Beeldmultiresolutiemodellen en schattingstechnieken

Multiresolution Image Models and Estimation Techniques

Bart Goossens

Promotoren: prof. dr. ir. W. Philips, prof. dr. ir. A. Pižurica
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: prof. dr. ir. H. Bruneel
Faculteit Ingenieurswetenschappen
Academiejaar 2009 - 2010

UNIVERSITEIT
GENT

ISBN 978-90-8578-351-0
NUR 959, 954
Wettelijk depot: D/2010/10.500/27

Members of the jury

prof. dr. Ingrid Daubechies (Princeton University)
prof. dr. Paul Scheunders (University of Antwerp)
prof. dr. ir. Luc Taerwe (Ghent University, chairman)
prof. dr. ir. Marc Moeneclaey (Ghent University, secretary)
prof. dr. ir. Wilfried Philips (Ghent University, promoter)
prof. dr. ir. Aleksandra Pižurica (Ghent University, promoter)
prof. dr. ir. Stefaan Vandenberghe (Ghent University)
dr. ir. Hiep Luong (Ghent University)
dr. lic. Ewout Vansteenkiste (Ghent University)

This research is supported by the special research fund (BOF) from Ghent University.

Affiliations

Research Group for Image Processing and Interpretation (IPI)
Interdisciplinary Institute for Broadband Technology (IBBT)
Department of Telecommunications and Information Processing (TELIN)
Faculty of Engineering
Ghent University

Sint-Pietersnieuwstraat 41
B-9000 Ghent
Belgium

Acknowledgements

I would like to gratefully acknowledge my advisors, prof. Wilfried Philips and prof. Aleksandra Pižurica, for following my research with great interest and for giving me constant feedback and support.

I would like to thank the members of my Ph.D. jury, prof. Luc Taerwe, prof. Ingrid Daubechies, prof. Paul Scheunders, prof. Marc Moeneclaey, prof. Stefaan Vandenberghe, dr. Hiêp Luong and dr. Ewout Vansteenkiste, for being in the jury, for reading the thesis text and for giving various comments and suggestions.

I am grateful to several colleagues who read draft versions of the text and who gave very useful feedback: Ljiljana Platiša (TELIN), Jan Aelterman (TELIN), Filip Rooms (TELIN) and Valérie De Witte (University of Antwerp). I would like to thank all the colleagues with whom I collaborated on a daily basis: Hiêp Luong, Jan Aelterman, Aleksandra Pižurica, Ewout Vansteenkiste, Filip Rooms, Johan De Bock, Ljiljana Platiša, Ivana Despotović, Vladimir Zlokolica, Tijana Ružić and Danilo Babin. Sincere thanks also to my other colleagues at the TELIN research department of the Ghent University, for the pleasant time at the department. Also many thanks to Annette Nevejans, Patrick Schaillée and Alice Verheylesonne for their administrative assistance and Philippe Serbruyns and Davy Moreels for their excellent IT support.

I would like to thank dr. Ewout Vansteenkiste and dr. Cédric Marchessoux (Barco) for the many interesting discussions and also for providing medical image data sets.

I would also like to acknowledge all the persons outside the university or the department for the nice discussions and correspondences. I would like to thank all the people that I met on the conferences, for the productive discussions and the great time at the conferences. Special thanks to my office mates: Hiêp Luong, Stefaan Lippens, Johan De Bock and Jorge Oswaldo Niño Castaneda for creating a nice atmosphere to work in. Last but not least, I would like to thank my parents for their support and patience during the work on this PhD.

ir. Bart Goossens

April, 2010.

Summary

In the last decades, the use of digital images has tremendously increased, due to the wide availability of imaging devices, such as digital still cameras and medical scanners (in hospitals). Simultaneously, image quality expectations of end users have augmented. To meet these expectations, manufacturers of imaging devices push the electronics in their devices up to their limits. However, physical limitations and cost restrictions pose stringent limits on the maximum achievable image quality. Also in medical imaging, it is desirable to lower the acquisition times, for patient comfort, less motion-induced artifacts and increased examination throughput. Especially important in x-ray imaging is the tendency to lower radiation doses in order to minimize health risks due to radiation exposure. In practice, a compromise between image quality, radiation dosis, acquisition time and other imaging parameters need to be made.

Consequently, acquired digital images are imperfect and often contain many degradations, such as noise and blur. It becomes increasingly important to extract more useful information out of these degraded images and to improve the quality of these images.

In this dissertation we develop digital post-processing techniques to restore the images after acquisition. We thereby employ a generic methodology that can easily be adapted to various practical applications. Based on improved multiresolution representations, we develop new and improved statistical models for images and image noise.

Correspondingly, the three pilars of this work are *multiresolution representations*, *image models* and *noise models*. Multiresolution representations describe images using a small number of significant coefficients, typically much less than the total number of pixels in the image. Statistical models describe the properties of ideal undegraded images and are used as prior knowledge for image restoration. Noise models characterize the statistical properties of image noise.

In the context of multiresolution representations, we improve the directional selectivity of the first-scale of the dual-tree complex wavelet transform, which results in a better reconstruction of fine details at arbitrary orientations. Next, we propose a new design for the discrete shearlet transform that offers low redundancy, shift invariance and high directional analysis properties, a combination of properties difficult to achieve with existing multiresolution representations.

We present two new statistical models for image multiresolution transform coefficients: a joint intra/inter-scale model, which incorporates in a novel way

both spatial correlations and inter-scale dependencies between transform coefficients, and an intra-scale model (MPGSM), which is a generalization and improved version of the well-known and popular GSM model. While these models attempt to fully exploit *local dependencies* between transform coefficients, we also investigate models for *non-local* dependencies. In particular, we introduce an improved image-domain non-local prior model for images. Experimental results show that the denoising techniques that make use of these two multiresolution models and non-local model are currently among the state-of-the-art methods in image denoising.

Because existing restoration techniques often make restrictive assumptions with respect to the noise statistics (e.g. they assume additive white Gaussian noise), these techniques generally perform poorly when applied to real images (as for example captured by a digital still camera). Therefore, we devote special attention to models for noise encountered in realistic scenarios. In particular, we study colored Gaussian noise, non-stationary noise and signal-dependent noise and we provide novel estimation techniques for estimating the parameters of these noise models. These noise models can easily be adapted to underlying processing techniques in the imaging devices.

In this respect, one of the contributions of this dissertation is a new statistical model for non-stationary noise in Computed Tomography (CT) images. Because of the signal-dependency of the projection data noise and because of the non-local character of the traditionally used *filtered backprojection* reconstruction algorithm, the statistics of noise in reconstructed images are very complicated. Our statistical model describes the position-dependent and orientation-dependent properties of the CT noise. We provide a parameter estimation technique for this model based on multiresolution concepts.

We also introduce a new complex-wavelet packet based demosaicing algorithm, which fully exploits the analyzing properties of the complex wavelets in order to reconstruct fine details and to avoid discoloration artifacts. This algorithm is particularly intriguing because it has a low computation complexity (which permits implementation on digital camera hardware) and can be easily extended to perform joint denoising and demosaicing.

Next, we explain how the "generic" *image models*, *noise models* and *multiresolution representations* can be jointly combined for a given restoration task. Consider for example the problem of noise reduction of images captured by a digital still camera. The photon energies measured by the sensor elements of the camera typically exhibit a Poisson distribution. By image reconstruction techniques and various post-processing techniques in the camera (such as gamma correction, color enhancement, digital zoom, demosaicing, ...) the statistics of the noise are significantly changed. Consequently, the image restoration problem becomes very difficult. Therefore, we illustrate how the Bregman optimization framework can be used to solve the more complicated image restoration problems. In particular, we present new techniques that exploit properties of the improved multiresolution transforms (e.g. shearlet transform) in order to perform joint denoising and deblurring, estimate and

remove correlated noise from images, and remove signal-dependent noise from images jointly with de-biasing.

Finally, we consider the automatic quality assessment of medical images. Here the goal is to objectively and automatically measure image quality. This is useful for validating restoration techniques for medical images and for improving future medical display systems. Traditionally, to assess the impact of new display technologies on the quality of the images as seen on the screen, time-consuming and costly psychovisual experiments involving human specialists need to be conducted. By automatically assessing the image quality, the need for these experiments is eliminated. Image quality is then measured by means of a "virtual specialist", which is a software program that is tuned to match human specialists in evaluating the images. In this dissertation, we present a novel framework for deriving model observers that detect abnormalities with random parameters. We find that model observers tuned for this detection task make use of multiresolution concepts, hence the quality assessment task is very related to the image restoration task. Because nowadays the use of 3D medical images becomes increasingly important, we also discuss novel extensions of the "virtual specialist" model to 3D image visualization.

Samenvatting

De laatste decennia is het gebruik van digitale beelden sterk toegenomen als gevolg van de ruime beschikbaarheid van beeldopnametoestellen, zoals digitale camera's en medische scanners (ondermeer in ziekenhuizen). Tegelijkertijd zijn ook de verwachtingen van de eindgebruikers over de beeldkwaliteit toegenomen. Om hieraan tegemoet te komen, trachten fabrikanten van beeldopnameapparatuur het maximale uit de elektronica in hun apparaten te halen. Maar omwille van fysische beperkingen en kostprijsredenen, is de maximaal haalbare beeldkwaliteit echter beperkt. Ook in de medische beeldvorming is het wenselijk om lage opnametijden te gebruiken, dit voor het comfort van de patiënt, minder bewegingsartefacten en kortere onderzoeken. Om gezondheidsrisico's te minimaliseren wenst men bij x-ray beelden een zo laag mogelijke stralingsdosis te gebruiken. In de praktijk moet er vaak een compromis tussen beeldkwaliteit, stralingsdosis, opnametijd en andere beeldvormingsparameters worden gemaakt.

Bijgevolg vertonen opgenomen digitale beelden vaak vele degradaties, zoals ruis en onscherpte. Het is dus belangrijk om de kwaliteit van de beelden te verbeteren om zo meer nuttige informatie te halen uit deze imperfecte beelden.

In dit proefschrift worden digitale naverwerkingstechnieken ontwikkeld om opgenomen beelden te verbeteren, aan de hand van een generieke methodologie die gemakkelijk kan worden aangepast voor diverse praktische toepassingen en die gebaseerd is op verbeterde multiresolutievoorstellingen. Verder worden nieuwe en verbeterde statistische modellen ontwikkeld voor beelden en de bijhorende ruis.

Het proefschrift steunt daarom op drie grote pijlers: *multiresolutievoorstellingen*, *statistische beeldmodellen* en *ruismodellen*. Multiresolutievoorstellingen beschrijven beelden met behulp van een klein aantal significante coëfficiënten, doorgaans is dit aantal veel minder dan het totale aantal pixels in het beeld. Statistische modellen beschrijven de eigenschappen van ideale, niet-gedegradeerde beelden en worden gebruikt als voorkennis voor de beeldrestauratie. Ruismodellen karakteriseren statistische eigenschappen van beeldruis.

In het kader van multiresolutievoorstellingen, wordt de richtingsgevoeligheid van de eerste schaal van de dual-tree complexe wavelettransformatie verbeterd. Hierdoor worden fijne details met willekeurige oriëntaties beter gereconstrueerd.

Vervolgens wordt een nieuw ontwerp ontwikkeld voor de discrete shearlettransformatie, dat een lage redundantie en verschuivingsinvariantie combineert met een fijne richtingsanalyse.

Verder worden twee nieuwe statistische modellen voorgesteld voor multiresolutiecoëfficiënten van beelden: een gezamenlijk intra-/inter-schaalmodel, waarin op een nieuwe manier spatiale correlaties en onderlinge afhankelijkheden tussen coëfficiënten in verschillende multiresolutieschalen in rekening worden gebracht en een intra-schaalmodel (MPGSM), dat een veralgemening en verbeterde versie is van het bekende en populaire GSM-model. Deze modellen benutten de *lokale* afhankelijkheden ten volle. In dit proefschrift worden ook *niet-lokale* afhankelijkheden behandeld aan de hand van een verbeterd niet-lokaal a priori model voor het beelddomein. Experimentele resultaten tonen aan dat de ruisonderdrukkingstechnieken die gebruik maken van deze modellen momenteel tot de state-of-the-art technieken behoren in de ruisonderdrukking voor beelden.

Omdat bestaande beeldrestauratietechnieken vaak beperkende veronderstelling maken ten aanzien van de ruiseigenschappen, zoals uitgaan van additieve witte Gaussiaanse ruis, presteren deze technieken in het algemeen slecht wanneer ze worden toegepast op beelden die opgenomen zijn met een digitale camera.

In dit proefschrift wordt daarom speciale aandacht besteed aan modellen voor ruis in realistische omstandigheden. In het bijzonder wordt gekleurde Gaussiaanse ruis, niet-stationaire ruis en signaal afhankelijke ruis bestudeerd. Nieuwe technieken voor het schatten van de parameters van deze ruismodellen komen aan bod. Deze ruismodellen kunnen gemakkelijk aangepast worden aan de onderliggende verwerkingsstappen in de beeldopnametoestellen.

In dit opzicht is één van de bijdragen van dit proefschrift een nieuw statistisch model voor niet-stationaire ruis in Computed Tomography (CT) beelden. Omwille van de signaalafhankelijkheid van de projectiedata en omwille van het niet-lokale karakter van het traditionele *"filtered backprojection"* reconstructiealgoritme, zijn de ruisstatistieken in gereconstrueerde CT beelden zeer ingewikkeld. Ons statistisch model beschrijft de positieafhankelijkheid en de oriëntatieafhankelijkheid van het lokale ruisvermogenspectrum van CT ruis. We bespreken een parameterschattingstechniek gebaseerd op multiresolutieconcepten.

Er wordt ook een nieuw demosaicing algoritme ontwikkeld, gebaseerd op complexe wavelet packets, dat volledig gebruik maakt van de eigenschappen van complexe wavelets om fijne details in de beelden te reconstrueren en om kleurartefacten te voorkomen. Dit bijzonder intrigerende algoritme heeft een lage rekencomplexiteit, wat een implementatie mogelijk maakt op digitale camerahardware. Het kan gemakkelijk worden uitgebreid om ruisonderdrukking en demosaicing gezamelijk uit te voeren.

Vervolgens wordt uitgelegd hoe de "generieke" beeldmodellen, ruismodellen en multiresolutievoorstellingen gecombineerd kunnen worden om een bepaald beeldrestauratieprobleem op te lossen. Neem bijvoorbeeld de ruisonderdrukking van beelden gemaakt met een digitale fotocamera. De fotonenergiëen gemeten door de sensorelementen van de camera vertonen algemeen beschouwd een Poissondistributie. Door beeldreconstructietechnieken en diverse naverwer-

kingsstappen in de camera (zoals gammacorrectie, kleurverbetering, digitale zoom, demosaicing, ...) worden de ruisstatistieken aanzienlijk beïnvloed. Het gebruik van het Bregmanoptimalisatieraamwerk wordt geïllustreerd om gecompliceerde beeldrestauratieproblemen op te lossen. Nieuwe technieken worden voorgesteld die de eigenschappen van de verbeterde multiresolutievoorstellingen ten volle benutten. Behandelde voorbeelden zijn gezamenlijke ruisonderdrukking en verscherping, schatting en verwijdering van gecorreleerde ruis in beelden, verwijdering van signaalafhankelijke ruis.

Tot slot wordt de automatische kwaliteitsbeoordeling van medische beelden behandeld. Het doel is hier om objectief en automatisch de beeldkwaliteit te bepalen. Dit is nuttig voor zowel het valideren van restauratietechnieken voor medische beelden als voor het verbeteren van toekomstige medische beeldschermen. Om de impact van nieuwe beeldschermtechnologiëen op de beeldkwaliteit vast te stellen, worden traditioneel tijdrovende en dure psychovisuele experimenten met medische specialisten uitgevoerd. Door de beeldkwaliteit automatisch te beoordelen, worden deze experimenten overbodig. De beeldkwaliteit wordt dan gemeten door middel van een "virtuele specialist." Dit is een softwareprogramma dat is afgestemd op artsen in het beoordelen van beelden. In dit proefwerk wordt een nieuw raamwerk voor virtuele specialistmodellen voor de detectie van afwijkingen met willekeurige parameters voorgesteld. De specialistmodellen, die afgestemd zijn op deze detectietaak, maken noodzakelijkerwijs gebruik van multiresolutieconcepten. De automatische kwaliteitsbeoordeling van medische beelden is dus sterk gerelateerd aan het probleem van beeldrestauratie. Omdat tegenwoordig het gebruik van 3D medische beelden steeds belangrijker wordt, worden ook nieuwe uitbreidingen van het virtuele specialistmodel voor 3D beeldvisualisatie besproken.

Contents

1 Introduction **1**
- 1.1 Problem statement and topical outline . . . 1
- 1.2 Contributions and list of publications . . . 3
- 1.3 Organization of this dissertation . . . 6

2 Multiresolution representations for images **9**
- 2.1 The discrete wavelet transform . . . 10
  - 2.1.1 A short introduction to wavelet theory . . . 10
  - 2.1.2 Multiresolution analysis and wavelet filters . . . 12
  - 2.1.3 The Fast DWT in higher dimensions . . . 14
  - 2.1.4 Problems with the DWT . . . 15
- 2.2 The dual-tree complex wavelet transform . . . 17
  - 2.2.1 One-dimensional complex wavelets . . . 18
  - 2.2.2 Higher dimensional complex wavelets: how directional selectivity is obtained. . . . 20
  - 2.2.3 Improving the Directional Selectivity of the First Scale . 23
- 2.3 The steerable pyramid transform . . . 30
  - 2.3.1 Steerability . . . 31
  - 2.3.2 Steerable filters . . . 36
  - 2.3.3 Architecture of the steerable pyramid transform . . . 38
- 2.4 Overview of related representations . . . 39
- 2.5 The shearlet transform . . . 42
  - 2.5.1 An introduction to shearlet theory . . . 43
  - 2.5.2 New design of the discrete shearlet transform . . . 48
- 2.6 Conclusion . . . 57

3 Statistical models for images **61**
- 3.1 Decomposition of images . . . 62
  - 3.1.1 Classical image model . . . 62
  - 3.1.2 Probabilistic PCA . . . 64
  - 3.1.3 Analysis of images in independent components . . . 67
  - 3.1.4 Related techniques . . . 71
- 3.2 Parametric densities . . . 72
  - 3.2.1 Generalized Laplace distribution . . . 73
  - 3.2.2 Weighted mixtures of two distributions . . . 74
  - 3.2.3 Elliptically symmetric distributions . . . 75

3.2.4 Gaussian Scale Mixtures 77
3.2.5 Bessel K Form density 78
3.2.6 Other densities 80
3.3 Joint statistics of subband coefficients 81
3.4 Models for intra-scale correlations 82
3.4.1 Markov Random Field models 82
3.4.2 Local spatial activity indicators 85
3.4.3 Mixtures of Gaussian Scale Mixtures 85
3.4.4 An improved model: MPGSM 91
3.5 Models for inter-scale dependencies 97
3.5.1 Hidden Markov Tree models 97
3.5.2 The Bivariate distribution of Şendur and Selesnick . . . 98
3.6 A novel joint inter/intra-scale model 99
3.7 Non-local image models 102
3.8 Conclusion 105

4 Noise modeling and estimation 107
4.1 Probability density functions for modeling noise 109
4.2 Second-order statistics of noise 112
4.2.1 From white to colored noise 112
4.2.2 Estimation of colored Gaussian noise 116
4.3 Modeling and estimation of non-stationary noise 121
4.3.1 Modeling of locally stationary Gaussian noise 123
4.3.2 Estimation of locally stationary Gaussian noise 124
4.4 Signal-dependent noise 130
4.4.1 Variance stabilization 131
4.4.2 Gaussian modeling of signal-dependent noise 132
4.4.3 Exact conditional moments $\mathrm{E}[y^n|x]$ 139
4.4.4 Possible extensions to the signal-dependent noise models 143
4.4.5 Estimation of signal-dependent noise 145
4.5 Conclusion 146

5 Digital image restoration 147
5.1 Image domain restoration 152
5.1.1 Overview of existing techniques 152
5.1.2 Improved Non-Local Means filter 154
5.2 Multiresolution image restoration 157
5.2.1 Existing multiresolution techniques for noise reduction . 157
5.2.2 Estimators for the univariate Bessel K distribution . . . 163
5.2.3 Vector-ProbShrink 166
5.2.4 MMSE estimation for MPGSM 169
5.2.5 Complex-wavelet based demosaicing 173
5.3 Bregman framework for image restoration 184
5.3.1 Splitting the Bregman iteration 185
5.3.2 Bayesian MAP estimation through Bregman optimization 188
5.3.3 Split Bregman based removal of correlated noise 190

5.3.4 Multiresolution joint denoising and deblurring 192
5.3.5 Joint signal-dependent noise and bias removal 194
5.4 Experimental results 197
5.4.1 Denoising results for white noise 198
5.4.2 Denoising results for colored noise 204
5.4.3 Demosaicing 207
5.4.4 Image Restoration using Split Bregman techniques . . . 209
5.5 Conclusion 216

6 Noise models for CT images 219
6.1 The Filtered Backprojection Algorithm 220
6.1.1 Parallel-beam CT 220
6.1.2 Discrete implementation 223
6.2 Sources of noise and imperfection in the measurement data . . 225
6.3 Signal-dependency characteristics of the projection data noise . 226
6.4 Noise modeling after FBP reconstruction 230
6.4.1 Existing models 230
6.4.2 Analytical formulation of the local NSD 232
6.5 The discrete NSD model and its relation to directional multiresolution representations 236
6.6 CT noise characteristics 239
6.7 Experimental results 240
6.8 Conclusion 244

7 Models for measuring medical image quality 245
7.1 Existing model observers for medical image quality assessment 247
7.1.1 Signal detection theory and the ideal observer 250
7.1.2 Background models 251
7.1.3 Signal models 252
7.1.4 Channelized Hotelling observers 253
7.2 New model observers for SKS tasks 258
7.2.1 A Variational approximation of the IO for SKS tasks . . 259
7.2.2 Model Observer based on Joint Detection and Estimation 262
7.3 Channelized Hotelling observers for SKS detection tasks 265
7.3.1 Detection of signals with random amplitude 268
7.3.2 Orientation-unaware detection 270
7.3.3 Scale-unaware detection 271
7.3.4 Location-unaware detection using a scanning CHO . . . 272
7.3.5 More complex detection tasks 273
7.4 Results 274
7.4.1 Detection performance experiment 274
7.4.2 Artificial asymmetric lung nodule detection experiment 275
7.5 Multi-slice Observer Models 276
7.6 Conclusion 277

8 Concluding remarks 285

A Appendix A: convergence of the EM algorithm for GSMs 289

B Appendix B: Derivation of the local NSD for parallel beam CT 293

Bibliography 295

List of Acronyms

AUC	:	Area Under the ROC Curve
AWGN	:	Additive White Gaussian Noise
BK	:	Background
BKF	:	Bessel K Form
CAD	:	Computer-assisted diagnosis
CFA	:	Color filter array
CGB	:	Correlated Gaussian background
CHO	:	Channelized Hotelling observer
CLB	:	Clustered lumpy background
CRF	:	Camera response function
CRT	:	Cathode Ray Tube
CST	:	Continuous shearlet transform
CT	:	Computed Tomography
CWT	:	Continuous wavelet transform
DDOG	:	Dense Difference-of-Gaussians
DRF	:	Detector response function
DSC	:	Digital still cameras
DST	:	Discrete shearlet transform
DTFT	:	Discrete Time Fourier transform
DWT	:	Discrete wavelet transform
EM	:	Expectation maximization
EPD	:	Exponential power distribution
FBP	:	Filtered backprojection
FC	:	Frequency coordinates
FPR	:	False positive rate
GRF	:	Gaussian Random Field
GSM	:	Gaussian Scale Mixture
HDR	:	High dynamic range
HMT	:	Hidden Markov Tree
HVS	:	Human visual system
ICA	:	Independent component analysis
ICLS	:	Iterative constrained least squares
ICM	:	Iterated conditional modes
ILO	:	Ideal linear observer

IO	:	Ideal observer
JDE	:	Joint detection and estimation
KLD	:	Kullback-Leibler divergence
KLT	:	Karhunen-Loèvre Transform
LB	:	Lumpy background
LDA	:	Linear Discriminant Analysis
LG	:	Laguerre-Gauss
LGN	:	Lateral geniculate nucleus
LRT	:	Likelihood ratio
LSAI	:	Local Spatial Activity Indicator
MAD	:	Median of Absolute Deviations
MAP	:	Maximum a posteriori
MC	:	Monte Carlo
MGSM	:	Mixture of Gaussian Scale Mixtures
ML	:	Maximum likelihood
MMSE	:	Minimum Mean Square Error
MPGSM	:	Mixtures of projected Gaussian Scale Mixtures
MRA	:	Multiresolution analysis
MRF	:	Markov Random Field
MRI	:	Magnetic Resonance Imaging
MRMC	:	Multiple reader/multiple case
MSE	:	Mean Squared Error
NEQ	:	Noise-equivalent quanta
NLF	:	Noise level function
NSD	:	Noise power spectral density
OAGSM	:	Orientation Adaptive GSM
PAL	:	Phase Alternating Line
PCA	:	Principal component analysis
PDF	:	Probability density function
PPCA	:	Probabilistic principal component analysis
PR	:	Perfect reconstruction
PSD	:	Power spectral density
PSF	:	Point Spread Function
PSNR	:	Peak Signal To Noise Ratio
RL	:	Richardson-Lucy
ROC	:	Receiver operating characteristic
ROI	:	Region of interest
RS	:	Rotationally symmetric
SAR	:	Synthetic aperture radar
SD	:	Semidefinite
SDP	:	Semidefinite programming
SKE	:	Signal known exactly

SKS	:	Signal known statistically
SNR	:	Signal-To-Noise-Ratio
SR	:	Superresolution
STP	:	Steerable pyramid
SVGSM	:	Spatially Variant Gaussian Scale Mixture
SVS	:	Space-varying spectrum
TIHP	:	Translation Invariant Haar Pyramid
TPR	:	True positive rate
TV	:	Total Variation
VS	:	Variance stabilizing
WGB	:	White Gaussian background
WPT	:	Wavelet packet transform

1

Introduction

In this dissertation, we examine and develop statistical multiresolution models for images and statistical models for noise in images. Using statistical estimation, we devise new processing techniques using these models. In particular, we focus on *image restoration* and *medical image quality assessment*. We thereby attempt to fill the gap between theory and practice.

1.1 Problem statement and topical outline

Because of the wide availability of imaging devices such as digital still cameras and medical scanners (in hospitals), there has been a tremendously growing use of digital images in the last decades. While the image quality expectations of the end users have significantly been increased, physical limitations and cost restrictions of the imaging acquisition devices often pose stringent limits on the maximum achievable image quality, e.g., in terms of image resolution, noise, ...

For example, there is an ongoing trend to increase the number of photosensitive elements in digital cameras: nowadays, cameras with 10 megapixel (10 million photosensitive elements) or more are very common. However, adding more photosensitive elements in the camera involves reducing the area of the sensor elements, which also increases the noise in the images. Furthermore, this reduces image resolution because of the non-negligible crosstalk between the sensor elements. Also in medical imaging, decreased acquisition times and the use of lower radiation doses (in x-ray imaging) are desired. This often results in a trade-off between image quality, radiation dosis and acquisition time and various other imaging parameters.

Because the acquired images are imperfect and always contain artifacts or degradations, it becomes increasingly important to improve the quality of these images, or to extract as much information from these images as possible. In this dissertation, we develop digital post-processing techniques to restore the images after acquisition. To build such processing techniques, we employ a generic methodology that can easily be adapted to various practical problems: we devise advanced statistical models for images and image noise based on multiresolution concepts and we explain a general framework in which these models

can be jointly used for a given image restoration task (e.g. noise reduction of digital still images). Hence, the three pilars of this work are *multiresolution representations*, *image models* and *noise models*. Multiresolution representations describe images using a small number of "significant" coefficients, typically much less than the total number of pixels in the image, thereby exploiting their correlation structure and redundancy. Statistical image models describe the properties of ideal, undegraded images and are used as prior knowledge for image restoration. Noise models characterize the statistical properties of noise in images.

In the context of *multiresolution representations*, we improve the directional selectivity of the first-scale basis elements of the dual-tree complex wavelet transform. We also propose a new design for the discrete shearlet transform. This new design offers low redundancy, shift invariance and high directional analysis properties, a combination of properties which is difficult to achieve with existing multiresolution representations.

We present two new *statistical image models* for multiresolution transform coefficients: a joint intra/inter-scale model, which incorporates both spatial correlations and inter-scale dependencies between transform coefficients, and an intra-scale model (MPGSM), which is a generalization and improved version of the well-known and popular GSM model. While these models attempt to fully exploit *local dependencies* between transform coefficients, we also investigate models for *non-local* dependencies. In particular, we introduce an improved image-domain non-local prior model for images. Experimental results show that the denoising techniques based on these two multiresolution models and the non-local model are currently among the state-of-the-art methods in image denoising.

We particularly devote attention to noise models that can take underlying processing techniques in the acquisition devices into account. By the image reconstruction and processing methods used by the devices, the noise characteristics are often seriously altered. Because traditional image restoration methods often make restrictive assumptions with respect to the noise statistics (e.g. they assume additive white Gaussian noise), these techniques generally give poor results or even fail when applied to images originating from real acquisition devices. In this work, we provide several novel approaches to solve this problem.

Next to *image restoration* applications, we also consider the problem of *quality assessment* of medical images. Here the goal is to automatically determine which images are "better" in terms of quality (i.e. clinical value) than other images and to objectively measure the quality. Medical image quality assessment is not only useful for optimizing restoration techniques dealing with medical images, but it is also of great importance for improving future medical display systems.

Traditionally, to assess the impact of new technologies (e.g. new types of backlights for an LCD display) on the quality of the images as seen on the screen, psychovisual experiments involving human specialists are performed.

A panel of experienced physicians is asked to view a set of medical images and to make a binary decision for each image: an abnormality is either present in the diagnostic image, or not. Based on the results of these experiments, the image quality is then measured objectively. Because psychovisual experiments are very time-consuming and costly, "virtual specialist" models (so-called model observers) have been developed in the last decades. The goal is to develop a model that well predicts the performance of a physician performing the same detection task.

In this dissertation, we present a novel framework for model observers for detecting abnormalities with random parameters in images. We will see that model observers that are optimized for this detection task make use of multiresolution concepts, hence quality assessment is very related to image restoration.

Finally, we provide practical results of the proposed image restoration and medical image quality assessment techniques, and when available, we compare to state-of-the-art techniques performing the same task.

1.2 Contributions and list of publications

This research has resulted in several contributions. These contributions are not only improvements to multiresolution representations, but also novel image and noise models and corresponding estimation techniques. The *key novelties*that will be discussed in detail in this dissertation, are:

- A specialized design technique for complex wavelet filters for the first scale of the dual-tree complex wavelet transform (DT-CWT). This method improves the directional properties of the basis functions of the first scale of the transform, which is crucial for many applications in which accurate representation of image details is desired. Published in ICIP conference [Goossens et al., 2009b].

- A novel design of a discrete shearlet transform with a low redundancy ratio (around 2.6) and many interesting properties, such as shift invariance and excellent directional selectivity of the basis functions. This transform is an optimally sparse multiresolution representation for smooth images containing smooth line-like and curve-like discontinuities. Published in LNLA'09 workshop [Goossens et al., 2009a].

- MPGSM, which is an improved intra-scale statistical model for images in a multiresolution transform domain. This model is a generalization of the existing Gaussian Scale Mixtures and Mixtures of Gaussian Scale Mixtures model and captures the spatial variability of the local covariance matrix of subband coefficients. The Bayesian Minimum Squared Error (MMSE) estimator for this model in a denoising task has a performance that is competitive to or better than many state-of-the-art image denoising methods. Published in IEEE Trans. Image Processing [Goossens et al., 2009c].

- Vector-ProbShrink, which is a vector-based extension of ProbShrink [Pižurica and Philips, 2006, Pižurica, 2002], particularly designed to deal with stationary correlated noise in images. With a small modification, the method can also be used for non-stationary noise. Furthermore, the underlying Vector-ProbShrink image model can be coupled to a hidden markov tree (HMT) model in order to take statistical dependencies between subband coefficients in different scales in to account. Published in IEEE Trans. Image Processing [Goossens et al., 2009d].

- An expectation maximization (EM) algorithm for the estimation of the noise covariance of stationary correlated noise in images, which leads - in combination with Vector-ProbShrink - to a "blind" restoration method for images corrupted with correlated noise. We also investigate the estimation of non-stationary noise with spatially variant correlation structure in images. As far as the authors are aware of, this problem has never been tackled before. We present an extension of the EM algorithm to deal with this type of noise. Published in ICIP conference [Goossens et al., 2008c].

- A novel non-stationary model for noise in computed tomography images (CT) reconstructed by the filtered backprojection algorithm. This model is of great importance not only for correct analysis of low-dose CT images (which suffer from noise streaking artifacts) but also for devising CT imaging denoising techniques. Submitted to IEEE Trans. Medical Imaging (under revision).

- An approximative analytical relationship between the camera response function and the noise level function that is important for the modeling and estimation of signal-dependent noise in images.

- An improved non-local means algorithm which exploits self-similarity in images. Here we made several quality and computation time improvements, leading to a denoising technique that again compares favorably to state-of-the-art approaches, both objectively and visually. Published in LNLA'08 workshop [Goossens et al., 2008a].

- A complex-wavelet based demosaicing method (developed in close collaboration with ir. J. Aelterman) that exploits the spatial and frequency localization properties of the wavelets. This technique is computationally simple, solves many discoloration problems and better reconstructs high frequency information in images than a very recent wavelet-based demosaicing method of Hirakawa. The method is particularly interesting because it can be easily combined with denoising.

- Applications and tuning of Bregman optimization for image restoration tasks in combination with novel multiresolution representations. These contributions resulted from a close collaboration with dr. Hiêp Luong and ir. Jan Aelterman. Combined with other image and noise models presented in this dissertation, this led to:

  - An iterative technique which makes use of the discrete shearlet transform in order to jointly perform denoising and deblurring.
  - An iterative method which estimates the Power Spectral Density of the noise in images and simultaneously restores these images.
  - An iterative method which estimates and removes signal-dependent noise from images. This technique is also able to remove the signal-dependent bias caused by the degradation process (e.g. due to clipping of the intensity range).

- A novel theoretical framework for deriving channelized Hotelling observers for assessing medical image quality, in medical signal detection tasks in which the signal is known *statistically*. The application of joint estimation and detection theory to this problem leads to model observers that are optimal in the maximum a posteriori sense and that make use of multiresolution representations.

- In close collaboration with ir. Ljiljana Platiša and the American Food and Drug Administration, we developed novel multi-slice Channelized Hotelling observers, intended for the quality assessment of volumetric medical images viewed in stack-browsing mode (i.e. slices of the medical volume are shown sequentially).

So far, our work resulted in 3 published journal publications in IEEE TRANSACTIONS (as first author), 5 submitted journal publications (of which 1 as first author) and 2 publications in book chapters (as second author). 22 papers are published in the proceedings of international conferences with peer review (of which 8 as first author). 11 papers have been published in conference proceedings without peer review (of which 5 as first author). 1 European patent application has been filed. A small selection of the key publications published during this research is given below:

- B. Goossens, A. Pizurica and W. Philips, "Removal of Correlated Noise by Modeling the Signal of Interest in the Wavelet Domain," *in IEEE Transactions on Image Processing*, vol 18 (6), p.1153–1165, june 2009.

- B. Goossens, A. Pizurica and W. Philips, "Image Denoising Using Mixtures of Projected Gaussian Scale Mixtures," *in IEEE Transactions on Image Processing*, vol 18 (8), p. 1689–1702, august 2009.

- B. Goossens, J. Aelterman, A. Pizurica and W. Philips, "A Recursive Scheme for Computing Autocovariance Functions of Decimated Complex Wavelet Subbands, " *in IEEE Trans. Signal Processing*, in press.

- J. Aelterman, B. Goossens, A. Pizurica and W. Philips, "Suppression of Correlated Noise", *in Recent Advances in Signal Processing*, 2010, IN-TECH, ISBN 978-953-307-002-5

- F. Rooms, B. Goossens, A. Pizurica and W. Philips, "Image restoration and application in biomedical processing, " *in Optical and Digital Image Processing*,

2010, Eds. G. Cristobal, P. Schelkens and H. Thienport, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, ISBN 13-978-3-527-31291-7

- B. Goossens, Lj. Platisa, E. Vansteenkiste and W. Philips, "The Use of Steerable Channels for Detecting Asymmetrical Signals with Random Orientations", *in Proc. of the SPIE: Medical Imaging 2010*, Feb. 2010, San Diego, CA, USA

- B. Goossens, J. Aelterman, H. Q. Luong, A. Pizurica and W. Philips, "Efficient Design of a Low Redundant Discrete Shearlet Transform, " *in Proc. 2009 International Workshop on Local and Non-Local Approximation in Image Processing (LNLA 2009)*, August 19-21, 2009, Tuusula, Finland (invited paper)

- B. Goossens, A. Pizurica and W. Philips, "A Filter Design Technique for Improving the Directional Selectivity of the First Scale of the dual-tree complex wavelet transform," *in Proc. of the IEEE Int. Conf. Image Processing (ICIP 2009)*, Nov. 7-11, Cairo, Egypt

- B. Goossens, A. Pizurica, and W. Philips, "EM-Based Estimation of Spatially Variant Correlated Image Noise," *in Proc. of the IEEE IEEE Int. Conf. Image Processing (ICIP2008)*, San Diego, California, USA, Oct. 2008, pp. 1744-1747

- B. Goossens, H.Q. Luong, A. Pizurica, W. Philips, "An improved non-local means algorithm for image denoising," *in Proc. of the 2008 International Workshop on Local and Non-Local Approximation in Image Processing (LNLA 2008)*, Lausanne, Switzerland, Aug. 25-29 (invited paper)

- B. Goossens, A. Pizurica, and W. Philips, "Removal of Correlated Noise By Modeling Spatial Correlations and Interscale Dependencies in the Complex Wavelet Domain," *in Proc. of the IEEE Int. Conf. Image Processing (ICIP 2007)*, San Antonio, Texas, USA, 16-19 Sept. 2007.

1.3 Organization of this dissertation

This thesis is organized as follows: in Chapter 2 we review a number of multiresolution representations for images. We describe the discrete wavelet transform, the complex wavelet transform, the steerable pyramid transform and the shearlet transform, as the concepts related to these transforms form the basis of our further developments. We also present our improvement to the dual-tree complex wavelet transform and our novel design for the discrete shearlet transform.

In Chapter 3 we discuss a number of statistical modeling strategies for ideal (i.e. degradation-free) images. Because multiresolution transforms do not fully "decorrelate" images, statistical dependencies between the multiresolution transform coefficients still exist. In particular, we devote attention to the intra-scale and inter-scale modeling of the dependencies between the transform coefficients. This leads to two novel statistical image models.

Next, we present novel statistical models and estimation techniques for noise in images in Chapter 4. In particular, we investigate spatially stationary colored Gaussian noise, spatially non-stationary colored Gaussian noise and

signal-dependent noise. We explain the Gaussian modeling of signal-dependent noise and discuss how the signal-dependency characteristics of the noise are changed by various post-processing operations in a digital still camera.

In Chapter 5, we use the models presented in the previous chapters for the purpose of image restoration. We consider three designs of image restoration algorithms: spatial domain techniques, multiresolution transform domain techniques and combined domain techniques. We solve various restoration problems by using Bayesian estimation techniques, resulting in algorithms for denoising, deblurring, demosaicing or combinations such as joint denoising and deblurring.

