 7.2. We have randomly picked 500 pixels of the image shown at the top and plotted them in linear sRGB space. The butterfly and the background are distinctly different processes, leading to less straightforward clustering than the example shown in Figure 7.1. (Changi Airport butterfly garden, Singapore, 2012)

Independent Component Analysis (ICA) [304], canonical correlation analysis [304, 734], linear discriminant analysis [505], Fisher's linear discriminant [221], topic models, and latent dirichlet allocation. Both PCA and ICA are special cases of blind source separation, which is a general collection of techniques that try to separate a set of signals from a mixed signal without prior knowledge of the source signals or the mixing process [3].

7.1 Principal Component Analysis

Principal component analysis (PCA—also known as the Karhunen-Loéve transform or the Hotelling transform) is a common data mining and dimensionality reduction technique that takes as input a series of n d-dimensional observations. Its output consists of a new set of eigenvectors (or basis vectors), as well as their eigenvalues (or weighting values). PCA is also often used in many different application areas beyond data mining and dimensionality reduction, including data visualization, variance calculations, factor analysis, perceptual experiment data analysis, and of course, image statistics. The eigenvectors point to directions of maximum variance in the data. We begin by showing how eigenvectors and eigenvalues are computed.

The first step in applying PCA consists of choosing an appropriate set of n d-dimensional input vectors. This can, for example, be a set of images for which the most important "eigenimages" are of interest. Alternatively, it would be possible to use small patches within a set of images as the input vectors. Finally, one could use individual pixels in an image, or set of images, as n three-dimensional observations.

In the standard definition of PCA, one then creates a $d \times n$ matrix $\mathbf{M}$ by concatenating all the input vectors. The covariance between two dimensions across all input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by:

$$\operatorname{cov}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{i=1}^{n} (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{2,i} - \bar{\mathbf{x}}_2)}{n-1} \tag{7.1}$$

To facilitate the computation of a covariance matrix, the data is first centered so that the mean value in each of the d dimensions is zero, i.e., $\bar{\mathbf{x}}_1 = \bar{\mathbf{x}}_2 = \ldots \bar{\mathbf{x}}_d = 0$. The covariance computation then simplifies to:

$$\operatorname{cov}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{i=1}^{n} \mathbf{x}_{1,i}\, \mathbf{x}_{2.i}}{n-1} \tag{7.2}$$

Note that the variance within a single dimension can be computed as $\text{var}(\mathbf{x}_i) = \text{cov}(\mathbf{x}_i, \mathbf{x}_i)$. The covariance matrix is then the covariance between each pair of dimensions in our matrix of input vectors M:

$$\mathbf{C} = \begin{bmatrix} \text{cov}(\mathbf{x}_1, \mathbf{x}_1) & \text{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_1, \mathbf{x}_d) \\ \text{cov}(\mathbf{x}_2, \mathbf{x}_1) & \text{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_2, \mathbf{x}_d) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\mathbf{x}_d, \mathbf{x}_1) & \text{cov}(\mathbf{x}_d, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_d, \mathbf{x}_d) \end{bmatrix} \tag{7.3}$$

As the data has zero mean, this can be efficiently computed as a matrix multiplication:

$$\mathbf{C} = \frac{1}{n-1} \mathbf{M}\mathbf{M}^T \tag{7.4}$$

We then wish to compute the eigenvectors and eigenvalues of this square matrix. Each of the d eigenvectors has an eigenvalue that gives the length of its corresponding eigenvector. It is then possible to sort the eigenvectors according to their length. In that case, the longest eigenvector will point in the direction of largest variability in the data. It is normally described as the first principal component. The second-longest eigenvector points in the direction that exhibits the second-most variability. In this manner, each subsequent eigenvector will explain less variability in the data. To achieve dimensionality reduction, it would be possible to ignore all dimensions of the data for which the corresponding eigenvalues are below some threshold.

The decomposition into eigenvalues and eigenvectors proceeds as follows. An eigenvector of a square matrix $\mathbf{C}$ is a nonzero vector $\mathbf{v}$ such that:

$$\mathbf{C}\,\mathbf{v} = \lambda\,\mathbf{v} \tag{7.5}$$

where λ is an eigenvalue. This can be rewritten as:

$$\mathbf{C}\,\mathbf{v} - \lambda\,\mathbf{v} = 0 \tag{7.6}$$

$$(\mathbf{C} - \lambda\,\mathbf{I})\,\mathbf{v} = 0 \tag{7.7}$$

where $\mathbf{I}$ is the identity matrix. There can only be a nonzero solution for $\mathbf{v}$ in the above equation if:

$$\det(\mathbf{C} - \lambda\,\mathbf{I}) = 0 \tag{7.8}$$

In other words, the determinant of $\mathbf{C} - \lambda\,\mathbf{I}$ should be zero. This equation, known as the *characteristic equation* (or secular equation), is essentially a polynomial equation of degree d in the variable λ. As we are applying eigenvalue decomposition on pixels, patches of pixels, or images, the matrix $\mathbf{C} - \lambda\,\mathbf{I}$ will be real-valued, meaning that there are at most n real-valued roots to the above polynomial equation. These could be computed numerically, giving a set of eigenvalues. By substituting these into Equation (7.5) the eigenvectors could be computed.

In practice, eigendecomposition is achieved via a different route. There exist numerical algorithms to directly compute eigenvectors and eigenvalues [434, 641], including the QR algorithm [242, 243, 429] and Arnoldi iteration [21].

Performing PCA directly on the covariance matrix of this dimensionality is often computationally infeasible. If small, say, 50×50 pixel, grayscale images are used as input to the PCA, each image is a 2500-dimensional vector, and accordingly the covariance matrix $\mathbf{C}$ has 6.25×10^6 elements.

There is, however, a trick based on linear algebra that can be used to simplify this: the rank (that is, the maximum number of linearly independent rows or columns) of the covariance matrix $\mathbf{C}$ is limited by the number of training examples. If there are n training examples, there will be at most $n-1$ eigenvectors with nonzero eigenvalues. Therefore, if the number of input feature vectors n is smaller than the dimensionality d of each vector, the principal component analysis can be performed on a relatively small $n \times n$ sized matrix, instead of the larger $d \times d$ sized one.

To achieve this, we calculate the covariance matrix $\mathbf{C} = \mathbf{M}^T \mathbf{M}$. The eigenvalue decomposition then creates n eigenvectors $\mathbf{v}_i$ and corresponding eigenvalues λ_i. It can be proven that if $\mathbf{v}$ is an eigenvector of $\mathbf{C} = \mathbf{M}^T \mathbf{M}$, then $\mathbf{v} = \mathbf{M}\mathbf{v}$ is an eigenvector of $\mathbf{C} = \mathbf{M}\mathbf{M}^T$, having the same eigenvalue λ. This can be seen as follows. Let us assume that $\mathbf{v}$ is an eigenvector of $\mathbf{M}^T \mathbf{M}$ with eigenvalue λ. We then have:

$$\left(\mathbf{M}^T \mathbf{M}\right) \mathbf{v} = \lambda \mathbf{v} \tag{7.9}$$

$$\mathbf{M}\left(\mathbf{M}^T \mathbf{M}\right) \mathbf{v} = \mathbf{M} \lambda \mathbf{v} \tag{7.10}$$

$$\left(\mathbf{M}\mathbf{M}^T\right)\left(\mathbf{M}\mathbf{v}\right) = \lambda\left(\mathbf{M}\mathbf{v}\right)\left(\mathbf{C}\right)\left(\mathbf{M}\mathbf{v}\right) = \lambda\left(\mathbf{M}\mathbf{v}\right) \tag{7.11}$$

This proves that $\mathbf{v} = \mathbf{M}\mathbf{v}$ is an eigenvector of covariance matrix $\mathbf{C}$.

As stated before, components with a larger eigenvalue correspond to more important features in that they explain more of the variance in the data. Conversely, lower eigenvalues signify less important components. In order to identify how many dimensions might be needed for dimensionality reduction, it is possible to sort the eigenvalues. The amount of explained variance up to a dimension $d_k < d$, then, is the sum of all eigenvalues up to that point:

$$\sum_{i=1}^{d_k} \lambda_i \tag{7.12}$$

Depending on the desired fidelity of the smaller, reconstructed space, it is then possible choose an appropriate cut-off point α that selects the k most important dimensions:

$$k \mid \sum_{i=1}^{d_k} \lambda_i < \alpha \sum_{i=1}^{d} \lambda_i \tag{7.13}$$

The original stimuli or feature vectors (as well as any new stimulus) can then be reconstructed using the reduced set of corresponding eigenvectors. This representation offers the advantage that a relatively small number of components is sufficient to encode most of the information in the images.

Nevertheless, although the computed eigenvectors are orthogonal, statistical independence cannot be guaranteed. Natural images tend to be non-Gaussian and thus the decomposition offered by PCA can only decorrelate the data [365]. The main consequence of this is that meaningful information is only captured by the first few components, while further ones mostly correspond to less important features. That this still affords useful information is demonstrated in the following paragraph.

7.1.1 Whitening

One of the strengths of PCA is that it can be used to decorrelate data. The eigenvalues are a measure of the variance along each dimension. If we were to scale decorrelated data to have unit variance in each dimension, we would obtain whitened data, i.e., data that is uncorrelated and of unit variance. This means that we would have removed all second-order effects from the data, “second-order” meaning variances and correlations. As such, we would be able to use the resulting images to study higher-order statistics. In Section 7.2, we demonstrate that removal of second-order statistics from the data is a necessary prerequisite for computing independent components.

7.1.2 PCA on Pixels

An example of PCA analysis of the pixels in an image is given in Figure 7.3. The three-dimensional pixel data was analyzed, leading to three eigenvectors that are plotted in the figure. Note that the point cloud is stretched most in the direction of the first principal component. In this figure, we adjusted the length of the axes for the purpose of visualization; the actual length of the second and third principal components is significantly shorter than seen in this figure.

Applying PCA to single pixels of either individual images or ensembles of images is useful in color applications. In essence, it is possible to decorrelate the three color channels by running PCA on images in this manner. As discussed in detail in Chapter 10, decorrelation of color information appears to happen in human vision, and it has direct applications in image processing.

7.1.3 PCA on Patches

We can also apply PCA to small patches, drawn randomly from one or more images. The output is then an ordered set of patches whereby the first principal components represent the most common modes of variation in local image regions. An example is shown in Figure 7.4, where the high dynamic range input

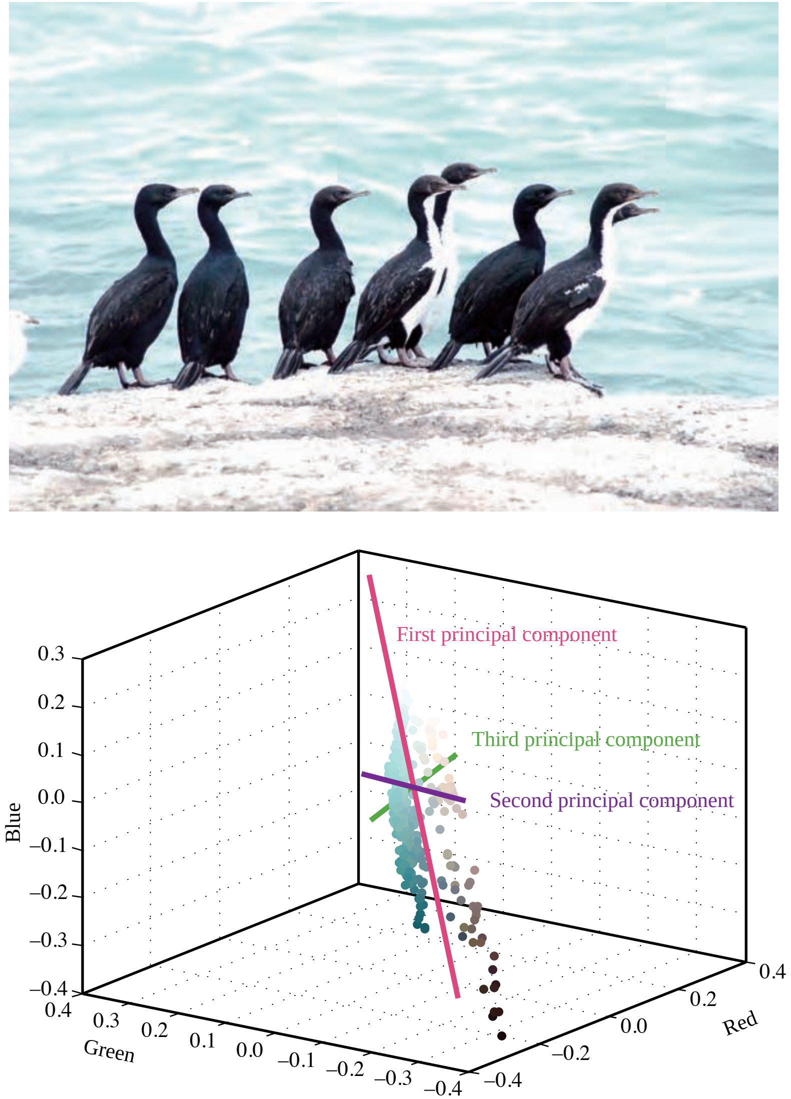
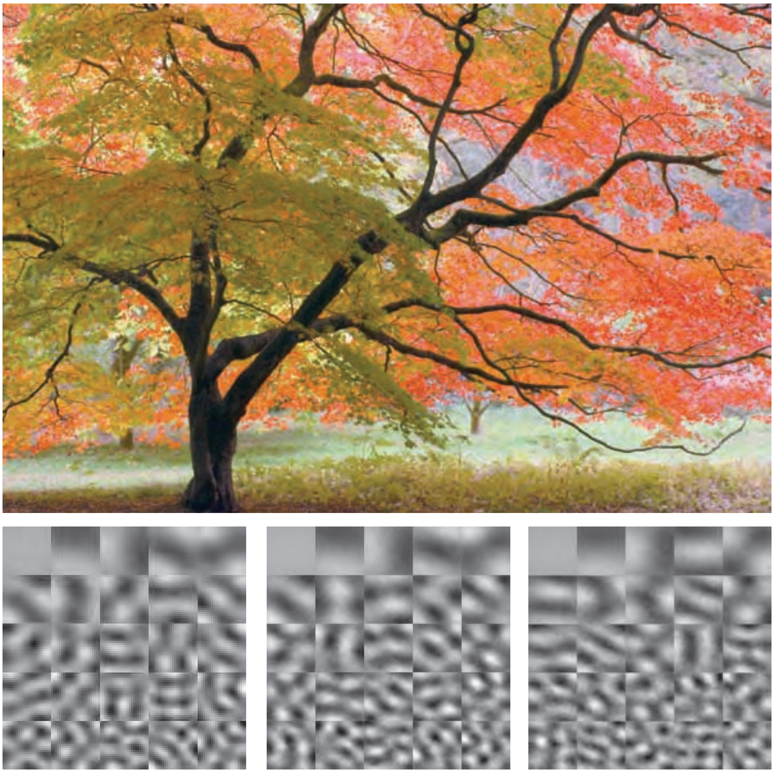


Figure 7.3. PCA analysis on three-dimensional pixel data produces three directions, which are ordered according to how much variation each direction explains. (Dunedin, New Zealand, 2012)



25 x 25 patches

50 x 50 patches

100 x 100 patches

Figure 7.4. PCA run on 5000 randomly chosen patches taken from the high dynamic range image at the top. The first 25 principal components are shown in reading order for patches of size 25 × 25 (bottom left), 50 × 50 (bottom middle), and 100 × 100 (bottom right). (Westonbirt Arboretum, UK, 2009)

image was transformed to a luminance-only image. We then subtracted the mean luminance and randomly selected 5000 patches of given size (see Figure 7.4). The patches are then linearly transformed into one-dimensional vectors by concatenating the rows of the patch. As before, the resulting 5000 vectors are placed in a matrix that can then be subjected to PCA. The resulting basis vectors $\mathbf{v}_i$ are then reshaped into 2D patches $\mathbf{V}_i(x,y)$ for visualization. Figure 7.4 shows the first 25 principal components for different patch sizes ranging from 25×25 pixels to 100×100 pixels.

Irrespective of how large the patches are, the first principal component gives an average response, being mostly uniform in luminance distribution. The second and third principal components tend to encode horizontal and vertical gradients/edges. The fourth and fifth principal components appear to encode horizontal

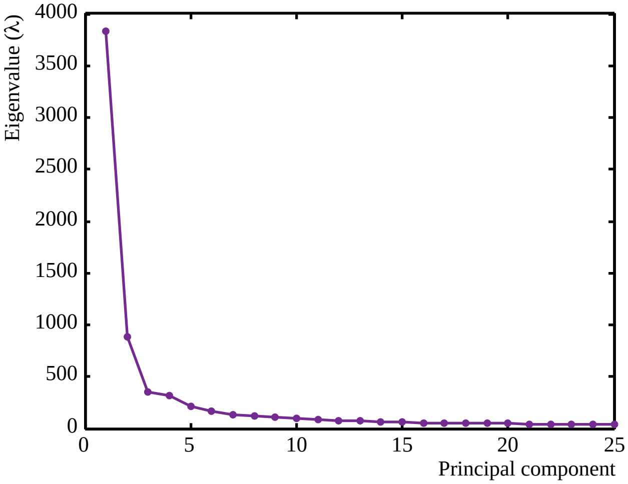


Figure 7.5. The first 25 eigenvalues corresponding to the eigenvectors of size 100 × 100, which are shown in the bottom right of Figure 7.4.

and diagonal lines. The remaining principal components show increasingly complex patterns, which are less meaningful.

The eigenvectors are only good features if the multidimensional data contains at least a few dimensions which carry a strong signal—if the data contains significant levels of noise, then the interpretation of the eigenvectors becomes difficult. This can usually be checked by investigating how quickly the eigenvalues fall off as the dimensionality becomes larger: if the falloff is steep in the beginning and then tapers off, then the first few eigenvectors are potential candidates for "good" features. The relative importance of the eigenvectors of Figure 7.4 are plotted in Figure 7.5. We note that the falloff is very significant for the first few components and then tapers off, which is consistent with our ability to assign meaning to the first few components.

Results such as these are also found in studies whereby ensembles of natural images are taken as input [299], although in this work the patches were windowed with a Gaussian window first. They observe that the first few principal components begin to resemble receptive fields of simple cells, for example in cats [548].

A consequence of applying PCA to image patches is that in principle an image patch can be represented to a reasonable degree by the first few components. Effectively, a patch $\mathbf{I}(x,y)$ could be represented by a linear superposition of basis functions. The principal components $\mathbf{V}_i(x,y)$ in such a case act as the basis functions [546]:

$$\mathbf{I}(x,y) = \sum_i w_i \, \mathbf{V}_i(x,y) \tag{7.14}$$

For human vision, the implication is that only a few different types of receptive fields can account for a relatively large proportion of our visual experience.

While it is unlikely that the human visual system applies PCA to its input, certain receptive fields do bear some resemblance to the patches obtained by this method.

PCA, however, assumes that either the data has a Gaussian distribution (which is already violated, for instance, in the image of Figure 7.2, where the point cloud is not ellipsoidal), or that linear pairwise correlations are the most important form of statistical dependence [546]. As many, if not all natural images have higher-order dependencies, PCA is limited in representing such images as basis functions, and the technique also does not fully explain the receptive field structure in the human visual system.

Thus, one of the main applications of PCA is in the realm of feature extraction in that it finds a good, new set of basis vectors (or features) for the data. For this interpretation to hold, however, the data must be distributed according to a single, Gaussian distribution. For non-Gaussian data, or for multimodal Gaussian data, PCA only decorrelates the axes. For highly clustered data, PCA may therefore not be a good method.

On the other hand, the results obtained with patch-based PCA do point into an interesting direction, which is that images may be analyzed with a small set of basis functions that capture the dominant trends in the image. This leads to an important concept in natural image statistics, which is that of sparse coding. It is now believed that the human visual system aims to preserve information, but to process and represent it sparsely, i.e., with as few neurons active as possible [217, 546]. If the variability in the input signal can be represented by as few neurons as possible, then this has several advantages for organism: such neuronal activity is metabolically efficient [29], it minimizes wiring length [235] and it increases capacity in associative memory [39].

The concept of sparse coding returns in the discussion of Independent Component Analysis (ICA) in Section 7.2, as well as wavelets in Chapter 8.

7.1.4 PCA on Images

So far, we have discussed the application of PCA on pixels as well as on patches of images. It is also possible to apply PCA on entire images. Each image is then considered a separate feature vector, and are known as eigenimages. Such an approach can be useful if the set of images contain carefully constrained exemplars. Variations in illumination, reflectance properties, location, and orientation or the objects in the image can complicate processing severely [507, 520]. Robustness is often a problem, in that the method does not handle outliers or noisy data very well [358, 636].

Nonetheless, various approaches have been proposed to overcome these limitations [449], which have led to various applications, including illumination planning [519], tracking of robot manipulators [528], visual inspection [813], and human face recognition [56, 756], the latter of which is discussed in the next section.

7.1.5 Eigenfaces

Perhaps the most famous example of PCA in the context of image analysis is the extraction of so-called *eigenfaces* within the computer vision domain [756], the principles of which are illustrated in Figure 7.6. The input consists of n images of $p \times p = d$ pixels. The eigenfaces are the eigenvectors resulting from the PCA decomposition of the input face images. Since PCA is a global method, each of the d pixels is treated as single dimension. This means that any difference in background, lighting, pose, and size of the faces will become encoded in the PCA eigenvectors, which usually is not desirable.

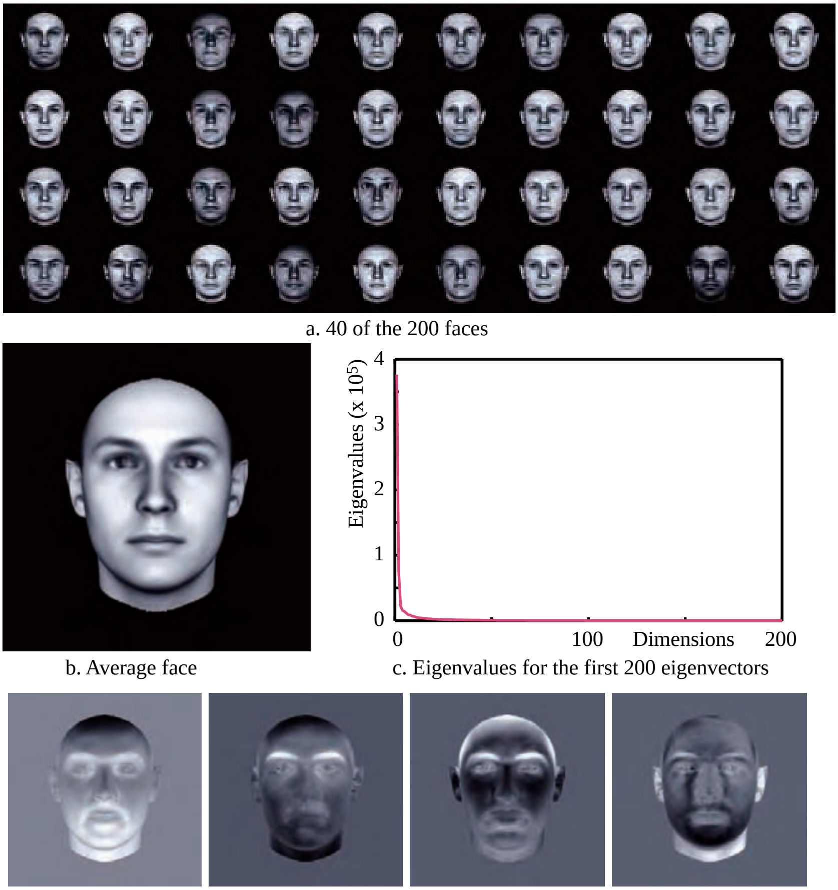
If one is interested only in the variation in appearance of the faces, then one possibility is to bring all faces into correspondence by warping them onto a common face coordinate system. Each face will then have the exact same shape with each pixel in the image exactly specifying the same facial features in all faces. The only variation left will then be the one according to the facial appearance (or texture) of the face or variations in illumination, which PCA can extract.

Figure 7.6 shows an example in which a database of 200 three-dimensional faces (100 male, 100 female) was brought into correspondence first [65]. The faces were then rendered onto the shape of the average face (Figure 7.6a). Figure 7.6b shows how fast the eigenvalues fall off, indicating that the first few eigenvectors explain a large amount of variance. Furthermore, Figures 7.6c–f show the first four eigenvectors, or eigenfaces: the first eigenface highlights the change in appearance from top to bottom, the second focuses on the eyebrows, the third on the cheeks, and the fourth on the chin and cheeks (the beard region). Although eigenfaces do not always correspond to such well-defined features, they are able to capture the main modes of variation in the images.

It has been shown that such a separation between form on the one hand and texture on the other hand corresponds well to human recognition performance characteristics [300]. Furthermore, the idea that faces are represented in a vector space (known as the *face space* [759]) might be compatible with a representation of the average face in a coordinate system defined by PCA dimensions [300, 450].

PCA has also been shown to be effective as a statistical approach in the reconstruction of 3D representations of human faces on the basis of a single 2D images [24]. In this work, a low-dimensional parameterization of head shape was learned from a large set of scans of 3D fases. A single 2D input image can then be mapped to 3D using this parameterization.

Finally, PCA has not only been applied to analysis of static faces, but also to the modeling of facial expressions. Indeed, applying PCA to a set of faces varying in identity and expressions yields an interesting separation of identity and expression axes, which is also compatible with behavioral and neurophysiological data, suggesting a large dissociation between the two types of information [101].



a. 40 of the 200 faces

b. Average face

c. Eigenvalues for the first 200 eigenvectors

d. The first 4 Eigenvectors (Eigenfaces)

Figure 7.6. PCA can be used to decompose a collection of face images into a set of components, which can be used for constructing novel faces or for face recognition. (a) Forty of the 200 faces from the MPI face database [755]—note that only the texture of the face varies but not its shape. (b) Average face of all 200 faces. (c) Plot of the eigenvalues of the PCA showing how quickly the values fall off as the number of dimensions increases. (d) First four eigenvectors (Principal components, or *eigenfaces*) showing where changes in the face texture happen. (Figure used with permission from the copyright holder Christian Wallraven.)

7.2 Independent Components Analysis

The orthogonality of the axes of PCA can be seen as a limitation. A more advanced, albeit computationally much more involved technique is Independent Component Analysis (ICA) [365, 364, 127, 128, 366]. Rather than producing decorrelated axes that are orthogonal as achieved with PCA, independent components analysis finds axes that are more or less independent, but which are not necessarily orthogonal, as shown in Figure 7.7.

Several ICA algorithms are known, including infomax [43], extended infomax [448], fastICA [363], Jade [363], kernel ICA [28], and RADICAL [446]. Although the implementations vary, their goals are the same: represent multivariate data as a sum of components in such a way that the components reveal the underlying structure that gave rise to the observed signal.

The number of ICA algorithms available and their computational and algorithmic complexity make an in-depth discussion of this class of algorithms beyond the scope of this book. However, we will discuss the general principle as well as some of their implications for natural image statistics as well as for human vision.

While PCA measures covariances of pixels separated by varying distances, the components that are found constitute a linear transform. This means that although the data may be decorrelated, there is no guarantee that the resulting components are in any way independent. In particular, only if the input data happens to have a Gaussian distribution, then the resulting principal components are both decorrelated and independent [640].

As we have seen, many of the statistics of natural images that are currently known point in the direction of high kurtosis, i.e., they are highly non-Gaussian. This has given rise to the use of ICA in examing natural image statistics, where image patches [366] or the color components of a set of pixels (as shown in Figure 7.7) are calculated. ICA could also conceivably be used on sets of entire images (although this would probably be prohibitively expensive computationally).

ICA algorithms tend to require a set of preprocessing steps to make the problem more tractable. In particular, the data need to be centred and whitened. In addition, data reduction is often applied. Data can be whitened by running the PCA algorithm first, as outlined in Section 7.1.1. By keeping only the first n components, data reduction is achieved. This ensures that only those components are computed that will be meaningful to the problem being solved. Moreover, it speeds up the computations required to determine independent components.

Note that if the multivariate data being analyzed has a Gaussian distribution along each of its dimensions, there would be no higher-order statistical correlations available in the data. In that case, applying PCA would not only decorrelate the data but would also lead to the computation of independent axes. Thus, for ICA to be meaningful, there is the requirement that the distribution of data points is non-Gaussian.

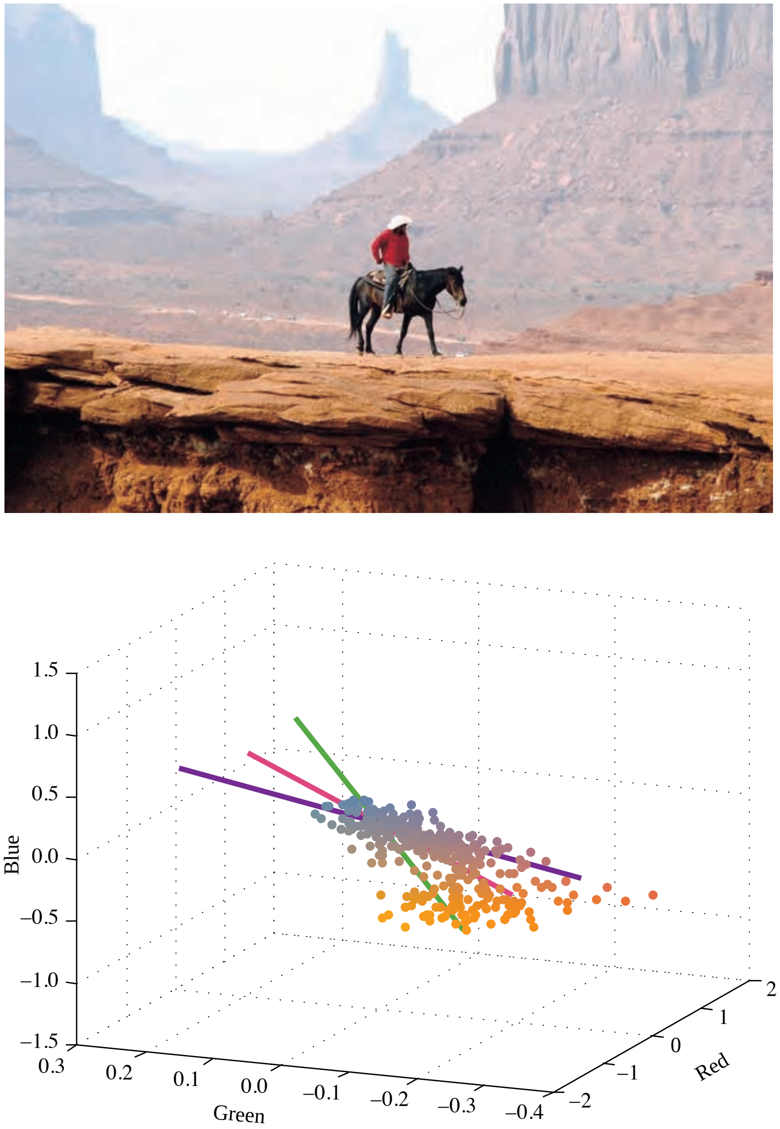


Figure 7.7. In this plot, we have chosen 50,000 points from the image at the top, and applied ICA to the color channels of these points, revealing three new axes that are not orthogonal—the main difference with PCA. Note that we only plot 500 randomly chosen points here. Further, the axes were scaled for visualization; the pink axis does not correspond to a meaningful basis, it being drawn more than 3000 times longer than the other two axes. (Monument Valley, Arizona, 2012)

Assuming for now that we are performing ICA on small image patches randomly drawn from a set of images, the pixels $\mathbf{I}(x,y)$ in the patch can be represented as the linear superposition of a weighted set of basis functions $\mathbf{V}_i$:

$$\mathbf{I}(x,y) = \sum_i w_i \, \mathbf{V}_i(x,y) \tag{7.15}$$

where w_i are the weights. This is the same equation as the one used for PCA (Equation (7.14)), albeit that the assumptions used for the construction of the basis functions are different. In particular, the assumptions made for ICA are [365]:

- The weights w_i can be seen as random variables that have a non-Gaussian distribution. They are also assumed to be statistically independent.
- The components $\mathbf{V}_i(x,y)$ are invertible.

Under these assumptions, and given a large enough set of image patches, it is possible to derive a set of components $\mathbf{V}_i$ without advance knowledge of any of the weights w_i. A property of linear systems is that the generative model described by Equation (7.15) is equivalent to:

$$w_i = \sum_x \sum_y \mathbf{V}_i^{-1}(x,y)\,\mathbf{I}(x,y) \tag{7.16}$$

$$= \sum_x \sum_y \mathbf{Q}_i(x,y)\,\mathbf{I}(x,y) \tag{7.17}$$

This means that either set of equations can be solved, and the components computed for either system can be converted into the system by matrix inversion. Similar to the argumentation in Section 7.1.3, we can concatenate the values in the patches $\mathbf{Q}_i$ to form vectors $\mathbf{q}_i$, and similarly rearrange the image patch $\mathbf{I}(x,y)$ to vector $\mathbf{k} = k_1, \ldots, k_m$ so that we can write:

$$w_i = \sum_{j=1}^{m} \mathbf{q}_{i,j}\,\mathbf{k}_j = \mathbf{q}^{\mathrm{T}}\,\mathbf{k} \tag{7.18}$$

We now have an expression that relates the independent components $\mathbf{q}_i$ to the weights w_i. Recall that these as yet unknown components could be combined by applying the weights to each corresponding component, and then we would obtain the original image patch, now represented by $\mathbf{k}$. The weights w_i can therefore be interpreted as random variables drawn from a certain probability density function (pdf). The probability density for each component $\mathbf{q}_i$ is then given by $p_i(w_i)$.

By the definition that the components we seek are independent, the multidimensional probability density of all the n weights combined is given by:

$$p(w_1, \ldots, w_n) = \prod_{i=1}^{n} p_i(w_i) \tag{7.19}$$

Unfortunately, we do not know the pdf of the weights w_i. Instead, we would like to find the pdf of the observed variable $\mathbf{k}$ (recall that this is a vector of pixel values that serves as input). Due to the linear transform applied in Equation (7.18), this pdf is given by:

$$p(\mathbf{k}) = \ \det(\mathbf{R}) \ \prod_{i=1}^{n} p_i\left(\mathbf{q}_i\,\mathbf{k}\right) \tag{7.20}$$

Here, the matrix $\mathbf{R}$ is the matrix that defines the linear transform in Equation (7.18), i.e., $\mathbf{R} = \ \ q_{i,j} \ \ $.

Equation (7.20) constitutes the statistical model for independent component analysis. To estimate parameters given a statistical model, a technique known as *maximum likelihood estimation* is normally used. The unknown parameters are the independent components $\mathbf{q}_i$. If we are using a large number A of image patches, then we can construct an equally large number of vectors $(\mathbf{k}_1, \ldots, \mathbf{k}_A)$. The likelihood is then the probability of observing these input vectors given the model parameters. The likelihood $L(\mathbf{q}_1, \ldots, \mathbf{q}_n)$ is given in log space by:

$$\log L(\mathbf{q}_1, \ldots, \mathbf{q}_n) = A \log \ \det(\mathbf{R}) \ + \sum_{i=1}^{n} \sum_{a=1}^{A} \log p_i(\mathbf{q}_i^{\mathrm{T}}\,\mathbf{k}) \tag{7.21}$$

It can be shown that if the features are uncorrelated, as they would be after whitening using PCA, then the determinant of $\mathbf{R}$ will be ± 1. As a result, the log-likelihood can be estimated as:

$$\log L(\mathbf{q}_1, \ldots, \mathbf{q}_n) = \sum_{i=1}^{n} \sum_{a=1}^{A} \log p_i(\mathbf{q}_i^{\mathrm{T}}\,\mathbf{k}) \tag{7.22}$$

Of course, we would like to choose model parameters such that this log-likelihood is maximized. Maximum likelihood estimation can be performed by standard numerical optimization techniques [193, 222, 220, 510], although specific optimization techniques have been developed in the context of ICA [365, 363].

7.3 ICA on Natural Images

One could ask what the independent components are in the context of natural images. It could be argued that these are the objects that comprise a scene [547]. It is exactly these that provide the structure that gives rise to the pixel array that is analyzed by computer algorithms as well as by the human visual system. Both PCA and ICA, however, are generative models that are based on linear superposition. Thus, they cannot take into account translation and rotation of objects [547].

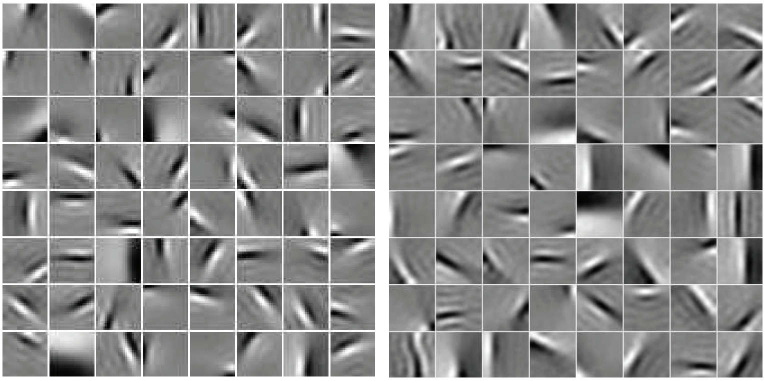
However, it has been found that small image patches of natural image ensembles can be analyzed using ICA. It should be noted here that we are still free to choose our assumed probability density functions $p_i(w_i)$. So far, ICA is a general technique for blind source separation. To make ICA relate to natural images, the

Figure 7.8. A set of 15 images used in the independent component analysis in Section 7.3. (Westonbirt Arboretum, UK, 2009 and 2011)

pdfs should be chosen such as to mimic a presumed feature of human vision. A reasonable assumption would be to argue that receptive fields in the human visual pathway are in some sense sparse, i.e., cells preferentially respond little if at all to most stimuli, and they respond vigorously to some stimuli to which they are tuned. Such sparse coding is likely to occur and helps, for instance, with metabolism: a cell that is dormant most of the time does not use precious energy.

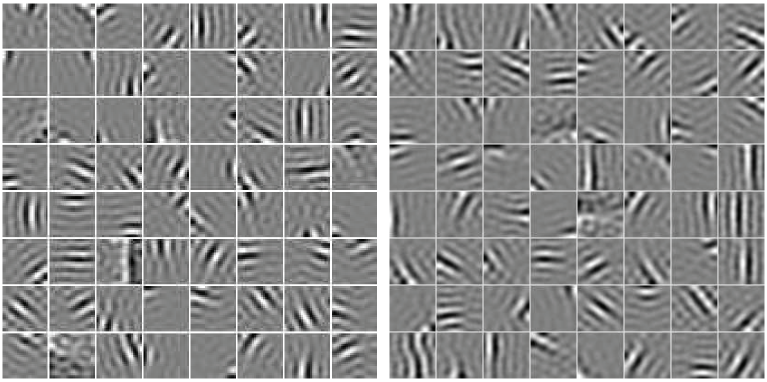
The sparseness of a probability distribution function could, for instance, be measured by computing the kurtosis of the observed variables. Kurtosis is, in fact, a direct measure of how far a distribution deviates from being Gaussian. Larger values of kurtosis indicate a sparser distribution. It would therefore be possible to choose pdfs that maximize kurtosis. Several alternative functions have been proposed, including [365]:

$$\log p_i(w_i) = -2 \log \cosh \left(\pi s / 2 \sqrt{3} \right) - 4 \sqrt{3} / \pi \tag{7.23}$$



Basis functions (25 × 25 pixel patches) Basis functions (50 × 50 pixel patches)

Figure 7.9. Basis functions $\mathbf{V}_i$ computed using 50,000 patches of size 25 $\times$ 25 (bottom left) and 50 $\times$ 50 (bottom right), drawn from the 15 images presented in Figure 7.8.



Feature vectors (25 × 25 pixel patches) Feature vectors (50 × 50 pixel patches)

Figure 7.10. Features $\mathbf{V}_i^{-1} = \mathbf{Q}_i$ computed using the 15 images presented in Figure 7.8.

As an example, we have selected 15 natural images, shown in Figure 7.8, and created a set of 50,000 randomly chosen patches of size 25×25 or 50×50 pixels. We then applied a fast ICA algorithm to this data [363]. A resulting set of 64 feature patches, corresponding to the basis functions $\mathbf{V}_i$ is shown in Figure 7.9.

The feature detector weights, i.e., those corresponding to $\mathbf{V}_i^{-1} = \mathbf{Q}_i$, are shown in Figure 7.10. As these are the patches that would be applied to the image to generate the encoding of an image, they could be compared to receptive fields in the human visual system. In that context, as argued by others [44], it is interesting to note that such basis functions reveal structures that resemble those found in the human visual system [44].

In particular, this technique yields Gabor-like patches—elongated structures that are localized in space, orientation, and frequency. Their size, aspect ratio, spatial frequency bandwidth, receptive field length, and orientation tuning bandwidth are similar to those measured in cortical cells [316]. These results lend credence to the argument that the human visual system appears to represent natural images with independent variables, each having a highly kurtotic distribution that leads to a metabolically efficient sparse coding. A number of researchers have shown that the statistics of art can also be captured with very similar sparse coding, especially if a compressive nonlinearity is applied as a tone mapper [360, 359, 410, 545].

As an example, ICA decompositions have been used in object classification, specifically on images of human faces as well as flowers [388].

7.4 Gaussian Mixture Models

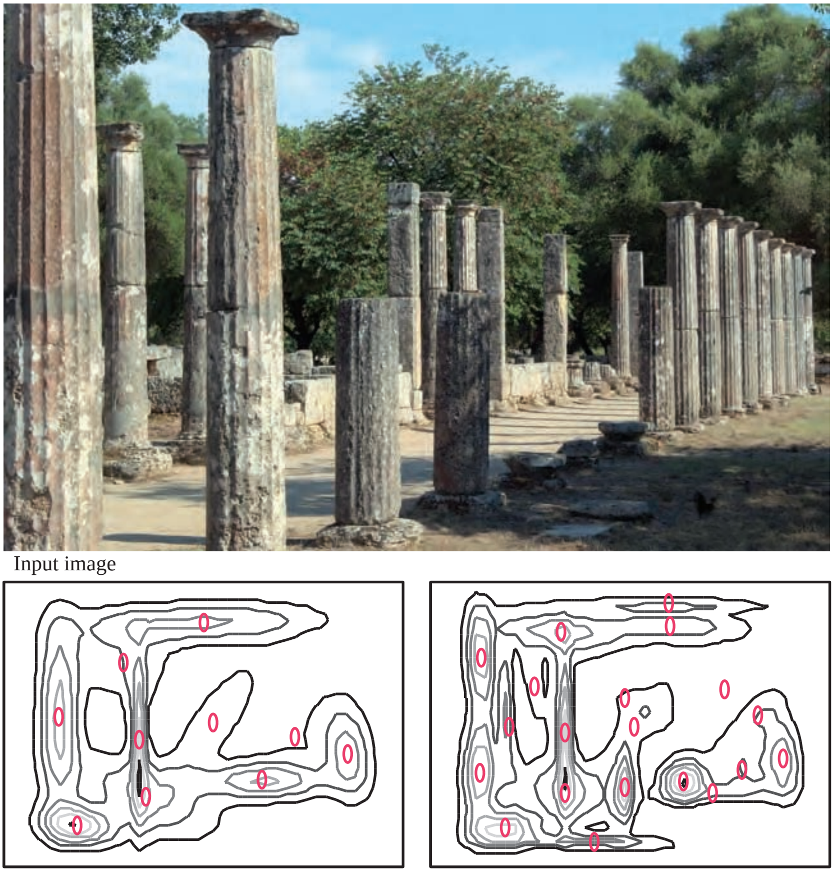
The analysis of image patches can be performed in many different ways, each revealing subtly different information about how the human visual system may function, and each offering different opportunities in imaging applications. We have seen in Section 5.2.8, for example, that image patches tend to recur nearby, both in scale and in space. In this chapter, we have shown that the human visual system appears to analyze images by removing correlations, as evidenced by the analyzers that are revealed by applying principal component analysis to individual patches. This was taken one step further in the section on independent components analysis, showing that independence is often achieved, rather than mere decorrelation.

An alternative way to analyze image patches is by applying Gaussian Mixture Models [831, 832]. A Gaussian mixture model is simply a sum of k (multivariate) Gaussian distributions, each with a weight factor w_j:

$$p(\mathbf{x}) = \sum_{i=1}^{k} w_j \, N(x \, \mathbf{mu}_j, \Sigma_j) \tag{7.24}$$

Here, the weights are prior probabilities and are therefore limited in range to $0 \leq w_j \leq 1$. Moreover, these weights sum to 1. See Appendix A.2 for the definition of multivariate Gaussian distributions. Given a set of observations, which can be individual pixels, patches, or entire images, a predetermined number of Gaussians can be fitted. This is accomplished with numerical optimization in the form of expectation maximization (EM) [166].

As an example, we have estimated an image with 10 and 20 Gaussians and plotted the resulting contours in Figure 7.11. Here, a single color channel is treated as a 2D dataset of observations. The Gaussians therefore, to some extent, represent the image. The computational complexity of the algorithm is such



Input image

10 Gaussians

20 Gaussians

Figure 7.11. The image at the top was fitted with 10 and 20 Gaussians. (Olympia, Greece, 2010)

that it would not be tractable to represent an image with sufficient Gaussians. This problem could be alleviated by employing Gaussian mixtures in a scale-space approach, however [583, 733].

It is entirely possible to compute a Gaussian mixture for image patches as well as for individual pixels. The latter case has proven a useful representation in color transfer applications [724]. This application is discussed further in Section 10.4.

Application of Gaussian mixture models to image patches drawn from a natural image ensemble has produced interesting insights as well as applications in image segmentation and image querying [113], as well as image restoration [831] and image denoising [832, 583].

One approach to image denoising would be to define the expected log-likelihood of all patches occurring in a given image $\mathbf{I}$ [831]. This is computed by summing the log-likelihood of each (possibly overlapping) patch $\mathbf{I}_i$, after applying a given prior p.

$$E_p(\mathbf{I}) = \sum_i \log(p(\mathbf{I}_i)) \tag{7.25}$$

Note that this does not give the log probability of a full image, as patches $\mathbf{I}_i$ may overlap. Given a noisy image $\mathbf{J}$, the corruption can generally be modeled with:

$$\mathbf{AJ} - \mathbf{I} \ \ ^2 \tag{7.26}$$

If we would wish to find an uncorrupted image $\mathbf{I}$ given a corrupted image $\mathbf{J}$, then this can be expressed as a minimization problem consisting of a likelihood term (Equation (7.26)) and a prior term (Equation (7.25)):

$$\arg\min \ \ \mathbf{AJ} - \mathbf{I} \ \ ^2 - \lambda \ E_p(\mathbf{I}) \tag{7.27}$$

where λ controls the relative weight of the prior term. The log-likelihood of a patch is given by:

$$\log p(\mathbf{I}_i) = \log\left(\sum_{k=1}^{K} w_k \ N(\mathbf{I}_j \ \mathbf{mu}_j, \Sigma_j)\right) \tag{7.28}$$

where the number of Gaussian mixture components is K. This approach has been shown to produce favorable results for both patch restoration and image restoration tasks, as compared to various other techniques that include PCA and ICA [831].

Chapter 8

Wavelet Analysis

The Fourier series for periodic signals, as discussed in Chapter 6, is used in many applications, including the analysis of natural images. It is the first example of a signal expansion [763], and as shown before, it uses sines and cosines as its basis functions. A Fourier expansion provides insight into which frequencies exist in an image by means of the amplitude spectrum, as well as where they exist through the phase spectrum.

However, a Fourier expansion does not allow an analysis of local structure. This means that it is not possible to use a Fourier expansion to understand which frequencies and locations exist in a local area around a given pixel of interest. In other words, Fourier basis functions are localized in frequency, but not in space. A further disadvantage is that discontinuities in images require a large number of basis functions to be adequately represented. Figure 6.8, for instance, shows that summing 500 sinusoids is only very roughly beginning to approximate a step edge. Representations that require only a small number of basis functions to represent the image are useful in, for instance, data compression. Wavelets, which are discussed next, are one such example.

As discussed in Section 6.4.1, Fourier series are intended for periodic signals. Boundary effects can be minimized by applying a window to the center of the image. This strategy can be extended to create local Fourier bases, effectively centering a set of windows on a grid, and applying a Fourier transform for each window. This is known as the Gabor transform, short-time Fourier transform (STFT), or short-term Fourier transform [422]. Such an approach gives information about each grid location of the image and could therefore be used to assess local information.

Given that images are typically nonstationary, i.e., their statistical structure can vary by image location, an assessment of local features has revealed additional patterns unique to natural images and has even revealed a correlation with saliency [741]. In this chapter, we discuss methods that have been used in the analysis of local structure in images—in particular, wavelets.

Wavelets are an alternative decomposition of a function into sets of basis functions. As opposed to the Fourier transform, it can localize the analysis in both space and frequency, often allowing for a sparser representation [153, 490, 711]. This means that fewer basis functions are needed to represent the original function, which is, for instance, directly exploited in image compression [668]. Wavelets are amenable to statistical analysis, and due to their spatially local form they are able to reveal statistical regularities of local structure. In this section, we first outline the mathematical background to wavelets, followed by a presentation of the discoveries made by analyzing wavelet decompositions of natural images.

8.1 Wavelet Transform

An alternative to the Gabor transform is the wavelet transform, which can compute a linear expansion of a signal or image by using different scalings and shifts of a prototype wavelet. As the scale factors tend to be powers of 2—i.e., each subsequent scale is a factor of two larger—the frequencies represented at each scale halve. This means that the frequency axis is logarithmic rather than linear, as is used in Fourier and Gabor transforms.

The aforementioned wavelet itself is a basis function, which is analogous to the sines and cosines that form the bases in Fourier series. There are, however, many different choices of basis functions, each with their own trade-offs. In all cases, though, the basis function needs to be well localized and integrate to zero in order to qualify as a potential wavelet.

Wavelets can be defined in terms of a single function, if suitably parameterized. This function is normally called the *mother wavelet* and in one dimension is denoted by $\psi(x)$. The parameter a indicates scaling (dilation), whereas b indicates translation:

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\,\psi\left(\frac{x-b}{a}\right) \tag{8.1}$$

where $(a,b) \quad \mathbb{R}^+ \times \mathbb{R}$. Examples of specific wavelets ψ are given in Section 8.4 and onwards. We can define the Fourier transform of this wavelet as:

$$\Psi(\omega) = \int_{\mathbb{R}} \psi(x)\, e^{-ix\omega}\, dx \tag{8.2}$$

Certain functions $f(x)$ can be represented by applying a wavelet transform to them. A requirement of such functions is that they are square-integrable, i.e.:

$$\int_{-\infty}^{\infty} \; f(x) \;^{2}\, dx < \infty \tag{8.3}$$

Square-integrable functions are said to be in $\mathbb{L}^2$. Pairs of such functions, $f(x)$ and $g(x)$, will have an inner product $\ f,g\ $ defined as:

$$\ f,g\ = \int_{-\infty}^{\infty} f(x)\,\overline{g(x)}\,dx \tag{8.4}$$

where we note that the inner product of a function with itself is less than infinite due to the requirement of square integrability: $\ f,f\ < \infty$. The wavelet transform of a function $f(x)\quad \mathbb{L}^2$ is given by $W_f(a,b)$:

$$W_f(a,b) = \ f,\psi_{a,b}\ = \int_{-\infty}^{\infty} f(x)\,\overline{\psi_{a,b}(x)}\,dx \tag{8.5}$$

showing that the wavelet itself needs to be square-integrable. Using Calderóns reproducing identity, it is possible to reconstruct the original function from the wavelet transform:

$$f(x) = C_{\psi}^{-1} \int_{0}^{\infty} \int_{-\infty}^{\infty} \frac{W_f(a,b)}{a^2}\ \psi_{a,b}(x)\,db\,da \tag{8.6}$$

where

$$C_{\psi} = \int_{0}^{\infty} \frac{\ \Psi(\omega)\ ^2}{\omega}\,d\omega \tag{8.7}$$

This inverse transform requires that the admissibility condition be satisfied, which means that $C_{\psi} < \infty$. This implies that wavelets integrate to 0, i.e., $\int \psi(x)\,dx = 0$.

The wavelet transform has several interesting properties, which include shifting and scaling properties, as well as localization. Here, we assume that functions $f(x)$ and $g(x)$ have wavelet transforms $W_f(a,b)$ and $W_g(a,b)$. The shifting property states that if $g(x) = f(x-k)$, we have $W_g(a,b) = W_f(a,b-k)$. The scaling property states that if $g(x) = f(x/k)/\ \overline{k}$, then $W_g(a,b) = W_f(a/k,b/k)$. If $f(x)$ is zero everywhere, except at a single point in space x_0, i.e., $f(x) = \delta(x-x_0)$, then $W_f(a,b) = \psi((x_0-b)/a)/\ \overline{a}$.

The wavelet transform is redundant in that a function of one variable x is transformed into a function of two variables (a,b). It is possible to select a subset of values for a and b and still have an invertible transformation. The coarsest sampling that can be chosen for a and b is called the *critical sampling*, and is given by:

$$a = 2^{-j} \tag{8.8}$$

$$b = k\,2^{-j} \tag{8.9}$$

where $k,j\quad \mathbb{Z}$ are suitably chosen integers. Any coarser sampling will not allow the original function $f(x)$ to be uniquely recovered. Any finer sampling will cause redundancy to remain. Note that a can be thought of as the reciprocal of frequency. It can be shown that under certain conditions this sampling produces an orthonormal basis:

$$\psi_{j,k}(x) = 2^{j/2}\,\psi(2^j x - k) \qquad\qquad j,k \quad \mathbb{Z} \tag{8.10}$$

where dilation is governed by parameter j and translation is given by k.

To simplify the subsequent discussion, we will now give an example of a wavelet, which is also the oldest and arguably the simplest wavelet transform. It is known as the *Haar wavelet* [295]:

$$\psi(x) = \begin{cases} 1 & x \quad [0,1/2) \\ -1 & x \quad [1/2,1) \\ 0 & \text{otherwise} \end{cases} \tag{8.11}$$

We see that this function is 0 everywhere, except between 0 and 1, i.e., it has local support. The set of functions $\psi_{j,k} \;\; j,k \quad \mathbb{Z}$ forms an orthogonal basis in $\mathbb{L}^2$, i.e., the space of all square-integrable functions.

8.2 Multiresolution Analysis

For the choice of a and b outlined above, wavelet transformations can be analyzed as follows [489, 488]. We would like to efficiently represent a function $f(x)$ in some hierarchical fashion. The space of functions that we are interested in is the set of square-integrable functions, i.e., $f(x) \quad \mathbb{L}^2$. We could impose a hierarchy on this space, creating nested subspaces:

$$V_0 \subset V_1 \subset V_2 \subset \ldots \subset \mathbb{L}^2 \tag{8.12}$$

Each of these spaces contain functions, starting with the smallest subspace V_0, the *reference space*, that contains only one function family. Each subsequent subspace then extends this space. The function in V_0 is known as the generating function or *father wavelet*, and is denoted by $\phi(x)$. Having a generating function allows the number of nested subspaces to be finite. In the case of the Haar wavelet, the generating function is given by:

$$\phi(x) = \begin{cases} 1 & x \quad [0,1) \\ 0 & \text{otherwise} \end{cases} \tag{8.13}$$

This is the box function, defined to be 1 between $x = 0$ and $x = 1$. It is also known as the *characteristic function* or the *indicator function*. It is possible to construct an orthonormal basis for V_0 by introducing translation:

$$\phi_{0,k}(x) = \phi(x-k) \qquad\qquad k \quad \mathbb{Z} \tag{8.14}$$

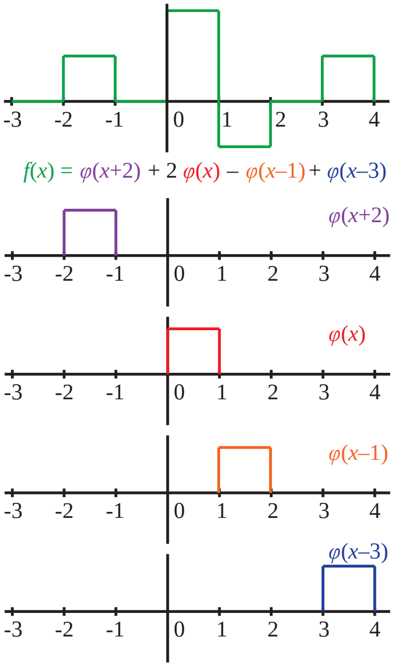


Figure 8.1. A stepped function $f(x)$ with steps at integer locations can be approximated by summing instances of basis functions $\phi_{0,k}$ in reference space V_0.

Here we have introduced the subscript 0 to indicate that this family of functions forms an orthonormal basis for the reference space V_0. With this family of functions we can approximate any function $f(x)$ with step functions which are defined at unit intervals by appropriately summing these basis functions. In particular, we can compute scaling coefficients $c(n)$ from the characteristic function, and wavelet coefficients from $d(j,n)$ from the wavelet function $\psi(x)$ for a signal $f(x)$:

$$c(n) = \int_{-infty}^{\infty} f(x)\ \phi(x-n)\ dt \tag{8.15}$$

$$d(j,n) = 2^{j/2} \int_{-\infty}^{\infty} f(x)\ \psi(2^j t - n)\ dt \tag{8.16}$$

An example for $c(n)$ is shown in Figure 8.1.

We cannot approximate any functions that step at places other than at integer positions within reference space V_0. However, we could augment V_0 with some additional functions that create a new subspace V_1. For instance, we could introduce a new subspace W_0 such that:

$$V_0 \oplus W_0 = V_1 \tag{8.17}$$

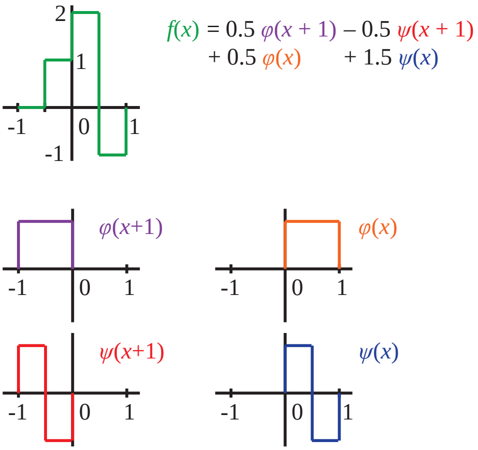


Figure 8.2. A stepped function $f(x)$ with steps at half-integer locations can be approximated by summing instances of basis functions $\phi_{1,k}$ in reference space V_1.

This means that W_0 is the complement of V_0 to V_1. To extend V_0 in a meaningful way, we could use translated versions of $\psi(x)$ as defined in (8.10) to form subspace W_0. Note that this function also has local support in that it is nonzero between 0 and 1. However, it makes a step from $+1$ to -1 at $x = 1/2$. It is therefore not a function that is already in V_0 and is orthogonal to any of the functions $\phi_{0,k}(x)$ in V_0. The translated versions of $\psi(x)$ are given by:

$$\psi_{0,k}(x) = \psi(x-k) \qquad\qquad k \quad \mathbb{Z} \tag{8.18}$$

The functions $f(x)$ that can be represented in V_1 are now all step functions that step at half-intervals. By going from V_0 to V_1 we have therefore refined our representation, admitting more detailed functions. An example of a function that can be represented by the basis functions in $V_1 = V_0 \oplus W_0$ is given in Figure 8.2.

We note that the principle of extending a subspace V_0 by W_0 to obtain subspace V_1 holds for all other subspaces V_j, i.e., in general we have $V_{j+1} = V_j \oplus W_j$. This also implies that by substitution V_j can be written as $V_0 \oplus W_0 \oplus W_1 \oplus \ldots \oplus W_j$. As a result, to represent a signal we only need the father wavelets $\phi_{0,k}(x)$ from reference space V_0 in addition to wavelet functions $\psi_{j,k}(x)$ from all subspaces V_j, although it is certainly possible (and often done) to construct a wavelet decomposition by means of a sequence of scaling functions $\phi_{j,k}(x)$ combined with a sequence of wavelet functions $\psi_{j,k}(x)$.

To extend the subspace V_1 to V_2 and beyond, the translation of functions by k is not sufficient. We also need to scale our basis functions, an operation known as *dilation*, to be able to represent all functions that step at intervals spaced by 2^{-j}. The amount by which each function is scaled depends on which subspace they are defined in. So, for subspace V_j and W_j, our basis functions $\phi_{j,k}(x)$ and $\psi_{j,k}(x)$ are written:

$$\phi_{j,k}(x) = 2^{j/2}\ \phi(2^j x - k) \qquad\qquad j,k \quad \mathbb{Z} \tag{8.19}$$

$$\psi_{j,k}(x) = 2^{j/2}\ \psi(2^j x - k) \qquad\qquad j,k \quad \mathbb{Z} \tag{8.20}$$

Note that only in the reference space V_0 and W_0 no dilations occur, as $2^0 = 1$. Assuming Haar wavelets, one way to think of this decomposition is that $\phi_{0,k}(x)$ represents averages, while $\psi_{j,k}(x)$ encodes deviations from the average at some scale. Alternatively, the scaling function acts as a low-pass filter, whereas the wavelet function acts as a band-pass filter.

Dilating by 2^j means that in each subsequent level the range of frequencies considered is halved/doubled. In other words, the band-pass filters with octave bandwidths and center frequencies that are one octave apart.

If enough scales are considered, in the limit continuous functions $f(x)$ could be approximated. In other words, the union of all subspaces is *dense*. However, as seen above, Haar wavelets do this by representing the function $f(x)$ as a sequence of piecewise constant segments. As the decomposition is into discontinuous functions, Haar wavelets are good at representing edges but not so good at representing smooth functions.

8.3 Signal Processing

Wavelets can be explained (and implemented) within a signal processing framework. Because V_0 is a subset of all functions in V_1 it should be possible to write the scaling functions in V_0 as a linear combination of the basis functions in V_1:

$$\phi(x) = \sum_{k \quad \mathbb{Z}} h_\phi(k) \quad \overline{2}\ \phi(2x - k) \tag{8.21}$$

where $\quad \overline{2}\ \phi(2x-k)$ are the basis functions in V_1. The coefficients $h_\phi(k)$ are defined as follows:

$$\phi(x), \quad \overline{2}\ \phi(2x - k) \tag{8.22}$$

However, note that as $\phi(x)$ is nonzero only between 0 and 1, in the case of Haar wavelets the number of nonzero coefficients $h_\phi(k)$ is limited to 2. Similar to the scaling function in V_0, the wavelet function can be defined as:

$$\psi(x) = \sum_{k \quad \mathbb{Z}} h_\psi(k) \quad \overline{2}\phi(2x - k) \tag{8.23}$$

The coefficients $h_\psi(k)$ and $h_\phi(k)$ are then related as follows:

$$h_\psi(k) = (-1)^k\ h_\phi(1 - k) \tag{8.24}$$

For Haar wavelets, it can be shown that:

$$\phi(x) = \phi(2x) + \phi(2x-1) \tag{8.25}$$

$$\psi(x) = \phi(2x) - \phi(2x-1) \tag{8.26}$$

As a result, the nonzero filter coefficients for $h_\phi(k)$ are $h_\phi(0) = 1/\sqrt{2}$ and $h_\phi(1) = 1/\sqrt{2}$. The nonzero coefficients for $h_\psi(k)$ are $h_\psi(0) = 1/\sqrt{2}$ and $h_\psi(1) = -1/\sqrt{2}$. It is therefore possible to compute coefficients for each of the (integer) elements of the input function $f(x)$ as follows:

$$c_\phi(x) = \frac{1}{\sqrt{2}} f(x) + \frac{1}{\sqrt{2}} f(x-1) \tag{8.27}$$

$$c_\psi(x) = \frac{1}{\sqrt{2}} f(x) - \frac{1}{\sqrt{2}} f(x-1) \tag{8.28}$$

where the coefficients $c_\phi(x)$ encode the average of $f(x)$ and $f(x-1)$ and the coefficients $c_\psi(x)$ represent deviations from this average for elements $f(x)$ and $f(x-1)$. To encode a signal of length n, the above coefficients would be computed for each pair of neighboring elements so that the total number of coefficients is $n/2$ for $c_\phi(x)$ as well as $n/2$ for $c_\psi(x)$. This could be written in matrix form as follows:

$$\left[\begin{array}{c} c^1_\phi(0) \\ c^1_\phi(1) \\ c^1_\phi(2) \\ \vdots \\ c^1_\phi(n/2) \\ \hline c^1_\psi(0) \\ c^1_\psi(1) \\ c^1_\psi(2) \\ \vdots \\ c^1_\psi(n/2) \end{array}\right] = \left[\begin{array}{cccccccccc} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \cdots & \cdots & 0 \\ & & & & & \ddots & \ddots & & & \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \hline \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & \cdots & \cdots & 0 \\ & & & & & \ddots & \ddots & & & \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{array}\right] \left[\begin{array}{c} f(0) \\ f(1) \\ f(2) \\ \vdots \\ f(n/2) \\ \hline f(n/2+1) \\ f(n/2+2) \\ f(n/2+3) \\ \vdots \\ f(n) \end{array}\right] \tag{8.29}$$

This matrix is set up such that the resulting vector contains all the averages in its first $n/2$ elements and all the differences in the remaining elements. An interesting observation is that on average we may expect that the elements $c^1_\phi(x)$ have values in the same range as the values in $f(x)$, whereas the elements $c^1_\psi(x)$ are on average much smaller. This has many implications for specific applications, which will be discussed later.

After this procedure, we have a signal that consists on average for the first half of large values and for the second half of small values. It is now possible to further encode the signal by repeating the procedure on $c^1_\phi(x)$:

$$\begin{bmatrix} c^2_\phi(0) \\ c^2_\phi(1) \\ c^2_\phi(2) \\ \vdots \\ c^2_\phi(n/4) \\ \hline c^2_\psi(0) \\ c^2_\psi(1) \\ c^2_\psi(2) \\ \vdots \\ c^2_\psi(n/4) \end{bmatrix} = \left[\begin{array}{cccccccccc} \frac{1}{2} & \frac{1}{2} & 0 & \multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & \multicolumn{4}{c}{\dots\dots\dots\dots\dots\dots} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & \multicolumn{2}{c}{\dots\dots\dots} & 0 \\ & & & & & \ddots & \ddots & & & \\ 0 & \multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots} & 0 & \frac{1}{2} & \frac{1}{2} \\ \hline \frac{-1}{2} & \frac{1}{2} & 0 & \multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots} & 0 \\ 0 & 0 & \frac{-1}{2} & \frac{1}{2} & 0 & \multicolumn{4}{c}{\dots\dots\dots\dots\dots\dots} & 0 \\ 0 & 0 & 0 & 0 & \frac{-1}{2} & \frac{1}{2} & 0 & \multicolumn{2}{c}{\dots\dots\dots} & 0 \\ & & & & & \ddots & \ddots & & & \\ 0 & \multicolumn{6}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots} & 0 & \frac{-1}{2} & \frac{1}{2} \end{array}\right] \begin{bmatrix} c^1_\phi(0) \\ c^1_\phi(1) \\ c^1_\phi(2) \\ \vdots \\ c^1_\phi(n/2) \end{bmatrix} \tag{8.30}$$

As a result, by repeating this process until we have only one average of the entire signal followed by $n-1$ differences, we can create a vector that represents the final wavelet encoding:

$$\begin{bmatrix} c^m_\phi(0) & c^m_\psi(0) & c^{m-1}_\psi(0) & c^{m-1}_\psi(1) & c^{m-2}_\psi(0) & c^{m-2}_\psi(1) & c^{m-2}_\psi(2) & c^{m-2}_\psi(3) & \dots \end{bmatrix} \tag{8.31}$$

where $m = \log_2(n)$. Note that it is possible to exactly reconstruct the original signal $f(x)$ from this encoding. The first step of the inverse is given by:

$$\begin{bmatrix} c^{m-1}_\phi(0) \\ c^{m-1}_\phi(1) \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{-1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} c^m_\phi(0) \\ c^m_\psi(0) \end{bmatrix} \tag{8.32}$$

Here, the two averages are reconstructed from the top-level average and difference. Repeating this process gives us four averages:

$$\begin{bmatrix} c^{m-2}_\phi(0) \\ c^{m-2}_\phi(1) \\ c^{m-2}_\phi(2) \\ c^{m-2}_\phi(3) \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & \frac{-1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{-1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} c^{m-1}_\phi(0) \\ c^{m-1}_\phi(1) \\ c^{m-1}_\psi(0) \\ c^{m-1}_\psi(1) \end{bmatrix} \tag{8.33}$$

This process can then be repeated a total of $m = \log_2(n)$ to reconstruct the original function $f(x)$. The final step in this reconstruction is:

$$\begin{bmatrix} f(0) \\ f(1) \\ \vdots \\ f(n) \end{bmatrix} = \mathbf{R} \begin{bmatrix} c^1_\phi(0) \\ c^1_\phi(1) \\ \vdots \\ c^1_\phi(n/2) \\ c^1_\psi(0) \\ c^1_\psi(1) \\ \vdots \\ c^1_\psi(n/2) \end{bmatrix} \tag{8.34}$$

where $\mathbf{R}$ is given by:

$$\mathbf{R} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \cdots & 0 & \frac{-1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{2}} & 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 & 0 & \frac{-1}{\sqrt{2}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 & 0 & 0 & \frac{-1}{\sqrt{2}} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ & & & \ddots & & & & & & & \ddots & \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 & \frac{-1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & 0 & \cdots & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \tag{8.35}$$

The wavelet-encoded signal consists of differences for all but the first element. As argued above, this means that most values in the encoded signal will be smaller than the signal's original values. For instance, the Haar wavelet-encoded signal of a sequence of eight numbers is given in Table 8.1. Although the average of the signal itself is 6.5 (first column on the $m = 3$ row), the average of the magnitude of the remaining coefficients is just 1.5. Another example is shown in Figure 8.3, where the signal in the left panel is encoded using Haar wavelets. The wavelet coefficients are then sorted by magnitude and plotted in the right panel. Note that most coefficients are very small.

As a result, it should be possible to encode each element with fewer bits than the original signal, leading to a simple form of data compression. If the signal is encoded as integers, many of these small differences would be the same value. The signal could therefore be encoded very efficiently with Huffman coding.

Moreover, it would be possible to set all elements that are smaller than a given threshold to zero, effectively compressing the data in a lossy manner. An example for Haar wavelets is given in Figure 8.4. Here, the signal was compressed by a ratio of around 4:1 as the signal $f(x)$ contained 4,096 samples, while the reconstruction used only 1,000 wavelets.

$f(x)$	**8**	**10**	**3**	**7**	**6**	**2**	**7**	**9**
$m = 1$	9	5	4	8	1	2	–2	1
$m = 2$	7	6	–2	2	1	2	–2	1
$m = 3$	6.5	–0.5	–2	2	1	2	–2	1

Table 8.1. The input signal $f(x)$ is successively decomposed into wavelet coefficients, leading to an average of 6.5 followed by a set of small coefficients. For clarity we have removed the factor of $1/\sqrt{2}$ so that $h_\phi = [1\ 1]$ and $h_\psi = [1\ {-1}]$.

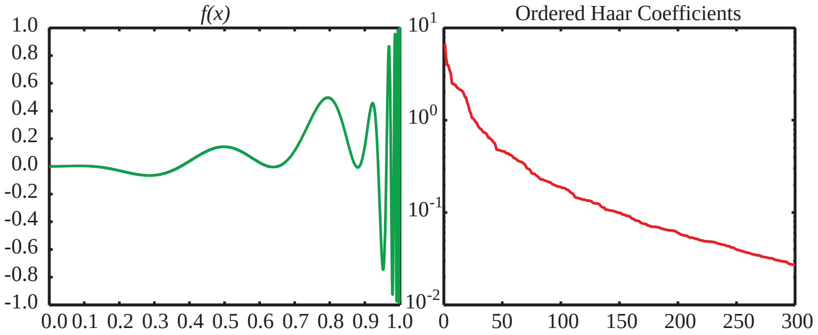


Figure 8.3. The signal on the left is encoded using Haar wavelets, after which the resulting coefficients are sorted by magnitude and plotted on the right. Note how the magnitude of the coefficients is small relative to the magnitude of the signal.

8.4 Other Bases

While the Haar wavelet basis is useful in many cases, in part due to its simplicity, and in part because it can represent sharp edges in the signal well, many coefficients are required to fully reconstruct a smooth signal, as seen in Figure 8.4. Thus, in addition to local support, smoothness is often a second requirement for effective wavelet decomposition. The machinery required for all wavelets is similar: the generating function $\phi(x)$ from Equation (8.13) will represent (weighted) averages, while the wavelet function $\psi(x)$ from Equation (8.11) will encode differences. To efficiently encode smoother functions, the generating and wavelet function pairs can be replaced with new basis functions. This has given rise to many alternatives to Haar wavelets, including Daubechies [152], Coiflet, and Symlet wavelet, as well as many others.

While the Haar wavelet can be implemented as a convolution with a filter that has two nonzero elements (or *taps*) per coefficient (see Equation (8.30)), all other wavelets require more taps. Daubechies, for instance, has developed a family of wavelets which have an increasing number of taps [152, 19], as shown in Figure 8.5. Daubechies weights at each tap cannot be generated algorithmically, but are typically produced numerically by the *cascade algorithm* [96]. Daubechies

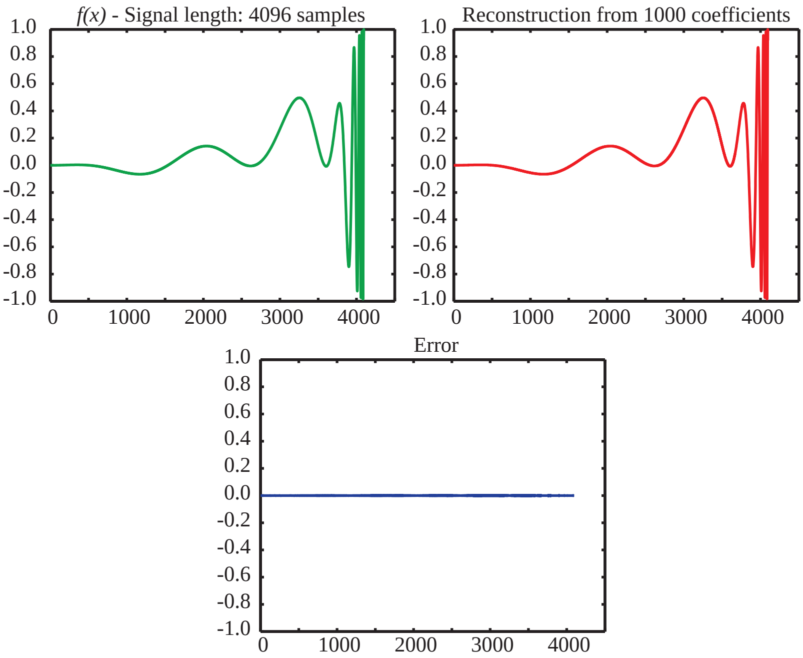


Figure 8.4. The signal on the top left is encoded using Haar wavelets, after which the signal is reconstructed using 1,000 coefficients. The original function contained 4,096 samples, so that lossy reconstruction is achieved with low error and a compression of around 4:1.

wavelets have a number of *vanishing moments* equal to half the number of taps. A Daubechies wavelet with four taps (D4) therefore has two vanishing moments.

The importance of vanishing moments lies in the fact that its value determines the smoothness of the functions that can be encoded. With n vanishing moments, polynomials of degree $n-1$ can be encoded. This means that a D4 wavelet can encode constant and linear polynomials. An example demonstrating the implications for reconstruction is given in Figure 8.6. As the input function is smooth, Daubechies D20 with ten vanishing moments can represent polynomials with degree 9 and is therefore able to reconstruct the function with only 50 wavelet coefficients. The Haar function can only represent piecewise constant functions and so does not perform well with only 50 coefficients. The Daubechies D20 wavelets have allowed a compression of around 82:1, which cannot be matched by the Haar wavelets.

The symlet wavelets improve symmetry properties relative to the Daubechies family of wavelets. The Coiflet wavelets [55, 153] have vanishing moments in both wavelet function as well as the scaling/generating function [353], while also being near-symmetric.

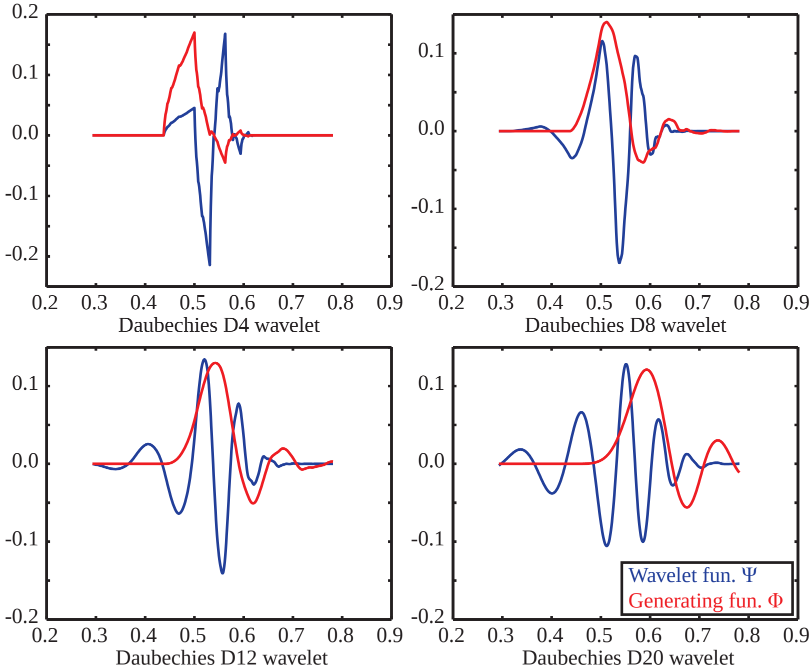


Figure 8.5. Daubechies wavelets are orthonormal and can be constructed with a different number of taps. The more taps, the smoother the wavelet. Shown here are the D4, D8, D12, and D20 wavelets.

8.5 2D Wavelets

So far we have considered functions of one variable. Images, however, are two-dimensional, and we therefore have to extend wavelet representations to two dimensions. The Haar wavelet is separable, which means that an image can be decomposed in one dimension first, followed by decomposition in the second dimension. In essence, the scaling function becomes:

$$\phi(x,y) = \phi(x)\,\phi(y) \tag{8.36}$$

The wavelet function is split into three wavelet functions, one each for the horizontal, vertical, and diagonal directions:

$$\psi^H(x,y) = \psi(x)\,\phi(y) \tag{8.37}$$

$$\psi^V(x,y) = \phi(x)\,\psi(y) \tag{8.38}$$

$$\psi^D(x,y) = \psi(x)\,\psi(y) \tag{8.39}$$

The only difference between the three functions $\psi^H(x,y)$, $\psi^V(x,y)$, and $\psi^D(x,y)$ lies, therefore, in which of the scaling and wavelet functions is applied and in

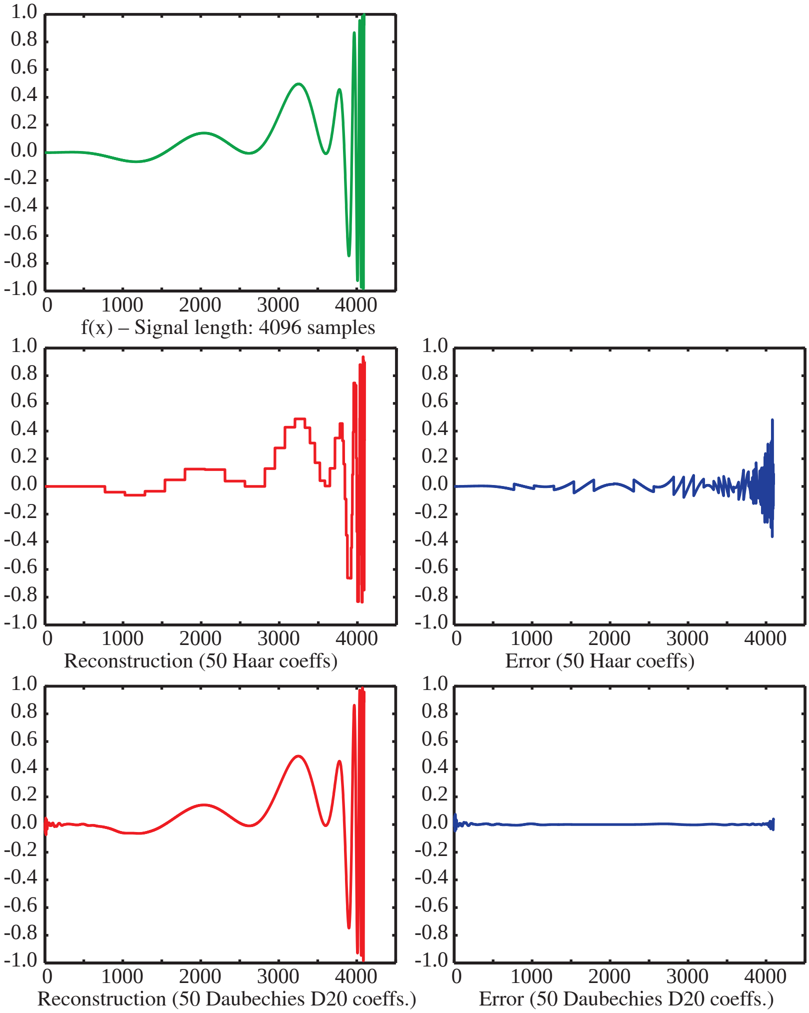
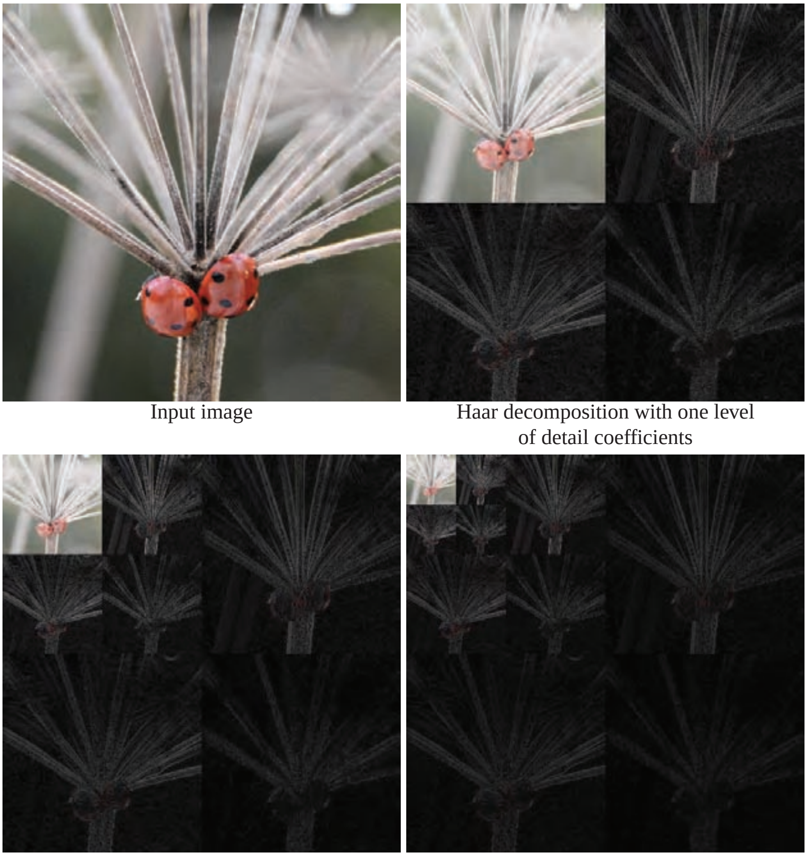


Figure 8.6. Reconstruction comparison between Haar and Daubechies D20 wavelets. The signal $f(x)$ has 4,096 samples, which are here reconstructed from 50 wavelet coefficients.

which dimension. Otherwise, the computations proceed as before. A single level decomposition can be created by applying the 1D Haar decomposition to all the rows of the image. As averages will be moved toward the left half of the image and differences to the right half, each row will consist of the elements of $\phi(x)$ followed by the elements of $\psi(x)$.

If the procedure is repeated, but now on the columns of the image, the resulting output will consist of four quadrants, as shown in the top-right panel of



Input image

Haar decomposition with one level of detail coefficients

Haar decomposition with two levels of detail coefficients

Haar decomposition with three levels of detail coefficients

Figure 8.7. One, two, and three levels of Haar decomposition. Note that we have passed the resulting images through a gamma curve (exponent 0.4) to help visualize the coefficients.

Figure 8.7. The top-left quadrant will contain the averages $\phi(x,y)$ which is a downsampled version of the input image. The top-right and bottom-left quadrants contain horizontal and vertical differences. Finally, the bottom-right quadrant contains the vertical differences of previously computed horizontal differences. These therefore represent diagonal differences. This decomposition is said to be separable, as horizontal and vertical calculations are carried out consecutively.

As before, these procedures can be applied recursively on the averages, leading to a hierarchical wavelet decomposition. Two- and three-level Haar decompositions are shown at the bottom of Figure 8.7. Of course, this procedure can be repeated until the top-left pixel contains the image average and all other pixels

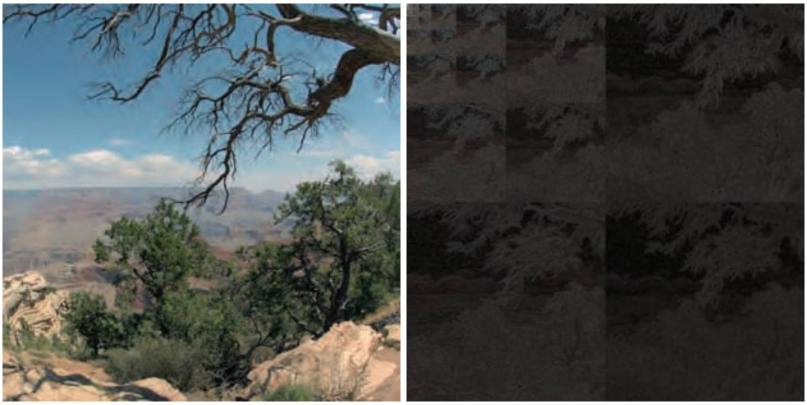


Figure 8.8. A full Haar decomposition. Note that we have passed the resulting images through a gamma curve (exponent 0.2) to help visualize the coefficients. (Grand Canyon, Arizona, 2012)

store differences. An example of a full decomposition is shown in Figure 8.8. Here, most coefficients in the image have become very small so that the encoded image appears mostly black.

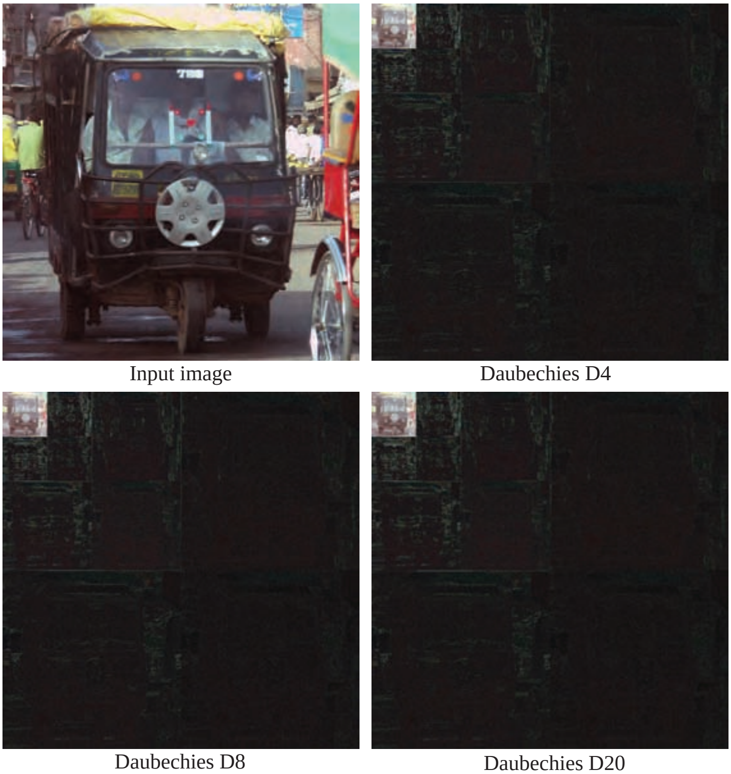
Although Haar wavelets are used extensively in image processing, they are by no means the only wavelets amenable to image encoding. Daubechies wavelets, for instance, can be used to encode images as shown in Figure 8.9, although note that as these wavelets use more than two taps, boundary effects can be more of a problem.

8.6 Contourlets, Curvelets, and Ridgelets

Further, wavelets have been extended to contourlets, which are more adaptable to the images being encoded [172], as well as curvelets [109] and ridgelets [107, 108], which are especially adapted to edges. Consider a curved edge in an image. In a conventional wavelet decomposition, the image would be partitioned as illustrated in the left part of Figure 8.10. By introducing more orientations, and by allowing the basis functions to follow image features more closely, a sparser decomposition can be achieved, as shown in the right part of Figure 8.10.

8.7 Coefficient Histograms

So far, we have shown that wavelet coefficients tend to lead to smaller numbers than the values found in the original signal/image. It has been shown that wavelet coefficients follow a highly kurtotic distribution [215, 489]. An example for a


Input image

Daubechies D4

Daubechies D8

Daubechies D20

Figure 8.9. Two levels of Daubechies decomposition for D4, D8, D12, and D20 wavelets. Note that we have passed the resulting images through a gamma curve (exponent 0.4) to help visualize the coefficients. (New Delhi, India, 2008)

single image is shown in Figure 8.11, where we have subjected a high dynamic range image to the Haar wavelet transform. The kurtosis observed in each channel is larger than 2,000, which is indeed pointing at very long tails. We have plotted the vertical axis on a log scale for the purpose of visualization.

The highly non-Gaussian histogram distribution of wavelet coefficients can be modeled with the *generalized Laplacian function* [489, 687, 85, 685] (also known as the *exponential power distribution* or the *generalized error distribution*), as given by:

$$P(x) = \frac{\beta}{2\alpha\Gamma\left(\frac{1}{\beta}\right)} e^{-\;(x-\mu)/\alpha\;^{\beta}} \tag{8.40}$$

where Γ is the gamma function, as defined by:

$$\Gamma(x) = \int_0^\infty e^{-t}\, t^{z-1}\, dt \tag{8.41}$$

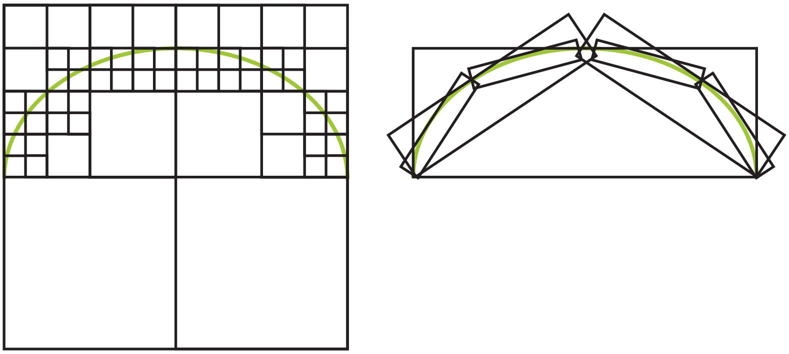


Figure 8.10. An illustration of a wavelet decomposition (left) and a contourlet decomposition (right).

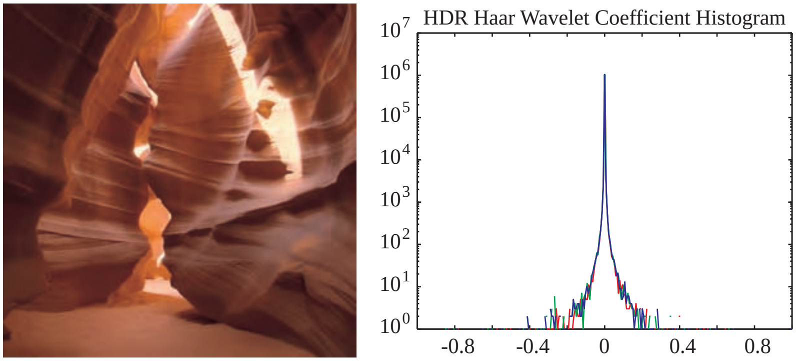


Figure 8.11. The Haar wavelet coefficients are collected into a histogram for each of the red, green, and blue channels of the high dynamic range image depicted on the left. The image is tonemapped for display [613], but coefficients are computed on the high dynamic range original. Note the highly kurtotic shape of the distribution. (Antelope Canyon, Arizona, 2012)

The parameters α, β, and μ can be used to fit the model to a specific image or image ensemble. For sparse distributions as seen with wavelet coefficients, the shape parameter β is less than 1. The parameter α scales the distribution, while μ is the mean. Note that to model the distribution of wavelet coefficients, the mean is zero. The variance of the distribution, as modeled with Equation (8.40), is given by:

$$\sigma^2 = \frac{\alpha^2 \, \Gamma(3/\beta)}{\Gamma(1/\beta)} \tag{8.42}$$

As the distribution is symmetric, the skewness is necessarily 0. The kurtosis of this distribution is:

$$\frac{\Gamma(5/\beta)\,\Gamma(1/\beta)}{\Gamma(3/\beta)^2} - 3 \tag{8.43}$$

For instance, if the function fit yields a value for β of 0.5, then the kurtosis of the distribution was $\Gamma(10)\,\Gamma(2)/\Gamma(6) - 3 > 22$. This would be a highly peaked function. For the distribution shown in Figure 8.11, the value of β is around 0.15, given that the pixel values in this image have a kurtosis of around $20{,}000$. Also note that the kurtosis of the original pixel values of this image is around 135.

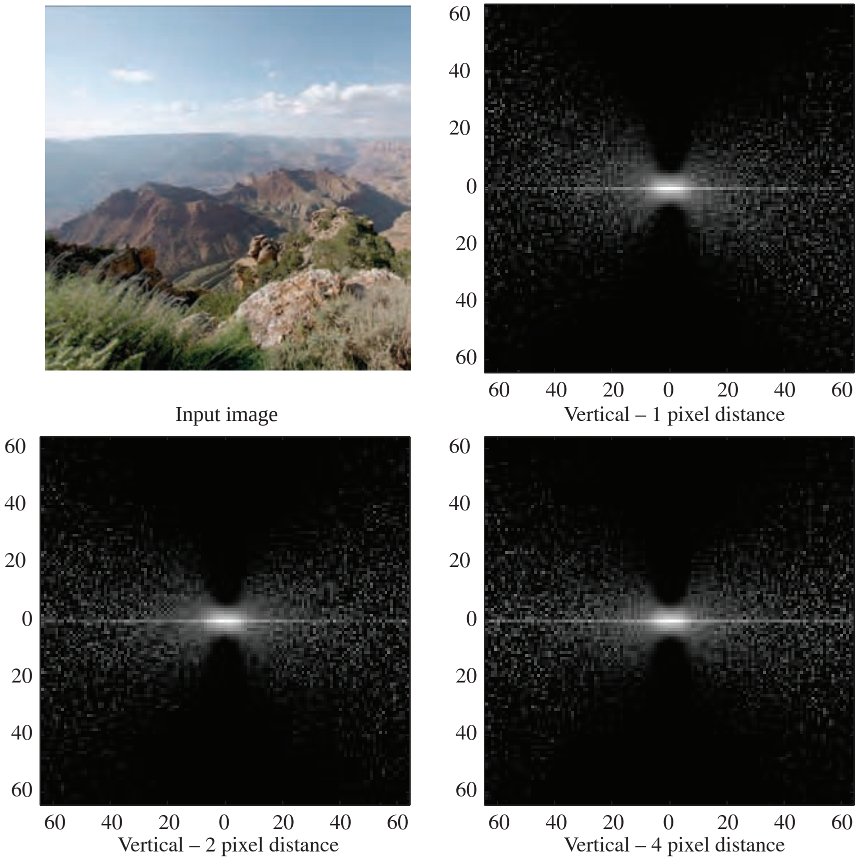
Wavelet encodings tend to be sparse. However, there will be cases where the encoding fails to be sparse. It has been shown that in such cases Markov random fields (discussed in Chapter 9) can be used to represent the local non-sparseness [803].

High values for the kurtosis of the distribution of wavelet coefficients are common in wavelet decompositions. As expected, they have also been found in contourlets [575] as well as curvelets and ridgelets [177]. In fact, it appears that high kurtosis appears whenever localized zero-mean linear kernels are used as bases [818, 704]. Sparseness of the representation, indicated by high kurtosis in wavelet coefficients, accounts for much of the success of wavelets in various applications. As an example, by explicitly estimating the parameters α and β, it is possible to construct an algorithm to denoise images [687] or to efficiently encode images [85]. Quality assessment algorithms have also been designed using the non-Gaussian distribution of wavelet coefficients [672, 671]. Wavelets have also been used in the authentication of art [278, 481], steganalysis [478, 480], and in the differentiation between photographs and photorealistic renderings [479].

8.8 Scale Invariance

One could analyze the distribution of wavelet coefficients at each scale of the wavelet tree separately. For natural images, the result would be that each distribution would have roughly the same, highly kurtotic shape [518]. This is the same as saying that the distribution of wavelet coefficients is invariant under scaling. If the histogram of a distribution remains invariant under scaling, then so are its associated moments [518].

It can be argued that image formation is driven by objects (which are possibly textured [824]). The analysis of amplitude spectra of natural image ensembles (Section 6.4.2) shows scale invariance as a result. Here, we have a different way of expressing the same scale invariance, albeit that we could go one step further and assess scale invariance locally in small image regions. Scale invariance returns in the analysis of the dead leaves model in Section 11.1, where scenes are assumed to be composed of objects that can (partially) occlude each other. For instance, the dead leaves model gives clues as to the size distribution of objects required to match the observed scale invariant statistics [12].



Input image

Vertical – 1 pixel distance

Vertical – 2 pixel distance

Vertical – 4 pixel distance

Figure 8.12. Conditional histograms for vertical wavelet coefficients. The distance between the pixels was 1, 2, and 4 pixels. (Grand Canyon, Arizona, 2012)

8.9 Correlations between Coefficients

A second observation that can be made about wavelet coefficients is that corresponding coefficients in either scale, space, or orientation tend to be highly correlated [468]. For instance, when taking neighboring coefficients at a particular scale and orientation, their values tend to be similar. This can be seen by computing a joint histogram of coefficients and their neighbors at some distance. For instance, Figure 8.12 shows the conditional histogram of the vertical coefficients for pixels separated by 1, 2, and 4 pixels in the vertical direction. The bow-tie shape of the conditional histogram shows that especially small values of

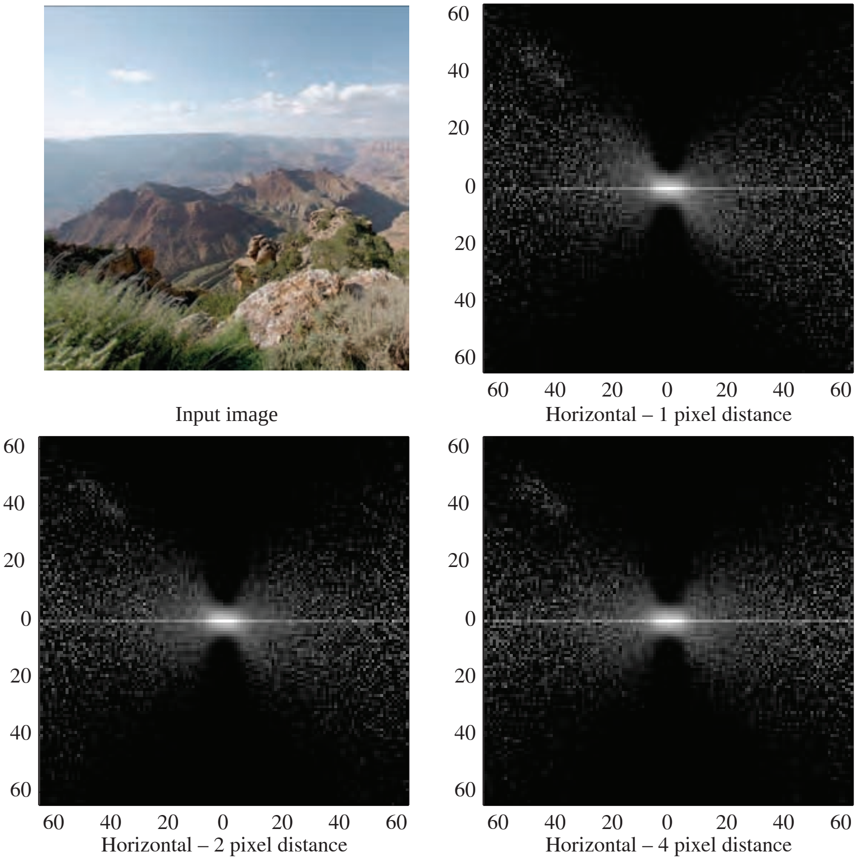


Figure 8.13. Conditional histograms for horizontal wavelet coefficients. The distance between the pixels was 1, 2, and 4 pixels. (Grand Canyon, Arizona, 2012)

one coefficient are good predictors for the value of coefficients some distance away. This holds for horizontal and diagonal coefficients as well, as shown in Figures 8.13 and 8.14, and also applies to contourlets [575].

This bow-tie shape suggests that there are still dependencies between the coefficients [767]. The variance scales with the absolute value on the abscissa. These dependencies cannot be further reduced with any linear transform. It is thought that occlusion of objects is one of the main processes giving rise to image formation [287] (see Chapter 11), which is nonlinear. This tends to give rise to relatively flat areas in images with sharp discontinuities. Such edge-like structures tend to have substantial power across scales, orientations, and local position. This is reflected in the conditional coefficient histograms.

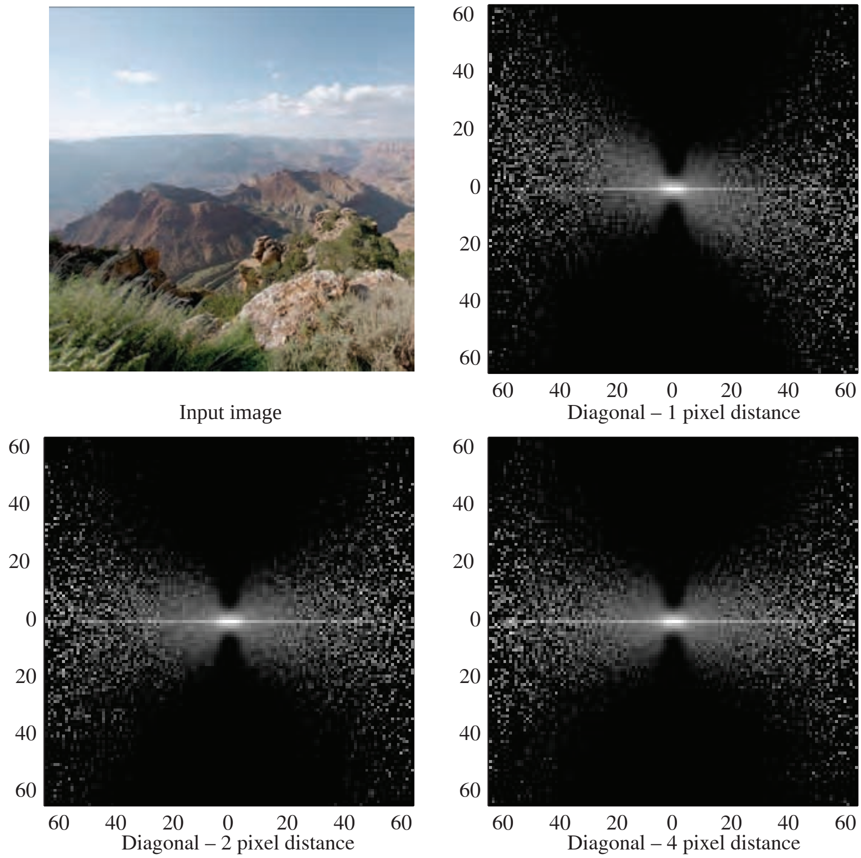


Figure 8.14. Conditional histograms for diagonal wavelet coefficients. The distance between the pixels was 1, 2 and, 4 pixels. (Grand Canyon, Arizona, 2012)

It is interesting to note that these dependencies can be reduced by introducing a divisive normalization akin to the Naka-Rushton equation, which models photoreceptors (see Section 2.3.2) [731, 659, 767]:

$$V_j = \frac{I_j^2}{\sum_k w_{jk} I_k^2 + \sigma_j^2} \tag{8.44}$$

where each coefficient I_j is replaced by V_j. Neighboring coefficients in scale, space, and orientation are indicated with I_k. The parameters σ_j and w_{jk} can be determined from the analysis of natural images. Note that σ_j represents the variance that cannot be accounted for by the neighboring coefficients.

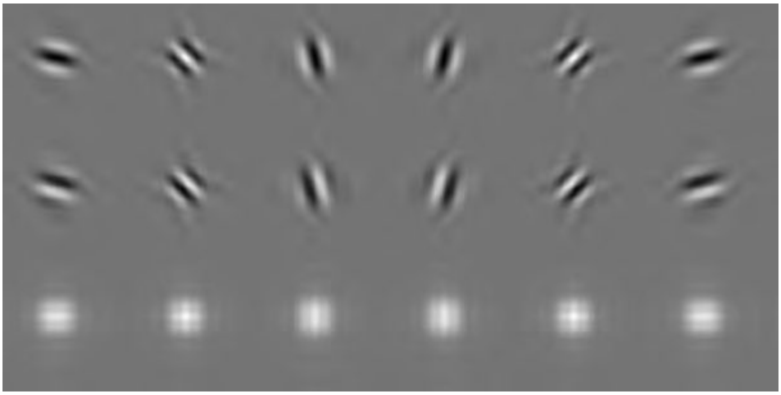


Figure 8.15. Complex wavelets. The top row shows the real part, the second row shows the imaginary part, and the bottom row shows the magnitude for each of six orientations.

Finally, we note that correlations between coefficients in a wavelet decompositions have been used in the field of steganography [480], image restoration [684], and image compression [86].

8.10 Complex Wavelets

The wavelets described so far produce one real-valued coefficient per pixel. In analogy to Fourier transforms, it is possible to design wavelet pairs that produce complex-valued coefficients, which can be reinterpreted as representing amplitude and phase. However, contrary to a Fourier decomposition, the amplitude and phase representations of complex wavelets are localized in space.

Real-valued wavelets as discussed so far have advantages in that they can have nonredundant orthonormal bases, allow for multiresolution decomposition and have fast linear time algorithms to decompose and reconstruct signals and images. However, they also have some disadvantages in image analysis and reconstruction tasks [403, 664]:

Sensitivity to shifts. If the input is shifted by small amounts, this can lead to major changes in the distribution and values of the coefficients. In many image processing tasks, this is an undesirable feature.

Poor directional selectivity. Separable and real wavelets tend to have poor selectivity for orientations other than horizontal and vertical.

No phase information. The discrete wavelet transform does not encode phase information and also tends to have poor frequency resolution.

Oscillations. Ideally wavelet coefficients are only large near edges. However, this is not necessarily the case with real-valued wavelets, which may make certain image processing tasks such as finding edges difficult [120].

These disadvantages can be overcome with complex wavelets. These in essence introduce some redundancy, as every real-valued sample in the input results in a complex-valued coefficient. Complex wavelets solve many of these problems akin to Fourier transforms. In particular, the amplitude of a Fourier transform is fully shift invariant. Further, in two dimensions complex wavelets are highly directionally sensitive.