In Chapter 6 we investigate the statistical properties of noise in CT images reconstructed using the traditional filtered backprojection method (which is used in almost all commercially available medical CT scanners [Hsieh, 2003]). Because of the signal-dependency of the noise captured by the detector elements and because of the non-local characteristics of the reconstruction algorithm, the statistics of the noise are much more complicated than in our chapter on general noise models. Nevertheless, by relying on the backprojection reconstruction formulas and using the theory presented in Chapter 4 we derive a novel statistical model for noise in CT images in which multiresolution concepts (in particular, steerability) plays an important role.

In Chapter 7 we present a new theoretical framework for mathematical model observers (*"virtual observers"*) for assessing medical image quality. We consider signal-known-statistically (SKS) detection tasks in which the signals have unknown parameters. We show that the optimal *linear* (or Channelized Hotelling) observer is not always adequate for these tasks. We develop a model observer that is optimal in the maximum a posteriori (MAP) sense based on joint estimation and detection theory. We show that this leads to Channelized Hotelling Observers that make use of steerable channels as used in the steerable pyramid multiresolution transform.

Finally, in Chapter 8 a general conclusion is given for our work and some directions for future research are being discussed.

2

Multiresolution representations for images

Taking advantage of the redundancy present in images is crucial for designing good image processing algorithms. The redundancy in images is caused by 1) *geometrical structures* in images (such as edges, contours, textures, ...), 2) *self-similarity* (or non-local redundancy) and 3) *color information*. For many applications, it is desirable to find an adapted representation of images that is suitable for the given problem, thereby taking advantage of the redundancy. For example, in image denoising applications it is benificial if this representation separates structural information from the noise in the images well. This is possible if the clean, noise-free image can be approximated with a small number of significant coefficients. The representation is then called a *sparse representation*.

Multiresolution transforms decompose an image in a natural way: the image is approximated by successively adding detail information to it in subsequent refinement steps. This approach is effective as natural images are often low-pass in nature (see further in Chapter 3). Classical tools in this respect are the Fourier transform and the Short-Time Fourier transform, however, these transforms do not allow fine localization of image features in space: it is for example not possible to determine the exact position of edges. The *wavelet transform* offers a compromise between spatial and frequency localization of image features as we will see further. The classical wavelet transform, while ideally suited for one-dimensional signals, turns out to be sub-optimal for representing images, because it can not entirely adapt to the image geometry. Moreover, the transform can not exploit the self-similarity in images.

Therefore, during the last decades, there has been a quest for finding multiresolution transforms that yield "sparser" representations for images than the wavelet transform. Most of these transforms attempt to adapt to the geometrical patterns in images, for example, by decomposing the images according to multiple analysis directions. In this chapter, we will review some of these transforms, where occasionally, we will improve some of their properties: in

Section 2.1 we briefly introduce the discrete wavelet transform and we discuss a number of problems with this transform for representing images. In Section 2.2, we explain the dual-tree complex wavelet transform (DT-CWT), which offers a solution to some of these problems. To further improve the directional selectivity of the transform, we will propose a new filter design for the first scale of the DT-CWT (Section 2.2.3). The steerable pyramid transform is discussed in Section 2.3, together with steerability concepts that will prove useful for the remainder of this dissertation. Section 2.4 gives a literature overview of related multiresolution representations for images. Finally, in Section 2.5 we present a novel design of the discrete shearlet transform. This transform design offers multidirectional analysis, low redundancy and shift-invariance properties, a combination that is difficult to achieve with existing multiresolution transforms.

2.1 The discrete wavelet transform

The first transform that we will discuss, is the discrete wavelet transform. Before going into the details of this discrete transform, we will give a brief introduction to wavelet theory and multiresolution analysis. These concepts will also be useful in the remainder of this chapter.

2.1.1 A short introduction to wavelet theory

The basic idea of the wavelet transform [Daubechies, 1992, Mallat, 1999] is to analyze signals according to different scales and at different points in time. Wavelets are functions that oscillate in a small portion of time (or space) and that decrease back to zero outside this portion. When applying a (continuous) wavelet transform to a signal, we start from a fixed wavelet $\psi(t)$, called *mother wavelet*, and we analyze correlations of the input signal with time-shifted and time-stretched (dilated) versions of this wavelet. Correlations with wavelets with a large dilation factor then give the coarse features of the image, while correlations with wavelets with small dilation factors indicate fine details in the image. In general, the dilation factor can be tuned continuously, e.g. to zoom in at specific features of the image. In signal processing terms, a wavelet is a band-pass filter that has either a finite impulse response or an impulse response with a fast decay. By changing the dilation factor, the band-pass center frequency can be controlled. The time-shift parameter allows to select the time portion of the signal we want to analyze.

By considering *all* possible values of the dilation factor and time-shift parameter, the wavelet transform provides a representation of the signal in terms of these time-shifted and time-stretched wavelet functions. More specifically, let us denote the time-shifted and dilated basis functions by:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$$

where a is the dilation factor, b is the time shift, and the normalization constant $1/\sqrt{a}$ is introduced to keep the energy of the wavelet constant. The continuous wavelet transform (CWT) of a signal of finite energy $f(t) \in L^2(\mathbb{R})$ is defined by:

$$\mathcal{W}f(a,b) = \int_{-\infty}^{+\infty} f(t)\overline{\psi_{a,b}(t)}\mathrm{d}t = \langle f, \psi_{a,b}\rangle . \tag{2.1}$$

with $\bar{\cdot}$ the complex conjugate. If the wavelet $\psi(t)$ satisfies the admissibility condition ($\int_{-\infty}^{+\infty} \left|\widehat{\psi}(\omega)\right|^2 / \omega \mathrm{d}\omega < +\infty$), with $\mathcal{F}\psi$ the Fourier transform of ψ, the original signal can be exactly reconstructed by [Daubechies, 1992]:

$$f(t) = \int_{-\infty}^{+\infty} \mathrm{d}a \int_{-\infty}^{+\infty} \mathrm{d}b \, \langle f, \psi_{a,b}\rangle \, \psi_{a,b}(t), \tag{2.2}$$

which constitutes the inverse continuous wavelet transform. To satisfy the admissibility condition, it is required that the mother wavelet averages to zero $\int_{-\infty}^{+\infty} \psi(t)\mathrm{d}t = 0$ and that the mother wavelet is well localized (i.e., $|\psi(t)| \to 0$ when $|t| \to +\infty$). By convention, the wavelet is centered around $t = 0$ and has unit norm ($\|\psi\|^2 = \langle \psi, \psi\rangle = 1$). In practice, one can choose the mother wavelet according to the application one has in mind: in general one has to trade off time (spatial) localization properties versus the frequency localization properties. For example, a wavelet that is bandlimited (e.g. the Meyer wavelet) will necessarily have an infinite support, hence a poor localization in time. On the other hand, wavelets that have a compact support in time domain are not bandlimited. For image restoration, wavelets with a good localization in time are preferable over bandlimited wavelets, because of spurious ringing (Gibbs) artifacts that might appear after applying the inverse transform due to estimation errors.

So far, we considered the dilation and time-shift parameters a and b to be continuous, however, in practice it is impossible to analyze the signal according to all corresponding wavelet coefficients. Therefore, a and b are usually restricted to discrete values: integer shifts $b \in \mathbb{Z}$ and dyadic scales $a = 2^i$ with $i \in \mathbb{Z}$. Mathematically speaking, if the set of wavelet functions $\left\{\psi_{2^i,b}(t) | i \in \mathbb{Z}, b \in \mathbb{Z}\right\}$ forms a basis for $L^2(\mathbb{R})$, any signal $f(t)$ of finite energy ($f \in L^2(\mathbb{R})$) can be reconstructed by:

$$f(t) = \sum_{i=-\infty}^{+\infty} \sum_{b=\infty}^{+\infty} \left\langle f, \psi_{2^i,b}\right\rangle \psi_{2^i,b}(t). \tag{2.3}$$

The corresponding transform $\mathcal{W}f\left(2^i, b\right)$ is then called the dyadic *discrete wavelet transform* (DWT). When we compare (2.2) to (2.3), we see that the integral has been replaced by a sum. Now, suppose that we want to approximate the signal $f(t)$ up to scale I. This can be accomplished by reducing the number of terms in the sum (2.3), as follows:

$$f(t) \approx \sum_{i=-\infty}^{I} \sum_{b=\infty}^{+\infty} \left\langle f, \psi_{2^i,b}\right\rangle \psi_{2^i,b}(t),$$

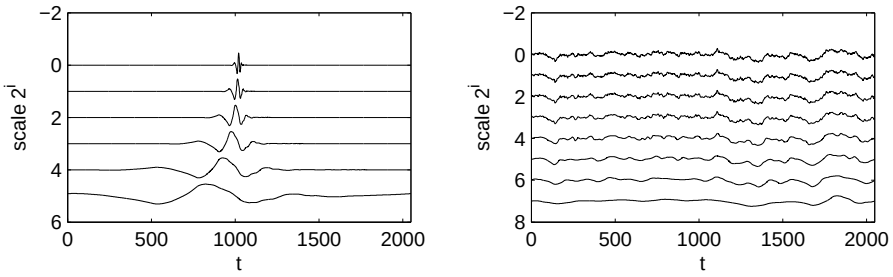


Figure 2.1: Illustration of wavelet analysis (a) wavelets at different scales, (b) wavelet approximation of a signal using the wavelets from (a): the smaller the scale index i, the more detail information that is added to the signal.

which only requires the computation of correlations $\langle f, \psi_{2^i,b} \rangle$ up to scale I. An illustration is given in Figure 2.1, where the maximum wavelet scale is subsequently increased.

Because the wavelet functions $\{\psi_{2^i,b}(t) | i \in \mathbb{Z}, b \in \mathbb{Z}\}$ need to form a basis, the choice of the wavelet is far more restrictive than for the CWT. For this reason, the mathematical concept of multiresolution analysis has been introduced.

2.1.2 Multiresolution analysis and wavelet filters

Multiresolution analysis (MRA) is the analysis of a function in successive approximation function spaces $\cdots V_2 \subset V_1 \subset V_0 \subset \widetilde{V_{-1}} \subset \cdots$ where the subspaces V_i form a partition of the $L^2(\mathbb{R})$-space: $\widetilde{\bigcup_{i\in\mathbb{Z}} V_i} = L^2(\mathbb{R})$ and $\bigcap_{i\in\mathbb{Z}} V_i = \{0\}$ [Mallat, 1989b, Daubechies, 1992]. Here, $\widetilde{V}$ denotes the closure of a set V. By definition, a number of additional properties are required from MRA:

1. When a function f belongs to a given subspace, it is demanded that a dilated version of this function is contained in the subsequent subspace:

   $$f(x) \in V_{i+1} \Leftrightarrow f(2x) \in V_i$$

   or in other words: all approximation spaces V_i are scaled versions of the central space V_0.

2. Furher, invariance under integer shifts is required ($n \in \mathbb{Z}$):

   $$f(x) \in V_i \Leftrightarrow f(x-n) \in V_i.$$

3. Finally, there must exist a *scaling* function $\phi \in V_0$ such that

   $$\{\phi_{0,n} | n \in \mathbb{Z}\} \text{ is an orthonormal basis in } V_0$$

   where

   $$\phi_{i,n}(t) = 2^{-i/2}\phi\left(2^{-i}t - n\right).$$

The merit of MRA is that when the subspaces V_i satisfy all of the above conditions, there exists an orthonormal wavelet basis $\{\psi_{i,n}|i,n\in\mathbb{Z}\}$ of $L^2(\mathbb{R})$ with $\psi_{i,n}(t)=2^{-i/2}\psi\left(2^{-i}t-n\right)$, such that for all $f\in L^2(\mathbb{R})$,

$$P_{i-1}f=P_if+\sum_{n\in\mathbb{Z}}\langle f,\psi_{i,n}\rangle\,\psi_{i,n}.$$

Consequently, instead of using infinitely many scales $i\in\mathbb{Z}$ as in (2.3), the decomposition can be stopped at a given scale I:

$$f(t)=\sum_{i=-\infty}^{I}\sum_{n=\infty}^{+\infty}\langle f,\psi_{i,n}\rangle\,\psi_{i,n}(t)+\sum_{n=\infty}^{+\infty}\langle f,\phi_{I,n}\rangle\,\phi_{I,n}(t),$$

which gives the representation of the signal $f(t)$ in terms of the scaling functions $\phi_{I,n}(t)$ and wavelet functions $\psi_{i,n}(t)$. Because the functions $\phi_{i,n}(t),n\in\mathbb{Z}$ form a basis for V_i, both the wavelet and scaling functions can be expanded in this basis, which gives the so-called wavelet and scaling equations:

$$\frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right) = \sum_{n\in\mathbb{Z}}h_n\phi\left(t-n\right), \tag{2.4}$$

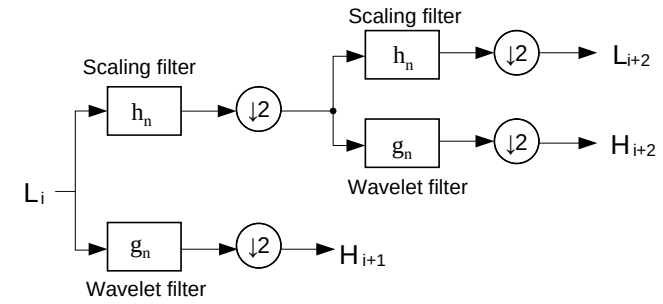
$$\frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right) = \sum_{n\in\mathbb{Z}}g_n\phi\left(t-n\right), \tag{2.5}$$

where $h_n=\langle\phi,\phi_{-1,n}\rangle$ and $g_n=\langle\psi,\phi_{-1,n}\rangle$. The sequences h_n and g_n are respectively called scaling and wavelet filters. The practical importance of these equations is that when we want to compute wavelet coefficients (i.e. the inner products of the signal $f(t)$ with the wavelet basis functions), this can be done in terms of scaling coefficients from a finer scale:

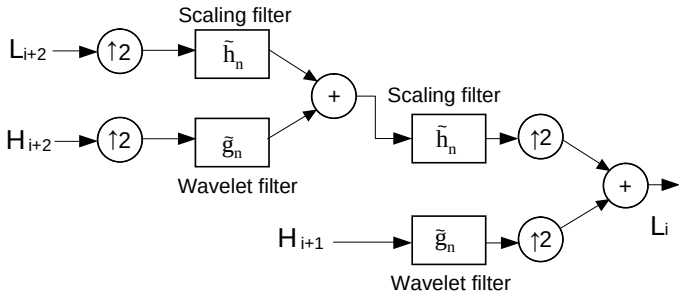
$$\langle f,\phi_{i,n}\rangle = \sum_k\overline{h}_{k-2n}\langle f,\phi_{i-1,n}\rangle, \tag{2.6}$$

$$\langle f,\psi_{i,n}\rangle = \sum_k\overline{g}_{k-2n}\langle f,\phi_{i-1,n}\rangle. \tag{2.7}$$

This immediately leads to a practical implementation for the DWT: starting from the finest scale coefficients $\langle f,\phi_{0,n}\rangle$, coefficients from coarser scales can be readily computed by recursively applying the filtering equations (2.6) and (2.7). The wavelet analysis and synthesis can be seen as a two-channel digital filter bank (see Figure 2.2), consisting of a low-pass (scaling) filter $\overline{h}$ and a high-pass (wavelet) filter $\overline{g}$, followed by decimation by factor 2. The filter bank is applied recursively on the low-pass output (i.e. scaling coefficients). Very useful for practical computation is that the filters $\overline{h}$ and $\overline{g}$ are FIR, such that the corresponding wavelet and scaling functions are compactly supported. The construction of wavelet bases relies on equations (2.4)-(2.5) and the reader can refer to [Daubechies, 1992, Mallat, 1989a] for more details on this topic.



(a)



(b)

Figure 2.2: Two analysis/synthesis steps of the fast (orthogonal) DWT in 1D. H_{i+1} and H_{i+2} are sequences of wavelet coefficients for respectively scale 1 and 2. L_{i+2} is a sequence of scaling coefficients. (a) Wavelet analysis (forward DWT), (b) wavelet synthesis (inverse DWT).

2.1.3 The Fast DWT in higher dimensions

The DWT can be generalized to any dimension. In the following we will briefly look at the 2D case. Most often, separable decompositions are used [Mallat, 1989b, Daubechies, 1992]. Here, the approximation spaces of the MRA are spanned by the shifts and dilations of three wavelets:

$$\begin{aligned} \psi^{\mathrm{LH}}(x,y) &= \phi(x)\psi(y), \\ \psi^{\mathrm{HL}}(x,y) &= \psi(x)\phi(y), \\ \psi^{\mathrm{HH}}(x,y) &= \psi(x)\psi(y) \end{aligned}$$

and a 2D scaling function is defined as:

$$\phi^{\mathrm{LL}}(x,y) = \phi(x)\phi(y).$$

Because of the separability of the wavelets, the 2D transform can be straightforwardly implemented, by applying the Fast DWT algorithm successively to the rows and to the columns of the image. The corresponding filter bank scheme for one decomposition step is illustrated in Figure 2.3. The decomposition is iterated on the scaling coefficients (i.e. LL_{i+1}).

In Figure 2.4, the 2D DWT of the picture of a mandrill is shown. It can be noted that the LL_2 subband contains most information about the input image, while other subbands only have a limited number of (significantly) non-zero coefficients. These significant wavelet coefficients correspond to detail information, mostly textures and edges. When a coefficient is non-zero in a coarse scale $i+1$, it is very likely to be non-zero in a finer scale i as well (*inter*scale dependency). Around vertical edges, HL_i subbands contain significant non-zero wavelet coefficients, while near horizontal edges, LH_i subband coeffients are typically non-zero. As we will explain later, the HH_i is not related to a specific orientation, but this subband contains information about features and edges with dominant orientation $+45°$ and $-45°$. To summarize, we observe that:

- The low-pass subband (that consists of the scaling coefficients) is a coarse approximation of the original input image.
- All other wavelet subbands are relatively *sparse*: many wavelet coefficients are (close to) zero, only few of them are significantly non-zero.
- The wavelet coefficients are not completely decorrelated: e.g. the occurrence of an edge results in many neighboring wavelet coefficients that are non-zero.
- There are *inter*scale correlations between wavelet coefficients in different scales.
- In presence of edges and textures there exist *inter*orientation dependencies between the wavelet coefficients in different orientation subbands (e.g. HL_i vs. LH_i): e.g. a vertical edge causes a non-zero response while in the LH-subbands the response is approximately zero in the HL-subbands.

These observations will form the basis of our statistical models for images. For more information about the dependencies that exist between wavelet coefficients and how these dependencies can be modeled, we refer to Chapter 3.

2.1.4 Problems with the DWT

Despite the efficiency and the sparsity of the DWT, there are a number of fundamental problems [Selesnick et al., 2005a]:

1. *Positive and negative oscillations* of the wavelet coefficients around singularities: practical applications that make use of the DWT need to take

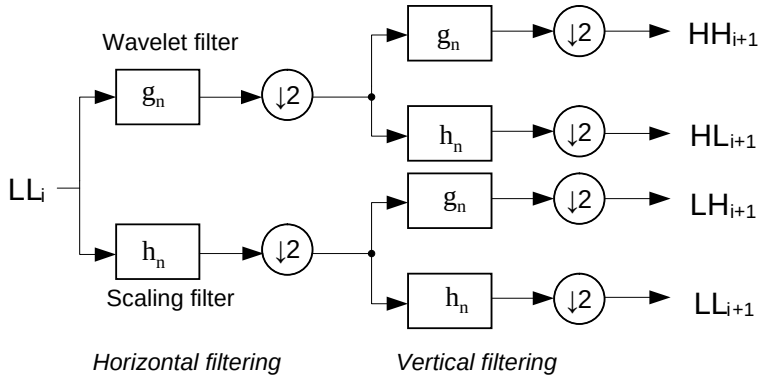
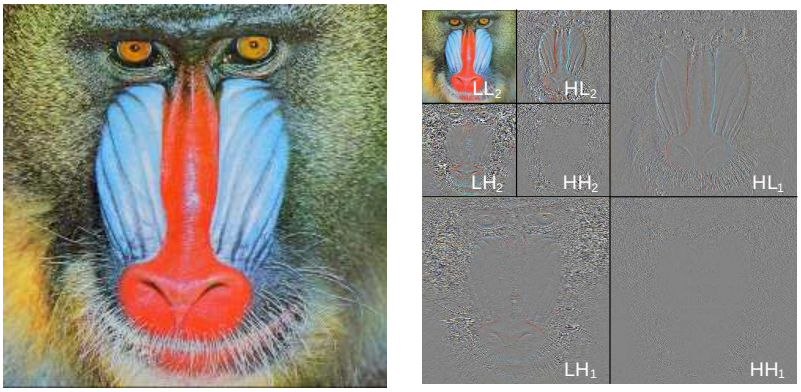


Figure 2.3: A decomposition step of the fast (orthogonal) DWT in 2D.



(a) (b)

Figure 2.4: (a) Image of a mandrill (b) 2D DWT subbands of the mandrill image (gray corresponds to 0).

into account that in presence of edges, the wavelet coefficients can be either significantly positive, negative or both. Often, it is desired to obtain an estimate of the local spectrum at a specific scale and position, which one can obtain by computing the wavelet coefficient magnitude. However, after reconstructing a wavelet coefficient (e.g. in a denoising task) it is not always clear whether to assign a positive or negative value to the processed wavelet coefficient.

2. *Shift variance*: the local energy signature of edges in the transform domain can be significantly disturbed by shifts of the input signal (or image). For statistical modeling tasks, it is convenient to have a representation that is invariant under translations of the input image. If this is not the

case, the (unknown) input shift would need to be treated as an extra hidden variable in the model, which makes the resulting statistical model more complicated. For example, to match keypoints in the wavelet domain in an image registration application, one can look for the keypoint that has a maximal similarity to the reference keypoint. However, if the similarity measure does not incorporate the shift dependence, either wrong matches will be found, or keypoints will be missed. In general, the DWT has been found to be unsuitable for certain applications, such as pattern recognition [Mallat, 1996].

3. *Aliasing* caused by decimation operations: the aliasing often causes disturbing artifacts in reconstructed images in which the wavelet coefficients have been processed. Obviously, the inverse DWT cancels this aliasing, however, this is only the case if the wavelet coefficients are unmodified.

4. *Poor directional selectivity*: the DWT allows to distinguish horizontal and vertical edges, however, the transform does not allow to make a distinction between features at $+45°$ and $-45°$ (also known as the *checkerboard* problem). This is due to the separability of the higher dimensional wavelets, which produces wavelets with a checkerboard pattern that does not have a dominant direction. The poor directional selectivity makes the modeling of edges more complicated.

Some of these issues can easily be alleviated, e.g. one can easily get rid of the shift variance and aliasing by *skipping the decimation steps* of the DWT (called undecimated DWT) [Mallat, 1999], or by *cyclically shifting* the input image (called cycle spinning) [Coifman and Donoho, 1995]. However, this comes at the cost of extra redundancy for the transform coefficients and some of the other problems still remain.

2.2 The dual-tree complex wavelet transform

In the last decades, many alternative multiresolution representations have been developed in order to give a solution to the previously mentioned problems with the DWT. One such transform is the dual-tree complex wavelet transform (DT-CWT) [Kingsbury, 2001], which is very related to the DWT and that also provides MRA (Section 2.1.2). As the name already says, the DT-CWT uses complex-valued wavelets instead of real-valued wavelets. When additionally these complex-valued wavelets fulfill a so-called Hilbert-transform pair property (see further), the transform performs a multi-directional analysis (which practically means that features that have a dominant direction, such as edges in images, can be more compactly represented). This is beneficial for analyzing higher-dimensional data in general, including images and image volumes. We will now explain the DT-CWT in more detail.

2.2.1 One-dimensional complex wavelets

In [Selesnick et al., 2005a], it has been noted that the Fourier transform does not suffer from the problems mentioned in Section 2.1.4: the magnitude of the Fourier coefficients does not oscillate but provides a smooth envelope in frequency domain. Next, the magnitude of the Fourier coefficients is shift-invariant, while the shift is encoded in the phases of the Fourier coefficients. Hence, one can see a difference between the DWT and the Fourier transform: while the DWT is based on real-valued wavelets, the Fourier transform is based on complex-valued oscillating functions:

$$\exp(j\omega t) = \cos(\omega t) + j\sin(\omega t)$$

where the imaginary part is 90° out of phase of the real part. Moreover, $\exp(j\omega t)$ is an analytic function, i.e. it is supported on one half of the frequency axis ($\omega > 0$). In analogy to the Fourier transform, we can devise a wavelet transform in which the wavelets are analytic, such that the magnitude of the wavelet coefficients is shift-invariant. It can be shown that this wavelet must be complex-valued and that the imaginary part of the wavelet should be the Hilbert transform of the real part [Selesnick et al., 2005a]:

$$\psi_c(t) = \psi(t) + j\mathcal{H}\{\psi\}(t). \tag{2.8}$$

This expression is also called the analytic representation of the wavelet.[1] Here, the Hilbert transform of $\psi(t)$ is given by:

$$\mathcal{H}\{\psi\}(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{\psi(t)}{t-\tau}\mathrm{d}\tau \tag{2.9}$$

or, in the Fourier domain:

$$\widehat{\mathcal{H}\{\psi\}}(\omega) = H_{\text{hil}}(\omega)\widehat{\psi}(\omega), \quad \text{with } H_{\text{hil}}(\omega) = \begin{cases} -j, & \omega > 0 \\ 0, & \omega = 0 \\ j, & \omega < 0 \end{cases}$$

To see the analogy with the Fourier transform, remark that $\mathcal{H}\{\cos\}(t) = \sin(t)$, such that substituting $\psi(t) = \cos(\omega t)$ yields $\psi_c(t) = \exp(j\omega t)$.[2] In Figure 2.5, the complex exponential function and an example of a complex wavelet are depicted. It can be seen that the complex wavelet has a smooth envelope that is well localized in time, while the complex exponential function has a constant magnitude and is not localized in time at all.

[1] We remark that (2.8) is not the only way to construct complex wavelets. More general complex wavelets $\psi_c(t)$ are investigated in [Belzer et al., 1995, Lina and Mayrand, 1995], however, these wavelets are not analytic.

[2] Note in this respect that $\exp(j\omega t)$ is not a wavelet - the complex exponential does not satisfy the admissibility condition $\int_{-\infty}^{+\infty}\left|\widehat{\psi}(\omega)\right|^2/\omega\mathrm{d}\omega < +\infty$.

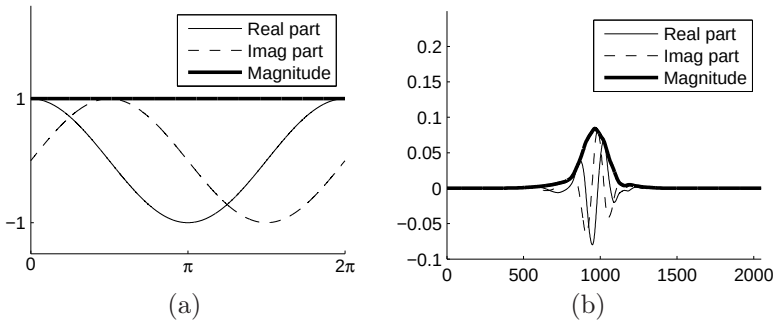


Figure 2.5: (a) The complex exponential function $\exp(j\omega t)$, (b) A complex wavelet $\psi_c(t)$.

Based on the definition of a complex analytic wavelet (2.8), it is easy to show that $\psi_c(t)$ is supported on positive frequencies ($\omega > 0$):

$$\widehat{\psi}_c(\omega) = \begin{cases} 2\widehat{\psi}(\omega) & \omega > 0 \\ 0 & \omega \leq 0 \end{cases}$$

and conversely, $\overline{\psi_c(t)}$ is supported only on negative frequencies ($\omega < 0$). By projecting a real-valued signal on the complex-valued wavelet functions:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi_c\left(\frac{t-b}{a}\right),$$

complex wavelet coefficients $\langle f, \psi_{a,b}\rangle$ are obtained, from which the magnitude and phase can be easily computed: a large magnitude then reveals the presence of a singularity (e.g. edge, texture, point, ...), while the phase indicates the position of the singularity within the support of the complex wavelet [Selesnick et al., 2005a].

Discrete implementations of the DT-CWT transform again make use of the fast DWT decomposition scheme: usually two fast DWTs are applied in parallel, one DWT for the real part of the complex wavelet, a second DWT for the imaginary part. Because there are two wavelet decomposition trees, the transform is called the *dual-tree* complex wavelet transform.

Unfortunately, a complication arises with the design of the complex wavelets: when the real part $\psi(t)$ is compactly supported, its Hilbert transform is infinitely supported, hence the wavelet filter coefficients cannot be compactly supported! For practical (time-domain) implementation, one often prefers compactly supported filters because of the computational efficiency. In the literature, two solutions are proposed to design complex wavelets:

1. One solution is to use dedicated complex wavelet design techniques to offer perfect reconstruction with compactly supported wavelets, compromising the analyticity of the wavelets. To name a few of these techniques:

the q-shift solution [Kingsbury, 2003], the Common Factor solution [Selesnick, 2002], Bernstein polynomial-based [Tay et al., 2006] and the semi-definite programming (SDP) approximation [Dumitrescu, 2009]. In all of these techniques, a trade-off between the degree of analyticity and the wavelet filter support size needs to be made: the longer the filter supports, the better the analyticity (but the higher the computational cost!). The main drawback of these techniques is that the analyticity only holds approximately, hence the shift-invariance property of the magnitudes of the complex wavelet coefficients is not exact. Fortunately, approximate analyticity is sufficient in practice as this does not pose many problems for most applications.

2. The second solution is to perform the wavelet filtering in the DFT domain [Chaux et al., 2006]. This has the advantage that the Hilbert transform relationship holds exactly, however, the wavelet filters are not compactly supported, which may cause more ringing artifacts to appear in the reconstructed images after processing of the DT-CWT coefficients.

A second complication that - as far as the author is aware of - has not been addressed in the literature, is that analyticity is difficult to fulfill for the first (finest) scale of the transform as the same input signal is used for every wavelet decomposition tree. We give a more detailed description of this problem and we propose a solution to it in Section 2.2.3.

Nevertheless, to summarize, the DT-CWT solves the problems from Section 2.1.4 in the following way:

- The positive and negative oscillation problem disappears when processing the complex wavelet coefficient *magnitudes,* which have been shown to be good estimates of the local spectrum at a given position and scale [Selesnick et al., 2005b].
- The magnitudes of the complex coefficients are *shift-invariant*, hence techniques that only process magnitude information in a coefficient-wise manner and that leave the phase information untouched, will automatically be shift-invariant and free of aliasing.
- The DT-CWT performs a directional analysis in 6 orientations in 2D and 28 orientations in 3D and does not suffer from the checkerboard problem of the DWT.

In the next section we will explain in more detail how the improved directional selectivity compared to the DWT is accomplished, as this is crucial for our later developments.

2.2.2 Higher dimensional complex wavelets: how directional selectivity is obtained.

As in the real-valued DWT, higher-dimensional wavelet can be formed as tensor-product of one-dimensional wavelets. For illustrational purposes, we will

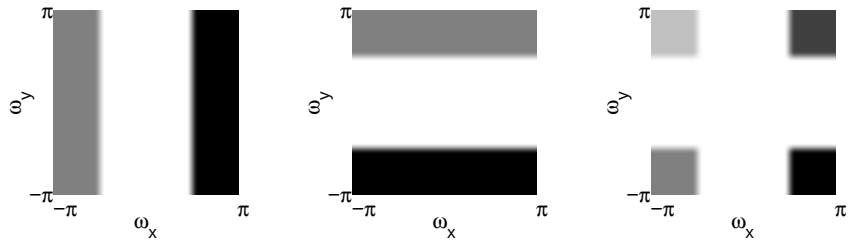


Figure 2.6: Illustration of how directional selectivity of the complex wavelets is achieved. Different gray-shades indicate the frequency spectra of different complex wavelets: (a) $\widehat{\psi}_{c,x}(\omega_x)$ and $\widehat{\psi}_{c,x}(-\omega_x)$, (b) $\widehat{\psi}_{c,y}(\omega_y)$ and $\widehat{\psi}_{c,y}(-\omega_y)$, (c) The resulting 2D complex wavelets $\widehat{\psi}_{c,k}(\omega_x, \omega_y)$ obtained as tensor products of 1D complex wavelets.

explain only the 2D case. Extensions to higher dimensions are straightforward. Let us start from a pair of complex wavelets $\psi_{c,x}(t) = \psi_x(t) + j\mathcal{H}\{\psi_x\}(t)$ and $\psi_{c,y}(t) = \psi_y(t) + j\mathcal{H}\{\psi_y\}(t)$ with Fourier transforms $\widehat{\psi}_{c,x}(\omega) = (1 + jH_{\text{hil}}(\omega))\,\widehat{\psi}_x(\omega)$ and $\widehat{\psi}_{c,y}(\omega) = (1 + jH_{\text{hil}}(\omega))\,\widehat{\psi}_y(\omega)$, respectively. The frequency response of a corresponding 2D complex wavelet is given by:

$$\begin{aligned}\widehat{\psi}_{c,1}(\omega_x, \omega_y) &= \widehat{\psi}_{c,x}(\omega_x)\,\widehat{\psi}_{c,y}(\omega_y) \\ &= \widehat{\psi}_x(\omega_x)\,\widehat{\psi}_y(\omega_y)\,(1 - H_{\text{hil}}(\omega_x)H_{\text{hil}}(\omega_y) + j\,(H_{\text{hil}}(\omega_x) + H_{\text{hil}}(\omega_y))) \\ &= 4\widehat{\psi}_x(\omega_x)\widehat{\psi}_y(\omega_y)S(\omega_x)\,S(\omega_y)\,, \end{aligned} \tag{2.10}$$

$$\text{with } S(t) = \begin{cases} 0 & t < 0 \\ 1/2 & t = 0 \\ 1 & t > 0 \end{cases}$$

a step function. From (2.10) it can be seen that $\widehat{\psi}_{c,1}(\omega_x, \omega_y)$ only passes positive horizontal and vertical frequencies, hence $\widehat{\psi}_{c,1}(\omega_x, \omega_y)$ is supported on the first quadrant of the 2D frequency plane. The same way, it is possible to design complex wavelets that pass frequencies in other quadrants of the frequency plane, by using conjugates of the one-dimensional complex wavelets, e.g.:

$$\psi_{c,2}(x, y) = \overline{\psi_{c,x}(x)}\psi_{c,y}(y)\,. \tag{2.11}$$

Using the same reasoning, it can be shown that the frequency response of $\psi_{c,2}(x, y)$ is given by:

$$\widehat{\psi}_{c,2}(\omega_x, \omega_y) = 4\widehat{\psi}_x(\omega_x)\widehat{\psi}_y(\omega_y)S(-\omega_x)\,S(\omega_y)\,, \tag{2.12}$$

which is supported in the second quadrant of the frequency plane (i.e. $\omega_x < 0$ and $\omega_y > 0$). A frequency domain illustration of this process is given in Figure 2.6.

Now, if we would look at the frequency response of the real part and imaginary part of the 2D complex wavelets, i.e.:

$$\widehat{\psi}_{c,k}(\omega_x,\omega_y) = \widehat{\psi}_{re,k}(\omega_x,\omega_y) + j\widehat{\psi}_{im,k}(\omega_x,\omega_y), \quad k=1,2 \tag{2.13}$$

then we find that, by the conjugate symmetry property of the Fourier transform, both parts have the same magnitude response:

$$\begin{aligned}
\left|\widehat{\psi}_{re,1}(\omega_x,\omega_y)\right| &= \left|\widehat{\psi}_{im,1}(\omega_x,\omega_y)\right| = \\
&2\left|\widehat{\psi}_x(\omega_x)\right|\left|\widehat{\psi}_y(\omega_y)\right|\left(S(\omega_x)S(\omega_y)+S(-\omega_x)S(-\omega_y)\right) \\
\left|\widehat{\psi}_{re,2}(\omega_x,\omega_y)\right| &= \left|\widehat{\psi}_{im,2}(\omega_x,\omega_y)\right| = \\
&2\left|\widehat{\psi}_x(\omega_x)\right|\left|\widehat{\psi}_y(\omega_y)\right|\left(S(-\omega_x)S(\omega_y)+S(\omega_x)S(-\omega_y)\right)
\end{aligned} \tag{2.14}$$

where the last factor selects two quadrants (as illustrated in Figure 2.7). Hence, by using a pair of complex wavelets: $\Psi_{c,x}(\omega_x)$, $\Psi_{c,y}(\omega_y)$ and their conjugates $\overline{\Psi_{c,x}(\omega_x)}$, $\overline{\Psi_{c,y}(\omega_y)}$, followed by taking the real parts of the resulting complex wavelet coefficients, we obtain two orientation bands with angles $\pm 45°$. This elegantly solves the directionality problem of the separable DWT, that cannot distinguish between orientation angles $+45°$ and $-45°$. When wavelet filters are used together with scaling filters in a multiresolution approach, an analysis is possible in 6 orientation angles instead of 3, as shown in Figure 2.8. Even a further decomposition in a higher number of orientation angles is possible by using complex wavelet packets [Bayram and Selesnick, 2008]. Also remarkable is that the real or imaginary parts of the complex wavelets are *not* separable. The DT-CWT provides directional selectivity using separable wavelet filters.