As argued in Section 6.2, the Fourier transform decomposes a signal into sine and cosine basis functions:

$$e^{i\omega x} = \cos(\omega x) + i \sin(\omega x) \tag{8.45}$$

As the cosine and sine components are 90 degrees out of phase with each other, they form a Hilbert transform pair. The background to complex wavelet transforms relates to the Hilbert transform in a similar fashion. A real signal $f(x)$ can be extended into the complex domain using [248]:

$$h(x) = f(x) + i\, g(x) \tag{8.46}$$

where $g(x)$ is the Hilbert transform of $f(x)$, denoted as:

$$g(x) = H \;\; f(x) \tag{8.47}$$

As mentioned above, the Hilbert transform is defined as a 90 degree rotation in the complex plane of the input signal $f(x)$. This makes $g(x)$ orthogonal to $f(x)$. Being out of phase by a quarter period of $f(x)$ and $g(x)$ is also known as being in *quadrature*. The Hilbert transform is given by [296]:

$$\begin{aligned} g(x) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{x - \tau}\, f(x)\, d\tau && (8.48) \\ &= f(x) \otimes \frac{1}{\pi x} && (8.49) \end{aligned}$$

The *magnitude* $h(x)$ and *phase* $\angle h(x)$ of the complex signal $h(x)$ are then given by:

$$\begin{aligned} h(x) \;\; &= \sqrt{f(x)^2 + g(x)^2} && (8.50) \\ \angle h(x) &= \tan^{-1}(g(x)/f(x)) && (8.51) \end{aligned}$$

Note that the magnitude is also known as the *envelope* of the combined response of $f(x)$ and $g(x)$.

As any signal can be extended into the complex domain in this manner, it is certainly also possible to create a complex scaling function (where the subscripts c, r, and i denote complex, real, and imaginary components) [664]:

$$\psi_c(x) = \psi_r(x) + i\,\psi_i(x) \tag{8.52}$$

Once again, the real and imaginary parts $\psi_i(x)$ and $\psi_r(x)$ should form a Hilbert transform pair. A 2D example is shown in Figure 8.15, where the top row shows the real part of six oriented wavelets, and the second row shows the imaginary part of the same six oriented wavelets. Note that the real part of this wavelet is an even-symmetric filter, whereas the imaginary part is odd-symmetric. The third row in this figure plots the envelope or magnitude of these six wavelets, which has an approximately Gaussian shape. It is then possible to compute wavelet coefficients for a specific signal $f(x)$ using complex wavelets [664]:

$$d(j,n) = 2^{j/2} \int_{-\infty}^{\infty} f(x)\ \psi_c(2^j t - n)\ dt \tag{8.53}$$

The magnitude and (phase) angle can be computed from these coefficients as before:

$$d_c(j,n) \ = \sqrt{d_r(x)^2 + d_i(x)^2} \tag{8.54}$$

$$\angle d_c(x) = \tan^{-1}(d_i(x)/d_r(x)) \tag{8.55}$$

On average, the magnitude will be small for most coefficients but may become large for pixels that are on or near a contrast feature in a signal or image. The value of the phase then indicates the position of the feature within the support of the wavelet.

As with real-valued wavelet decompositions, complex wavelet decompositions can be either redundant or nonredundant. In the latter case, it is possible to construct complex wavelets that have orthonormal or biorthogonal bases [23, 46, 213, 463, 701, 768]. It can be argued, however, that a redundant representation would be better able to overcome the shortcomings of real-valued wavelets [664], as outlined on page 197. The resulting wavelet decomposition would be 2× redundant in one dimension [152], as the real-valued samples in the original signal are replaced by complex-valued wavelet coefficients. Such redundant wavelet decompositions can be efficiently constructed by means of the dual-tree complex wavelet transform [403, 404, 405, 406, 663, 662, 664]. They can also be naturally extended to two dimensions, allowing images to be transformed into the complex wavelet domain. This, however, incurs a 4× redundancy.

Complex wavelets have advantages in practical applications as they are nearly shift and rotation invariant. In particular, they have been shown to be more effective than real-valued wavelets in image and video denoising [120, 628, 811, 665, 675]. Other applications in which complex wavelets have proven useful include image segmentation [666], feature extraction [415, 469], texture analysis, and synthesis [620, 317, 329], image classification [629], deconvolution [621, 372], watermarking [189, 473], image sharpening [676], motion estimation [497], and image and video coding [606, 695, 775]. In addition, complex wavelets are studied in the context of natural image statistics, having revealed one regularity that is not observed in the analysis of real-valued wavelets. This is discussed in the following section.

8.11 Correlations between Scales

Given that edges tend to persist across scales, it is to be expected that if we were to compute correlations between scales, the structure in scenes should cause such correlations to be high. In other words, we might expect there to be some redundancy between neighboring scales in a wavelet decomposition.

However, as the filters applied in subsequent scales of a wavelet transform are normally orthogonal, measuring the correlation of wavelet coefficients in subsequent scales proves problematic. This can be seen as follows. Assume that at scales i and $i+1$ we have wavelet filters $f_i(x)$ and $f_{i+1}(x)$, which are orthogonal by definition, so that:

$$\int f_i(x)\; f_{i+1}(x - x_0)\, dx = 0 \qquad\qquad x_0 \qquad\qquad (8.56)$$

This means that the convolution of two such filters with a given signal will also be orthogonal. As a result, the correlation between wavelet coefficients will be zero if the two filters are truly orthogonal [216]. An example of the correlations measured with a real-valued wavelet decomposition is shown in Figure 8.16.

Thus, a different measure is required to assess the correlations that may occur between scales. In particular, complex wavelets offer a unique opportunity to understand how edges produce correlations (or redundancy) between scales. Although the real and complex wavelets by themselves show the same problems as outlined above, i.e., they are to a large extent orthogonal between scales, the complex wavelet decomposition can be transformed from a Euclidian coordinate space to a polar space, whereby real and complex values transform into magnitudes and angles.

It turns out that the amplitudes of complex wavelet coefficients tend to be correlated across scales [216]. In fact, if the phases of image features are aligned at similar scales in two frequency bands, the correlation between the magnitude of complex wavelet coefficients will be high. An example of the correlations found across scale in a single image are shown in Figure 8.17, allowing a comparison with the real-valued wavelet composition of Figure 8.16.

We can show that these correlations are largely due to information encoded in the phase spectrum (whereby we mean the phase spectrum as available in Fourier

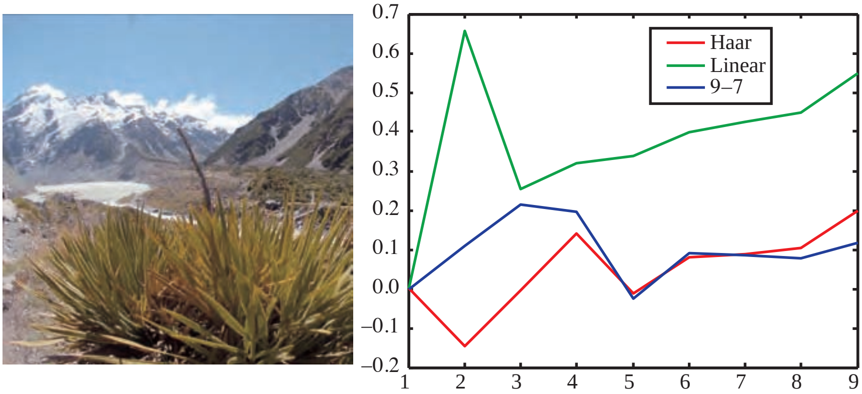


Figure 8.16. The magnitude of real wavelets shows comparatively weak correlations across scale. For each of three different types of real-valued wavelets, the correlation between two neighboring scales are shown. The smaller scales represent detail coefficients, whereas larger scales represent coarser features. (Mt. Cook, New Zealand, 2012)

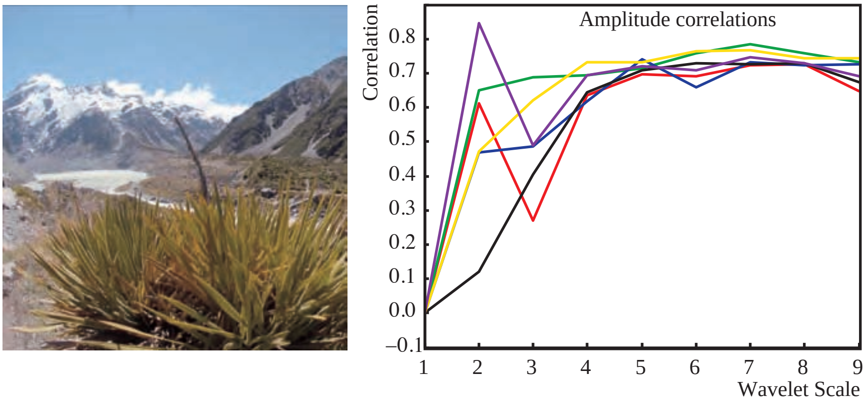


Figure 8.17. The magnitude of complex wavelets shows strong correlations across scale. For each of six orientations, the correlation between two neighboring scales are shown. The smaller scales represent detail coefficients, whereas larger scales represent coarser features. (Mt. Cook, New Zealand, 2012)

space). To this end, we have computed the Fourier transform of an image (the one shown in Figures 8.16 and 8.17), permuted the phase spectrum, and converted back to the spatial domain. We then computed the complex wavelet transform as well as the correlations across scales of its amplitude spectrum. The results are shown in Figure 8.18. Note that by removing structure from the Fourier domain phase spectrum, correlations across scale vanish for the most part. This means

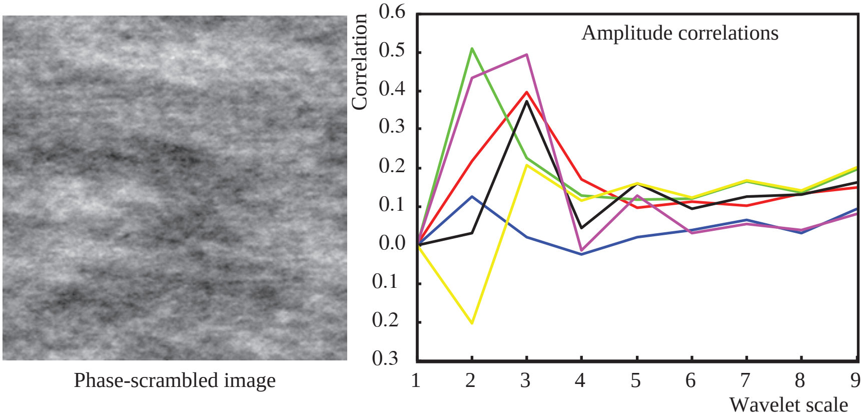


Figure 8.18. If the phase spectrum of an image is permuted (left), then the amplitude of a complex wavelet decomposition does not show strong correlations across scale (right).

that local structure that persists across scale in a complex wavelet decomposition is due to an image's phase, rather than its amplitude [216].

8.12 Application: Image Denoising

Wavelets can be used for image denoising. This can, for instance, be achieved by applying soft-thresholding on the coefficients $c_{j,k} = \; f, \psi_{j,k} \;$, leading to new coefficients $c_{j,k}$ [782, 758]:

$$c_{j,k} = \begin{cases} c_{j,k} - t_j & c_j, k \geq t_j \\ 0 & c_{j,k} \leq t_j \\ c_{j,k} + t_j & c_j, k \leq -t_j \end{cases} \tag{8.57}$$

This method is essentially the wavelet shrinkage denoising method [178, 176, 171, 758]. It can be extended by replacing coefficients according to some function, i.e., $c_{j,k} = g(c_{j,k})$. Assuming that the coefficients $c_{j,k}$ are polluted by additive noise:

$$c_{j,k} = c^s_{j,k} + n(i,k) \tag{8.58}$$

where $c^s_{j,k}$ is the noise-free coefficient and $n(i,k)$ is an independent noise component. It is then possible to replace the observed coefficients $c_{j,k}$ by an optimal linear estimate [53]:

$$c_{j,k} = c_{j,k} \, \frac{E(c^2_{j,k})^2}{E(c^s_{j,k} + n(j,k))^2} \tag{8.59}$$

8.13 Application: Progressive Reconstruction

Wavelets can be configured to allow progressive reconstruction. Here, the wavelet coefficients can be sorted according to magnitude. If such sorted wavelet coefficients are then transmitted, the receiver can start reconstructing a crude image with few coefficients and refine the image as more coefficients are received [19, 49, 115, 614]. For Haar, Coiflet, and Daubechies wavelets, an example is shown in Figure 8.19.

8.14 Application: Texture Synthesis

This book is predominantly about natural image statistics. We see textures as a specific form of images, namely those that have stationary statistics. In other words, each local region of a texture has the same statistics, whereas this is not necessarily true for natural images. As an example, a natural image depicting a forest with a sky above it would have very different gradient statistics in the sky than in the forest. This means that textures are in some sense more amenable to statistical modeling than arbitrary natural images.

Textures have been described in terms of n^{th}-order joint empirical densities of image pixels [380] and subsequently as statistical interactions in local neighborhoods by means of Markov random fields [310, 136, 254, 167] (see Chapter 9). Inspired by knowledge of early visual processing in human and mammalian vision, textures can also be analyzed and represented at multiple spatial scales with oriented linear kernels [757, 486]. These, in turn, have given rise to the use of wavelet representations for the purpose of texture synthesis [110, 579, 578, 323, 580, 823, 157].

Statistics of complex wavelet decompositions have also proven successful in the area of texture synthesis. Here, correlations between pairs of coefficients across space, scale, and orientation in an example texture form constraints, which are augmented with a selection of additional constraints, including the expected product of the magnitudes of coefficient pairs, marginal statistics of image pixels, and low-pass coefficients [581]. An efficient sampling algorithm then creates a new texture subject to these constraints. Thus, the new texture will have the same correlations as those measured in the wavelet decomposition of the example texture.

A texture can be represented as a two-dimensional homogeneous random field defined on integer positions.[1] Its statistical properties can be connected to human visual perception by making the basic assumption that a set of functions can be defined such that if we were to draw samples from two random fields X and Y

[1] Note that an image is not necessarily homogeneous, while a texture is assumed to be statistically homogeneous.

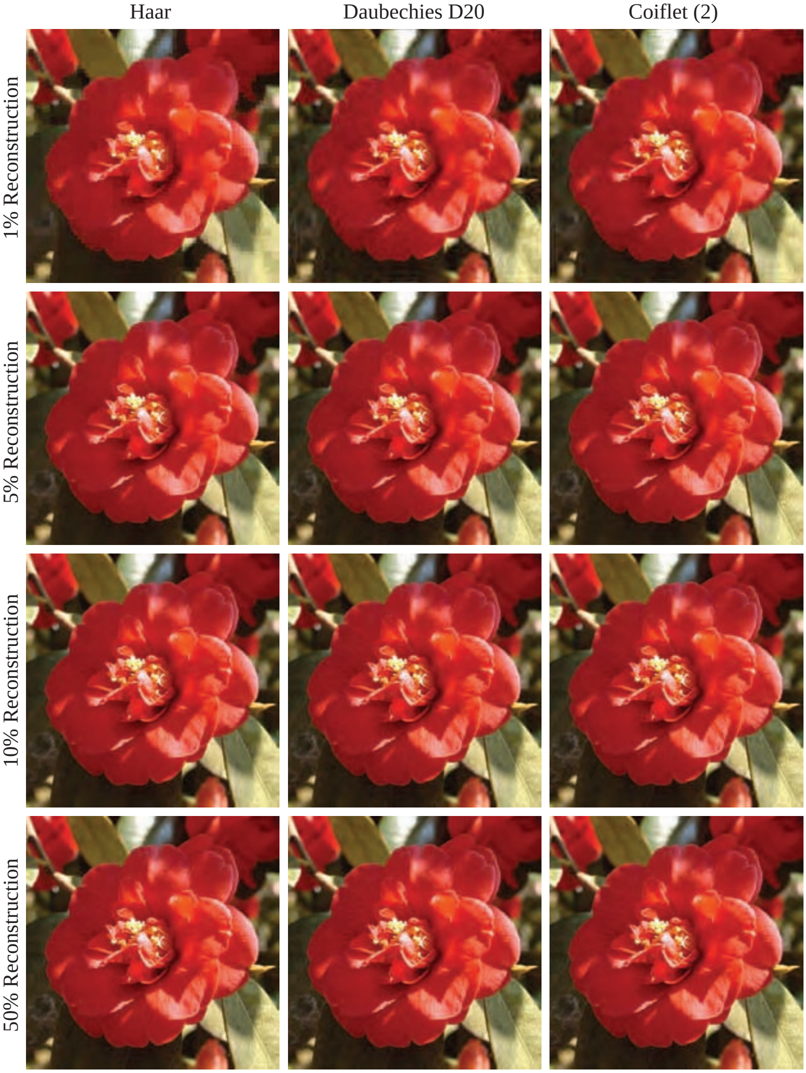
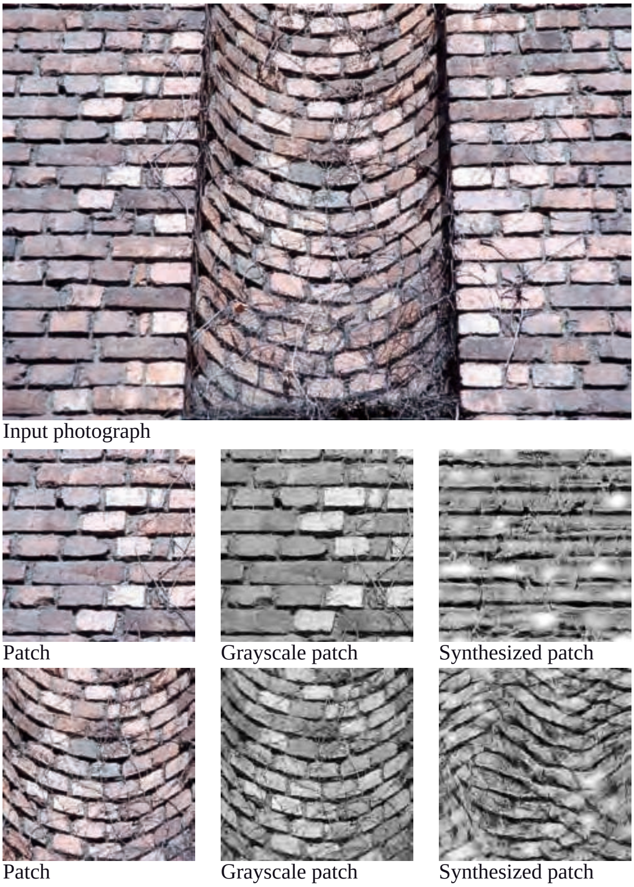


Figure 8.19. Wavelets allow images to be progressively reconstructed, which can be used in transmission. Here, we use the best 1%, 5%, 10%, and 50% of the wavelet coefficients (top to bottom), using three different wavelets (left to right): Haar, Daubechies D20, and Coiflet (2).


Input photograph

Patch

Grayscale patch

Synthesized patch

Patch

Grayscale patch

Synthesized patch

Figure 8.20. Texture synthesis examples, using analysis/synthesis on complex wavelets [581]. (Stockholm, Sweden, 2009)

that are equal in expectation over these functions, they would be visually indistinguishable [380]. This is known as the Julesz conjecture. Assuming we have such a set of functions ϕ_k, then perceptual equivalence would be obtained if:

$$E(\phi_k(X)) = E(\phi_k(Y)) \qquad\qquad k \tag{8.60}$$

In a synthesis-by-analysis approach, one would analyze an example texture to find that the expected value for each of these functions is a specific value c_k. The idea is then to generate a new random field such that it has the same expected values for each of the constraint functions:

$$E(\phi_k(X)) = c_k \qquad\qquad k \tag{8.61}$$

Portilla and Simoncelli have developed an efficient sampling algorithm that can generate such textures [581]. Rather than aim to satisfy all k constraints simultaneously, they treat each constraint sequentially.

Of interest, then, is the set of constraint functions that would lead to perceptual equivalence. It can be argued that to achieve this goal, constraint functions that emulate the early stages of human vision can be chosen. Portilla and Simoncelli have demonstrated that a viable route would be to linearly decompose an image into a multiscale representation such that the basis functions are localized, oriented, and about one-octave in bandwidth [581]. Their steerable pyramid [690, 689] includes complex filters to allow local phase information to be incorporated in the analysis.

The statistical constraints are then chosen on this basis using a technique akin to greedy algorithms: one by one a new constraint is added that captures the visual quality most noticeably missing while verifying that previously added constraints are still relevant. The resulting constraint set contains marginal statistics on texture pixels, coefficient correlations, magnitude correlation, and cross-scale phase statistics. An example of texture synthesis achieved with this method is shown in Figure 8.20.

Chapter 9

Markov Random Fields

There are many thousands of visual illusions, most of which serve to show that the human visual system does not, in fact, perform some form of inverse physics but instead uses a wide variety of assumptions, simplifications, and heuristics (see Section 2.5.5). Perhaps the most common feature of nearly all visual processing—from simple edge detection through object recognition and up to aesthetic judgment—is that these processes are *context dependent*. That is, the same visual signal can be interpreted in many different ways depending on the information right next to it (either spatially or temporally), on our previous experience, and on a host of other factors.

For example, our recognition of an edge is based not just on the local luminance change from one pixel to the next but on the other luminance changes in the image (i.e., only maxima in the gradient field are edges; see Chapter 5). Context dependency is encoded at a very early stage in visual processing, as the center-surround nature of receptive fields shows (see Chapter 2). In addition to this local context dependency, there is also evidence that global context dependency is encoded early: it has been repeatedly demonstrated that information that is clearly outside of a cell's receptive field strongly influences the response of that cell, even though by definition a cell should not be able to react to information outside its receptive fields (see, for example, [658]).

For edges and many other low-level processes, the "outside influences" come largely from cells that are immediate neighbors (the so-called *near-field*), again a decent reflection of real-world statistics. This influence is often implemented with simple lateral inhibitions. Some influence, however, comes from cells that are very far away, and this is believed to occur either through the massive feedback projections from higher visual areas down to lower visual areas or through chains of lateral connections (i.e., propagated influence). In short, to fully describe the statistics of real-world scenes, as well as how humans process the real world, one cannot look solely at individual cells, regions, or pixels. It is critical that at least the local neighborhood be examined. In the ideal case, global influences must also be taken into account.

One excellent method for embedding this context dependency in visual computing applications is through *Markov random fields* (MRFs). MRFs are based on Ising's work in the physics of condensed matter in the 1920s [369, 539]. After Ising, MRFs continued to be developed and expanded, at first mostly by physicists and mathematicians. The seminal paper from Geman and Geman in 1984 [254] brought MRFs to the field of image processing (for a historical review of MRFs, please see [399]).

Since the 1980s, MRFs have been successfully applied to an incredible range of problems in computer graphics and computer vision, including denoising and general image restoration, image segmentation, edge detection, optical flow, shape detection, depth segmentation, and texture analysis (for more on these and other examples, see, for example, [460]). Although most MRFs focus on highly local context dependencies similar to the effect of retinal cells on their immediate neighbors and other low-level visual tasks, modeling global dependencies is also possible. Further, it is also possible to use MRFs to learn the statistical dependencies in image sets (including natural image sets). This often takes the form of estimating the parameters of some experimenter-specified model (generally using Maximum Pseudo Likelihood Estimation or contrastive divergence).

In this chapter, we focus on the basic concepts necessary to understand MRFs and provide a few insights into several statistical regularities that can be detected with them. For further reading, we direct the reader to any of the many excellent books an Markov random fields and their myriad applications, such as [64] and [460].

9.1 Image Interpretation

In some areas of psychology, perception is thought of as a form of hypothesis testing (see, for example, [282]). That is, given (ambiguous) information at a specific spot on the retina (sometimes in vision research, this is called the *proximal stimulus*), the visual systems poses hypotheses or guesses as to what the real-world object or event (called the *distal stimulus*) might be. If necessary, the visual system then searches for information. If there is not enough information in the current proximal stimulus to confirm or disprove the hypothesis, then new information is acquired (e.g., the person moves to get a better view). In other words, vision is sometimes thought of as the process of inference where a set of proximal stimuli (the retinal excitations) are assigned a value from a (limited) set of interpretations. For example, the visual system might try to determine if the information at a given cell came from an edge or not. Likewise, cells higher in the visual system might try to determine if the stimulation at a given cell came from a tree trunk, from leaves, from the ground, or from the sky.

Within the field of image analysis using Markov random fields, this process of inference is often called *labeling*. More formally, let S be a set of m observations or *sites*:

$$S = \quad 1, \ldots, m \tag{9.1}$$

where the numbers are indexes to individual observations. These might be cells on the retina, pixels in an image, segmented regions in an image, and so on. In a two-dimensional $n_1 \times n_2$ image, the set can also be written as:

$$S = \quad (i,j) \ 1 \le i \le n_1, 1 \le j \le n_2 \tag{9.2}$$

Note that connectivity between sites is important in MRFs, but this is usually represented separately. The global order of the sites, on the other hand, is not important. Thus, 2D images can be represented with a single parameter h, where $h = \quad 1, 2, \ldots, m \quad$ with $m = n_1 \times n_2$.

Labeling also requires the definition of a set of possible interpretations that a site may have. This list of categories or *labels* can be nominal (and therefore contain only information about identity), such as edge versus non-edge; ordinal (containing information about the identify and ranking of categories), such as relative depth ordering (closest surface, next closest surface,. . ., most distant surface); or even interval (containing information about the identity, ranking, and relative distance between categories), such as the intensity at a pixel (0,. . ., 255).

The class of operations that are possible on the labels is strongly dependent on the nature of the labels. For nominal label sets, it is only possible to determine if two sites have the same label or not. For ordinal and interval label sets, it is possible to define operations like similarity and distance. If the labels are discrete, the set of M labels can be written as:

$$\mathscr{L} = \quad 1, \ldots, M \tag{9.3}$$

with the indices pointing to individual categories or labels. If the labels are continuous (such as a floating point representation of luminance), then the set $\mathscr{L}$ can be thought of as a line in $\mathbb{R}$ or some portion of it:

$$\mathscr{L} = [X_1, X_2] \subset \mathbb{R} \tag{9.4}$$

with X_1 and X_2 being the lower and upper bounds of the set, respectively.

We can then interpret the image. More specifically, we assign each site $i \quad S$ a value f from the set of labels $\mathscr{L}$. Commonly the label at a given site i is denoted with f_i. The set $f = \quad f_1, f_2, \ldots, f_m \quad$ of labels for all m sites can be called a labeling of the sites in S in terms of the labels in $\mathscr{L}$. If all sites can be assigned a single label then we have a *mapping* f, which can be written as:

$$f : S \quad \mathscr{L} \tag{9.5}$$

Depending on the nature of the set of labels, we can create a wide variety of maps. If the labels are edge, non-edge then the mapping f_i is an *edge map*. If $\mathscr{L}$ consists of distances from the viewer, then f is a *depth map*. In the area of Markov random fields, a map is often called a *configuration*.

If all sites have the same set of labels $\mathscr{L}$, then the set of all possible configurations $\mathscr{F}$—which is also called the *configuration space*—is the Cartesian product of the set of labels $\mathscr{L}$ by itself m times, with m being the number of sites. If $\mathscr{L}$ is discrete, then the configuration space is combinatorial: with m sites and M labels, there are M^m possible mappings. For example, if there is only one label, then there is only one configuration (or interpretation of the data) regardless of the number of sites: $1^m = 1$. If there are two labels, then with two sites there are four logical possible configurations: both sites have label 1 ($f = \quad 1,1 \quad$), both sites have label 2 ($f = \quad 2,2 \quad$), site 1 has label 1 and site 2 has label 2 ($f = \quad 1,2 \quad$), or site 1 has label 2 and site 2 has label 1 ($f = \quad 2,1 \quad$).

9.2 Graphs

As was mentioned earlier, most visual processes are context-dependent and can be modeled by looking at the influence of sites that are immediately adjacent to the site of interest. This, of course, requires some formal notion of adjacency. In essence, i and j can be thought of as being connected to one another. One way of explicitly representing this is with a graph $G = (S, E)$, where E is the set of edges (i.e., the a set of adjacent sites). If two sites $i, j \quad V$ are neighbors, then they will have an edge $(i, j) \quad E$ between them. One important property of adjacency in the context of Markov random fields is that, like neighbors in real life, the relationship is mutual. If the site i is adjacent to the site j, then j must be adjacent to i. This is represented with an *undirected* graph, which is equivalent to saying if $(i, j) \quad E$, then $(j, i) \quad E$. Note that, as in real life, a site cannot be its own neighbor $(i, i) \ / E$.

9.2.1 Neighborhood Systems

Now that we have a method for formalizing context, we need to define what we mean by immediate or local context. The neighborhood $\mathscr{N}_i$ of a site i is defined as the set of all sites that share an edge with i:

$$\mathscr{N}_i = \quad j \quad j \quad S, (i, j) \quad E \tag{9.6}$$

The complete set of all neighborhoods, then, is a *neighborhood system*:

$$\mathscr{N} = \quad \mathscr{N}_i \quad i \quad S \tag{9.7}$$

In practice, this is made easier by the fact that most images are spatially regular (in fact, they form a lattice). Thus, the set of neighbors for a given site

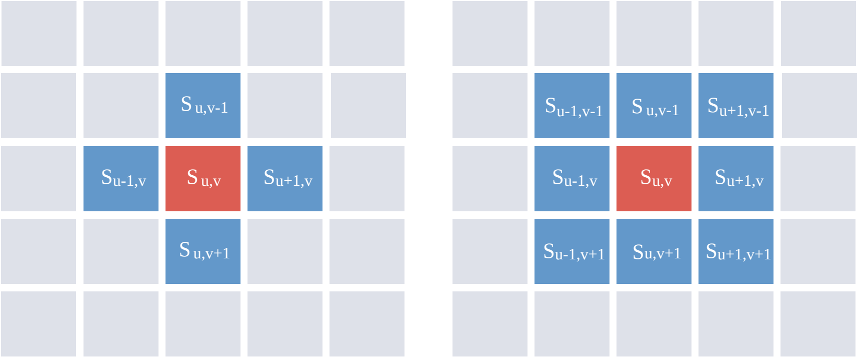


Figure 9.1. Neighborhoods: first-order or 4-neighborhood (left) and second-order or 8-neighborhood (right).

is visually obvious. There are two common neighborhood systems in lattices, depending on whether diagonals are allowed. The simplest case is a first-order neighborhood (also called a 4-neighborhood), where we do not look at diagonals (see Figure 9.1). For example, the 4-neighborhood of the site $S_{u,v}$ consists of:

$$\mathcal{N}^1_{u,v} = \quad S_{u-1,v}, S_{u+1,v}, S_{u,v-1}, S_{u,v+1} \tag{9.8}$$

The superscript denotes that we are using a first-order neighborhood. In a second-order neighborhood (also called an 8-neighborhood), diagonal neighbors are also considered (see Figure 9.1). Thus, the 8-neighborhood of site $S_{u,v}$ consists of:

$$\mathcal{N}^2_{u,v} = \quad S_{u-1,v}, S_{u+1,v}, S_{u,v-1}, S_{u,v+1}, S_{u-1,v-1}, S_{u-1,v+1}, S_{u+1,v-1}, S_{u+1,v+1} \tag{9.9}$$

It should be mentioned that pixels at the edges of an image have fewer neighbors. Often, this is avoided in practice by "wrapping" an image (so that in an $n_1 \times n_2$ image, pixel $I(0,y)$ is connected to $I(n_1,y)$).

If the graphs are irregular and there is no explicit concept of connection (for example, as might occur in an edge map), we need to define some distance function $dist(i,j)$ between sites i and j. All sites that fall within a given threshold distance are included in the neighborhood. For example, if the location of the sites can be interpreted spatially, then an acceptable distance function would be to take a circle centered on the site of interest and all cells that are within a threshold radius are considered to be within the local neighborhood. Of course, care needs to be taken when defining both the distance function and the threshold to ensure that the neighborhood range completely includes the initial site itself (for example, since most edges are spatially extended, it is important to ensure that a given edge in an edge map is within its own neighborhood calculations). If the graph cannot be spatially interpreted, then some other form of distance function needs to be defined.

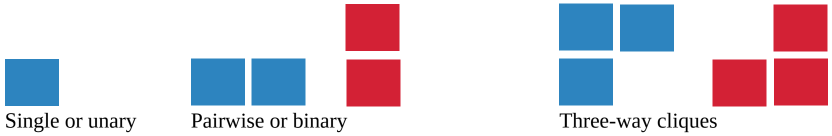


Figure 9.2. The possible cliques in a 4-neighborhood: single or unary cliques (left), pairwise or binary cliques (middle), and three-way cliques (right).

9.2.2 Cliques

The final term that is needed is a *clique*, which is a subgraph of G in which all the elements are neighbors of each other. Cliques may be unary (containing just the site itself), binary (pairs of neighbors), triplets of neighbors, and so on. The set of all unary cliques will be referred to as C_1, the set of all binary cliques as C_2, and so on. The set of all cliques C is given by:

$$C = C_1 \quad C_2 \quad \ldots \tag{9.10}$$

Critically, cliques are ordered, so that (i, j) is not the same as (j, i). Additionally, a site can be in more than one clique. Thus, a single site is in its own unary clique as well as possibly in one or more binary cliques. Possible cliques within a 4-neighborhood in a lattice are shown in Figure 9.2. These include the sites themselves (unary cliques), horizontal and vertical pairs (binary cliques), and the three-way cliques. A *maximal clique* is a clique where it is not possible to add any more sites while still retaining complete connectedness.

9.3 Probabilities and Markov Random Fields

Some of all the possible configurations f in the configuration space $\mathscr{F}$ are more likely than others. Markov random fields capture this intuition more concretely and allow us to model contextual dependencies. For an image I whose pixels make up the set S of m sites, we can define a set $F = \quad F_1, F_2, \ldots, F_m \quad$ of random variables that are the actual mappings from S to $\mathscr{L}$. Each site's random variable F_i takes one of the possible mappings f_i. The probability that the actual mapping at a specific site F_i is a particular possible mapping for that site f_i is given by $p(F_i = f_i)$, or $p(f_i)$ for short. Following standard notation, the $(F = f)$ term will be shortened everywhere to f. The set or family F is called a random field.

If there were no context dependencies, then the probability that a given site has a given interpretation $p(f_i)$ would depend only on the site itself. That is, this probability would be conditionally independent of all other sites j:

$$p(f_i \; f_j) = p(f_i) \tag{9.11}$$

where f_j is the mapping at all other sites $j = i$. As we have seen, however, this is generally not the case and therefore context dependencies need to be modeled.

The random field F is a Markov random field if the following two conditions are met:

Positivity. The probability of a mapping occurring is greater than zero for all possible mappings in the configuration space:

$$p(f) > 0 \qquad\qquad f \quad \mathscr{F} \tag{9.12}$$

Markovianity. The probability that a site i takes a specific value f_i is dependent only on the local neighborhood of that site $\mathscr{N}_i$

$$p(f_i \; f_{\; S-i} \;) = p(f_i \; \mathscr{N}_i) \tag{9.13}$$

where $\; S-i \;$ is all of S except i.

9.3.1 Gibbs Distributions

The Hammersley-Clifford theorem [297] states that if the above two assumptions are met then we can write the joint probability $p(f)$ as a Gibbs distribution:

$$p(f) = \frac{1}{Z} e^{-\frac{1}{T}U(f)} \tag{9.14}$$

where Z is the normalizing constraint called the partition function, T is a global parameter called the *temperature* (often assumed to be 1), and $U(f)$ is the prior energy. The partition function is given by:

$$Z = \sum_{f \quad \mathscr{F}} e^{-\frac{1}{T}U(f)} \tag{9.15}$$

and ensures that the sum of all probabilities equals unity:

$$\sum_{f \quad \mathscr{F}} p(f) = 1 \tag{9.16}$$

The prior energy $U(f)$ is given by:

$$U(f) = \sum_{c \quad C} V_c(f) \tag{9.17}$$

where $V_c(f)$ denotes the prior potentials (the values of subgraphs of G), which depend only on f_i, with $i \quad c$ (the labels for sites within a clique). Lower energy states $U(f)$ are more likely. Note that the temperature T controls the sharpness of the distribution. The higher T is, the more evenly distributed all configurations are. As T approaches zero, the distribution concentrates around the global energy minima.

In some cases, it can be convenient to rewrite the energy function as a sum of terms, with each term relating to a clique size (single, pair, triple, etc.):

$$U(f) = \sum_{i \ C_1} V_1(f_i) + \sum_{i,j \ C_2} V_2(f_i, f_j) + \sum_{i,j,k \ C_3} V_3(f_i, f_j, f_k) + \dots \tag{9.18}$$

where C_1 contains the unary cliques, C_2 contains the binary cliques, C_3 contains the triple cliques, and so on. Likewise, V_1 denotes the potentials of the unary cliques, and so on.

9.3.2 Auto-Models

The simplest MRF model would be to restrict our considerations solely to cliques up to size two. If the set of sites S is a two-dimensional image and we use only a first-order neighborhood, then the auto-model is the same as the Ising model. In these cases, Equation (9.18) can be rewritten as:

$$U(f) = \sum_{i \ S} V_1(f_i) + \sum_{i \ S} \sum_{j \ \mathcal{N}_i} V_2(f_i, f_j) \tag{9.19}$$

Note that here we sum over *all* unary and binary cliques, and that each site is in a unary clique and four binary cliques. Combining Equation (9.19) with Equations (9.14) and (9.15), and assuming T is 1 gives the conditional probability of a site based on its neighbors:

$$P(f_i \ f_{\mathcal{N}_i}) = \frac{\exp\left(-V_1(f_i) - \sum_{j \ \mathcal{N}_i} V_2(f_i, f_j)\right)}{\sum_{f_i \ \mathcal{L}} \exp\left(-V_1(f_i) - V_2(f_i, f_j)\right)} \tag{9.20}$$

In the original Ising model, there were two states $\mathcal{L} = \ -1, +1$. The Potts model provides an important extension to the Ising model by allowing the set of labels to have more than two states. Ising also ignored unary cliques (for physically based reasons that were relevant to the specifics of the phenomenon he was studying), leaving only binary term:

$$U(f) = \sum_{c \ C} V_c(f) = -\beta \sum_{(i,j) \ C_2} f_i f_j \tag{9.21}$$

where β captures the interaction between neighbors. If $\beta > 0$, then neighbors will tend to have the same value. One can generate images using an Ising model, with different values for β giving different images. It is also possible to use different β for horizontal and vertical cliques, yielding anisotropic images. Interestingly, as was shown in Chapter 5, the difference between pairs of neighbors is an approximation of the gradient, so the binary terms are essentially modeling the first derivative properties of the image.

Since the joint probability $p(f)$ measures the probability that a given configuration occurs, we can bias the outcome towards specific configurations by setting the clique potentials carefully. Thus, much of the interest in MRFs is based on the choice of potential functions (using prior knowledge about interactions between neighboring sites and labels) so that the desired behavior arises.

9.4 MAP-MRF

MRFs and their connections with Gibbs distributions provide us with the ability to calculate the probability of different configurations. As mentioned, this is interesting for texture generation. A simple Ising or Potts model can produce convincing textures for different settings of the available parameters.

Within image analysis, however, MRFs present some limitations. As stated earlier, problems in low-level vision or image analysis can be thought of as inference or hypothesis testing problems: we have a given image and we want to interpret it. The probability of a given configuration $p(f)$ simply tells us how likely any given configuration is in principle (with lower energy configurations being more likely). Often, however, we would like to know which of the many possible configurations applies to our current image. MRFs by themselves have no means for incorporating real observations and thus are insufficient to answer this question. For this reason MRFs are often embedded within a Bayesian framework.

Although a full treatment of Bayesian statistics is beyond the scope of this book,[1] a few comments are required. An example of how Bayesian inference can be used in the context of MRFs is as follows. Assume that there is a real-world from which we take some picture or observation. Due to the process of image capture, however, the information from the real-world gets slightly corrupted. If we know what the real-world scene might have contained and have some idea of how the information was corrupted, we could make some plausible guesses about what the scene really contained based on the (corrupted) image. In other words, there is a known dataset (e.g., the observed pixel values) d and the possible interpretations of that dataset (the labels $\mathscr{L}$). The knowledge about how likely different sets of interpretations or configurations are captured by $p(f)$, as we saw earlier. In Bayesian statistics, this is called the *a priori* or prior. Knowledge about how the information was corrupted during the observation process is captured by the *likelihood* $p(d\ f)$. Note that this is very much application-dependent. Given this information we can calculate how likely any given interpretation or configuration

[1] The interested reader is directed to [460] for more on Bayesian statistics in image processing. See also the appendix for a brief discussion on Bayes' rule.

f is given the observed data:

$$p(f\ d) = \frac{p(d\ f)p(f)}{p(d)} \tag{9.22}$$

This is called the *a posterior* or posterior. Note that the denominator $p(d)$ is a constant since the image is known, but its calculation is generally intractable. Fortunately, this is usually not a problem as we shall see. Since we want to know which interpretation is most likely, it is common to calculate the maximum *a priori* or MAP estimate:

$$f* = argmax_{f\ \mathscr{F}} p(f\ d) \tag{9.23}$$

Since $p(d)$ is constant, it is not needed for the MAP, which therefore becomes:

$$f* = argmax_{f\ \mathscr{F}} p(d\ f)p(f) \tag{9.24}$$

In sum, when using MRFs in a Bayesian framework, we need to define a neighborhood system, then define the allowable cliques, and then define the clique potentials. This gives us the prior $p(f)$. We then derive the likelihood energy based on assumptions on the underlying distributions and processes involved in generating the dataset. Finally, we derive the posterior energy and find the maximum.

9.5 Applications

We present two examples where MRFs have been successfully used to detect statistical regularities in the image. For more on what regularities MRFs have been used to model, please see [460].

9.5.1 Image Restoration

One of the classic examples of how an MRF can be used is simple image restoration. There are many, many different versions of this procedure. The one we present here is based on [255]. We assume that we have an image, represented as a two-dimensional spatial lattice, and that we only look at the first-order neighborhood. The set of sites $S = 1, \ldots, m$ contains the pixel locations, while the set of labels $\mathscr{L} = 0, \ldots, 255$ contains possible 8-bit pixel luminances. The possible cliques are the sites themselves and pairs of vertically or pairs of horizontally adjacent pixels. The binary clique potential is $V_2(f_i, f_j) = \beta g(f_i - f_j)$ where β is a scalar and $g(.)$ is a function that penalizes differences between f_i and f_j (i.e., it penalizes smoothness violations). Further, the assumption is made that the "correct" image was corrupted by adding white noise[2] $d_i = F_i + \mu_i$ where d_i is the

[2] Note that the noise in real imaging systems is often not white. In real restoration processes usually a Poisson distribution is assumed [254]. Interestingly, there is some evidence that electronic

observed pixel value, F_i is the correct pixel value, and μ_i is a zero-mean Gaussian distribution with variance σ^2. Thus, the likelihood is:

$$p(d\ \ f) = \left(\frac{1}{2\pi\sigma^2}\right)^{\ S\ /2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i\ \ S}(d_i - f_i)^2\right) \tag{9.25}$$

where S is the number of data points. Using an Ising model and ignoring the unary terms gives:

$$p(f) = \frac{1}{Z}\exp\left(-\beta\sum_{i,j} f_i f_j\right) \tag{9.26}$$

which gives the *a posteriori* distribution of:

$$p(f\ \ d) = \frac{1}{Z}e^{-(\beta\sum_{i,j} f_i f_j)-(\frac{1}{2\sigma^2})\sum_{i\ \ S}(d_i-f_i)^2} \tag{9.27}$$

To find the solution the energy function in the exponent is minimized:

$$U(f) = -\left(\beta\sum_{i,j} f_i f_j\right) - \left(\frac{1}{2\sigma^2}\right)\sum_{i\ \ S}(d_i - f_i)^2 \tag{9.28}$$

9.5.2 Object Segmentation

Another classic application of MAP-MRF is to segment an object from the rest of an image (part of the classic graph cut problem). Here, we present Boykov and Jolly's algorithm [70], which uses an auto-model. Specifically, the user selects a few pixels that are a part of the object of interest and a few pixels that are definitely not on the object of interest. The unary term focuses on the how well the remaining pixels match a description of the object or of the background. Although a number of potential functions are named, the one used in this implementation assigns a value to each pixel based on how well it matches the log of two histograms (one each for object and background pixels, respectively). The binary term governs smoothness, again with a list of potentially useful functions. In the specific implementation, a simple penalty was used:

$$V_2(f_i) = \exp\left(-\frac{(I_i - I_j)^2}{2\sigma^2}\cdot\frac{1}{\mathrm{dist}(i,j)}\right) \tag{9.29}$$

for $i = j$. This gives us the following energy function:

$$U(f) = \lambda\cdot\sum_{i\ \ S} V_1(f_i) + \sum_{i\ \ S}\sum_{j\ \ \mathcal{N}_i} V_2(f_i, f_j) \tag{9.30}$$

where $\lambda \geq 0$ is a constant that determines the relative weight of the unary term. Minimizing the resulting energy function yields a rather successful segmentation of the object.

noise—and thus the corrupting factor between real scene and observation—is not white, but the power spectrum actually follows a power law [396, 647]). Using a power law rather than white noise in the likelihood might improve the performance of MRFs in image restoration.

9.6 Complex Models and Patch-Based Regularities

Most of the MRF models contain relatively simple underlying functions. There are, however, several methods for combining multiple simple models to make a more powerful one. By combining different classes of models and learning their parameters, it should be possible to ensure that we can adequately represent real-world statistics. Perhaps the most promising approach to combining simple models is the *Products of Experts* (PoE) technique. Here, we present the approach, followed by how it can be used with MRFs (a procedure called *Field of Experts*) to learn the statistical regularities of image patches.

9.6.1 Products of Experts

High-dimensional datasets can be modeled with PoE [331]. The basic idea is to create several straightforward models called experts, each of which is trained to model a part of the dataset (such as one of the many dimensions in a high-dimensional dataset) and then multiply the experts together (and then normalizing) to model the full dimensionality of the dataset. The multiplication overcomes some limitations of mixture models (see Chapter 7). The experts essentially model how likely a given image is given a specific set of models $p(d\ \theta_n)$ where d is the observed data (i.e., the image) and θ_n represents the parameters of model n. The joint probability is then given by:

$$p(d\ \theta_1,...,\theta_n) = \frac{\prod_n p(d\ f_n,\theta_n)}{\sum_c \prod_n p(c\ f_n,\theta_n)} \tag{9.31}$$

where c indexes all possible vectors in the dataspace (e.g., all possible images). Using machine-learning techniques (including sampling and contrastive divergence to make the problem computationally tractable), it is possible to learn the parameters of the individual experts on a training dataset. The trained experts can be used (with some limitations) to process new data. For example, a large number of simple experts, each of which is a mixture of a uniform distribution and a Gaussian aligned to a single axis in the dataset, can give rise to edge detectors when trained on a training set of images that consist of a single edge each [331].

Since natural image statistics tend to be highly kurtotic, using Student-t distributions as the basis of the experts is a viable alternative [793]. This procedure is called a *Product of T-Distribution* (PoT) model. Learning the parameters of a set of Student-t-based experts on 5 × 5 pixel image patches from a database of natural images has produced filters that look much like those obtained in sparse encoding approaches.

9.6.2 Fields of Experts

Most MRFs focus on first-order neighborhoods and use simple models, meaning they can only explicitly capture simple, very local image structures [631, 634]. It is clear, however, that both longer-range and more complex interactions are very important in images and need to be modeled. One approach to include larger ranges of context using more complex models is to use Product of Experts in a Markov random field, which is called a *Field of Experts* (FoE) for short. In essence, the experts of a PoE are learned, as mentioned earlier, and then used as priors in MRFs with very large, overlapping cliques. Student-*t* distributions and a variant on the L1 norm have both been shown to be effective as the basis of such experts [631, 634].

The Fields of Experts were tested on the Berkeley Segmentation Benchmark [501]. For computational reasons, the size of the image was limited to being between three and five times as large as the patches. To test the approach, 20,000 image patches were randomly extracted (usually each 5×5 pixels large). These were subsequently converted to grayscale. They then created subsets of 200 patches each (again, for computational reasons) and used a different subset at each training step. There were usually 5,000 training steps in the experiments. Visual inspection of the recovered filters does not immediately lead to any recognizable structure. They do not look like edge detectors or any of the usual filters found in sparse encoding approaches (or even in PoE approaches). Roth et al. suggest that this is due to the overlap of the cliques.

A direct comparison of the models obtained under different circumstances (e.g., clique sizes, training steps, etc.) is not possible since the partition function cannot be evaluated. Instead, the effectiveness of different FoEs on different tasks, such as image denoising and inpainting can be assessed. The FoE model performs very well, showing that these models can capture some regularities in an image.

9.7 Statistical analysis of MRFs

Using MRFs within a Bayesian framework is an elegant way to perform many computer vision tasks or even to model the human perceptual system. It is amazingly flexible and can, in principle, measure a truly astonishing variety of statistical regularities. A part of this flexibility is due to the fact that the researcher is free to specify the underlying model. Once a model is specified, it is then possible to use a set of images as a training dataset to estimate the parameters of the model underlying the MRF.

In most cases, however, it is unclear how well the model has represented the images or what features it has modeled [249]. This is largely because the partition function is computationally intractable. Thus, in most cases, the success of learning image parameters is tested by using the MRF to perform some task such as image denoising or image segmentation.

Zhu and Mumford [822], however, suggested that an appropriate way to see what the models are modeling is to take advantage of the generative nature of MRFs. That is, as mentioned earlier, it is possible to use a (trained) MRF to synthesize textures. To examine the statistics of their MRF, Zhu and Mumford trained it on 44 real-world images and then generated several images from their MRF—a process they called sampling the MRF. Finally, they examined the statistical moments of the marginal distributions of both real and synthesized images. As usual, the intensity histograms of real-world images had high kurtosis and heavy tails and the statistics were (mostly) scale invariant. The synthesized images neither visually resembled real-world images nor did their statistics. Although the intensity histograms of the synthetic images had flat tails, they had only small spikes. Zhu and Mumford suggested that this was due to the size of the filters used. Adjusting the filter size improved the similarity to real-world images somewhat.

Later, several researchers simultaneously revived the idea of using a trained MRF to synthesize images and then examine their statistical moments [391, 452, 482, 600, 549, 656]. As an example, Lyu and Simoncelli [482] trained an MRF based on Fields of Gaussian Scale Mixtures using five images (each 512 × 512 pixels) using 5 × 5 pixel neighborhoods. The marginal statistics of the samples turn out to be clearly non-Gaussian but not as heavy-tailed as real-world images. The joint statistics were also close but note quite the same as real-world images.

Ranzato and colleagues [600] augmented the MRF to have two latent variables rather than one. The first set of latent variables models pixel intensities. The second set models image-specific pixel covariances. They call this a gated MRF. They use a Product of Student-*t* model (PoT) [730] as the base MRF and they also add the ability to model mean pixel intensities and not just pixel pair differences. After training, they sampled the model (with patches of 160 × 160 pixels) and assessed how similar the intensity histograms of synthesized images and random images were to either random images or real-world images using the Kullback-Leibler divergence. This is a measure of how different two probability distribution functions are, as explained in the appendix. They found that adding mean pixel intensities greatly improves the quality of the generated images. Interestingly, not accounting for the mean pixel intensity yields intensity histograms that are closer to random images than real ones.

Schmidt and colleagues suggest that a more flexible underlying model, a variant of Gaussian scale mixtures, might yield better results than traditional ones [656]. They examined intensity histograms and multiscale derivatives using the Kullback-Leibler divergence. They found that for both pairwise MRFs and Fields of Experts, this model yields a higher kurtosis than the older models. We used this procedure on a set of 15 images (shown in Figure 7.8) using pairwise MRF and obtained the results shown in Figure 9.3. We also used a 3 × 3 clique Fields

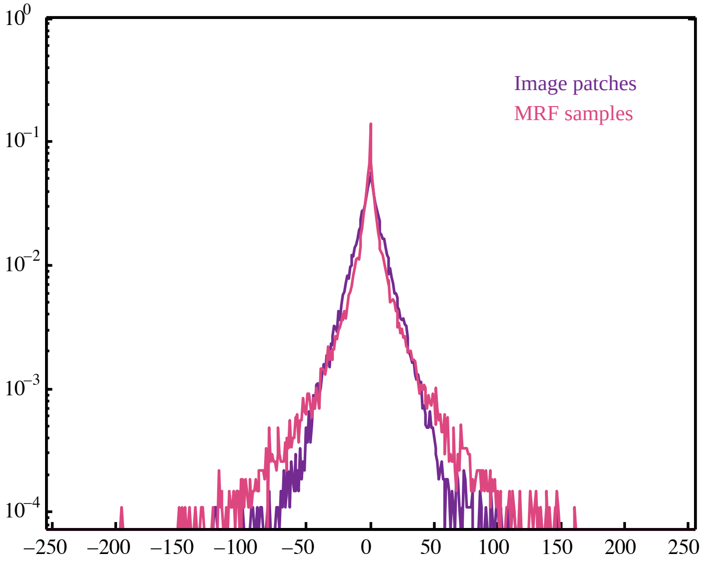


Figure 9.3. Filter marginals computed on an MRF model using a variant of Gaussian scale mixtures as the underlying model [656]. They are compared with the derivative statistics of the image patches themselves. The images used to generate these results are shown in Figure 7.8.

of Experts on the same set of images and obtained the results shown in Figure 9.4. Note that the experts shown in this figure have a broad base and a narrow peak, showing heavy-tailed behavior. On of the applications of this approach is image inpainting, a result of which is shown in the bottom-right of Figure 9.5. Note that the Fields of Experts improve the quality of the result somewhat relative to pairwise MRFs.

Schmidt and colleagues also find that the Bayesian minimum mean squared error estimate (MMSE) was better than MAP. As their approach lends itself not only to analysis but also applies in a generative setting, Schmidt et al. showed its applicability in several scenarios, including image inpainting (a result obtained with this method is given in the bottom-left of Figure 9.5). Subsequently, they further extended the model and showed that the synthesized images more accurately matched the multiscale derivative statistics, random filter statistics, and joint feature statistics of of real-world images [249].

Several other researchers have shown that MRFs trained on patches do reasonably well but are not perfect. Karklin and Lewicke [391] examined the properties of MRFs based on multivariate Gaussians trained on 20 × 20 pixel patches from real-world images and compared the responses to those of simulated neurons. They found that the properties are similar to those in visual areas V1 and V2 in the human visual system. Likewise, Osindero and Hinton [549] examined the

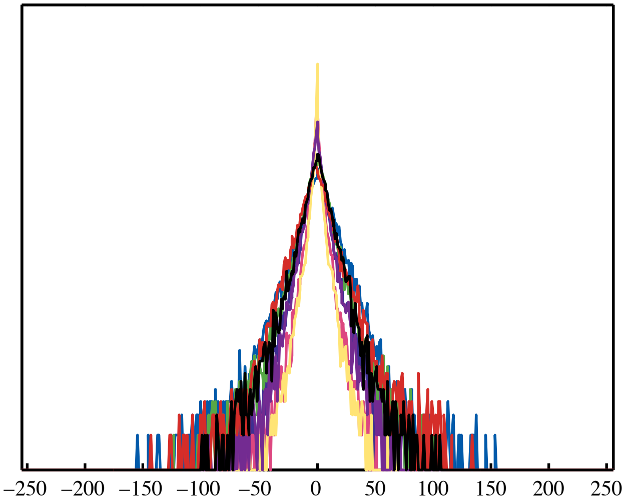


Figure 9.4. Eight filter marginals computed on a 3 × 3 clique FoE model [656].