Finally, it is worth to point out that the 2D DT-CWT is easily implemented using 4 DWT transforms (or 4 wavelet trees) in parallel. Each DWT then uses real-valued wavelets: $\psi_x(t)\psi_y(t), \psi_x(t)\mathcal{H}\{\psi_y\}(t), \mathcal{H}\{\psi_x\}(t)\psi_y(t)$ and $\mathcal{H}\{\psi_x\}(t)\mathcal{H}\{\psi_y\}(t)$. Note that each of the four DWTs has a frequency tiling as in Figure 2.8a. To arrive at *oriented* complex wavelet subbands as explained above (Figure 2.8), an extra operation needs to be performed to the output of each tree. To see this, we write the real and imaginary part of the 2D complex wavelets in terms of the real-valued wavelets $\psi_x(t)$ and $\psi_y(t)$:

$$\begin{aligned}
\psi_{re,1}(x,y) &= 2^{-1/2}\left(\psi_x(x)\psi_y(y) - \mathcal{H}\{\psi_x\}(x)\mathcal{H}\{\psi_y\}(y)\right), \\
\psi_{im,1}(x,y) &= 2^{-1/2}\left(\mathcal{H}\{\psi_x\}(x)\psi_y(y) + \psi_x(x)\mathcal{H}\{\psi_y\}(y)\right), \\
\psi_{re,2}(x,y) &= 2^{-1/2}\left(\psi_x(x)\psi_y(y) + \mathcal{H}\{\psi_x\}(x)\mathcal{H}\{\psi_y\}(y)\right), \\
\psi_{im,2}(x,y) &= 2^{-1/2}\left(-\mathcal{H}\{\psi_x\}(x)\psi_y(y) + \psi_x(x)\mathcal{H}\{\psi_y\}(y)\right),
\end{aligned} \tag{2.15}$$

where $2^{-1/2}$ is an energy normalization constant. Hence, when we want to compute the inner products of an image $f(x,y)$ with $\psi_{re,1}(x,y)$, we need to subtract wavelet coefficients as follows:

$$\begin{aligned}
\langle f,\psi_{re,1}\rangle &= 2^{-1/2}\left(\langle f,\psi_x(x)\psi_y(y)\rangle - \langle f,\mathcal{H}\{\psi_x\}(x)\mathcal{H}\{\psi_y\}(y)\rangle\right) \\
\langle f,\psi_{re,2}\rangle &= 2^{-1/2}\left(\langle f,\psi_x(x)\psi_y(y)\rangle + \langle f,\mathcal{H}\{\psi_x\}(x)\mathcal{H}\{\psi_y\}(y)\rangle\right)
\end{aligned} \tag{2.16}$$

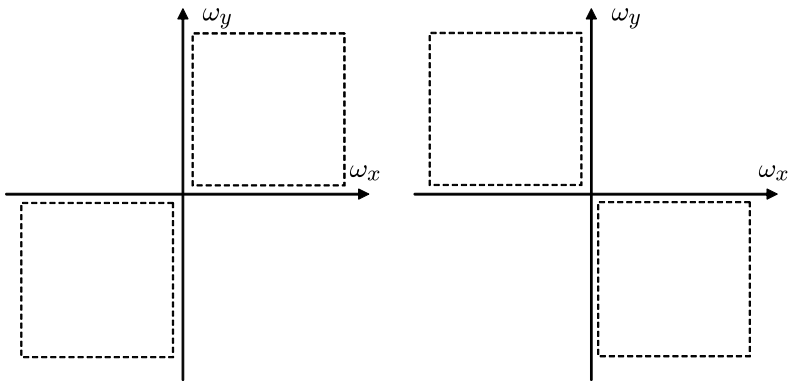


Figure 2.7: Frequency supports of the real (or imaginary) parts of the complex wavelets (a) $\psi_{re,1}(\omega_x, \omega_y)$ or $\psi_{im,1}(\omega_x, \omega_y)$ (b) $\psi_{re,2}(\omega_x, \omega_y)$ or $\psi_{im,2}(\omega_x, \omega_y)$

and with similar expressions for the other equations. The calculation in (2.16) is even more conveniently written in matrix form:

$$\begin{pmatrix} \langle f, \psi_{re,1} \rangle \\ \langle f, \psi_{re,2} \rangle \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \langle f, \psi_x(x)\psi_y(y) \rangle \\ \langle f, \mathcal{H}\{\psi_x\}(x)\mathcal{H}\{\psi_y\}(y) \rangle \end{pmatrix}. \quad (2.17)$$

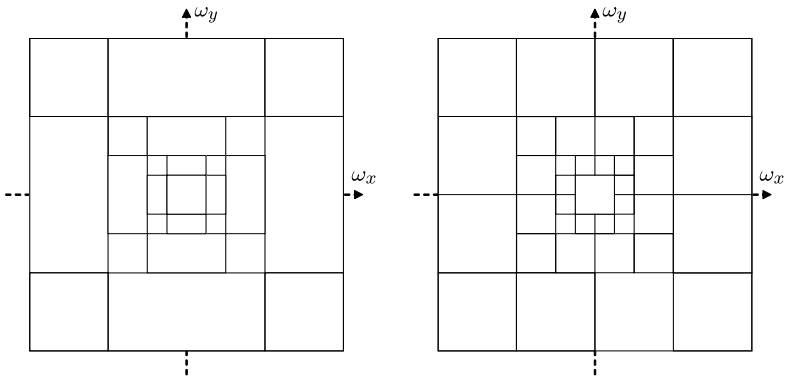
Since the matrix in (2.17) is invertible, the inverse DT-CWT can be implemented first by inverting (2.17) and next by applying inverse DWTs to the four trees.

Finally, as an illustration, the 2D DT-CWT of *barbara* is shown in Figure 2.9. The stripes in the trousers of *Barbara* have orientations $+45°$ (left leg) and $-45°$ (right leg). These features can be distinguished from each other in the subbands corresponding to these orientations.

2.2.3 Improving the Directional Selectivity of the First Scale

In this section, we present our novel extension to the DT-CWT, which improves the directional selectivity of the first scale of the transform. We shall start from a thorough analysis and explanation of the problem.

As explained before, to obtain a good directional selectivity, we must have that $\phi_2(t) = \mathcal{H}\{\phi_1\}(t)$ and $\psi_2(t) = \mathcal{H}\{\psi_1\}(t)$. A reasonable question is: how do the filter coefficients $h_2(n)$ and $g_2(n)$ relate to $h_1(n)$ and $g_1(n)$ such that the two wavelets form a Hilbert pair?



(a) (b)

Figure 2.8: Frequency domain tiling of (a) the 2D DWT, with 3 orientation bands (b) 2D DT-CWT, with 6 orientation bands.

Figure 2.9: The 2D DT-CWT of *Barbara*: magnitude of the complex wavelet coefficients (black indicates a high magnitude, white corresponds to zero).

Let $H_k(z)$ and $G_k(z)$ denote the z-transforms of respectively $h_k(n)$ and $g_k(n)$, $k = 1, 2$. In [Selesnick, 2001] it was shown that the sufficient condition for Hilbert pair of wavelets is that

$$H_2(z) = H_1(z)z^{-1/2}, \tag{2.18}$$

or informally, $h_2(n) \approx h_1(n - 0.5)$, which means that there must be a *half-sample delay* between both scaling filters. This condition we will further call the *half-sample delay condition*. Given that the wavelets are orthogonal, the condition for the wavelet filters on the unit circle is given by [Dumitrescu, 2009]:

$$G_2(e^{j\omega}) = G_1(e^{j\omega})e^{j\omega/2}H_{\text{hil}}(\omega). \tag{2.19}$$

In (2.19), we can also recognize the half sample delay $e^{j\omega/2}$, although the delay is now negative.

In order to have analyticity (and hence a good directional selectivity and shift invariance) at every scale, each decomposition stage of the DT-CWT needs to satisfy the following condition: if the input signal of the second tree is one half-sample delayed from the input signal of the first tree, after one wavelet decomposition stage (i.e. filtering and decimation), there must also be a half-sample delay between the output signals of both trees.

According to this criterion, for the *first* (finest) scale of the DT-CWT, the input samples of each tree should also be delayed 1/2 sample. For example in 2D, there are 4 trees (2 horizontally, 2 vertically) and ideally, the input image should be shifted first by the offsets $(0, 0)$, $(0, 1/2)$, $(1/2, 0)$, $(1/2, 1/2)$, depending on the filters in each tree (whether we use $G_2(z)$ or $G_1(z)$ horizontally or vertically), as illustrated in Figure 2.10. A practical implementation of the scheme would require *interpolation* of the input image, as concerning the input, there is no information about samples that have been delayed 1/2 sample!

To get around this problem, in [Kingsbury, 2001, Selesnick et al., 2005b], it is proposed to use wavelet filters that are delayed with 1 sample (or $H_2(z) = H_1(z)z^{-1}$) such that starting from the second scale, the half-sample delay condition is fulfilled. This is equivalent to using an undecimated DWT for the first scale (with cycle spinning) and has the advantage of being shift-invariant. The downside is that the wavelet pairs are *no longer Hilbert pairs* and we can expect a directional selectivity that is no better than the directional selectivity of a DWT (causing *checkerboard* problems)!

Let us analyze the severity of this problem. Figure 2.11 shows the magnitude responses of the 2D DT-CWT basis functions for different orientation subbands at the same scale and as function of the polar angle (this is done by integrating the 2D squared frequency response over line integrals that go to the origin of frequency space). The more concentrated the frequency response around the center peak, the better the directional selectivity. In Figure 2.11(a)-(b), i.e. for the second and third scales, the directional selectivity is quite good. The magnitude response for orientations at 45° and 135° contains sharper peaks and smaller side-lobes than the response for other orientations. This effect is caused by the rectangular tiling of the frequency domain (see Figure 2.8(b)): the DT-CWT basis functions are supported (approximately) on squares in frequency domain and integrating the squared magnitude response in the radial direction inherently results in asymmetry of the responses.[3] However, Figure 2.11(c) reveals the energy leakage for the orientations at 45° and 135° for the *first* scale of the DT-CWT. Here the checkerboard problem of the DWT remains, as predicted.

To have a consistent directional selectivity of the basis functions for *all scales* of the DT-CWT, we investigate the design of first-scale wavelet filters that satisfy the half-sample delay condition, such that there is one sample delay

[3] In [Kingsbury, 2006] the rotational symmetry of the basis functions is improved for analysis tasks.

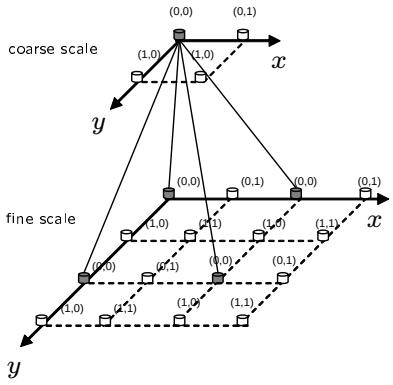


Figure 2.10: Illustration of the half-sample delay condition: each pair $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$ signifies a tree of the DT-CWT. The samples of each tree are located midway the samples of the other trees, e.g. between the samples of tree $(0,0)$ and $(0,1)$ there is a half-sample delay of $(0, \frac{1}{2})$. This recursive property is retained when zooming in to a more coarse scale. For the first (finest) scale of the DT-CWT transform, the same input image is used for each tree, which means that the samples are coincident and there is no half-sample delay, hence a correction is needed.

between the corresponding scaling filters. To do so, we use an undecimated wavelet transform for the first scale and subsequently impose the half-sample delay condition (2.19) to the wavelet filters, while leaving the scaling filters untouched (i.e. one sample delay between the scaling filters):

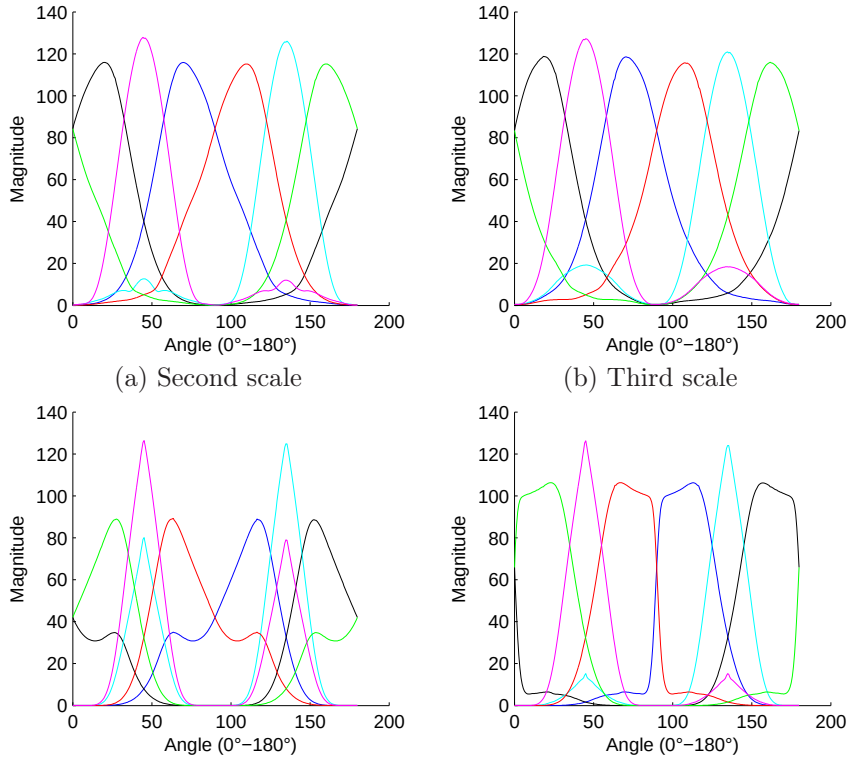
$$\begin{aligned} H_2(z) &= H_1(z)z^{-1} && (2.20)\\ G_2(z) &= G_1(z)z^{-1}Q(z) && (2.21) \end{aligned}$$

where $Q(z)$ is a filter that we will design to satisfy (2.19). Obviously, by modifying the wavelet filters, the perfect reconstruction property of the wavelet scheme will be lost. To prevent this, we will choose $Q(z)$ appropriately, such that this is not the case. The wavelet and scaling filters satisfy the following perfect reconstruction (PR) conditions [Daubechies, 1992]:

$$\begin{aligned} G_1(z)\tilde{G}_1(z^{-1}) + H_1(z)\tilde{H}_1(z^{-1}) &= 2 \quad \text{and} \\ G_1(z)\tilde{G}_1(-z^{-1}) + H_1(z)\tilde{H}_1(-z^{-1}) &= 0. \end{aligned} \tag{2.22}$$

For the modified wavelet filters in (2.20)-(2.21), the perfect reconstruction conditions are given by:

$$\begin{aligned} G_1(z)\tilde{G}_1(z^{-1})Q(z)\tilde{Q}(z^{-1}) + H_1(z)\tilde{H}_1(z^{-1}) = 2 \quad \text{and} \\ G_1(z)\tilde{G}_1(-z^{-1})Q(z)\tilde{Q}(-z^{-1}) + H_1(z)\tilde{H}_1(-z^{-1}) = 0. \end{aligned} \tag{2.23}$$



(a) Second scale (b) Third scale

(c) First scale (without corrections) (d) First scale (improved)

Figure 2.11: Directional selectivity of various scales of the DT-CWT. In the plots, the magnitude response of the wavelet basis functions is shown as function of the polar angle, for different orientation bands at the same scale.

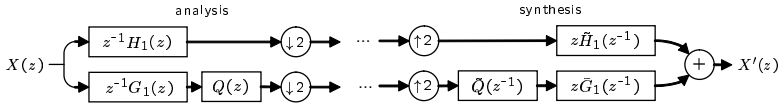


Figure 2.12: Proposed analysis and synthesis stages for the second tree of the DT-CWT (Solution using IIR filters): the scaling filters are delayed one sample, such that the half-sample delay condition is satisfied starting from the second scale. The wavelet filters are filtered by the allpass filter $Q(z)$.

Identification with the PR conditions (2.22) for the first QMF pair, immediately gives the following design constraints for $Q(z)$:

$$Q(-z)\tilde{Q}(z^{-1}) = 1 \quad \text{and} \quad Q(z)\tilde{Q}(z^{-1}) \;=\; 1. \tag{2.24}$$

These conditions are satisfied if $Q(z) = Q(-z)$ and $\tilde{Q}(z) = Q^{-1}(z^{-1})$, hence $Q(z)$ can only have terms in even powers of z and $\tilde{Q}(z)$ is the inverse of the

time reversed version of $Q(z)$. To impose the half-sample delay condition on $Q(z)$, equation (2.19) requires that $Q(e^{j\omega}) \approx e^{3j\omega/2} H_{\text{hil}}(\omega)$, which gives us the following design constraints for $Q(z)$:

$$\begin{aligned} Q(e^{j\omega}) &\approx e^{3j\omega/2} H_{\text{hil}}(\omega) && (2.25) \\ Q(z) &= Q(-z) && (2.26) \end{aligned}$$

Following (2.25), we find $\big|Q(z)\big| = \big|e^{3j\omega/2}\big|\,|H_{\text{hil}}(\omega)| = 1$, such that $Q(z)$ passes all frequencies equally. This leads us to the allpass filter design $Q(z) = B(z)/A(z)$ [Laakso et al., 1996, Baher, 2001] with

$$\begin{aligned} B(z) &= 1 + \sum_{m=1}^{M} q_m z^{2m} \\ A(z) &= B(1/z) \\ &= 1 + \sum_{m=1}^{M} q_m z^{-2m} && (2.27) \end{aligned}$$

where M is the filter order. With this choice, $Q(z) = B(z)/A(z)$ is an allpass filter. Because $A(z) = A(-z)$ and $B(z) = B(-z)$, the requirement $Q(z) = Q(-z)$ is also automatically fulfilled. Now, from equation (2.25) it follows that $B(e^{j\omega}) \approx A(e^{j\omega}) e^{3j\omega/2} H_{\text{hil}}(\omega)$. We therefore define the following error function:

$$E_Q = \int_{-\pi}^{+\pi} \Big| B(e^{j\omega}) - A(e^{j\omega}) e^{3j\omega/2} H_{\text{hil}}(\omega) \Big|^2 \,\mathrm{d}\omega. \qquad (2.28)$$

Minimizing E_Q is a least-squares problem, which leads to a linear system of equations. The linear system to solve is described by:

$$\begin{pmatrix} \frac{1}{5} & \frac{1}{9} & \frac{1}{13} & \cdots & \frac{1}{1+4M} \\ \frac{1}{9} & \frac{1}{13} & \frac{1}{17} & \cdots & \frac{1}{5+4M} \\ \frac{1}{13} & \frac{1}{17} & \frac{1}{21} & \cdots & \frac{1}{9+4M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1+4M} & \frac{1}{5+4M} & \frac{1}{9+4M} & \cdots & \frac{1}{1+8M} \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_M \end{pmatrix} - \frac{\pi}{2} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_M \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{5} \\ \frac{1}{9} \\ \vdots \\ \frac{1}{4M-3} \end{pmatrix}$$

where the matrix on the left hand side is a Hankel matrix. An efficient Matlab program to design the filter $Q(z)$ is given in Table 2.1. For $M = 1$, the solution $Q(z)$ is given by:

$$Q(z) = \tilde{Q}(z) = \frac{1 - 0.5647176836 z^2}{1 - 0.5647176836 z^{-2}} \qquad (2.29)$$

and the second order solution is ($M = 2$):

$$Q(z) = \tilde{Q}(z) = \frac{1 - 0.5594687532 z^2 - 0.0836530813 z^4}{1 - 0.5594687532 z^{-2} - 0.0836530813 z^{-4}}$$

Because $z^{-2M}\tilde{Q}(z)$ is causal and infinite, the anticausal filter $z^{2M}\tilde{Q}(z^{-1})$ can only be realized by filtering the time-reversed input signals with $\tilde{Q}(z)$, followed by time-reversal of the result. This practically corresponds to filtering from right to left (future to past) instead of from left to right (past to future).

A wavelet constructed using the modified filtering scheme is shown in Figure 2.13: on the top, the Daubechies wavelet with two vanishing moments is shown. In the middle, the Hilbert transform of this wavelet is drawn (such that both wavelets form a Hilbert filter pair, and consequently $\psi_c(t)$ is analytic). On the bottom, the IIR-allpass based approximation of the Hilbert transform of this wavelet (according to (2.21)) is shown. In Figure 2.13b, the frequency response of the complex wavelet formed by the Daubechies wavelet and its Hilbert transform approximation is visualized and it can be seen that there is a good suppression of negative frequencies. The improved analyticity leads to a better directional selectivity that is consistent to the other scales of the DT-CWT, which is shown in Figure 2.11(d).

Experimental results

As an illustration, we test the improved filters in a practical application: the estimation of the dominant orientation of edges in an image. If we denote the complex wavelet coefficients for the first scale of the DT-CWT as $x_{k,i}$, with $k = 1, ..., 6$ the orientation index and with i a one-dimensional index of the position in the wavelet subband (e.g. obtained using raster scanning), the dominant orientation ϑ_i at position i can be estimated as:

$$\hat{\vartheta}_i = \frac{1}{2}\angle\left\{\sum_{k=1}^{6}|x_{k,i}|^2\exp\left(-2j\vartheta_k^{(\mathrm{ref})}\right)\right\}, \tag{2.30}$$

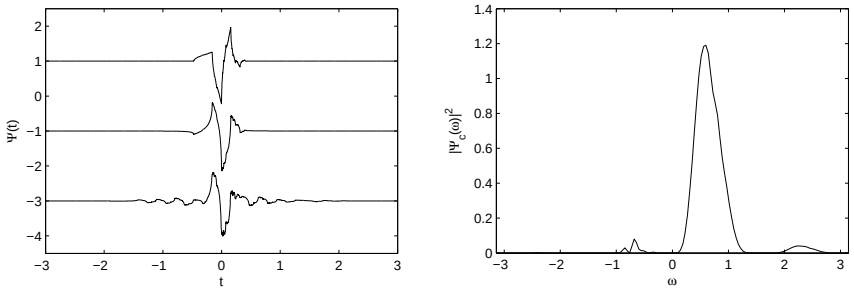
with $\angle z$ the argument of the complex number z and with $\vartheta_k^{(\mathrm{ref})}$ the reference dominant orientation of the complex wavelet at orientation k. In this experiment, we used the simple formula $\vartheta_k^{(\mathrm{ref})} = 2.548 + \pi(k-1)/6$, assuming that the dominant orientation angles of the complex wavelets are equally spaced. The estimated dominant orientation is shown in Figure 2.14 for the *zoneplate* image.[4] Because of the rotational symmetry of this image, it is expected that the orientation angle is approximately constant on radial lines that intersect with the center of the image. In Figure 2.14(b), the orientation angles are clearly misestimated; this effect is the most prominent in the high frequencies (i.e. corners of the image) at angles $\pm 45°$. In fact, the misestimation can be explained by the checkerboard problem of the unmodified first-scale DT-CWT basis functions, which was mentioned earlier. With the proposed complex wavelet filters, this problem is solved, as can be seen in Figure 2.14(c).

Finally, we also analyze the improvement in directional selectivity in 3D. In Figure 2.15, frequency responses of the real parts of the complex wavelet basis

[4] Visit `http://telin.ugent.be/~bgoossen/wavelets.htm` for testing the effects of the wavelet being used and for estimating the edge orientation in different images.

Table 2.1: Matlab program for designing the filter $Q(z)$.

```
b=-1./(1:4:-3+4*M)';
C=hankel(1./(5:4:-3+8*M));
C=C(1:M,1:M)+pi/2*eye(M);
q=[1; C\b];
B=upsample(q(end:-1:1),2); % B(z)
A=upsample(q,2); % A(z)
```



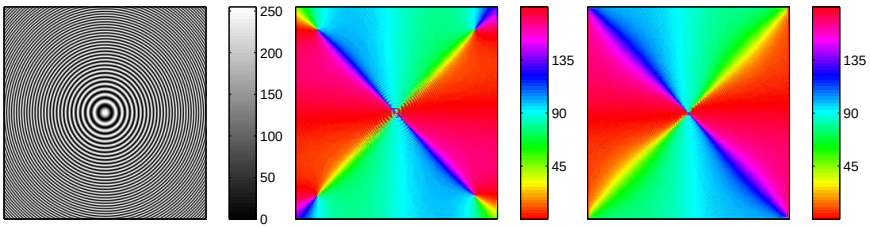
(a) (b)

Figure 2.13: (a) Plots of different wavelets: *(top)* Daubechies wavelet with two vanishing moments (Daub2) *(middle)* Hilbert transform of the Daub2-wavelet (bottom) approximation of the Hilbert transform of the Daub2-wavelet, obtained using the IIR wavelet filter solution (with $M = 4$). (b) Squared magnitude response of $\Psi_c(\omega)$, for the IIR wavelet filter solution, illustrating the near-analyticity of the complex wavelet.

functions are given and compared to the original wavelet filters (Figure 2.15(a)-(b)). In Figure 2.15(a)-(b) there is leakage of the filter energy to different orientations, which results in a poor directional selectivity. For the proposed filters, the frequency responses are well-localized and as expected.

2.3 The steerable pyramid transform

The steerable pyramid (STP) transform [Simoncelli et al., 1992, Simoncelli and Freeman, 1995] is a 2D multiresolution transform that, like the DT-CWT, has been introduced to overcome limitations of the DWT. Detailed information about this transform can be found in [Freeman and Adelson, 1991, Simoncelli et al., 1992, Simoncelli and Freeman, 1995], here we will only review the basic concepts behind the STP that are of importance for the remainder of this work.



(a) (b) (c)

Figure 2.14: Edge orientation estimation of the "zoneplate" image: (a) the "zoneplate" image, (b) estimated edge orientation (in degrees) *without* the proposed filter design, (c) estimated edge orientation (in degrees) *with* the proposed filter design.

The STP transform is based on the concept of "steerable filters", which has been introduced for image analysis tasks as an efficient way to do scale-space analysis [Freeman and Adelson, 1991]. In many image processing applications, one needs to apply the same filter several times, but each time under a different rotation angle. An example could be the estimation of the dominant orientation of an edge. Instead of having to investigate e.g. 360 different orientations to have an estimation accuracy of $\pm 0.5°$, steerable filters allow to use only a *small* number of filters and by making linear combinations of the basis filter responses, the filter response in *any* direction can be obtained. This not only results in a huge savings in computation time, but also facilitates rotationally invariant processing [Simoncelli, 1996].

2.3.1 Steerability

While steerability is usally explained for designing oriented filters, we will introduce this concept from a slightly broader perspective, to illustrate that steerability is even applicable to a wide range of problems. Therefore, we introduce steerability already for 1D functions, which will also raise some interesting thoughts.

We say that a function $f(t)$ is steerable in t, if any shifted version of this function $f(t-t_0)$ with $t_0 \in \mathbb{R}$, can be written as a linear sum of a fixed number of shifted functions $f(t-t_k)$, $k = 1, ..., K$, for a given K [5] and with t_k constant and independent of $f(t)$.

For example, let $f(t) = \cos t$, then

$$\begin{aligned} f(t-t_0) &= \cos(t-t_0) \\ &= \cos t \cos t_0 + \sin t \sin t_0 \\ &= \cos t \cos t_0 + \cos\left(t - \frac{\pi}{2}\right) \sin t_0 \\ &= b_1(t_0) \cos t + b_2(t_0) \cos\left(t - \frac{\pi}{2}\right). \end{aligned}$$

[5] Note that a function may be steerable for different values of K. In this case, we will take the smallest value of K under which this function is steerable.

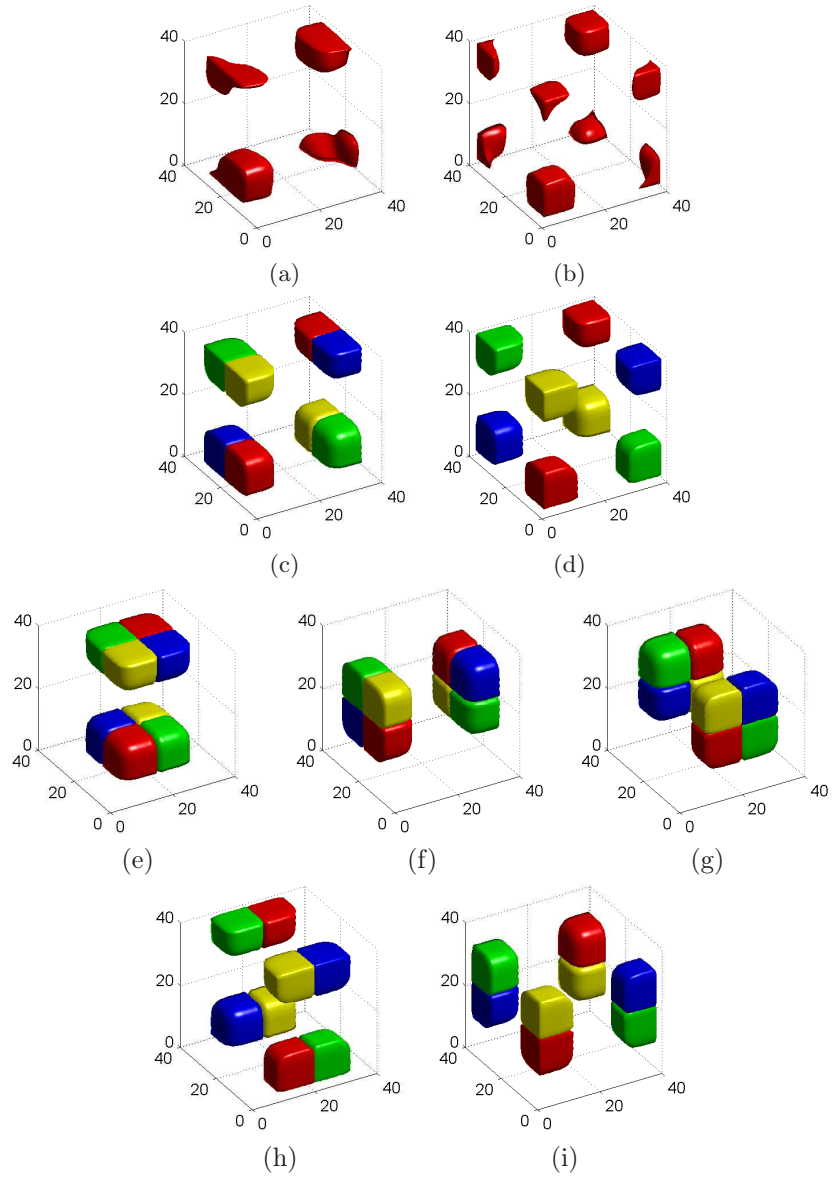


Figure 2.15: Iso-energy plot of the frequency responses of the real parts of the basis functions for the first scale of the 3D DT-CWT (different colors signify different orientations), for (a)-(b) The original wavelet filters (*Db8*), in *one orientation* only (corresponding to the red surfaces in (c)-(d)). (c)-(d) The proposed wavelet filters (*Db8*), for the first order $M = 1$, and in *four orientations*. For the original wavelet filters (top row), there is leakage of the basis function energy to different orientations, indicating a poor directional selectivity. With the proposed filters (bottom row), the leakage is well suppressed. (e)-(i) The remaining directional filters (in 3D there are 28 orientations).

From our definition, it follows that $f(t)$ is steerable in t, with $K = 2$, $t_1 = 0$ and $t_2 = \frac{\pi}{2}$. The functions $b_1(t) = \cos t$ and $b_2(t) = \sin t$ are called interpolation functions (the reason for this name will be made clear later).

At first sight, a definition of steerability does not seem to be very useful, because the function $f(t)$ is completely known and shifted versions can be readily obtained by simply shifting the argument t. To recognize the primary advantage of steerability, consider $f(t)$ as a basis function and project an arbitrary function $g \in L^2(\mathbb{R})$ onto a fixed number of shifted versions of this basis function. Let us denote the corresponding projections by $c(t_k)$, $k = 1, ..., K$, then:

$$c(t_k) = \langle g, f(\cdot - t_k) \rangle = \int_{-\infty}^{+\infty} g(t)\overline{f(t - t_k)}\mathrm{d}t,\ k = 1, ..., K.$$

Based on this small set of projections, we can compute the projection on any shifted version of the basis function:

$$c(t_0) \quad = \quad \int_{-\infty}^{+\infty} g(t)\overline{f(t - t_0)}\mathrm{d}t = \sum_{k=1}^{K} c(t_k)\, b_k(t_0), \tag{2.31}$$

which results in a linear combination of the projections on the basis functions (or *interpolation*). Instead of having to project $g(t)$ onto a continuous range of basis functions, the same result can be obtained by projecting onto a very small number (e.g. $K = 2$) of basis functions and by making a linear combination of the projection coefficients.

In the previous example, we dealt with a very simple function, $f(t) = \cos t$, and this gave us the interpolation functions $b_1(t_0)$ and $b_2(t_0)$. Note however that $b_k(t_0)$ and the shifts t_k are not uniquely defined according to our definition of steerability. For example, if we had chosen $t_0 = 0$ and $t_1 = \pi/3$, we would have found:

$$\begin{aligned} b_1(t_0) &= -1 + 2\cos t_0, \\ b_2(t_0) &= \frac{2}{\sqrt{3}} \sin t_0. \end{aligned}$$

In other words: the sine basis functions do not need to be in quadrature phase! It turns out that t_0 and t_1 can be freely specified (as long as they are not equal), as we will explain further on.

Also, we are interested in investigating the steerability of an arbitrary function instead of only sine (or cosine) functions. In the following, we will first consider real-valued functions $f(t) \in L^2([0, 2\pi])$ that are 2π periodic. The shifting operation is then interpreted modulo 2π. These functions can be expanded into a converging complex Fourier series:

$$f(t) = \sum_{k=-\infty}^{+\infty} a_k e^{jkt} \tag{2.32}$$

Again, we can formulate the steerability condition:

$$f(t-t_0) = \sum_{k=-\infty}^{+\infty} a_k e^{jk(t-t_0)} = \sum_{k=-\infty}^{+\infty} \sum_{l=1}^{K} a_k e^{jk(t-t_l)} b_l(t_0) \tag{2.33}$$

For real-valued functions, we have the conjugate symmetry property $a_k = \overline{a_{-k}}$. Now, if the function only consists of the first K Fourier terms (i.e. if $a_k = 0$ for $|k| \geq K$), the above equation amounts to a linear system of K equations and unknowns $b_k(t_0), k = 1, ..., K$:

$$\begin{pmatrix} 1 \\ e^{jt_0} \\ \vdots \\ e^{j(K-1)t_0} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ e^{jt_1} & e^{jt_2} & \cdots & e^{jt_K} \\ \vdots & \vdots & \ddots & \vdots \\ e^{jt_1(K-1)} & e^{jt_2(K-1)} & \cdots & e^{jt_K(K-1)} \end{pmatrix} \begin{pmatrix} b_l(t_0) \\ b_2(t_0) \\ \vdots \\ b_K(t_0) \end{pmatrix}, \tag{2.34}$$

which has a unique solution. Hence, for functions consisting of K Fourier terms, K basis functions $f(t-t_k)$ are needed and the interpolation functions are solutions of (2.34). This is known as Steering Theorem 1 in [Freeman and Adelson, 1991]. This also implies that, independent of the exact values of a_k, the interpolation functions only depend on the shifts t_k. In practice, uniform shifts in the interval $[0, 2\pi]$ are mostly used:

$$t_k = \frac{2\pi}{K}(k-1), \quad k = 1, ..., K \tag{2.35}$$

For this choice, the matrix in (2.34) amounts to the Fourier transform matrix and the interpolation functions are uniquely given by:

$$b_k(t_0) = \frac{1}{K} \sum_{l=-(K-1)/2}^{(K-1)/2} \cos(2(t_k - t_0)l) \tag{2.36}$$

$$= \frac{1}{K} \frac{\sin((t_k - t_0)K)}{\sin(t_k - t_0)} \tag{2.37}$$

where in (2.36) we use the convention that for even K the summation takes place over non-integer values of l. The function in (2.37) is commonly known as the Dirichlet function (see Figure 2.16). The *Dirichlet* function is very similar to the *sinc* function, with the main difference that the Dirichlet function is periodic. If we now turn again to (2.31), unsurprisingly we find the result:

$$c(t_0) = \sum_{k=1}^{K} c(t_k) \frac{1}{K} \frac{\sin((t_k - t_0)K)}{\sin(t_k - t_0)}, \quad \text{with } t_0 \in \mathbb{R} \tag{2.38}$$

This is basically the sampling theorem for 2π periodic functions. Recall that we have called $b_k(t_0)$ interpolation functions. This is clear now: "intermediate" samples of $c(t)$ can be obtained by interpolating the samples $c(t_k), k = 1, ..., K$.

The basis process to design a steerable function is hence as follows:

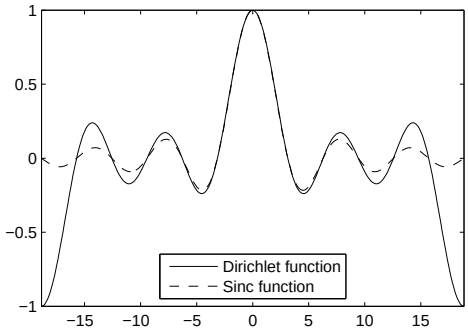


Figure 2.16: The dirichlet function (with $K = 6$) and the sinc function.

- First select the number of Fourier terms K for $f(t)$.
- Choose the shifts t_k, $k = 1, ..., K$, for example use uniform shifts as in (2.35).
- Obtain the interpolation functions $b_k(t_0)$ by solving (2.34). For uniform shifts this gives (2.37).
- Determine the coefficients a_k.

The last step deserves some extra attention. In essence, we are free to choose the coefficients $a_k, k = 0, ..., K-1$. In practice, we want $f(t)$ to be well-localized in time (to extract a certain frequency band or orientation, see further in Section 2.3.2). Ideally, a rectangular function can be used, however, this would require an infinite number of Fourier terms. Instead, we can approximate the rectangular function by a bandlimited function by taking the K most significant Fourier terms:

$$f_{\text{square}}(t) = \frac{1}{K}\left(1 + \sum_{k=1,3,5,\ldots}^{K-1} \frac{1}{k}\sin(kt)\right), \tag{2.39}$$

with K even. If the time shifts t_k are uniformly spaced ($t_k = 2\pi(k-1)/K$), the interpolation functions are given by (2.36)-(2.37). Moreover, it can be checked that $f(t)$ as defined in (2.39) satisfies other properties for analysis:

$$\begin{aligned} f(t) &\geq 0 \\ \sum_{k=1}^{K} f(t - t_k) &= 1, \quad t \in [0, 2\pi] \end{aligned}$$

On the other hand, in [Portilla et al., 2003], the following steerable function is used:

$$f_{\cos}(t) = \frac{2^{K-1}(K-1)!}{\sqrt{K(2K-2)!}}\cos^{K-1}(t/2) \tag{2.40}$$

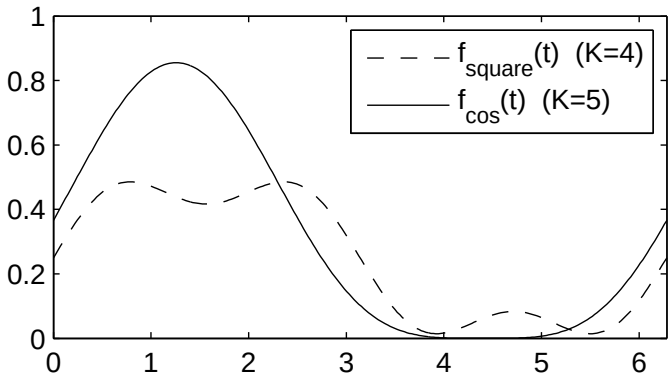


Figure 2.17: Illustration of two steerable functions: $f_{\text{square}}(t)$ and $f_{\cos}(t)$.

with time shifts again uniformly spaced but on the interval $[0, 2\pi]$: $t_k = 2\pi(k-1)/K$. By iteratively using the formula $\cos^2 t = (1 + \cos 2t)/2$, it can be seen that $f(t)$ consists of K Fourier terms. The magnitude of the function is maximal for $t = 0$ and $t = \pi$. Furthermore, the function has the following properties:

$$\begin{aligned} \sum_{k=1}^{K} f(t - t_k) &= \text{constant} \quad \text{if } K \text{ is odd} \\ \sum_{k=1}^{K} |f(t - t_k)|^2 &= 1 \quad \text{if } K > 0 \end{aligned} \tag{2.41}$$

The property (2.41) is very useful in the design of the steerable pyramid transform, as we will see further on. The functions (2.39) and (2.40) are illustrated in Figure 2.17.

2.3.2 Steerable filters

Based on the concept of "steerable functions", a steerable filter $G(\omega, \vartheta)$ can be easily designed in polar (frequency) coordinates [Freeman and Adelson, 1991]:

$$G(\omega, \vartheta) = R(\omega) f(\vartheta) \tag{2.42}$$

where ω is the radial frequency, ϑ is the angle, $R(\omega)$ is the frequency response of a radially symmetric filter and $f(\vartheta)$ is a steerable function, as defined in Section 2.3.1. Although the term "steerable filters" is typically reserved for filters that can be steered in *orientation*, we remark that steering in *scale* is also possible: in essence, it suffices to use a second steerable function for $R(\omega)$. This will be of importance in Section 7.3.3.

As a simple example, consider again $f(\vartheta) = \cos\vartheta$ for $K = 2$ and $\vartheta_1 = 0$,

$\vartheta_2 = \pi/2$. This gives the steerable filters:

$$\begin{aligned} G_1(\omega,\vartheta) = G(\omega,\vartheta-\vartheta_1) &= R(\omega)\cos\vartheta \text{ and} \\ G_2(\omega,\vartheta) = G(\omega,\vartheta-\vartheta_2) &= R(\omega)\sin\vartheta. \end{aligned}$$

To steer the filters for an arbitrary angle of ϑ_0 radians, we just need to apply the steering matrix $\begin{pmatrix} \cos\vartheta_0 & -\sin\vartheta_0 \\ \sin\vartheta_0 & \cos\vartheta_0 \end{pmatrix}$. The filters rotated by ϑ_0 radians are:

$$\begin{pmatrix} \cos\vartheta_0 & -\sin\vartheta_0 \\ \sin\vartheta_0 & \cos\vartheta_0 \end{pmatrix} \begin{pmatrix} G_1(\omega,\vartheta) \\ G_2(\omega,\vartheta) \end{pmatrix} = \begin{pmatrix} R(\omega)\cos(\vartheta-\vartheta_0) \\ R(\omega)\sin(\vartheta-\vartheta_0) \end{pmatrix}. \tag{2.43}$$

In Cartesian frequency coordinates, the filters can also be written as follows:

$$\begin{aligned} G_1(\omega_x,\omega_y) &= R\left(\sqrt{\omega_x^2+\omega_y^2}\right)\frac{\omega_x}{\sqrt{\omega_x^2+\omega_y^2}} \text{ and} \\ G_2(\omega_x,\omega_y) &= R\left(\sqrt{\omega_x^2+\omega_y^2}\right)\frac{\omega_y}{\sqrt{\omega_x^2+\omega_y^2}}. \end{aligned}$$

Because of the steerability of $f(\vartheta)$, all properties relating to steerability also apply to $G(\omega,\vartheta)$. For example, if we use $f(\vartheta)$ from (2.40) with orientation angles $\vartheta_k = \pi(k-1), k = 1,...,K$, property (2.41) amounts to:

$$\sum_{k=1}^{K} |G(\omega,\vartheta-\vartheta_k)|^2 = |R(\omega)|^2, \tag{2.44}$$

which means that the sum of the power spectral densities of the steerable filters $G(\omega,\vartheta-\vartheta_k)$ only depends on the power spectral density of the radially symmetric filter. Now, suppose we have a number of radially symmetric filters $R_i(\omega), i = 1,...,I$ with a similar property:

$$\sum_{i=1}^{I} |R_i(\omega)|^2 = 1,$$

and with corresponding steerable filters $G_{i,k}(\omega,\vartheta) = R_i(\omega)f(\vartheta-\vartheta_k)$, we could immediately obtain a subband decomposition scheme in I scales and K orientations. For this scheme, the analysis and synthesis filters are given by respectively $G_{i,k}(\omega,\vartheta)$ and $\overline{G_{i,k}(\omega,\vartheta)}$. The perfect reconstruction of this scheme can readily be checked:

$$\begin{aligned} \sum_{i=1}^{I}\sum_{k=1}^{K} G_{i,k}(\omega,\vartheta)\overline{G_{i,k}(\omega,\vartheta)} &= \sum_{i=1}^{I}\sum_{k=1}^{K} |G_{i,k}(\omega,\vartheta)|^2 \\ &= \sum_{i=1}^{I} |R_i(\omega)|^2 = 1 \end{aligned}$$

In other words: analyzing an image using the set of filters $\{G_{i,k}(\omega,\vartheta)\}$ and resynthesizing the image using the complex conjugates of the filters and subsequently summing the results is equivalent to using a filter with frequency response 1, hence the original image is reconstructed. The steerable pyramid transform, which can be seen as a refinement of this scheme, will be explained next.

2.3.3 Architecture of the steerable pyramid transform

The STP performs a decomposition of the image in a variable number of scales (I) and orientations (K). As with the DWT, this process is done recursively by splitting the input image in a number of oriented subbands and a low-pass subband that is decimated by a factor of two in both dimensions. Unlike the DWT, the oriented subbands are not decimated. To obtain shift-invariance of the transform, the STP filters are designed to be bandlimited. The major benefit of the STP is that orientation subbands in *any* orientation $\alpha \in [0,\pi]$ can be synthesized by linear combinations of the STP subbands, computed for a *fixed* number of orientations. As the steering operation (see (2.43)) can be expressed as an orthonormal transformation in orientation space, the squared norm of the STP coefficients across orientation will be invariant to rotations. This is very useful for certain applications, such as image retrieval [Simoncelli, 1996].

Furthermore, the STP does not suffer from the problems mentioned in Section 2.1.4, at the cost of a high redundancy factor: the redundancy factor (i.e. the number of transform coefficients divided by the number of input pixels) of the transform is $7K/3$ for an infinite number of scales $I \to \infty$. For example, for 6 orientations, the redundancy factor of the STP is 14, compared to 4 for the DT-CWT. The frequency tiling of the STP is shown in Figure 2.18.

The frequency responses of the steerable STP filters are defined as [Portilla et al., 2003]:

$$H(\omega,\vartheta) = \begin{cases} \cos\left(\frac{\pi}{2}\log_2\left(\frac{4\omega}{\pi}\right)\right), & \frac{\pi}{4} < \omega < \frac{\pi}{2} \\ 1, & \omega \le \frac{\pi}{4} \\ 0, & \omega > \frac{\pi}{2} \end{cases} \tag{2.45}$$

$$G_k(\omega,\vartheta) = \begin{cases} f_k(\vartheta)\cos\left(\frac{\pi}{2}\log_2\left(\frac{4\omega}{\pi}\right)\right), & \frac{\pi}{4} < \omega < \frac{\pi}{2} \\ f_k(\vartheta), & \omega > \frac{\pi}{2} \\ 0, & \omega \le \frac{\pi}{4} \end{cases} \tag{2.46}$$

$$f_k(\vartheta) = \frac{(K-1)!}{\sqrt{K\,(2K-2)!}}\left(2\cos\left(\vartheta - \frac{\pi k}{K}\right)\right)^{K-1} \tag{2.47}$$

with $k = 1,...,K$. The radial response of $H(\omega,\vartheta)$ and $G_k(\omega,\vartheta)$ is a raised cosine function that is logarithmically warped to simulate dyadic scales. The angular response (2.47) is a *steerable* trigonometric function from (2.40). In Figure 2.18(b), a number of basis elements of the STP transform are depicted,

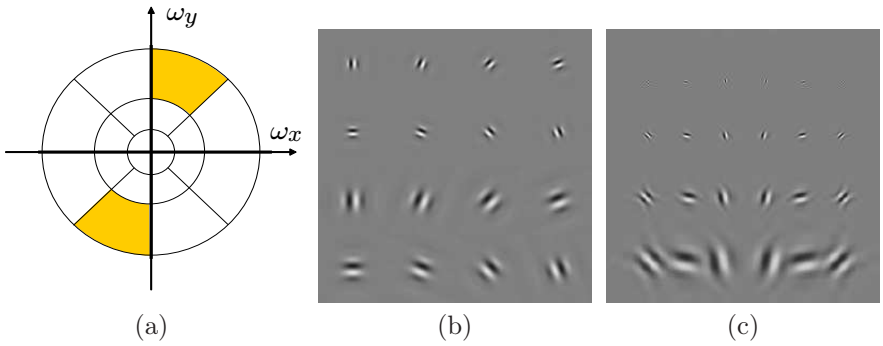


Figure 2.18: (a) Frequency tiling of a STP transform with two scales ($I = 2$) and eight orientations ($K = 8$), (b) Example of a set of basis function of the STP transform, (c) Basis functions of the DT-CWT.

for two different scales of the transform. As expected, the basis elements are rotated versions of each other (rotated by multiples of 22.5° in this case).

For comparison, we also computed basis elements of the DT-CWT (see Figure 2.18(c)). Despite the fact that the DT-CWT basis elements contain "staircasing artifacts" (this is due the approximate analyticity of the complex wavelets - see Section 2.2), there is a good resemblance between the complex wavelets and the STP basis elements. We remark that by the use of separable filters, the implementation of the DT-CWT is often computationally much more efficient than the STP, also the redundancy factor of the DT-CWT is lower (4 compared to $7K/3$). Therefore, we can consider the DT-CWT as a more practical transform with STP-like properties. Nevertheless, exact "steerability" properties only hold for the STP transform.

2.4 Overview of related representations

Due to the success of wavelets, complex wavelets and steerable pyramids in many areas of image processing, other related multiresolution transforms have been proposed as well, to further improve the properties and to obtain even more sparse representations for images. We make a distinction between *adaptive representations* (which are optimized with respect to the image being analyzed), and *non-adaptive representations*.

Adaptive representations

Many adaptive representations attempt to exploit the geometry that is present in images. Instead of decomposing the image in a fixed basis, adaptive techniques modify the representation dynamically based on the geometry computed from the image. In [Philips, 1996], bases of time-warped and spatial-warped orthonormal polynomials are used to efficiently adapt the polynomial basis to the

local bandwidth of the image. The *wedgelet* transform [Donoho, 1999, Romberg et al., 2002] divides the image into dyadic blocks at different scales and projects these blocks onto piecewise constant functions with linear discontinuities. The direction of the edges is locally estimated from the image. [Shukla et al., 2005] generalizes this approach by considering piecewise polynomial functions.

The *bandelet* transform [LePennec and Mallat, 2005, Mallat and Gabriel, 2007] further improves the sparsity of DWT coefficients by applying polynomial approximation to the wavelet coefficients, thereby constructing orthogonal vectors that are elongated along the edges in the image. The *grouplet* transform [Mallat, 2009] exploits non-local redundancy in images by further decomposing the wavelet subbands using the lifting scheme [Daubechies and Sweldens, 1996] of the Haar wavelet transform applied to "similar" wavelet coefficients. These similar coefficients are found using block-matching techniques.

The *directionlet* transform [Velisavljevic et al., 2006] partitions the image domain into integer lattices, where 1D filtering is performed along lines of the lattice yielding directional and anisotropic basis functions. The filtering direction is thereby adapted to the dominant orientation of each block of the partition.

Nonadaptive representations

Next to adaptive representations, a number of image-independent (nonadaptive) transforms have been proposed. *Brushlets* [Coifman and Meyer, 1997] improve the angular selectivity of wavelet packets by expanding the Fourier plane into a windowed Fourier basis. Brushlets are well localized in frequency and are good for representing texture-rich images.

The *contourlet* transform [Do and Vetterli, 2003a, Do and Vetterli, 2005] provides a multidirectional analysis for the *Laplacian* pyramid [Burt and Adelson, 1983, Do and Vetterli, 2001], a multiscale transform that was very successful in the 80s for e.g. image coding. While the *Laplacian* pyramid decomposes an image into multiple scales by successively filtering and downsampling and by computing the prediction error, the *contourlet* transform adds a second layer in which a directional filter bank is used to further analyze the frequency bands in multiple orientations. *Directional filter banks* [Bamberger and Smith, 1992] further make use of quincunx downsampling, and critical sampling is almost achieved: the redundancy factor is approximately 4/3.

A shortcoming of the contourlet transform is its shift-variance (as the DWT), therefore a nondecimated version of the *contourlet* transform has been proposed in [da Cunha et al., 2006], yielding excellent results in denoising at the cost of a relatively high redundancy factor ($1 + I \cdot K$, where I is the number of scales and where K is the number of orientations). *Surfacelets* [Lu and Do, 2007] are related to *contourlets*, but use N-D directional filterbanks based on pyramidal and wedge shaped filters to achieve an efficient multidirectional decomposition of higher dimensional data.

Phaselets [Gopinath, 2003] are a generalization of dual-tree complex wavelets and are designed to obtain better shift invariance and directional properties.

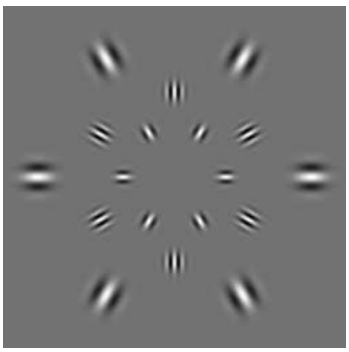


Figure 2.19: Illustration of 2D-Gabor functions.

While the *DT-CWT* uses scaling filters that are shifted one half sample (see Section 2.2), the *Phaselet transform* considers arbitrary shifts. Phaselets are also specific cases of *framelets* [Daubechies et al., 2003], which emerge from MRA and frame theory.

2D Gabor functions [Lee, 1996] (see Figure 2.19) are Gaussian functions multiplied by a complex exponential (or in the frequency domain - shifted Gaussian functions), well-known for their excellent localization in position, frequency and orientation. Furthermore, several neuroscience studies have shown that the responses of simple cells in the Primary Visual Cortex (V1) can be well modeled by Gabor functions [Field, 1987]. Consequently, *Gabor multiresolution analysis* has been successfully used in image analysis tasks, but somewhat less effective in image restoration tasks: the design of Gabor transforms with exact reconstruction is rather difficult, because the Fourier domain is not uniformly covered [Fischer et al., 2007]. Instead, *log-Gabor filters* tend to bring a better coverage of the frequency space and a self-invertible multiresolution transform based on log-Gabor filters has been introduced in [Fischer et al., 2007]. We remark that the Gabor functions have a very similar appearance as the STP basis functions (see Figure 2.18(b)).

Other recent additions to the (complex) wavelet family are: *polyharmonic wavelets* [Van De Ville et al., 2005], *Marr-like wavelet pyramids* [Van De Ville and Unser, 2008] and *steerable wavelet frames* [Unser and Van De Ville, 2009].

Related research focused on the study of multiscale geometrical transforms with a typically very high number of analysis orientations, such as *ridgelets* [Candès, 1998, Do and Vetterli, 2003b], *curvelets* [Starck et al., 2000, Candès et al., 2006] and *shearlets* [Guo and Labate, 2007]. The *ridgelet* transform achieves a multi-directional decomposition through the use of the *Radon* transform. Next, a 1D DWT is performed in the Radon domain, to obtain a further analysis in multiple scales. The *ridgelet* transform is well suited for representing discontinuities along *straight* lines. This is in contrast to the *curvelet* and *shearlet* transforms, which can represent discontinuities along curves with bounded curvatures. More specifically, it has been shown that if each basis

Table 2.2: Overview of the properties of a number of multiresolution transforms (I=number of scales, K=number of orientations).

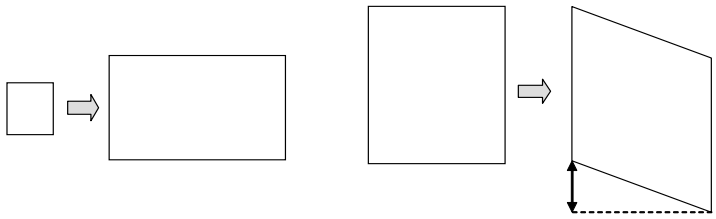
Representation	Sep. filtering	Phase	# orient.	Red. factor	Shift invariance
DWT	yes	no	2 (hor+vert)	1	no
dual-tree DWT	yes	no	6	2	no
DT-CWT	yes	yes	6	4	approx.
Undec. DWT	yes	no	2 (hor+vert)	$1+3I$	yes
Contourlet	no	no	power of 2	$\approx 4/3$	no
Unsubs. Contourlet	no	possible	variable	$1+I\cdot K$	yes
STP	no	possible	variable	$1.33K$	yes
log-Gabor	no	yes	variable	$4.66K$	yes
Curvelets	no	yes	variable	≈ 7.2	yes
Shearlets	no	possible	variable	≈ 2.6	yes

element has a frequency support that is contained in a rectangle of size proportional to $2^i \times 4^i$ (or $4^i \times 2^i$), or in other words if the length of the frequency support is approximately the squared width of the frequency support (called *parabolic scaling* property), we obtain a sparse representation that is optimal for representing images that contain *edges* [Candès et al., 2006, Guo and Labate, 2007]. This practically means that images with edges can be represented in these transform domains with much less significant coefficients than with other transforms.

A few properties of some of the above mentioned multiresolution transforms are listed in Table 2.2: *sep. filtering* indicates whether the implementation can be based on separable filtering, *phase* indicates whether phase information of the transform coefficients is available (which is useful for a number of applications), *# orient* gives the number of analysis orientations for the transform, *Red. factor* is the redundancy factor of the transform (the number of transform coefficients divided by the number of image pixels). Finally, *shift invariance* indicates whether processing of the transform coefficients can be done in a shift-invariant manner. The last transform, i.e. the shearlet transform, will now be discussed somewhat in more detail.

2.5 The shearlet transform

The shearlet transform [Guo and Labate, 2007], is a very recent sibling in the family of geometric image representations and provides a traditional multiresolution analysis (see Section 2.1.2). By a specific design of the discrete shearlet transform that we will explain in this section, a lower redundancy factor is possible than with most other multiresolution representations, while offering an excellent directional analysis and shift invariance. We will show that a design is possible with a redundancy factor as low as $8/3 \approx 2.6$, independent of the number of analysis orientations. This property, together with the optimality results for images that contain edges, make the shearlet transform an attractive candidate for image representation.



(a) (b)

Figure 2.20: Geometric transformations used by the shearlet transform (a) anisotropic dilation (matrix $\mathbf{A}$). (b) shear (matrix $\mathbf{B}$).

2.5.1 An introduction to shearlet theory

The continuous shearlet transform (CST) is a generalization of the continuous wavelet transform (Section 2.1) with basis functions well localized in *space*, *frequency* and *orientation*. Let $\psi_{i,k,\mathbf{l}}(\mathbf{p})$ denote the shearlet basis functions (or in the remainder simply called shearlets), then the CST of an image $f(\mathbf{p}) \in L^2\left(\mathbb{R}^2\right)$ is defined by [Guo et al., 2009, Yi et al., 2009]:

$$\left[\mathcal{SH}_\psi f\right](i,k,\mathbf{l}) = \int_{\mathbb{R}^2} f(\mathbf{p})\psi_{i,k,\mathbf{l}}(\mathbf{l}-\mathbf{p})\mathrm{d}\mathbf{p} \tag{2.48}$$

where $i \in \mathbb{R}$, $k \in \mathbb{R}$ and $\mathbf{l} \in \mathbb{R}^2$ denote the scale, orientation and the spatial location, respectively. The idea behind the continuous shearlet transform (CST) is to combine geometry and multiscale analysis [Easley et al., 2008]: shearlets are formed by dilating, shearing and translating a mother shearlet function $\psi \in L^2\left(\mathbb{R}^2\right)$, as follows:

$$\psi_{i,k,\mathbf{l}}(\mathbf{p}) = \left|\det\,\mathbf{A}\right|^{i/2}\psi\left(\mathbf{B}^k\mathbf{A}^i\mathbf{p}-\mathbf{l}\right) \tag{2.49}$$

where $\mathbf{A}$ and $\mathbf{B}$ are invertible 2×2 matrices, with $\det\,\mathbf{B}=1$. As with wavelets, the normalization factor $\left|\det\,\mathbf{A}\right|^{i/2}$ has been chosen such that the norm $\|\psi\|_2 = \|\psi_{i,k,\mathbf{l}}\|_2$ for all $i,k,\mathbf{l}$. The basis functions are subject to a composite dilation $\mathbf{A}^i$ and geometrical transform $\mathbf{B}^k$. For the shearlet analysis, the following transform matrices are being used:

$$\mathbf{A} = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}. \tag{2.50}$$

Here, $\mathbf{A}$ is an anisotropic scaling matrix (in the x-direction, the scaling is twice the scaling in the y-direction) and $\mathbf{B}$ is a geometric shear matrix. These transforms are illustrated in Figure 2.20.

The shearlet mother function is a composite wavelet that satisfies appropriate admissibility conditions [Guo et al., 2009], and that is defined in the Fourier transform domain as:

$$\widehat{\psi}(\boldsymbol{\omega}) = \widehat{\psi}_1\left(\omega_x\right)\widehat{\psi}_2\left(\frac{\omega_y}{\omega_x}\right) \tag{2.51}$$

with $\boldsymbol{\omega} = [\omega_x\ \omega_y]$, $\widehat{\psi}_1(\omega_x)$ the Fourier transform of a wavelet function and $\widehat{\psi}_2(\omega_y)$ a compactly supported bump function:

$$\widehat{\psi}_2(\omega_y) = 0 \Leftrightarrow \omega_y \notin [-1, 1]. \tag{2.52}$$

By this condition, the mother shearlet function is bandlimited in a diagonal band of the 2D frequency spectrum. Because the basis functions are obtained through shears and dilations of the mother shearlet function, this bandlimited property also directly controls the directional sensitivity of the basis functions. To see this, let us investigate the effect of a shear operation on the mother shearlet function. For the shear transform in (2.50), we have: [6]

$$\widehat{\psi}\left(\mathbf{B}^k\boldsymbol{\omega}\right) = \widehat{\psi}_1\left(\omega_x\right)\widehat{\psi}_2\left(k - \frac{\omega_y}{\omega_x}\right), \tag{2.53}$$

which means that a shear operation results in a shift in the argument of $\widehat{\psi}_2(\omega_y/\omega_x)$, hence the orientation of the basis function is controlled by the shear parameter k (see Figure 2.21(b)). Similarly, the anisotropic scaling leads to:

$$\widehat{\psi}\left(\mathbf{A}^i\boldsymbol{\omega}\right) = \widehat{\psi}_1\left(4^{-i}\omega_x\right)\widehat{\psi}_2\left(2^{-i}\frac{\omega_y}{\omega_x}\right). \tag{2.54}$$

We see that changing the scale parameter i results in a scaling in the argument of the wavelet $\widehat{\psi}_1$, but it also affects the support of the bump function $\widehat{\psi}_2$. More concretely, when the scale parameter is increased by 1, the bandwidth of the shearlet is halved (hence the shearlet has a finer directional selectivity).

Hence, by changing the shear and scale parameters k and i, arbitrary wedges of the frequency plane can be selected, as shown in Figure 2.21(b).

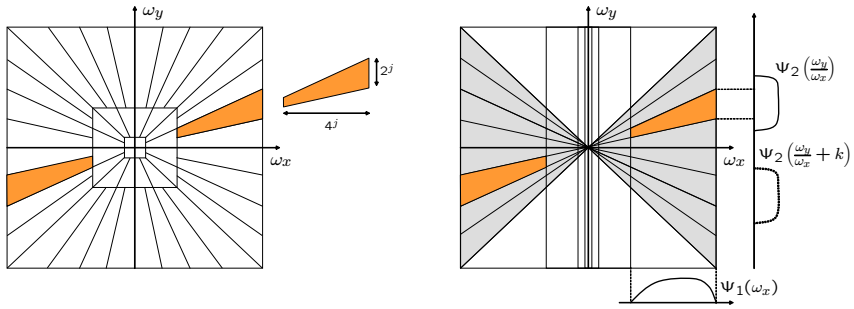
Shearlets on the cone

So far, we considered shear operations in the vertical direction and anisotropic dilation, with a larger scaling factor in the x-direction than in the y-direction. To obtain a more equal treatment of the horizontal and vertical directions, the frequency plane is split into two cones (for the high frequency band) and a square at the origin (for the low frequency band), see Figure 2.22 [Guo and Labate, 2007]:

$$\begin{aligned} C_1 &= \left\{(\omega_x, \omega_y) \in \mathbb{R}^2 |\, |\omega_x| \geq \omega_0,\, |\omega_y| \leq |\omega_x|\right\}, \\ C_2 &= \left\{(\omega_x, \omega_y) \in \mathbb{R}^2 |\, |\omega_y| \geq \omega_0,\, |\omega_y| > |\omega_x|\right\}, \\ C_3 &= \left\{(\omega_x, \omega_y) \in \mathbb{R}^2 |\, |\omega_x| < \omega_0,\, |\omega_y| < \omega_0\right\}, \end{aligned}$$

with ω_0 the maximal frequency of the the center square C_3. This square is added to be able to construct a shearlet tight frame [Guo and Labate, 2007, Easley et al., 2008]. To treat horizontal and vertical frequencies equally, in

[6] Here, we rely on the fact that the Fourier transform of a geometrically transformed function $f(\mathbf{Ap})$ is given by $|\det \mathbf{A}|^{-1}\,\mathcal{F}\{f\}\left(\mathbf{A}^{-T}\boldsymbol{\omega}\right)$, with $\mathcal{F}\{f\}$ the Fourier transform of f.



(a) (b)

Figure 2.21: (a) Frequency tiling of the shearlet transform in trapezoidal shaped tiles (wedges) [Guo and Labate, 2007]. (b) Individual components $\widehat{\psi}_1(\omega_x)$ and $\widehat{\psi}_2(\omega_y/\omega_x)$ of the Fourier transform of the shearlet mother function and the selection of orientations by the parameter k.

cone C_2, the x- and y-components for $\mathbf{p}$ need to be switched before applying geometric transforms. This comes down to using the following dilation and shear matrices in both cones:

$$\begin{aligned} \mathbf{A}_1 &= \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, \mathbf{B}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ \mathbf{A}_2 &= \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}, \mathbf{B}_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \end{aligned} \tag{2.55}$$

Consequently, the horizontal cone is dilated horizontally by a factor 4 per scale, while the vertical cone is dilated vertically by factor 4. In the following, we distinguish between both cones explicit by assigning different shearlet basis functions to each cone $d = 1, 2$:

$$\psi^{(d)}_{i,k,\mathbf{l}}(\mathbf{p}) = |\det\, \mathbf{A}_d|^{i/2}\, \psi\left(\mathbf{B}_d^k \mathbf{A}_d^i \mathbf{p} - \mathbf{l}\right). \tag{2.56}$$

Analogously to the DWT (Section 2.1), it is natural to discretize the scale, orientation and position indices. In the remainder, we will therefore restrict $i, k, \mathbf{l}$ to discrete (integer) values. The resulting frequency tiling is illustrated in Figure 2.21(a).

Tight frames of shearlets

Next, we want to represent an arbitrary function $f \in L^2\left(\mathbb{R}^2\right)$ by a set of projections of this function onto the shearlet basis elements, $\left\langle f, \psi^{(d)}_{i,k,\mathbf{l}} \right\rangle$. Unfortunately, it is still an open question whether it is possible to design *bases* of shearlet functions. A shearlet basis would give a complete (non-redundant) representation for e.g. images. Instead, we rely on frame theory [Daubechies,

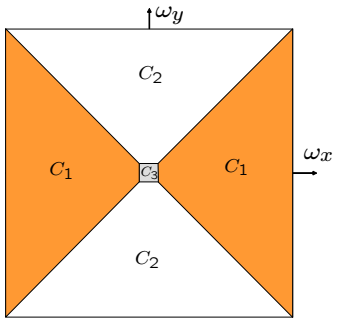


Figure 2.22: Partitioning of the 2D frequency plane into two cones (C_1 and C_2) and a square (C_3) at the origin.

1992], which has been developed as a general theory for overcomplete (redundant) representations. The family of functions

$$\left\{\psi^{(1)}_{i,k,\mathbf{l}}(\mathbf{p}), \psi^{(2)}_{i,k,\mathbf{l}}(\mathbf{p}) | i \in \mathbb{Z},\, k \in \mathbb{Z},\, \mathbf{l} \in \mathbb{Z}^2, i \geq 0\right\} \tag{2.57}$$

is called *frame* if there exists positive constants A and B, such that for all $f \in L^2\left(\mathbb{R}^2\right)$:

$$A \left\|f\right\|_2^2 \leq \sum_{i,k,\mathbf{l},d} \left|\left\langle f, \psi^{(d)}_{i,k,\mathbf{l}} \right\rangle\right|^2 \leq B \left\|f\right\|_2^2, \tag{2.58}$$

where A and B are frame bounds. The analysis of the frame bounds of often useful to investigate the numerical stability of a multiresolution transform. Additionally, a frame is called *tight frame* (or Parseval frame) if the frame bounds are equal ($A = B$). By proper normalization, the frame bound can be chosen to be equal to 1, such that the Parseval relationship holds:

$$\sum_{i,k,\mathbf{l},d} \left|\left\langle f, \psi^{(d)}_{i,k,\mathbf{l}} \right\rangle\right|^2 = \left\|f\right\|_2^2. \tag{2.59}$$

The Parseval relationship implies that any function in $L^2\left(\mathbb{R}^2\right)$ can be expanded into a set of functions [Daubechies, 1992]:

$$f = \sum_{i,k,\mathbf{l},d} \left\langle f, \psi^{(d)}_{i,k,\mathbf{l}} \right\rangle \psi^{(d)}_{i,k,\mathbf{l}}. \tag{2.60}$$

This equation is in fact similar to (2.3), the difference here is that the set of functions (2.57) do not form a basis but a frame. For the shearlet functions,

equation (2.59) imposes specific constraints to the functions $\widehat{\psi}_1(\omega)$ and $\widehat{\psi}_2(\omega)$:

$$\sum_{i\geq 0}\left|\widehat{\psi}_1(4^{-i}\omega)\right|^2 = 1 \quad \text{for } |\omega| \geq \omega_0 \tag{2.61}$$

$$\sum_{k=-2^i}^{2^i-1}\left|\widehat{\psi}_2(2^i\omega - k)\right|^2 = 1 \quad \text{for } |\omega| \leq \pi \tag{2.62}$$

which practically means that the sum of the energies of $\widehat{\psi}_1(\omega)$ and $\widehat{\psi}_2(\omega)$ for respectively scaled frequencies and shifted frequencies must be one. The Parseval relationship holds for the part of the frequency plane that excludes the center square (see Figure 2.22), although this can be trivially extended to the complete frequency plane by adding a bandlimited scaling function [Easley et al., 2008]:

$$\widehat{\phi}(\omega_x, \omega_y) = \begin{cases} \tilde{\Phi}(\omega_x) & |\omega_y| \leq |\omega_x| < \omega_0 \\ \tilde{\Phi}(\omega_y) & |\omega_x| \leq |\omega_y| < \omega_0 \\ 0 & \text{else} \end{cases} \tag{2.63}$$

with $\tilde{\Phi}(\omega)$ a 1D scaling function that satisfies:

$$\sum_{i\geq 0}\left|\widehat{\psi}_1(4^{-i}\omega)\right|^2 + \left|\tilde{\Phi}(\omega)\right|^2 = 1 \quad \text{for } |\omega| < \omega_0 \tag{2.64}$$

By comparing equation (2.49) to equation (2.63), it can be noted that the scaling function is more or less isotropic. This behavior resembles the isotropy of the scaling functions in the DWT, with the main difference that in the DWT, 2D scaling functions are formed by a tensor products of one-dimensional scaling functions, instead of being defined per cone.

Shearlets or curvelets?

Shearlets are very similar to curvelets in the sense that both perform a multiscale and multidirectional analysis, and both transforms obey the parabolic scaling property. Both transforms have very similar asymptotic approximation properties: for images $f(\mathbf{p})$ that are C^2 everywhere except near edges, where $f(\mathbf{p})$ is piecewise C^2, the approximation error of a reconstruction with the N-largest coefficients ($f_N(\mathbf{p})$) in the shearlet/curvelet expansion is given by [Candès et al., 2006, Guo and Labate, 2007]:

$$\|f - f_N\|_2^2 \leq B \cdot N^{-2}(\log N)^3, \quad N \to \infty$$

with B a constant. Because this is the optimal approximation rate for this type of functions [Guo and Labate, 2007], this property is often referred to as *optimal sparsity*. Still, there are a number of differences between shearlets and curvelets [Easley et al., 2008]:

- Shearlets are generated by applying a family of operators to a single function, while curvelet basis elements are not in the form of equation (2.49).
- Shearlets are associated to a fixed translation lattice, while curvelets are not. This is of importance for applications: when combining information from multiple scales and orientations (e.g. to model inter- or intrascale dependencies, see Chapter 3), curvelet techniques need to take into account that the translation lattice is not fixed.
- In the construction of the shearlet tight frame above, the number of orientations doubles at every scale, while in the curvelet frame, this number doubles at every *other* scale.
- Shearlets are associated to a multiresolution analysis, while curvelets are not.

Perhaps the most primary advantage that we want to point out, is that shearlets allow for a much less redundant sparse tight frame representation, while offering shift invariance.

2.5.2 New design of the discrete shearlet transform

In analogy to the fast DWT (see Section 2.1) for computing the DWT from an image, it is desirable to have an efficient decomposition and synthesis scheme for shearlets as well. Recently, a number of possible designs of the discrete shearlet transform (DST) have been proposed. A first approach, used e.g. in [Easley et al., 2009], is to apply a direct discretization to the shearlet basis functions $\psi_{i,k,\mathbf{l}}^{(d)}(\mathbf{p})$, which leads to shearlet filters implemented in Fourier space. Although such a scheme is simple and shift-invariant, the redundancy factor is high due to the lack of downsampling operations in the decomposition. More specifically, the redundancy factor is:

$$1 + \sum_{i=1}^{I} 2^{i+1} = 2^{I+2} - 3,$$

when choosing 2^{I+1} orientations for the first scale. For example, using 3 scales gives redundancy factor 29. In analogy to the undecimated DWT, we will call this transform the *undecimated DST* because of the lack of decimations.

[Easley et al., 2008] propose a discrete implementation with one of the main applications in image denoising. In their work, a Laplacian pyramid is followed by windowing filters in the Pseudo-Polar DFT domain. By including decimations in the Laplacian pyramid, the redundancy of the transform is reduced. Because the redundancy factor per scale of the transform increases linearly with the number of orientations for that scale, the overall redundancy factor is still high. We remark that the Laplacian pyramid representation of [Burt and Adelson, 1983] that is used in [Easley et al., 2008] is not a tight

frame in its standard form, however, a tight frame can be constructed by using orthogonal pyramid filters [Do and Vetterli, 2001].

[Yi et al., 2009] outline a different implementation for edge detection and analysis. In their implementation, there is an explicit distinction between horizontal and vertical shearlets. However, the authors do not take further steps to reduce the redundancy, as they choose to stay faithful to the CST in terms of edge analysis. Further, it is not clear which cascade algorithm would be the best to do the inverse transform of this scheme, as no reconstruction algorithm is proposed.

Next, we will present a design that can be used in a wide range of applications and that has a lower redundancy factor than the above implementations of the DST. As we explained in Section 2.5.1, for the CST there is an explicit separation of the horizontal cone C_1 and the vertical cone C_2. An obvious discrete realization would be to use hourglass-shaped filters. We prefer not to do this, as this either increases the redundancy factor by 2, or causes angular aliasing when including decimations in the angular filtering.[7] The presence of angular aliasing is very cumbersome in practical applications as it severely degrades the directional selectivity of the basis functions. Instead, we apply only one angular filtering stage at each scale to directly split up all orientation subbands, which also has the advantage that the corresponding filterbank is conceptually more clean.

To proceed, we will define shearlet filters in pseudo-polar frequency coordinates (FC). Every shearlet filter will extract a wedge-shaped region of the 2D frequency plane; these wedge filters can be easily described in pseudo-polar FC.

Pseudo-polar coordinate system

We use FC (ω_r, ϑ) in a pseudo-polar grid [Averbuch et al., 2006] that is consistent to a polar grid, in the sense that the pseudo-angle is in the range $\vartheta \in [-\pi, \pi]$. The corresponding conversion from Cartesian coordinates (ω_x, ω_y) to pseudo-polar FC (ω_r, ϑ) is given by:

$$\omega_r(\omega_x, \omega_y) = \sqrt{\frac{1 + \max\left(|\omega_x|^2, |\omega_y|^2\right)}{1 + \pi^{-2}}} \tag{2.65}$$

$$\vartheta(\omega_x, \omega_y) = \begin{cases} \frac{\pi}{4}\left(\frac{\omega_y}{\omega_x}\right), & \text{if } |\omega_x| > |\omega_y| \text{ and } \omega_x \geq 0, \\ \frac{\pi}{4}\left(2 - \frac{\omega_x}{\omega_y}\right), & \text{if } |\omega_x| < |\omega_y| \text{ and } \omega_y \geq 0, \\ \frac{\pi}{4}\left(4 + \frac{\omega_y}{\omega_x}\right), & \text{if } |\omega_x| > \omega_y \geq 0 \text{ and } \omega_x < 0, \\ \frac{\pi}{4}\left(-4 + \frac{\omega_y}{\omega_x}\right), & \text{if } |\omega_x| > \text{-}\omega_y > 0 \text{ and } \omega_x < 0, \\ \frac{\pi}{4}\left(-2 - \frac{\omega_x}{\omega_y}\right) & \text{else,} \end{cases} \tag{2.66}$$

[7] Such an angular filterbank with decimation is used in e.g. the contourlet transform [Do and Vetterli, 2005].

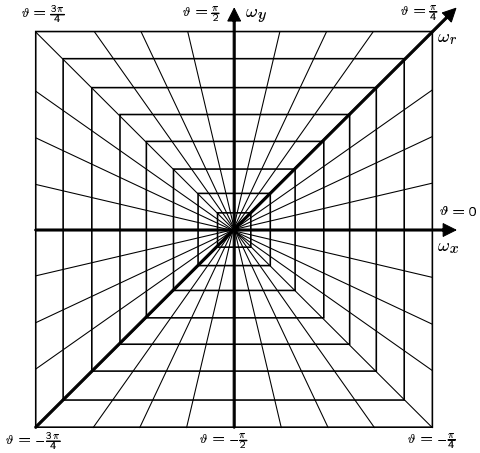


Figure 2.23: Pseudo-polar coordinate system.

where we replace the fractions ω_x/ω_y and ω_y/ω_x by 0 whenever the denominator becomes 0. The denominator in (2.65) has been chosen such that $\omega_r(\pm\pi, \pm\pi) = \pi$. Important to note is that adding a constant to the pseudo-angle ϑ corresponds to a vertical shear transform if both $(\omega_x, \omega_y) \in C_1$ and the transformed point also belong to C_1. Equivalently, if $(\omega_x, \omega_y) \in C_2$, adding a constant to ϑ corresponds to a horizontal shear transform if the transformed point also belongs to C_2. Transiting from C_1 to C_2 (or vice versa) can be done using a cascade of a horizontal and vertical shear transform. The pseudopolar grid defined above is illustrated in Figure 2.23. It can be noted that contours of equal pseudo-radial frequencies ω_r define concentric squares around the origin, instead of circles as is the case with a polar grid.

Filter bank

The filter bank design that we propose is mostly related to the design of the steerable pyramid transform (see Section 2.3.3), in the sense that we use a frequency partitioning into $K_i \geq 2$ orientation subbands and a low-pass subband at each scale. In our scheme, the specific way of decimating the horizontal and vertical orientation subbands, is different, as well as the pseudo-polar grid for defining the filters and the filters being used. This will allow us to further subsample the orientation subbands.

At each scale of our DST, we use a $(K_i + 1)$-band recursive decomposition into K_i (band-pass or high-pass) orientation bands and a low-pass band. Let us denote the scaling analysis filters as $H(\boldsymbol{\omega})$ and the shearlet analysis filters as $G_k(\boldsymbol{\omega})$, with $k = 1, ..., K_i$ the index of the orientation band. The scaling synthesis filters and shearlet synthesis filters are $\tilde{H}(\boldsymbol{\omega})$ and $\tilde{G}_k(\boldsymbol{\omega})$, respectively. The analysis and synthesis filter bank is shown in Figure 2.24. In our filter

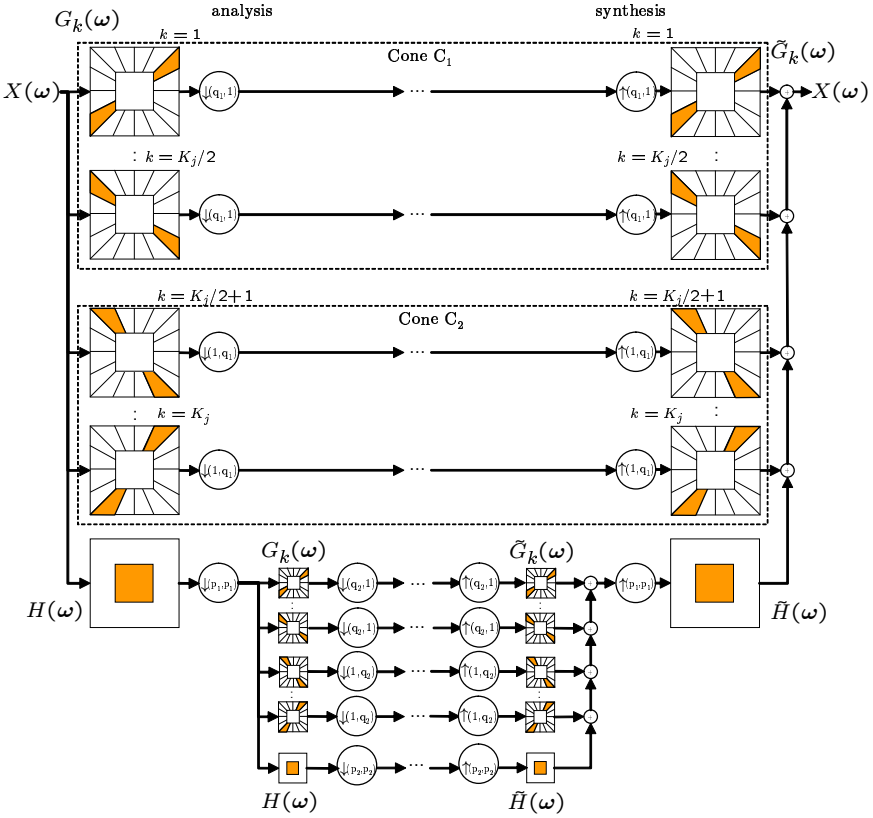


Figure 2.24: Shearlet analysis and synthesis filterbank.

bank, the image is first filtered using the oriented shearlet filters $G_k(\boldsymbol{\omega})$, subsequently the result is decimated with a scale-dependent factor q_i, in the direction orthogonal to the main filter orientation (i.e. horizontal or vertical). Next, the filter bank is iterated on the decimated output of the scaling filters, where the decimation factor for the scaling step i is denoted by p_i. The synthesis filter bank is entirely analogous, hence when designing the filters appropriately, the filter bank can be made to be self-inverting.