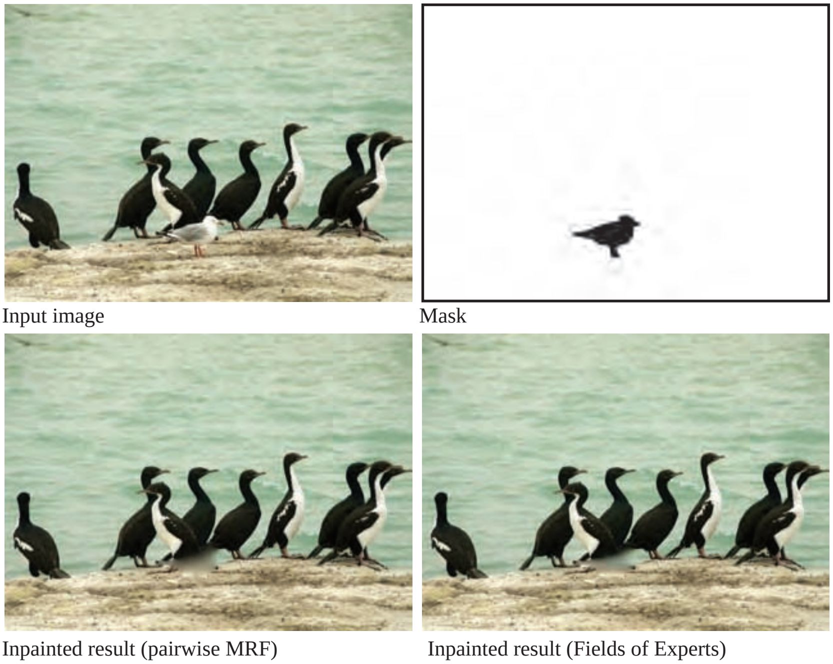

Input image
Mask
Inpainted result (pairwise MRF)
Inpainted result (Fields of Experts)

Figure 9.5. Inpainting result using pairwise MRFs (bottom left) and FoE (bottom right), generated with the method of Schmidt et al. [656]. (Dunedin, New Zealand, 2012)

performance of a model that has several hidden layers, each with its own MRF (conditioned on the variables in the layer above). They used 150,000 patches each of 20×20 pixels to train their model. They generated 10,000 sample images and found that the intensity histograms for the real-world patches have a high kurtosis (8.3). The version of their model that includes lateral connections yields images with a kurtosis of 7.3, while the version without lateral connections has a kurtosis of 3.4.

Part III

Beyond Two Dimensions

Chapter 10

Color

We have so far considered statistical regularities either without specifically including the notion of color, or deliberately ignoring color and assuming a single luminance channel. Although this is sufficient for studying many of the regularities within natural scenes, color deserves special treatment as it is one of the most informative aspects of the visual world. The use of color in images and art can be used to create a specific mood and to induce particular emotions, it can serve as a signal (e.g., to indicate fruit ripeness in nature or whether we can cross the road in manmade environments) and even give indication about the type of environement or time of the year [786].

In general, images are formed as light is transduced to electric signals. This happens both in the retina as well as in electronic image sensors. Before that, light is emitted and usually reflected several times, encountering several different surfaces before it reaches an electronic or physiological sensor. This process is modeled by the rendering equation, given by:

$$L_o(\lambda) = L_e(\lambda) + \int_{\Omega} L_i(\lambda)\ f_r(\lambda)\ \cos(\Theta) d\omega \tag{10.1}$$

This equation models the behavior of light at a surface: light L_i impinging on a surface point from all directions Ω is reflected into an outgoing direction of interest, weighted by the bi-directional reflectance distribution function (BRDF) f_r. The amount of light L_o going into that direction is then the sum of the reflected light and the light emitted by the surface point (which is zero unless the surface is a light source). Here, we have specifically made all relevant terms dependent on wavelength λ to indicate that this behavior happens at all wavelengths, including all visible wavelengths. This means that L_o can be seen as a spectral distribution.

Light reaching our eyes is therefore carrying information about the intensity and direction of illumination but also about the materials and colors of objects that it has interacted with in its path. For a sensor—be it biological or electronic—to be able to perceive color, special adaptations are necessary. Ideally, to accurately represent color, the full spectrum of light should be captured. This, however, would require a prohibitive amount of processing.

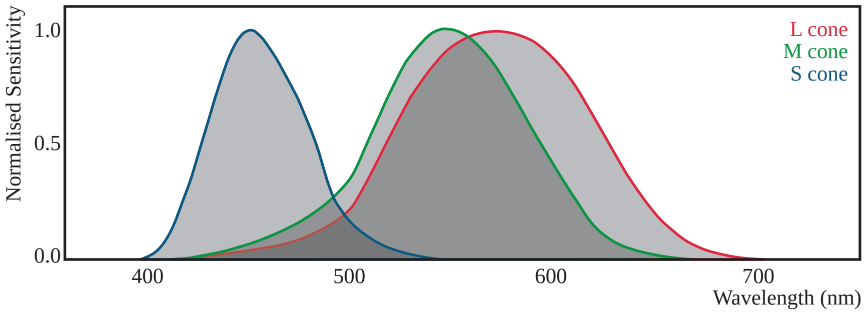


Figure 10.1. The three cone photoreceptor sensitivity curves.

To reduce the amount of data, the light spectrum can instead be sampled through cells with different sensitivities. The human retina, for instance, is populated by three types of cone photoreceptors, while the eyes of mantis shrimp contain at least eight visual pigments with narrowly tuned sensitivities [500] and many species of birds are tetrachromatic [47]. In all of these cases, although the precise form of sampling of the spectrum differs, the purpose is the same: to reduce the amount of data to be transmitted and processed while minimizing loss of information.

This remains an important goal in further stages of the visual system, with many processes relying on particular regularities in the color information of the environment, allowing us to quickly and effortlessly understand our environment. In this chapter, we will review the main color processes of the human visual system and discuss how they relate to certain statistical regularities and patterns in natural environments. We will also look at the implications and applications of these regularities in imaging disciplines.

10.1 Trichromacy and Metamerism

The human visual system is well equipped for viewing color, and several different processes take place in our visual pathways that mediate its perception. Several models were proposed in the last centuries to account for the different aspects of color vision. One of the earliest theories suggests that new colors can be created by mixing three other colors [814, 326]. This was determined by a color matching experiment where participants were asked to match given colors by mixing a set of light sources with adjustable intensities. Helmholtz found that three light sources were necessary and sufficient for matching any given color. This led to what is now known as the *Young-Helmholtz trichromatic theory*.

Within the human visual system, photopic vision is indeed mediated by three types of cone photoreceptors, each with a different peak sensitivity. Each cone type integrates the incident light according to a different weight function:

$$L = \int_{\lambda} L_o(\lambda)\, \bar{l}(\lambda)\, d\lambda \tag{10.2a}$$

$$M = \int_{\lambda} L_o(\lambda)\, \bar{m}(\lambda)\, d\lambda \tag{10.2b}$$

$$S = \int_{\lambda} L_o(\lambda)\, \bar{s}(\lambda)\, d\lambda \tag{10.2c}$$

where $\bar{l}(\lambda)$, $\bar{m}(\lambda)$, and $\bar{s}(\lambda)$ are the weight functions (responsivities), which peak at wavelengths that are perceived roughly as red, green, and blue (Figure 10.1), specifically at 565 nm, 545 nm, and 440 nm [68, 707]. The letters L, M, and S stand for "long," "medium," and "short" wavelengths. A spectral distribution $L_o(\lambda)$ therefore gives rise to a triple of numbers (L, M, S), which represent the signal that is passed on from the photoreceptors to the remainder of the human visual system. Such triples are called *tristimulus values*.

An important property of this behavior is that different spectral distributions can integrate to the same tristimulus values since each receptor type responds to a wide range of wavelengths. Such spectral distributions are then necessarily perceived to be identical, which is known as *metamerism* [204]. Although metameric surfaces do not often occur in nature [239], a crucial implication of metamerism is that with three carefully controlled light sources, the full spectrum (or nearly) can be simulated. Because of that, we are able to build display devices that emit light using three primaries rather than emulating the full spectral distribution, which has shaped the way digital images are stored and managed. Digital imaging relies on this precise process too, which is why digital images are also typically encoded using a triplet of values per pixel. Similar to our visual system, this allows images to be stored using a relatively small amount of data compared to what would be required for storing their full spectral distribution.

10.2 Color as a 3D Space

The primaries used in display devices or digital imaging need not be the same as the responsivities of the cones: tristimulus values for one choice of primaries can be converted to tristimulus values for a different set of primaries, and although they do not represent the same spectral distributions, through metamerism they lead to the same percept. Consequently, the same color can be described in infinitely many ways by changing the set of primaries that it is defined upon. In effect, each tristimulus value corresponds to a point in a three-dimensional space, known as a *color space*, and each axis defines a channel. By shifting, scaling, and rotating the axes defining a color space, a different space can be constructed to achieve different goals.

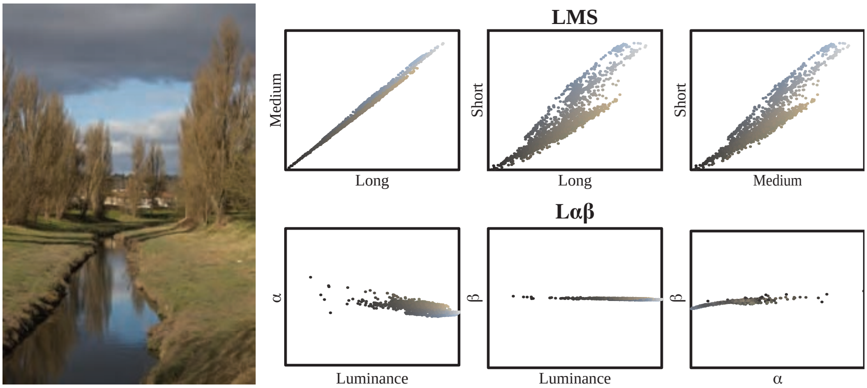


Figure 10.2. For the image shown, corresponding pixel values are plotted for all pairs of channels in the LMS cone space (top) and the $l\alpha\beta$ opponent space (bottom). Values in the LMS space lie along the diagonal, indicating high correlation, while this is much less so the case with $l\alpha\beta$. (Seamills, UK, 2009)

This treatment of color allows for two different ways of studying and using chromatic information in images. A set of pixels in an image can be seen as a set of coordinates within the color space of choice, which can be manipulated as a three-dimensional manifold. Although this would ensure that relations between pixels remain unchanged, it bears no resemblance to the workings of the visual system and would require more complex imaging algorithms.

On the other hand, each channel can be treated as a separate entity, resulting in three one-dimensional distributions. This, of course, is a much simpler problem compared to its 3D counterpart, but it comes with its own set of limitations. If, for instance, we look at the response curves of the L, M, and S cones, it is easy to see that they have a large overlap. This, albeit necessary for the visual system, leads to a highly correlated color space. Effectively, a given value for one channel becomes a good predictor for the values of other channels. This is illustrated in Figure 10.2 (top row) where the pixel values for pairs of channels are plotted for the LMS cone space. Most values sit near the diagonal, indicating that changes in one channel will affect the behavior of the other two as well.

10.3 Opponent Processing

Fortunately, the issue of correlation between color channels can be easily resolved. If an appropriate transformation is chosen, the axes of the color space may be rotated such that data in one channel no longer predict the values in other channels, thus reducing correlation.

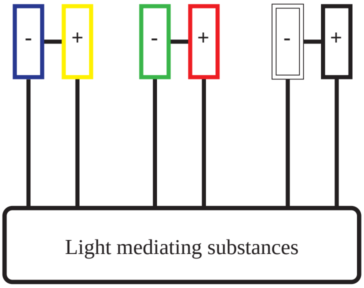


Figure 10.3. The opponent processing scheme proposed by Hering [328] and later verified by Hurvich and Jameson [362].

Further stages of the visual system employ one such solution known as the *opponent processing theory* [328, 362]. It states that as proposed in the trichromatic theory, there are indeed three variables mediating color vision, but in this case, instead of assigning unique color sensations to each of them, pairs of opponent sensory qualities are assigned. These sensory qualities form pairs of red-green and blue-yellow, which are responsible for color information, while a black-white pair allows for luminance perception, illustrated in Figure 10.3. This opponent nature of the visual system leads to sensations that are paired and mutually exclusive, explaining both why color blindness removes pairs of colors as well as why subjective experiences of reddish-green and yellowish-blue are not possible.

An additional implication of opponent processing in the visual system is that it reduces correlation for the input that it typically encounters. The LMS cone responses and in particular the L and M cones (approximately red and green) have a large overlap, meaning that a particular signal is likely to excite both cone types. It has been shown that an eigenvector decomposition of the LMS cone responses leads to a more efficient representation, with less redundancy. In addition, this representation matches the opponency in the visual system, suggesting that dimensionality reduction is its main purpose [88].

To further link the opponent processing stages of the HSV to regularities in natural environments, Ruderman perfomed an interesting statistical experiment [640]. A set of spectral images of natural scenes was captured and converted to log LMS space, effectively simulating the photoreceptor responses to the same scenes. The transformed images were then subjected to Principal Components Analysis (PCA), which was applied on the color data. Each image pixel in this analysis was treated as a point in a 3D color space (log LMS) while spatial information was ignored. The results were striking: finding that the eigenvectors resulting from PCA closely resemble the opponency in the visual system.

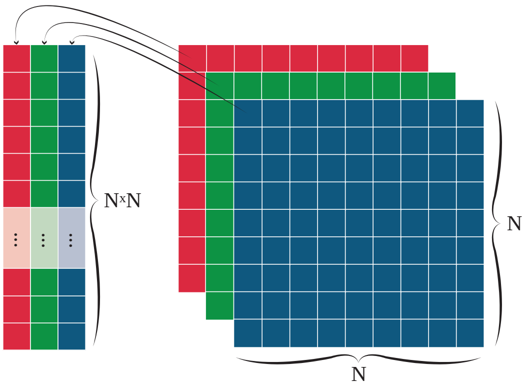


Figure 10.4. Before processing image data with principal component analysis, images are reshaped such that each channel is now a single vector.

The process followed for this experiment can be summarized as follows:

1. The input spectral images are converted to the LMS cone space.
2. A square patch of 128 × 128 pixels is selected from the center of each of the images.
3. Images are logarithmically compressed to ensure that values are more symmetrically spread within their coordinate system.
4. Values in each of the patches are centered around their mean. This operation is applied on each channel separately and it simulates a simplified von Kries adaptation step [804].

$$\mathbf{L} = \log L - < \log L > \tag{10.3a}$$
$$\mathbf{M} = \log M - < \log M > \tag{10.3b}$$
$$\mathbf{S} = \log S - < \log S > \tag{10.3c}$$

5. Each of the three channels of each image patch is reshaped to a single column vector. These vectors are concatenated into a three-column matrix, containing 16,384 LMS triples for each image. This process is repeated for all images in an ensemble, and all subsequent matrices are appended to the end of the initial matrix. See Figure 10.4 for an illustration.
6. Finally, principal component analysis (PCA) is performed on the resulting matrix, rotating the axes so that they are maximally decorrelated, producing a set of three components (eigenvectors) sorted according to the decreasing order of their respective eigenvalues. The mathematical details of that process are given in Section 7.1.

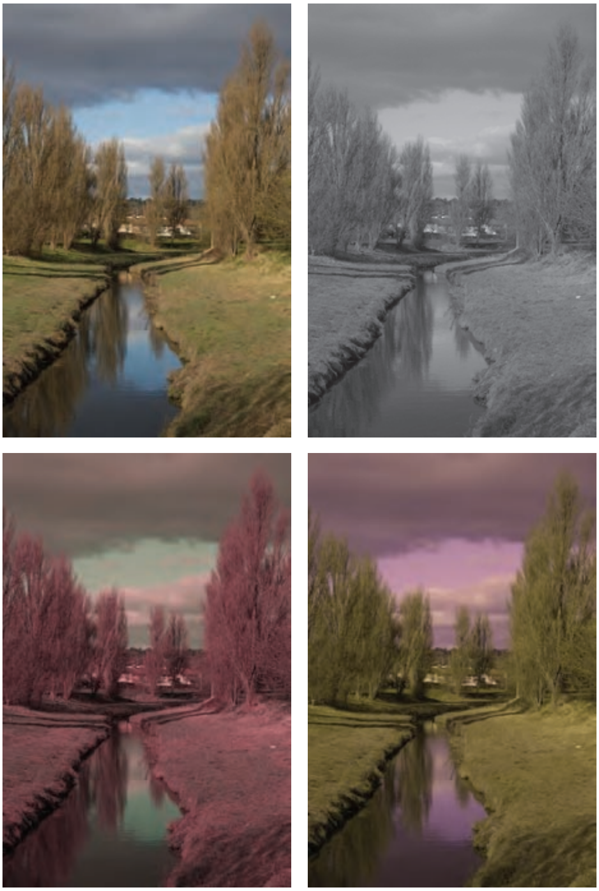


Figure 10.5. The top-left image is decomposed into the l channel of the $l\alpha\beta$ color space, shown in the top-right image, as well as l + α and l + β channels in the bottom-left and bottom-right images. (Seamills, UK, 2009)

By starting in LMS cone space, the rotation yields a new color space, which surprisingly corresponds closely to the color opponency found in further stages of the visual system. Color opponent spaces are characterized by a single achromatic channel, typically encoding luminance or lightness, and two chromatic axes, which can be thought of as spanning red-green and yellow-blue (although the axes do not precisely correspond to these perceptual hues). The chromatic axes can have positive and negative values; a positive value can, for instance, denote the degree of redness, whereas a negative value in the same channel would denote the degree of greenness. A consequence of color opponency is that we are not able to simultaneously perceive an object to have green and red hues. Although we may describe objects as reddish-blue or yellowish-green, we never describe them as reddish-green. The same holds for yellow and blue.

The color opponency of the $l\alpha\beta$ space is demonstrated in Figure 10.5, where the image is decomposed into its separate channels. The image representing the α channel has the β channel reset to zero and vice versa. We have retained the luminance variation here for the purpose of visualization. The image showing the luminance channel only was created by setting both the α and β channels to zero.

The conversion from LMS to the aforementioned decorrelated opponent color space $l\alpha\beta$ is given by:

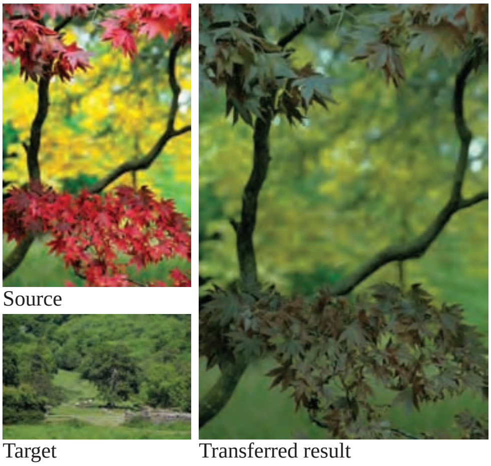
$$\begin{bmatrix} L \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \log L \\ \log M \\ \log S \end{bmatrix} \tag{10.4}$$

while the inverse transform is given by:

$$\begin{bmatrix} \log L \\ \log M \\ \log S \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -2 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 & 0 \\ 0 & \frac{\sqrt{6}}{6} & 0 \\ 0 & 0 & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} L \\ \alpha \\ \beta \end{bmatrix} \tag{10.5}$$

Returning to Figure 10.2, it is readily visible that such a transform is quite successful at reducing correlation in the color data. The bottom row shows plots of corresponding pixel values for all pairs of channels in the $l\alpha\beta$ color space, resulting in point clouds that are much less diagonal compared with their LMS counterparts, shown in the top row. Values in one channel no longer predict the behavior of the other two. In addition, the transform to $l\alpha\beta$ not only decorrelates data, but it effectively makes it independent [640]. These properties have an important implication in imaging: each channel of an image can now be processed separately without affecting the others. One particular application of this observation is discussed in the following section.

It is worth noting here that $l\alpha\beta$ is not the only color space that exhibits such statistical properties. Although the $l\alpha\beta$ space was explicitly constructed through statistical analysis and has strong links with the opponent processing steps in the visual system, other color spaces that are often used in imaging applications have similar properties. For instance, both CIELab and IPT color spaces have a nonlinear compression (albeit not logarithmic), and both encode color data in luminance channel and two opponent channels [190, 610]. The opponency and decorrelation properties of these and several other color spaces will be discussed in Section 10.5.


Source

Target

Transferred result

Figure 10.6. A color transfer example. Here the green colors of the target image (bottom left) are transferred to the source autumn scene (top left) to obtain the result on the right. (Westonbirt Arboretum, Bristol, 2009)

10.4 Color Transfer

Images can convey information not only through the depicted objects but also through the particular mood, color scheme, and composition of the scene. Artists can manipulate the color palette manually to change the appearance of an image and achieve specific effects but that can be a time-consuming process, requiring advanced image manipulation skills. An alternative solution for editing a given image is to find another image with the desired look and transfer properties from that image to our original one. If the property being transferred is color, this process is known as *color transfer*.

This of course is a non-obvious problem: the color in an image could be taken to mean anything from the overall appearance or color palette (warm or cool colors, dark or light mood, and so on) to the full color distribution in a chosen color space. Matters are further complicated since for any chosen definition for describing color within an image, an almost infinite number of ways to transfer it to another exist. Let us consider the images in Figure 10.6, where the color palette of the green landscape is transferred to an autumn foliage scene. The approach taken here makes the whole image green overall but maintains a hint of red on the leaves. Another possible interpretation would be to make the red leaves green but preserve the yellow background and yet another would be to manipulate the source such that it has exactly the same color distribution as the target.

The lack of a strict definition of what constitutes a successful transfer has led to an expansive collection of techniques, each approaching the problem in a slightly different way and focusing on slightly different application areas. Many of them though share a particular characteristic, which makes them interesting for us in the context of color statistics. Specifically, many of the existing color transfer techniques rely on the decorrelation afforded by opponent spaces such as the $l\alpha\beta$ space, discussed earlier, or CIELab to transfer properties between single channels at a time, rather than having to consider the full 3D color distribution. We will look at two examples of color transfer methods to demonstrate the flexibility that this fortunate regularity of the color content of images allows.

10.4.1 Color Transfer through Simple Moments

One of the earliest color transfer techniques takes advantage of the decorrelation property of $l\alpha\beta$ and shifts and scales the pixel values of the source image to match the mean and standard deviation from the target [607]. Each channel of the image is processed separately in $l\alpha\beta$, as it allows the transfer to take place independently in each channel, turning a potentially complex 3D problem into three much simpler 1D problems.

Input images are assumed to be given in the sRGB space and are first converted to $l\alpha\beta$. The colors of the source image can now be altered to approximate those of the target. This is achieved by shifting and scaling the source distribution such that it is aligned with that of the target. Figure 10.7 shows an example pair of source and target images as well as their corresponding histograms before and after transferring the colors between them.

To achieve this, the values of each channel of the source image are shifted to a zero mean by subtracting the mean of the source from each pixel. They are then scaled by the standard deviations of both images such that they acquire that of the target image, and finally, they are shifted to the mean of the target by adding its mean instead of the source mean that was originally subtracted:

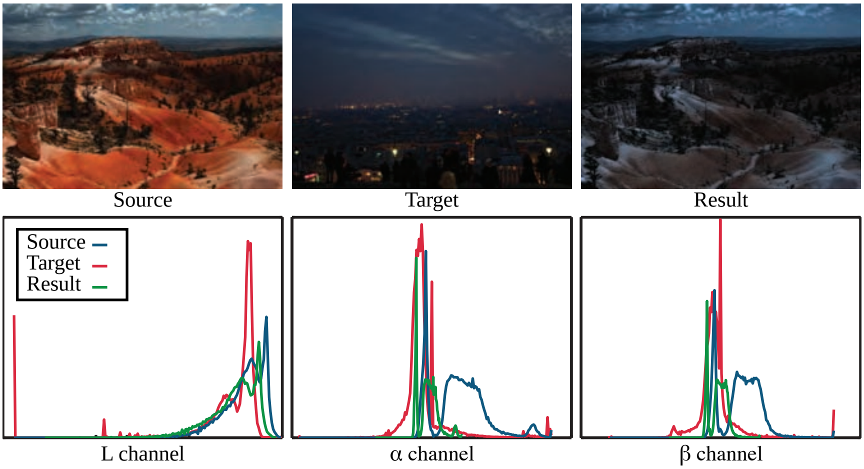
$$l \;= l_s - \mu_s \tag{10.6a}$$

$$l \;= \frac{\sigma_{t,l}}{\sigma_{s,l}} l \tag{10.6b}$$

$$l_o = l \;+ \mu_t \tag{10.6c}$$

Here, the subscripts s, t, and o correspond to the source, target, and output images and μ and σ are their respective means and standard deviations. Equations (10.6a)–(10.6c) describe the distribution transfer between source and target for the luminance channel only. This process effectively ensures that the distributions of the luminance values for the two images have the same first and second moment. The same process is repeated for the two chromatic channels of the $l\alpha\beta$ space to complete the color transfer. The images can then be converted back to their original display space.

Although this technique can be successful for a wide range of images, the quality of the results largely depends on the composition of the source and target


Source Target Result

L channel α channel β channel

Figure 10.7. An example pair of source and target images as well as their resulting output after using the color transfer technique by Reinhard et al. [607] are shown on the top row. The corresponding histograms are shown on the bottom row for the three channels of the $l\alpha\beta$ color space. (Left: Bryce Canyon, Utah, 2009; right: Paris, France, 2008)

images. Swatches may be used to supplement this technique and allow more control over color correspondences if the two images are too different. The user can in this case select pairs of swatches from the source and target images to indicate corresponding regions between them. The pixels in these swatches form clusters in the $l\alpha\beta$ space, allowing statistics to be computed separately for each of the swatches. Source pixels are transformed according to the statistics of each pair of swatches, leading to a number of possible renditions of the source image. To construct the final image, these different renditions are blended by considering the distance of each pixel's color to the center of each cluster. Distances are divided by the standard deviation of each cluster to account for different cluster sizes, and pixels are blended using inversely proportional weights to the normalized distances.

Despite the increased control that these swatches allow, in both the local and global versions of this technique the transfer of colors between the two images relies on first-order statistics, namely the mean and standard deviation of each channel. Such statistics can provide useful information about overall tendencies in a distribution but, as argued in Chapter 4, more information and ideally higher-order analysis are necessary to capture more subtle variations in images.

An obvious next step is to use higher statistical moments in addition to the mean and standard deviation to achieve a better result. For instance, the skew and kurtosis of the distribution can be transferred to reshape the histogram of the

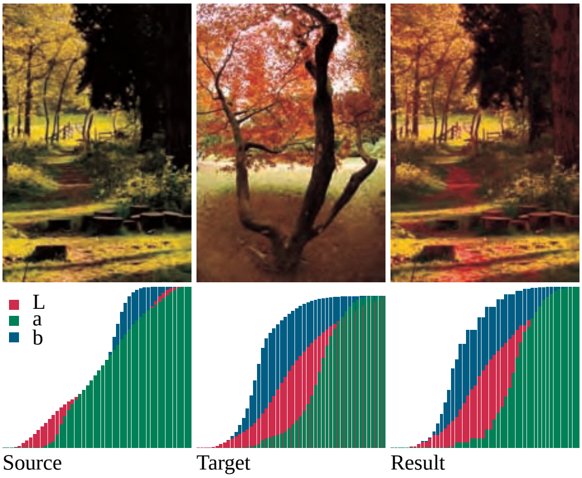


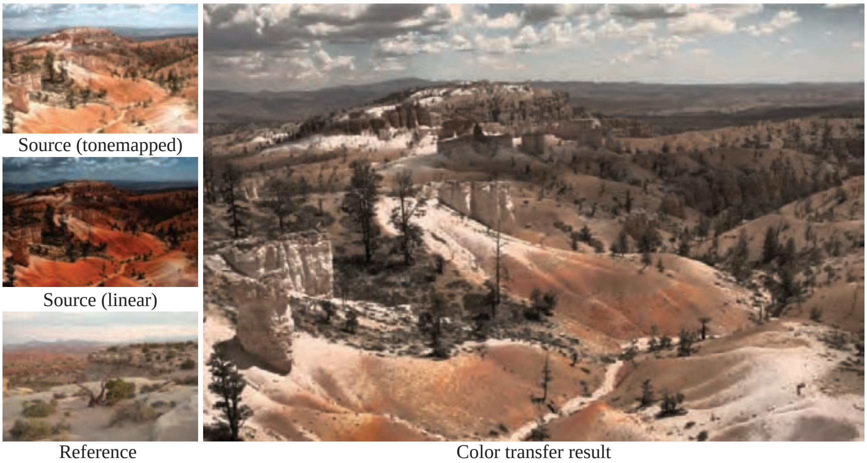
Figure 10.8. Histogram matching between two color images. (Westonbirt Arboretum, Bristol, 2009)

source image to be closer to that of the target. Although such a process is conceptually elegant, unfortunately it is not as effective in practice. The mean and standard deviation of a distribution can easily be transferred analytically (Equations (10.6a), (10.6b), and (10.6c)), while higher moments such as the skew and kurtosis require more complex solutions. Although a distribution can be uniquely described by its moments, arbitrary moments cannot be changed on demand, thus requiring optimization-based solutions. Even so, the improvements afforded by transferring further moments in this way are very minor.

10.4.2 Color Transfer through Higher-Order Manipulation

In many cases a small set of statistical moments is not sufficient to adequately describe subtleties of the color distribution of images. A much more faithful representation can be achieved by considering the full distribution of values in each channel. In Chapter 4, we showed how a histogram distribution can be fully matched to another one. This process can be easily extended to a color image by repeating the histogram matching process for each of the three channels to achieve a result as shown in Figure 10.8.

This process ensures that the transformed image will have exactly the same distribution as the selected target and therefore the same colors. Occasionally, however, the resulting image may be too harsh as the transfer can amplify artifacts that were previously invisible, indicating that higher-order properties of the image may need to be matched or preserved to achieve a successful result. Based on this observation, a wealth of color transfer methods have been developed, each relying



Source (tonemapped)

Source (linear)

Reference

Color transfer result

Figure 10.9. An example result obtained with the histogram scale-space approach discussed here [587]. The input image (high dynamic range original—shown both tonemapped and linearly normalized) is transformed to match the color palette of the given reference. (Utah, USA, 2009)

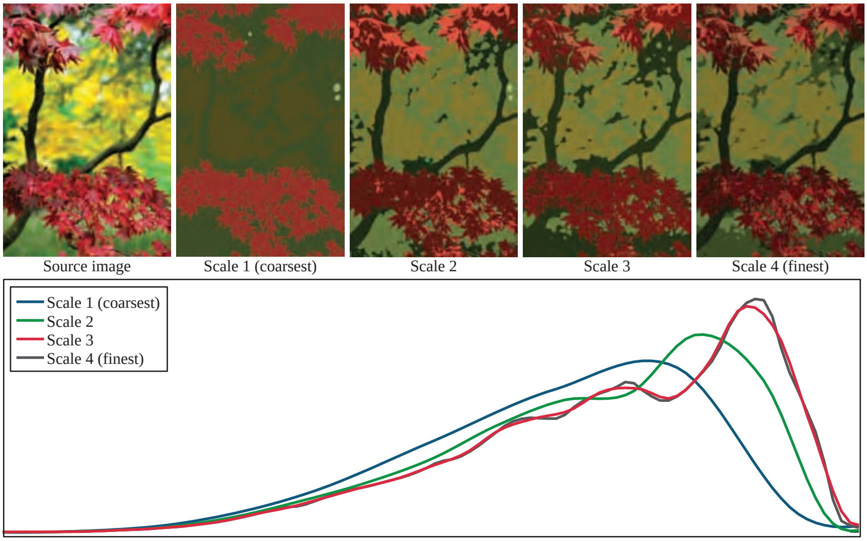
on a different set of statistical properties of the images to transfer the color palette between them without otherwise affecting the appearance of the image.

10.4.3 Histogram Features at Different Scales

Since a full histogram match is likely to push the image appearance too far, one possibility is to only partially match the two histograms by taking advantage of their local structure [587]. One recent solution achieves that by considering features of the histograms in different scales. An example result from this method is shown in Figure 10.9.

Consider, for instance, the image in Figure 10.10; a pyramid of increasingly coarse scales can be constructed by filtering the original histogram with a Gaussian kernel of increasing size. Similar to image scale-space filtering, coarser levels preserve only large scale features of the histogram, which represent larger segments of the image. On the other hand, finer scales contain more details but each of the histogram features at those scales will correspond to smaller regions in the image.

The color palette between the two images is transferred by reshaping the source histogram so that it approaches that of the target. Rather than aim for an accurate match, though, where every subtlety of the target is matched, the source can be progressively transformed to the target by matching more and more scales, starting from the coarsest. When all scales including the finest are matched, then the result will be identical to the simpler histogram matching discussed earlier.


Source image
Scale 1 (coarsest)
Scale 2
Scale 3
Scale 4 (finest)

Histograms for different scales

Figure 10.10. The histogram of the top-left image, shown at the bottom, is progressively filtered to remove detail. The coarser features of the histogram correspond to large coherent regions of the image with similar colors. When examining the finer scales of the histogram, though, peaks in the histogram only contain a small number of pixels and therefore correspond to small image segments. (Westonbirt Arboretum, Bristol, 2009)

This leads to an interesting observation: such an approach is possible because nearby pixels in images tend to be similar. This is a recurring property of natural scenes, which we have discussed in the context of simple luminance statistics, edges, and higher-order properties throughout this book (e.g., see Chapters 5 or 8). In this case, it means that by treating each feature of the filtered histograms separately, we are inherently treating the image as a collection of segments, without ever having to explicitly segment it.

Although this approach is successful for many images, some transfers require relatively flat areas to acquire much more extreme color transitions, which may amplify noise or compression artifacts or simply change the appearance of the scene in terms of contrast. As discussed in Chapter 5, edges are one of the main streams of information within visual data, and consequently, in an application such as color transfer it would be desired to preserve them. This, however, is not entirely straightforward as edge information is intricately linked to the pixel values in the image and therefore to color as well. Whether edges are recovered by adding local contrast back to the image after the transfer has taken place [587, 588] or through an optimization scheme that aims to preserve the gradient distribution of the source image [807], such methods present a tradeoff between accurate matching of the color distribution and preservation of edge content.

10.4.4 Color Transfer as a 3D Problem

Referring back to Section 10.2, we have seen how color spaces describe a three-dimensional space. Pixels in an image are in that case given as a triplet, denoting a point within that space. In all our discussion on color transfer so far, we have focused only on techniques that manipulate each of the image channels independently. Although this description of color provides the obvious advantage of simplifying a potentially complex 3D problem to a set of three 1D problems, it cannot capture all subtleties in the color distribution of images.

To maintain local color information and interrelations, the 3D color distribution of the two images needs to be treated as a whole; this is done so that the source 3D distribution will be reshaped to match or approximate that of the target. Unfortunately, translating processes, such as histogram matching or histogram reshaping, to more than one dimension is not straightforward and either requires an optimization-based solution or a way to simplify the problem to fewer dimensions.

In the latter case, a general approach that iteratively matches the two distributions through 1D projections may be employed, proposed by Pitié et al. [572, 573]. Specifically, at each iteration step, the 3D distributions of the source and target images are rotated using a random 3D rotation matrix and projected to the axes of their new coordinate system. Each 1D projection of the source is then matched to that of the target and the data is transformed back to its original coordinate system. This process is repeated with different rotations until convergence. An example result and the effect of an increasing number of iterations can be seen in Figure 10.11.

Whether the color distribution is matched by manipulating each of the color channels separately or by treating the image as a 3D color distribution, a few important observations arise. Although all the techniques discussed have at a high level the same overall goal, they employ widely different approaches to get there, offering a unique view into color processing.

Another issue of note is the lack of consensus on the choice of color space to be used for such processing. The simple moment transfer technique discussed earlier [607] as well as many other color transfer techniques [281, 744, 812, 461, 808, 776, 474, 820, 807, 805] operate in Ruderman's $l\alpha\beta$ space [640], while the histogram reshaping process uses *CIELab* [588].

On the other hand, the iterative projection of 1D marginals begins in RGB and selects a random coordinate system in each iteration [573], effectively transforming the image to a randomly selected color space at each step, while a similar dimensionality reduction process has been demonstrated in the *LCh* space [534].

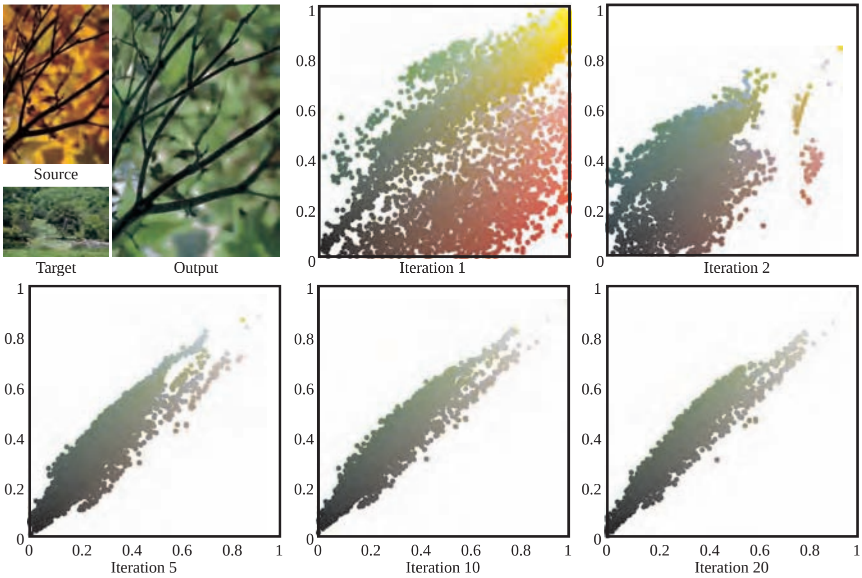


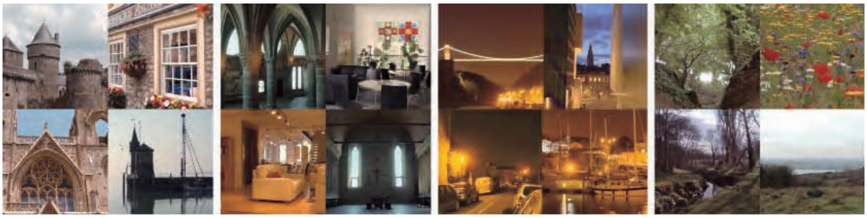
Figure 10.11. The colors of the source image are progressively mapped to those of the target by iteratively rotating the 3D color distributions of both images to a randomly chosen coordinate system and matching the 1D marginal distributions for each of the axes.

Other color spaces implicated in color transfer are YC_bC_r [430] and Yuv [454, 779], as well as color appearance models [514]. Finally, several authors suggest to compute a dedicated color space for each transfer, based on PCA [419, 1, 418, 806, 2]. The following section will attempt to shed some light on the issue of color space choice.

10.5 Color Space Statistics

Any RGB-like color space tends to be significantly correlated. This means that a high value in, say, the red channel suggests a good probability of finding high values in the green and blue channels for those pixels. In essence, this is a different way of saying that colors in natural scenes tend to be desaturated, or that images on average tend towards gray. This feature of natural images is explicitly exploited in photography, where the gray-world assumption is made to infer the color of the illumination in a scene, which is discussed in more detail in the following section.

The high correlation of RGB-like color spaces also means that for imaging applications such as color transfer, as we saw in the previous section, a different color space may be necessary to ensure that changes in one channel do not



Manmade Day (MD) Manmade Indoors (MI) Manmade Night (MN) Natural Day (ND)

Figure 10.12. Sample images from the four image categories used for the color space analysis and the computation of PCA-based spaces.

inadvertently affect the other channels and therefore lead to unpredictable results. A decorrelated space, such as the $l\alpha\beta$ space proposed by Ruderman et al. [640] allows images to be transformed to a coordinate space where values in one of their channels predict those in other channels as little as possible. Decorrelated color spaces are frequently used in image editing applications for this precise reason.

This is of course a desirable property as it reduces the dimensionality of the problem at hand as we have seen, but such spaces have been found to not fully decorrelate all images. The $l\alpha\beta$ space, for instance, is designed around the principal axes arising from the decorrelation of natural scenes. Consequently, converting an image to that space is equivalent to rotating its data so that it is aligned to the axes defining the $l\alpha\beta$ space. If the content of this image is significantly different to the types of images used for the initial derivation of that space, these axes may or may not correctly describe most of the variation in the new data. It is not clear, however, whether different scene categories or images possessing a higher dynamic range would lead to a different color space or how such differences might impact applications such as color transfer.

To answer this question, a study was done comparing various color spaces in terms of their decorrelation properties [611]. Images from different scene categories were analyzed to assess whether color spaces such as $l\alpha\beta$ are in fact better at decorrelating particular scene types. Each of the image categories was also used to construct a custom PCA-based space following a similar process to that the $l\alpha\beta$ space discussed in Section 10.1, based on the hypothesis that a color space constructed through decorrelation of a particular type of image would be better at reducing correlation when processing similar images. The study focused on both manmade and natural imagery, specifically categorizing images as daytime manmade scenes denoted by MD, indoors manmade scenes (MI), nighttime manmade scenes (MN), and daytime natural scenes (ND). Figure 10.12 shows a few sample images from each of the categories used.

To compare both existing and the per-scene type color spaces, correlation measures were employed as well as a short experiment where pairs of images were processed using a linear color transfer method [607] and the color trans-

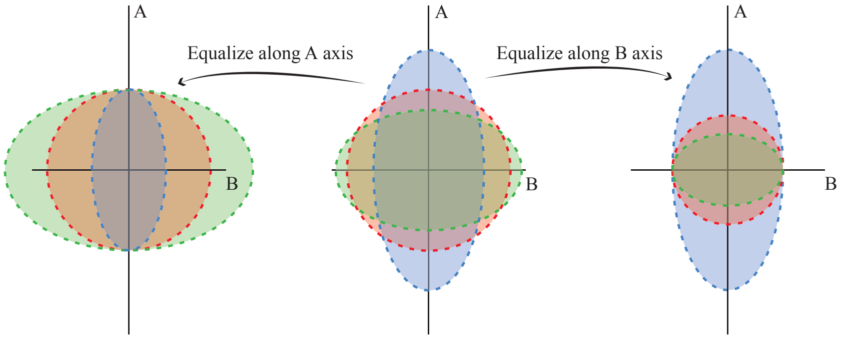


Figure 10.13. Since different color spaces are defined along different ranges, normalizing along a single axis is unlikely to correctly allow for comparisons between the different spaces.

fer result was ranked as successful or not. In addition to $l\alpha\beta$ and the PCA-based spaces discussed, the following color spaces were also considered: CIELab (using both illuminants D65 and E), Yxy, *Yuv*, *Yu v* , XYZ, RGB, and HSV [610].

10.5.1 Color Space Normalization

Different color spaces are defined for different ranges. For instance, an RGB image may be defined between 0 and 1, while CIELab is not explicitly constrained. To ensure that results from different color spaces are comparable in such an analysis, data needs to be normalized. This, however, is not straightforward since the meaning of different channels, as well as the relations between them, vary with color space. Consequently, simply normalizing along each of the channels will lead to incorrect results, as illustrated in Figure 10.13.

To resolve this issue, data can be normalized according to the volume of each color space, which is computed according to the range that each channel or axis is expected to acquire. To approximate the volume of each of the color spaces, the maximum and minimum achievable values for each channel are determined by using input RGB values constrained between 0 and 1. The computed volumes used for the normalization step are shown in Table 10.1.

Color Space	Volume	Color Space	Volume
$l\alpha\beta$	6.4480	*Yu v*	0.1318
CIELab **(E)**	6.0447e+03	*HSV*	0.8333
CIELab **(D65)**	8.8841e+03	*XYZ*	1.0351
Yxy	0.2646	*RGB*	1.0000
Yuv	0.0879		

Table 10.1. The volumes of different color spaces, approximated by a cuboid.

10.5.2 Correlation Analysis

In Chapter 4, we saw how statistical moments can be computed for distributions. The second moment corresponds to the variance of a dataset I of size N and is computed as follows:

$$\sigma^2(I) = \sum_{p=1}^{N} \frac{(I(p) - c)^2}{N - 1} \tag{10.7}$$

where c is a constant around which the data is considered. Typically this can be taken as 0, if statistics of the data are computed with respect to the origin, or more frequently, c is set to the mean of the data μ.

The variance of a dataset provides information about the spread of the values within that set. To measure how more than one set varies with respect to each other, a more general formulation of this measure is necessary, known as *covariance*.

$$\text{Cov}(I, J) = \sum_{p=1}^{N} \frac{(I(p) - \mu_I)(J(p) - \mu_J)}{N - 1} \tag{10.8}$$

For each color space, covariance matrices were computed corresponding to the four datasets shown in Figure 10.12. Values along the diagonals of the covariance matrices contain the variance within each channel, while elements off the diagonal capture covariance values for pairs of channels:

$$\mathbf{Cov}_s = \begin{bmatrix} \text{Cov}(s_1, s_1) & \text{Cov}(s_1, s_2) & \text{Cov}(s_1, s_3) \\ \text{Cov}(s_2, s_1) & \text{Cov}(s_2, s_2) & \text{Cov}(s_2, s_3) \\ \text{Cov}(s_3, s_1) & \text{Cov}(s_3, s_2) & \text{Cov}(s_3, s_3) \end{bmatrix} \tag{10.9}$$

where $s = \ l\alpha\beta$, CIELab (E), CIELab (D65), Yuv, Yu'v', HSV, XYZ, RGB and s_i refers to the i^{th} channel of that space, with $i = \ 1,2,3$. The covariance values for each channel pair were then averaged to a single number per image set, representing the ability of each color space to decorrelate different classes. These results are summarized in Figure 10.14.

As would be expected, opponent spaces lead to the lowest covariance values, while the RGB and XYZ spaces are much more correlated. Interestingly, the choice of white point used for the conversion to the *CIELab* color space was found to drastically matter when assessing color spaces with the covariance measure. When the achromatic illuminant E was used, *CIELab* resulted to much lower covariance values compared to using the bluish D65 illuminant.

In addition to standard color spaces, ensemble-specific spaces similar to $l\alpha\beta$ were constructed following the procedure detailed in Section 10.3 for each of the image ensembles shown in Figure 10.12. Since these color spaces are computed from specific datasets representing different images categories, one would expect that they would lead to minimal covariance for the categories they were derived from.

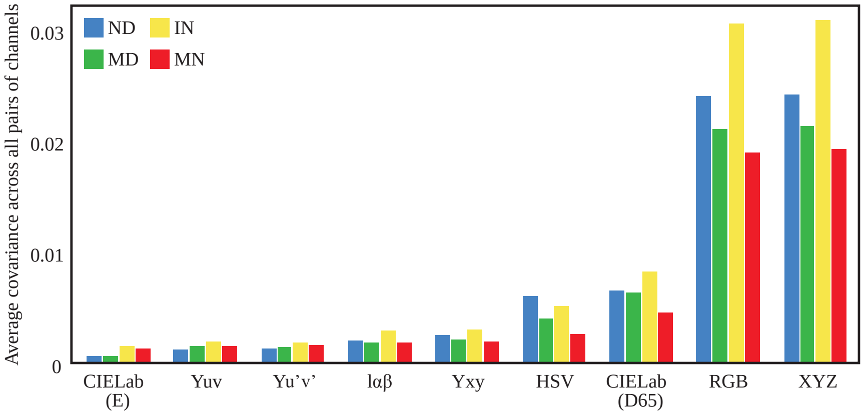


Figure 10.14. Average covariance values for the color spaces tested for each image ensemble. Higher values suggest higher correlation and therefore indicate spaces less suitable for edits where each channel is processed separately.

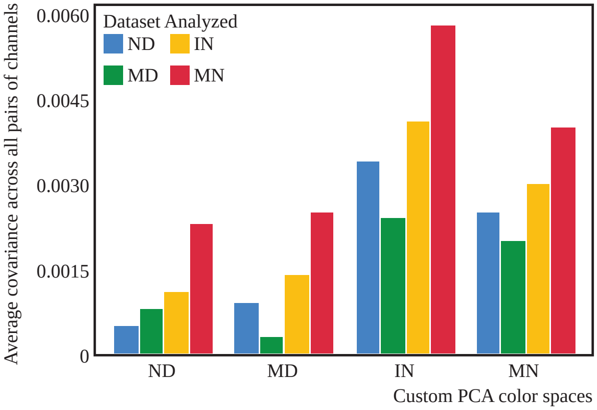


Figure 10.15. Average covariance for the four PCA-based color spaces computed from different image ensembles. Images from each of the ensembles were transformed with all spaces to assess whether better decorrelation can be achieved when the color space is computed from similar images.

Figure 10.15 summarizes the results for the covariance analysis of these ensemble-specific spaces. Although all four spaces lead to lower covariance overall compared to standard color spaces, only for the "natural day" (ND) and "manmade day" (MD) sets does the corresponding color space lead to better performance. On the other hand, the PCA-based space constructed from the "manmade night" (MN) set led to the worst performance overall. One hypothesis explaining these results is that more varied image sets offer a better sampling of the color gamut, therefore leading to a more descriptive decomposition.

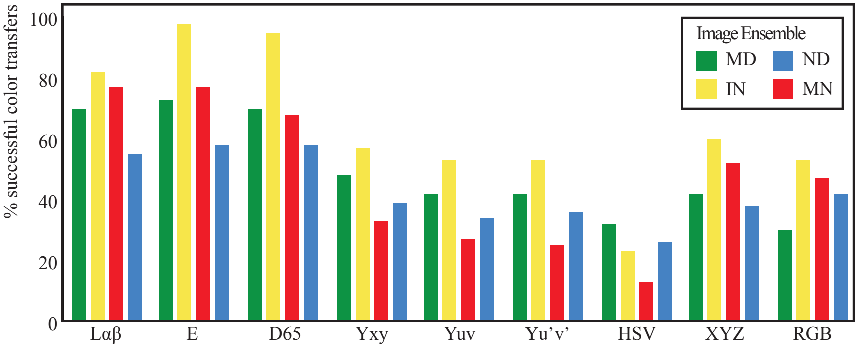


Figure 10.16. Percentage of successful color transfers for each color space for each of the ensembles. E and D65 refer to CIELab with illuminants E and D65, respectively.

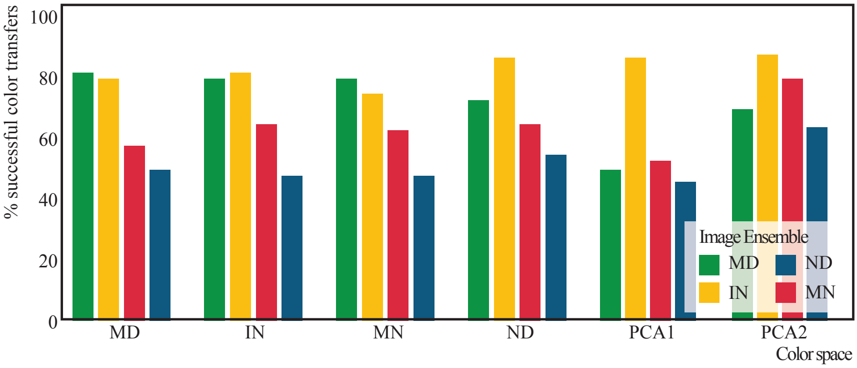


Figure 10.17. Percentage of successful color transfers for each ensemble-specific color space for each of the ensembles. Results are also shown for per-image PCA, carried out in the color space derived from the source image (PCA1) or the reference image (PCA2).

To confirm whether the findings from the statistical analysis correspond with viewers' preferences, a short study was also conducted. Pairs of images from the four ensembles were transformed to one of the color spaces discussed and then used as input for the color transfer algorithm presented in Section 10.4.1. In addition to the standard and PCA-based color spaces discussed in the previous section, PCA spaces were this time also computed from each of the individual input images. Participants were then shown the pair of input images as well as the color transferred result and asked to classify it as successful or not.

Interestingly, the color space rankings from this study, shown in Figures 10.16 and 10.17 for the standard and PCA spaces, respectively, were very similar to the results of the previous analysis. The CIELab space led to the most successful color transfer, followed by the PCA-based space computed from the reference image for each transfer. Unlike the covariance analysis results, though, the choice of white point for the color space conversion seemed to matter less.

In summary, although one may argue that decorrelating individual images would be the most fine-grained approach to color transfer, it was found that similarly excellent results can be obtained by simply choosing the CIELab space. Thus, it does not appear to be necessary to expend computational resources to deriving either category-based color spaces or image-specific color spaces for the purpose of color transfer.

10.6 Color Constancy and White Balancing

The color of the light reaching our retinas is determined by the spectral distribution of the illumination in the scene and the reflectance properties of the surfaces it encounters in its path to our eyes. An intriguing property of our visual system is its ability to correctly assess the color of objects under varying illumination. A banana will be perceived as yellow whether it is seen under the orange cast of tungsten light or outside in sunlight, even though the precise properties of the light reflected from its surface will be very different in these two situations. In other words, the visual system is able to separate the effect of surface reflectance and the prevalent illumination in a scene, effectively discounting the color of the light source when assessing the color of an object.

This property is known as *color constancy* and was first experimentally demonstrated by Edwin H. Land in his “Color Mondrian” experiment [435, 238]. A Mondrian-like display was constructed out of colored pieces of paper, which were illuminated by three narrow-spectrum red, green, and blue projectors that could be individually controlled, as illustrated in Figure 10.18. In a typical experiment, the tristimulus values of light reflected from a patch of a particular color, such as the green patch (a) in the figure, were first measured and the projected light was then adjusted so that a different patch, such as the orange patch (b), reflected the same tristimulus values. Although in both cases the tristimulus values of the light reaching the retina were identical, viewers would consistently see the green patch as green and the orange one as orange when viewed as part of the complete display.