To design the filters, we express the perfect reconstruction equations for Figure 2.24, and try to find filters that satisfy these equations. The first perfect reconstruction (PR) condition for this filter bank is given by:

$$H(\boldsymbol{\omega})\tilde{H}(\boldsymbol{\omega}) + \sum_{k=1}^{K_i} G_k(\boldsymbol{\omega})\tilde{G}_k(\boldsymbol{\omega}) = 1,\ \boldsymbol{\omega} \in \Omega \tag{2.67}$$

with $\Omega = [-\pi, \pi] \times [-\pi, \pi]$. For a decimated transform, other PR conditions are needed to state that the aliasing caused by the downsampling operations should

cancel itself. We will call these conditions the *aliasing canceling* conditions. Note that for some combinations of (p, q) PR is not even possible (e.g. $(p, q) = (4, 2)$), in that case, the PR conditions are conflicting.

We investigate $p = 4$ and $q = 1$ with the anisotropic dilation matrices from (2.55). The aliasing canceling PR conditions are:

$$H\left(\omega_x, \omega_y\right) \tilde{H}\left(\omega_x + \frac{m\pi}{2}, \omega_y + \frac{n\pi}{2}\right) = 0, \tag{2.68}$$

with $m = 0, ..., 3$, $n = 0, ..., 3$ and $(m, n) \neq 0$. Notably, (2.68) only affects the scaling filter and not the shearlet filters: the scaling filter $H(\boldsymbol{\omega})$ must have frequency support $[-\frac{\pi}{4}, \frac{\pi}{4}] \times [-\frac{\pi}{4}, \frac{\pi}{4}]$. Consequently, $H(\boldsymbol{\omega})$ cannot have compact support in spatial domain (see [Daubechies, 1992]). Nevertheless, because of the lack of aliasing, the transform can be made to be *shift-invariant*, in a similar way as done for the steerable pyramid transform. To do so, we define the filters in separable pseudo-polar frequency coordinates:

$$\begin{aligned}
H(\omega_r, \vartheta) &= H_0(\omega_r), \\
G_k(\omega_r, \vartheta) &= G_0(\omega_r) \sum_{i=-\infty}^{+\infty} R\left(\frac{(\vartheta + i\pi)K_i}{\pi} - k + 1\right), \\
\tilde{H}(\omega_r, \vartheta) &= \tilde{H}_0(\omega_r), \\
\tilde{G}_k(\omega_r, \vartheta) &= \tilde{G}_0(\omega_r) \sum_{i=-\infty}^{\infty} \tilde{R}\left(\frac{(\vartheta + i\pi)K_i}{\pi} - k + 1\right)
\end{aligned} \tag{2.69}$$

with $H_0(\omega_r)$ the frequency response of a 1D scaling filter, $G_0(\omega_r)$ the frequency response of a 1D wavelet filter and $R(\vartheta)$ a real-valued compactly supported bump function. In (2.69), the bump function is periodized in ϑ with period π to construct filters with real-valued impulse responses. In practice we can assume that $\vartheta \in [-\pi, \pi]$, by the construction of the pseudo-polar grid. Consequently, the summation in (2.69) only needs to iterate over a finite number of values for i.

Using the above filters, the PR conditions come down to:

$$H_0(\omega_r)\tilde{H}_0(\omega_r) + G_0(\omega_r)\tilde{G}_0(\omega_r) = 1, \quad \omega_r \in [-\pi, \pi]$$

$$\sum_{k=1}^{K} \sum_{i=-\infty}^{+\infty} R\left(\frac{(\vartheta + i\pi)K_i}{2\pi} - k + 1\right) \cdot \tilde{R}\left(\frac{(\vartheta + i\pi)K_i}{2\pi} - k + 1\right) = 1, \quad \vartheta \in [-\pi, \pi] \tag{2.70}$$

$$H_0(\omega_r) = \tilde{H}_0(\omega_r) = 0, \quad |\omega_r| > \frac{\pi}{4} \tag{2.71}$$

In Figure 2.26(a), examples of radial filters satisfying these equations are shown. It can be seen that the scaling filter has a band center frequency $\sim \pi/8$, as a result the frequency resolution of this DST may be rather poor (for the second

scale i	p_i	q_i
1	2	1
2	4	1
3	4	2
4	4	4
5	4	4

Table 2.3: Proposed decimation factors for the DST with anisotropic dilation (also see Figure 2.24).

scale, this frequency becomes $\sim \pi/32$; hence much of the frequency content of the image is contained in the first scale). Therefore, we replace (2.71) by a less strong condition:

$$H_0(\omega_r) = \tilde{H}_0(\omega_r) = 0, \quad |\omega_r| > \frac{\pi}{2}$$

and modify the decimation operations appropriately, such that there is no information loss and hence PR is still possible. For the first scale, we set $p_1 = 2$, $q_1 = 1$ and starting from the second scale ($i > 1$), we use $p_i = 4$ and $q_i = 2$. In Table 2.3, the decimation factors p_i and q_i are listed per scale. The modified Meyer wavelet and scaling filters with adjusted frequency scaling is shown in Figure 2.26(b). The use of the Meyer wavelet here is an appealing choice due to its excellent localization properties in both time and frequency and also because the filters are defined directly in frequency domain [Daubechies, 1992]:

$$H_0(\omega_r) = \begin{cases} 1 & |\omega_r| < \frac{\pi}{4} \\ \cos\left(\frac{\pi}{2} v\left(\frac{4|\omega|}{\pi} - 1\right)\right) & \frac{\pi}{4} \le |\omega_r| \le \frac{\pi}{2} \\ 0 & \text{else} \end{cases},$$

$$G_0(\omega_r) = \begin{cases} 0 & |\omega_r| < \frac{\pi}{4} \\ \sin\left(\frac{\pi}{2} v\left(\frac{4|\omega|}{\pi} - 1\right)\right) & \frac{\pi}{4} \le |\omega_r| \le \frac{\pi}{2} \\ 1 & \text{else} \end{cases},$$

$$\tilde{H}_0(\omega_r) = H_0(\omega_r),$$

$$\tilde{G}_0(\omega_r) = G_0(\omega_r).$$

where the interpolation function $v(x)$ satisfies $v(x) = 1 - v(1-x)$ [Daubechies, 1992]. Here, we use a third order polynomial for the range $x \in [0, 1]$ (see Figure 2.25):

$$v(x) = \begin{cases} 3x^2 - 2x^3 & 0 \le x \le 1 \\ 0 & x < 0 \\ 1 & 1 < x \end{cases} \tag{2.72}$$

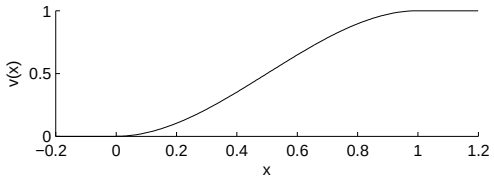


Figure 2.25: The interpolation function $v(x)$.

Similarly, angular filters satisfying (2.71) are given by:

$$R(x) = \tilde{R}(x) = \begin{cases} 0 & x < -\frac{1+\alpha}{2} \\ \sin\left(\frac{\pi}{2} v\left(\frac{\alpha+2x+1}{2\alpha}\right)\right) & \left|x+\frac{1}{2}\right| \leq \frac{\alpha}{2} \\ 1 & |x| < \frac{1-\alpha}{2} \\ \cos\left(\frac{\pi}{2} v\left(\frac{\alpha+2x-1}{2\alpha}\right)\right) & \left|x-\frac{1}{2}\right| \leq \frac{\alpha}{2} \\ 0 & \text{else} \end{cases}$$

with $\alpha \in \left[0, \frac{1}{2}\right]$ a constant parameter that determines the bandwidth of the angular filters. In Figure 2.26(c), $R(x)$ is depicted for different values of α. Higher values of α correspond to a slower decay of the transition bandwidth. The corresponding filters $\tilde{G}_k(\omega_r, \vartheta)$ for $\omega_r = \frac{\pi}{2}$ and $\alpha = \frac{1}{2}$ are shown in Figure 2.26(d).

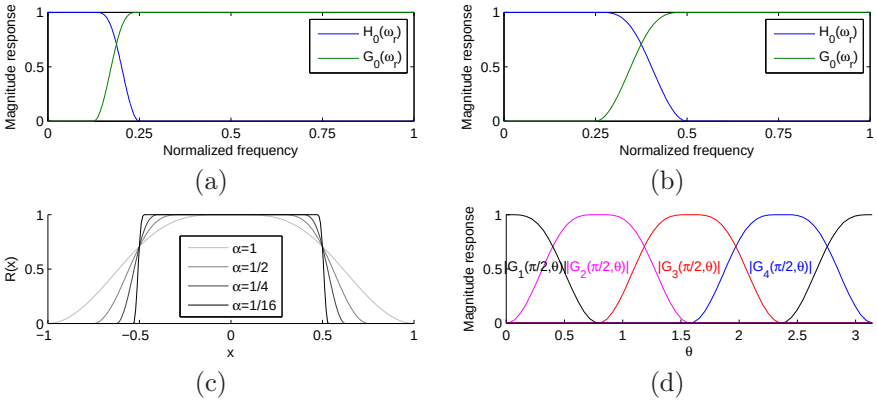


Figure 2.26: (a) Shearlet radial magnitude responses for dilation factor 4 (using the Meyer wavelet), (b) Shearlet filter radial magnitude responses with proposed adjustment to increase the low-pass center band frequency (using the Meyer wavelet), (c) Angular response $R(x)$, (d) Shearlet filter magnitude responses for the constant radial frequency $\omega_r = \pi/2$.

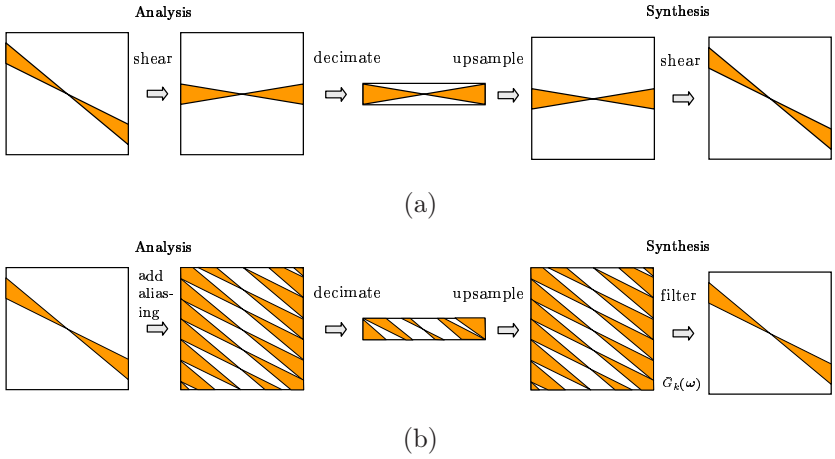


Figure 2.27: Strategies to reduce the redundancy of the DST. (a) Perfect reconstruction by shear operations and decimating *(Folding)*, (b) Perfect reconstruction by decimating without shearing *(Non-folding)*. See text also.

Folding and angular decimation

By the compact support of $R(x)$, the filters $G_k(\omega_r, \vartheta)$ are supported on trapezoidal wedges in the frequency plane. In case of more than two orientations ($K > 2$), we can partially get rid of the extra redundancy in two different ways (see Figure 2.27):

1. (*Folding*) The filtered subbands can be sheared such that the frequency support is fully contained in the central rectangles as shown in Figure 2.27(a). Subsequently, a vertical decimation can be applied to the subbands in cone C_1 and a horizontal decimation to the subbands in cone C_2. Note that a suitable (possibly non-integer) decimation factor needs to be chosen, we will explain this further on. For the shear transform, we rely on bandlimited interpolation (most efficiently implemented in the DFT domain). For the exact details of the shear transform implementation, we refer to [Condat et al., 2008].

2. (*Non-folding*) On the other hand, perfect reconstruction is possible even without folding. Therefore we need to make sure that the aliasing caused by the decimations does not contaminate the content of the wedges of interest. This can again be done by chosing the decimation factor suitably (actually the same as in the folding strategy). This approach is illustrated in Figure 2.27(b). Even though many aliasing copies are produced during decimation, the original wedges can be perfectly reconstructed after applying the reconstruction filter $\tilde{G}_k(\omega_r, \vartheta)$.

Both schemes have the same redundancy factor. The difference is that in the *folding* strategy, the translation lattice is sheared, while without folding, the

translation lattice remains Cartesian, which can be an advantage in certain applications. Additionally, the *non-folding* strategy heavily relies on aliasing and it is easy to show that the non-folding strategy is *not* shift-invariant, whereas the folding strategy *is* shift-invariant.

For a squared subband at scale i of size N_i, we compute the decimation factor from Figure 2.27 as follows:

$$d_i = \max\left(1, \frac{N_i}{\lceil (1+2\alpha)N_i / (K_i/2) \rceil}\right) \tag{2.73}$$

In our implementation, the folding is performed in the DFT domain; N_i/d_i then determines the integer number of DFT coefficients to keep per row or column. For this reason a ceiling function is present in (2.73).

More importantly, because we have K_i orientation subbands per scale, we see that the redundancy for scale i of the transform, which is proportional to $K_i/d_i \approx 2(1+2\alpha)$, becomes independent of K_i! [8]

Because the filters are defined in frequency domain and because the shearing operations are based on DFT transforms, our current implementation of this filterbank makes use of FFTs. Because the filters $H_0(\omega_r)$, $G_0(\omega_r)$, $\tilde{H}_0(\omega_r)$ and $\tilde{G}_0(\omega_r)$ are bandlimited, the filters do not have compact support in spatial domain. Nevertheless, it is possible to approximate the impulse response by truncation, as proposed e.g. in [Castleman et al., 1998] for the steerable pyramid filters.

Based on the filter bank scheme from Figure 2.24 and equation (2.73), a recursive formula can be written for the redundancy factor of our scheme:

$$\begin{aligned} R &\approx \left[\frac{2}{q_1} + \frac{1}{p_1^2}\left(\frac{2}{q_2} + \frac{1}{p_2^2}\left(\frac{2}{q_3} + \frac{1}{p_3^2}\left(\frac{2}{q_4} + \cdots\right)\right)\right)\right](1+2\alpha) + 2^{-2\sum_{i=1}^{I} p_i} \\ &\approx \left[2 + \frac{1}{4}\left(2 + \frac{1}{16}(1+...)\right)\right](1+2\alpha) + 2^{-2\sum_{i=1}^{I} p_i}. \end{aligned} \tag{2.74}$$

with p_i and q_i as listed in Table 2.3. In Table 2.4, redundancy factors of the transform are given with respect to the number of scales I and the parameter α.

In Figure 2.28, shearlet basis functions are shown for different scales and orientations. Even though the size of the support of these basis functions is not finite, these functions have a fast decay and are well localized in space, frequency and orientation.

As an example, we investigate the approximation quality of different multiresolution transforms. We start from a test image, apply a given wavelet or shearlet transform to this image and we reconstruct the image from the 2.5% largest wavelet or shearlet coefficients (in magnitude). In Figure 2.29, the results are given for the zone plate image and the barbara image, for the decimated DWT, the undecimated DWT, the DT-CWT and the DST. We also list the redundancy ratios for each transform in the figure, because this ratio

[8] Up to small deviations caused by the ceiling operation, but this is usually neglectible.

number of scales I	α		
	$\frac{1}{32}$	$\frac{1}{8}$	$\frac{1}{2}$
1	2.19	2.56	4.06
2	2.66	3.13	5.00
3	2.67	3.14	5.03
4	2.67	3.15	5.03
5	2.67	3.15	5.03

Table 2.4: Redundancy factors for I scales, computed using (2.74).

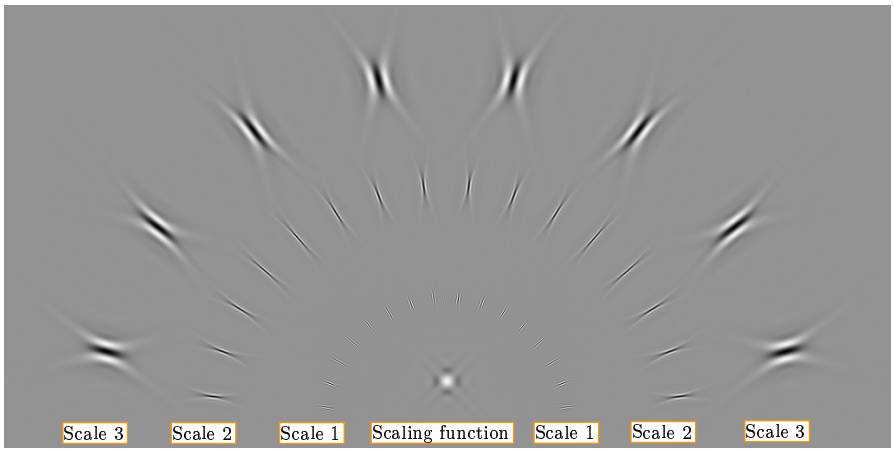
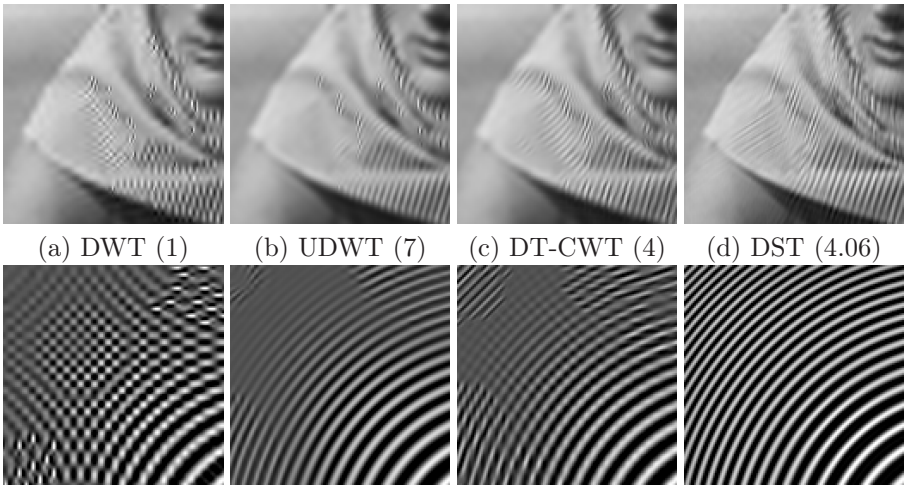


Figure 2.28: Shearlets basis elements for $\alpha = \frac{\pi}{2}$. For illustration purposes, we used $K_1 = 16$ (instead of 32) orientations for the first scale.

plays a big role here. The DWT is a *shift variant* transform, and the aliasing creates *disturbing artifacts* in the end result. The undecimated DWT has the largest number of coefficients retained in absolute terms, however, the basis functions of this transform corresponding to the HH_i subbands have a poor directional selectivity, which causes here the blurring of some of the edges. The DT-CWT basis functions have an excellent spatial localization, but are only able to distinguish 6 orientations, also causing a fair amount of blurring here (see Figure 2.29(c)/(g)). The DST gives here the best visual result, mainly because of its excellent directional selectivity and shift-invariance.

2.6 Conclusion

In this chapter, we presented a number of multiresolution representations for images. The discrete wavelet transform (DWT) offers a multiresolution analysis by successively approximating the original image with a coarser version of this image, thereby extracting detail information. Unfortunately, there are a number of fundamental problems with the DWT: most notably are shift vari-



(a) DWT (1) (b) UDWT (7) (c) DT-CWT (4) (d) DST (4.06)

(e) DWT (1) (f) UDWT (7) (g) DT-CWT (4) (h) DST (4.06)

Figure 2.29: Reconstruction from 2.5% of the x-let coefficients, for (a),(e) DWT with 2 scales, (b),(f) Undecimated DWT with 2 scales, (c),(g) DT-CWT with 2 scales, (d),(h) DST with 1 scale (because of anisotropic dilation with factor 4). *Top row*: crop out of the Barbara image, *bottom row*: crop out of the zone plate image. Between parentheses is the redundancy factor of each transform in this experiment.

ance, aliasing and the poor representation of directional features in images, such as edges. The dual-tree complex wavelet transform brings an appealing solution here, however, in the current discrete implementation of the transform, the analyticity properties of the complex basis functions for the first scale of the transform are rather poor, which makes the first scale of the transform not better than an undecimated version of the DWT. As a solution, we presented a specific filter design technique for improving the first scale complex wavelet filters, without affecting the subsequent scales. This gives a vast improvement in directional selectivity properties, which is benificial for many applications.

Further, we explained the steerable pyramid transform, which offers a "steerable" or rotational invariant representation for images. Steerability properties will be of use later, in Chapter 6 and Chapter 7. While the steerable pyramid transform in its standard version, considers a fixed number of analysis orientations per scale, in the shearlet transform, the number of orientations doubles at every scale. This allows for an optimally sparse representation for images containing edges with bounded curvatures. Furthermore, the shearlet transform allows for a much less redundant sparse representation, compared to its predecessors. To achieve this, we presented a novel discrete implementation of the shearlet transform that makes use of digital shearing operations and subsequent "angular" decimation. The redundancy factor of the transform thereby becomes independent of the number of analysis orientations and can be as low as ~ 2.66, while still offering shift invariance. This design makes

the shearlet transform an effective candidate representation for many practical image processing applications.

The contributions of this chapter so far have led to two publications in proceedings of international conferences [Goossens et al., 2009b,Goossens et al., 2009a]. One journal paper is in preparation [Goossens et al., 2010b].

3

Statistical models for images

Many digital imaging applications, such as image analysis, restoration and compression, require a mathematical representation of images. For example, to restore a degraded image, it is very useful to know how typical hypothetical ideal images "look like", i.e. which features they exhibit. This kind of information can be encoded in the form of pre-knowledge in the application. Prior knowledge can be obtained from a large class of images, by a learning process where patterns and tendencies in images are being isolated and extracted. The development of statistical models for images is then a result of the efforts to explain patterns found in natural images.

To understand more about the statistics of images, we first make a few observations:

- The class of natural images that we encounter in daily life is only a very small subset of the set of all possible images. Some authors call this subset a natural image *manifold* [Srivastava et al., 2003].
- All images, are *not equally likely* to occur. Digital imaging applications can hence benefit by concentrating on classes of images that occur most frequently.

Although the image formation can be seen a consequence of exact physical processes corresponding with deterministic mathematical models (e.g. based on geometry, material properties, lighting models, ...), working with these kind of models is often overly complicated and not very practical.[1] Instead, there has been a lot of interest over several last decades in the statistical modeling of images, where images are treated as the realization of a random process. Statistical models are built to explain image variability that occurs in observed images. Usually the models are designed to be practical in terms of computations.

[1] Moreover it becomes increasingly difficult to use these models for images containing degradations, such as noise, blur, ...

A first possible approach toward statistical image modeling is to isolate the image manifolds (e.g. by using a multiresolution representation as introduced in Chapter 2). Subsequently statistical properties are investigated and statistical distributions (e.g. univariate Gaussian, generalized Laplace distributions) are fitted to the manifold. Many wavelet-based techniques belong to this category (e.g. [Chang et al., 1998,Portilla et al., 2003,Fadili and Boubchir, 2005,Pižurica and Philips, 2006, Selesnick, 2008]), and some of them will be briefly outlined later in this chapter.

Another approach is to impose a probability density function (PDF) to the image space and to discover the manifold and to assign most of the mass of the PDF to it [Srivastava et al., 2003]. To these class belong techniques such as density estimation [Comaniciu and Meer, 2002], local linear embedding [Roweis et al., 2002a], probabilistic PCA [Tipping and Bishop, 1999] and Markov Random Field models [Geman and Geman, 1984, Li, 1995].

In this chapter, we will first review existing image decomposition schemes (Section 3.1). In Section 3.2 we discuss a number of parametric densities for modeling the marginal histograms of multiresolution transform subbands. In Section 3.3, we will investigate the joint statistics of subband coefficients. Statistical models for intra-scale and inter-scale dependencies will be presented in Section 3.4 and Section 3.5, respectively. In particular, we introduce our new intra-scale model MP-GSM in Section 3.4.4 and we present our novel joint inter/intra-scale model in Section 3.6. Finally, we discuss the use of non-local image models in Section 3.7.

3.1 Decomposition of images

As a starting point, we consider the problem of decomposing an image in a set of independent elements (or building blocks). In general, this problem is very difficult because we have to deal with the high dimensionality of the data, for example when computating joint histograms of images. To circumvent this problem, additional assumptions are usually made.

3.1.1 Classical image model

In classical image analysis, the image is modeled as a second order statistical process. This involves taking mean and correlations of the input image into account. A restriction that is often made is to consider only linear decompositions. One solution to find uncorrelated components is given by principal component analysis (PCA) [Hotelling, 1933, Anderson, 1963, Jolliffe, 1986], also known as the Karhunen-Loèvre Transform (KLT). PCA computes the eigenvectors of the sample covariance matrix. The eigenvectors corresponding to the largest eigenvalues are called principal components. These are a set of orthogonal axes along which the components are decorrelated. In case the im-

age statistics are Gaussian[2], the PCA solution also gives components that are *statistically independent*.

Often, an additional assumption is made that the statistical properties of the image are translation invariant (also called *spatial stationarity*). For second order processes this means that the correlation between two pixel intensities in the image only depends on the difference between the positions, and not on the absolute positions of the pixels. In that case, the sample covariance matrix becomes *circulant*, and possible principal component axes are the basis vectors of the Fourier transform. Let us denote by $R(\mathbf{p}, \mathbf{q})$ the autocorrelation function of a signal in the spatial domain, where $\mathbf{p}$, $\mathbf{q}$ are two-dimensional vectors representing the spatial position in the image. Then the correlation between two pixel intensities depends only on the difference between their positions if:

$$R(\mathbf{p}, \mathbf{q}) = R(\mathbf{0}, \mathbf{q} - \mathbf{p}). \tag{3.1}$$

The power spectral density (PSD) describes how the energy (or variance) of an image is distributed in frequency space and according to the Wiener−Khintchine theorem [Baher, 2001], the PSD is obtained as Fourier transform of the autocorrelation function $R(\mathbf{0}, \mathbf{q})$:

$$P(\boldsymbol{\omega}) = \int_{\mathbb{R}^2} R(0, \mathbf{q}) \exp\left(-j\boldsymbol{\omega}^T \mathbf{q}\right) \mathrm{d}\mathbf{q} \tag{3.2}$$

If the image statistics are Gaussian, images can be completely represented by their Fourier coefficients, which are uncorrelated, and statistical models can be build to allow for some randomization in the Fourier coefficients. A number of studies have indicated that the PSD for natural images obeys a *power law* property [Field, 1987, Srivastava et al., 2003] : the power decays as:

$$P(\boldsymbol{\omega}) \propto \frac{1}{\|\boldsymbol{\omega}\|^{2-\eta}} \tag{3.3}$$

where $\|\boldsymbol{\omega}\|$ is the magnitude of the spatial frequency and η is a constant that varies with the image type but that is usually small [Mumford and Gidas, 2001]. According to the power law, the spectra of images show highest amplitudes for low frequencies and the amplitude decreases as the frequency increases. This characteristic is very different from white noise, which has a flat spectrum ($P(\boldsymbol{\omega}) \propto 1$).

In Figure 3.1, the PSD is computed for a set of 12 standard test images. Figure 3.1(a) shows iso-power contours of the average PSD over the images. As a rough approximation, the PSD can be considered to be radially symmetric, with a slightly higher power concentration near the frequency coordinate axes. In Figure 3.1(b), the PSD of these images is shown after averaging over all orientations. It can be seen that in all of these images, the power law is applicable.

[2]Note that this assumption is too simplistic for most applications, as we will see in the remainder of this chapter.

In the past, researchers have searched for a plausible explanation for the $1/\omega^2$ law and this has been topic of speculation and debate [Simoncelli and Olshausen, 2001]. One of the most common beliefs is that it is due to the scale invariance of the visual world [Ruderman, 1997, Zhu et al., 1997]. Scale invariance means that the statistical properties do not depend on the scale at which the observation was made. For example, let us compute the variance of an image in a one octave frequency band $[\omega_0, 2\omega_0]$. The central band frequency $1.5\omega_0$ determines the zoom factor at which the observation was made. The variance in this band can be computed by integrating the PSD:

$$\sigma^2 = \int_{\omega_0}^{2\omega_0} P(\omega)\mathrm{d}\omega = \int_1^2 P\left(\omega_0\omega\right)\omega_0\mathrm{d}\omega \tag{3.4}$$

Upon a constant factor, by the power law, (3.4) amounts to $\int_1^2 P\left(\omega\right)\mathrm{d}\omega$. Hence the variance does not depend on the center band frequency, or the zoom factor at which the observation was made.

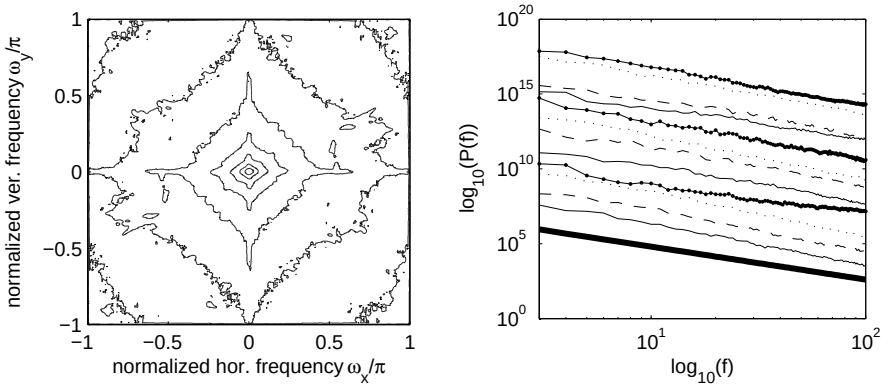
In addition, it is important to remark that the power law is applicable for a whole *ensemble* of images, but not necessarily for one particular image from this ensemble. For one particular image, (3.4) may not hold at all! On the other hand, (3.4) indicates another useful property for designing subband decomposition schemes: if subbands are used with a constant bandwidths in logarithmic scale, the variance of the image in each subband will be roughly equal [Field, 1987, Burton and Moorhead, 1987].

A second proposed explanation for the $1/\omega^2$ law is that it is caused by the presence of edges, which possess this frequency characteristic on their own. [Simoncelli and Olshausen, 2001]. Other authors have argued that the spatial distribution of sizes of objects and distances between objects causes this phenomenon [Ruderman, 1997, Lee and Mumford, 1999].

Although the PSD provides very useful "global" information about images, a description of the frequency distribution alone is not sufficient for representing various features in images, such as edges that are much more localized and often exhibit fluctuating local time-frequency characteristics. In general, PCA is found to be non-suitable for non-Gaussian random vectors (such as images), as the computed components are *not* independent.

3.1.2 Probabilistic PCA

Because PCA only defines a *linear* projection under which the projection components are uncorrelated, the applicability of PCA is somewhat limited [Tipping and Bishop, 1999], not only for *image modeling,* but also in the general context of data representation and visualisation. This lead to various schemes for *non-linear* PCA, in an attempt to model the non-linear structure (manifold) in which the data resides, such as Principal Curves [Hastie and Stuetzle, 1989], Generative Topographic Mapping [Bishop et al., 1998] and Locally Linear Embedding [Roweis et al., 2002a].



(a) (b)

Figure 3.1: (a) Iso-power contours of the PSD (averaged over 12 test images) (b) Logarithmic plot of the (radial) PSD of 12 test images, obtained by radially averaging. For clarity, the PSDs have been shifted vertically. The bold line indicates the power law $P(\boldsymbol{\omega}) \propto 1/\left\|\boldsymbol{\omega}\right\|^{2-\eta}$, where $\eta = 0.2$.

An alternative to non-linear projections is the use of mixtures of local linear models. Locally, the data is then assumed to be well described by a linear model (or a combination of a number of linear models), while globally, the data is contained in a non-linear manifold. In [Tipping and Bishop, 1999], probabilistic PCA (PPCA) is proposed for this task. PPCA is a latent variable model that also has an associated likelihood function. A latent variable model [Bartholomew, 1987] describes the set of observed data vectors $\mathbf{x}_j$ in a d-dimensional vector space $\mathcal{W}$ in terms of a set of q-dimensional latent (unobserved) variables $\mathbf{t}_j$, according to:

$$\mathbf{x}_j = \mathbf{h}(\mathbf{t}_j) + \mathbf{g}_j \tag{3.5}$$

where $\mathbf{h}(\cdot)$ is a function of random variable $\mathbf{t}_j$, and $\mathbf{g}_j$ is a residual process, independent of $\mathbf{t}_j$. In general, $q < d$, such that a lower dimensional description of the observed image is obtained. These models are sometimes also called *generative* [Tipping and Bishop, 1999], in the sense that a high-dimensional vector $\mathbf{y}_j$ can be obtained by mapping a low-dimensional vector $\mathbf{t}_j$ to a higher dimensional space, followed by adding a residual $\mathbf{g}_j$. For the task of image modeling, one observation vector $\mathbf{y}_j$ can (in principle) be either a complete image (e.g. reshaped to a vector using raster scanning) or a fixed-sized patch of an image.

The PPCA model introduced in [Tipping and Bishop, 1999] is a specific latent variable model in which the residual is assumed to be Gaussian with zero mean and (diagonal) covariance matrix $\sigma^2\mathbf{I}$, $\mathbf{t}_j$ is Gaussian with mean $\mathbf{m}$ and (diagonal) covariance $\mathbf{C}_t$. Further, the function $\mathbf{h}(\mathbf{t})$ is linear:

$$\mathbf{h}(\mathbf{t}) = \mathbf{V}\mathbf{t} \tag{3.6}$$

with $\mathbf{V}$ a $d \times q$-projection matrix. To build a mixture model, we can allow for multiple projection matrices for $\mathbf{V}$, by introducing K hypotheses:

$$\begin{cases} H_1, & \text{if } \mathbf{V} = \mathbf{V}_1 \\ H_2, & \text{if } \mathbf{V} = \mathbf{V}_2 \\ \vdots & \vdots \\ H_K & \text{if } \mathbf{V} = \mathbf{V}_K \end{cases}$$

Each hypothesis H_k corresponds to one suitable projection $\mathbf{V}_k$ that will yield uncorrelated components, i.e. if H_k holds for observation $\mathbf{x}_j$, the corresponding low-dimensional vector $\mathbf{x}_j$ will be uncorrelated. The model likelihood function is fairly easy to compute. Let $\mathbf{m}_k$ and $\mathbf{C}_{t|k}$ respectively denote the mean and covariance matrix of $\mathbf{t}$ under hypothesis H_k.[3] The conditional likelihood function is then given by:

$$f_{\mathbf{x}|H}\left(\mathbf{x}|H_k\right) \quad = \quad N\left(\mathbf{x}; \mathbf{m}_k, \mathbf{V}_k \mathbf{C}_{t|k} \mathbf{V}_k^T + \sigma^2 \mathbf{I}\right) \tag{3.7}$$

where $N\left(\mathbf{x}; \mathbf{m}, \mathbf{C}\right)$ is the Gaussian PDF (with mean $\mathbf{m}$ and covariance $\mathbf{C}$), evaluated in $\mathbf{y}$. From (3.7), the likelihood function can be computed using the law of total probability:

$$f_{\mathbf{x}}\left(\mathbf{x}\right) = \sum_{k=1}^{K} f_{\mathbf{x}|H}\left(\mathbf{x}|H_k\right) \mathrm{P}\left(H_k\right). \tag{3.8}$$

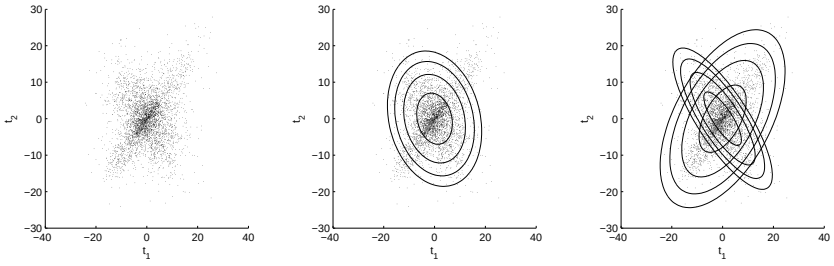
The PPCA model contains a set of model parameters

$$\mathbf{\Theta} = \left\{\mathbf{m}_k, \mathbf{V}_k, \mathbf{C}_{t,k}, \sigma, \mathrm{P}\left(H_k\right) k = 1, ..., K\right\},$$

which need to be trained either from the observed image itself, or from a training set of images. Because of the presence of unobserved, hidden variables, the direct training by maximizing the likelihood function is difficult. Instead, the expectation maximization (EM) algorithm [Dempster et al., 1977] is used. The EM algorithm is a general method for finding the maximum likelihood (ML) estimate of the model parameters $\mathbf{\Theta}$, when the data has missing values. The EM algorithm iteratively updates the posterior probabilities of the unobserved variables and the estimates of the model parameters. It can be shown [Tipping and Bishop, 1999] that a straightforward (but not the most efficient) implementation of the EM algorithm for PPCA is very similar to the EM algorithms for Gaussian mixtures, with one little modification that applies PCA to the activity[4] weighted sample covariance matrices. The algorithm is briefly summarized in Algorithm 3.1.

[3] For completeness, the description of PPCA in [Tipping and Bishop, 1999] is slightly different from the description here, in the sense that non-orthogonal projection matrices $\mathbf{V}$ are allowed such that $\mathbf{C}_{t|k} = \mathbf{I}$. The restriction to orthogonal projection matrices is without loss of generality and will offer some advantages later in Section 3.4.4. In our description, these "non-orthogonal" projection matrices are simply given by: $\mathbf{V}\mathbf{C}_{t|k}^{1/2}$.

[4] *Activities* are posterior probabilities $f_{H|\mathbf{x}}\left(H_k|\mathbf{x}\right)$.



(a) (b) (c)

Figure 3.2: Illustration of PPCA. (a) Scatter plot of two-dimensional data (neighboring pixel intensities), (b) Isoprobability contours of a fitted bivariate Gaussian PDF, (c) Isoprobability contours of the mixture components $f_{y|k}(\mathbf{y}|k)$ for a fitted PPCA model.

Remark that for $q = d$, the PPCA model specializes to a Gaussian Mixture model.

In Figure 3.2, the PPCA model is trained on a two-dimensional data (obtained by taking pairs of neighboring pixel intensities of an image as observation vectors) in Figure 3.2(a). Figure 3.2(b-c) show isoprobability contours of respectively a fitted Gaussian PDF and mixture components of the PPCA trained model. It is clear that the isoprobability contours for PPCA better coincide with the data than in case of a simple Gaussian PDF.

However, there are a few limitations of using PPCA:

1. The mixture model count K needs to be known in advance. "Greedy" EM algorithms [Vlassis and A., 2002, Verbeek et al., 2003] offer a solution here by starting with one mixture component and iteratively adding other mixture components when a certain criterion is fulfilled).

2. The negative log-likelihood function ($-\log f_{\mathbf{x}}(\mathbf{x})$) is non-convex. Consequently, the solution found by the EM algorithm is not necessarily the global optimum [Tipping and Bishop, 1999] and can depend on the chosen initial parameter values.

In the remainder, we will use PPCA as a reference method for comparison. In Section 3.4 we will further extend PPCA to mixtures of Gaussian Scale Mixture models.

3.1.3 Analysis of images in independent components

Another alternative to PCA for decomposing images is independent component analysis (ICA) [Bell and Sejnowsky, 1997]. ICA was originally introduced to separate mixed audio signals, for example voice recordings of two speakers that are talking simultaneously. Instead of directly decorrelating the components (which is not guaranteed to yield statistical independence), ICA aims at

Algorithm 3.1 A straightforward EM algorithm for PPCA (as implemented in *Netlab* [Nabney, 2001]).

1. Initialization: choose initial values for the parameters $\mathbf{\Theta}^{(1)} = \left\{\mathbf{m}_k^{(1)}, \mathbf{V}_k^{(1)}, \mathbf{C}_{t,k}^{(1)}, \sigma^{(1)}, \mathrm{P}^{(1)}\left(H_k\right) k = 1, ..., K\right\}$

2. Compute the posterior probabilities for iteration $i+1$ (starting from $i = 1$):

$$\mathrm{P}\left(H_k \left| \mathbf{y}_j, \mathbf{\Theta}^{(i+1)}\right.\right) = \frac{f_{\mathbf{y}|H,\mathbf{\Theta}}\left(\mathbf{y}_j \left| H_k, \mathbf{\Theta}^{(i)}\right.\right) \mathrm{P}\left(H_k \left| \mathbf{\Theta}^{(i)}\right.\right)}{\sum_{k=1}^{K} f_{\mathbf{y}|H,\mathbf{\Theta}}\left(\mathbf{y}_j \left| H_k, \mathbf{\Theta}^{(i)}\right.\right) \mathrm{P}\left(H_k \left| \mathbf{\Theta}^{(i)}\right.\right)}$$

3. Update the model parameters:

$$\begin{aligned}
\mathrm{P}\left(H_k \left| \mathbf{\Theta}^{(i+1)}\right.\right) &= \frac{1}{N}\sum_{j=1}^{N} \mathrm{P}\left(H_k \left| \mathbf{y}_j, \mathbf{\Theta}^{(i)}\right.\right) \\
\mathbf{m}_k^{(i+1)} &= \frac{\sum_{j=1}^{N} \mathbf{y}_j \mathrm{P}\left(H_k \left| \mathbf{y}_j, \mathbf{\Theta}^{(i)}\right.\right)}{\sum_{j=1}^{N} \mathrm{P}\left(H_k \left| \mathbf{y}_j, \mathbf{\Theta}^{(i)}\right.\right)} \\
\mathbf{C}_k^{(i+1)} &= \frac{\sum_{j=1}^{N} \mathbf{y}_j \mathbf{y}_j^T \mathrm{P}\left(H_k \left| \mathbf{y}_j, \mathbf{\Theta}^{(i)}\right.\right)}{\sum_{j=1}^{N} \mathrm{P}\left(H_k \left| \mathbf{y}_j, \mathbf{\Theta}^{(i)}\right.\right)}
\end{aligned}$$

4. PPCA-step: the projection bases $\mathbf{V}_k^{(i+1)}$ are obtained as the eigenvectors of $\mathbf{C}_k^{(i+1)}$ corresponding to the q most dominant eigenvalues (i.e. with largest magnitude). Therefore, let $\mathbf{C}_k^{(i+1)} = \mathbf{V}_k^{(i+1)} \mathbf{\Lambda}^{(i+1)} \left(\mathbf{V}_k^{(i+1)}\right)^T$ be the SVD of the positive definite matrix $\mathbf{C}_k^{(i+1)}$, where the eigenvalues (diagonal elements of $\mathbf{\Lambda}^{(i+1)}$) are sorted according to descending magnitude. Subsequently,

$$\begin{aligned}
\left(\sigma^2\right)^{(i+1)} &= \frac{1}{d-q}\sum_{n=q+1}^{d} \left[\mathbf{\Lambda}\right]_{nn} \\
\mathbf{C}_{t,k}^{(i+1)} &= \left(\mathbf{V}_k^{(i+1)}\right)^T \mathbf{C}_k^{(i+1)} \mathbf{V}_k^{(i+1)} - \left(\sigma^2\right)^{(i+1)} \mathbf{I}
\end{aligned}$$

5. Increase the iteration index i by 1 and go to step 2 until convergence (e.g. when the increment of the loglikelihood function is smaller than a predefined threshold).

maximizing the higher order statistical moments (e.g. kurtosis) of the components. This typically results in components that have very non-Gaussian

characteristics but that are (approximately) statistically independent. Because of a relatively high computational cost of ICA algorithms, the analysis is usually performed on small patches of an image (e.g. 15×15). More specifically, ICA decomposes an image patch into a sum

$$\mathbf{x} = \sum_{k=1}^{K} x_k \mathbf{u}_k, \tag{3.9}$$

where $\mathbf{u}_k$ denotes an ICA basis element and where $x_k, k = 1, ..., K$ are "independent" components of $\mathbf{x}$.

We remark that ICA on its own can not be regarded as a complete image model: the superposition in (3.9) can for example not account for occlusion of objects in images [Donoho and Flesia, 2001]. Nevertheless, the most important result is that the ICA yields edge filters: the basis elements are spatially localized, have a clear orientation and a band-pass frequency characteristic. Further, the computed independent components are sparse, which is very useful for many applications. Because the filters $\mathbf{Z}_k$ found by ICA are by assumption independent, the probability density function of the considered patch $\mathbf{x}$ can be found by multiplying the marginal distributions of the filter responses:

$$f_{\mathbf{x}}(\mathbf{x}) = \prod_{k=1}^{K} f_{x_k}(x_k). \tag{3.10}$$

However, in general, the independence assumption holds only approximately, which makes this density model only an approximation.

As an illustration, we computed principal components from overlapping patches of size 13×13 extracted from the *Lena* image (see Figure 3.3(a)). In Figure 3.3(b), PPCA basis vectors are shown, for patches of size 7×7, using 40 mixture components ($K = 40$) and with $q = 1$. Figure 3.3(c) shows independent components obtained from the same image and using the same patch size. To compute the independent components we used the *Fast-ICA* algorithm [Hyvärinen, 1999] with PCA-based prewhitening enabled. It can be seen that the PCA components resemble Fourier basis functions, while the ICA components are edge filters that have much better localization and orientation properties. We also computed the ICA basis elements optimized for a number of images, where we made a distinction between texture-rich images (*Barbara*, *Baboon*) and edge-rich images (*Lena*, *Boats*, *Man*). The results are shown in Figure 3.3(c)-(d): for *edge-rich* images, the ICA basis elements are even more dominantly edge filters, while for *texture-rich* images, the ICA basis elements are edge filters that are shifted and superposed on top of each other.

Applying ICA results in band-pass subbands with a non-Gaussian behavior. To test this, we carefully selected a PCA basis element, an ICA basis element and a shearlet basis element (see Chapter 2), such that the three filters have "similar" orientation and frequency characteristics. Next, we computed subband coefficients by applying the filters to the input image. As a measure of the non-Gaussianity, we use the sample kurtosis [Donoho and Flesia, 2001].

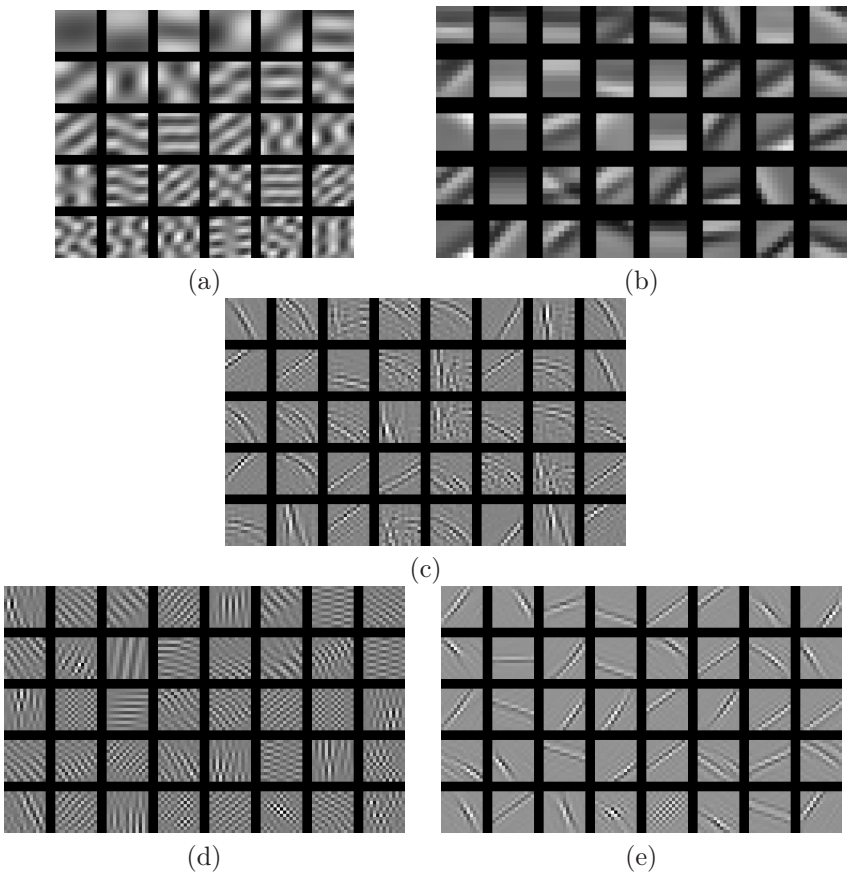


Figure 3.3: Basis elements computed using (a) PCA (*Lena* image), (b) PPCA (*Lena* image), (c) ICA (*Lena* image), (d) ICA (2 texture-rich images), (e) ICA (3 edge-rich images).

We computed the histogram and the kurtosis of the subband coefficients. The results are shown in Figure 3.4: for ICA, the kurtosis is the highest (38.21). In contrast to ICA, shearlet filters are image-independent and not optimized to the image structures, the kurtosis of the shearlet filtered subbands is also quite high (34.66). Finally, PCA also gives kurtotic filter responses, but the kurtosis is significantly lower (9.62). This provides evidence that the shearlet transform, while being image-independent, does allow to reveal the non-Gaussian structure in the images, in a similar way as ICA.

Although projections on ICA components give very sparse subbands, the redundancy (or the number of subband coefficients) is relatively high and depends on the number of ICA components chosen (which is typically the number of pixels in a patch). Also, ICA does not provide a means to estimate the actual number of independent components in advance.

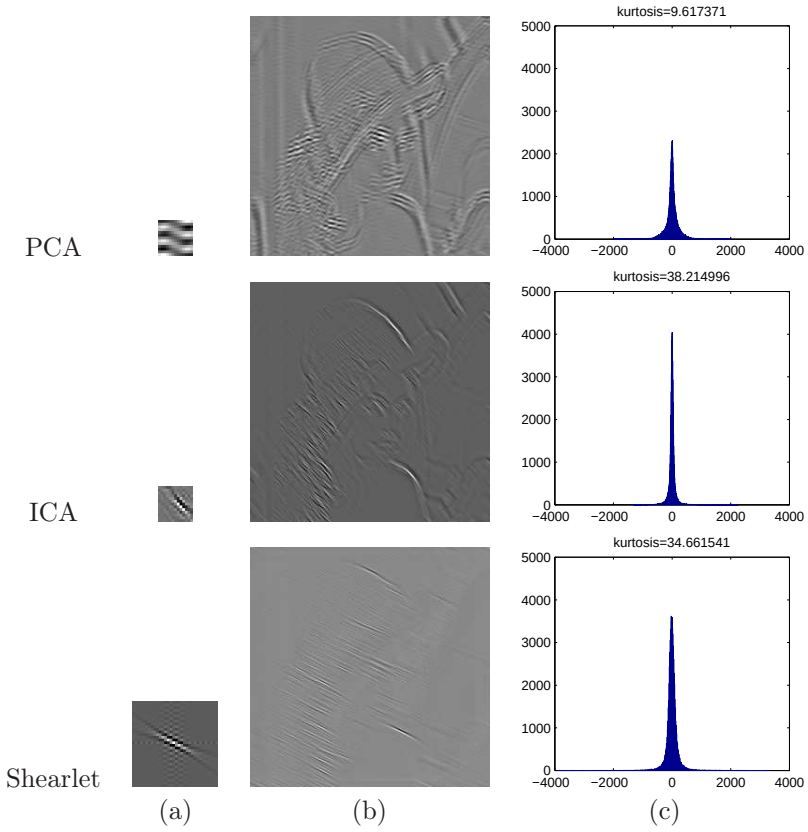


Figure 3.4: Illustration of the non-Gaussianity of band-pass filtered images. (a) Basis elements, (b) Filter responses, (c) Histogram of the filter responses from (b).

3.1.4 Related techniques

Next to ICA, several other authors investigated the construction of "optimal" bases by maximizing measures that depend on the higher order statistical moments. Examples of such measures are "Minimum Entropy" coding [Banham et al., 1994], which maximizes the sparsity of the filter responses. This has also led to projection pursuit, best orthogonal basis, matching pursuit and basis pursuit techniques. Projection pursuit [Friedman, 1987] iteratively searches for projections that lead to projection coefficients with a highly non-Gaussian behavior. When such a projection is found, the dimensionality of the image is reduced by removing the component along that projection. Matching pursuit [Mallat and Zhang, 1993] is a greedy solution technique for finding the best matching projections out of an overcomplete dictionary. Best orthogonal basis [Coifman and Wickerhauser, 1992] adaptively picks a single orthogonal basis that is the "best basis" out of many bases, thereby yielding near-optimal sparsity representations. Basis pursuit [Chen et al., 1998] decomposes an image

into an "optimal" superposition of basis elements, optimal in the sense that the L1-norm of the resulting coefficients is minimized.

Also related to ICA are sparse coding techniques inspired by the visual cortex (V1). In [Olshausen and Field, 1996, Olshausen and Field, 1997], a linear generative model is used, in which each patch is modeled in terms of a linear sum of basis functions, in such a way that for every patch, either only a few basis functions have non-zero weight in the sum (*sparsity*). These sparse coding objective are very related to non-Gaussianity criterion used in ICA.

Most of the above techniques focus on image patches, but not on statistics of the whole image. Therefore, [Sallee and Olshausen, 2003] propose a probability density function for the *whole* image. This prior consists of a mixture of a Gaussian distribution and a Dirac delta function. The model parameters are computed using Gibbs sampler techniques [Li, 1995]. The obtained basis is very similar to the steerable pyramid basis (see Chapter 2), which also indicates that image-independent bases often yield sparse representations. [Elad and Aharon, 2006a, Elad and Aharon, 2006b] train a dictionary based from an observed image, such that the dictionary yields a sparse representation for that image. This leads to a prior distribution for the whole image that enforces sparsity for overlapping small patches in the image.

To overcome some of the limitations of the ICA probability density model in (3.10), products of experts (PoE) [Hinton, 1999] have been used for modeling image patches in [Teh et al., 2003]. PoE is a special case of fields of experts (FoE) [Roth and Black, 2009] and models the high-dimensional PDF $f_{\mathbf{x}}(\mathbf{x})$ as products of several "expert" distribution, where each expert works on a low dimensional subspace that is much easier to model. Usually, one-dimensional subspaces are used (similar to the sparse coding models). The projection onto this subspace is performed using a linear filter, as in ICA approaches. [Teh et al., 2003] proposes to model the highly kurtotic data in these one-dimensional subspaces using a Student-t distribution. Similar to [Olshausen and Field, 1997], PoE allows for an overcomplete representation.

3.2 Parametric densities

Classical techniques (such as PCA), do not work well in practice because these methods assume that images are secondary order statistical processes. In parallel to the development of ICA-type techniques discussed in the previous section, researchers studied the higher order statistics of images and this led to some interesting results. Studies from [Field, 1987, Mallat, 1989a] were one of the first to point out the highly kurtotic shape of the histograms of band-pass filtered images. [Mallat, 1989a] proposed the generalized Laplacian distribution for modeling these highly kurtotic histograms. The non-Gaussian behavior has been investigated further by [Simoncelli and Adelson, 1996, Simoncelli, 1999], [Moulin and Liu, 1999], [Chang et al., 1998], [Srivastava et al., 2002], [Fadili and Boubchir, 2005] and [Boubchir, 2007]. Additionally, joint statistics of filter responses have been studied: the presence of local structures

such as edges and textures results in correlations between band-pass coefficients. Joint histograms of wavelet coefficients reveal dependencies across position, scales and even orientations. Shapiro exploited these dependencies for compression [Shapiro, 1993], with remarkable results. [Simoncelli, 1999] found that joint histograms often show correlations between the amplitudes of the band-pass filter responses, even when their signed responses are uncorrelated. Inside local neighborhoods of band-pass coefficients, joint histograms are typically ellipsoidal [Portilla et al., 2003].

In this section, we review several of parametric densities that have proven to be useful for modeling band-pass coefficients in various sparse representations (such as ICA, wavelets, shearlets...). We make a distinction between *univariate* densities and *multivariate* densities. For *univariate* densities, wavelet coefficients are assumed to be statistically independent and identically distributed. That means that each wavelet coefficient is treated as being drawn from a given (known) univariate distribution. Often the parameters of this distribution are estimated empirically from the complete set of wavelet coefficients for that particular subband.

In case of *multivariate* densities, a neighborhood of a given size is defined around every wavelet coefficient. Typically, squared neighborhoods of a rather small size (e.g. 3×3 or 5×5) are being used. The neighborhoods are overlapping. Next, every neighborhood can be represented by a vector (e.g. using column-stacking) that follows a multivariate density. The assumptions made are typically the same as in the case of univariate densities. Here individual neighborhoods are ussumed to be mutually independent, despite the overlap. This is mostly to keep the computations tractable and to keep the resulting algorithms practical, although recently it has been shown that modeling dependencies across neighborhoods (e.g. by Markov random field type approaches) can increase the performance of the model in e.g. denoising [Lyu and Simoncelli, 2008].

Besides the local spatial neighborhood, it is also possible to include wavelet coefficients from other scales and/or orientations: for example, in [Portilla et al., 2003] better denoising results were obtained in some cases by including the parent wavelet coefficient (in the wavelet tree) to the local neighborhood. Later (in Section 3.3) we will investigate which information to include in the spatial neighborhood, in the remainder of this section, we will concentrate on the involved parametric densities.

3.2.1 Generalized Laplace distribution

For natural images, the marginal histograms of wavelet coefficients (as shown in Figure 3.4(c)) typically have a long tail and a sharp peak at 0. Several authors proposed to use a *generalized Laplace* distribution (also known as *generalized Gaussian* distribution, GGD) to model this behavior [Mallat, 1989b, Antonini et al., 1992, Simoncelli and Adelson, 1996, Chang et al., 1998, Moulin and Liu,

1999, Pižurica and Philips, 2006]:

$$f_x(x) = \frac{\nu}{2s\Gamma(1/\nu)} \exp\left(-\left|\frac{x}{s}\right|^{\nu}\right) \tag{3.11}$$

where $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t$ is the Gamma function. The parameter s is a scale parameter that is related to the variance of the GGD, which is given by

$$\sigma^2 = s^2\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}. \tag{3.12}$$

The parameter ν is a shape parameter of the GGD, which controls the kurtosis of the distribution. More specifically, the kurtosis is given by:

$$\kappa = \frac{\Gamma(5/\nu)\Gamma(1/\nu)}{\Gamma(3/s)^2} - 3. \tag{3.13}$$

For the specific case $\nu = 1$, a *Laplace* distribution (also called *double exponential* distribution) is obtained. For $\nu = 2$, (3.11) amounts to a Gaussian distribution. For wavelet, shearlet or STP coefficients, ν is typically in the range $[0.5, 1]$ and increasing with the support size of the basis functions, i.e. lower frequency subbands tend to have higher ν. The parameters are usually estimated by the method of moments (i.e. by replacing the variance σ^2 and kurtosis κ in (3.12) and (3.13) by respectively the sample variance and sample kurtosis estimated from the observed wavelet subbands).

3.2.2 Weighted mixtures of two distributions

An alternative to (generalized) Laplacian distributions are mixtures of two distributions, where one distribution models the "significant" coefficients (i.e. coefficients with a large magnitude) and where the other distribution models the "non-significant" coefficients (i.e. coefficients with a small magnitude) [Vidakovic, 1998b, Leporini et al., 1999, Abramovich et al., 1998, Chipman et al., 1997, Clyde et al., 1998, Crouse et al., 1998, Romberg et al., 2001b, **?**, Pižurica et al., 2002, Pižurica and Philips, 2006, Pižurica and Philips, 2007, Shi and Selesnick, 2006]. A Bernoulli (*hidden*) random variable is used as the mixing parameter. Examples are:

- Mixtures of a Gaussian distribution and a point mass at zero [Abramovich et al., 1998, Clyde et al., 1998].
- Mixtures of two Gaussian distributions [Crouse et al., 1998, Romberg et al., 2001b, ?] (see Figure 3.5(a)).
- Mixtures of two (truncated) Laplace distributions [Pižurica et al., 2002, Pižurica and Philips, 2006, Pižurica and Philips, 2007] (see Figure 3.5(b)).

The merit of mixture distributions is that the probability density functions of the mixture components can be relatively simple, which facilitates statistical

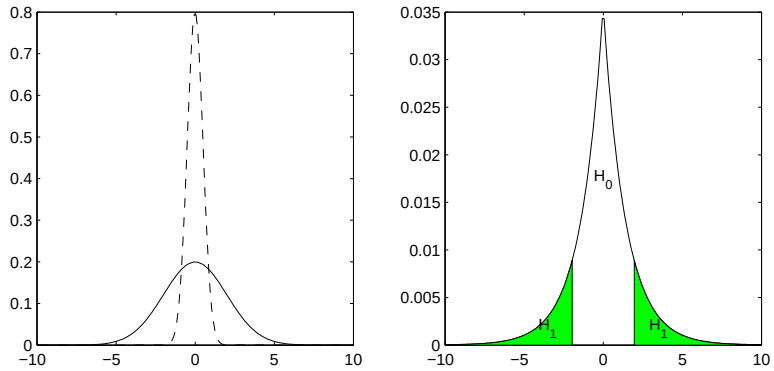


Figure 3.5: Weighted mixture models of two distributions (a) Mixture of Gaussians (b) Mixture of two truncated Laplace distributions.

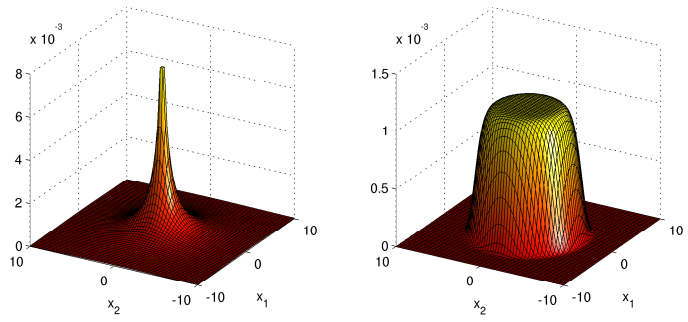
estimation techniques (e.g. denoising) and makes the computations analytically tractable. Further, the mixing variable can be used to model interdependencies between coefficients. Let H_0 and H_1 denote the hypotheses that a given coefficient is respectively "non-significant" and "significant", then the model likelihood function can be written as:

$$f_x(x) = \mathrm{P}(H_0)\, f_{x|H}(x|H_0) + \mathrm{P}(H_1)\, f_{x|H}(x|H_1),$$

where the conditional densities $f_{x|H}(x|H_k)$ are chosen as listed above. The mixture model depends on the prior probabilities $\mathrm{P}(H_0)$ and $\mathrm{P}(H_1)$. When the mixture model is used to model coefficients *within the same* subband, prior probabilities are normally estimated per subband [Chipman et al., 1997]. Techniques that exploit dependencies between coefficients, estimate $\mathrm{P}(H_0)$ and $\mathrm{P}(H_1)$ *adaptively* per coefficient, in a Hidden markov tree framework (for taking into account interscale dependencies, see further in Section 3.5), in a Markov Random Field (MRF) model (to model intrascale dependencies, see Section 3.4.1), or by conditioning the hypothesis to a local spatial activity indicator (see further in Section 3.4.2).

3.2.3 Elliptically symmetric distributions

A number of recent statistical studies (e.g. [Chang et al., 2000a], [Portilla et al., 2003]) have shown that distributions of noise-free wavelet coefficients are typically symmetric around the mode, have a highly kurtotic non-Gaussian behavior and exhibit strong correlations, especially in areas with edges and textures. For many natural images, the histograms of the wavelet coefficients reveal elliptical contours, which suggests the use of the elliptically symmetric family for modeling this behavior. The family of *elliptically symmetric* distributions



(a) (b)

Figure 3.6: The multivariate exponential power (EPD) distribution (a) $\nu = 0.5$, (b) $\nu = 10$.

is defined by the following class of densities [Kotz and Kozubowski, 2001]:

$$f_{\mathbf{X}}(\mathbf{x}) = k_d |\mathbf{C}_x|^{-1/2} g\left(\left| (\mathbf{x}-\mathbf{m})^T \mathbf{C}_x^{-1} (\mathbf{x}-\mathbf{m}) \right|^{1/2} \right) \tag{3.14}$$

where $\mathbf{m}$ is the mean of the distribution, $g(u)$ is a one dimensional real-valued function (called *density generator function*), d is the number of dimensions of $\mathbf{x}$ and k_d is a proportionality constant.

An example is the multivariate extension of the generalized Laplace distribution (usually known and referred to as the multivariate *exponential power* distribution, EPD) [Gómez et al., 1998, Kotz et al., 2000], which has the folowing density generator function:

$$g(x) = \exp\left(-|x|^{\nu}\right).$$

where ν is a shape parameter. The EPD density function is depicted in Figure 3.6 for both leptokurtic and platykurtic shapes. However, there are a few issues related to the multivariate EPD when modeling neighborhoods of wavelet coefficients:

- When a random vector has an EPD, its marginal distributions do generally not belong to the EPD class (except for the case of a Gaussian distribution, $\nu = 2$).
- The distribution often leads to expressions that are analytically intractable. An example is the MAP or MMSE estimator for an EPD random variable corrupted with Gaussian noise.

Fortunately, the Bessel K Form (BKF) has similar properties as the GGD/EPD and does not suffer from these problems. The BKF distribution belongs to the class of the Gaussian Scale Mixtures, which will be discussed next.

3.2.4 Gaussian Scale Mixtures

Wainwright and Simoncelli [Wainwright and Simoncelli, 2000] used the property that when the band-pass filter responses are normalized by dividing by the square root of their local variance, the statistics of the normalized coefficients are approximately Gaussian. For this reason, the Gaussian Scale Mixtures (GSM) was proposed, to account for second order statistics and for the variability in the local variance of the wavelet coefficients.

A random variable $\mathbf{x}$ conforms to a GSM model [Andrews and Mallows, 1974] if it can be written as the product of a zero mean Gaussian random vector $\mathbf{u}$ (with covariance $\mathbf{C}_u$) and a scalar random variable $z^{1/2}$ where $z \geq 0$:

$$\mathbf{x} \stackrel{\mathrm{d}}{=} z^{1/2}\mathbf{u} \tag{3.15}$$

where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution. The scalar random variable z is often called 'hidden' multiplier (or mixing variable) because it is not observed. Given that the expected value of z exists, the covariance matrix of $\mathbf{x}$ is related to the covariance matrix $\mathbf{C}_u$ by

$$\mathbf{C}_x = \mathrm{E}[z]\,\mathbf{C}_u. \tag{3.16}$$

To avoid scaling ambiguity between $\mathbf{u}$ and z, the convention $\mathrm{E}[z] = 1$ is often used such that $\mathbf{C}_x = \mathbf{C}_u$. Prior distributions $f_z(z)$ for the hidden variable z include Jeffrey's non-informative prior [Portilla et al., 2003], the *log-normal* prior [Portilla and Simoncelli, 2001], the *exponential* distribution [Selesnick, 2006] and the *Gamma* distribution (see e.g. [Srivastava et al., 2002, Fadili and Boubchir, 2005]).

Like the EPD, the GSM also belongs to the family of elliptically symmetric distributions, where the density generator function is in this case given by:

$$g(x) = \int_0^{+\infty} f_z(z)\, z^{-\frac{d}{2}} \exp\left(-\frac{1}{2z}x^2\right) \mathrm{d}z \tag{3.17}$$

where a marginalization takes place over the hidden multiplier. In some specific cases, a closed form expression can be found for $g(x)$, although most often, the integration is performed numerically over a closed interval. To have the best approximation with the fewest count of integration points possible, integration points for z are chosen to be linear in a logarithmic scale, i.e. $z_k \propto \exp(ak)$ with a a constant [Portilla et al., 2003].

Recently, it has been shown that the multivariate exponential power distribution is also a Gaussian Scale Mixture distribution [Gómez et al., 2008], *for some values* of the shape parameter $\nu \in]0, 1]$, i.e. if the kurtosis of the distribution is higher than the kurtosis of the Laplace distribution. However, the distribution of the hidden multiplier depends on d and has a rather complicated analytical expression (see [Gómez et al., 2008]).

The hidden multiplier z can also be estimated directly from the data, e.g. using a maximum likelihood estimate [Wainwright and Simoncelli, 2000]:

$$\hat{z} = \mathbf{x}^T \mathbf{C}_u^{-1} \mathbf{x}^T / d, \tag{3.18}$$

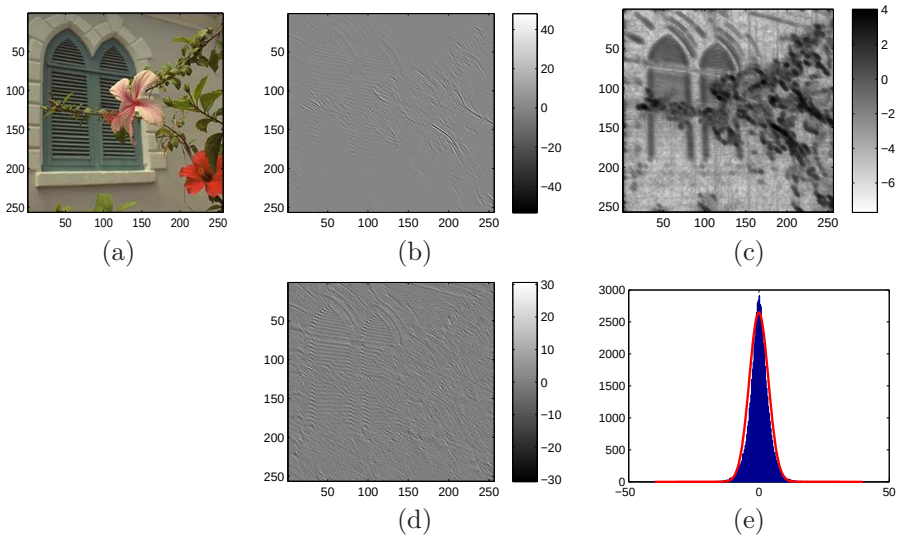


Figure 3.7: Illustration of the hidden multiplier z, estimated from a shearlet subband using (3.18). (a) *Window* image, (b) Shearlet subband, (c) Logarithm of the estimated local variance ($\log \hat{z}$). *Black* corresponds to a high local variance, *white* to low local variance. (d) Normalized shearlet subbands $\mathbf{x}/\sqrt{\hat{z}}$, (e) Histogram of (d) and fitted Gaussian distribution (thick line).

which is basically a measure for the local variance of the data $\mathbf{x}$. In Figure 3.7(b), a shearlet subband of the *window* image (Figure 3.7(a)) is depicted, together with its estimated local variance (Figure 3.7(c)). In this example, a 3×3 neighborhood was used.

3.2.5 Bessel K Form density

An alternative to the multivariate EPD is the *Bessel K Form (BKF)* prior [Kotz and Kozubowski, 2001,Srivastava et al., 2002,Fadili and Boubchir, 2005], which is again an elliptically distributed distribution with density generator:

$$g(u) = \left(\frac{u}{2}\right)^{\frac{\tau}{2}-\frac{d}{4}} K_{\tau-d/2}(\sqrt{2u}), \text{ and } k_d = \frac{2(2\pi)^{-d/2}}{\Gamma(\tau)} \tag{3.19}$$

where $K_i(u)$ is the modified Bessel function of the second kind and order i (see [Kotz and Kozubowski, 2001]) and $\Gamma(\tau) = \int_0^\infty z^{\tau-1}e^{-z}dz$ is the Gamma function. In [Srivastava et al., 2002], it has been shown that the marginals of this distribution fit well with the observed histograms for a wide variety of images. The BKF distribution is also a specific case of the GSM, in which the hidden multiplier has a Gamma distribution:

$$f_Z(z) = \frac{1}{\Gamma(\tau)} z^{\tau-1}e^{-z}. \tag{3.20}$$

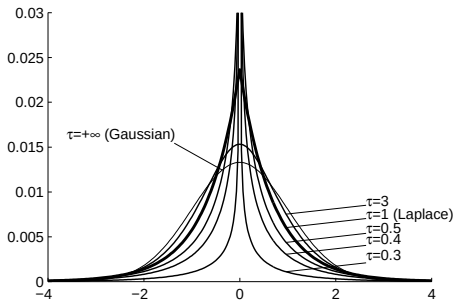


Figure 3.8: The univariate Bessel K form density, for different values of τ. Special value $\tau = 1$ gives the symmetric Laplace distribution, for $\tau \to +\infty$ the Bessel K form approaches the Gaussian density. Values $\tau < 1$ result in a high kurtosis (sharp peak).

The use of the BKF density is attractive because it generalizes both the multivariate Laplace distribution (for $\tau = 1$ and z is *exponentially* distributed [Kotz and Kozubowski, 2001, Selesnick, 2006]) and the Gaussian distribution (For $\tau \to \infty$ [Kotz and Kozubowski, 2001]), just as the EPD. In contrast to the EPD, conditional density functions of $\mathbf{x}$ are also BKF distributed. For practical applications, the BKF is computationally and analytically much more tractable because it consists of a mixture of Gaussian distributions. We also note that the Bessel K Form corresponds to the symmetrized Gamma family proposed in [Wainwright and Simoncelli, 2000]. The kurtosis is given by $\kappa = 3 + 3/\tau$, thus for small positive τ, we obtain a highly leptokurtic prior. Furthermore, the parameter τ depends on the frequency of occurrence (or sparsity) of particular features in the image, like edges, bands, textures [Srivastava et al., 2002]. In Fig. 3.8 the univariate marginals of the Bessel K Form density are plotted for different values of τ.

Compared to GSM distributions with other hidden multiplier priors (log-normal, Jeffrey and exponential), the BKF is the only one that offers explicit control of the kurtosis, which is advantageous when modeling wavelet subbands of natural images (see [Srivastava et al., 2002]). The parameter τ is usually estimated through the method of Matching Cumulants (see [Fadili and Boubchir, 2005]), however, we remark that the maximum likelihood estimate (obtained through an EM algorithm, see [Boubchir, 2007, p. 85]) is often more accurate.

In [Fadili and Boubchir, 2005] the Bessel K Form prior is compared to the α-stable prior and Generalized Gaussian Distribution in modeling observed histograms by means of the Kullback-Leibler divergence (KLD). The authors conclude that the Bessel K prior performs at least as well as the GGD for modeling the statistics of wavelet coefficients of a test set of natural images. In Figure 3.9 we performed a similar experiment for modeling shearlet coefficients. It can be seen that the GGD and BKF density fits are considerably better in modeling the highly kurtotic behavior of the shearlet coefficients than the Laplace density fit. The Kullback-Leibner divergences for the GGD are respectively: 0.1008, 0.0960, 0.0269 while for the BKF density: 0.0866, 0.1086,

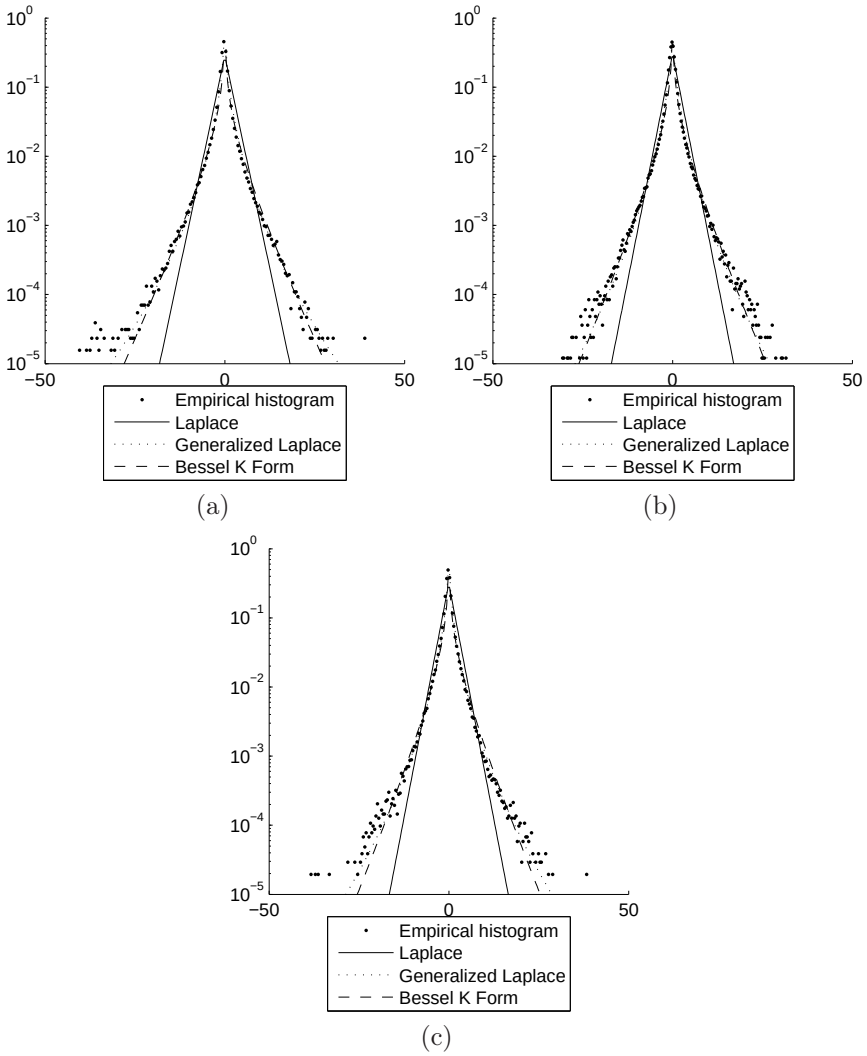


Figure 3.9: Different empirical marginal histograms of shearlet coefficients (for one image) with fitted distributions. A logarithmic scale is used for the y-axis.