Unlike the visual system, though, cameras simply capture light exactly as it reaches the sensor. If a scene has a color cast due to the illumination, this cast will remain visible and noticeable on the image since the conditions where the image was captured are unlikely to match the display and viewing environment where it is then shown. Consequently, the ability to discount or remove the prevalent illumination from images would be a very desirable one for cameras. This

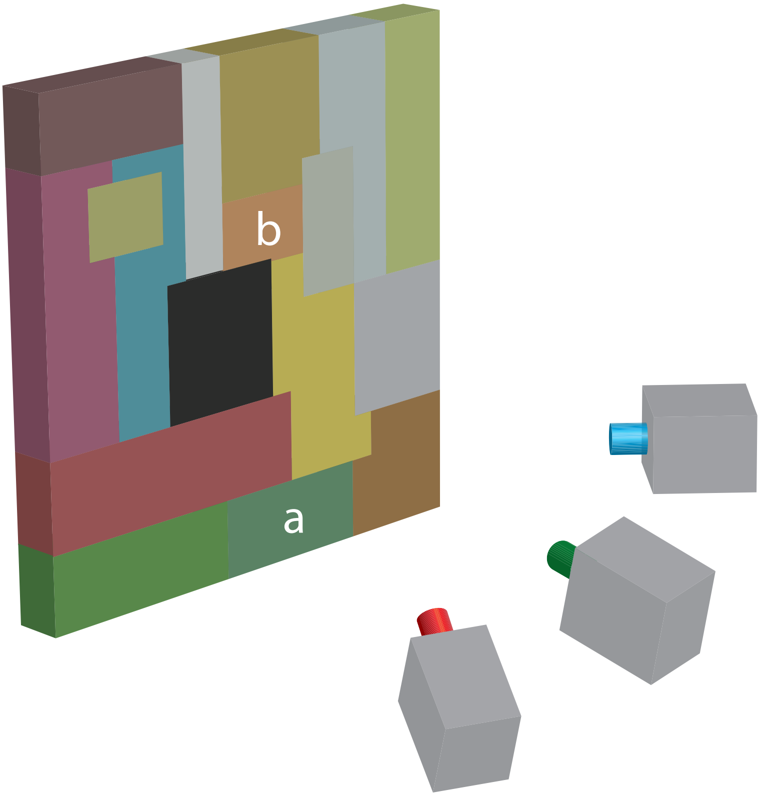


Figure 10.18. An illustration of the experimental setup used to demonstrate color constancy.

is typically implemented in a process known as *white-balancing*, which aims to simulate the color constancy properties of the visual system [191, 268, 238].

In color spaces such as the aforementioned LMS space, equal values of the three components denote achromatic colors. One way to achieve such neutral colors is to start with an equal energy spectrum, i.e., a spectral distribution which has the same value L_o for each wavelength λ. This could happen if a scene was illuminated by an equal-energy light source, a source that emitted the same energy at all wavelengths. The colors that a camera will capture, then, from such a scene can be attributed to the reflectance properties of the objects and surfaces in it.

In practice, a scene is illuminated by only one or at most a few light sources, with an emission spectrum that is off-white. In natural scenes, for instance, the main source of illumination is likely to be the sun, with evidence suggesting that

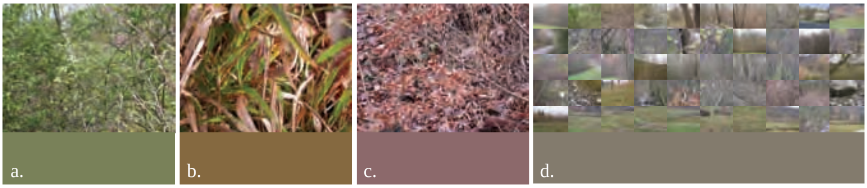


Figure 10.19. The white points shown in the bottom row were estimated for three individual images, as well as for a collection of 128 images using the gray-world assumption. Although when a large set of images is averaged, the result will be approximately gray, this is not the case with individual images as the dominant colors of the scene are likely to bias the estimation (images taken from the Zurich Natural Images database; see Chapter 3 for more details).

this may even be encoded as a prior in human color constancy [165]. The reflectance properties in a scene, on the other hand, are much more variable, suggesting that the local variations in color in an image are likely due to different materials and reflectance properties. Despite this variation, a surprising finding is that when a collection of reflectances are averaged, the result ends up being a distribution function that is close to gray. This is known as the *gray-world assumption* [87].

The implication of this finding is that if we were to average the colors of all pixels in an image, and the gray-world assumption holds, the average tristimulus value $(\bar{L}, \bar{M}, \bar{S})$ would be a good indicator of the color of the light source. Unfortunately, the gray-world assumption does not always hold. Figure 10.19a–c shows the result of averaging pixel values over single natural image, while 10.19d shows the average pixel value over a hundred natural images. As can be seen, in single images this assumption may or may not hold. In particular, if the surface colors in the scene are biased towards some specific saturated color, the average reflectance will not tend toward gray, and therefore extracted information about the illumination will likely also be biased.

The gray-world assumption is often used to aid white-balancing in imaging applications. After all, if we know the color of the illuminant, then we can correct all pixels by simply dividing all pixels by the average image value. Moreover, if we know that the display has a different white point, say $(L_{d,w}, M_{d,w}, S_{d,w})$, white-balancing can be implemented as follows:

$$L_{\mathrm{wb}} = L\,\frac{L_{d,w}}{\bar{L}} \tag{10.10a}$$

$$M_{\mathrm{wb}} = M\,\frac{M_{d,w}}{\bar{M}} \tag{10.10b}$$

$$S_{\mathrm{wb}} = S\,\frac{S_{d,w}}{\bar{S}} \tag{10.10c}$$

The gray-world assumption is a statistical argument that is necessary to perform white-balancing on images in the absence of further information about the illuminant, given that white-balancing is by itself an underconstrained problem [246]. Note that this procedure is best applied in a perceptual color space, such as LMS, thereby mimicking chromatic adaptation processes that occur in the human visual system [610]. If an image is given in a different color space, most likely the RGB space, the image should first be converted to LMS.

The approximation of the illuminant can be improved by excluding the most saturated pixels from the estimation [4]. Alternatively, the image can be subjected to further statistical analysis to determine if the color distribution is due to colored surfaces or a colored light source [251]. Here the image is first converted to the CIELab color opponent space. Ignoring the lightest and darkest pixels, since they do not contribute to a reliable estimate, the remaining pixels are used to computed a two-dimensional histogram $F(a,b)$ on the two chromatic channels a and b. In each channel, the chromatic distributions are modeled with:

$$\mu_k = \int_k k F(a,b)\, dk \tag{10.11}$$

$$\sigma_k^2 = \int_k (\mu_k - k)\ F(a,b)\, dk \tag{10.12}$$

where $k = a,b$. These are the mean and variances of the histogram projections onto the a and b axes. In CIELab neutral colors lie around the $(a,b) = (0,0)$ point. To assess how far the histogram lies from this point, the distance D can be computed:

$$D = \mu - \sigma \tag{10.13}$$

where $\mu = \sqrt{\mu_a^2 + \mu_b^2}$ and $\sigma = \sqrt{\sigma_a^2 + \sigma_b^2}$.

The measure $D_\sigma = D/\sigma$ can be used to assess the strength of the cast. If the spread of the histogram is small and lies far away from the origin, the image is likely to be dominated by strong reflectances rather than illumination.

10.6.1 Computational Color Constancy as the Minkowski Norm

It was found that while gray-world algorithms work well on some images, alternate solutions, such as the *white-patch algorithm* [435], perform better on texture-rich images. The white-patch algorithm assumes that the lightest pixels in an image depict a surface with neutral reflectance, so that its color represents the illuminant.

Both the gray-world and white-patch algorithms are special instances of the Minkowski norm [219]:

$$L_p = \left(\frac{\int f^p(x)dx}{\int dx} \right)^{1/p} = ke \tag{10.14}$$

Algorithm	Symbol	Description
Gray-World	$e^{0,1,0}$	The average reflectance in a scene is assumed to be achromatic; therefore, any shifts from an achromatic average are due to illumination [87].
White-Patch (Max-RGB)	$e^{0,\infty,0}$	The brightest patch in the scene is assumed to belong to a surface with neutral reflectance and therefore the color of that patch represents the color of the illumination.
Shades of Gray	$e^{0,p,0}$	A more general formulation of the gray-world and max-RGB algorithms that assumes that the p^{th} norm of the scene is achromatic. Shown to perform best when $p = 6$ [219].
General Gray-World	$e^{0,p,\sigma}$	A more general version of the above, where a local region of scale σ is considered, which can be achieved by first filtering the image.
Gray-Edge	$e^{1,p,\sigma}$	The average of the derivative of the image is assumed to be achromatic [789].
Max-Edge	$e^{1,\infty,\sigma}$	The maximum difference of reflectances in the scene is assumed to be achromatic.
Second-Order Gray-Edge	$e^{2,p,\sigma}$	The average of higher-order derivatives (second in this case) is assumed to be achromatic [790].

Table 10.2. Different formulations based on the Minkowski norm can be used to obtain different color constancy algorithms (adapted from [790]).

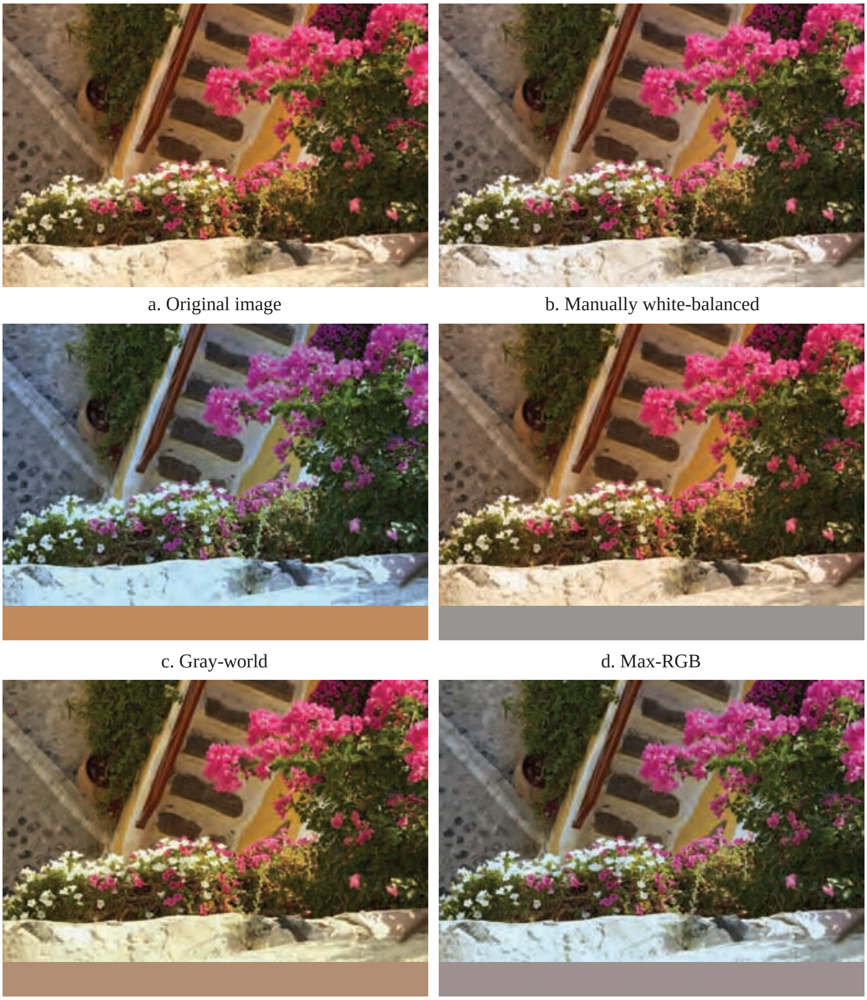
where $f(x)$ denotes the image at pixel x. The average of the image is computed for $p = 1$, thereby implementing the gray-world assumption. The maximum value of the image is computed by substituting $p = \infty$, which represents the white-patch algorithm, also known as Max-RGB.

A further generalized assumption can be made about images, which is that the average difference between two pixels evaluates to gray. This is known as the gray-edge assumption [789, 790] and can be formulated as follows [266]:

$$\left(\int \left| \frac{\partial^n f^\sigma(x)}{\partial x^n} \right|^p dx \right)^{1/p} = ke^{n,p,\sigma} \tag{10.15}$$

Here, n is the order of the derivative, p is the Minkowski-norm, and $f^\sigma(x) = f \otimes G^\sigma$ is the Gaussian-blurred image where the size of the filter kernel is given by σ. With this formulation several color constancy algorithms can be constructed, shown in Table 10.2. Examples of some of these methods can be seen in Figure 10.20.

Although different algorithms rely on different and increasingly complex statistics to correctly white-balance images, no single method has yet emerged as a universally effective solution [268]. This suggests that a single aspect of images may not be sufficient to determine the color of the illumination, especially in ambiguous scenarios. An overall white scene illuminated with a blue light may look substantially similar to a blue scene illuminated with a white light. Yet



a. Original image

b. Manually white-balanced

c. Gray-world

d. Max-RGB

e. Shades of gray

f. Gray-edge

Figure 10.20. Different color constancy algorithms based on the Minkowski norm were used to white-balance the image shown at the top left (a). The algorithms used are given in Table 10.2. The small rectangle below each image shows the white point estimated by each algorithm and the manually white-balanced image is given in (b). (Santorini, Greece, 2009)

most color constancy algorithms will find it hard to distinguish between them. On the other hand, our visual system is capable of such a distinction in most cases, perhaps by relying on a variety of higher-order statistical information in scenes [272, 271, 280].

10.6.2 White-Balance Algorithm Selection

Having noted that many color constancy and white-balancing algorithms exist, with none of them universally applicable, Gijsenij and Gevers use natural image statistics to classify an image, and then they use this classification to select the most appropriate white-balancing algorithm for the image at hand [266, 267]. In particular, they use the finding that the distribution of edge responses in an image can be modeled by means of a Weibull distribution [260]:

$$f(x) = \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} \exp\left(\frac{x}{\beta}\right)^{\gamma} \tag{10.16}$$

The parameters β and γ have meaning in this context. The contrast of an image (or image ensemble) is given by β, whereas γ is an indicator of grain size (i.e., the peakedness of the distribution). This means that higher values for β represent images with more contrast. Higher values for γ indicate finer textures.

Fitting a Weibull distribution involves computing a Gaussian derivative filter in both x and y directions. This results in the $(\beta_x, \beta_y, \gamma_x, \gamma_y)$ set of parameters for each color channel. The Gaussian derivative filter can be first-, second-, or third-order. However, it was found that the order of the chosen filter is relatively unimportant: high correlations exist between the fitted parameters [260].

It is now possible to fit Weibull parameters to the derivatives of a large number of images and correlate their values with the white-balancing algorithm that performs best for each image. The parameter space tends to form clusters where a specific algorithm tends to produce the most accurate result. This means that the Weibull distribution can be used to select the most appropriate white-balancing algorithm for each image individually [266, 267].

10.7 Summary

In this chapter, we looked at some of the regularities of color information in images and discussed a number of imaging applications that rely on them. The human visual system has special adaptations allowing it to take advantage of the color distribution in nature. One important such example is color opponency (Section 10.3). The visual system re-encodes information from the photoreceptors into three opponent channels, which effectively reduce redundancy. The same principle has been applied to images, leading to a number of color opponent spaces that rotate image data such that correlation between the three channels is minimized.

A consequence of the decorrelation abilities of such color spaces is that each channel can be processed separately without affecting the other two. One imaging application that relies on this particular property of color opponent spaces is color transfer, which was discussed in Section 10.4. Given a pair of source and reference images, color transfer methods aim to re-color the source image using the color palette or distribution of the reference.

In Section 10.5, we described a set of experiments for guiding the selection of a particular color space for color transfer as well as other image editing applications. Several standard color spaces were compared by means of a correlation analysis as well as a simple study. In addition, custom spaces were created through principal component analysis of large sets of images of the same category or even single images, finding that in most cases, the CIELab color space offers good decorrelation of color information in images.

Another important mechanism of the HVS that relies on color regularities of natural images is color constancy, which allows us to assess the colors of objects despite changing illumination. In the case of images, algorithms that are functionally similar to the color constancy of the visual system can be used to remove color shifts due to illumination. This process is then known as white-balancing. In Section 10.6 we discussed a number of such methods as well as a general formulation for many of them, which is based on the Minkowski norm. Finally, we showed how image statistics can guide the selection of white-balancing algorithm, increasing the likelihood of a correct white-point estimation.

Chapter 11

Depth Statistics

The natural world has three spatial dimensions, and we need to percieve all three to function within it. Although the images formed on our retinas are two-dimensional spatial representations of the world, we employ a variety of cues in order to determine the geometry and depth of the environment around us [343]. Some of these cues, such as binocular disparity or motion parallax, require us to be able to move and perceive the 3D environment directly. Yet our ability to understand depth relationships in scenes is still present even when viewing a photograph, indicating that at least a subset of the necessary information is still present in the 2D representation.

Statistics of 2D images can lead to powerful tools for image analysis and understanding, as we have seen throughout this book. Strong links have been found between many statistical regularities in scenes and human perception (see Chapter 1 in particular). To form a 2D image, 3D information is projected onto a plane—be it our retinas or a camera sensor, effectively flattening depth information and removing empty space. In 3D environments, in contrast, most space is empty, making the information within them very sparse. Consequently, it is reasonable to expect that statistical analysis of the 3D information will give rise to even stronger regularities describing the structure within scenes [584, 713, 773].

To study 3D scenes, the first step is to capture a representation of depth information within them. Typically, range scanners are used for this task, as they measure depth for each point in the scene. Figure 11.1 shows an example range image for a natural scene taken from the Brown Range Image Database [350, 447], visualized with a heat map (see Section 3.1.5 for a more detailed discussion on range and depth capture). Once range data is captured, it can be analyzed much in the same way as intensity images. By analyzing depth rather than intensity, the statistical properties of scene structures may be determined.

One of the earlier such studies analyzed images from the Brown Range Image Database using a series of simple statistical tools, including gradient and wavelet analysis [350]. Although many of the findings were similar to previously reported

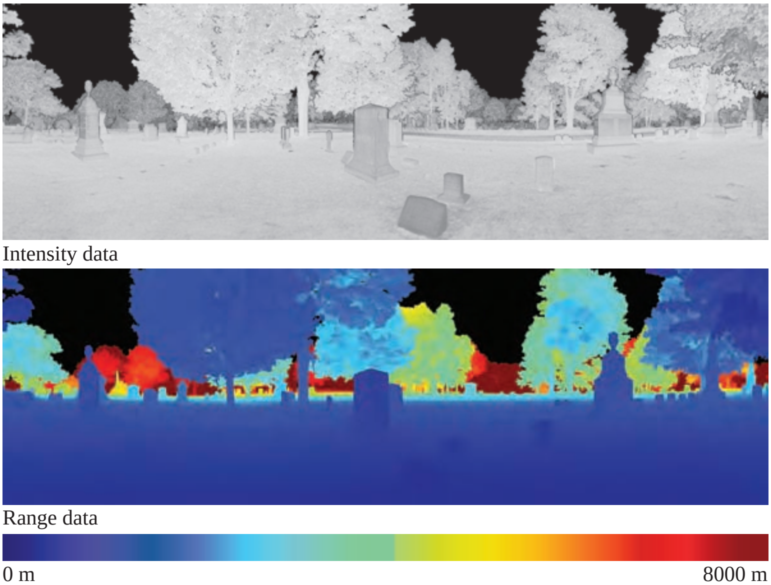


Figure 11.1. An example image from the Brown Range Image Database [350, 447]. The top image shows intensity data for the scene while the bottom image shows depth, here visualized with a heat map. Note that for both intensity and depth images, black pixels indicate areas where there is no information, such as sky areas or highly reflective surfaces.

statistics from analyses of 2D image collections (e.g., [352]), some interesting features were observed.

Gradient analysis of range images suggested that the structure of scenes as captured by such data is even sparser than would be obtained with 2D photographs. Figure 11.2 shows the logarithmic gradient distribution of the images from the Brown database for both the range and the intensity data. The log gradient distributions shown have a kurtosis of 5.1 for the vertical and 5.4 for the horizontal gradients, while the intensity distributions lead to kurtosis values of 3.2 and 3.3, respectively.

More striking results were found in the analysis of wavelet coefficients and bivariate statistics computed from the range data, further supporting the hypothesis that natural scenes are sparse and scale-invariant. Bivariate statistics can be used to study the correlations between pairs of pixels at a given distance [350]:

$$K(a,b\quad x) = Pr\quad I(x_1) = a, I(x_2) = b\qquad x_1 - x_2\quad = r\quad , \tag{11.1}$$

where x_1 and x_2 represent two different positions within the image I, which are at a distance r apart. Since images consist of mostly flat regions with discontinuities between them, it can be expected that nearby pixels are more likely to be corre-

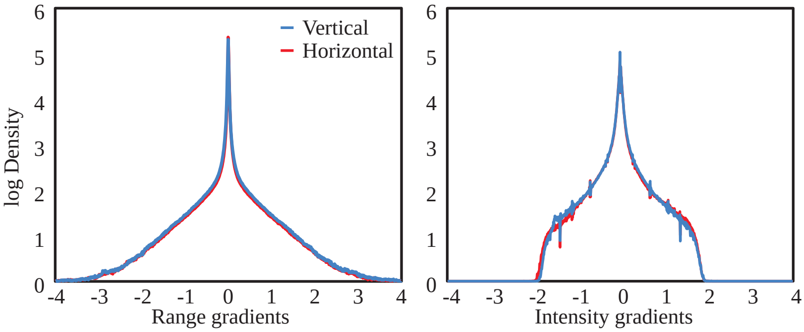


Figure 11.2. Log gradient distributions for the range (left) and corresponding intensity data (right) for images in the Brown Range Image Database. The increased kurtosis in the range distributions suggests that depth information is sparser than 2D intensities of the same scenes.

lated. In addition to the qualitative differences between the two representations, the bivariate distributions of the range data proved to be a better fit for a model formalizing the sparseness and scale invariance properties of natural scenes. This model will be discussed in the following section.

11.1 The “Dead Leaves” Model

One of the main goals of natural image statistics is to derive models that can robustly describe and predict images with natural characteristics. An important property of natural images is that they are highly non-Gaussian. Many of the statistical regularities of natural scenes are characterized by this recurring theme. For instance, in Chapter 5 we saw that gradient distributions both of individual scenes and image collections are highly kurtotic, with a sharp peak at zero. Similar distributions arise if we analyze wavelet coefficients (Chapter 8) as well as virtually any other transform we may apply to image data.

Given the non-Gaussian nature of images, what model can adequately describe natural scenes and their structure? A prevalent example is known as the “dead leaves” model, which has been repreatedly explored in the context of natural image formation [638, 12, 447]. Based on this model, images can be seen as consisting of sets of elementary objects (“leaves”), whose properties such as position and scale are drawn from a Poisson distribution, but which are independent of each other. Objects appear in layers and partially occlude each other, like fallen leaves would.

Figure 11.3 shows two example images generated with the dead leaves model using circular and square primitives and an image generated using a Gaussian model. All images are approximately scale-invariant, but the Gaussian model

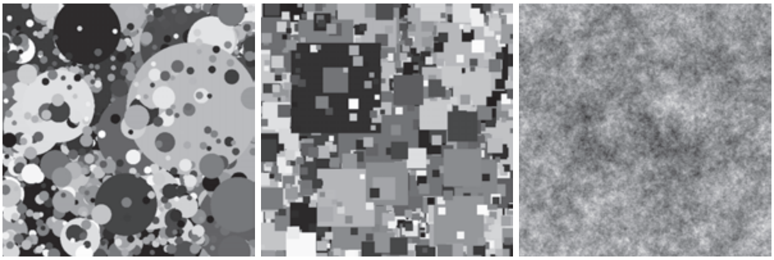


Figure 11.3. The first two images were generated with the dead leaves model [447] using circular and square primitives, while the third image was created using a Gaussian model. All images were constructed such that they are approximately scale invariant.

cannot generate the distinct structures that characterize natural scenes. Lee et al. showed that by using different primitives as in the two examples shown, or by varying additional parameters (e.g., using ellipsoid primitives with varying widths) this model can generate images that replicate the statistics of specific image classes, such as natural or manmade image categories [447].

In the case of bivariate statistics (see Equation (11.1)), the dead leaves model can be expressed as:

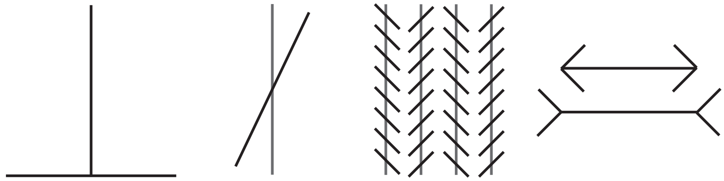
$$K(a,b\quad x) = [1-\lambda(x)]\, q(a)\, q(b) + 2\lambda(x)\, h_x(a+b)\, g_x(b-a) \tag{11.2}$$

where q represents the marginal distribution for a pixel, and h_x, g_x are distributions with predefined shapes. Specifically, h_x is similar to q, while g_x is concentrated at 0. The model effectively uses the parameter λ to control the probability of pixels a and b being part of the same object or belonging to different objects. If the bivariate distributions considered are looking at nearby pixels, a high λ value will be necessary, while larger separations are modeled with a lower value for this parameter. Despite its simplicity, the model can predict and replicate the statistical properties of natural scenes very effectively. More strikingly, when range data is analyzed instead of intensities, an even better fit can be achieved.

11.2 Perception of Scene Geometry

In addition to providing useful insight into scene formation, statistics of range data can be linked to the way we perceive scene geometry. Figure 11.4 shows some visual illusions where properties such as length and orientation are often misperceived. Scene statistics have been related to increasingly complex perceptual oddities and visual illusions such as these [344, 347, 346, 345, 348].

Although we live and function in a 3D world, we perceive the environment around us through 2D projections formed on our retinas. The loss of a dimension



a. Horizontal/vertical illusion

b. Tilt illusion

c. Zöllner illusion

d. Müller-Lyer illusion

Figure 11.4. Several oddities can be observed concerning the perception of lines of different orientations and configurations, many of which can be related to the statistics of 3D scenes. Some example illusions illustrating these mismatches between perception and actual scene configuration are shown here. (a) Horizontal/vertical illusion: both lines are the same length but typically the vertical line is perceived as longer. (b) Tilt illusion: the vertical line may appear to be slightly tilted counterclockwise. (c) Zöllner illusion: the vertical lines are parallel but appear tilted. (d) Müller-Lyer illusion: both horizontal lines are the same length but appear longer or shorted depending on the orientation of the arrowheads.

inevitably poses many ambiguitites that our visual system needs to resolve. However, in several cases, some aspects of the scene (such as distances or orientations) may be overestimated or underestimated. An interesting theory, which has gained growing support in recent years, explains these mismatches between real scene properties and their perceived counterparts by hypothesizing that we rely on statistical priors that relate the appearance of scenes in 2D projections with real 3D measurements [409, 601, 592]. A given 3D scene can lead to an infinite number of different 2D projections, all of which are mathematically valid. However, some are statistically more likely to occur in natural scenes than others, leading to a probability distribution for different aspects of scenes, which in turn may induce perceptual biases.

11.2.1 Length Perception

One example is the perception of length and its relation to orientation: if the orientation of a stimulus in a scene is changed, so is its perceived length (see Figure 11.4a). Thus, vertical distances or lengths appear longer than horizontal, even if they are identical, with the maximum perceived length occuring at 20 to 30 degrees from the vertical axis [577, 132, 655]. This relation is shown in blue in Figure 11.5.

Howe et al. showed that the relation between lengths in 2D images and corresponding distances in 3D data correlates very strongly with perceived length [344]. To study this phenomenon, a set of range images was captured and 2D projections of parts of the 3D scenes were formed. Pairs of points were then randomly selected from within the 2D images and their corresponding distance in the range

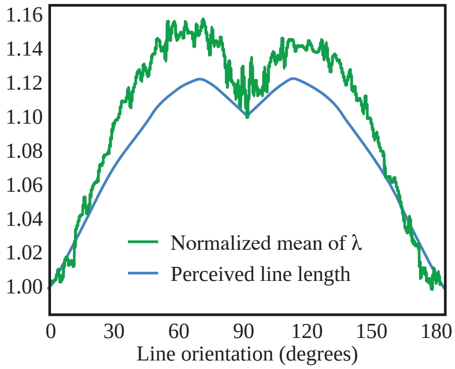


Figure 11.5. The blue line shows the relation between the perceived length of lines of the same actual length in images and their orientation. Observers generally perceive vertical lines as longer, with the maximum perceived length at 20 to 30 degrees from the vertical axis. The red line shows the statistical relationship uncovered by Howe et al. [344] between orientation of edges and the ratio of their length in 2D images versus their actual 3D distance. (Figure adapted from [344].)

scenes was measured. These measurements were used to compute a ratio λ between the projected length l and the real 3D distance. Figure 11.5 shows in green the relation between λ and the orientation of the segments, which very closely approximates the perceptual relation (shown in blue).

Other relations between real scene properties and 2D projections have also been studied, further supporting the hypothesis that the visual system may internally employ probability distributions of scene properties to make judgments about them [347, 346, 345, 348, 89]. Many visual illusions indicate that intersecting lines at different orientations pose a challenge to our visual system [622].

11.2.2 Orientation and Angle Perception

It has long been known that human observers tend to misjudge angles in images [326, 328]. Specifically, acute angles are overestimated while obtuse angles are underestimated [540]. The magnitude of misperception for different angles is shown in Figure 11.6.

The tilt illusion shown in Figure 11.4b is one of the simplest line configurations that can cause orientation misperception consistent with these angle mismatches. In such a configuration, one of the lines (in this case, the vertical gray line) is considered in the context of another (the slanted black line). Because the acute angle subtended by the two lines is perceived as larger than it really is, the vertical gray line in this case will likely appear to be slightly tilted away from the obliquely oriented black line.

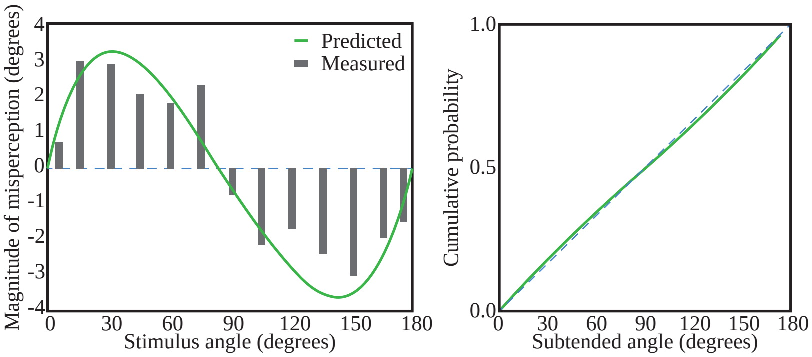


Figure 11.6. Left: human observers tend to misperceive angles. Acute angles are overestimated while obtuse ones are underestimated. The gray bars show psychophysical measurements [540], while the green line shows the predicted misperception, based on the statistical analysis in [346]. Right: the green line shows the cumulative probability of different angles occuring in physical, 3D scenes. If plotted in degrees rather than probability, the green line in the left plot is obtained. (Figure adapted from [346].)

Similar to perceived length mismatches, evidence from studies of range images suggests that this perceptual oddity can also be explained by the statistics of natural scenes [346]. To study the relation between angles in 2D projections and their corresponding physical sources in the range data, intersecting line segments were detected in images using a template-based approach. Given that data, the probability of different configurations of intersecting lines were computed, revealing that 90-degree intersections were less likely than smaller or larger angles subtended by the two lines. Figure 11.6 (right) shows the cumulative probability of different angles as detected in this study.

Interestingly, the probability of different physical line configurations occurring in the range data was shown to be a strong predictor of the perceptual misjudgments of angles in images [346]. At a high level, this can be understood as having a prior expectation of what line configurations we are likely to encounter in our visual environment in order to resolve the ambiguity of the 2D projections on our retina. If this is true, then projected angles would be misperceived such that they fit with the probability of a given physical angle causing them. As can be seen in Figure 11.6 (left), statistical findings support this hypothesis.

Specific mismatches like the ones discussed so far are not the only perceptual aspects that can be explained by the statistical structure in natural scenes. Statistical analyses of scenes have repeatedly provided evidence that a probabilistic model may underlie many perceptual oddities, especially in the context of scene geometry [810, 809, 345, 348, 472]. We refer the reader to [347] for a more detailed discussion of these phenomena.

11.3 Correlations between 2D and Range Statistics

The previous section looked at the links between statistical properties of range data and the way we perceive scenes. The goal of these studies is ultimately to understand human vision by determining what priors, if any, the visual system may use to resolve ambiguities between 3D physical scenes and their 2D projections. A similar problem is faced by computer vision and image processing: algorithms need to make decisions about the content of scenes based on 2D representations of them. Since a given image may be generated by an infinite number of different configurations, some priors or assumptions are necessary to guide algorithms through this inherent ambiguity of 2D projections.

Throughout previous chapters, we have discussed several cases where 2D image statistics over large image sets have been used as priors for particular algorithms (e.g., see Sections 5.6 or 5.8). However, in these examples, priors serve as a representation of the structure that we would expect to find in a natural 2D image. On the other hand, the studies discussed in the earlier parts of this chapter have focused on statistics of range data describing the 3D structure of scenes.

In contrast to either of these approaches, a small set of studies has focused on correlations between 2D statistics and their 3D counterparts. Kalkan et al. analyzed patches in 2D and 3D scenes in order to determine the physical causes of different types of structure in images [384, 385]. Structures in 2D images can be categorized according to the following classes (from [384]):

- Homogeneous patches,
- Edge-like structures,
- Corners, and
- Texture.

Figure 11.7 shows examples of these structures in an image.

Although it is possible to find patches that exemplify the characteristics of these different categories, most image patches are likely to display mixed characteristics and are therefore better described in a continuum, such as using the *intrinsic dimensionality* scheme [210, 426, 209]. This scheme was first introduced in image processing applications by Zetsche & Barth [817] and it classifies image patches according to the shape of their Fourier spectrum. A patch is categorized as:

- **i0D** if the spectrum is concentrated at the origin, as would be the case for homogeneous patches,
- **i1D** if the spectrum forms a line, which is likely to occur if the patch contains an edge, and

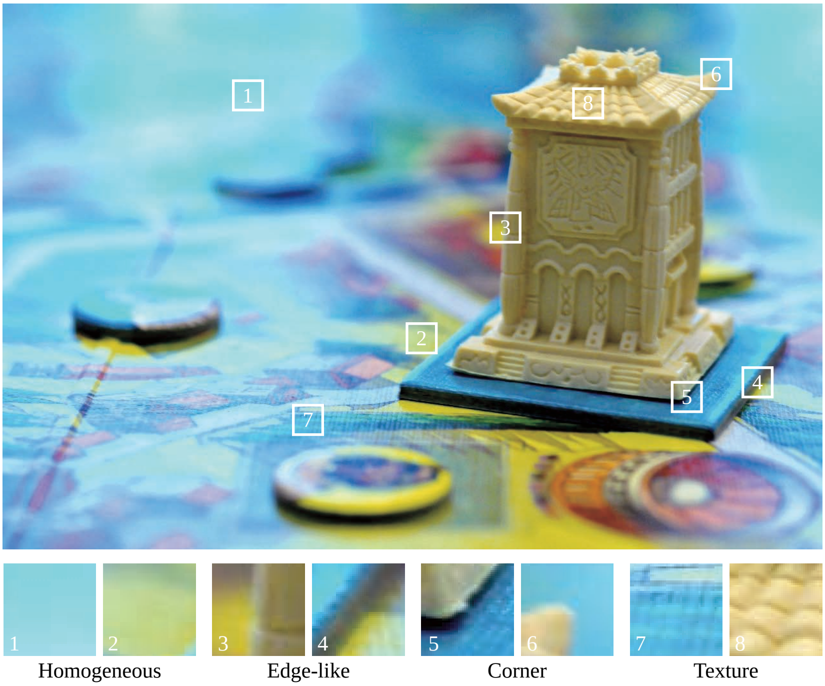


Figure 11.7. Examples of the four types of image structure identified by Kalkan et al. [384].

- **i2D** if the spectrum is neither focused at the origin nor forming a line, which would be the case with corner or junction structures.

As most patches will not perfectly fit within a single category, the intrinsic dimensionality of a patch can be defined using barycentric coordinates as a point in a triangle, as illustrated in Figure 11.8. Please refer to Felsberg and Krüger [210] for details regarding the computation of this space.

Similar to 2D images, patches of 3D geometry also exhibit different types of structure. In some cases, such as for continuous surfaces, 3D structures can be directly correlated to corresponding 2D patches: a homogeneous patch in an image is likely to be caused by a smooth, continuous surface. When discontinuities occur, however, correlations between different patches are less straightforward. For instance, an edge in a 2D patch may be caused by a depth discontinuity where one object partially occludes another. It may also be due to texture on an otherwise continuous surface, or it may be caused by orientation discontinuities where two surfaces meet.

To determine whether different types of 2D structure can predict the underlying physical geometry, Kalkan et al. analyzed scenes where both range and

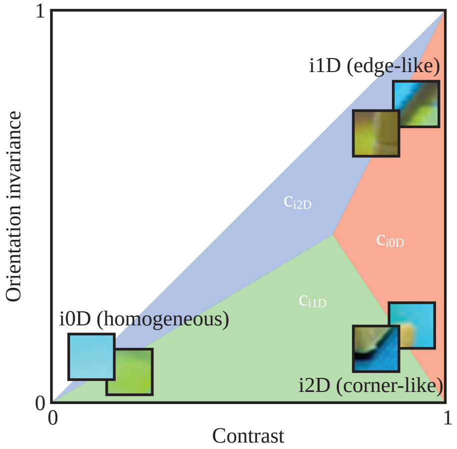


Figure 11.8. A visualization of the intrinsic dimensionality space for image patches. The intrinsic dimensionality of a patch is given by barycentric coordinates ($c_{i0D} - c_{i1D}$). The axes can be understood as a representation of contrast of a patch (x-axis) and orientation invariance on the y-axis. (Figure adapted from [210, 384].)

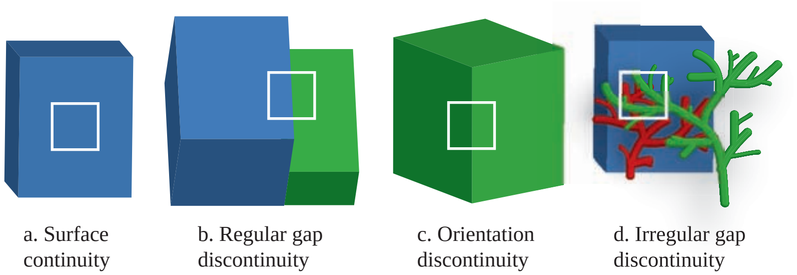

a. Surface continuity
b. Regular gap discontinuity
c. Orientation discontinuity
d. Irregular gap discontinuity

Figure 11.9. Different types of patch structures in 3D geometry as identified by Kalkan et al. [384].

chromatic data was captured [384]. Four types of geometry discontinuities were identified:

Surface continuity, where the underlying surface is homogeneous and does not change.

Regular gap discontinuity, occurring where a small set of surfaces meet or overlap.

Orientation discontinuity, occurring at corners where two surfaces with different orientations meet.

Irregular gap discontinuity, where the 3D structure cannot be described with a small set of surfaces and more complex interactions take place, e.g., tree branches or leaves.

These types of surface discontinuities are illustrated in Figure 11.9 using simple surfaces.

Based on this classification and the intrinsic dimensionality scheme described earlier, corresponding patches in the range and 2D images were selected and categorized, allowing for correlations between 2D and 3D structures to be computed. As would be expected, continuous surface patches in 3D correlated strongly with homogeneous regions, and orientation discontinuities appeared as edges in images. But some less obvious correlations were found with other structure classes as well. Regular gap discontinuities were correlated with edges, and more so, with corner structures. Additionally, irregular gap discontinuities appeared mostly correlated with texture structures in their 2D counterparts [384, 385].

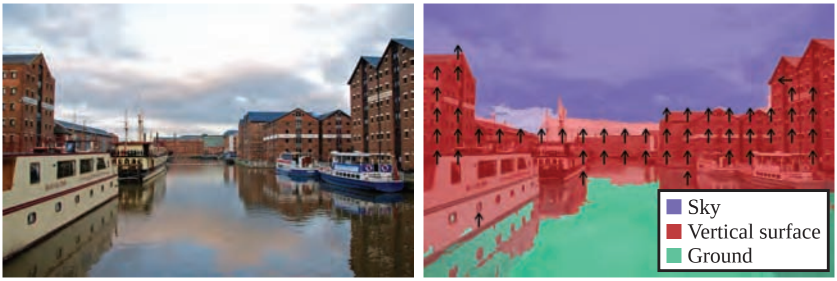
11.4 Depth Reconstruction

In Section 4.5, the "Dark-Is-Deep" paradigm was discussed. Extracting depth from two-dimensional projections of a scene is and has been an actively researched problem [413, 589, 819]. Since this is an underconstrained problem, statistical priors can be employed.

Human vision often relies on contextual information to resolve underconstrained problems such as recognizing objects or estimating the shape of an item. By considering not only the object in question but also the environment in which it is placed, contextual information can help constrain the space of possible solutions [553, 158]. Recently, statistical priors have served to provide context in the problem of both object recognition [749, 748, 752] and geometric reconstruction of scenes from 2D images [750, 336, 337].

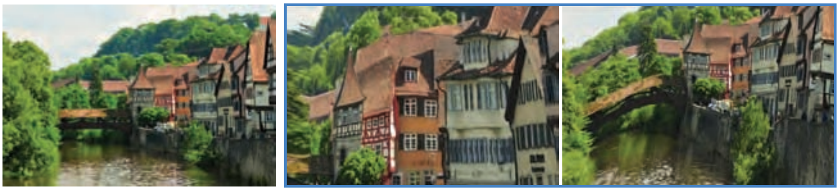
In an attempt to estimate surface orientation in scenes, Hoiem et al. rely on statistical learning to divide the scene into oriented surfaces. Their method relies on two observations. First, most surfaces in outdoors images can be categorized as either sky, ground, or vertical surfaces perpendicular to the ground. Second, the orientation of surfaces in 3D can be determined from the appearance of corresponding image regions in 2D [336, 337]. Although this approach does not recover accurate depth for each pixel in the scene, it successfully creates a geometric context from a single image as shown in Figure 11.10.

Alternatively, instead of attempting to understand the scene as a whole, depth can be recovered from images by learning the relations between image structures and their corresponding depth. As we saw in Section 11.3, even when considered in isolation, image patches have been found to correlate to particular types of depth discontinuities [385, 384]. Although these correlations are not



a. Original image b. Geometric context reconstruction

Figure 11.10. Using a statistical learning approach, Hoiem et al. recover the geometric context of the scene, classifying surfaces in the image as sky, ground, and oriented surfaces [337]. (Gloucester, UK, 2011)



a. Original image b. Views from reconstructed virtual environment

Figure 11.11. Using global and local features, Saxena et al. trained an MRF to reconstruct 3D geometry from a single image [652]. The image (a) can be mapped on the reconstructed geometry to form a virtual environment (b) that can be navigated interactively. (Schwäbisch Hall, Germany, 2012)

strong enough to be used directly for reconstructing depth from an image, probabilistic learning models, such as the Markov random fields (MRF) discussed in Chapter 9, have been used to achieve reasonably accurate reconstructions of 3D scenes [750, 652, 597, 596].

Saxena and colleagues combined both contextual information and local priors in a supervised learning approach for depth recovery from a single, monocular image [649, 651, 650, 652]. Based on the observation that local features cannot be accurately interpreted without context, they used a set of multiscale features to train an MRF. Images were divided into patches, for which a depth value was computed from ground-truth 3D scanned data. Patches were then analyzed using two types of features that looked at absolute depth in each patch and relative depth relations between two patches [649].

These examples demonstrate that even a single image contains sufficient information to recover a representation of the underlying scene geometry. Although the methods discussed do not aim to compute accurate depth maps, they show that well chosen statistical priors carry significant descriptive power.

Chapter 12

Time and Motion

The world is ever changing. Heraclitus of Ephesus stated that “Everything flows, nothing stands still.”[1] He is also the source of the more well-known saying, “You could not step twice into the same river; for other waters are ever flowing on to you.”[2] This constant flux and change of the world is an important source of information.

Despite the impressive range of visual abilities in different animals, there are no known motion-blind species [72]. It seems that, biologically, motion perception is one of the most basic visual abilities. This can be especially seen in those species with limited spatial vision: they use motion to compensate for the lack of retinal receptors. The *Copilia quadrata*, for instance, moves its spotlight retina in a particular pattern, giving it one-dimensional vision. Similarly, the carnivorous sea snail and the jumping spider move their eyes so that the one-dimensional strip of visual receptive cells can sense two spatial dimensions.

Since many of the changes that occur in the world are not random, temporal flux is a potential source of information not just for organic visual systems, but also both the synthesis and computational analysis of image sequences. This chapter examines some temporal properties with a focus on motion. It also introduces some scenarios where these findings have been applied in visual computing.

12.1 The Statistics of Time

All of the statistics in this book so far have been applied to one or more spatial locations at a given point in time. Other than technical reasons, there is nothing preventing us from performing the temporal equivalent: comparing different points in time for a given spatial location. Although such purely temporal statistics are certainly possible, it is much more common to examine both space and time simultaneously.

[1] As quoted by Plato [379], p. 344.

[2] As quoted by Plato [379], pp. 344–345.

We could compare just two points in time. Rather than subtracting two spatially neighboring pixels to get a spatial gradient (see Chapter 5), it is possible to subtract the intensity value at one pixel for two temporally neighboring images in an image sequence. This would give a measurement of the temporal contrast modulation. Likewise, as was shown in Chapter 6, by varying the separation (either in space or time) of the pairs of pixels, we could examine the statistical regularities at different frequencies (spatial or temporal, respectively).

In perhaps the first spatiotemporal analysis of image sequences, a series of works examined the spatiotemporal properties of natural image sequences that were not photographs. Specifically, they examined multispectral satellite image sequences (LANDSAT multispectral scanner sequences) and were able to detect such regularities as seasonal changes [367, 470, 471, 98].

For video sequences of terrestrial-based natural scenes, the first spatiotemporal analysis seems to be that of Watson and Ahumada [781], who suggested that time be considered like space, and thus image sequences should be represented as a spatiotemporal volume $I(x,y,t)$, where $I(x,y)$ is a given image and t is the time at which that image was taken. They then proposed that the frequency transform of these spatiotemporal volumes be examined.

Starting with an analysis from first principles, they showed that different forms of motion within an image sequence should trace very specific paths within the spatiotemporal volume. For example, a vertical line moving horizontally (a common stimulus in psychophysics) should trace a diagonal path through the spatiotemporal volume. Interestingly, in the frequency domain, the path lies along a straight line in f_s, f_t (where f_s is the frequency domain transform of x and y and f_t is the frequency domain transform of t). The slope of this line is then $-1/r$ where r is the horizontal speed. Thus, higher spatial frequencies will have higher temporal frequencies, and higher velocities will have shallower slopes. Moreover, leftwards motion will show up in the odd quadrants while rightwards motion will show up in the even quadrants in Fourier space. Physical analysis of real images containing a vertical line moving horizontally confirmed the theoretical analysis. Moreover, the critical sampling frequency for which discretely presented motion should be seen as continuous was shown to be a linear function of velocity.

Finally, Watson and Ahumada also demonstrated that the spatiotemporal volume can be used to create motion detectors [781]. The idea is to use pairs of simple cells with one of the cell's responses being adjusted by a hyperbolic filter. The response is modified once in the spatial domain and once in the temporal domain. The signals of the two cells are then added. Note that the sensor only detects direction of motion, not speed. The speed is obtained by examining the temporal frequency response of the sensors. Eckert and colleagues examined the spatiotemporal power spectra of 14 image sequences (256×256 pixels $\times$ 64 frames at 30 frames per second, with no scene cuts), and found that all spectra were separable (with an index of separability of 0.98) [192]. The power spectra fit the form:

$$P(k,f) = \frac{(ab)/(4\pi^3)}{((a/2\pi)^2+k^2)^{3/2}\,((b/2\pi)^2+f^2)} \tag{12.1}$$

with k representing the radial spatial frequency, f representing temporal frequency, and a and b representing model parameters that define the spatiotemporal bandwidth of the signal.

Similarly, Adelson and Bergen [6] suggested that the statistical regularities of spatiotemporal sequences can be examined with any traditional volume analysis. They also created a motion detector (the energy detector) that is formally identical to the elaborate Reichardt Detector (see Section 12.2.1). Many others have since used spatiotemporal volumes to analyze or manipulate image sequences, examining slices, tunnels, and other features in the volumes (see, e.g., [535, 536, 537, 619]).

Dong and Atick examined the spatiotemporal properties of a set of 1,049 commercial movie clips (each clip was 64 × 64 pixels large and 64 frames long, with a temporal resolution of 24 frames per second) and 320 clips from homemade movies (each at 64 × 64 × 256 with a temporal resolution of 60 frames per second[3] [174]). Their basic measure was the spatiotemporal correlation $R(x,y,t)$ between two pixels. In a first analysis, the correlation between two pixels at the same point in space but at two different points in time (a separation of 33 ms) was examined. Notice that this correlation is related to a pure temporal gradient. They found that the value of the pixel at the two times was highly correlated ($R = 0.9$). In other words, whatever the intensity was at that pixel at time 1, there was a high probability it was the same 33 ms later. The greater the difference in intensity between two frames was, the lower the probability for that change to happen. In fact, a visual inspection of the probability distribution suggests that it is strikingly similar to the spatial gradient distribution, once again revealing a high kurtosis (see Chapter 5).

Dong and Atick then examined the two-point correlation matrix for many spatial and temporal separations, as well as the results in the frequency domain [174]. They found that for any given temporal frequency, the spatial frequency power spectrum follows the usual power law:

$$1/f_s^a \tag{12.2}$$

where f_s is the spatial frequency. Interestingly, the value of the slope a varied as a function of temporal frequency. At low temporal frequencies they found that the exponent a tends to 2 while at higher temporal frequencies a approaches 1.

The temporal frequencies showed the same behavior: for a given spatial frequency, the temporal frequency power spectrum also follows a power law:

$$1/f_t^b \tag{12.3}$$

[3] They also tried to avoid scene cuts in the clips, as these do not represent natural motion.

where f_t is the temporal frequency. Interestingly, it was found that the exponent b varies as function of spatial frequency. The slope b tends to 2 for low spatial-frequencies and b approaches 1 for higher spatial frequencies. The temporal slope can be thought of as being related to the persistance or purposefulness of motion [60]. A slope of 0.5 represents pure Brownian motion: the changes from frame 1 to frame 2 are completely uncorrelated with the changes from frame 2 to frame 3. Slopes greater than 0.5 represent persistent motion: any change that occurred from frame 1 to frame 2 is likely to occur again from frame 2 to frame 3. Exponents smaller than 0.5 are anti-persistent (the opposite motion is likely to occur).

Critically, if the correlation is multiplied by a power of f_s (such as f_s^{m+1}) and plotted as a function of the ratio of temporal to spatial frequency f_t/f_s, all the curves line up [174]. Otherwise stated:

$$R(f_s, f_t) \approx \frac{1}{f_s^{(m+1)}} F(f_t/f_s) \tag{12.4}$$

where m is a constant and $F(f_t/f_s)$ is a function of the ratio of spatial and temporal frequencies. In both the theoretical analysis and the experimental estimates (using real image sequences), the value of m is around 2.3, which is consistent with the slope of the power spectra for static images. Theoretically this function, including the value of m, can be derived by assuming a specific distribution of relative velocities from objects at many depths [174]. This result can be linked to human vision, in that the receptive fields of neurons are often coupled in space and time, thereby reflecting the real-world spatiotemporal statistics. As such, receptive fields should not be characterized by their spatial and temporal properties separately but by the ratio of temporal to spatial frequency f_t/f_s [175].

The properties of natural image sequences have been compared to the visual systems of both humans as well as flies. In one study, the spatial properties of 117 photographs, using patches of 128 × 128 pixels, were analyzed, revealing the typical $1/f_s^b$ power spectrum, with $b = 2.13$ [311]. Note that if an observer were to move in a straight line through a static environment that was uniformly cluttered with objects, the predicted power spectrum would be $1/f$ for a range of observer speeds [311], which is at the lower end of the range measured by Dong and Atick [174]. The spatiotemporal statistics were used to derive the receptive field structure of cells early in the fly's visual system.

Alternatively, it is possible to capture digital recordings of what a single retinal cell would see by having people wear a specially designed photosensor-based recording system on their head as they walk around [313, 315]. Signals captured in this manner were also found to conform to the $1/f$ property of the temporal power spectrum, confirming the predicted value for observers moving through a static environment. In addition, by processing the captured signal, the responses of various stages of the visual system were simulated. Figure 12.1 shows the differences between direct measurements of this time sequence and (simulated)

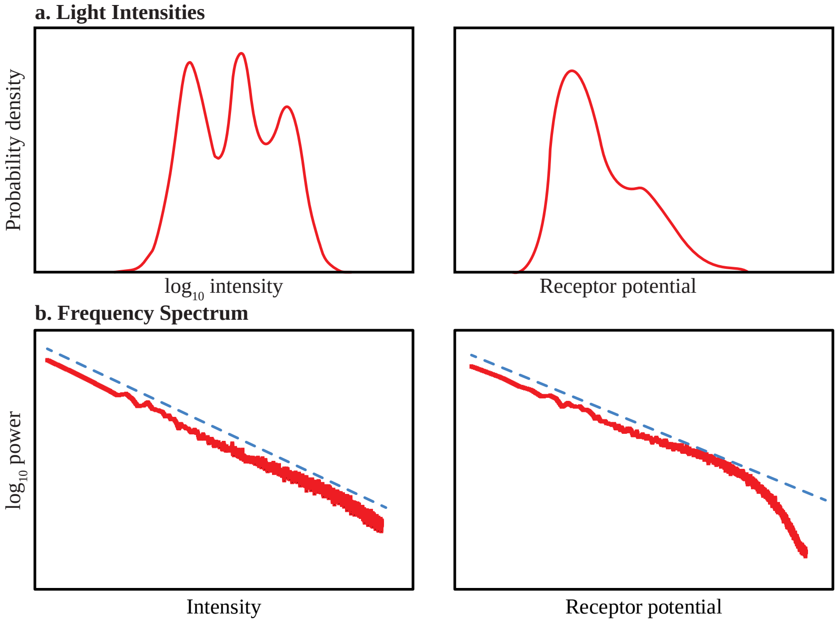


Figure 12.1. (a) Probability distributions of the light intensity (left) and the corresponding photoreptor responses for a time sequence simulating the input to a single photoreceptor. (b) Average power spectrum for consecutive sections from the same time sequence, following approximately a $1/f^t$ behavior, with $t = -1$. In the photoreceptor-processed spectrum, higher frequencies are filtered by the low-pass behavior of the photoreceptors and as such deviate from $1/f$. (Adapted from [315].)

photoreceptor responses for the same data in terms of intensity distribution and power spectrum [315].

Independent component analysis (see Chapter 7) can also be performed on image sequences [314]. A database of 216 clips taken from television signals, each 128 × 128 pixels large and 4,800 frames long (192 seconds at 25 frames per second) was used for this purpose. The first 288 independent components (ICs) for a set of patches, each 12 × 12 × 12 large, were calculated. The ICs resemble edges (or bars) moving at a constant velocity, with the direction being perpendicular to the orientation of the bar. Interestingly, this is very similar to the spatiotemporal receptive fields of LGN neurons [162]. The related independent component filters (ICFs; which can be used to filter an image sequence) also resemble other aspects of the human visual cortex (especially in their spatial properties). Interestingly, they found a significant correlation between spatial frequency and velocity, with low spatial frequency ICFs being tuned to faster movement than higher spatial-frequency ICFs and fast-moving ICFs mainly being encoded at low spatial frequencies.

12.2 Motion

One of the strongest statistical regularities that arise when we look at spatiotemporal changes is that of motion. Motion has been the subject of intense interdisciplinary study for at least the last 150 years. A large number of theories exist about what kinds of motion there are, what causes motion, what properties motion has, and how motion can be processed. For a recent overview on the perception of motion and some of the related theoretical and philosophical issues, the reader is directed to [95, 125, 270, 736].

Before we can examine the statistical regularities of motion, we first need to extract it. That is, we need some form of spatiotemporal transform that takes image sequences and extractes or emphasizes motion information. We have already mentioned the one from Watson and Ahumada [781] as well as the one from Adelson and Bergen [6]. The very first motion detector, though, was developed a century earlier by Sigmund Exner [201]. Although Exner's model is no longer used, it provides an excellent (and remarkably accurate) starting point for subsequent work on motion detection. Here, we present the basics of the two most common forms of motion detection. Nearly all analyses of the statistical regularities of motion start with one of these. We then examine a number of motion regularities.

12.2.1 Correlation-Based Motion Detection

In the 1950s, Werner Reichardt examined the behavior and underlying neurophysiology of the beetle and the fly, with a focus on motion perception [309]. He suggested that a good motion detector should see a given object as it moves through two different points in space at two different times. Thus, it needs two temporal gradient sensors separated by a specific distance in space and one of them should respond with a specific temporal delay. The two signals are multiplied together to obtained the motion signal. Since both receptors must see the *same* change in the environment, this form of motion detector is called a *correlator type*.

The basic version of the Reichardt motion detector allows detection of motion at a specific speed in a specific direction [309]. If an object is moving in the right direction at the right speed, then the change in illumination caused by the object passing over the first receptor cell will hit the multiplication point just after the object passes the second receptor, causing the motion detector to emit a signal. By adding a second set of pathways connecting the same two sensors to a second multiplier, and then subtracting the signals from the two multipliers, we can construct an elaborate Reichardt detector, which is capable of detecting motion at a given speed in the two opposed directions. Technically, the Reichardt detector measures temporal frequency [125], but Dror has shown that due to the statistics of natural images this is a very reliable and consistent estimation of velocity [183].

It has been shown that humans use correlator-type detectors [322, 594] (so do flies [294]), that the tuning is done by changing spatial preferences [16, 93], and that the detector can explain many low-level visual illusions [92, 94]. There exist several computational implementations of the elaborate Reichardt detector (e.g., [137]).

12.2.2 Gradient-Based Motion Detection

In examining the encoding of television signals, Lamb and colleagues [462] noted that the correlation approach requires many multiplications, and that any increase in the number of velocities that should be detected required the addition of a new multiplication unit.[4] To create a more efficient encoding of television signals, Lamb and colleagues developed a model for measuring velocity directly from spatial and temporal gradients. First, the temporal derivative of local luminance ($\partial I(x,t)/\partial t$) and the spatial derivative of local luminance ($\partial I(x,t)/\partial x$) are calculated. Then the two signals are divided. Gradient-based detectors have the advantage that they are not dependent on pattern properties. The disadvantages include the fact that for low spatial derivatives, any noise in temporal derivatives is amplified.

12.3 Applications Using Statistical Motion Regularities

Once motion has been extracted, a number of statistical regularities are noticeable. The perception literature is full of many cases of motion-based perceptual processes and motion illusions, such as structure-from-motion, color-from-motion, depth-from-motion (e.g., motion paralax), heading and spatial orientation from optical flow (see Section 12.4.2), and so on. For an overview of the perception of motion and some of the many motion-based statistical regularities that people can sense, please see [736].

There are also a number of applications that take advantage of spatiotemporal regularities. These applications use a wide variety of approaches from simple gradients up to scale space and Markov random fields (for example, noise reduction [187, 828], feature tracking [247], camera motion estimation [333], video retrieval [334], motion interpretation [442, 443, 444], and object segmentation [702]). Several others are discussed in this section.

[4] It has subsequently been shown that the multiplication phase can be implemented using a combination of linear and nonlinear filters and does not explicitly require multiplication. For more, please see [125].

12.3.1 Specular Highlights

It has been shown that the pattern of change over time can be used to extract specular highlights. Using knowledge of how caustics change in real-world scenes and some simple geometry, specular highlights can be removed from video sequences [716]. Similarly, a “temporal color space” can be defined, where each pixel in an image sequence is represented in a four-dimensional space, with three color dimensions (R, G, and B) and one temporal dimension [646]. Given knowledge of how specular reflections change as a function of time, and that the specular highlights generally have the same spectral distribution as the illuminant, specular highlights can then be removed and the color of the illuminant can be estimated in an image sequence.

12.3.2 Markov Random Fields

Many researchers use Markov random fields (MRFs; see Chapter 9) with temporal data. For example, it can be shown that with three assumptions, namely that all luminance changes are due to translational motion, that Gaussian noise corrupts the observation, and that slower motions are more likely to occur than faster ones, the trained MRF model interprets several motion illusions in a manner similar to humans [792].

MRFs can be used to track people in videos using three underlying models (one each for edges, ridges, and optical flow) [682]. After training the models on 150 images and short sequences, results were shown to be better when all three models were used than for any of the three models individually.

Black and colleagues used four generative models within an MRF context to detect motion in general [62, 63]. The four models reflect changes in form (i.e., including translational and rotational motion), changes in illumination, changes in specular reflections, and a catch-all for all other changes (which they call iconic changes). Note that the iconic changes are domain-dependent and thus require a separate model for each set of domain-specific change (e.g., a model for detecting eye blinks needs to be represented separately from a model for mouth motion). They tested the models on complex, real-world scenes and showed that once again it works best when all four models are used simultaneously.

12.3.3 People Detection and Biological Motion

People are remarkably sensitive to the motion of other people. In fact, Johansson showed that by placing dots on the joints of people and showing just the motion of those dots (usually a set of about 15 points) is sufficient for people to recognize how many people are present in the video and what they are doing [374, 375]. It has since been shown that these motion signals are also sufficient to tell many other things, including what gender a walking person is (for an overview, please see [736]). This field has come to be known as biological motion. A number of models of how people process biological motion signals have been proposed (see, e.g., [114, 145, 265, 754]).

In an interesting variation, it was suggested that crowds of people can be tracked by following corners [10]. In essence, once corners are detected, it is assumed that only the corners on people move (background corners do not move) and that each person has on average the same number of corners. Those assumptions allow humans to detect how many people are present. A measurement of the average speed for walking in a sequence allows the detection of running later on.

12.3.4 Motion Blur

All of the photons that hit a given cell within a specific temporal window are integrated and are perceptually treated as if they all arrived at exactly the same instant in time [94, 270]. The temporal integration window is believed to be about 80 ms in bright conditions and around 150 ms when we are in the dark and our eyes are fully dark-adapted. When an object is stationary, all photons from a specific location on the object hit the same receptor cell for the duration of the temporal integration window. The result is that the retinal cell receives enough photons to respond and the retinotopic location of the spot on the object is unambiguous.

If the object is moving fast, however, light from a given spot on its surface will sequentially hit a number of neighboring retinal cells during the temporal integration window. This means that this spot will be seen by the retina as existing at one instant in time at multiple locations. Thus, directly after processing by the retinal receptors, the object will seem to be longer in the direction of motion than it should be. Moreover, the visual system will not be able to resolve the higher spatial frequencies, since they are spread out. Cameras exhibit similar phenomena relating to blur since the shutter speed of modern cameras is not zero.

The human visual system, however, is built such that it can at least partially compensate for this motion-based spreading. Once the visual system detects that the object is moving, it automatically performs some motion deblurring [270]. Several computational approaches to removing motion blur from images have been developed (see Section 5.6). Motion blur is also routinely applied to computer-generated sequences, often with the aid of high dynamic range imaging [164].

12.3.5 Shape from Dynamic Occlusion

As discussed by Reichardt, a moving object disappears from one location and appears at another location. Obviously, when the object disappears from a location that location is not empty. One of two things can happen: either another part of the surface of the moving object shows up (which consequently correlates with a disappearance elsewhere) or some background texture suddenly appears.

Likewise, at the leading edge of the figure, background texture will suddenly disappear. The specific pattern of background texture appearances and disappearances at the edges of a moving object is referred to as *dynamic occlusion*. A moment of thought will show that the appearanace and disappearance of the background will not have a match elsewhere and thus will not be picked up as a motion signal.

These temporal contrast modulations that do not come from motion provide a wealth of information. Obviously, they indicate that there is or was an occluding object at that location [508, 509]. It also provides reliable information that disappearing background still exists [264]. Finally, the collection of dynamic occlusion signals in an image sequence specify the shape, velocity, and degree of transparency of the moving object [15, 73, 81, 82, 83, 121, 122, 131, 140, 141, 264, 330, 392, 393, 552, 590, 624, 679, 680, 681, 705]. For a review, see [678]. The visual process responsible for seeing a form from dynamic occlusion changes is referred to as Spatiotemporal Boundary Formation (SBF). The class of changes that can result in the perception of a moving bounded form extends beyond the simple accretion and deletion of texture elements; the class includes changes in an element's color, orientation, or location [680].

Shipley and Kellman [681] provided a mathematical proof showing that the orientation of a small portion of the moving object's contour (a *local edge segment* or LES) may be recovered from changes of three texture elements, as long as the elements are not spatially collinear and all three changes occur within the temporal integration window. If the velocity of the LES is known, only two element changes are needed. Cunningham and colleagues [139, 131] modified and extended the proof to show how the global shape and the global velocity can also be recovered from six element transformations. They then tested the model on synthetic and real-world images, with decent results. Most of the failures of the algorithm on real-world images seems to lie in the implementation of the Reichard detectors, which are needed to separate the motion-based and non-motion-based temporal gradients.

12.4 Optical Flow

Perhaps the most well-known and widely used motion-based technique in visual computing is optical flow. Optical flow has its origins in James J. Gibson's work describing the information that a moving animal receives [261]. Optical flow is in principle described by the temporal gradient field for an image sequence. Essentially, the change in location of each point on each object in a scene is measured. In practice, this can be performed on an image directly, rather than on the real-world scene. If both the observer and the real-world scene are perfectly stationary, then there will be no difference between the locations of any objects (or their two-

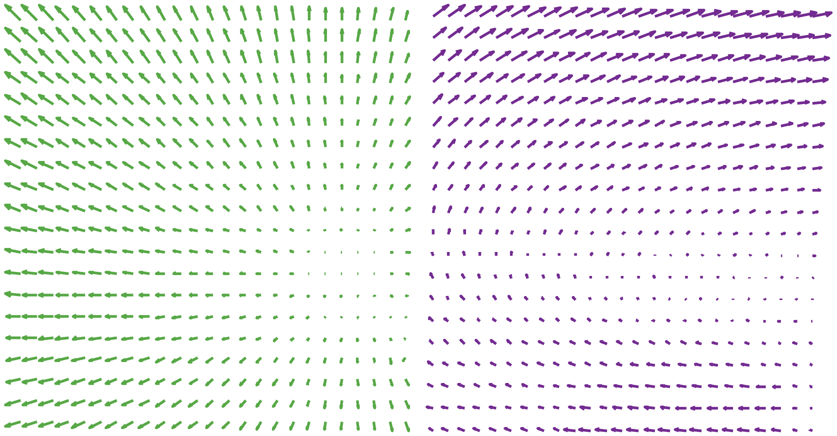


Figure 12.2. Two examples of flow fields. The field on the left shows a radially expanding flow, which is likely to be due to forward motion within the scene. The movement within the right field is most likely due to camera rotation.

dimensional image projections) at any two points in time and the vector field will be zero everywhere.

If the observer is moving forward (while looking forward), every point in the image will move towards the borders of the image (see the left panel of Figure 12.2). This yields a radially expanding vector field; the center of the field is the point towards which we are traveling [263]. If the observer is looking forward but traveling to the left, all points in the image will move to the right of the image, creating a vector field with a large global translation component. In principle, optical flow can be used to determine the direction and speed of motion of a moving observer, since the same pattern of optical flow is obtained as when the scene is moving and the observer is static. All that is necessary is relative motion between the observer and the scene. Although humans can use optical flow for determining direction of travel, they often do not seem to use it in real-world conditions [306].

Optical flow is also useful in object perception. If an otherwise static scene has a single moving object in it, this will show up as a local patch of high vector values in an empty optical flow vector field. Likewise, if an object within a scene has a different motion than its surrounding scene, this will also show up as a characteristic pattern within the optical flow field [261] (see [574] for an example of how this can be used in a computational framework). Of particular interest is the fact that at the edge of a moving object there will be discontinuities in the flow field. Humans can use these discontinuities to detect the shape, location, relative depth, global motion, and transparency of the moving object (as discussed in Section 12.3.5).

Many optical flow algorithms are gradient-based. Given that an image sequence is represented as a spatiotemporal volume $I(x,y,t)$, where (x,y) are spatial indices and t is the temporal index (see Section 12.1), it is assumed that the total derivative of the image intensities in both space and time is zero at all times, i.e.:

$$I_x(x,y,t)\,v_x + I_y(x,y,t)\,v_y + I_t(x,y,t) = 0 \tag{12.5}$$

Here, the subscripts indicate partial derivatives and $\mathbf{v} = (v_x, v_y)$ is the optical flow, which is always in the image plane and normal to the spatial image orientation [688]. The implicit assumption made here is that any changes in intensity are due to translation and not to changes in illumination and reflectance. This equation would allow one to estimate an optical flow field $\mathbf{v}$, for instance by defining an error term:

$$E(\mathbf{v}) = \left(I_x(x,y,t)\,v_x + I_y(x,y,t)\,v_y + I_t(x,y,t)\right)^2 \tag{12.6}$$

It is then possible to compute a linear least-squares estimate of $\mathbf{v}$ by setting the gradient of this error term to zero [688]:

$$\nabla E(\mathbf{v}) = \mathbf{A}\mathbf{v} + \mathbf{b} = \mathbf{0} \tag{12.7}$$

In this formulation, the matrix $\mathbf{A}$ is defined as:

$$\mathbf{A} = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \tag{12.8}$$

and vector $\mathbf{b}$ is given by:

$$\mathbf{b} = \begin{bmatrix} I_x I_t \\ I_y I_t \end{bmatrix} \tag{12.9}$$

This would give the following solution:

$$\mathbf{v} = -\mathbf{A}^{-1}\,\mathbf{b} \tag{12.10}$$

However, computing the inverse of $\mathbf{A}$ is problematic, as this matrix is singular and therefore has a determinant that is always zero. To allow optimization using this function, additional constraints can be added to the error term. The problem can for instance be regularized by adding a global smoothness constraint, as suggested by Horn and Schunck [341].

An alternative solution is to consider a neighborhood of pixels and select the velocity that is most consistent with the translations observed for all pixels within the patch. Given a patch with n pixels, the error term then becomes [475]:

$$E(\mathbf{v}) = \sum_{k=1}^{n} w_i\left(I_x(x_k,y_k,t)\,v_x + I_y(x_k,y_k,t)\,v_y + I_t(x_k,y_k,t)\right)^2 \tag{12.11}$$

Following the same procedure as above, the error term can be written as:

$$\nabla E(\mathbf{v}) = \sum_{i=1}^{n} w_i\,\left(\mathbf{A}_i\mathbf{v} + \mathbf{b}_i\right) \tag{12.12}$$

The solution to this problem is then:

$$\mathbf{v} = -\left(\sum_{i=1}^{n} w_i \mathbf{A}_i\right)^{-1} \sum_{i=1}^{n} w_i \mathbf{b}_i \tag{12.13}$$

Note that there is still a chance that the sum of matrices $\mathbf{A}_i$ is singular. In the following, we discuss a probabilistic method that would reduce the chances of creating a singular matrix.

12.4.1 Probabilistic Optical Flow

There are many sources of variation that cause spatiotemporal gradients in images. These include noise, quantization, lack of precision of the gradient operators (see Section 5.2 for a discussion), illumination variations, the possibility of multiple motions within a region, and areas of low contrast, as well as the aperture problem. As mentioned earlier, the latter is an issue that relates to both gradient-based algorithms as well as motion-sensitive neurons in the human visual system [769]. A single neuron receives input over a spatially localized receptive field (an aperture). Within this aperture, the direction of the motion is necessarily ambiguous (see Section 12.3.5 for more on this topic). There will be many patterns, directions, and speeds of movement that will elicit the same response within the neuron.

Despite all these sources of spatiotemporal variation, optical flow algorithms are intended to detect motion. Gradient-based algorithms will suffer from low accuracy whenever gradients are due to any of the other sources. To account for this, optical flow detection can be restated in terms of a probabilistic framework [688]. In this case, the probability of a vector field $\mathbf{v}$ given an observed image gradient $\nabla I = (I_x, I_y, I_t)$ can be constructed:

$$P(\mathbf{v} \ \nabla I) \tag{12.14}$$

The different types of uncertainty can be accounted for by the introduction of Gaussian noise terms bfn_1 and n_2, accounting effectively for spatial and temporal sources of uncertainty. Including these noise terms in to Equation (12.5), and denoting (I_x, I_y) as $\mathbf{I_s}$, we can write:

$$\mathbf{I}_s(\mathbf{v} - \mathbf{n}_1) + I_t = n_2 \tag{12.15}$$

This describes the conditional probability $P(I_t \ \mathbf{v}, \mathbf{I}_s)$. Using Bayes' rule (see the Appendix), we obtain:

$$P(\mathbf{v} \ \mathbf{I}_s, I_t) = \frac{P(I_t \ \mathbf{v}, \mathbf{I}_s)\, P(\mathbf{v})}{P(I_t)} \tag{12.16}$$

Simoncelli and colleagues choose a zero-mean Gaussian as the prior distribution of $P(I_t)$ [688] with covariance Λ_p. If the variance of $\mathbf{n}_1$ is set to (σ_1, σ_1) and the variance of n_2 is σ_2, then the covariance and mean of this distribution are given by:

$$\Lambda_{\mathbf{v}} = \left(\frac{\mathbf{A}}{\sigma_1 \;\; \mathbf{I}_s \;\; {}^2 + \sigma_2} + \Lambda_p^{-1} \right)^{-1} \tag{12.17}$$

$$\mu_{\mathbf{v}} = -\frac{\Lambda_{\mathbf{v}}\,\mathbf{b}}{\sigma_1 \;\; \mathbf{I}_s \;\; {}^2 + \sigma_2} \tag{12.18}$$

Since this distribution is Gaussian, the mean $\mu_{\mathbf{v}}$ is also the maximum *a posteriori* (MAP) estimate. This approach ensures that the matrix inversion is possible. This statistical approach can be extended to image patches, in which case the covariance and mean become:

$$\Lambda_{\mathbf{v}} = \left(\sum_{i=1}^{n} \frac{w_i\,\mathbf{A}}{\sigma_1 \;\; \mathbf{I}_s \;\; {}^2 + \sigma_2} + \Lambda_p^{-1} \right)^{-1} \tag{12.19}$$

$$\mu_{\mathbf{v}} = -\Lambda_{\mathbf{v}} \sum_{i=1}^{n} \frac{w_i\,\mathbf{b}}{\sigma_1 \;\; \mathbf{I}_s \;\; {}^2 + \sigma_2} \tag{12.20}$$

Finally, note that the division by $\sigma_1 \;\; \mathbf{I}_s \;\; {}^2 + \sigma_2$ acts as a nonlinear gain control, which has been shown to improve performance [688]. A similar normalization was discussed in the context of the reduction of dependencies between wavelet coefficients in Section 8.9.

12.4.2 Statistics of Real-World Motion and Retinal Flow

Optical flow in image sequences, as well as optical flow of retinal projections, can arise due to self-motion of the observer/camera or due to objects moving within the environment. Optical flow due to self-motion generates much information about the direction of movement and the structure of the environment, as well as the presence of possible objects and obstacles [261, 262]. The speed of the flow is also important. As an example, consider moving directly forward. In this case, the optical flow of the projected environment on the retina (or the camera sensor) will radiate outward from the point toward which the observer is traveling. If any object within this expanding field generates a stationary flow field, then that object is on a collision course with the observer/camera. A car coming out of a side street at a speed that would lead to a collision, for instance, would remain stationary on the retina.

It appears that such information is used by animals [441]. Further, there are specialized regions in the brain that process optical flow. These regions take their input from simple motion detectors earlier in the human visual system. It can be hypothesized that statistical regularities may play a role in the analysis taking place in these brain regions [102, 103]. This leads to two ways in which the statistics of optical flow can be assessed. The first is to directly study motion signals [370, 102, 632, 103], whereas the second is to pass motion signals through basic motion detectors first and then study the emerging statistical regularities [212, 383, 816].

To study motion signals directly, it is possible to derive flow fields from range databases [102, 103, 632], such as the Brown Range Image Database (see Section 3.3.13) [350, 447]. In this manner average velocities can be analyzed. As it turns out, looking forward in natural scenes causes the horizon to roughly separate images into upper and lower halves. On average, the distance between the viewer and the nearest point will be farther in the upper part of the image than in the lower part. As a consequence, egocentric motion will cause objects located in the upper half of the scene to move more slowly across the retina than the lower half, an effect which can be detected in optical flows derived from range data [102]. The variance in retinal speed is lower in the lower half of the visual field, indicating more uniform retinal speed, whereas this variance is higher in the upper half of the visual field. However, it appears that the distribution of retinal speed can be modeled well with a log-Gaussian distribution [103].

The distribution of retinal speed depends to a large extent on the distribution of depth values in a natural scene, showing a near-inverse relationship. However, this changes dramatically for scenes that are non-natural [103]. For natural scenes, the statistical relation between depth and retinal speed may help inform the human visual system about relative depth in the scene, especially in the lower half of the visual field.

The distribution of the directions of retinal flow is close to Gaussian for positions in the lower half of the visual field, whereas extreme non-Gaussian distributions occur near the horizontal median. In this region, the distribution becomes heavy-tailed, i.e., with a high kurtosis [103]. The direction of the optical flow as projected on the retina is strongly dependent on the direction of movement of the observer, especially in the lower half of the visual field.

Finally, it was found that the retinal speed and the direction of the optical flow are largely uncorrelated [103], especially after gaze stabilization. This allows the human visual system to encode direction and speed independently in cortical area MT [625]

12.4.3 Statistics of Optical Flow

Rather than estimate the optical flow as it would arise after projection onto the retina, it would be possible to directly assess the statistics of optical flow, which is for instance of interest in the application of video retrieval [202]. This is for instance possible by subjecting optical flow to principal components analysis [224] (discussed previously in Section 7.1). In such a model, motion is decomposed into a set of basis motions, which can then be ordered in order of prevalence. General motion can then be described with a small number of basis motions. In addition, the model coefficients can be derived directly from image derivatives, and there is therefore no requirement to compute dense image motion first. Such a model can be applied effectively to domain-specific problems. An example is the analysis and representation of the motion of mouths or human gait [224]. Moreover, it is possible to incorporate such models within a Bayesian inference framework to reliably estimate image motion in the vicinity of occlusion boundaries [223].

The statistics of optical flow are interesting also for the purpose of designing optical flow algorithms. As argued earlier in Section 12.4, gradient-based optical flow algorithms require a matrix inversion, which requires the matrix to be nonsingular. This has given rise to the formulation of smoothness constraints (see Roth and Black for a discussion [633]), formulating optical flow as a Bayesian inference problem using an appropriate posterior distribution. Using Bayes' rule, this can be split into a data term and a smoothness term (see Section 12.4.1).

Many smoothness constraints have been formulated, usually enforcing local smoothness by being based on nearest-neighbor differences or other local measures of flow gradient [633]. Roth and Black argue that such smoothness contraints can model piecewise smooth flow, but that they cannot not model more complex flows [633]. Global constraints, on the other hand, could account for more arbitrary flows. The overall insight appears to be that, dependent on the stimuli chosen, motion as captured in optical flow fields can be characterized as "slow and smooth" [791] or "mostly slow and smooth, but sometimes fast and discontinuous" [633].

Analyzing optical flow data derived from the Brown Range Image Database, Roth and Black [633] come to several conclusions regarding the velocity of optical flow:

- Vertical velocity follows a roughly Laplacian distribution.
- Horizontal velocity is more prevalent than vertical motion and therefore leads to a broader distribution.
- The heavy-tailed nature of motion is due to camera translations rather than rotations.

The distribution of flow orientations also tends to favor horizontal directions.

Spatial derivative histograms of optical flow show the characteristic central peak and long tails associate with sparseness and high kurtosis. This suggests that optical flow is often smooth, but on occasion it exhibits much stronger motions. Horizontal and vertical flow gradients are largely independent, as can be determined by analyzing joint statistics of horizontal and vertical flow gradients. This independence can be confirmed by analyzing the flow gradients with PCA [224, 633]. Temporal derivative histograms show similar heavy-tailed behavior, which can be attributed to the translational component of motion [633].

These statistical findings have been used to form priors for the calculation of optical flow with the aid of Markov random fields, and especially using the Fields of Experts model [633], which was discussed in Section 9.6.2.

Appendix A
Basic Definitions

In this appendix, we list a few basic definitions that are used throughout the book.

A.1 Probabilities and Bayes' Rule

To begin, the probability of an event A occurring is often denoted $p(A)$. The probability of an event A occurring, given that another event B has happened, is written as $p(A\ B)$. This is known as a conditional probability. Two events occurring simultaneously is given by:

$$p(A \quad B) = \frac{p(A\ B)\,p(B)}{p(B)} \tag{A.1}$$

If the conditional probability of event A occuring, given that event B has already occurred, and given the probabilities of A and B happening in isolation, we can compute the conditional probability of B occurring, given event A. This is known as Bayes' rule:

$$p(B\ A) = \frac{p(A\ B)\,p(B)}{p(A)} \tag{A.2}$$

For a set of N events B_i, if we have:

$$\bigcup_{i=1}^{N} B_i = \Omega \tag{A.3}$$

$$i = j : B_i \bigcup B_j = \emptyset \tag{A.4}$$

then we can compute the following probability:

$$p(A) = \sum_{i=1}^{N} p(A \quad B_i) \tag{A.5}$$

A.2 Gaussian Distribution

A univariate Gaussian probability distribution (also known as the univariate normal distribution) is given by:

$$N(x\ \mu,\sigma) = \frac{1}{\sigma\ \ \overline{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{A.6}$$

This can be extended to multiple dimensions d, leading to the multivariate Gaussian distribution:

$$N(\mathbf{x}\ \mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{\Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{A.7}$$

Here, μ is a d-dimensional vector of means and Σ is the covariance matrix.

A.3 Kullback-Leibler Divergence

The Kullback-Leibler divergence is an asymmetric measure of how different two probability distributions are. It is also known as KL-divergence, information divergence, information gain, or relative entropy. For two discrete probability distributions P and Q, this divergence is given by:

$$D_{\mathrm{KL}}(P\ \ Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i) \tag{A.8}$$

For continuous random variables p and q, the Kullback-Leibler divergence is given by:

$$D_{\mathrm{KL}}(p\ \ q) = \int_{-\infty}^{\infty} \ln\left(\frac{p(x)}{q(x)}\right) p(x)\,dx \tag{A.9}$$

Bibliography

[1] A. Abadpour and S. Kasaei. A fast and efficient fuzzy color transfer method. In *Proceedings of the 4th IEEE International Symposium on Signal Processing and Information Technology*, pages 491–494, 2004. 242

[2] A. Abadpour and S. Kasaei. An efficient PCA-based color transfer method. *Journal of Visual Communication and Image Representation*, 18(1):15–34, 2007. 242

[3] R. Acharyya. *A New Approach for Blind Source Separation of Convolutive Sources: Wavelet Based Separation Using Shrinkage Function*. VDM Verlag Dr. Müller e.K, Saarbrücken, Germany, 2008. 156

[4] J. Adams, K. Parulski, and K. Spaulding. Color processing in digital cameras. *IEEE Micro*, 18(6):20–30, 1998. 75, 251

[5] E. H. Adelson. Saturation and adaptation in the rod system. *Vision Research*, 22(10):1299–312, 1982. 22

[6] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, 1985. 271, 274

[7] K. M. Ahmad, K. Klug, S. Herr, P. Sterling, and S. Schein. Cell density ratios in a foveal patch in macaque retina. *Visual Neuroscience*, 20(2):189–209, 2003. 25

[8] P. K. Ahnelt. The photoreceptor mosaic. *Eye*, 12(3b):531–540, 1998. 24

[9] M. Al-Ayyou, M. T. Irfan, and D. G. Stork. Boosting multi-feature visual texture classifiers for the authentication of Jackson Pollock's drip paintings. In *SPIE: Computer Vision and Image Analysis of Art II*, volume 7869, page 1, 2011. 152

[10] A. Albiol, M. J. Silla, A. Albiol, and J. M. Mossi. Video analysis using corner motion statistics. In *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 31–38, 2009. 277

[11] S. R. Allred, A. Radonjić, A. L. Gilchrist, and D. H. Brainard. Lightness perception in high dynamic range images: Local and remote luminance effects. *Journal of Vision*, 12(2), 2012. 24

[12] L. Alvarez, Y. Gousseau, and J.-M. Morel. The size of objects in natural and artificial images. *Advances in Imaging and Electron Physics*, 111:167–242, 1999. 193, 259

[13] J. Alvarez-Ramirez, C. Ibarra-Valdez, E. Rodriguez, and L. Dagdug. $1/f$-noise structures in Pollocks's drip paintings. *Physica A: Statistical Mechanics and Its Applications*, 387(1):281–295, 2008. 152

[14] B. L. Anderson and J. Kim. Image statistics do not explain the perception of gloss and lightness. *Journal of Vision*, 9(11), 2009. 81

[15] G. J. Anderson and J. M. Cortese. 2-D contour perception resulting from kinetic occlusion. *Perception & Psychophysics*, 46(1):49–55, 1989. 278

[16] S. J. Anderson and D. C. Burr. Spatial and temporal selectivity of the human motion detection system. *Vision Research*, 125(8):1147–1154, 1985. 275

[17] T. J. Andrews, S. D. Halpern, and D. Purves. Correlated size variations in human visual cortex, lateral geniculate nucleus, and optic tract. *The Journal of Neuroscience*, 17(8):2859–2868, 1997. 3

[18] A. Angelucci and P. C. Bressloff. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate v1 neurons. *Progress in Brain Research*, 154:93–120, 2006. 109

[19] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding using the wavelet transform. *IEEE Transactions on Image Processing*, 1(2):205–220, 1992. 185, 203

[20] Anyhere Software. Photosphere. http://www.anyhere.com/. 46

[21] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of Applied Mathematics*, 9(1):17–29, 1951. 158

[22] N. Asada, A. Amano, and M. Baba. Photometric calibration of zoom lens systems. In *Proceedings of the 13th IEEE International Conference on Pattern Recognition*, volume 1, pages 186–190, 1996. 59

[23] H. F. Ates and M. T. Orchard. A nonlinear image representation in wavelet domain using complex signals with single quadrant spectrum. In *Proceedings of the Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1966–1970, 2003. 199

[24] J. J. Atick, P. A. Griffin, and A. N. Redlich. Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images. *Neural Computation*, 8(6):1321–1340, 1996. 164

[25] J. J. Atick and A. N. Redlich. Towards a theory of early visual processing. *Neural Computation*, 2(3):308–320, 1990. 111

[26] J. J. Atick and N. A. Redlich. What does the retina know about natural scenes? *Neural Computation*, 4(2):196–210, 1992. 4

[27] P. Axelsson. Processing of laser scanner data—Algorithms and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54(2):138–147, 1999. 51

[28] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002. 166

[29] R. J. Baddeley. An efficient code in V1? *Nature*, 381(6583):560–561, 1996. 163

[30] R. J. Baddeley, L. F. Abbott, C. A. Michael, C. A. Booth, F. Sengpiel, T. Freeman, E. A. Wakeman, and E. T. Rolls. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society B*, 264(1389):1775–1783, 1997. 4, 140, 146

[31] V. Balasubramanian and P. Sterling. Receptive fields and functional architecture in the retina. *Journal of Physiology*, 587(12):2753–2767, 2009. 113

[32] R. M. Balboa and N. M. Grzywacz. Occlusions and their relationship with the distribution of contrasts in natural images. *Vision Research*, 40(19):2661–2669, 2000. 97

[33] R. M. Balboa and N. M. Grzywacz. Power spectra and distribution of contrasts of natural images from different habitats. *Vision Research*, 43(24):2527–2537, 2003. 146

[34] H. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241–253, 2001. 12

[35] H. Barlow and P. Földiák. Adaptation and decorrelation in the cortex. In G. Mitchinson, editor, *The Computing Neuron*, chapter 4, pages 54–72. Addison-Wesley Longman Publishing Co., Inc., New York, 1989. 24

[36] H. B. Barlow. Possible principles underlying the transformations of sensory images. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge, MA, 1961. 111

[37] H. B. Barlow. A theory about the functional role and synaptic mechanism of visual after-effects. In C. Blakemore, editor, *Vision: Coding and Efficiency*, pages 363–375. Cambridge University Press, Cambridge, UK, 1990. 24

[38] P. Barone, A. Batardiere, K. Knoblauch, and H. Kennedy. Laminar distribution of neurons in extrastriate areas projecting to visual areas V1 and V4 correlates with the hierarchical rank and indicates the operation of a distance rule. *Journal of Neuroscience*, 20(9):3263–3281, 2000. 29

[39] E. B. Baum, J. Moody, and F. Wilczek. Internal representations for associative memory. *Biological Cybernetics*, 59(4–5):217–228, 1988. 163

[40] B. E. Bayer. Color imaging array, 1976. U.S. Patent No 3,971,065. 49

[41] J. Beck and S. Prazdny. Highlights and the perception of glossiness. *Attention, Perception, & Psychophysics*, 30(4):407–410, 1981. 81

[42] E. D. Becker and T. Farrar. Fourier transform spectroscopy. *Science*, 178(4059):361–368, 1972. 50

[43] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):217–234, 1995. 166

[44] A. J. Bell and T. J. Sejnowski. The independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997. 171

[45] R. J. Bell. *Introductory Fourier Transform Spectroscopy*. Academic Press, New York, 1972. 50

[46] B. Belzer, J. M. Lina, and J. Villasenor. Complex, linear-phase filters for efficient image coding. *IEEE Transactions on Signal Processing*, 40(4):2425–2427, 1995. 199

[47] A. Bennett and I. Cuthill. Avian color vision and coloration: Multidisciplinary evolutionary biology. *American Naturalist*, 169(S1):1–6, 2007. 228

[48] J.-A. Beraldin, F. Blais, L. Cournoyer, M. Rioux, S. El-Hakim, R. Rodella, F. Bernier, and N. Harrison. Digital 3D imaging system for rapid response on remote sites. In *Proceedings of the 2nd International Conference on 3-D Digital Imaging and Modeling*, pages 34–43, 1999. 52

[49] D. F. Berman, J. T. Bartell, and D. H. Salesin. Multiresolution painting and compositing. In *SIGGRAPH 94: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Systems*, pages 85–90, Orlando, FL, 1994. 203

[50] F. Bernardini and H. Rushmeier. The 3D model acquisition pipeline. *Computer Graphics Forum*, 21(2):149–172, 2002. 51

[51] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *SIGGRAPH 00: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 417–424, 2000. 116, 118

[52] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing*, 12(8):882–889, 2003. 118

[53] O. Bertrand, J. Bohorquez, and J. Pernier. Time-frequency digital filtering based on an invertible wavelet transform: An application to evoked potentials. *IEEE Transactions on Biomedical Engineering*, 41(1):77–88, 1994. 202

[54] P. J. Bex, S. G. Solomon, and S. C. Dakin. Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure. *Journal of Vision*, 9(10):1–19, 2009. 32

[55] G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms I. *Comunications on Pure and Applied Mathematics*, 44(2):141–183, 1991. 186

[56] D. Beymer and T. Poggio. Face recognition from one example view. In *Proceedings of the 5th IEEE International Conference on Computer Vision*, pages 500–507, 1995. 163

[57] V. A. Billock. Neural acclimation to $1/f$ spatial frequency spectra in natural images transduced by the human visual system. *Physica D*, 137(3):379–391, 2000. 144, 145

[58] V. A. Billock, D. W. Cunningham, P. R. Havig, and B. H. Tsou. Perception of spatiotemporal random fractals: An extension of colorimetric methods to the study of dynamic texture. *Journal of the Optical Society of America A*, 18(10):2404–2413, 2001. 144, 146

[59] V. A. Billock, D. W. Cunningham, and B. H. Tsou. What visual discrimination of fractal textures can tell us about discrimination of camouflaged targets. In D. H. Andrews, R. P. Herz, and M. B. Wolf, editors, *Human Factors Issues in Combat Identification*, pages 99–112. Ashgate, 2008. 132, 140, 144, 145

[60] V. A. Billock, G. C. D. Guzman, and J. A. S. Kelso. Fractal time and $1/f$ spectra in dynamic images and human vision. *Physica D*, 148(1):136–146, 2001. 140, 272

[61] V. A. Billock and T. H. Harding. Evidence of a colour appearance model for colour management systems spatial and temporal channels in the correlational structure of human spatiotemporal contrast sensitivity. *Journal of Physiology*, 490(2):509–517, 1996. 144

[62] M. J. Black, D. J. Fleet, and Y. Yacoob. A framework for modeling appearance change in image sequences. In *Proceedings of the 6th IEEE International Conference on Computer Vision*, pages 660–667, 1998. 276

[63] M. J. Black, D. J. Fleet, and Y. Yacoob. Robustly estimating changes in image appearance. *Computer Vision and Image Understanding*, 78(1):8–31, 2000. 276

[64] A. Blake, P. Kohli, and C. Rother. *Markov Random Fields for Vision and Image Processing*. MIT Press, Cambridge, MA, 2011. 208

[65] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH 99: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 164

[66] C. Bloch. *The HDRI Handbook 2.0: High Dynamic Range Imaging for Photographers and CG Artists*. Rocky Nook Inc., Santa Barbara, CA, 2012. 45, 46, 47, 48

[67] T. Bonhoeffer and A. Grinvald. Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature*, 353(6343):429–431, 1991. 29

[68] J. K. Bowmaker and H. J. Dartnall. Visual pigments of rods and cones in a human retina. *Journal of Physiology*, 298(1):501–511, 1980. 21, 229

[69] I. Boyadzhiev, K. Bala, S. Paris, and F. Durand. User-guided white balance for mixed lighting conditions. *ACM Transactions on Graphics*, 31(6):200, 2012. 10

[70] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *Proceedings of the 8th IEEE International Conference on Computer Vision*, volume 1, pages 105–112, 2001. 217

[71] R. M. Boynton. *Human Color Vision*. Holt, Rinehart, and Winston, New York, 1979. 33

[72] O. Braddick. Motion perception. In E. B. Goldstein, editor, *Encyclopedia of Perception*, pages 572–578. SAGE Publications, Inc., Thousand Oaks, CA, 2010. 269

[73] D. R. Bradley and K. Lee. Animated subjective contours. *Perception & Psychophysics*, 32(4):393–395, 1982. 278

[74] M. Brady and D. J. Field. Local contrast in natural images: Normalisation and coding efficiency. *Perception*, 29(9):1041–1055, 2000. 78

[75] M. Brady and G. Legge. Camera calibration for natural image studies and vision research. *Journal of the Optical Society of America A*, 26(1):30–42, 2009. 55, 57, 58, 59

[76] N. Brady, P. J. Bex, and R. E. Fredericksen. Independent coding across spatial scales in moving fractal images. *Vision Research*, 37(14):1873–1883, 1997. 147

[77] D. H. Brainard, A. Roorda, Y. Yamauchi, J. B. Calerone, A. Metha, M. Neitz, J. Neitz, D. R. Williams, and G. H. Jacobs. Functional consequences of the relative numbers of L and M cones. *Journal of the Optical Society of America A*, 17(3):607–614, 2000. 24

[78] G. J. Brelstaff, A. Parraga, T. Troscianko, and D. Carr. Hyperspectral camera system: acquisition and analysis. In *Satellite Remote Sensing II*, pages 150–159. International Society for Optics and Photonics, 1995. 50, 66

[79] K. Brodmann. *Vergleichende Localisationslehre der Großhirnrhinde*. Barth Verlag, Leipzig, 1909. Translated by L. J. Gaery, *Localisation in the Cerebral Cortex*, London, 1994. 29

[80] M. Brown and D. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007. 48

[81] N. Bruno and M. Bertamini. Identifying contours from occlusion events. *Perception & Psychophysics*, 48(4):331–342, 1990. 278

[82] N. Bruno, M. Bertamini, and F. Domini. Amodal completion of partly occluded surfaces: Is there a mosaic stage? *Journal of Experimental Psychology: Human Perception and Performance*, 23(5):1412–1426, 1997. 278

[83] N. Bruno and W. Gerbino. Illusory figures based on local kinematics. *Perception*, 20(2):259–274, 1991. 278

[84] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 60–65, 2005. 99

[85] R. W. Buccigrossi and E. P. Simoncelli. Progressive wavelet image coding based on a conditional probability model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 2957–2960, Munich, Germany, 1997. 191, 193

[86] R. W. Buccigrossi and E. P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–1701, 1999. 197

[87] G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1):1–26, 1980. 250, 252

[88] G. Buchsbaum and A. Gottschalk. Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London B*, 220(1218):89–113, 1983. 231

[89] J. Burge, C. C. Fowlkes, and M. S. Banks. Natural-scene statistics predict how the figure–ground cue of convexity affects human depth perception. *The Journal of Neuroscience*, 30(21):7269–7280, 2010. 262

[90] J. Burge and W. S. Geisler. Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences*, 108(40):16849–16854, 2011. 61

[91] C. J. C. Burges. Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning*, 2(4):275–365, 2009. 153

[92] D. C. Burr. Motion perception, elementary mechanisms. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA, 2003. 275

[93] D. C. Burr and J. Ross. Contrast sensitivity at high velocities. *Vision Research*, 22(4):479–484, 1982. 275

[94] D. C. Burr and J. Ross. Visual processing of motion. *Trends in Neuroscience*, 9:304–306, 1986. 275, 277

[95] D. C. Burr and P. Thompson. Motion psychophysics: 1985–2010. *Vision Research*, 51(13):1431–1456, 2011. 274

[96] C. S. Burrus, R. A. Gopinath, and H. Guo. *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice-Hall, Upper Saddle River, NJ, 1998. 185

[97] G. J. Burton and I. R. Moorhead. Color and spatial structure in natural scenes. *Applied Optics*, 26(1):157–170, 1987. 133, 135

[98] G. F. Byrne, P. F. Crapper, and K. K. Mayo. Monitoring land-cover change by principal component analysis of multitemporal landsat data. *Remote Sensing of Environment*, 10(3):175–184, 1980. 270

[99] M. Cadík, M. Wimmer, L. Neumann, and A. Artusi. Evaluation of HDR tone mapping methods using essential perceptual attributes. *Computers & Graphics*, 32(3):330–349, 2008. 84

[100] T. Caelli. *Visual Perception: Theory and Practice*. Pergamon Press, Oxford, 1981. 119, 145

[101] A. J. Calder and A. W. Young. Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, 6(8):641–651, 2005. 164

[102] D. Calow, N. Krüger, F. Wörgötter, and M. Lappe. Statistics of optic flow for self-motion through natural scenes. *Dynamic Perception*, pages 133–138, 2004. 282, 283

[103] D. Calow and M. Lappe. Local statistics of retinal optic flow for self-motion through natural sceneries. *Network: Computation in Neural Systems*, 18(4):343–374, 2007. 282, 283

[104] F. W. Campbell and D. G. Green. Optical and retinal factors affecting visual resolution. *Journal of Physiology*, 181(3):576–593, 1965. 30

[105] F. W. Campbell, E. R. Howell, and J. R. Johnson. A comparison of threshold and suprathreshold appearance of gratings with components in the low and high spatial frequency range. *Journal of Physiology*, 284(1):193–201, 1978. 145

[106] F. W. Campbell and J. G. Robson. Application of Fourier analysis to the visibility of gratings. *Journal of Physiology*, 197(3):551–566, 1968. 32

[107] E. Candés. *Ridgelets: Theory and Applications*. PhD thesis, Department of Statistics, Stanford University, 1998. 190

[108] E. Candés and D. L. Donoho. Ridgelets: The key to high-dimensional intermittency. *Philosophical Transactions of the Royal Society A*, 357(1760):2495–2509, 1999. 190

[109] E. Candés and D. L. Donoho. Curvelets: A surprisingly effective nonadaptive representation of objects with edges. In A. Cohen, C. Rabut, and L. L. Schumaker, editors, *Curve and Surface Fitting: Saint Malo 1999*. Vanderbilt University Press, Nashville, TN, 2000. 190

[110] D. Cano and T. H. Minh. Texture synthesis using hierarchical linear transforms. *Signal Processing*, 15(2):131–148, 1988. 203

[111] F. Canters and H. Decleir. *The World in Perspective: A Directory of World Map Projections*. John Wiley & Sons, New York, 1989. 49

[112] J. N. Caron, N. M. Namazi, and C. J. Rollins. Noniterative blind data restoration by use of an extracted filter function. *Applied Optics*, 41(32):6884–6889, 2002. 115

[113] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002. 173

[114] A. Casile and M. A. Giese. Critical features for the recognition of biological motion. *Journal of Vision*, 5(4):348–360, 2005. 277

[115] A. Certain, J. Popovįć, T. DeRose, T. Duchamp, D. Salesin, and W. Stuetzle. Interactive multiresolution surface viewing. In *SIGGRAPH 96: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pages 91–98, New Orleans, 1996. 203

[116] A. Chakrabarti and T. Zickler. Statistics of real-world hyperspectral images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 193–200, 2011. 50, 65

[117] T. Chan and J. Shen. Local inpainting models and TV inpainting. *SIAM Journal on Applied Mathematics*, 62(3):1019–1043, 2001. 118

[118] S. E. Chen. QuickTime VR: An image-based approach to virtual environment navigation. In *SIGGRAPH 95: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pages 29–38, New York, 1995. ACM. 47

[119] M. E. Chevreul. *The Principles of Harmony and Contrast of Colors and Their Applications to the Arts*. Schiffer, West Chester, PA, 1987. 35

[120] H. Choi, J. Romberg, R. G. Baraniuk, and N. Kingsbury. Hidden Markov tree modeling of complex wavelet transforms. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 133–136, 2000. 198, 199

[121] C. M. Cicerone and D. D. Hoffman. Color from motion: Dichoptic activation and a possible role in breaking camouflage. *Perception*, 26(11):1367–1380, 1997. 278

[122] C. M. Cicerone, D. D. Hoffman, P. D. Gowdy, and J. S. Kim. The perception of color from motion. *Perception & Psychophysics*, 57(6):761–777, 1995. 278

[123] J. Clark, C. Zhang, and A. Wallace. Image aquisition using fixed and variable triangulation. In *Proceedings of the 5th IET International Conference on Image Processing and Its Applications*, pages 539–543, 1995. 52

[124] D. Claus and A. Fitzgibbon. A rational function lens distortion model for general cameras. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 213–219, 2005. 58

[125] C. W. G. Clifford and M. R. Ibbotson. Fundamental mechanisms of visual motion detection: Models, cells and functions. *Progress in Neurobiology*, 68(6):409–437, 2003. 274, 275

[126] J. Coddington, J. Elton, D. Rockmore, and Y. Wang. Multifractal analysis and authentication of Jackson Pollock paintings. In D. G. Stork and J. Coddington, editors, *Proceedings of the SPIE: Computer Image Analysis in the Study of Art*, volume 6810, page 68100F, 2008. 152

[127] P. Comon. Independent component analysis: A new concept? *Signal Processing*, 36(3):287–314, 1994. 166

[128] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press / Elsevier, Oxford, UK, 2010. 166

[129] B. R. Conway, S. Chatterjee, G. D. Field, G. D. Horwitz, E. N. Johnson, K. Koida, and K. Mancuso. Advances in color science: From retina to behavior. *Journal of Neuroscience*, 30(45):14955–14963, 2010. 26

[130] R. L. Cook, L. Carpenter, and E. Catmull. The Reyes image rendering architecture. In *SIGGRAPH 87: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, pages 95–102, 1987. 148

[131] T. Cooke, D. W. Cunningham, and C. Wallraven. Local processing in spatiotemporal boundary formation. In *Proceedings of the 7th Tübingen Perception Conference*, page 65, 2004. 278

[132] E. O. Cormack and R. H. Cormack. Stimulus configuration and line orientation in the horizontal-vertical illusion. *Attention, Perception, & Psychophysics*, 16(2):208–212, 1974. 261

[133] T. Cornsweet. *Visual perception*. Academic Press, New York, 1970. 36, 145

[134] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004. xviii, 116, 117, 118

[135] A. Criminisi and D. G. Stork. Did the great masters use optical projections while painting? Perspective comparison of paintings and photographs of renaissance chandeliers. In *Proceedings of the IEEE International Conference on Pattern Recognition*, volume 4, pages 645–648, 2004. 84

[136] G. R. Cross and A. K. Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(1):25–39, 1983. 203

[137] N. A. Crowder, M. Dawson, and D. Wylie. Temporal frequency and velocity-like tuning in the pigeon accessory optic system. *Journal of Neurophysiology*, 90(3):1829–1841, 2003. 275

[138] D. Cunningham and C. Wallraven. *Experimental Design: From User Studies to Psychophysics*. AK Peters, Natick, MA, 2011. 4, 6, 17, 42, 43

[139] D. W. Cunningham, A. B. A. Graf, and H. H. Bülthoff. A relative encoding approach to modeling spatiotemporal boundary formation. *Journal of Vision*, 2(7):704, 2002. 278

[140] D. W. Cunningham, T. F. Shipley, and P. J. Kellman. The dynamic specification of surfaces and boundaries. *Perception*, 27(4):403–416, 1998. 278

[141] D. W. Cunningham, T. F. Shipley, and P. J. Kellman. Interactions between spatial and spatiotemporal information in spatiotemporal boundary formation. *Perception & Psychophysics*, 60(5):839–851, 1998. 278

[142] B. Curless. From range scans to 3D models. In *SIGGRAPH 99: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 38–41. ACM, 1999. 51, 52

[143] I. Cuthill, J. Partridge, A. Bennett, S. Church, N. Hart, and S. Hunt. Ultraviolet vision in birds. *Advances in the Study of Behavior*, 29:159–214, 2000. 49

[144] J. Cutting and J. Garvin. Fractal curves and complexity. *Perception and Psychophysics*, 42:365–370, 1987. 136, 146

[145] J. E. Cutting, D. R. Proffitt, L. T. Kozlowski, et al. A biomechanical invariant for gait perception. *Journal of Experimental Psychology: Human Perception and Performance*, 4(3):357–372, 1978. 277

[146] F. Cutzu, R. I. Hammoud, and A. Leykin. Distinguishing paintings from photographs. *Computer Vision and Image Understanding*, 100(3):249–273, 2005. 84

[147] D. M. Dacey and B. B. Lee. Cone inputs to the receptive field of midget ganglion cells in the periphery of the macaque retina. *Investigative Ophthalmology and Visual Science*, 38:S708, 1997. 29

[148] D. M. Dacey and B. B. Lee. Functional architecture of cone signal pathways in the primate retina. In K. R. Gegenfurtner and L. Sharpe, editors, *Color Vision: From Genes to Perception*, pages 181–202. Cambridge University Press, Cambridge, UK, 1999. 29

[149] D. M. Dacey, B. B. Peterson, F. R. Robinson, and P. D. Gamlin. Fireworks in the primate retina: In vitro photodynamics reveals diverse LGN-projecting ganglion cell types. *Neuron*, 37(1):15–27, 2003. 26

[150] S. C. Dakin and P. J. Bex. Natural image statistics mediate brightness filling in. *Proceedings of the Royal Society of London, B*, 270(1531):2341–2348, 2003. 37, 112, 145

[151] Y. Dan, J. J. Atick, and R. C. Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *Journal of Neuroscience*, 16(10):3351–3362, 1996. 4

[152] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, 1990. 185, 199

[153] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, 1992. 176, 186

[154] J. G. Daugman. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Transactions on Biomedical Engineering*, 36(1):107–114, 1989. 4

[155] J. Davis and X. Chen. A laser range scanner designed for minimum calibration complexity. In *Proceedings of the 3rd International Conference on 3-D Digital Imaging and Modeling*, pages 91–98, 2001. 52

[156] J. Davis, S. R. Marschner, M. Garr, and M. Levoy. Filling holes in complex surfaces using volumetric diffusion. In *Proceedings of the 1st IEEE International Symposium on 3D Data Processing Visualization and Transmission*, pages 428–441, 2002. 53