0.0450. Hence the GGD is marginally better in terms of KLD than BKF for these examples, although we found that this is generally image-dependent.

3.2.6 Other densities

Some authors have used a number of related probability density functions for modeling subband coefficients. Examples are: the Student-T distribution [Tzikas et al., 2007], Alpha-stable distributions [Nikias and Shao, 1995, Achim

et al., 2001b] and the Cauchy distribution [Rabbani et al., 2006]. All these heavy tailed distributions have a Gaussian Scale Mixture representation, hence studying general GSMs automatically covers all of these distributions. Further, a complex extension of the Gaussian Scale Mixture density has been proposed for modeling complex-valued wavelet coefficients in [Vo et al., 2007]. This complex GSM distribution is a special case of the GSM distribution, with a special condition imposed to the covariance matrix of the distribution.

3.3 Joint statistics of subband coefficients

In general, multiresolution representations do not fully decorrelate the signal, and noise-free coefficients often exhibit strong local correlations. To illustrate this, we applied a dyadic shearlet transform with two scales and four orientations to an image of two squirrels (Figure 3.10, left image of the bottom row). Next, we computed the joint histograms of shearlet coefficients:

1. within the same scale (called *intra-scale* histograms),
2. between adjacent scales (called *inter-scale* histograms) and
3. between adjacent orientations (called *inter-orientation* histograms).

The results are shown in Figure 3.10. The first row shows the basis elements for the second scale of the used shearlet transform, to indicate the analysis orientation. The second row (Figure 3.10(a)-(d)) shows equiprobability contours of the joint histogram of neighboring shearlet coefficients within the *same subband*, horizontally next to each other. It can be seen that the joint histograms reveal approximately elliptical contours, which can be well modeled using elliptically symmetric densities as explained in Section 3.2. Also, the correlation is maximal (in absolute sense) in the direction *orthogonal* to the filter direction (see Figure 3.10(d)). The third row (Figure 3.10(e)-(h)) shows equiprobability contours for the joint histogram between coefficients residing in *adjacent scales*. Here, elliptical symmetry is only present up to a certain degree: in general the correlations seem to be much more complex and cannot be described well by elliptically symmetric densities. Figure 3.10(j)-(l) illustrates the joint histograms for shearlet coefficients that belong to different *orientation bands*: the equiprobability contours are *rhombus*-shaped. This phenomenon can be explained by the fact that natural images contain many oriented edges. Suppose that at a given position in the image we find an edge with angle 45°, then the shearlet subbands with orientation 45° will reveal a high amplitude at that position, while in the other orientation subbands, the shearlet coefficient amplitude at the same position will be very low.

Several authors have pointed out that by exploiting dependencies between these coefficients, improvements can be achieved in various applications, such as denoising [Chang et al., 2000a, Şendur and Selesnick, 2002a, Portilla et al., 2003, Crouse et al., 1998] and compression [Shapiro, 1993]. One way to deal with these dependencies, is to model the joint statistics using multivariate densities

Table 3.1: Correlation coefficients for shearlet coefficients

Type of correlation	*Lena*	*Barbara*	*Peppers*	*Man*	*House*
Inter-scale correlation	0.60	0.67	0.49	0.58	0.59
Inter-orientation correlation	0.23	0.23	0.18	0.24	0.23
Intra-subband	**0.81**	**0.83**	**0.79**	**0.79**	**0.83**

that are presented above. Despite the fact that elliptically symmetrical densities offer many benefits, most empirical joint histograms in Figure 3.10 do not have elliptical iso-probability contours. To model these joint histograms, one needs to use probability densities with more parameters to estimate. However, when the number of parameters becomes too large, the estimation from one observed image becomes unreliable. Hence we need to strike a balance between the number of model parameters and the number of exploited dependencies and we need to find out which dependencies are stronger than others. For this task, we performed a simple experiment using the shearlet transform, similar to the experiment in [Tessens et al., 2008] for the curvelet transform. First we transformed a number of images to the shearlet domain. Next, we computed the average correlation coefficients between shearlet coefficients residing in different subbands and within the same neighborhood (intra-subband). The results are shown in Table 3.1. For computing the inter-scale and inter-orientation correlation coefficients, we used an *undecimated* shearlet transform (obtained by skipping the decimation steps of the transform). The intra-subband correlation was computed between neighboring shearlet coefficients, in the direction orthogonal to the filtering direction (this is the direction in which the correlations are *maximal*) and for a *decimated* shearlet transform (to reduce the amount of correlations introduced by the transform itself). Although correlation coefficients cannot fully capture all dependencies, the table gives a good idea of which dependencies are the most relevant.

3.4 Models for intra-scale correlations

3.4.1 Markov Random Field models

In [Malfait and Roose, 1997, Jansen and Bultheel, 1999, Pižurica et al., 2002], Markov Random Field (MRF) models are developed for wavelet-domain image denoising. These models encode the "geometry" of subbands by giving preference to spatially connected configurations of significant wavelet coefficients. Similar to weighted mixtures of Gaussians (Section 3.2.2), a hidden variable $k_j \in \{0, 1\}$ is attached to each coefficient x_j (or node from V), where again $k_j = 1$ denotes that the coefficient is "significant" and $k_j = 0$ denotes the opposite. The vector of binary labels $\mathbf{k} = [k_1 \cdots k_N]$ for a given subband, consisting of N-coefficients, is called a mask. Each mask is assumed to be a realization of a Markov Random Field k. A set of neighboring pixels in a predefined configuration is called a *clique*. Let $\mathcal{C}$ denote the set of all considered cliques, then

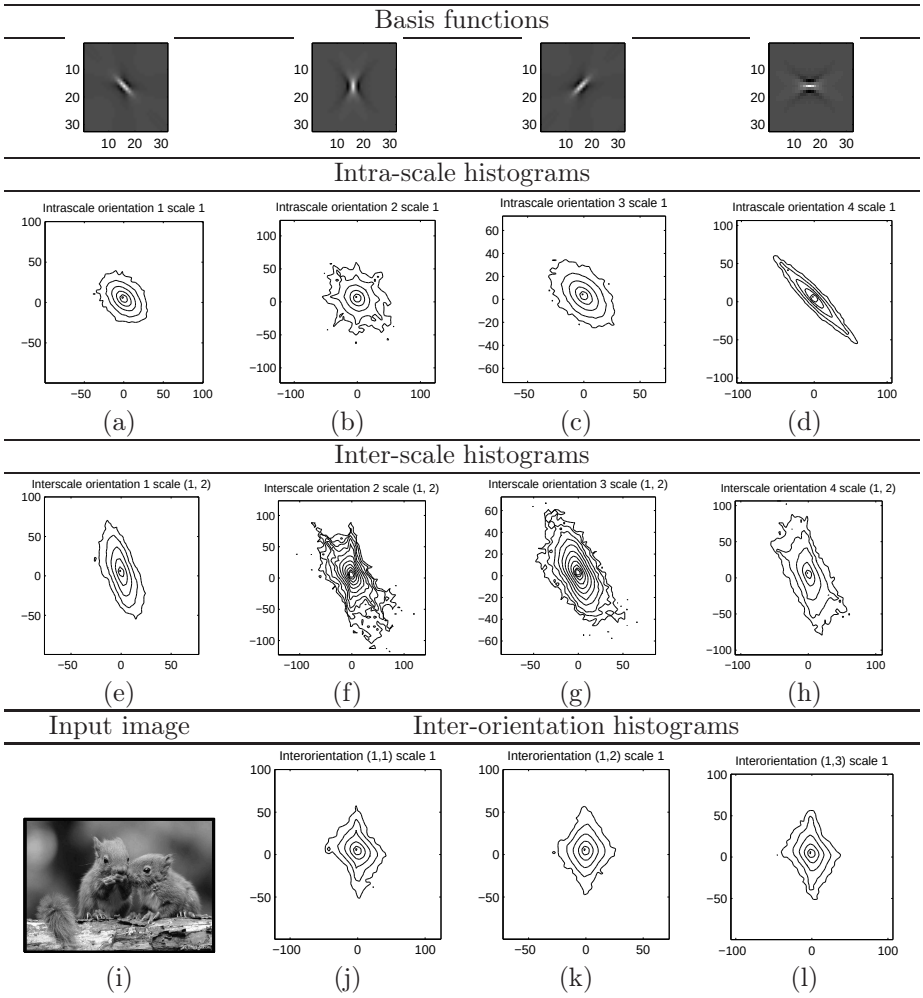


Figure 3.10: Equiprobability contours of the joint histograms of shearlet coefficients.

by the Hammersley-Clifford theorem [Moussouris, 1974], the joint probability of $\mathbf{x}$ is given by the Gibbs distribution:

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{A} \exp\left(-\frac{1}{T} \sum_{C \in \mathcal{C}} V_C(\mathbf{x})\right), \tag{3.21}$$

where A is a probability density normalization constant, T is called "temperature" constant and $V_C(\mathbf{x})$ is a clique potential function. The clique potential function is defined to give preference to certain local spatial dependencies. In many cases, the cliques are chosen a priori by hand, according to some regular neighborhood structure [Roth and Black, 2009]. Often, pairwise cliques are chosen, in which each label is connected to one of its four direct neighbors to

the left, top, right and bottom. This is then called a pairwise MRF model (the maximal cliques are pairs of neighboring labels). Further, a MRF model is called *homogeneous* if the clique potential functions are invariant to the spatial position (hence yielding a shift-invariant representation). *Isotropic* MRF models (e.g. [Malfait and Roose, 1997]) treat all spatial directions equally, while *anisotropic* MRF models [Pižurica et al., 2002] are slightly more complex, but are generally better in modeling directional features, such as edges. Another interesting case of a MRF model is a Gaussian MRF [Chellappa, 1985]:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}) \tag{3.22}$$

where elements of the precision matrix $\mathbf{C}^{-1}$ are zero ($\left[\mathbf{C}^{-1}\right]_{j,j'} = 0$) if the pixels at positions j, j' are unconnected.

Using Markov Random Fields, it is often desirable to compute the posterior probabilities that a given coefficient x_j is either significant or non-significant:

$$\begin{aligned} \mathrm{P}(k_j = 1|\mathbf{x}) &= \int_{\mathbf{k}\in[0,1]^N : k_j = 1} \mathrm{P}(\mathbf{k}|\mathbf{x})\, \mathrm{d}\mathbf{k} \\ &= \frac{\int_{\mathbf{k}\in[0,1]^N : k_j = 1} f_{\mathbf{x}|\mathbf{k}}(\mathbf{x}|\mathbf{k})\, \mathrm{P}(\mathbf{k})\, \mathrm{d}\mathbf{k}}{\mathrm{P}(\mathbf{x})} \end{aligned} \tag{3.23}$$

and an equivalent expression can be written for $\mathrm{P}(k_j = 0|\mathbf{x})$. This process is called *Bayesian inference*, however, the exact computation of $\mathrm{P}(k_j|\mathbf{x})$ is *intractable*, because as (3.23) shows, it requires the summation over all possible configurations $\mathbf{k}$. In practice, the summation therefore takes place over a set of "important" configurations [Pižurica and Philips, 2003], which is called *importance sampling*. In [Malfait and Roose, 1997, Pižurica et al., 2002], the Metropolis sampler is used for this task.

Recently, the Fields of Gaussian Scale Mixtures (FoGSM) model has been proposed in [Lyu and Simoncelli, 2008]. In FoGSM, the coefficient subbands are modeled as the product of a homogeneous Gaussian MRF (hGMRF) and a second independent *positive-valued* MRF. The former MRF models the second-order statistics of the coefficients, while the latter captures the variability of the local variance. Individual coefficients of the FoGSM model are marginally GSM distributed, while the global MRF structure generates dependencies across local neighborhoods (where GSM models typically assume that different local neighborhoods are statistically independent). Unfortunately, by the homogenity assumptions, FoGSM is not able to capture long-distance interactions that are present in images around captures and in textures. Nevertheless, the FoGSM model gives a denoising performance comparable to state-of-the-art approaches [Lyu and Simoncelli, 2008].

The Fields of Experts (FoE) model [Roth and Black, 2009] is a higher-order MRF model that uses clique potentials that represent products of experts (see Section 3.1.4). All parameters, including the filters of the PoE model, are trained from the observed images. The FoE model is, compared to other MRF models, directly applicable to many image processing problems, such as

denoising, impainting. The drawback is that this inference in this higher-order MRF model is considerably more difficult than for more simple MRF models, for this reason only small clique sizes have been used so far.

3.4.2 Local spatial activity indicators

In [Pižurica and Philips, 2006], it is proposed to estimate the probability that a given coefficient is significant given its value and knowing the marginal distribution of the noise-free coefficients. This probability was used in a denoising application, as a suppression factor for the wavelet coefficients in the *Prob-Shrink* estimator. A locally adaptive version of this approach was also introduced in [Pižurica and Philips, 2006] which attempts at making use of spatial correlations that exist between the coefficients within the same subband. The marginal distribution is a weighted mixture of truncated Laplace distributions (see Section 3.2.2) that are conditioned on a local spatial activity indicator v_j:

$$\begin{aligned} f_x(x_j) &= \int_0^{+\infty} f_{x|v}(x_j|v_j)\, f_v(v_j)\, \mathrm{d}v \quad \text{with} \\ f_{x|v}(x_j|v_j) &= \mathrm{P}(H_0|v_j)\, f_{x|v,H}(x_j|v_j, H_0) + \mathrm{P}(H_1|v_j)\, f_{x|v,H}(x_j|v_j, H_1), \end{aligned}$$

where the Local Spatial Activity Indicator (LSAI) v_j is computed as the locally averaged coefficient magnitude.

The rationale behind this approach is: if a wavelet coefficient is large (small) in magnitude then the majority of the neighboring coefficients within a local window is also likely to be large (small) because true image discontinuities typically result in spatially clustered wavelet coefficients. To illustrate this rationale, we computed equiprobability contours of the joint histogram (averaged over 7 test images) of a shearlet coefficient and its LSAI, for two different window sizes (see Figure 3.11). According to the rationale, the equiprobability contours need to be radial lines, with both positive and negative slopes, that pass through the origin $x = v = 0$. From the figure, it can be seen that this is satisfied to some extent. Furthermore, for a fixed value of v, the coefficient x is approximately equally likely to be in the range $[-v, v]$.

3.4.3 Mixtures of Gaussian Scale Mixtures

A recent extension to the GSM model for modeling spatial correlations, are Mixtures of GSMs [Guerrero-Colón et al., 2008b]. First, consider a square $\sqrt{d} \times \sqrt{d}$ local (overlapping) neighborhoods of coefficients residing within the same subband (where $\sqrt{d}$ is integer). Each of these neighborhoods can be seen as a realization of a d-dimensional random vector $\mathbf{x}$. The covariance matrix $\mathbf{C}_x$ of this vector has size $d \times d$, the matrix is symmetrical and contains $d(d+1)/2$ independent parameters. We further denote by

$$R(\mathbf{p}, \mathbf{q}) = (\mathbf{C}_x)_{\mathbf{p},\mathbf{q}} \tag{3.24}$$

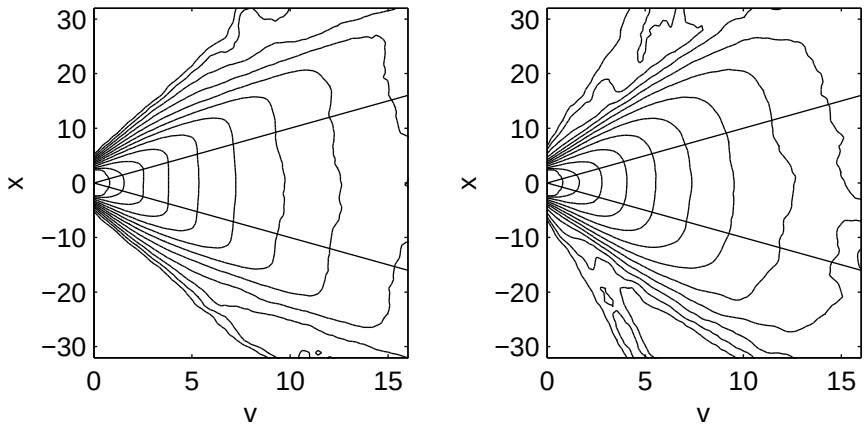


Figure 3.11: Equiprobability contours of the joint histogram of a coefficient x and its LSAI v for (a) 3×3-neighborhood, (b) 7×7-neighborhood. The lines $x = \pm v$ are also shown (straight lines).

the covariance between the components corresponding to the positions $\mathbf{p} = [p_1, p_2] \in [0, \sqrt{d} - 1]^2$ and $\mathbf{q} = [q_1, q_2] \in [0, \sqrt{d} - 1]^2$ of the local window, i.e., the element at row $(p_1 + \sqrt{d}p_2 + 1)$ and column $(q_1 + \sqrt{d}q_2 + 1)$ of $\mathbf{C}_y$. Here p_i and q_i are the ith component of respectively $\mathbf{p}$ and $\mathbf{q}$. When either $\mathbf{p}$ or $\mathbf{q}$ are *outside* the local window, we assume that the corresponding covariance $R(\mathbf{p}, \mathbf{q}) = 0$, thus we only consider correlations between coefficients inside the local window.

When assuming spatial stationarity of the observed wavelet coefficients, the covariance two coefficients at positions $\mathbf{p}$ and $\mathbf{q}$ only depends on the difference in location between both positions (see Section 3.1):

$$(\mathbf{C}_x)_{\mathbf{p},\mathbf{q}} = R(\mathbf{p}, \mathbf{q}) = R(\mathbf{0}, \mathbf{q} - \mathbf{p}). \tag{3.25}$$

In Figure 3.12, the autocorrelation function is shown for the high-pass bands of the shearlet transform of the House test image. In Figure 3.12 we notice that the spatial correlations are typically the strongest in the directions *orthogonal* to the filter direction.

These second order statistics can be taken into account by modeling $\mathbf{x}$ using elliptically symmetric distributions, such as the GSM (Section 3.2.4). The traditional GSM model, as employed in [Portilla et al., 2003], assumes that the signal covariance matrix is constant within each subband. Ideally, this would mean that if we would divide the subband into different regions, the covariance matrix of the coefficients in every region would be approximately[5] the same, up to a constant scaling factor. Despite the fact that this holds for large ensembles of images, it turns out that for individual images this is not always the case and

[5] Due to estimation errors.

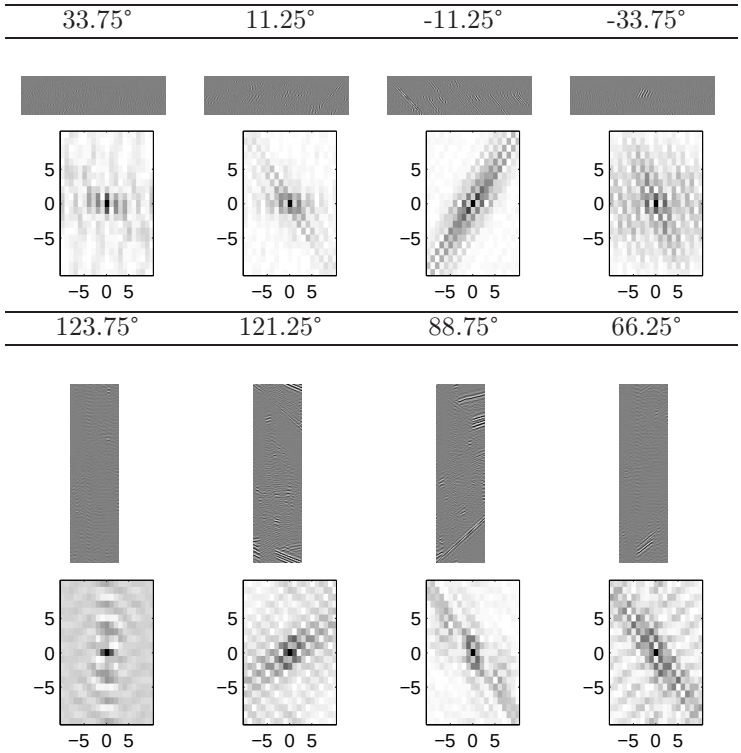


Figure 3.12: High-pass shearlet subbands bands of the House image, for the different orientations. Below each band is the *autocorrelation function* $R(\mathbf{0}, \mathbf{p})$ for that band cropped to an 21×21 window (*black* corresponds to large correlation magnitudes, *white* to small correlations).

that the subband statistics may vary with the spatial position. An example is given in Figure 3.13 for the *zebra*-texture: even though we would expect that the statistics in different regions of the image would be the same (due to the self-similarity in this texture), this is not true at all! It can be even noted that in the third part of the image, the equiprobability contours are not even elliptical[6]. In this example, the variability of the covariance matrix can be explained by the fact that the correlation between neighboring coefficients depends on the orientation of the present edges in the given region. Non-elliptical equiprobability contours arise when the given region contains edges with different orientations (in this case $+45°$ and $-45°$).[7] A poor directional selectivity is not the only cause of spatial variability of the covariance matrix:

[6]Because this effect can also be caused by the shift variance of the transform (see Section 2.1.4), we used the undecimated wavelet transform in this example, which is shift-invariant.

[7]In this particular example, this could alternatively be solved by using the DT-CWT instead of the undecimated DWT because the DT-CWT allows to distuingish features oriented at $+45°$ from features oriented at $-45°$ (see Section 2.2).

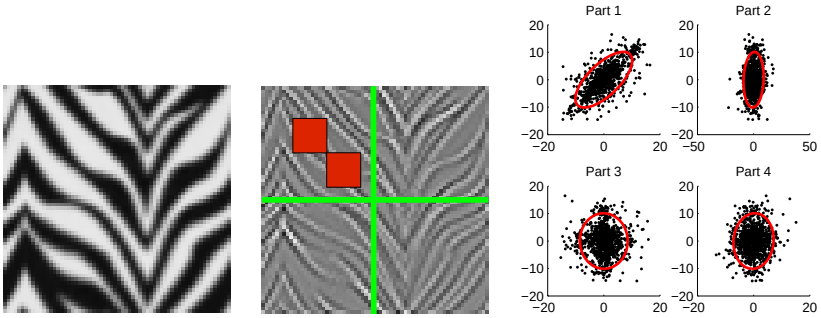


Figure 3.13: Illustration of the spatial variability of the covariance matrix. (a) *Zebra*-texture, (b) HH undecimated wavelet subband, (c) Scatter plots of neighboring coefficients (as indicated in (b)) for different parts of the wavelet subband.

in [Guerrero-Colón, 2008] it is noted that multiresolution transforms with a poor frequency localization often exhibit more variability in the local spectra of the transform subbands than transforms with a good frequency localization. This was illustrated for the translation invariant Haar pyramid [Guerrero-Colón et al., 2007], which uses Haar wavelet filters as radial filter responses.

By taking the variability of the local covariance matrix into account, improvements to the GSM model were obtained. In Spatially Variant GSM (SVGSM) [Guerrero-Colón et al., 2008a], the covariance matrix is estimated *locally* in non-overlapping regions. In Orientation Adaptive GSM (OAGSM) [Hammond and Simoncelli, 2008], the local covariance matrix is adapted to the local dominant orientation. In [Portilla and Guerrero-Colón, 2007], the spatial variability of the covariance matrix is attributed to the fact that texture boundaries in natural images are not sharply defined and that textures may blend into each other. To obtain adaptability, a mixture of Gaussian Scale Mixtures (MGSM) models is proposed in [Portilla and Guerrero-Colón, 2007, Guerrero-Colón, 2008]. By clustering the local covariance matrices globally, the model can also exploit non-local redundancy in images, to some extent.[8] Let $k = 1, ..., K$ again denote the mixture component index and let H_k denote the hypothesis that mixture component k is "the correct GSM model". Then, the likelihood function is given by:

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) &= \sum_{k=1}^{K} f_{\mathbf{x}|H}(\mathbf{x}|k)\,\mathrm{P}(H_k) \quad \text{with} \\ f_{\mathbf{x}|H}(\mathbf{x}|H_k) &= \pi^{-d/2}|\mathbf{C}_{x|k}|^{-1/2} g\left(\left|\mathbf{x}^T\mathbf{C}_{x|k}^{-1}\mathbf{x}\right|^{1/2}\right) \end{aligned} \tag{3.26}$$

[8] We will go deeper into this in Section 3.7.

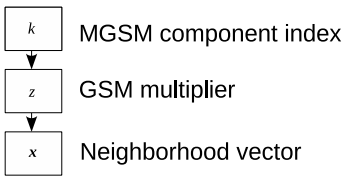


Figure 3.14: Bayesian network structure of MGSM: the MGSM likelihood function can be factored as: $f_{\mathbf{x}}(\mathbf{x}) = f_{\mathbf{x}|z,H}(\mathbf{x}|z,H_k)\, f_{z|H}(z|H_k)\, \mathrm{P}(H_k)$.

where $g(\cdot)$ is given by (3.17). The model parameters:

$$\mathbf{\Theta} = \left\{\mathbf{C}_{x|k}, \mathrm{P}(H_k), \text{ with } k = 1, ..., K\right\},$$

can be trained again from the observed subband, using an EM algorithm (see [Guerrero-Colón et al., 2008b]).

MGSM has two hidden variables per local neighborhood: the scaling variable z, which is proportional to the local variance and the mixture component index, which determines the local covariance matrix. Consequently, MGSM has a causal Bayesian network structure as shown in Figure 3.14, consisting of two layers of hidden variables. Both variables determine the local covariance matrix: suppose that for observation $\mathbf{x}_j$ we have estimates $\hat{z}_j$ and $\hat{k}_j$, then the local covariance matrix at position j, according to the MGSM model, can be computed as:

$$\hat{\mathbf{C}}_{x,j} = \hat{z}_j \mathbf{C}_{x|\hat{k}_j}. \tag{3.27}$$

Similar as for GSM, maximum likelihood estimates can be used to obtain $\hat{z}_j$ and $\hat{k}_j$:

$$\begin{aligned}\left(\hat{z}_j, \hat{k}_j\right) &= \arg\max_{(z,k)} \log f_{\mathbf{x}|z,H}(\mathbf{x}_j|z,H_k) \\ &= \arg\min_{(z,k)} \left[d\log z + \frac{1}{z}\mathbf{x}_j^T \mathbf{C}_{x|k}^{-1}\mathbf{x}_j\right]\end{aligned} \tag{3.28}$$

Solving this optimization problem gives the following global solution:

$$\hat{k}_j = \arg\min_k \mathbf{x}_j^T \mathbf{C}_{x|k}^{-1}\mathbf{x}_j^T \quad \text{and} \quad \hat{z} = \frac{1}{d}\mathbf{x}^T \mathbf{C}_{x|\hat{k}_j}^{-1}\mathbf{x}^T \tag{3.29}$$

The first estimate $\hat{k}_j$ is basically a classifier that selects which spatial covariance matrix $\mathbf{C}_{x|k}$ best describes the data (locally at position j), while the second estimate $\hat{z}$ uses this covariance matrix to compute the local variance.

The number of mixture components K needs to be selected in advance or chosen using greedy EM-type algorithms (e.g. [Vlassis and A., 2002, Verbeek et al., 2003]). In general, K needs to be sufficiently large to capture most variability of the spatial covariance matrix. On the other hand, the number of model parameters increases linear with K, which can cause problems due

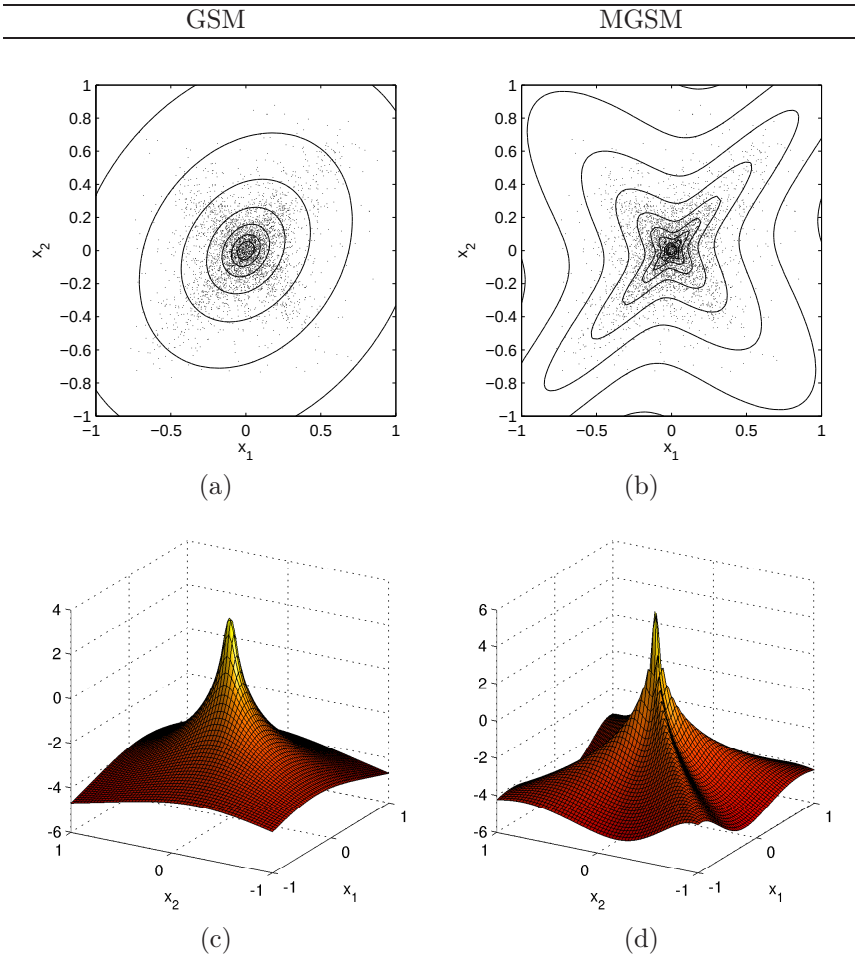


Figure 3.15: Illustration of GSM and MGSM models fitted to the wavelet coefficients from Figure 3.13. (a)-(b) Equiprobability contours of respectively GSM, MGSM. (c)-(d) Logarithm of the pdfs.

to the *"curse of dimensionality"* [Bellman, 1961]. This frequently occurs in smaller subbands of decimated multiresolution transforms, because the number of neighborhood vectors is too small. Also, the computation time for the EM algorithm is relatively large: for example, in [Guerrero-Colón et al., 2008b] it is reported that it takes six hours to denoise a 512×512 image using MGSM with $K = 10$ in a 5-scale Translation Invariant Haar Pyramid (TIHP) transform (*Matlab* implementation).

3.4.4 An improved model: MPGSM

While MGSM is potentially very powerful, as already mentioned there is are two important issues: the unreliable model training due to the excessive number of parameters and the high computation time. For example, for a 5×5 local neighborhood and $K = 10$ MGSM components, the number of parameters is $25 \times 26 \times 10/2 = 3250$!

In this dissertation, we will address these issues by introducing dimension reduction through linear projections in the MGSM model and we will call this model the mixtures of *projected* GSM models (MPGSM). We show that the use of linear projections not only significantly reduces the number of model parameters but also allows us to design fast training algorithms. In this sense, we build upon the MGSM model from [Portilla and Guerrero-Colón, 2007, Guerrero-Colón, 2008, Guerrero-Colón et al., 2008b]. Further, the resulting model can be interpreted as a generalized MGSM model that unifies the SVGSM and OAGSM methods. To reduce the number of free parameters of the MGSM model, we use dimension reduction through linear projections, which is similar to PPCA Section 3.1.2.

In our application, we consider the following linear latent variable model:

$$\mathbf{x}_j = \mathbf{V}\mathbf{t}_j + \bar{\mathbf{V}}\mathbf{r}_j \tag{3.30}$$

where $\mathbf{t}_j$ is a q-dimensional zero mean *GSM random vector*, with covariance $\mathbf{C}_t$, $\mathbf{r}_j$ is $(d-q)$ dimensional zero mean Gaussian distributed residual vector, with diagonal covariance $\mathbf{\Psi}$ and independent of $\mathbf{t}_j$ and $\bar{\mathbf{V}}\mathbf{r}_j = \mathbf{g}_j$. $\mathbf{V}$ is a $d \times q$ matrix, the columns of which are orthonormal basis vectors of the low-dimensional space $\mathcal{V}$. $\bar{\mathbf{V}}$ is a $d \times (d-q)$ matrix, containing the orthonormal basis vectors of the orthogonal complementary subspace $\mathcal{V}^\perp$, such that $\mathcal{W} = \mathcal{V} \oplus \mathcal{V}^\perp$ (here "$\oplus$" denotes the orthogonal direct sum). We remark that $\mathbf{r}_j$ is *not* the image noise, but the approximation error in the complementary space $\mathcal{V}^\perp$.

Using equation (3.30), we can write the covariance matrix of $\mathbf{x}$ as:

$$\mathbf{C}_x = \mathbf{V}\mathbf{C}_t\mathbf{V}^T + \bar{\mathbf{V}}\mathbf{\Psi}\bar{\mathbf{V}}^T \tag{3.31}$$

where $\mathbf{C}_t = \mathrm{E}[z]\,\mathbf{C}_u$ and where $\mathbf{\Psi}$ is assumed to be a diagonal matrix. Because of this diagonality assumption, the number of free parameters in $\mathbf{C}_x$ is significantly reduced when $q \ll d$. Compared to the MGSM model and the GSM model, the proposed MPGSM model adds a third layer of adaptation as depicted in Fig. 3.16. In this conceptual scheme, the first layer is the GSM scaling factor that provides adaptation to the local signal amplitude or variance. The second layer is the MGSM component index, which provides adaptation to signal covariance (textural and edge characteristics). The third layer is added by the proposed model, and it encodes the information inside the covariance matrix more efficiently. In the following paragraphs, we will investigate two different types of projection bases for MPGSM (i.e. the matrices $\mathbf{V}$ and $\bar{\mathbf{V}}$).

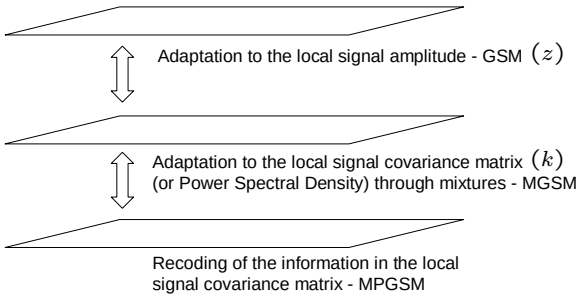


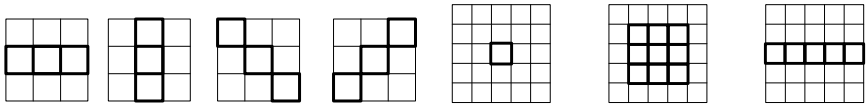
Figure 3.16: Three layers of the MPGSM model

Data-independent bases for MPGSM

A first choice of bases are data-independent bases, which do not depend on the image. The spatial autocorrelation functions from subbands of natural images (see Figure 3.12) reveal that the strongest correlations are along straight lines passing through the center $(0, 0)$, and the direction of these lines is orthogonal to the filtering direction. Our goal is to construct data-independent bases that have a large proportion of the signal in the latent space. A computationally attractive choice are bases made of unit vectors consisting of $d-1$ zeros. This results in simple neighborhood structures, as illustrated in Figure 3.17. For the 3×3 neighborhood structure of Figure 3.17a (left), $\mathbf{V}$ and $\bar{\mathbf{V}}$ are given by:

$$\mathbf{V} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}^T, \bar{\mathbf{V}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}^T \tag{3.32}$$

The main benefit of these simple projection bases is that the dimension reduction is very fast. The covariance matrix $\mathbf{C}_t$ is obtained from $\mathbf{C}_y$ using $\mathbf{C}_t = \mathbf{V}^T \mathbf{C}_x \mathbf{V}$ (see (3.31)), which comes down to simply extracting elements of $\mathbf{C}_x$. Analogously, the diagonal elements of $\mathbf{\Psi}$ are computed as $\mathbf{\Psi}_{ii} = \left[\bar{\mathbf{V}}^T \mathbf{C}_x \bar{\mathbf{V}}\right]_{ii}$. An attractive feature of these projection bases is that one is not limited to neighborhoods of the same size. As illustrated in Figure 3.17, one could e.g. use a 1×1 neighborhood for wavelet coefficients with small (negligible) magnitudes, a 3×3 neighborhood for modeling textures and a 5×1 neighborhood for edges. This limits the number of model parameters but at the same time allows to retain a 5×5 window size globally.



(a) (b)

Figure 3.17: (a) A set of four simple neighborhood structures representing bases of three unit vectors in the Cartesian coordinate system. Each structure models correlations in a specific direction, e.g. the first structure is sensitive to *horizontal* edges, the second structure to *vertical* edges, etc. (b) A set of neighborhood structures with variable sizes and orientations.

Bases of Principal Components for MPGSM

A second choice is to estimate the projection bases from the observed data, e.g., using PPCA. The matrix $\mathbf{V}$ then contains the eigenvectors of the covariance matrix $\mathbf{C}_x$ that correspond to the largest eigenvalues of $\mathbf{C}_x$. The matrix $\mathbf{C}_t$ is diagonal and has the largest (most dominant) q eigenvalues of $\mathbf{C}_x$ as diagonal elements.

To estimate the dimensionality q of the model in a data-driven way, we consider the cumulative proportion of the variance explained by the first q Principal Components [Jolliffe, 1986]:

$$\alpha_q = \sum_{i=1}^{q} \lambda_i / \sum_{i=1}^{d} \lambda_i = \sum_{i=1}^{q} \lambda_i / \mathrm{tr}\left(\mathbf{C}_\mathrm{x}\right) \tag{3.33}$$

where λ_i is the i-th eigenvalue of the covariance matrix $\mathbf{C}_y$. To determine q we select a proportion of the total variance and solve this equation to q numerically. In Fig. 3.18 it can be seen that for common test images, this yields dimension reduction parameters $q \ll d$. For example, if we select $\alpha_q = 88\%$ for the Lena image, we obtain $q = 15 \ll 49$, as illustrated by the solid lines in Fig. 3.18. Other approaches estimate the dimensionality q by looking for a drop in the decrease of the reconstruction error when q increases [Tenenbaum et al., 2000], are based on the eigenvalues of the covariance matrix of samples in a local neighborhood [Verveer and Duin, 1995], or determine q by comparing distances between data vectors [Verveer and Duin, 1995].