[157] J. S. De Bonet and P. A. Viola. A non-parametric multi-scale statistical model for natural images. In *NIPS-11: Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems*, pages 773–779, 1998. 9, 203

[158] P. De Graef, D. Christiaens, and G. d'Ydewalle. Perceptual effects of scene context on object identification. *Psychological Research*, 52(4):317–329, 1990. 267

[159] R. L. De Valois and K. K. De Valois. A multi-stage color model. *Vision Research*, 33(8):1053–1065, 1993. 27

[160] R. L. De Valois and K. K. De Valois. On a three-stage color model. *Vision Research*, 36(6):833–836, 1996. 27

[161] R. L. De Valois, K. K. De Valois, E. Switkes, and L. Mahon. Hue scaling of isoluminant and cone-specific lights. *Vision Research*, 37(7):885–897, 2000. 27

[162] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Receptive-field dynamics in the central visual pathways. *Trends in Neurosciences*, 18(10):451–458, 1995. 29, 273

[163] P. Debevec. HDRShop. http://www.hdrshop.com/. 46

[164] P. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH 97: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, pages 369–378, New York, 1997. ACM. 44, 45, 46, 277

[165] P. B. Delahunt and D. H. Brainard. Does human color constancy incorporate the statistical regularity of natural daylight? *Journal of Vision*, 4(2), 2004. 250

[166] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977. 172

[167] H. Derin and H. Elliott. Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):39–55, 1987. 203

[168] N. G. Deriugin. The power spectrum and the correlation function of the television signal. *Telecomunications*, 1(7):1–12, 1956. 4

[169] O. Deussen, C. Colditz, L. Coconu, and H. Hege. Efficient modeling and rendering of landscapes. In I. Bishop and E. Lange, editors, *Visualization in Landscape and Environmental Planning*. Taylor & Francis, Oxon, UK, 2005. 145

[170] O. Deussen and B. Lintermann. *Digital Design of Nature: Computer Generated Plants and Organics*. Springer-Verlag, New York, 2005. 145, 148

[171] R. A. DeVore and B. J. Lucier. Fast wavelet techniques for near-optimal image processing. In *Proceedings of the IEEE Military Communications Conference*, pages 48.3.1–48.3.7, New York, 1992. 202

[172] M. N. Do and M. Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106, 2005. 190

[173] K. R. Dobkins. Moving colors in the lime light. *Neuron*, 25(1):15–18, 2000. 29

[174] D. W. Dong and J. J. Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6(3):345–358, 1995. 133, 135, 140, 271, 272

[175] D. W. Dong and J. J. Atick. Temporal decorrelation: A theory of lagged and non-lagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6(2):159–178, 1995. 29, 272

[176] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995. 202

[177] D. L. Donoho and A. Flesia. Can recent innovations in harmonic analysis “explain” key findings in natural image statistics? *Network: Computation in Neural Systems*, 12(3):371–393, 2001. 193

[178] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. 202

[179] J. E. Dowling. *The Retina: An Approachable Part of the Brain*. Belknap Press, Cambridge, MA, 1987. 22, 24

[180] J. E. Dowling and H. Ripps. Adaptation in skate photoreceptors. *Journal of General Physiology*, 60(6):698–719, 1972. 22

[181] R. O. Dror, E. H. Adelson, and A. S. Willsky. Surface reflectance estimation and natural illumination statistics. In *Proceedings of the IEEE Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, 2001. 146

[182] R. O. Dror, T. K. Leung, E. H. Adelson, and A. S. Willsky. Statistics of real-world illumination. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 164–171, Kauai, HI, 2001. 49, 79, 97

[183] R. O. Dror, D. C. O’Carroll, and S. B. Laughlin. The role of natural image statistics in biological motion estimation. In *Proceedings of the IEEE Workshop on Biologically Motivated Computer Vision*, pages 492–501, Seoul, Korea, 2000. 146, 274

[184] R. O. Dror, D. C. O’Carroll, and S. B. Laughlin. Accuracy of velocity estimation by Reichardt correlators. *Journal of the Optical Society of America A*, 18(2):241–252, 2001. 146

[185] R. O. Dror, A. S. Willsky, and E. H. Adelson. Statistical characterization of real-world illumination. *Journal of Vision*, 4(9):821–837, 2004. 49, 79, 97, 133, 135

[186] H. Du, X. Tong, X. Cao, and S. Lin. A prism-based system for multispectral video acquisition. In *Proceedings of the 12th IEEE International Conference on Computer Vision*, pages 175–182, 2009. 50

[187] E. Dubois and S. Sabri. Noise reduction in image sequences using motion-compensated temporal filtering. *IEEE Transactions on Communications*, 32(7):826–831, 1984. 275

[188] DxO Image Science. Automated optical corrections dedicated to your lenses. http://www.dxo.com/us/photo/dxo_optics_pro/optics_geometry_corrections. 56

[189] J. W. Earl and N. G. Kingsbury. Spread transform watermarking for video sources. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 491–494, 2001. 200

[190] F. Ebner and M. D. Fairchild. Development and testing of a color space (IPT) with improved hue uniformity. In *Proceedings of the 6th IS&T/SID Color Imaging Conference: Color Science, Systems and Applications*, pages 8–13, 1998. 234

[191] M. Ebner. *Color Constancy*. John Wiley & Sons, West Sussex, England, 2007. 249

[192] M. P. Eckert, G. Buchsbaum, and A. B. Watson. Separability of spatiotemporal spectra of image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(12):1210–1213, 1992. 140, 270

[193] F. Y. Edgeworth. On the probably errors of frequency-constants. *Journal of the Royal Statistical Society*, 71:381–397, 499–512, 651–678, 72:81–90, 1908. 169

[194] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *SIGGRAPH 01: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 341–346, 2001. 118

[195] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the 7th IEEE International Conference on Computer Vision*, volume 2, pages 1033–1038, 1999. 9, 118

[196] M. El-Melegy and A. Farag. Nonmetric lens distortion calibration: Closed-form solutions, robust estimation and model selection. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, pages 554–559, 2003. 57

[197] J. H. Elder. Are edges incomplete? *International Journal of Computer Vision*, 34(2/3):97–122, 1999. 36, 112

[198] ePaperPress. PTLens. http://epaperpress.com/ptlens/. 56

[199] E. Erwin, F. H. Baker, W. F. Busen, and J. G. Malpeli. Relationship between laminar topology and retinotopy in the rhesus lateral geniculate nucleus: Results from a functional atlas. *Journal of Computational Neurology*, 407(1):92–102, 1999. 29

[200] D. C. van Essen, C. H. Anderson, and D. J. Felleman. Information processing in the primate visual system: An integrated systems perspective. *Science*, 255(5043):419–423, 1992. 17, 29

[201] S. Exner. *Entwurf zu Einer Physiologischen Erklärung der Psychischen Erscheinungen. I. Teil [Draft of a Physiological Explanation of Mental Impressions. Part I.]*. Deutike, Leipzig, Wein, 1894. 274

[202] R. Fablet and P. Bouthemy. Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1619–1624, 2003. 283

[203] M. D. Fairchild. The HDR photographic survey. In *Proceedings of the 15th IS&T/SID Color Imaging Conference*, volume 15, pages 233–238. The Society for Imaging Science and Technology, 2007. 62, 63, 79, 81

[204] M. D. Fairchild. *Color Appearance Models*. John Wiley & Sons, Chichester, UK, 3rd edition, 2013. 229

[205] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Transactions on Graphics*, 27(3):67, 2008. 106

[206] R. Fattal. Image upsampling via imposed edge statistics. *ACM Transactions on Graphics*, 26(3):95, 2007. 116

[207] R. Fattal. Edge-avoiding wavelets and their applications. *ACM Transactions on Graphics*, 28(3):22, 2009. 106

[208] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *Proceedings of the Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004. 67

[209] M. Felsberg, S. Kalkan, and N. Krüger. Continuous dimensionality characterization of image structures. *Image and Vision Computing*, 27(6):628–636, 2009. 264

[210] M. Felsberg and N. Krüger. A probabilistic definition of intrinsic dimensionality for images. In *Pattern Recognition*, pages 140–147. Springer, 2003. 264, 265, 266

[211] R. Fergus, B. Singh, A. Hertzmann, S. Roweis, and W. Freeman. Removing camera shake from a single photograph. *ACM Transactions on Graphics*, 25(3):787–794, 2006. 6, 7, 115

[212] C. Fermüller, D. Shulman, and Y. Aloimonos. The statistics of optical flow. *Computer Vision and Image Understanding*, 82(1):1–32, 2001. 283

[213] F. Fernandez, M. Wakin, and R. G. Baraniuk. Non-redundant, linear-phase, semi-orthogonal, directional complex wavelets. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 953–956, 2004. 199

[214] J. A. Ferwerda, S. Pattanaik, P. Shirley, and D. P. Greenberg. A model of visual adaptation for realistic image synthesis. In *SIGGRAPH 96: Proceedings of the 23th Annual Conference on Computer Graphics and Interactive Techniques*, pages 249–258, 1996. 20

[215] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987. 4, 97, 133, 135, 190

[216] D. J. Field. Scale-invariance and self-similar "wavelet" transforms: An analysis of natural scenes and mammalian visual systems. In M. Farge, J. C. R. Hunt, and J. C. Vassilicos, editors, *Wavelets, Fractals and Fourier Transforms*, pages 151–193. Clarendon Press, Oxford, UK, 1993. 133, 134, 135, 200, 202

[217] D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, 1994. 163

[218] D. J. Field and N. Brady. Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes. *Vision Research*, 37(23):3367–3383, 1997. 133, 135, 136, 146

[219] G. Finlayson and E. Trezzi. Shades of grey and colour constancy. In *Proceedings of the 12th IS&T/SID Color Imaging Conference*, pages 37–41, 2004. 251, 252

[220] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222:309–368, 1922. 169

[221] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–199, 1936. 156

[222] R. A. Fisher. On an absolute criterion for fitting frequency curves. *Statistical Science*, 12(1):39–41, 1997. 169

[223] D. J. Fleet, M. J. Black, and O. Nestares. Bayesian inference of visual motion boundaries. In G. Lakemeyer and B. Nebel, editors, *Exploring Artificial Intelligence in the New Millennium*, pages 139–174. Morgan Kaufmann, San Francisco, 2002. 284

[224] D. J. Fleet, M. J. Black, Y. Yacoob, and A. D. Jepson. Design and use of linear models for image motion analysis. *International Journal of Computer Vision*, 36(3):171–193, 2000. 283, 284

[225] R. W. Fleming, R. O. Dror, and E. H. Adelson. Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, 3(5), 2003. 81

[226] R. W. Fleming, H. W. Jensen, and H. H. Bülthoff. Perceiving translucent materials. In *Proceedings of the First ACM Symposium on Applied Perception in Graphics and Visualization*, pages 127–134, 2004. 81

[227] R. W. Fleming, A. Torralba, and E. H. Adelson. Specular reflections and the perception of shape. *Journal of Vision*, 4(9):798–820, 2004. 81

[228] J. Flusser, J. Kautsky, and F. Šroubek. Implicit moment invariants. *International Journal of Computer Vision*, 86(1):72–86, 2010. 75

[229] J. Flusser and T. Suk. Pattern recognition by affine moment invariants. *Pattern Recognition*, 26(1):167–174, 1993. 75

[230] J. Flusser and T. Suk. Affine moment invariants: A new tool for character recognition. *Pattern Recognition Letters*, 15(4):433–436, 1994. 75

[231] J. Flusser and T. Suk. Degraded image analysis: An invariant approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):590–603, 1998. 75

[232] J. Flusser, T. Suk, and S. Saic. Recognition of blurred images by the method of moments. *IEEE Transactions on Image Processing*, 5(3):533–538, 1996. 75

[233] J. Flusser, B. Zitova, and T. Suk. *Moments and Moment Invariants in Pattern Recognition*. Wiley, 2009. 74, 75, 76, 84

[234] P. Földiák. Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64(2):165–170, 1990. 4

[235] P. Földiák and M. P. Young. Sparse coding in the primate cortex. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 895–989. MIT Press, Cambridge, MA, 1995. 163

[236] J. Forrester, A. Dick, P. McMenamin, and W. Lee. *The Eye: Basic Sciences in Practice*. W B Saunders, London, 2001. 20

[237] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, 2nd edition, 2011. 102

[238] D. H. Foster. Color constancy. *Vision Research*, 51(7):674–700, 2011. 248, 249

[239] D. H. Foster, K. Amano, S. Nascimento, and M. J. Foster. Frequency of metamerism in natural scenes. *Journal of the Optical Society of America A*, 23(10):2359–2372, 2006. 64, 229

[240] D. H. Foster, S. M. C. Nascimento, and K. Amano. Information limits on neural identification of colored surfaces in natural scenes. *Visual Neuroscience*, 21(3):331–336, 2004. 64

[241] A. Fournier, D. Fussell, and L. Carpenter. Computer rendering of stochastic models. *Communications of the ACM*, 25(6):371–384, 1982. 117, 148

[242] J. G. F. Francis. The QR transformation, I. *The Computer Journal*, 4(3):265–271, 1961. 158

[243] J. G. F. Francis. The QR transformation, II. *The Computer Journal*, 4(4):332–345, 1962. 158

[244] R. A. Frazor and W. S. Geisler. Local luminance and contrast in natural images. *Vision Research*, 46(10):1585–1598, 2006. 114

[245] G. Freedman and R. Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics*, 30(2):12, 2011. 99

[246] B. Funt, K. Barnard, and L. Martin. Is machine colour constancy good enough? In *Proceedings of the 5th European Conference on Computer Vision*, pages 445–459, Freiburg, Germany, 1998. 251

[247] A. Fusiello, E. Trucco, T. Tommasini, and V. Roberto. Improving feature tracking with robust statistics. *Pattern Analysis & Applications*, 2(4):312–320, 1999. 275

[248] D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers (JIEE)*, 93(III):429–457, 1946. 198

[249] Q. Gao and S. Roth. How well do filter-based MRFs model natural images? In A. Pinz, T. Pock, H. Bischof, and F. Leberl, editors, *Pattern Recognition*, volume 7476 of *Lecture Notes in Computer Science*, pages 62–72. Springer, Berlin, 2012. 219, 221

[250] J. Gårding and T. Lindeberg. Direct estimation of local surface shape in a fixating binocular vision system. In J.-O. Eklund, editor, *Proceedings of the 3rd European Conference on Computer Vision*, volume 800 of *Lecture Notes in Computer Science*, pages 365–376. Springer, Berlin, 1994. 110

[251] F. Gasparini and R. Schettini. Color balancing of digital photos using simple image statistics. *Pattern Recognition*, 37(6):1201–1217, 2004. 251

[252] W. S. Geisler and J. S. Perry. Statistics for optimal point prediction in natural images. *Journal of Vision*, 11(12), 2011. 61

[253] W. S. Geisler, J. S. Perry, B. J. Super, and D. P. Gallogly. Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–724, 2001. 61, 103, 104

[254] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984. 203, 208, 216

[255] S. Geman and C. Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, volume 1, page 2. AMS, Providence, RI, 1986. 216

[256] T. Georgiev, C. Intwala, S. Babakan, and A. Lumsdaine. Unified frequency domain analysis of lightfield cameras. In *Proceedings of the 13th European Conference on Computer Vision*, pages 224–237, 2008. 100

[257] H. E. Gerhard, F. A. Wichmann, and M. Bethge. How sensitive is the human visual system to the local statistics of natural images? *PLoS Computational Biology*, 9(1), 2013. 5

[258] G. A. Gescheider. *Psychophysics: The Fundamentals*. Lawrence Erlbaum Associates, New Jersey, 3rd edition, 1997. 17

[259] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005. 66

[260] J.-M. Geusebroek and A. W. M. Smeulders. A six-stimulus theory for stochastic texture. *International Journal of Computer Vision*, 62(1-2):7–16, 2005. 254

[261] J. J. Gibson. *The Perception of the Visual World*. Houghton Mifflin, Boston, 1950. 278, 279, 282

[262] J. J. Gibson. *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston, 1966. 282

[263] J. J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum, Hillsdale, NJ, 1979. 90, 279

[264] J. J. Gibson, G. A. Kaplan, H. N. Reynolds, Jr., and K. Wheeler. The change from visible to invisible: A study of optical transitions. *Perception & Psychophysics*, 5(2):113–116, 1969. 278

[265] M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4(3):179–192, 2003. 277

[266] A. Gijsenij and T. Gevers. Color constancy using natural image statistics. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 252, 254

[267] A. Gijsenij and T. Gevers. Color constancy using natural image statistics and scene semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):687–698, 2011. 254

[268] A. Gijsenij, T. Gevers, and J. van De Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 20(9):2475–2489, 2011. 249, 252

[269] D. Goldman and J.-H. Chen. Vignette and exposure calibration and compensation. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, volume 1, pages 899–906, 2005. 59

[270] E. Goldstein. *Sensation and Perception*. Wadsworth Cengage Learning, Belmont, CA, 9th edition, 2013. 116, 274, 277

[271] J. Golz. The role of chromatic scene statistics in color constancy: Spatial integration. *Journal of Vision*, 8(13), 2008. 253

[272] J. Golz and D. I. MacLeod. Influence of scene statistics on colour constancy. *Nature*, 415(6872):637–640, 2002. 253

[273] D. J. Graham and D. J. Field. Statistical regularities of art images and natural scenes: Spectra, sparseness and nonlinearities. *Spatial Vision*, 21(1–2):1–2, 2007. 83, 84, 85, 148

[274] D. J. Graham and D. J. Field. Global nonlinear compression of natural luminances in painted art. In *Proceedings of IS&T/SPIE Electronic Imaging*, page 68100K, 2008. 83, 84, 151

[275] D. J. Graham and D. J. Field. Variations in intensity statistics for representational and abstract art, and for art from the eastern and western hemispheres. *Perception*, 37(9):1341–1352, 2008. 83, 84, 148, 151

[276] D. J. Graham, J. D. Friedenberg, C. H. McCandless, and D. N. Rockmore. Preference for art: Similarity, statistics, and selling price. In *Proceedings of IS&T/SPIE Electronic Imaging*, page 75271A. International Society for Optics and Photonics, 2010. 83, 148

[277] D. J. Graham, J. D. Friedenberg, and D. N. Rockmore. Efficient visual system processing of spatial and luminance statistics in representational and non-representational art. In *Proceedings of IS&T/SPIE Electronic Imaging*, volume 72401N, 2009. 83, 84, 104, 148, 151

[278] D. J. Graham, J. D. Friedenberg, D. N. Rockmore, and D. J. Field. Mapping the similarity space of paintings: Image statistics and visual perception. *Visual Cognition*, 18(4):559–573, 2010. 83, 104, 148, 193

[279] D. J. Graham and C. Redies. Statistical regularities in art: Relations with visual coding and perception. *Vision Research*, 50(16):1503–1509, 2010. 6

[280] J. J. Granzier, E. Brenner, F. W. Cornelissen, and J. B. Smeets. Luminance—Color correlation is not used to estimate the color of the illumination. *Journal of Vision*, 5(1), 2005. 253

[281] G. Greenfield and D. House. Image recoloring induced by palette color associations. *Journal of the WSCG*, 11(1):189–196, 2003. 241

[282] R. L. Gregory. Perceptual illusions and brain models. *Proceedings of the Royal Society of London B*, 171(1024):279–296, 1968. 208

[283] R. L. Gregory. *Eye and the Brain: The Psychology of Seeing*. McGraw-Hill, New York, 3rd edition, 1978. 34

[284] R. L. Gregory. Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London B*, 352(1358):1121–1128, 1997. 35

[285] R. L. Gregory. *Seeing Through Illusions*. Oxford University Press, Oxford, UK, 2009. 34, 35

[286] R. L. Gregory and P. Heard. Border locking and the café wall illusion. *Perception*, 8(4):365–380, 1979. 35

[287] U. Grenander and A. Srivastava. Probability models for clutter in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):424–429, 2001. 195

[288] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007. 67

[289] W. Grimson. Surface consistency constraints in vision. *Computer Vision, Graphics, and Image Processing*, 24(1):28–51, 1983. 111

[290] M. D. Grossberg and S. K. Nayar. What is the space of camera response functions? In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume II, pages 602–609, 2003. 54

[291] S. Grossberg and E. Mingolla. Neural dynamics of form perception: Boundary adaptation, illusory figures, and neon color spreading. *Psychological Review*, 92(2):173–211, 1985. 36, 112

[292] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan. Structured light 3D scanning in the presence of global illumination. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 713–720, 2011. 52

[293] M. Gur, I. Kagan, and D. M. Snodderly. Orientation and direction selectivity of neurons in V1 of alert monkeys: Functional relationships and laminar distributions. *Cerebral Cortex*, 15(8):1207–1221, 2005. 29

[294] J. Haag, W. Denk, and A. Borst. Fly motion vision is based on Reichardt detectors regardless of the signal-to-noise ratio. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16333–16338, 2004. 275

[295] A. Haar. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69(3):331–371, 1910. 178

[296] S. Hahn. *Hilbert Transforms in Signal Processing*. Artech House, Boston, 1996. 198

[297] J. M. Hammersley and P. E. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971. 213

[298] S. T. Hammett and P. J. Bex. Motion sharpening: Evidence for the addition of high spatial frequencies to the effective neural image. *Vision Research*, 36(17):2729–2733, 1996. 145

[299] P. J. B. Hancock, R. Baddeley, and L. Smith. The principal components of natural images. *Network: Computation in Neural Systems*, 3(1):61–70, 1992. 162

[300] P. J. B. Hancock, A. M. Burton, and V. Bruce. Face processing: Human perception and principal components analysis. *Memory and Cognition*, 24(1):26–40, 1996. 164

[301] M. Hansard, S. Lee, O. Choi, and R. P. Horaud. *Time of Flight Cameras: Principles, Methods, and Applications*. SpringerBriefs in Computer Science. Springer, 2012. 53

[302] J. Y. Hardeberg, F. J. Schmitt, and H. Brettel. Multispectral image capture using a tunable filter. In *Proceedings of IS&T/SPIE Electronic Imaging*, pages 77–88. International Society for Optics and Photonics, 1999. 50

[303] J. Y. Hardeberg, F. J. Schmitt, and H. Brettel. Multispectral color image capture using a liquid crystal tunable filter. *Optical Engineering*, 41(10):2532–2548, 2002. 50

[304] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to machine learning methods. *Neural Computation*, 16(12):2639–2664, 2004. 156

[305] F. J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–84, 1978. 131

[306] J. M. Harris and W. Bonas. Optic flow and scene structure do not always contribute to the control of human walking. *Vision Research*, 42(13):1619–1626, 2002. 279

[307] W. M. Hart, Jr. The temporal responsiveness of vision. In R. A. Moses and W. M. Hart, editors, *Adler s Physiology of the Eye, Clinical Application*. The C. V. Mosby Company, St. Louis, 1987. 31

[308] R. Hartley and S. B. Kang. Parameter-free radial distortion correction with center of distortion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1309–1321, 2007. 58

[309] B. Hassenstein and W. Reichardt. Systemtheoretische Analyse der Zeit, Reihenfolgen, und Vorzeichenauswertung bei der Bewegungsperzepion des Rüsselkäfers Chlorophanus. *Zeitschrift für Naturforschung*, 11:513–524, 1956. 274

[310] M. Hassner and J. Sklansky. The use of Markov random fields as models of texture. *Computer Graphics and Image Processing*, 12(4):357–370, 1980. 203

[311] J. H. van Hateren. Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation. *Journal of Comparative Physiology A*, 171:151–170, 1992. 272

[312] J. H. van Hateren. A theory of maximizing sensory information. *Biological Cybernetics*, 68(1):23–29, 1992. 133, 135

[313] J. H. van Hateren. Processing of natural time series of intensities by the visual system of the blowfly. *Vision Research*, 37(23):3407–3416, 1997. 272

[314] J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265(1412):2315–2320, 1998. 273

[315] J. H. van Hateren and A. van der Schaaf. Temporal properties of natural scenes. In *Electronic Imaging: Science & Technology*, pages 139–143, 1996. 272, 273

[316] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, 265(1394):359, 1998. 42, 60, 77, 132, 172

[317] S. Hatipoglu, S. K. Mitra, and N. G. Kingsbury. Image texture description using complex wavelet transform. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 530–533, 2000. 200

[318] S. Hattar, R. J. Lucas, N. Mrosovsky, S. Thompson, R. H. Douglas, M. W. Hankins, J. Lem, M. Biel, F. Hofmann, R. G. Foster, and K.-W. Yau. Melanopsin and rod-cone photoreceptive systems account for all major accessory visual functions in mice. *Nature*, 424(6944):76–81, 2003. 19

[319] M. J. Hawken, A. J. Parker, and J. S. Lund. Laminar organization and contrast sensitivity of direction-selective cells in the striate cortex of the old world monkey. *Journal of Neuroscience*, 8(10):3541–3548, 1988. 29

[320] HDR Soft. Photomatix. http://www.hdrsoft.com/. 46

[321] G. Healey and R. Kondepudy. Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267–276, 1994. 60

[322] D. Heeger, G. Boynton, J. Demb, E. Seidemann, and W. Newsome. Motion opponency in visual cortex. *Journal of Neuroscience*, 19(16):7162–7174, 1999. 275

[323] D. J. Heeger and J. R. Bergen. Pyramid-based texture analysis/synthesis. In *SIGGRAPH 95: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pages 229–238, New York, 1995. ACM. 117, 203

[324] M. Heiler and C. Schnorr. Natural image statistics for natural image segmentation. *International Journal of Computer Vision*, 63(1):5–19, 2005. 115

[325] R. N. Helga Kolb, Eduardo Fernandez. Webvision: The organization of the retina and visual system. http://webvision.med.utah.edu/, 2010. 18, 22

[326] H. L. F. von Helmholtz. *Handbuch der Physiologischen Optik*. Leopold Voss, Leipzig, Germany, 1867. 32, 228, 262

[327] S. H. C. Hendry and R. C. Reid. The koniocellular pathway in primate vision. *Annual Review of Neuroscience*, 23(1):127–153, 2000. 29

[328] E. Hering. *Outlines of a Theory of the Light Sense (Translation from German: Zur Lehre vom Lichtsinne, 1878)*. Harvard University Press, Cambridge, MA, 1920. 33, 231, 262

[329] P. R. Hill, D. R. Bull, and C. N. Canagarajah. Rotationally invariant texture features using the dual-tree complex wavelet transform. In *Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages 901–904, 2000. 200

[330] T. Hine. Subjective contours produced purely by dynamic occlusion of sparse-points array. *Bulletin of the Psychonomic Society*, 25(3):182–184, 1987. 278

[331] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. 218

[332] A. N. Hirani and T. Totsuka. Combining frequency and spatial domain information for fast interactive image noise removal. In *SIGGRAPH 96: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pages 269–276, 1996. 118

[333] H. Hirschmuller, P. R. Innocent, and J. M. Garibaldi. Fast, unconstrained camera motion estimation from stereo without tracking and robust statistics. In *ICARCV 02: Proceedings of the 7th IEEE International Conference on Control, Automation, Robotics and Vision*, volume 2, pages 1099–1104, 2002. 275

[334] Y.-H. Ho, C.-W. Lin, J.-F. Chen, and H.-Y. Liao. Fast coarse-to-fine video retrieval using shot-level spatio-temporal statistics. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(5):642–648, 2006. 275

[335] E. W. Hobson. *The Theory of Spherical and Ellipsoidal Harmonics*. Chelsea, New York, 1965. 49

[336] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, volume 1, pages 654–661, 2005. 267

[337] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007. 267, 268

[338] G. Hong, M. R. Luo, and P. A. Rhodes. A study of digital camera colorimetric characterization based on polynomial modeling. *Color Research & Application*, 26(1):76–84, 2001. 54

[339] D. C. Hood and M. A. Finkelstein. Comparison of changes in sensitivity and sensation: implications for the response-intensity function of the human photopic system. *Journal of Experimental Psychology: Human Perception and Performance*, 5(3):391–405, 1979. 22

[340] B. K. P. Horn. *Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View*. PhD thesis, Massachusetts Institute of Technology, 1970. 85

[341] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artifical Intelligence*, 17(1):185–203, 1981. 280

[342] R. Hosseini, F. Sinz, and M. Bethge. Lower bounds on the redundancy of natural images. *Vision Research*, 50(22):2213–2222, 2010. 12

[343] I. P. Howard and B. J. Rogers. *Perceiving in Depth*. Oxford University Press, New York, 2012. 257

[344] C. Q. Howe and D. Purves. Range image statistics can explain the anomalous perception of length. *Proceedings of the National Academy of Sciences*, 99(20):13184–13188, 2002. 260, 261, 262

[345] C. Q. Howe and D. Purves. The Müller-Lyer illusion explained by the statistics of image–source relationships. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4):1234–1239, 2005. 260, 262, 263

[346] C. Q. Howe and D. Purves. Natural-scene geometry predicts the perception of angles and line orientation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4):1228–1233, 2005. 260, 262, 263

[347] C. Q. Howe and D. Purves. *Perceiving Geometry: Geometrical Illusions Explained by Natural Scene Statistics*. Springer, New York, 2005. 260, 262, 263

[348] C. Q. Howe, Z. Yang, and D. Purves. The Poggendorff illusion explained by natural scene geometry. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7707–7712, 2005. 260, 262, 263

[349] M.-K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962. 75, 76

[350] J. Huang, A. Lee, and D. Mumford. Statistics of range images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 324–331, Washington, DC, 2000. 67, 97, 133, 257, 258, 283

[351] J. Huang and D. Mumford. Image statistics for the British Aerospace segmented database. Technical report, Division of Applied Math, Brown Univeristy, 1999. 135, 136

[352] J. Huang and D. Mumford. Statistics of natural images and models. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, Washington, DC, 1999. 42, 60, 77, 78, 97, 136, 146, 258

[353] S.-J. Huang and C.-T. Hsieh. Coiflet wavelet transform applied to inspect power system disturbance-generated signals. *IEEE Transactions on Aerospace and Electronic Systems*, 38(1):204–210, 2002. 186

[354] D. H. Hubel. *Eye, Brain, and Vision*. Scientific American Library Series. Henry Holt and Company, New York, 1995. 4

[355] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of the monkey striate cortex. *Journal of Physiology*, 195(1):215–243, 1968. 29, 103

[356] D. H. Hubel and T. N. Wiesel. Laminar and columnar distribution of geniculocortical fibers in the macaque monkey. *Journal of Comparative Neurology*, 146(4):421–450, 1972. 29

[357] D. H. Hubel and T. N. Wiesel. Ferrier lecture: Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London B*, 198(1130):1–59, 1977. 17

[358] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981. 163

[359] J. M. Hughes, D. J. Graham, C. R. Jacobsen, and D. N. Rockmore. Comparing higher-order spatial statistics and perceptual judgements in the stylometric analysis of art. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 1244–1248, 2011. 172

[360] J. M. Hughes, D. J. Graham, and D. N. Rockmore. Quantifiction of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the elder. *Proceedings of the National Academy of Sciences*, 107(4):1279–1283, 2010. 172

[361] P. Hung. Colorimetric calibration in electronic imaging devices using a look-up-table model and interpolations. *Journal of Electronic Imaging*, 2(1):53–61, 1993. 54

[362] L. M. Hurvich and D. Jameson. The opponent process theory of color vision. *Psychological Review*, 64(6):384–404, 1957. 33, 231

[363] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999. 166, 169, 171

[364] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2(4):94–128, 1999. 166

[365] A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer, London, 2009. 159, 166, 168, 169, 170

[366] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, New York, 2001. 166

[367] S. Ingebritsen and R. Lyon. Principal components analysis of multitemporal image pairs. *International Journal of Remote Sensing*, 6(5):687–696, 1985. 270

[368] M. Irfan and D. G. Stork. Multiple visual features for the computer authentication of Jackson Pollock's drip paintings: Beyond box counting and fractals. In *SPIE: Image Processing: Machine Vision Applications II*, volume 7251, pages 1–11, 2009. 152

[369] E. Ising. *Beitrag zur Theorie des Ferro- und Paramagnetismus*. PhD thesis, University of Hamburg, 1924. 208

[370] J. Ivins, J. Porrill, J. Frisby, and G. Orban. The “ecological" probability density function for linear optic flow: Implications for neurophysiology. *Perception*, 28(1):17–32, 1999. 283

[371] A. Jalobeanu, L. Blanc-Féraud, and J. Zerubia. Estimation of blur and noise parameters in remote sensing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 249–256, 2002. 115

[372] A. Jalobeanu, N. G. Kingsbury, and J. Zerubia. Image deconvolution using hidden Markov tree modeling of complex wavelet packets. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 201–204, 2001. 200

[373] R. Jarvis. A laser time-of-flight range scanner for robotic vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(5):505–512, 1983. 53

[374] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. 276

[375] G. Johansson. Visual motion perception. *Scientific American*, 232(6):76–88, 1975. 276

[376] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, New York, 1994. 74

[377] I. Jolliffe. Principal component analysis. In B. Everitt and D. Howell, editors, *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, New York, 2005. 153

[378] D. G. Jones and J. Malik. A computational framework for determining stereo correspondence from a set of linear spatial filters. *Image and Vision Computing*, 10(10):395–410, 1992. 110

[379] B. Jowett, editor. *The Dialogues Of Plato, Translated into English with Analyses and Introductions*, volume 1. Oxford University Press, London, 3rd edition, 1892. 269

[380] B. Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8(2):84–92, 1962. 203, 205

[381] B. Julesz and T. Caelli. On the limits of Fourier decompositions in visual texture perception. *Perception*, 8(1):69–73, 1979. 119, 145

[382] D. C. Kale and D. G. Stork. Estimating the position of illuminants in paintings under weak model assumptions: An application to the works of two Baroque masters. In *Proceedings of IS&T/SPIE Electronic Imaging*, page 72401, 2009. 84

[383] S. Kalkan, D. Calow, F. Wörgötter, M. Lappe, and N. Krüger. Local image structures and optic flow estimation. *Network: Computation in Neural Systems*, 16(4):341–356, 2005. 283

[384] S. Kalkan, F. Wörgötter, and N. Krüger. Statistical analysis of local 3D structure in 2D images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1114–1121, 2006. 264, 265, 266, 267

[385] S. Kalkan, F. Wörgötter, and N. Krüger. First-order and second-order statistical analysis of 3D and 2D image structure. *Network: Computation in Neural Systems*, 18(2):129–160, 2007. 264, 267

[386] M. Kalloniatis and C. Luu. Temporal resolution. In *Webvision: The Organization of the Retina and Visual System*. http://webvision.med.utah.edu/book/part-viii-gabac-receptors/temporal-resolution/, Salt Lake City, UT, 2007. 31, 33

[387] M. Kalloniatis and C. Luu. Visual acuity. In *Webvision: The Organization of the Retina and Visual System*. http://webvision.med.utah.edu/book/part-viii-gabac-receptors/visual-acuity/, Salt Lake City, UT, 2011. 30

[388] C. Kanan and G. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2472–2479, 2010. 172

[389] S. B. Kang. Automatic removal of chromatic aberration from a single image. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 56

[390] S. B. Kang and R. Weiss. Can we calibrate a camera using an image of a flat, textureless lambertian surface? In *Proceedings of the 6th European Conference on Computer Vision*, pages 640–653, 2000. 59

[391] Y. Karklin and M. S. Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225):83–86, 2008. 220, 221

[392] P. J. Kellman and M. H. Cohen. Kinetic subjective contours. *Perception & Psychophysics*, 35(3):237–244, 1984. 278

[393] P. J. Kellman, E. M. Palmer, and T. F. Shipley. Effects of velocity in dynamic object completion. *Investigative Ophthalmology and Visual Science Supplement*, 39:S855, 1998. 278

[394] J. N. D. Kerr and W. Denk. Imaging *in vivo*: Watching the brain in action. *Nature Reviews Neuroscience*, 9(3):195–205, 2008. 17

[395] D. Kersten. Predictability and redundancy of natural images. *Journal of the Optical Society of America A*, 4(12):2395–2400, 1987. 4

[396] M. Keshner. $1/f$ noise. *Proceedings of the IEEE*, 70(3):212–218, 1982. 217

[397] A. Khintchine. Korrelationstheorie der stationären stochastischen Processe. *Mathematische Annalen*, 109(1):604–615, 1934. 128

[398] B. Kim and R. Park. Automatic detection and correction of purple fringing using the gradient information and desaturation. In *Proceedings of the 16th European Signal Processing Conference*, 2008. 56

[399] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*, volume 1. American Mathematical Society, Providence, RI, 1980. 208

[400] G. Kindlmann, E. Reinhard, and S. Creem. Face-based luminance matching for perceptual colormap generation. In *Proceedings of IEEE Visualization*, pages 309–406, 2002. 33, 34

[401] F. A. A. Kingdom and B. Moulden. Border effects on brightness: A review of findings, models and issues. *Spatial Vision*, 3(4):225–262, 1988. 36

[402] F. A. A. Kingdom and N. Prins. *Psychophysics: A Practical Introduction*. Academic Press / Elsevier, London, 2009. 17

[403] N. G. Kingsbury. Image processing with complex wavelets. *Philosophical Transactions of the Royal Society A*, 357(1760):2543–2560, 1999. 197, 199

[404] N. G. Kingsbury. Shift invariant properties of the dual-tree complex wavelet transform. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1221–1224, 1999. 199

[405] N. G. Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Journal of Applied and Computational Harmonic Analysis*, 10(3):234–253, 2001. 199

[406] N. G. Kingsbury. Design of Q-shift complex wavelets for image processing using frequency domain energy minimization. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 1013–1016, 2003. 199

[407] D. A. Kleffner and V. S. Ramachandran. On the perception of shape from shading. *Perception & Psychophysics*, 52(1):18–36, 1992. 3

[408] D. C. Knill, D. J. Field, and D. Kersten. Human discrimination of fractal images. *Journal of the Optical Society of America A*, 7(6):1113–1123, 1990. 144

[409] D. C. Knill and W. Richards. *Perception as Bayesian inference*. Cambridge University Press, Cambridge, UK, 1996. 261

[410] M. Koch, J. Denzler, and C. Redies. $1/f^2$ characteristics and isotropy in the Fourier power spectra of visual art, cartoons, comics, mangas, and different categories of photographs. *PLoS ONE*, 5(8):e12268, 2010. 151, 152, 172

[411] J. J. Koenderink. The structure of images. *Biological Cybernetics*, 50(5):363–370, 1984. 105, 108, 110

[412] J. J. Koenderink. The brain a geometry engine. *Psychological Research*, 52(2–3):122–127, 1990. 109

[413] J. J. Koenderink and A. J. van Doorn. Shape from shading. In L. M. Chalupa and J. S. Werner, editors, *The Visual Neurosciences*, pages 1090–1105. MIT Press, Cambridge, MA, 2003. 34, 267

[414] W. Köhler. *Gestalt Psychology*. Liveright, New York, 1929. 103

[415] M. Kokare, P. K. Biswas, and B. N. Chatterji. Rotation invariant features using rotated complex wavelet for content based image retrieval. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 393–396, 2004. 200

[416] N. Komodakis and G. Tziritas. Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Transactions on Image Processing*, 16(11):2649–2661, Nov. 118

[417] S. Konishi, A. Yuille, J. Coughlan, and S.-C. Zhu. Statistical edge detection: learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):57–74, 2003. 103

[418] H. Kotera. A scene-referred color transfer for pleasant imaging on display. *Proceedings of the IEEE International Conference on Image Processing*, 2:5–8, 2005. 242

[419] H. Kotera, T. Morimoto, and R. Saito. Object-oriented color matching by image clustering. In *Proceedings of the 6th IS&T/SID Color Imaging Conference*, pages 154–158, 1998. 242

[420] D. Koubaroulis, J. Matas, and J. Kittler. Evaluating colour-based object recognition algorithms using the soil-47 database. In *Asian Conference on Computer Vision*, volume 2002, page 2, 2002. 66

[421] N. Kouyama and D. W. Marshak. Bipolar cells specific for blue cones in the macaque retina. *Journal of Neuroscience*, 12(4):1233–1252, 1992. 25

[422] J. Kovacević, V. K. Goyal, and M. Vetterli. *Signal Processing Fourier and Wavelet Representations*. fourierandwavelets.org, 2012. 175

[423] J. Krauskopf. Effect of retinal stabilization on the appearance of heterochromatic targets. *Journal of the Optical Society of America*, 53(6):741–744, 1963. 36, 112

[424] E. R. Kretzmer. Statistics of television signals. *Bell Systems Technical Journal*, 31(4):751–763, 1952. 4

[425] D. Krishnan and R. Fergus. Fast image deconvolution using hyper-Laplacian priors. In *NIPS-23: Proceedings of the 2009 Conference on Advances in Neural Information Processing Systems*, volume 22, pages 1–9, 2009. 115

[426] N. Krüger and M. Felsberg. A continuous formulation of intrinsic dimension. In *Proceedings of the British Machine Vision Conference*, volume 2, 2003. 264

[427] N. Krüger and F. Wörgötter. Multi-modal estimation of collinearity and parallelism in natural image sequences. *Network: Computation in Neural Systems*, 13(4):553–576, 2002. 103, 104

[428] P. Kube and A. Pentland. On the imaging of fractal surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):704–707, 1988. 146

[429] V. N. Kublanovskaya. On some algorithms for the solution of the complete eigenvalue problem. *USSR Computational Mathematics and Mathematical Physics*, 1(3):637–657, 1961. 158

[430] R. Kumar and S. K. Mitra. Motion estimation based color transfer and its application to color video compression. *Pattern Analysis and Applications*, 11(2):131–139, 2007. 242

[431] T. Kunkel and E. Reinhard. A reassessment of the simultaneous dynamic range of the human visual system. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, pages 17–24, New York, 2010. ACM. 22, 43

[432] K. N. Kutulakos and E. Steger. A theory of refractive and specular 3D shape by light-path triangulation. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, volume 2, pages 1448–1455, 2005. 52

[433] D. Lamming. Contrast sensitivity. In J. Cronly-Dillon, editor, *Vision and Visual Dysfunction*, volume 5. Macmillan Press, London, 1991. 33

[434] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of the Research of the National Bureau of Standards*, 45(4):255–282, 1950. 158

[435] E. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–128, 1977. 248, 251

[436] R. Lange and P. Seitz. Solid-state time-of-flight range camera. *IEEE Journal of Quantum Electronics*, 37(3):390–397, 2001. 53

[437] M. S. Langer. Large-scale failures of $f^{-\alpha}$ scaling in natural image spectra. *Journal of the Optical Society of America A*, 17(1):28–33, 2000. 141

[438] M. S. Langer and H. H. Bülthoff. Depth discrimination from shading under diffuse lighting. *Perception*, 29(6):649–660, 2000. 85

[439] M. S. Langer and H. H. Bülthoff. A prior for global convexity in local shape-from-shading. *Perception*, 30(4):403–410, 2001. 86

[440] V. Laparra, S. Jiménez, G. Camps-Valls, and J. Malo. Nonlinearities and adaptation of color vision from sequential principal curves analysis. *Neural Computation*, 24(10):2751–2788, 2012. 63

[441] M. Lappe, editor. *Neuronal processing of optic flow*, volume 44. Academic Press, San Diego, 1999. 282

[442] I. Laptev. *Local spatio-temporal image features for motion interpretation*. PhD thesis, KTH, 2004. 275

[443] I. Laptev and T. Lindeberg. Space-time interest points. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 432–439, 2003. 275

[444] I. Laptev and T. Lindeberg. Velocity adaptation of space-time interest points. In *Proceedings of the 17th IEEE International Conference on Pattern Recognition*, pages 52–56, 2004. 275

[445] S. B. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung C*, 36:910–912, 1981. 4

[446] E. G. Learned-Miller and J. W. Fisher III. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003. 166

[447] A. B. Lee, D. Mumford, and J. Huang. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision*, 41(1):35–59, 2001. 67, 257, 258, 259, 260, 283

[448] T.-W. Lee, M. Girolami, and T. J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources. *Neural Computation*, 11(2):417–441, 1999. 166

[449] A. Leonardis and H. Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding*, 78(1):99–118, 2000. 163

[450] D. A. Leopold, A. O'Toole, T. Vetter, and V. Blanz. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, 4(1):89–94, 2001. 164

[451] L. Leslie, T.-S. Chua, and R. Jain. Annotation of paintings with high-level semantic concepts using transductive inference and ontology-based concept disambiguation. In *Proceedings of the 15th ACM Conference on Multimedia*, pages 443–452, 2007. 84

[452] E. Levi. Using natural image priors—Maximizing or sampling? Master's thesis, The Hebrew University, 2009. 220

[453] A. Levin. Blind motion deblurring using image statistics. In *NIPS-20: Proceedings of the Conference on Advances in Neural Information Processing Systems*, volume 19, pages 841–848, 2006. 6, 96, 115

[454] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. *ACM Transactions on Graphics*, 24(3):689–694, 2004. 242

[455] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1647–1654, 2007. 115

[456] A. Levin, A. Zomet, S. Peleg, and Y. Weiss. Seamless image stitching in the gradient domain. In T. Pajdla and J. Matas, editors, *Proceedings of the 10th European Conference on Computer Vision*, volume 3024 of *Lecture Notes in Computer Science*. Springer, Berlin, 2004. 48

[457] A. Levin, A. Zomet, and Y. Weiss. Learning to perceive transparency from the statistics of natural scenes. In *NIPS-15: Proceedings of the 2002 Conference on Advances in Neural Information Processing Systems*, Cambridge, MA, 2002. MIT Press. 97

[458] A. Levin, A. Zomet, and Y. Weiss. Learning how to inpaint from global image statistics. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, pages 305–312, Washington, DC, 2003. 118

[459] M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, and J. Ginsberg. The digital Michelangelo project: 3D scanning of large statues. In *SIGGRAPH 00: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 131–144. ACM Press/Addison-Wesley Publishing Co., 2000. 52

[460] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, 2009. 208, 215, 216

[461] Z. Li, Z. Jing, X. Yang, and S. Sun. Color transfer based remote sensing image fusion using non-separable wavelet frame transform. *Pattern Recognition Letters*, 26(13):2006–2014, 2005. 241

[462] J. Limb and J. Murphy. Estimating the velocity of moving images in television signals. *Computer Graphics and Image Processing*, 4(4):311–327, 1975. 275

[463] J. M. Lina and M. Mayrand. Complex Daubechies wavelets. *Applied and Computational Harmonic Analysis*, 2(3):219–229, 1995. 199

[464] T. Lindeberg. Linear scale space. In ter Haar Romeny, editor, *Geometry-Driven Diffusion in Computer Vision*, pages 1–77. Kluwer Academic Publishers, Dordrecht, Netherlands, 1994. 105, 109, 110

[465] T. Lindeberg and L. Florack. Foveal scale-space and the linear increase of receptive field size as a function of eccentricity. Technical report, ISRN KTH NA/P–94/27–SE, 1994. 109

[466] T. Lindeberg and J. Gårding. Shape from texture from a multi-scale perspective. In *Proceedings of the IEEE International Conference in Computer Vision*, pages 683–691, 1993. 110

[467] A. Litvinov and Y. Y. Schechner. Radiometric framework for image mosaicking. *Journal of the Optical Society of America A*, 22(5):839–848, 2005. 59

[468] J. Liu and P. Moulin. Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients. *IEEE Transactions on Image Processing*, 10(11):1647–1658, 2001. 194

[469] E. Lo, M. Pickering, M. Frater, and J. Arnold. Scale and rotation invariant texture features from the dual-tree complex wavelet transform. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 227–230, 2004. 200

[470] G. Lodwick. Measuring ecological changes in multitemporal landsat data using principal components. In *Proceedings of the International Symposium on Remote Sensing of the Environment*, pages 1131–1141, 1979. 270

[471] G. Lodwick. A computer system for monitoring environmental change in multitemporal landsat data. *Canadian Journal of Remote Sensing*, 7:24, 1981. 270

[472] F. Long, Z. Yang, and D. Purves. Spectral statistics in natural scenes predict hue, saturation, and brightness. *Proceedings of the National Academy of Sciences*, 103(15):6013–6018, 2006. 263

[473] P. Loo and N. G. Kingsbury. Digital watermarking using complex wavelets. In *Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages 29–32, 2000. 200

[474] Q. Luan, F. Wen, and Y.-Q. Xu. Color transfer brush. In *Proceedings of Pacific Graphics*, pages 465–468, 2007. 241

[475] B. D. Lucas and T. Kanade. An iterative image registration technique with an application in stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679, 1981. 280

[476] J. S. Lund. Organization of neurons in the visual cortex, area 17, of the monkey (Macaca mulatta). *Journal of Comparative Neurology*, 147(4):455–496, 1973. 29

[477] J. S. Lund and R. G. Boothe. Interlaminar connections and pyramidal neuron organization in the visual cortex, area 17, of the macaque monkey. *Journal of Comparative Neurology*, 159(3):305–334, 1975. 29

[478] S. Lyu and H. Farid. Detecting hidden messages using higher-order statistics and support vector machines. In *Information Hiding*, pages 340–354, 2003. 193

[479] S. Lyu and H. Farid. How realistic is photorealistic? *IEEE Transactions on Signal Processing*, 53(2):845–850, 2005. 193

[480] S. Lyu and H. Farid. Steganalysis using higher-order image statistics. *IEEE Transactions on Information Forensics and Security*, 1(1):111–119, 2006. 193, 197

[481] S. Lyu, D. Rockmore, and H. Farid. A digital technique for art authentication. *Proceedings of the National Academy of Sciences of the United States of America*, 101(49):17006–17010, 2004. 193

[482] S. Lyu and E. P. Simoncelli. Modeling multiscale subbands of photographic images with fields of Gaussian scale mixtures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):693–706, 2009. 220

[483] E. Mach. *The Analysis of Sensations and the Relation of the Physical to the Psychical*. Dover, New York, 1959. 35

[484] T. MacRobert and I. Sneddon. *Spherical Harmonics. An Elementary Treatise on Harmonic Functions*. Pergamon Press, Oxford, UK, 3rd edition, 1967. 49

[485] B. Mahdian and S. Saic. Detection of copy–move forgery using a method based on blur moment invariants. *Forensic Science International*, 171(2):180–189, 2007. 75

[486] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A*, 7(5):923–932, 1990. 203

[487] J. Malik and R. Rosenholtz. A differential method for computing local shape-from-texture for planar and curved surfaces. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 267–273, 1993. 110

[488] S. G. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Transactions of the American Mathematical Society*, 315(1):69–87, 1989. 178

[489] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. 97, 178, 190, 191

[490] S. G. Mallat. *A Wavelet Tour of Signal Processing*. Elsevier, Amsterdam, 3rd edition, 2009. 176

[491] J. Mallon and P. Whelan. Calibration and removal of lateral chromatic aberration in images. *Pattern Recognition Letters*, 28(1):125–135, 2007. 56

[492] P. Mamassian and R. Goutcher. Prior knowledge on the illumination position. *Cognition*, 81(1):B1–B9, 2001. 86

[493] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W.H. Freeman and Co., New York, 1983. 136, 145, 146

[494] S. Mann and R. Picard. On being “undigital” with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proceedings of the IS&T 48th Annual Conference*, pages 422–428, 1995. 44, 46

[495] V. Mante, R. A. Frazor, V. Bonin, W. S. Geisler, and M. Carandini. Independence of luminance and contrast in natural scenes and in the early visual system. *Nature Neuroscience*, 8(12):1690–1697, 2005. 114

[496] Y. Marchenko, T.-S. Chua, and I. Aristarkhova. Analysis and retrieval of paintings using artistic color concepts. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1246–1249, 2005. 84

[497] J. F. A. Margarey and N. G. Kingsbury. Motion estimation using a complex-valued wavelet transform. *IEEE Transactions on Signal Processing*, 46(4):1069–1084, 1998. 200

[498] D. Marr. *Vision, A Computational Investigation into the Human Representation and Processing of Visual Information*. W H Freeman and Company, San Fransisco, 1982. 4, 103

[499] D. Marr and E. C. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London B*, 207(1167):187–217, 1980. 147

[500] N. J. Marshall, M. F. Land, C. A. King, and T. W. Cronin. The compound eyes of mantis shrimps (crustacea, hoplocarida, stomatopoda). II. Colour pigments in the eyes of stomatopod crustaceans: Polychromatic vision by serial and lateral filtering. *Philosophical Transactions of the Royal Society of London B*, 334(1269):57–84, 1991. 49, 228

[501] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference in Computer Vision*, volume 2, pages 416–423, 2001. 219

[502] G. Mather. The use of image blur as a depth cue. *Perception*, 26(9):1147–1158, 1997. 147

[503] C. McCamy, H. Marcus, and J. Davidson. A color-rendition chart. *Journal of Applied Photographic Engineering*, 2(3):95–99, 1976. 54

[504] M. McCotter, F. Gosselin, P. Sowden, and P. Schyns. The use of visual information in natural scenes. *Visual Cognition*, 12(6):938–953, 2005. 142

[505] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2004. 156

[506] I. C. McManus, J. Buckman, and E. Woolley. Is light in pictures presumed to come from the left side? *Perception*, 33(12):1421–1436, 2004. 86

[507] B. W. Mel. SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, 9(4):777–804, 1997. 163

[508] A. Michotte. A propos de la permanence phénoménale: Faits et théories (On phenomenal permanence: Facts and theories). *Acta Psychologica*, 7:298–322, 1950. 278

[509] A. Michotte, G. Thinès, and G. Crabbé. Les compléments amodaux des structures perceptives [amodal completion and perceptual organization]. In *Studia Psychologica*. Publications Universitaires de Louvain, Louvain, Belgium, 1964. 278

[510] R. B. Millar. *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB*. John Wiley & Sons, Chichester, UK, 2011. 169

[511] T. Mitsunaga and S. K. Nayar. Radiometric self calibration. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 374–380, 1999. 44, 46

[512] J. D. Mollon and J. K. Bowmaker. The spatial arrangement of cones in the primate retina. *Nature*, 360(6405):677–679, 1992. 24

[513] R. A. Morano, C. Ozturk, R. Conn, S. Dubin, S. Zietz, and J. Nissano. Structured light using pseudorandom codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):322–327, 1998. 52

[514] J. Morovic and P.-L. Sun. Accurate 3D image colour histogram transformation. *Pattern Recognition Letters*, 24(11):1725–1735, 2003. 242

[515] I. Mosseri, M. Zontak, and M. Irani. Combining the power of internal and external denoising. In *Proceedings of the IEEE International Conference on Computational Photography*, 2013. 99

[516] I. Motoyoshi and E. H. Adelson. Luminance re-mapping for the control of apparent material. In *Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization*, pages 165–165, 2005. 81

[517] I. Motoyoshi, S. Nishida, L. Sharan, and E. H. Adelson. Image statistics and the perception of surface qualities. *Nature*, 447(7141):206–209, 2007. 81, 82, 83

[518] D. Mumford and B. Gidas. Stochastic models for generic images. *Quarterly of Applied Mathematics*, 59(1):85–112, 2001. 136, 193

[519] H. Murase and S. K. Nayar. Illumination planning for object recognition using parametric eigenspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(12):1219–1227, 1994. 163

[520] H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995. 163

[521] J. R. Mureika, C. C. Dyer, and G. C. Cupchik. Multifractal structure in nonrepresentational art. *Physical Review E*, 72(4):281–295, 2005. 152

[522] J. R. Mureika, M. S. Fairbanks, and R. P. Taylor. Multifractal comparison of the painting techniques of adults and children. In D. Stork, J. Coddington, and A. Bentkowska-Kafel, editors, *SPIE: Computer Vision and Image Analysis of Art*, volume 7531, pages 1–6, 2010. 152

[523] J. R. Mureika and R. P. Taylor. The abstract expressionists and les automatistes: A shared multi-fractal depth? *Signal Processing*, 93(3):573–578, 2013. 152

[524] S. Murray and P. J. Bex. Perceived blur in naturally contoured images depends on phase. *Frontiers in Psychology*, 1(185):1–12, 2010. 147

[525] K. I. Naka and W. A. H. Rushton. S-potentials from luminosity units in the retina of fish (cyprinidae). *Journal of Physiology*, 185(3):587–599, 1966. 22, 84

[526] J. Nakamura. *Image Sensors and Signal Processing for Digital Still Cameras*. CRC Press, Boca Raton, FL, 2005. 49

[527] S. Nascimento, F. P. Ferreira, and D. H. Foster. Statistics of spatial cone-excitation ratios in natural scenes. *Journal of the Optical Society of America A*, 19(8):1484–1490, 2002. 64

[528] S. K. Nayar, H. Murase, and S. A. Nene. Learning, positioning, and tracking visual appearance. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3237–3244, 1994. 163

[529] S. K. Nayar, S. A. Nene, and H. Murase. Columbia object image library (coil-100). Technical Report CUCS-006-96, Department of Comp. Science, Columbia University, 1996. 66

[530] R. Neelamani, H. Choi, and R. Baraniuk. ForWaRD: Fourier-wavelet regularized deconvolution for ill-conditioned systems. *IEEE Transactions on Signal Processing*, 52(2):418–433, 2004. 115

[531] J. Neitz, J. Carroll, Y. Yamauchi, M. Neitz, and D. R. Williams. Color perception is mediated by a plastic neural mechanism that is adjustable in adults. *Neuron*, 35(4):783–792, 2002. 25

[532] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (COIL-20). Technical Report CUCS-006-96, Columbia University Computer Science, 1996. 66

[533] H. Neumann, L. Pessoa, and T. Hansen. Visual filling-in for computing perceptual surface properties. *Biological Cybernetics*, 85(5):355–369, 2001. 112

[534] L. Neumann and A. Neumann. Color style transfer techniques using hue, lightness and saturation histogram matching. In *Proceedings of the 1st Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, pages 111–122, 2005. 241

[535] C.-W. Ngo, T.-C. Pong, and R. T. Chin. Detection of gradual transitions through temporal slice analysis. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, 1999. 271

[536] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang. Motion analysis and segmentation through spatio-temporal slices processing. *IEEE Transactions on Image Processing*, 12(3):341–355, 2003. 271

[537] C.-W. Ngo, T.-C. Pong, H.-J. Zhang, and R. T. Chin. Motion characterization by temporal slices analysis. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 768–773, 2000. 271

[538] C. Nikias and A. Petropolu. *Higher-Order Spectra Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1993. 128

[539] M. Niss. History of the Lenz-Ising model 1920–1950: From ferromagnetic to cooperative phenomena. *Archive for History of Exact Sciences*, 59(3):267–318, 2005. 208

[540] S. Nundy, B. Lotto, D. Coppola, A. Shimpi, and D. Purves. Why are angles misperceived? *Proceedings of the National Academy of Sciences*, 97(10):5592–5597, 2000. 262, 263

[541] V. O'Brien. Contour perception, illusion and reality. *Journal of the Optical Society of America*, 48(2):112–119, 1958. 36

[542] K. Ohki, S. Chung, Y. H. Ch'ng, P. Kara, and R. C. Reid. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature*, 433(7026):597–603, 2005. 17

[543] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 140, 146

[544] A. Olmos. A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, 33(12):1463, 2004. 64

[545] B. A. Olshausen and M. R. DeWeese. The statistics of style. *Nature*, 463(7284):1027–1028, 2010. 172

[546] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. 4, 162, 163

[547] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997. 169

[548] G. A. Orban. *Neuronal Operations in the Visual Cortex*. Springer-Verlag, Berlin, 1984. 162

[549] S. Osindero and G. E. Hinton. Modeling image patches with a directed hierarchy of Markov random fields. In *NIPS-22: Proceedings of the 2008 Conference on Advances in Neural Information Processing Systems*, volume 20, pages 1121–1128, 2008. 220, 221

[550] C. Owsley, R. Sekular, and D. Siemsen. Contrast sensitivity throughout adulthood. *Vision Research*, 23(7):689–699, 1983. 144

[551] J. Pages, J. Salvi, R. Garcia, and C. Matabosch. Overview of coded light projection techniques for automatic 3D profiling. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 1, pages 133–138, 2003. 52

[552] E. Palmer, P. J. Kellman, and T. F. Shipley. Spatiotemporal relatability in dynamic object completion. *Investigative Ophthalmology & Visual Science*, 38(4):256, 1997. 278

[553] S. E. Palmer. The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3(5):519–526, 1975. 267

[554] S. E. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, MA, 1999. 4, 17, 19, 51

[555] S. Panda, I. Provencio, D. C. Tu, S. S. Pires, M. D. Rollag, A. M. Castrucci, M. T. Pletcher, T. K. Sato, T. Wiltshire, and M. Andahazy. Melanopsin is required for non-image-forming photic responses in blind mice. *Science Signalling*, 301(5632):525, 2003. 72

[556] M. Pappas and I. Pitas. Digital color restoration of old paintings. *IEEE Transactions on Image Processing*, 9(2):291–294, 2000. 84

[557] S. Paris and F. Durand. A fast approximation of the bilateral filter using a signal processing approach. *International Journal of Computer Vision*, 81(1):24–52, 2009. 106

[558] S. Paris, S. W. Hasinoff, and J. Kautz. Local Laplacian filters: Edge-aware image processing with a Laplacian pyramid. *ACM Transactions on Graphics*, 30(4):68, 2011. 106

[559] C. A. Párraga, R. Baldrich, and M. Vanrell. Accurate mapping of natural scenes radiance to cone activation space: A new image dataset. In *Proceedings of the 5th European Conference on Colour in Graphics, Imaging, and Vision*, 2010. 62

[560] C. A. Párraga, G. Brelstaff, T. Troscianko, and I. R. Moorehead. Color and luminance information in natural scenes. *Journal of the Optical Society of America A*, 15(3):563–569, 1998. 66, 133, 135

[561] C. A. Párraga, J. Vazquez-Corral, and M. Vanrell. A new cone activation-based natural images dataset. *Perception*, 36:180, 2009. 62

[562] D. Pascale. RGB coordinates of the Macbeth Color Checker. Technical report, The BabelColor Company, 2006. 54

[563] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901. 153

[564] K. S. Pedersen and A. B. Lee. Toward a full probability model of edges in natural images. In *Proceedings of the 7th European Conference on Computer Vision*, pages 328–342. Springer, 2002. 103

[565] K. S. Pedersen and A. B. Lee. Toward a full probability model of edges in natural images. In *Proceedings of the 8th European Conference on Computer Vision*, pages 328–342, Heidelberg, 2002. Springer. 110

[566] H.-O. Peitgen and D. Saupe. *The Science of Fractal Images*. Springer-Verlag, New York, 1988. 145, 148

[567] E. Peli. Contrast in complex images. *Journal of the Optical Society of America A*, 7(10):2032–2040, 1990. 32, 114

[568] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003. 6, 118

[569] K. Perlin. An image synthesizer. In *SIGGRAPH 85: Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, pages 287–296, 1985. 148

[570] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990. 106

[571] D. H. Peterzell, J. S. Werner, and P. S. Kaplan. Individual differences in contrast sensitivity functions: Longitudinal study of 4-, 6- and 8-month-old human infants. *Vision Research*, 35(7):9651–979, 1995. 144

[572] F. Pitié, A. Kokaram, and R. Dahyot. N-dimensional probability density function transfer and its application to colour transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pages 1434–1439, Washington, DC, 2005. 241

[573] F. Pitié, A. Kokaram, and R. Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(2):1434–1439, 2007. 241

[574] R. Pless, T. Brodsky, and Y. Aloimonos. Detecting independent motion: The statistics of temporal continuity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):768–773, 2000. 279

[575] D.-Y. Po and M. N. Do. Directional multiscale modeling of images using the contourlet transform. *IEEE Transactions on Image Processing*, 15(6):1610–1620, 2006. 193, 195

[576] J. Pokorny, V. C. Smith, and M. Wesner. Variability in cone populations and implications. In A. Valberg and B. B. Lee, editors, *From Pigments to Perception*, pages 23–34. Plenum, New York, 1991. 24

[577] W. T. Pollock and A. Chapanis. The apparent length of a line as a function of its inclination. *Quarterly Journal of Experimental Psychology*, 4(4):170–178, 1952. 261

[578] K. Popat and R. W. Picard. Novel cluster-based probability model for texture synthesis, classification, and compression. In *Visual Communications 93*, pages 756–768. International Society for Optics and Photonics, 1993. 203

[579] M. Porat and Y. Y. Zeevi. Localized texture processing in vision: Analysis and synthesis in the Gaborian space. *IEEE Transactions on Biomedical Engineering*, 36(1):115–129, 1989. 203

[580] J. Portilla, R. Navarro, O. Nestares, and A. Tabernero. Texture synthesis-by-analysis method based on a multiscale early-vision model. *Optical Engineering*, 35(8):2403–2417, 1996. 203

[581] J. Portilla and E. P. Simoncelli. Texture modeling and synthesis using joint statistics of complex wavelet coefficients. In *Proceedings of the IEEE Workshop on Statistical and Computational Theories of Vision*, 1999. 9, 203, 205, 206

[582] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–71, 2000. 9

[583] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 12(11):1338–1351, 2003. 173

[584] B. Potetz and T. S. Lee. Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. *Journal of the Optical Society of America A*, 20(7):1292–1303, 2003. 257

[585] T. Pouli, D. Cunningham, and E. Reinhard. Statistical regularities in low and high dynamic range images. In *APGV 10: Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, pages 9–16, New York, 2010. 78, 79, 80, 135, 136, 137, 138

[586] T. Pouli, D. W. Cunningham, and E. Reinhard. Image statistics and their applications in computer graphics. In *Eurographics State-of-the-Art Reports*, 2010. 97

[587] T. Pouli and E. Reinhard. Progressive histogram reshaping for creative color transfer and tone reproduction. In *NPAR 10: Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, pages 81–90, 2010. 239, 240

[588] T. Pouli and E. Reinhard. Progressive color transfer for images of arbitrary dynamic range. *Computers and Graphics*, 35(1):67–80, 2011. 240, 241

[589] E. Prados and O. Faugeras. Shape from shading. In N. Paragios, Y. Chen, and O. Faugeras, editors, *Handbook of Mathematical Models in Computer Vision*, pages 375–388. Springer, New York, 2006. 34, 267

[590] K. Prazdny. Illusory contours from inducers defined solely by spatiotemporal correlation. *Perception & Psychophysics*, 39(3):175–178, 1986. 278

[591] P. Prusinkiewicz and A. Lindenmayer. *The Algorithmic Beauty of Plants*. Springer-Verlag, 1990. 147

[592] D. Purves and R. B. Lotto. *Why We See What We Do Redux: A Wholly Empirical Theory of Vision*. Sinauer Associates, Inc., Sunderland, MA, 2011. 261

[593] D. Purves, A. Shimpi, and R. B. Lotto. An empirical explanation of the Cornsweet effect. *The Journal of Neuroscience*, 19(19):8542–8551, 1999. 37, 38, 39

[594] N. Qian and R. Andersen. Transparent motion perception as detection of unbalanced motion signals. II. Physiology. *Journal of Neuroscience*, 14(12):7367–7380, 1994. 275

[595] S. J. M. Rainville and F. A. A. Kingdom. Spatial-scale contribution to the detection of mirror symmetry in fractal noise. *Journal of the Optical Society of America A*, 16(9):2112–2123, 1999. 144

[596] A. Rajagopalan and S. Chaudhuri. Optimal recovery of depth from defocused images using an MRF model. In *Proceedings of the 6th IEEE International Conference on Computer Vision*, pages 1047–1052, 1998. 268

[597] A. Rajagopalan and S. Chaudhuri. An MRF model-based approach to simultaneous recovery of depth and restoration from defocused images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):577–589, 1999. 268

[598] V. S. Ramachandran. Perceiving shape from shading. *Scientific American*, 256(8):76–83, 1988. 3

[599] V. S. Ramachandran. Perception of shape from shading. *Nature*, 331(6152):163–166, 1988. 85

[600] M. Ranzato, V. Mnih, and G. E. Hinton. Generating more realistic images using gated MRF's. In *NIPS-23: Proceedings of the 2010 Conference on Advances in Neural Information Processing Systems*, pages 2002–2010, 2010. 220

[601] R. P. Rao, B. A. Olshausen, and M. S. Lewicki. *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press, Cambridge, MA, 2002. 261

[602] C. Redies. A universal model of esthetic perception based on the sensory coding of natural stimuli. *Spatial Vision*, 21(1–2):1–2, 2007. 6, 151, 152

[603] C. Redies, S. A. Amirshahi, M. Koch, and J. Denzler. PHOG-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects. In *European Conference on Computer Vision, Workshops and Demonstrations*, pages 522–531, 2012. 104

[604] C. Redies, J. Hänisch, M. Blickhan, and J. Denzler. Artists portray human faces with the Fourier statistics of complex natural scenes. *Network: Computation in Neural Systems*, 18(3):235–248, 2007. 152

[605] C. Redies, J. Hasenstein, and J. Denzler. Fractal-like image statistics in visual art: Similarity to natural scenes. *Spatial Vision*, 21(1–2):97–117, 2007. 151

[606] T. H. Reeves and N. G. Kingsbury. Overcomplete image coding using iterative projection-based noise shaping. In *Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages 597–600, 2002. 200

[607] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 76, 133, 135, 236, 237, 241, 243

[608] E. Reinhard, A. Efros, J. Kautz, and H.-P. Seidel. On visual realism of synthesized imagery. *Proceedings of the IEEE*, 101(9), 2013. 34

[609] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski. *High Dynamic Range Imaging: Acquisition, Display and Image-Based Lighting*. Morgan Kaufmann, 2nd edition, 2010. 44, 46, 47

[610] E. Reinhard, E. A. Khan, A. O. Akyüz, and G. M. Johnson. *Color Imaging: Fundamentals and Applications*. A K Peters, Wellesley, 2008. 18, 54, 55, 57, 60, 75, 90, 234, 244, 251

[611] E. Reinhard and T. Pouli. Colour spaces for colour transfer. In R. Schettini, S. Tominaga, and A. Trémeau, editors, *IAPR Computational Color Imaging Workshop*, volume 6626 of *Lecture Notes in Computer Science*, pages 1–15. Springer, Berlin, 2011 (invited paper). 243

[612] E. Reinhard, P. Shirley, M. Ashikhmin, and T. Troscianko. Second order image statistics in computer graphics. In *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*, pages 99–106, New York, 2004. ACM. 132, 133, 139

[613] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267–276, 2002. 192

[614] A. G. Rempel. Fast progressive transmission of images using wavelets with sorted coefficients. Master's thesis, University of Saskatchewan, Canada, 1993. 203

[615] W. Richards. A lightness scale from image intensity distributions. *Applied Optics*, 21(14):2569–2604, 1982. 24

[616] F. Riecke, D. A. Bodanr, and W. Bialek. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London B*, 262(1365):259–265, 1995. 4

[617] L. A. Riggs. Visual acuity. In C. H. Graham, editor, *Vision and Visual Perception*. John Wiley & Sons, Inc., New York, 1965. 31

[618] D. L. Ringach, R. M. Shapley, and M. J. Hawken. Orientation selectivity in macaque V1: Diversity and laminar dependence. *Journal of Neuroscience*, 22(13):5639–5651, 2002. 29

[619] M. Ristivojevic and J. Konrad. Joint space-time image sequence segmentation: Object tunnels and occlusion volumes. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 9–12, 2004. 271

[620] P. F. C. de Rivaz and N. G. Kingsbury. Complex wavelet features for fast texture image retrieval. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 109–113, 1999. 200

[621] P. F. C. de Rivaz and N. G. Kingsbury. Bayesian image deconvolution and denoising using complex wavelets. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 273–276, 2001. 200

[622] J. O. Robinson. *The Psychology of Visual Illusion*. Dover Publications, Mineola, NY, 1998. 262

[623] C. Rocchini, P. Cignoni, C. Montani, P. Pingi, and R. Scopigno. A low cost 3D scanner based on structured light. *Computer Graphics Forum*, 20(3):299–308, 2001. 52

[624] I. Rock and F. Halper. Form perception without a retinal image. *American Journal of Psychology*, 82(4):425–440, 1969. 278

[625] H. R. Rodman and T. D. Albright. Coding of visual stimulus velocity in area MT of the macaque. *Vision Research*, 27(12):2035–2048, 1987. 283

[626] H. R. Rodman, K. M. Sorenson, A. J. Shim, and D. P. Hexter. Calbindin immunoreactivity in the geniculo-extrastriate system of the macaque: Implications for the heterogeneity in the koniocellular pathway and recovery from cortical damage. *Journal of Comparative Neurology*, 431(2):168–181, 2001. 29

[627] B. Rogowitz and R. Voss. Shape perception and low-dimension fractal boundaries. *Proceedings of the SPIE*, 1249:387–394, 1990. 145, 146

[628] J. K. Romberg, H. Choi, and R. G. Baraniuk. Multiscale edge grammars for complex wavelet transforms. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 614–617, 2003. 199

[629] J. K. Romberg, H. Choi, R. G. Baraniuk, and N. G. Kingsbury. Multiscale classification using complex wavelets and hidden Markov tree models. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 371–374, 2000. 200

[630] A. Roorda and D. R. Williams. The arrangement of the three cone classes in the living human eye. *Nature*, 397(6719):520–522, 1999. 24

[631] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 860–867, Washington, DC, 2005. 7, 219

[632] S. Roth and M. J. Black. On the spatial statistics of optical flow. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, volume 1, pages 42–49, 2005. 283

[633] S. Roth and M. J. Black. On the spatial statistics of optical flow. *International Journal of Computer Vision*, 74(1):33–50, 2007. 284

[634] S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205–229, 2009. 219

[635] A. R. Rouse and A. F. Gmitro. Multispectral imaging with a confocal microendoscope. *Optics Letters*, 25(23):1708–1710, 2000. 50

[636] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, Hoboken, NJ, 1987. 163

[637] D. L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994. 4, 97, 133, 135

[638] D. L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, 1997. 136, 140, 146, 259

[639] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994. 4, 97, 133, 135

[640] D. L. Ruderman, T. W. Cronin, and C. C. Chiao. Statistics of cone responses to natural images: Implications for visual coding. *Journal of the Optical Society of America A*, 15(8):2036–2045, 1998. 42, 49, 166, 231, 234, 241, 243

[641] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, Manchester, UK, 1992. 158

[642] A. H. Salden, B. M. ter Haar Romeny, and M. A. Viergever. Differential and integral geometry of linear scale-spaces. *Journal of Mathematical Imaging and Vision*, 9(1):5–27, 1998. 110

[643] D. Salomon. *Transformations and Projections in Computer Graphics*. Springer, 2006. 49

[644] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado. A state of the art in structured light patterns for surface profilometry. *Pattern Recognition*, 43(8):2666–2680, 2010. 52

[645] T. D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2(6):459–473, 1989. 4

[646] Y. Sato and K. Ikeuchi. Temporal-color space analysis of reflection. *Journal of the Optical Society of America A*, 11(11):2990–3002, 1994. 276

[647] M. Savilli, G. Lecoy, and J. Nougier. *Noise in Physical Systems and $1/f$ Noise*. Elsevier, New York, 1983. 217

[648] H. Sawhney and R. Kumar. True multi-image alignment and its application to mosaicing and lens distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(3):235–243, 1999. 58

[649] A. Saxena, S. H. Chung, and A. Ng. Learning depth from single monocular images. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *NIPS-20: Proceedings of the 2006 Conference on Advances in Neural Information Processing Systems*, pages 1161–1168. MIT Press, Cambridge, MA, 2006. 268

[650] A. Saxena, S. H. Chung, and A. Y. Ng. 3-D depth reconstruction from a single still image. *International Journal of Computer Vision*, 76(1):53–69, 2008. 268

[651] A. Saxena, M. Sun, and A. Y. Ng. Learning 3-D scene structure from a single still image. In *Proceedings of the 11th IEEE International Conference on Computer Vision*, pages 1–8, 2007. 268

[652] A. Saxena, M. Sun, and A. Y. Ng. Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009. 268

[653] A. van der Schaaf. *Natural Image Statistics and Visual Processing*. PhD thesis, Rijksuniversiteit Groningen, The Netherlands, 1998. 129, 132

[654] A. van der Schaaf and J. H. van Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770, 1996. 42, 133, 135, 140, 146

[655] H. R. Schiffman and J. G. Thompson. The role of figure orientation and apparent depth in the perception of the horizontal-vertical illusion. *Perception*, 4(1):79, 1975. 261

[656] U. Schmidt, Q. Gao, and S. Roth. A generative perspective on MRFs in low-level vision. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1751–1758, 2010. 220, 221, 222

[657] P. Schröder and W. Sweldens. Spherical wavelets: Efficiently representing functions on the sphere. In *SIGGRAPH 95: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pages 161–172. ACM, 1995. 49

[658] L. Schwabe, K. Obermayer, A. Angelucci, and P. C. Bressloff. The role of feedback in shaping the extra-classical receptive field of cortical neurons: A recurrent network model. *Journal of Neuroscience*, 26(36):9117–9129, 2006. 207

[659] O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, 2001. 196

[660] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 519–528, 2006. 51

[661] R. Sekuler, H. R. Wilson, and C. Owsley. Structural modeling of spatial vision. *Vision Research*, 24(7):689–700, 1984. 144

[662] I. W. Selesnick. The design of Hilbert transform pairs of wavelet bases via the flat delay filter. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3673–3676, 2001. 199

[663] I. W. Selesnick. Hilbert transform pairs of wavelet bases. *IEEE Signal Processing Letters*, 8(6):170–173, 2001. 199

[664] I. W. Selesnick, R. G. Baraniok, and N. G. Kingsbury. The dual-tree complex wavelet transform: A coherent framework for multiscale signal and image processing. *IEEE Signal Processing Magazine*, 22(6):123–151, 2005. 197, 199

[665] I. W. Selesnick and K.-L. Li. Video denoising using 2D and 3D dual-tree complex wavelet transforms. In *Proceedings of Wavelet Applications and Signal Processing X*, pages 607–618, 2003. 199

[666] C. W. Shaffrey, N. G. Kingsbury, and I. H. Jermyn. Unsupervised image segmentation via Markov trees and complex wavelets. In *Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages 801–804, 2002. 200

[667] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. *ACM Transactions on Graphics*, 27(3):1–10, 2008. 6, 7, 115

[668] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993. 176

[669] L. Shapiro and G. C. Stockman. *Computer Vision*. Prentice Hall, Upper Saddle River, NJ, 2001. 102

[670] L. Sharan, Y. Li, I. Motoyoshi, S. Nishida, and E. H. Adelson. Image statistics for surface reflectance perception. *Journal of the Optical Society of America A*, 25(4):846–865, 2008. 82

[671] H. R. Sheikh, A. C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, 2005. 193

[672] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, 2005. 193

[673] J. Shen, X. Jin, C. Zhou, and C. C. L. Wang. Gradient based image completion by solving Poisson equation. *Computers and Graphics*, 31(1):119–126, 2007. 118

[674] S. L. Sherwood, editor. *The Nature of Psychology: A Selection of Papers, Essays and Other Writings by Kenneth J W Craik*. Cambridge University Press, Cambridge, UK, 1966. 36

[675] F. Shi and I. W. Selesnick. Video denoising using oriented complex wavelet transforms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 949–952, 2004. 199

[676] F. Shi, I. W. Selesnick, and S. Cai. Image sharpening via image denoising in the complex wavelet domain. In *Proceedings of Wavelet Applications and Signal Processing X*, pages 467–474, 2003. 200

[677] S. Shih, Y. Hung, and W. Lin. When should we consider lens distortion in camera calibration? *Pattern Recognition*, 28(3):447–461, 1995. 55

[678] T. F. Shipley and D. W. Cunningham. Perception of occluding and occluded objects over time: Spatiotemporal segmentation and unit formation. In T. F. Shipley and P. J. Kellman, editors, *From Fragments to Objects: Segmentation and Grouping in Vision*, pages 557–585. Elsevier Science, Oxford, UK, 2001. 278

[679] T. F. Shipley and P. J. Kellman. Optical tearing in spatiotemporal boundary formation: When do local element motions produce boundaries, form, and global motion? *Spatial Vision*, 7(4):323–339, 1993. 278

[680] T. F. Shipley and P. J. Kellman. Spatiotemporal boundary formation: Boundary, form, and motion perception from transformations of surface elements. *Journal of Experimental Psychology: General*, 123(1):3–20, 1994. 278

[681] T. F. Shipley and P. J. Kellman. Spatiotemporal boundary formation: The role of local motion signals in boundary perception. *Vision Research*, 37(10):1281–1293, 1997. 278

[682] H. Sidenbladh and M. J. Black. Learning image statistics for Bayesian tracking. In *Proceedings of the 8th IEEE International Conference on Computer Vision*, volume 2, pages 709–716, 2001. 276

[683] F. Sigernes, M. Dyrland, N. Peters, D. Lorentzen, T. Svenøe, K. Heia, S. Chernouss, C. Deehr, and M. Kosch. The absolute sensitivity of digital colour cameras. *Optics Express*, 17(22):20211–20220, 2009. 49, 50

[684] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *Proceedings of the 31st Asilomar Conference on Signals, Systems and Computers*, pages 673–678, 1997. 97, 197

[685] E. P. Simoncelli. Modeling the joint statistics of images in the wavelet domain. In *Proceedings of the SPIE 44th Annual Meeting*, volume 3813, pages 188–195, 1999. 191

[686] E. P. Simoncelli. Statistical modeling of photographic images. In A. Bovic, editor, *Handbook of Video and Image Processing*. Academic Press, Burlington, MA, 2nd edition, 2005. 7

[687] E. P. Simoncelli and E. H. Adelson. Noise removal via Bayesian wavelet coring. In *Proceedings of the 3rd IEEE International Conference on Image Processing*, volume I, pages 379–392, 1996. 191, 193

[688] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger. Probability distributions of optical flow. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 310–315, 1991. 280, 281, 282

[689] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proceedings of the 2nd IEEE International Conference on Image Processing*, pages 444–447, 1995. 206

[690] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, 1992. 206

[691] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Reviews of Neuroscience*, 24(1):1193–1216, 2001. 4

[692] E. P. Simoncelli and J. Portilla. Texture characterization via joint statistics of wavelet coefficient magnitudes. In *Proceedings of the 5th IEEE International Conference of Image Processing*, volume I, pages 62–66, 1998. 9

[693] L. C. Sincich, K. F. Park, M. J. Wohlgemuth, and J. C. Horton. Bypassing V1: A direct geniculate input to area MT. *Nature Neuroscience*, 7(10):1123–1128, 2004. 29

[694] F. Sinz and M. Bethge. Temporal adaptation enhances efficient contrast gain control on natural images. *PLoS Computational Biology*, 9(1), 2013. 24

[695] K. Sivaramakrishnan and T. Nguyen. A uniform transform domain video codec based on dual tree complex wavelet transform. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1821–1824, 2001. 200

[696] G. Smith and D. A. Atchison. *The Eye and the Visual Optical Instruments*. Cambridge University Press, New York, 1997. 30

[697] V. C. Smith and J. Pokorny. Spectral sensitivity of the foveal cone photopigments between 400 and 500 nm. *Vision Research*, 15(2):161–171, 1975. 62

[698] B. Smits, P. Shirley, and M. M. Stark. Direct ray tracing of displacement mapped triangles. In *Proceedings of the Eurographics Workshop on Rendering*, pages 307–318, 2000. 148

[699] R. Snowden, P. Thompson, and T. Troscianko. *Basic Vision: An Introduction to Visual Perception*. Oxford University Press, Oxford, UK, 2nd edition, 2012. 17

[700] S. G. Solomon and P. Lennie. The machinery of colour vision. *Nature Reviews Neuroscience*, 8(4):276–286, 2007. 21, 24

[701] R. van Spaendonck, T. Blu, R. Baraniuk, and M. Vetterli. Orthogonal Hilbert transform filter banks and wavelets. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 6, pages 505–508, 2003. 199

[702] P. Spagnolo, T. Orazio, M. Leo, and A. Distante. Moving object segmentation by background subtraction and temporal analysis. *Image and Vision Computing*, 24(5):411–423, 2006. 275

[703] M. Srinivasan, S. Laughlin, and A. Dubs. Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London B*, 216(1205):427–459, 1982. 111

[704] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, 2003. 5, 11, 193

[705] P. J. Stappers. Forms can be recognized from dynamic occlusion alone. *Perceptual and Motor Skills*, 68(1):243–251, 1989. 278

[706] G. Stein. Lens distortion calibration using point correspondences. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 602–608, 1997. 58

[707] A. Stockman, D. MacLeod, and N. Johnson. Spectral sensitivities of the human cones. *Journal of the Optical Society of America A*, 10(12):2491–2521, 1993. 21, 229

[708] A. Stockman and L. T. Sharpe. The spectral sensitivities of the middle-and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research*, 40(13):1711, 2000. 62

[709] D. G. Stork. Computer vision and computer graphics analysis of paintings and drawings: An introduction to the literature. In X. Jiang and N. Petkov, editors, *Computer Analysis of Images and Patterns*, volume 5702 of *Lecture Notes in Computer Science*, pages 9–24. Springer, 2009. 6

[710] D. G. Stork and M. K. Johnson. Lighting analysis of diffusely illuminated tableaus in realist paintings: An application to detecting “compositing" in the portraits of Garth Herrick. In *Media Forensics and Security*, page 72540, 2009. 84

[711] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley, MA, 2nd edition, 1996. 176

[712] P. Sturm. Multi-view geometry for general camera models. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 206–212, 2005. 58

[713] C.-C. Su, A. C. Bovik, and L. K. Cormack. Natural scene statistics of color and range. In *Proceedings of the IEEE International Conference on Image Processing*, pages 257–260, 2011. 257

[714] T. Suk and J. Flusser. Combined blur and affine moment invariants and their use in pattern recognition. *Pattern Recognition*, 36(12):2895–2907, 2003. 75

[715] T. Suk and J. Flusser. Affine moment invariants of color images. In *Computer Analysis of Images and Patterns*, pages 334–341, Berlin, 2009. Springer. 75

[716] R. Swaminathan, S. B. Kang, R. Szeliski, A. Criminisi, and S. K. Nayar. On the motion and appearance of specularities in image sequences. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proceedings of the 8th European Conference on Computer Vision*, volume 2350 of *Lecture Notes in Computer Science*, pages 508–523. Springer, Berlin, 2002. 276

[717] E. Switkes, M. J. Mayer, and J. A. Sloan. Spatial frequency analysis of the visual environment: Anisotropy and the carpentered environment hypothesis. *Vision Research*, 18(10):1393–1399, 1978. 136, 140, 146

[718] R. Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2(1):1–104, 2006. 48

[719] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2010. 51, 54, 57, 59

[720] R. Szeliski and H. Shum. Creating full view panoramic image mosaics and environment maps. In *SIGGRAPH 97: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, pages 251–258. ACM Press/Addison-Wesley Publishing Co., 1997. 47

[721] Y. Tadmor and D. J. Tolhurst. Both the phase and the amplitude spectrum may determine the appearance of natural images. *Vision Research*, 33(1):141–145, 1993. 140

[722] Y. Tadmor and D. J. Tolhurst. Discrimination of changes in the second order statistics of natural and synthetic images. *Vision Research*, 34(4):541–554, 1994. 144

[723] Y. Tadmor and D. J. Tolhurst. Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. *Vision Research*, 40(22):3145–3157, 2000. 111

[724] Y. Tai, J. Jia, and C. Tang. Local color transfer via probabilistic segmentation by expectation-maximization. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 747–754, Washington, DC, 2005. 173

[725] H. Tang, H. Shu, J.-L. Dillenseger, X. D. Bao, and L. M. Luo. Moment-based metrics for mesh simplification. *Computers & Graphics*, 31(5):710–718, 2007. 75

[726] M. F. Tappen, B. C. Russell, and W. T. Freeman. Exploiting the sparse derivative prior for super-resolution and image demosaicing. In *Proceedings of the 3rd International Workshop on Statistical and Computational Theories of Vision at ICCV*, 2003. 7, 115, 116

[727] R. P. Taylor, R. Guzman, T. P. Martin, G. D. R. Hall, A. P. Micolich, D. Jonas, B. C. Scannell, M. S. Fairbanks, and C. A. Marlow. Authenticating Pollock paintings using fractal geometry. *Pattern Recognition Letters*, 28(6):695–702, 2007. 152

[728] R. P. Taylor, A. P. Micolich, and D. Jonas. Fractal analysis of Pollock's drip paintings. *Nature*, 399(6735):422, 1999. 152

[729] R. P. Taylor, B. Spehar, J. A. Wise, C. W. Clifford, B. R. Newell, C. Hagerhall, T. Purcell, and T. P. Martin. Perceptual and physiological responses to the visual complexity of fractal patterns. *Nonlinear Dynamics, Psychology and Life Sciences*, 9:89–114, 2005. 146

[730] Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton. Energy-based models for sparse overcomplete representations. *The Journal of Machine Learning Research*, 4:1235–1260, 2003. 220

[731] P. C. Teo and D. J. Heeger. Perceptual image distortion. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 982–986, 1994. 196

[732] B. M. ter Haar Romeny. *Front-End Vision and Multi-Scale Image Analysis*. Kluwer Academic Publisher, Dordrecht, Netherlands, 2011. 105, 109

[733] L. Theis, R. Hosseini, and M. Bethge. Mixtures of conditional Gaussian scale mixtures applied to multiscale image representations. *PLoS ONE*, 7(7):e39857, 2012. 173

[734] B. Thompson. Canonical correlation analysis. In B. Everitt and D. Howell, editors, *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, 2005. 156

[735] R. F. Thompson. *The Brain: A Neuroscience Primer*. Worth Publishers, New York, 3rd edition, 2000. 103

[736] W. B. Thompson, R. W. Fleming, S. H. Creem-Regehr, and J. K. Stefanucci. *Visual Perception from a Computer Graphics Perspective*. CRC Press/AK Peters, Boca Raton, FL, 2011. 30, 274, 275, 276

[737] M. G. A. Thomson. Higher-order structure in natural scenes. *Journal of the Optical Society of America A*, 16(7):1549–1553, 1999. 142, 143

[738] M. G. A. Thomson. Beats, kurtosis and visual coding. *Network: Computation in Neural Systems*, 12(3):271–287, 2001. 143

[739] M. G. A. Thomson and D. H. Foster. Role of second- and third-order statistics in the discriminability of natural images. *Journal of the Optical Society of America A*, 14(9):2081–2090, 1997. 133, 135

[740] G. Tkacik, P. Garrigan, C. Ratliff, G. Milcinski, J. M. Klein, L. H. Seyfarth, P. Sterling, D. H. Brainard, and V. Balasubramanian. Natural images from the birthplace of the human eye. *PLoS ONE*, 6(6):e20409, 2011. 61

[741] G. Tkacik, J. S. Prentice, J. D. Victor, and V. Balasubramanian. Local statistics in natural scenes predict the saliency of synthetic textures. *Proceedings of the National Academy of Sciences*, 107(42):18149–18154, 2010. 12, 111, 175

[742] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen. A versatile HDR video production system. *ACM Transactions on Graphics*, 30(4):41, 2011. 45, 46

[743] J. T. Todd and E. Mingolla. Perception of surface curvature and direction of illuminant from patterns of shading. *Journal of Experimental Psychology: Human Perception and Performance*, 9(4):583–595, 1983. 85

[744] A. Toet. Natural colour mapping for multiband nightvision imagery. *Information Fusion*, 4(3):155–166, 2003. 241

[745] D. J. Tolhurst and Y. Tadmor. Discrimination of changes in the slopes of the amplitude spectra of natural images: Band-limited contrast and psychometric functions. *Perception*, 26(8):1011–1025, 1997. 147

[746] D. J. Tolhurst, Y. Tadmor, and T. Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992. 133, 135, 140

[747] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 839–846, Washington, DC, 1998. 106

[748] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003. 267

[749] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS-17: Proceedings of the 2004 Conference on Advances in Neural Information Processing Systems*, pages 1401–1408, 2004. 267

[750] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1226–1238, 2002. 267, 268

[751] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003. 43, 133, 135, 136, 140, 146

[752] A. Torralba and P. Sinha. Statistical context priming for object detection. In *Proceedings of the 8th IEEE International Conference on Computer Vision*, volume 1, pages 763–770, 2001. 267

[753] M. Trentacoste, R. Mantiuk, W. Heidrich, and F. Dufrot. Unsharp masking, countershading and halos: Enhancements or artifacts? *Computer Graphics Forum*, 31(2):555–564, 2012. 106

[754] N. F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5), 2002. 277

[755] N. F. Troje and H. H. Bülthoff. Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36(12):1761–1771, 1996. 165

[756] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. 163, 164

[757] M. R. Turner. Texture discrimination by Gabor functions. *Biological Cybernetics*, 55(2):71–82, 1986. 203

[758] M. Unser. Wavelets, statistics, and biomedical applications. In *Proceedings of the 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, pages 244–249, 1996. 202

[759] T. Valentine. *Cognitive and Computational Aspects of Face Recognition: Explorations in Face Space*. Routledge, London, 1995. 164

[760] J. M. Valeton and D. van Norren. Light adaptation of primate cones: An analysis based on extracellular data. *Vision Research*, 23(12):1539–1547, 1983. 22

[761] J. Vazquez-Corral, C. Párraga, R. Baldrich, and M. Vanrell. Color constancy algorithms: Psychophysical evaluation on a new dataset. *Journal of Imaging Science*, 53(3):31105–31105, 2009. 62

[762] F. A. Vera-Diaz, R. L. Woods, and E. Peli. Shape and individual variability of the blur adaptation curve. *Vision Research*, 50(15):1452–1461, 2010. 147

[763] M. Vetterli and J. Kovacević. *Wavelets and Subband Coding*. Prentice Hall, Englewood Cliffs, NJ, 2005. 175

[764] W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000. 4

[765] Vision Club of Finland. Electronic shuttering for high speed CMOS machine vision applications. http://www.automaatioseura.fi/jaostot/mvn/mvn2007/parameter.html, 2007. 59

[766] R. Voss. Random fractal forgeries. In R. A. Earnshaw, editor, *Fundamental Algorithms for Computer Graphics*, pages 805–835. Springer, Berlin, 1985. 145

[767] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli. Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons. In R. Rao, B. Olshausen, and M. Lewicki, editors, *Statistical Theories of the Brain*, page 203. MIT Press, Cambridge, MA, 2002. 24, 195, 196

[768] M. Wakin, M. Orchard, R. G. Baraniuk, and V. Chandrasekaran. Phase and magnitude perceptual sensitivities in nonredundant complex wavelet representations. In *Proceedings of the Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1413–1417, 2003. 199

[769] H. Wallach. Über visuell warhgenomme Bewegungsrichtung. *Psychologische Forshcung*, 20(1):325–380, 1935. 281

[770] C. Wallraven, R. Fleming, D. Cunningham, J. Rigau, M. Feixas, and M. Sbert. Categorizing art: Comparing humans and computers. *Computers & Graphics*, 33(4):484–495, 2009. 84

[771] G. Walls. The filling-in process. *American Journal of Optometry*, 31(7):329–340, 1954. 36, 112

[772] P. van Walree. Spherical aberration. http://toothwalker.org/optics/spherical.html. 56

[773] W. Wan and Z. Yang. Statistics of three-dimensional natural scene structures. *Journal of Vision*, 12(9):1203–1203, 2012. 257

[774] B. A. Wandell. *Foundations of Vision*. Sinauer Associates, Inc., Sunderland, MA, 1995. 22, 49

[775] B. Wang, Y. Wang, I. W. Selesnick, and A. Vetro. An investigation of 3D dual-tree wavelet transforms for video coding. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 1317–1320, 2004. 200

[776] C. Wang, Y. Huang, and M. Huang. An effective algorithm for image sequence color transfer. *Mathematical and Computer Modelling*, 44(7–8):608–627, 2006. 241

[777] C. Wang and H. Yao. Sensitivity of V1 neurons to direction of spectral motion. *Cerebral Cortex*, 21(4):964–973, 2011. 29

[778] J. Wang and M. M. Oliveira. A hole-filling strategy for reconstruction of smooth surfaces in range images. In *XVI Brazilian Symposium on Computer Graphics and Image Processing*, pages 11–18. IEEE, 2003. 53

[779] L. Wang, Y. Zhao, W. Jin, S. Shi, and S. Wang. Real-time color transfer system for low-light level visible and infrared images in YUV color space. In *Proceedings of the SPIE*, volume 6567, page 65671G, 2007. 242

[780] H. Wässle and B. B. Boycott. Functional architecture of the mammalian retina. *Physiological Reviews*, 71(2):447–480, 1991. 25

[781] A. B. Watson and A. Ahumada. *A Look at Motion in the Frequency Domain*, volume 84352. National Aeronautics and Space Administration, Ames Research Center, Moffett Field, CA, 1983. 270, 274

[782] J. B. Weaver, X. Yansun, D. M. Healy Jr., and L. D. Cromwell. Filtering noise from images with wavelet transforms. *Magnetic Resonance in Medicine*, 21(2):288–295, 1991. 202

[783] M. A. Webster. Adaptation and visual coding. *Journal of Vision*, 11(5):1–23, 2011. 21

[784] M. A. Webster, M. A. Georgeson, and S. M. Webster. Neural adjustments to image blur. *Nature Neuroscience*, 5(9):839–840, 2002. 147

[785] M. A. Webster and E. Miyahara. Contrast adaptation and the spatial structure of natural images. *Journal of the Optical Society of America A*, 14(9):2355–2366, 1997. 133, 135, 136, 146

[786] M. A. Webster, Y. Mizokami, and S. M. Webster. Seasonal variations in the color statistics of natural images. *Network: Computation in Neural Systems*, 18(3):213–233, 2007. 227

[787] L.-Y. Wei and M. Levoy. Texture synthesis over arbitrary manifold surfaces. In *SIGGRAPH 01: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 355–360, 2001. 9

[788] J. Weickert, S. Ishikawa, and A. Imiya. Linear scale-space has first been proposed in Japan. *Journal of Mathematical Imaging and Vision*, 10(3):237–252, 1999. 105

[789] J. van de Weijer and T. Gevers. Color constancy based on the grey-edge hypothesis. In *Proceedings of the IEEE International Conference on Image Processing*, pages 722–725, 2005. 252

[790] J. van de Weijer, T. Gevers, and A. Gijsenij. Edge-based color constancy. *IEEE Transactions on Image Processing*, 16(9):220–2214, 2007. 252

[791] Y. Weiss and E. H. Adelson. Slow and smooth: A Bayesian theory for the combination of local motion signals in human vision. Technical report, AI Memo 1624, MIT AI Lab, Cambridge, MA, 1998. 284

[792] Y. Weiss, E. P. Simoncelli, and E. H. Adelson. Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604, 2002. 276

[793] M. Welling, G. E. Hinton, and S. Osindero. Learning sparse topographic representations with products of student-t distributions. In *NIPS-15: Proceedings of the 2002 Conference on Advances in Neural Information Processing Systems*, pages 1359–1366, 2002. 218

[794] J. Weng, P. Cohen, and M. Herniou. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(10):965–980, 1992. 57

[795] M. White. A new effect of patterns on perceived lightness. *Perception*, 8(4):413–416, 1979. 145

[796] N. Wiener. Generalized harmonic analysis. *Acta Mathematica*, 55(1):117–258, 1930. 128

[797] D. R. Williams and N. J. Coletta. Cone spacing and the visual resolution limit. *Journal of the Optical Society of America A*, 4(8):1514–1523, 1987. 31

[798] D. R. Williams and A. Roorda. The trichromatic cone mosaic in the human eye. In K. R. Gegenfurtner and L. T. Sharpe, editors, *Color Vision: From Genes to Perception*, pages 113–122. Cambridge University Press, Cambridge, 1999. 24

[799] R. G. Willson. *Modeling and Calibration of Automated Zoom Lenses*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1994. 55

[800] A. P. Witkin. Scale-space filtering. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 1019–1022, 1983. 105, 106

[801] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. 153

[802] W.-H. Wong, W.-C. Siu, and K.-M. Lam. Generation of moment invariants and their uses for character recognition. *Pattern Recognition Letters*, 16(2):115–123, 1995. 75

[803] Y. N. Wu, C. E. Guo, and S. C. Zhu. From information scaling of natural images to regimes of statistical models. *Quarterly of Applied Mathematics*, 66(1):81–122, 2008. 193

[804] G. Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley & Sons, New York, 2nd edition, 2000. 232

[805] Y. Xiang, B. Zou, and H. Li. Selective color transfer with multi-source images. *Pattern Recognition Letters*, 30(7):682–689, 2009. 241

[806] X. Xiao and L. Ma. Color transfer in correlated color space. In *Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and Its Applications*, pages 305–309, New York, 2006. ACM. 242

[807] X. Xiao and L. Ma. Gradient-preserving color transfer. *Computer Graphics Forum*, 28(7):1879–1886, 2009. 241

[808] S. Xu, Y. Zhang, S. Zhang, and X. Ye. Uniform color transfer. In *Proceedings of the IEEE International Conference on Image Processing*, pages 940–943, 2005. 241

[809] Z. Yang and D. Purves. Image/source statistics of surfaces in natural scenes. *Network: Computation in Neural Systems*, 14(3):371–390, 2003. 263

[810] Z. Yang and D. Purves. A statistical explanation of visual space. *Nature Neuroscience*, 6(6):632–640, 2003. 263

[811] Z. Ye and C.-C. Lu. A complex wavelet domain Markov model for image denoising. In *Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages 365–368, 2003. 199

[812] J. Yin and J. R. Cooperstock. Color correction methods with applications to digital projection environments. *Journal of the WSCG*, 12(1–3), 2004. 241

[813] S. Yoshimura and T. Kanade. Fast template matching based on the normalized correlation by using multiresolution eigenimages. In *Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems (Advanced Robotic Systems and the Real World, IROS 94)*, volume 3, pages 2086–2093, 1994. 163

[814] T. Young. The Bakerian lecture: On the theory of light and colors. *Philosophical Transactions of the Royal Society of London*, 92:12–48, 1802. 32, 228

[815] M. Yukie and E. Iwai. Direct projection from the dorsal lateral geniculate nucleus to the prestriate cortex in macaque monkeys. *Journal of Comparative Neurology*, 201(1):81–97, 1981. 29

[816] J. Zanker and J. Zeil. Movement-induced motion signal distributions in outdoor scenes. *Network: Computation in Neural Systems*, 16(4):357–376, 2005. 283

[817] C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Research*, 30(7):1111–1117, 1990. 264

[818] C. Zetzsche and G. Krieger. Exploitation of natural scene statistics by orientation selectivity and cortical gain control. *Perception, ECVP Supplement*, 27:154, 1998. 193

[819] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999. 34, 85, 267

[820] Y. Zhao and R. S. Berns. Image-based spectral reflectance reconstruction using the matrix R method. *Color Research & Application*, 32(5):343–351, 2007. 241

[821] Y. Zheng, S. Lin, and S. Kang. Single-image vignetting correction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 461–468, 2010. 58, 59

[822] S. Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):1236–1250, 1997. 220

[823] S. C. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998. 203

[824] Z. Zhu, S. Wu, S. Rahardja, and P. Fränti. Real-time ghost removal for composing high dynamic range images. In *Proceedings of the 5th IEEE International Conference on Industrial Electronics and Applications*, pages 1627–1631, 2010. 193

[825] C. Ziegaus and E. W. Lang. Statistics of natural and urban images. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Proceedings of Artificial Neural Networks – ICANN 97*, volume 1327 of *Lecture Notes in Computer Science*, pages 219–224. Springer, Berlin, 1997. 42

[826] C. Ziegaus and E. W. Lang. Statistical invariances in artificial, natural and urban images. *Zeitschrift für Naturforschung A*, 53:1009–1021, 1998. 84

[827] R. E. Ziemer, W. H. Tranter, and D. R. Fannin. *Signals and Systems: Continuous and Discrete*. Prentice Hall, Upper Saddle River, NJ, 4th edition, 1993. 125

[828] V. Zlokolica, A. Pizurica, and W. Philips. Noise estimation for video processing based on spatio-temporal gradients. *IEEE Signal Processing Letters*, 13(6):337–340, 2006. 275

[829] M. Zontak and M. Irani. Internal statistics of a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 977–984, 2011. 6, 98, 99

[830] D. Zoran and Y. Weiss. Scale invariance and noise in natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2209–2216, 2009. 141

[831] D. Zoran and Y. Weiss. From learning models of natural image patches to whole image restoration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 479–486, 2011. 6, 172, 173, 174

[832] D. Zoran and Y. Weiss. Natural images, Gaussian mixtures and dead leaves. In *NIPS-25: Proceedings of the 2012 Conference on Advances in Neural Information Processing Systems*, pages 1745–1753, 2012. 172, 173

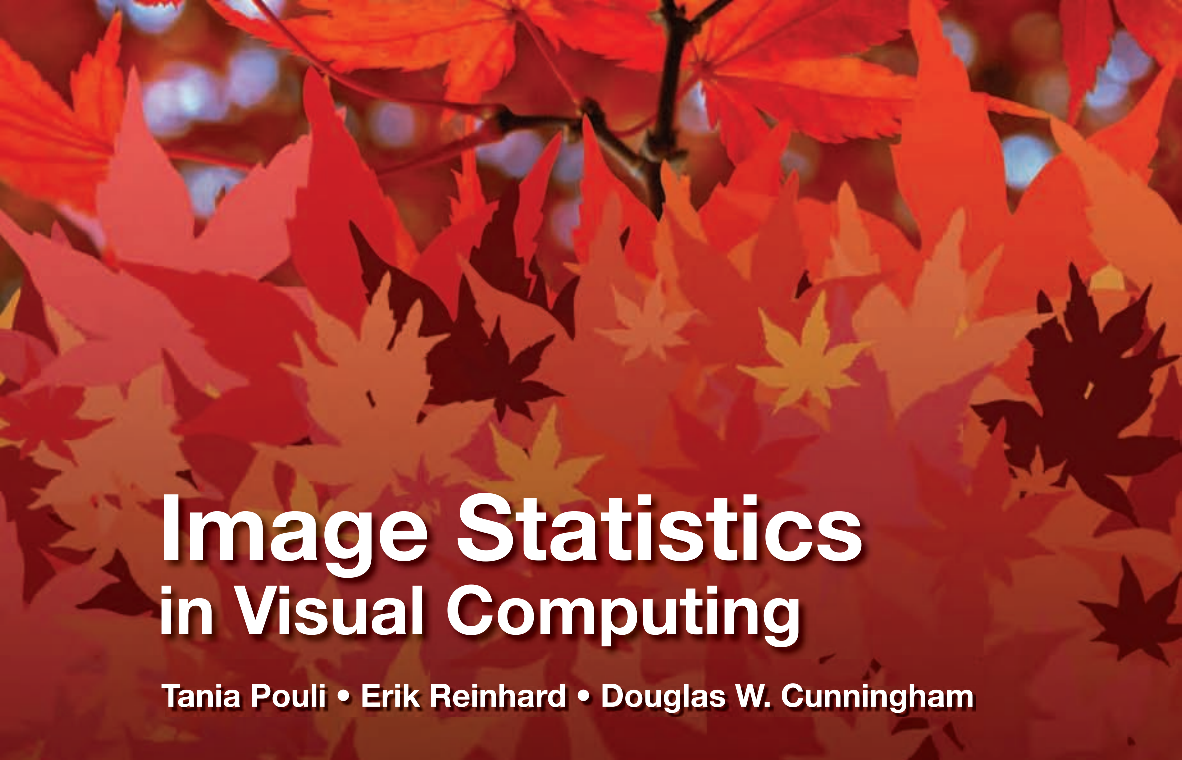


Image Statistics
in Visual Computing

Tania Pouli • Erik Reinhard • Douglas W. Cunningham

To achieve the complex task of interpreting what we see, our brains rely on statistical regularities and patterns in visual data. Knowledge of these regularities can also be considerably useful in visual computing disciplines, such as computer vision, computer graphics, and image processing. The field of natural image statistics studies the regularities to exploit their potential and better understand human vision. With numerous color figures throughout, **Image Statistics in Visual Computing** covers all aspects of natural image statistics, from data collection to analysis to applications in computer graphics, computational photography, image processing, and art.

The authors keep the material accessible, providing mathematical definitions where appropriate to help you understand the transforms that highlight statistical regularities present in images. The book also describes patterns that arise once the images are transformed and gives examples of applications that have successfully used statistical regularities. Numerous references enable you to easily look up more information about a specific concept or application. A supporting website also offers additional information, including descriptions of various image databases suitable for statistics.

Collecting state-of-the-art, interdisciplinary knowledge in one source, this book explores the relation of natural image statistics to human vision and shows how natural image statistics can be applied to visual computing. It encourages you to develop novel insights and applications in all disciplines that relate to visual computing.

K13473

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

Computer Graphics

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

ISBN: 978-1-56881-725-5
90000
9 781568 817255