MPGSM Model training

As for PPCA Section 3.1.2, we use the EM algorithm to estimate the model parameters. If we denote the mixing weights as $\alpha_k = \mathrm{P}\left(H_k\right)$, the set of model parameters is given by $\mathbf{\Theta} = \{\pi_k, \mathbf{V}_k, \bar{\mathbf{V}}_k, \mathbf{C}_{t,k}, \mathbf{\Psi}_k, k = 1, ..., K\}$ with the constraints $\sum_{k=1}^{K} \pi_k = 1$ and $\mathbf{\Psi}_k$ diagonal. In [Goossens et al., 2009c], it is shown

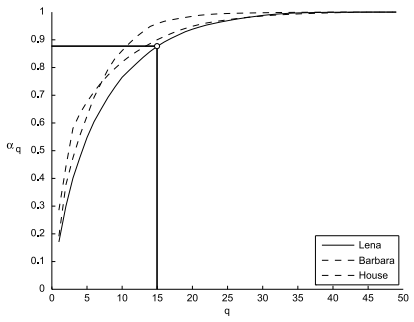


Figure 3.18: The cumulative proportion of the variance α_q explained by the first q Principal Components for all samples in a 7×7 window. In this example, we use the first high-pass band of a steerable pyramid with 8 orientations, for each image in the legend.

that for iteration i, the model parameters can be estimated as follows:

$$\hat{\alpha}_k^{(i)} = \frac{1}{N} \sum_{j=1}^{N} \mathrm{P}\left(H_k | \mathbf{x}_j, \boldsymbol{\Theta}^{(i-1)}\right) \text{ and } \mathbf{S}_k^{(i)} = \frac{\sum_{j=1}^{N} \mathrm{P}\left(H_k | \mathbf{x}_j, \boldsymbol{\Theta}^{(i-1)}\right) \mathbf{x}_j \mathbf{x}_j^T}{\sum_{j=1}^{N} \mathrm{P}\left(H_k | \mathbf{x}_j, \boldsymbol{\Theta}^{(i-1)}\right)}$$

where the posterior probabilities $\mathrm{P}\left(H_k | \mathbf{x}_j, \boldsymbol{\Theta}^{(i-1)}\right)$ (or activities) are computed using Bayes' rule:

$$\mathrm{P}\left(H_k | \mathbf{x}_j, \boldsymbol{\Theta}^{(i-1)}\right) = \frac{\alpha_k^{(i-1)} f_{\mathbf{x}|H,\boldsymbol{\Theta}}(\mathbf{x}_j | H_k, \boldsymbol{\Theta}^{(i-1)})}{\sum_{l=1}^{L} \alpha_l^{(i-1)} f_{\mathbf{x}|H,\boldsymbol{\Theta}}(\mathbf{x}_j | H_l, \boldsymbol{\Theta}^{(i-1)})}. \tag{3.34}$$

For the MPGSM model, the conditional likelihood function $f_{\mathbf{x}|H,\boldsymbol{\Theta}}(\mathbf{x}_j | H_k, \boldsymbol{\Theta}^{(i-1)})$ can be factored as:

$$\begin{aligned} f_{\mathbf{x}|H}(\mathbf{x}_j | H_k) &= f_{\mathbf{t},\mathbf{r}|H}(\mathbf{V}_k^T \mathbf{x}_j, \bar{\mathbf{V}}_k^T \mathbf{x}_j | H_k) \\ &= f_{\mathbf{t}|H}(\mathbf{V}_k^T \mathbf{x}_j | H_k) f_{\mathbf{r}|H}(\bar{\mathbf{V}}^T \mathbf{x}_j | H_k) \qquad (3.35) \\ &= f_{\mathbf{r}|H}(\bar{\mathbf{V}}^T \mathbf{x}_j | H_k) \int_{-\infty}^{+\infty} f_{\mathbf{t}|z,H}(\mathbf{V}^T \mathbf{x}_j | z, H_k) f_z(z) \mathrm{d}z \qquad (3.36) \end{aligned}$$

where $\mathbf{r}|H_k \sim N(\mathbf{0}, \boldsymbol{\Psi}_k)$ and $\mathbf{t}|z, H_k \sim N(\mathbf{0}, z\mathbf{C}_{u,k})$. Next, the ML estimates for $\mathbf{V}_k, \bar{\mathbf{V}}_k, \mathbf{C}_{t,k}$ and $\boldsymbol{\Psi}_k$ are obtained through a diagonalization of the *local activity-weighted* covariance matrix $\mathbf{S}_k^{(i)}$, similar to the straightforward implementation of PPCA (see Algorithm 3.1).

As it is common for most EM algorithms, the algorithm above may converge to poor non-global maxima of the objective function. Therefore, careful parameter initialization of the initial projection bases is required. In [Roweis et al., 2002a, Verbeek, 2006], other non-linear dimension reduction methods are

used to obtain these initial estimates, like the Local Linear Embedding algorithm [Roweis et al., 2002b]. In our experiments (see Chapter 5), we initialize the parameters using a uniform distribution for the mixture weights $\hat{\pi}_k^{(0)} = 1/K$ and initialize the sample covariance matrices heuristically as follows:

$$\hat{\mathbf{S}}_k^{(0)} = \mathrm{E}\left[z\right] \hat{\mathbf{C}}_u \frac{2k}{K+1} + \mathbf{C}_n, \tag{3.37}$$

with the scaling factor $2k/(K+1)$ chosen such that $\sum_{k=1}^{K} \hat{\pi}_k^{(0)} \hat{\mathbf{S}}_k^{(0)} = \mathrm{E}\left[z\right] \hat{\mathbf{C}}_u$, the expected covariance matrix of the signal.

To speed up the EM-algorithm, we investigated approximations. One way to speed up the training phase is by maximizing the log-likelihood for the expected value of the hidden variable z, instead of numerically integrating over all possible z-values (as explained in the Appendix). We also found that an additional significant improvement in computation time can be realized by using a "winner-takes-all" variant of the EM algorithm [Neal and Hinton, 1998]. This comes down to replacing the local activities (3.34) with binary values:

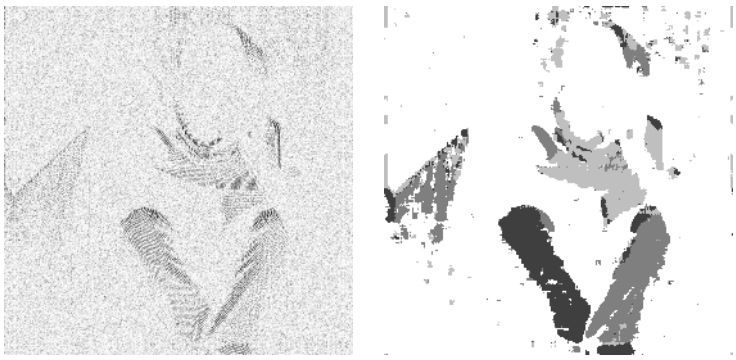
$$\mathrm{P}\left(H_k|\mathbf{x}_j, \boldsymbol{\Theta}^{(i-1)}\right) \approx \begin{cases} 1 & k = \arg\max_{k\in\{1,\ldots,K\}} \mathrm{P}\left(H_k|\mathbf{x}_j, \boldsymbol{\Theta}^{(i-1)}\right) \\ 0 & \text{else} \end{cases} \tag{3.38}$$

Sadly, in the EM context the "winner-takes-all" variant does not necessarily converge to a maximum of the log-likelihood function. However, we can still apply this technique during the first iterations and use the standard approach (3.34) only when the winner-takes-all variant has converged [Neal and Hinton, 1998].

Another advantage of the "winner-takes-all" approach is that the MAP classification in (3.38) can be optimized as follows:

$$\begin{aligned} &\arg \max_{k\in\{1,\ldots,K\}} \mathrm{P}\left(H_k|\mathbf{x}_j, \boldsymbol{\Theta}^{(i-1)}\right) \\ =&\arg \max_{k\in\{1,\ldots,K\}} \pi_k \mathrm{P}\left(\mathbf{x}_j|H_k, \boldsymbol{\Theta}^{(i-1)}\right) \\ =&\arg \max_{k\in\{1,\ldots,K\}} \left(\log \pi_k + \log \mathrm{P}\left(\mathbf{x}_j|H_k, \boldsymbol{\Theta}^{(i-1)}\right)\right) \\ =&\arg \max_{k\in\{1,\ldots,K\}} \left(\log \pi_k^{'} - \sum_{m=1}^{q} \frac{\left(\mathbf{V}_k^T \mathbf{x}_j\right)_m^2}{\left(\mathbf{C}_{t,k}\right)_{m,m}} - \sum_{m=1}^{d-q} \frac{\left(\bar{\mathbf{V}}_k^T \mathbf{x}_j\right)_m^2}{\left(\boldsymbol{\Psi}_k\right)_{m,m}}\right) \end{aligned} \tag{3.39}$$

with $\log \pi_k^{'} = 2 \log \pi_k - \sum_{m=1}^{q} \log\left(\mathbf{C}_{t,k}\right)_{m,m} - \sum_{m=1}^{d-q} \log\left(\boldsymbol{\Psi}\right)_{m,m}$. We particularly note that the terms in the summations in (3.39) are positive. While evaluating this equation, the computations can be stopped whenever the current accumulated sum becomes smaller than the last maximum. In this case, we would never be able to improve the last maximum. To get the most benefit of this trick as possible, we first completely evaluate (3.39) for the mixture component $k^\star$ that we predict to be the most likely. For EM iteration $i = 1$,



(a) (b)

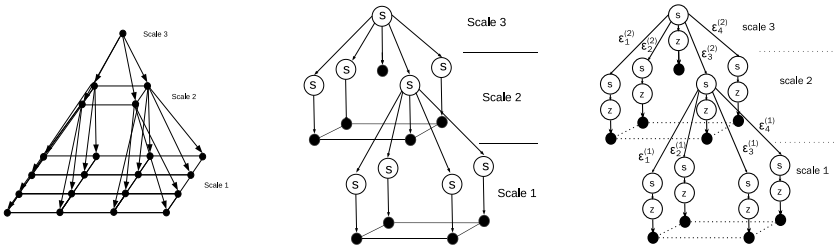
Figure 3.19: (a) Magnitude of a noisy wavelet subband of Barbara. Black corresponds to high magnitudes (b) Label image of the most dominant MPGSM component. The number of MPGSM components is 4 and the size of the neighborhood is 5×5. We used $q = 20$.

we therefore use $k^\star$ that has the highest $\pi_{k^\star}$. For subsequent iterations $i > 1$ we reuse the classification result from the previous estimate. Moreover, we can expect the most benefit if the terms in the summations (3.39) are ordered such that they are decreasing and such that the current maximum is attained as quickly as possible. Because $\mathbf{C}_{t,k}$ and $\mathbf{\Psi}$ usually are obtained from a SVD algorithm that orders the eigenvalues in decreasing order, this automatically is the case.

This way, the EM algorithm fully takes advantage of the linear projections. This technique finds the desired maximum $\mathrm{P}\left(H_k|\mathbf{x}_j, \mathbf{\Theta}^{(i-1)}\right)$ exactly, but in a reduced number of computations. In the best case scenario, it is K times faster; in the worst case (if $\mathbf{y}_j = \mathbf{0}$ and $\pi_k^{'} = \pi_1^{'}, k = 1, ..., K$, which never occurs in practical situations) the computation time remains the same.

To illustrate the effectiveness of the "winner-takes-all" variant of the EM-algorithm for this task, we applied the MPGSM EM algorithm to the Barbara image corrupted with additive white Gaussian noise (with standard deviation $\sigma = 25$). Noise was added to make the training task somewhat more difficult than the noise-free case. In Figure 3.19(a), the magnitude of a wavelet subband of the noisy image is depicted. Figure 3.19(b) shows the index k of the most dominant MPGSM component of each position in the wavelet subband. Even though the noise level is quite high, the method is able to capture the repetitivity in the image: neighborhoods that are similar are also classified as such.

In Chapter 5, we will validate the MPGSM model compared to other GSM models in denoising applications.



(a) (b) (c)

Figure 3.20: Schematic representation of (a) Quad-tree multiscale stochastic process (e.g. [Basseville et al., 1992]), (b) Hidden Markov Tree structure [Crouse et al., 1998], (c) Hidden Markov Tree structure with extra hidden layer, for the proposed joint inter/intra scale model (Section 3.6). Black nodes are coefficients, z-nodes and s-nodes represent respectively the hidden multiplier (local variance) and the significance associated with the coefficients.

3.5 Models for inter-scale dependencies

3.5.1 Hidden Markov Tree models

Multiscale stochastic processes ([Basseville et al., 1992, Wainwright et al., 2001, Banham and Katsaggelos, 1996]) have been studied for modeling these inter-relationships between coefficients in a quad-tree structure of many decimated representations such as the DWT, DT-CWT, STP...). By the decimations, each coefficient at scale 2^i is related to four coefficients at the finer scale 2^{i-1}: the spatial support of a basis element that produced a coefficient at scale 2^i overlaps with the spatial support of four bases elements at the finer scale 2^{i-1}. Because of the overlap of the spatial and frequency supports, dependencies will exist among the coefficients. Often, the studied multiscale processes rely on a Markov relationship between parent and child coefficients in a quad-tree, thereby reducing the number of number of parameters (compared to models that do not use the Markovian assumption).

The Hidden Markov Tree (HMT) [Crouse et al., 1998, Nowak, 1999a, Choi et al., 2000a, Romberg et al., 2001b, Fan and Xia, 2001a, Fan and Xia, 2001b] establishes relationships between hidden state variables rather than between the coefficients themselves. An illustration of different multiscale stochastic processes is shown in Figure 3.20. In Figure 3.20(a), a quad-tree multiscale stochastic process is depicted. Every node (black disks) corresponds to a coefficient; an arrow represents the dependency between two coefficients. The HMT structure (Figure 3.20(b)) is very similar to the quad-tree structure, however, for every scale there is an extra layer of *hidden* state variables (white disks), and the arrows connect the hidden variables rather than the coefficients.

Next, we will briefly summarize the HMT model. Let $\mathbf{x}_j^{(i)}$, $j = 1, ..., N_i$ represent the coefficients of a local window at position j and scale i (here, N_i is the number of coefficients at scale i). The HMT model is characterized by:

1. Two possible states for each scale i: $H_0^{(i)}$ and $H_1^{(i)}$, and their corresponding state probabilities $\mathrm{P}\left(H_0^{(i)}\right)$ and $\mathrm{P}\left(H_1^{(i)}\right) = 1 - \mathrm{P}\left(H_0^{(i)}\right)$. As in Section 3.2.2, $H_1^{(i)}$ denotes the hypothesis that a coefficient is *significant* (and has a large magnitude), while $H_0^{(i)}$ signifies the hypothesis that the coefficient is *non-significant* (with a small magnitude).

2. Two conditional densities for each scale: $f(\mathbf{x}^{(i)}|H_0^{(i)})$ and $f(\mathbf{x}^{(i)}|H_1^{(i)})$. This results in the overall PDF:

$$f(\mathbf{x}^{(k)}) = \mathrm{P}\left(H_0^{(i)}\right) f(\mathbf{x}^{(k)}|H_0^{(i)}) + \mathrm{P}\left(H_1^{(i)}\right) f(\mathbf{x}^{(k)}|H_1^{(i)}).$$

   The densities $f(\mathbf{x}^{(k)})$ are assumed to be mutually independent for each scale; their dependency is imposed through the hidden state variables.

3. The state transition probability distributions $\boldsymbol{\epsilon}^{(k)} = \{\epsilon_{m,n}^{(k)}\}$:

$$\epsilon_{m,n}^{(i)} = \mathrm{P}\left(H_n^{(i+1)}|H_m^{(i)}\right) \quad \text{with } m = 0, 1, \quad n = 0, 1.$$

   The parameters $\epsilon_{0,0}^{(i)}$ and $\epsilon_{1,1}^{(i)}$ are called *persistency* probabilities; $\epsilon_{0,1}^{(i)}$ and $\epsilon_{1,0}^{(i)}$ are *novelty* probabilities. These parameters express the probabilities that the hidden state values will change from one scale to the next.

The parameters $\mathrm{P}\left(H_0^{(i)}\right), \mathrm{P}\left(H_1^{(i)}\right), \boldsymbol{\epsilon}^{(i)},\ i = 2, ..., I$ are estimated iteratively using the Baum-Welch algorithm (also known as the Expectation Maximization (EM) algorithm for HMM's) [Crouse et al., 1998, Rabiner, 1989].

Commonly mentioned problems are [Pižurica and Philips, 2003]: (1) a large number of model parameters (despite the Markovian assumption) (2) model training using Baum-Welch can be relatively slow [Romberg et al., 2001b] and (3) the lack of spatial adaptation. In this respect, a local contextual HMT model of Fan et al. [Fan and Xia, 2001a] is an improvement: the authors attach an additional hidden state to every coefficient; this additional hidden variable is the local average energy of the surrounding coefficients.

3.5.2 The Bivariate distribution of Şendur and Selesnick

[Şendur and Selesnick, 2002b] focus on the dependencies between a coefficient and its parent in detail. Based on empirical joint histograms of parent and child coefficients (as in Figure 3.10), the authors propose an elliptically symmetric bivariate probability density function to model the dependency between a parent and child coefficient:

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{3}{2\pi\sigma_1\sigma_2} \exp\left(-\sqrt{3}\left(\left(\frac{x_1}{\sigma_1}\right)^2 + \left(\frac{x_2}{\sigma_2}\right)^2\right)^{1/2}\right), \tag{3.40}$$

where $\mathbf{x}$ here denotes a two-dimensional vector that contains both parent and child coefficients ($\mathbf{x} = [x_1, x_2]$). In this model, x_1 and x_2 are uncorrelated but not statistically independent. We remark in this respect that (3.40) is a special case of the EPD distribution from Section 3.2.3, with $\nu = 1$ and a diagonal covariance matrix $\mathbf{C}_x$. Further, the authors derive the MAP estimator in a denoising application, for additive white Gaussian noise. Despite the simplicity of the model, a better denoising performance was demonstrated for some images in the DT-CWT domain compared to the HMT model. We will use this bivariate method as a reference for comparison in Chapter 5.

3.6 A novel joint inter/intra-scale model

In the HMT model of Crouse et al. [Crouse et al., 1998], later extended by Romberg et al. [Romberg et al., 2001b], use a weighted mixture of two *univariate* Gaussian distributions (Section 3.2.2) for the marginal densities of the coefficients. The number of mixture components is directly related to the number of states: one mixture component corresponds to each state. To describe the leptokurtotic behavior of the coefficients more accurately, a larger number of Gaussian mixture components (e.g. 8) may be necessary.[9] However, this inherently increases the number of model parameters and subsequently the computational complexity. In [Kivinen et al., 2007], non-parametric HMT models, connecting discrete GSM distributions across states, are trained using a Monte Carlo learning algorithm. The number of states is also learned from the training Markov Chain Monte Carlo methods can be designed to escape from local maxima and saddle points of the likelihood function (see e.g. [Gamerman, 1997]). However, the computational cost is often significant, which makes these methods less practical.

The HMT models from [Crouse et al., 1998, Nowak, 1999a, Choi et al., 2000a, Romberg et al., 2001b] and the bivariate model from [Şendur and Selesnick, 2002b] do not capture *intra*-scale correlations of coefficients. More importantly, the models assume that the coefficients within the same subband are *statistically independent*, despite the fact that multiresolution transforms are not capable to *fully* decorrelate images (see Section 3.3). To further improve the results, [Fan and Xia, 2001a] and [Şendur and Selesnick, 2002a] independently include a spatial activity indicator in their models. The spatial activity indicator being used is in both cases an estimate of the local variance.

In this section, we present a *new* joint inter/intra-scale statistical model. The main idea is to model the probability of signal presence (i.e. significance of a coefficient) given a *vector* of surrounding coefficients, i.e., given a *structure* of the local neighborhood, thereby taking the true correlations between the coefficients into account. This is achieved by introducing an extra hidden parameter in the GSM model, that models signal presence. Further, the approach is combined with a HMT model to capture the inter-scale coefficient

[9] Note that for an infinite number of mixture components, the mixture becomes a GSM, see Section 3.2.4.

dependencies, yielding a joint inter/intra-scale model.

Our approach is on the one hand an improvement and generalization of the main ideas of [Pižurica and Philips, 2006] where the estimation of probability of signal presence is now improved, and where the estimator is combined with a powerful HMT model. On the other hand, this work can also be seen as an improvement and generalization of the HMT approaches of [Crouse et al., 1998, Choi et al., 2000b, Romberg et al., 2001b, Fan and Xia, 2001a], where we now employ a better likelihood model and a better estimation of the involved state probabilities.

Alternatively, our joint inter/intra-scale model may be combined with the MPGSM model from Section 3.4.4 as well. Because the extension of our approach to mixtures of GSMs is straightforward, we only discuss the GSM model here.

Modeling the signal of interest

Significant coefficients can be characterized by means of a *significance measure*, based on the magnitude of the considered wavelet coefficient [Pižurica and Philips, 2006]:

$$S(x) = \mathrm{I}\left(|x|/\sigma_w \geq T\right) \tag{3.41}$$

where σ_w is the noise standard deviation, T is a given threshold and $\mathrm{I}(x)$ is the indicator function. In our work, we extend (3.41) to vectors of neighboring coefficients, by the following generalization:

$$S(\mathbf{x}) = \mathrm{I}\left(\left\|\mathbf{C}_w^{-1/2}\mathbf{x}\right\| \geq T\right) \tag{3.42}$$

where $\mathbf{C}_w^{1/2}$ is the square root of the positive definite matrix $\mathbf{C}_w$ and $||\mathbf{x}||$ is the norm of $\mathbf{x}$. By the positive-definiteness of $\mathbf{C}_w$, $\left\|\mathbf{C}_w^{-1/2}\mathbf{x}\right\|^2 = \mathbf{x}^T\mathbf{C}_w^{-1}\mathbf{x} = T^2$ represents the equation of an ellipsoid in a d-dimensional space. The significance measure (3.42) then tests whether $\mathbf{x}$ is inside or outside the ellipsoid. This is illustrated in Fig. 3.21.

According to the significance measure (3.42), the conditional density $f_{\mathbf{x}|H}\left(\mathbf{x}|H_0\right)$ is given by:

$$f_{\mathbf{x}|H}\left(\mathbf{x}|H_0\right) = \frac{f_{\mathbf{x}}\left(\mathbf{x}\right)}{\mathrm{P}\left(H_0\right)}\mathrm{I}\left(\left\|\mathbf{C}_w^{-1/2}\mathbf{x}\right\| < T\right) \tag{3.43}$$

and an analogous expression can be given for $f_{\mathbf{x}|H}\left(\mathbf{x}|H_1\right)$. Now, to take intra-scale correlations into account, we use a GSM distribution, or more specifically, the multivariate Bessel K Form distribution (see Section 3.2.5) as PDF for $\mathbf{x}$. The resulting conditional densities are depicted in Figure 3.22 for a diagonal matrix $\mathbf{C}_w$.

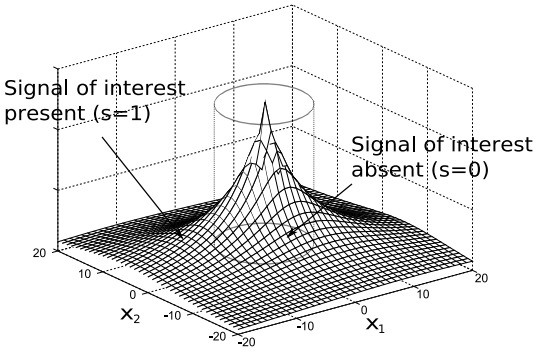


Figure 3.21: Illustration of "signal of interest" on the spatial prior $f_{\mathbf{X}}(\mathbf{x})$. The ellipse $\mathbf{x}^T\mathbf{C}_w^{-1}\mathbf{x} = T^2$ is extruded to a cylinder, for visibility. Samples $\mathbf{x}$ outside the cylinder are regarded as *significant*. *Non-significant* samples are contained in the cylinder.

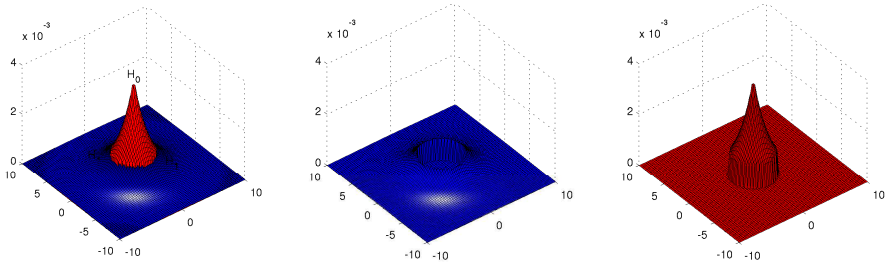


Figure 3.22: Multivariate mixture models: (a) $f_{\mathbf{x}}(\mathbf{x}) = f_{\mathbf{x}|H}(\mathbf{x}|H_0)\,\mathrm{P}(H_0) + f_{\mathbf{x}|H}(\mathbf{x}|H_1)\,\mathrm{P}(H_1)$, (b) Signal present $f_{\mathbf{x}|H}(\mathbf{x}|H_1)$, (c) Signal absent $f_{\mathbf{x}|H}(\mathbf{x}|H_0)$.

Hidden Markov Tree model based on the signal of interest modeling

In our joint inter/intra-scale model, the significance measures $S(\mathbf{x})$ are used as hidden nodes for the HMT model. Because this only requires two states, independent of the number of Gaussian mixture components, this reduces the computational complexity of the HMT training procedure while retaining a highly kurtotic distribution $f_{\mathbf{x}}(\mathbf{x})$. We use independent HMT models for the different *orientations* of the DT-CWT. The HMT structure of our model comprises two layers of hidden variables (a first layer for the significance variables, a second layer for the GSM multipliers), see Figure 3.20(c).

For the HMT model, the significance of the coefficients at a given scale i can be estimated through the posterior probabilities $f_{H|\mathbf{x}^{(i)},\ldots,\mathbf{x}^{(I)}}\left(H_k|\mathbf{x}^{(i)},\ldots,\mathbf{x}^{(I)}\right)$, $k \in \{1,2\}$, through the MAP estimate, as follows:

$$\hat{S}_{\mathrm{HMT}}\left(\mathbf{x}^{(i)}\right) = \arg\max_{k\in\{1,2\}} f_{H|\mathbf{x}^{(i)},\ldots,\mathbf{x}^{(I)}}\left(H_k|\mathbf{x}^{(i)},\ldots,\mathbf{x}^{(I)}\right). \tag{3.44}$$

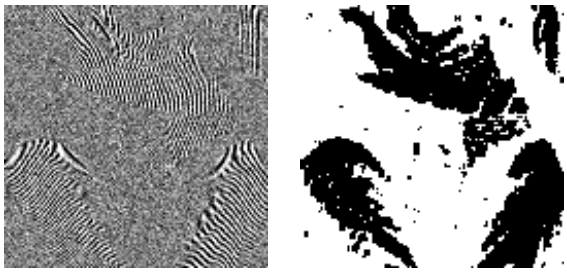


Figure 3.23: Detection of "signal presence". (*Left*) Cropout of a noisy wavelet band (HH_1) of the Barbara image (with added Gaussian noise with covariance $\mathbf{C}_w = 25^2\mathbf{I}_d$) (*Right*) MAP estimate of the significance.

Here $\mathbf{x}_j^{(i)}$ signifies the coefficients of a local window centered at position $j \in \{1, ..., N_i\}$ and scale $i \in \{1, ..., I\}$. Clearly, the estimate of the significance for a coefficient vector at scale i is conditioned on all coefficients on parent (coarser) scales. In Figure 3.23, the significance of *noise-free* wavelet coefficients is estimated for one subband of the DT-CWT transform, in the presence of additive Gaussian noise. Although for this example, the observation density is in fact the convolution of the PDF of $\mathbf{x}$ with the noise PDF. Therefore we applied the same principle as in (3.44), but we conditioned on the observed *noisy* variables instead of $\mathbf{x}^{(i)}, ..., \mathbf{x}^{(I)}$. It can be seen that despite the high variance of the added noise, the significance estimates are quite accurate. This is because the combination of the spatial GSM model and the HMT tree model allows us to capture *both* spatial and interscale dependencies between wavelet coefficients. In Section 5.2.3 we will present a denoising method based on this joint inter/intra-scale model.

3.7 Non-local image models

A completely different approach for modeling images is to exploit their self-similarity: many details, patterns or features occur several times in the same image. An example is given in Figure 3.24 where details of the windows and exterior shutters are present multiple times.

Already since the 80s, the self-similarity of images has played an important role in image processing, and has led to the development of *fractal*-based compression schemes (see e.g. [Jacquin, 1992]). Similarity between consecutive frames in a video sequence is also the basis of many motion estimation schemes for video compression (e.g. [Zhang and Zafar, 1992, Stiller, 1997, Wang and Ostermann, 1998]).

Concerning the content of images, we can distinguish 1) similaries between *patterns* or *features* within the same object or in different objects, 2) similarities *across edges* of objects and 3) similarities in *uniform* regions of an image [Luong, 2009]. In a multiresolution representation of the image, similaries can

Figure 3.24: Illustration of the self-similarity in images. The small crosses indicate patches that are similar to the 8×8 patch with center marked with the big cross, in the mean squared difference sense.

occur within the same scale or across similar scales (e.g. similar objects of different sizes), within the same analysis orientation or across different analysis orientations (e.g. similar rotated objects).

Recently, a number of non-local methods have been developed to take advantage of the similarity of small patches in the image at the same scale [Elad and Aharon, 2006b, Dabov et al., 2007, Luong et al., 2006], often yielding better results than existing local methods. Most notably is the non-local means (NLMeans) filter [Buades. et al., 2005, Buades et al., 2008], in which a pixel intensity is estimated as a weighted average of *all* pixels in the image, where the weights are proportional to the similarity between the local neighborhood of the pixel being processed and local neighborhoods of the surrounding pixels. Although the NLMeans filter has traditionally been introduced in the context of image denoising, we will review a number of aspects of the underlying image model that the filter relies on, in this section. Studying this non-local image model also allows to envisage extensions of the filter to other image restoration applications (see Chapter 5).

Let us denote $\mathbf{x}_j$ as the vector of pixel intensities of a local $\sqrt{d} \times \sqrt{d}$-neighborhood centered at position j, where we will typically use overlapping neighborhoods. Here, we are considering neighborhoods in the image domain. Recall that in the previous sections, we often assumed that $\mathbf{x}_j$ at different positions j are statistically independent. To take advantage of the "possible" similarity between different neighborhoods, non-local methods need to drop

Table 3.2: Overview of multivariate robust M-functions, with h a smoothing parameter.

Type	Weighting function $g(\mathbf{x})$	Robust function $\rho(\mathbf{x})$
Cauchy	$1/\left(1+\left\|\mathbf{x}\right\|^2/h^2\right)$	$\frac{1}{2}h^2\log\left(h^2+\left\|\mathbf{x}\right\|^2\right)$
Tukey	$\begin{cases}(1-\frac{\left\|\mathbf{x}\right\|^2}{h^2})^2 & \left\|\mathbf{x}\right\|\leq h\\ 0 & \left\|\mathbf{x}\right\|>h\end{cases}$	$\begin{cases}\frac{h^2}{6}\left[1-\left(1-\frac{\left\|\mathbf{x}\right\|^2}{h^2}\right)^3\right] & \left\|\mathbf{x}\right\|\leq h\\ \frac{h^2}{6} & \left\|\mathbf{x}\right\|>h\end{cases}$
Andrews	$\begin{cases}\frac{\sin(\pi\left\|\mathbf{x}\right\|/h)}{\pi\left\|\mathbf{x}\right\|/h} & \left\|\mathbf{x}\right\|\leq h\\ 0 & \left\|\mathbf{x}\right\|>h\end{cases}$	$\begin{cases}\frac{h^2}{\pi^2}\left(1-\cos\left(\pi\left\|\mathbf{x}\right\|/h\right)\right) & \left\|\mathbf{x}\right\|\leq h\\ \frac{h^2}{\pi^2} & \left\|\mathbf{x}\right\|>h\end{cases}$
Leclerc	$\exp\left(-\frac{\left\|\mathbf{x}\right\|^2}{2h^2}\right)$	$h^2-h^2\exp\left(-\frac{\left\|\mathbf{x}\right\|^2}{2h^2}\right)$
Bisquare	$\begin{cases}\left(1-\frac{\left\|\mathbf{x}\right\|^2}{h^2}\right)^2 & \left\|\mathbf{x}\right\|\leq h\\ 0 & \left\|\mathbf{x}\right\|>h\end{cases}$	$\begin{cases}\frac{h^2}{6}\left(\frac{\left\|\mathbf{x}\right\|}{h}-1\right)^3\left(\frac{\left\|\mathbf{x}\right\|}{h}+1\right)^3 & \left\|\mathbf{x}\right\|\leq h\\ 0 & \left\|\mathbf{x}\right\|>h\end{cases}$
Modified Bisquare	$\begin{cases}\left(1-\frac{\left\|\mathbf{x}\right\|^2}{h^2}\right)^8 & \left\|\mathbf{x}\right\|\leq h\\ 0 & \left\|\mathbf{x}\right\|>h\end{cases}$	$\begin{cases}\frac{h^2}{18}\left(\frac{\left\|\mathbf{x}\right\|}{h}-1\right)^9\left(\frac{\left\|\mathbf{x}\right\|}{h}+1\right)^9 & \left\|\mathbf{x}\right\|\leq h\\ 0 & \left\|\mathbf{x}\right\|>h\end{cases}$

this statistical independence assumption. However, because similarities occur across the whole image, further assumptions are needed to arrive at a tractable probability density model in which the parameters can be estimated from the image itself. As we already encountered in previous sections, the number of parameters can be reduced by either an independence assumption (e.g. MRF models) or by imposing a particular structure to the covariance matrix of the model (e.g. MPGSM). In this case, both approaches are not an option, since we really want to take these dependencies into account. Consequently, proposing a *correct* PDF model capturing non-local dependencies while allowing for easy parameter estimation from a single image is a challenging task (as far as we are aware of, such a model has not been proposed in the literature so far). Instead, a number of methods rely on clustering techniques, based on block-matching [Dabov et al., 2007], k-means or k-svd clustering [Aharon et al., 2006]. Here, we will impose a specific likelihood function to the neighborhood vectors, similar as in MRF models (Section 3.4.1):

$$f_{\mathbf{x}_1,...,\mathbf{x}_N}\left(\mathbf{x}_1,...,\mathbf{x}_N\right)=\frac{1}{A}\exp\left(-\frac{1}{T}\sum_{j,j'=1}^{N}\rho\left(\mathbf{x}_{j'}-\mathbf{x}_j\right)\right) \tag{3.45}$$

where A is a PDF normalization constant, T is a "temperature" constant and $\rho(\cdot)$ is a multivariate robust loss function. The robust loss function assigns a cost to the difference $\mathbf{x}_{j'}-\mathbf{x}_j$: typically, the lower the norm of the difference $\left\|\mathbf{x}_{j'}-\mathbf{x}_j\right\|$ is, the higher the likelihood function (3.45) and vice versa. A few robust loss functions are tabulated in Table 3.2.

In [Goossens et al., 2008a], we have shown that for the Leclerc robust loss function, the use of the distribution (3.45) for modeling a noisy image in a denoising application, directly leads to an iterative estimator that corresponds to the NLMeans filter from [Buades. et al., 2005]. The first iteration of this

estimator is given by:

$$\widehat{[\mathbf{x}_j]}_c = \frac{\sum_{j'=1}^{N} g\left(\mathbf{y}_{j'} - \mathbf{y}_j\right) \mathbf{y}_{j'}}{\sum_{j'=1}^{N} g\left(\mathbf{y}_{j'} - \mathbf{y}_j\right)} \tag{3.46}$$

with $\mathbf{y}$ is an observed degraded image, $[\cdot]_c$ denoting the central coefficient of the neighborhood and with the weighting function $g(\cdot)$ defined by $\mathbf{x}g(\mathbf{x}) = \partial\rho/\partial\mathbf{x}$. Hence, the "denoised" neighborhood $\hat{\mathbf{x}}_j$ is simply the weighted average of all neighborhoods in the image. As can be noted from Table 3.2, the weighting function is characterized by a parameter h that in some cases (e.g. bisquare) also serves as a threshold: if the Euclidean distance between two neighborhoods is smaller than h, the two neighborhoods are considered to be similar, otherwise the neighborhoods are dissimilar. In that case, only a limited number of terms will have a non-zero weight in (5.5), which is beneficial 1) to avoid the contribution of many dissimilar neighborhoods in the averaging and 2) to reduce the computation time [Dauwe et al., 2008]. We will discuss our improvements to the NLMeans filter in Chapter 5.

An interesting question is how to incorporate *multiresolution concepts* into the non-local model. Answers to this question are part of recent ongoing research. For example, in [Hammond et al., 2009] it is proposed to use graph theory: the authors associate a graph $G = (V, E, g)$ with each image, where the vertices are given by $V = \{\mathbf{x}_j, j = 1, ..., N\}$, the edges connect different nodes $E \subset V \times V$ and $g(\cdot)$ is the weighting function. Associated to the graph is the function $f : V \to \mathbb{R}$, which defines the pixel intensity for every vertex of the graph. Next, the unnormalized discrete graph Laplacian operator of f is defined as:

$$\left(\Delta f\right)(j) = \sum_{j'=1}^{N} g\left(\mathbf{x}_{j'} - \mathbf{x}_j\right)\left(\mathbf{x}_{j'} - \mathbf{x}_j\right). \tag{3.47}$$

Based on the spectral decomposition of the discrete graph Laplacian, the graph analogue of a Fourier transform is defined. Next, a "graph wavelet transform" is obtained by modulating the eigenvalues of the Laplacian operator (which can be seen as applying wavelet filtering in the graph Fourier domain). It was found that the resulting "graph" wavelet basis functions are generally localized both in frequency and in space, and are approximately radially symmetric. Based on the weighting function $g(\cdot)$, the basis functions effectively capture the similarity of features and patterns in images.

3.8 Conclusion

Despite all the efforts that have been done in the past for modeling the statistics of images, building an accurate probability density model for the general class of *natural* images remains a challenging task. In this chapter, we have first reviewed a number of image decomposition techniques (that are due to computational constraints mostly applied to patches): PCA, PPCA and ICA.

Marginal parametric densities directly model the highly kurtotic behavior of band-pass filtered coefficients, for either projections onto ICA components or for image-independent multiresolution transforms from Chapter 2. Next, we have shown that linear filters are not able to completely decorrelate the images, hence incorporating intra-scale or inter-scale correlations into the model is crucial. We have discussed a number of intra-scale models, such as Markov Random fields, GSM, MGSM and MPGSM and inter-scale models such as HMTs that can be used to this end. Then, we have presented a new joint inter/intra-scale model that combines ideas from earlier MRF and HMT approaches and jointly models the second order statistics of the coefficients and the dependencies between different multiresolution scales. Finally, we investigated the use of non-local information to exploit the self-similarity in images.

Another important aspect, for statistical image models, is that the model should be easily applicable as "prior knowledge" in practical circumstances. To demonstrate that this is the case for the novel models, we will further use these models in image restoration applications in Chapter 5.

The contributions of this chapter already resulted in the following publications: [Goossens et al., 2007b, Goossens et al., 2009c] (on MPGSM), [Goossens et al., 2007c, Goossens et al., 2009d] (on the joint inter/intra-scale model), [Dauwe et al., 2008, Goossens et al., 2008a] (on the non-local image model).

4

Noise modeling and estimation

Many digital imaging devices often produce a substantial amount of noise. The noise is originating from the analog circuitry (sensors, amplifiers) in the devices. Digital imaging techniques need to deal with the noise present in the images, which occasionally can lead to failure of these techniques. A first solution to this problem is to make the methods *more robust* against noise; a second solution is to apply *noise suppression* (colloquially known as denoising) as pre-processing step. In both solutions, an accurate noise model is indispensable: the more pre-knowledge about the noise we can build into the imaging technique, the better its performance will be.

In contrast to all the research that describes properties of natural images (see Chapter 3), relatively little research has gone into describing noise or noise properties. Also, the majority of digital image processing techniques are only capable of efficiently dealing with white Gaussian noise in its basic form.

First, let us consider noise in digital still cameras (DSC). The imaging pipeline of a DSC is shown in Figure 4.1.The incident light reaches an array of CCD or CMOS sensor elements through the lens. Subsequently, the measured light intensity signals are amplified electronically and converted to digital signals in the analog-to-digital (A/D) convertor. Next, the DSC performs a number of post-processing techniques, mainly to increase the quality of the images: white balance correction, gamma correction, color enhancement, contrast enhancement, digital zoom and noise reduction... Finally, the image is compressed to be put on a storage medium (e.g. flash).

Because the noise originates from the image sensor elements, it is clear that every component in this pipeline will influence the noise characteristics in the final (reconstructed) image. We remark that the noise reduction techniques integrated in DSCs are of low complexity (because of the limited amount of resources on the camera, e.g. memory and power usage) and consequently, the quality improvement of the processed images may be limited.

Furthermore, there is a trend to increase the number of image pixels and images consisting of 10 million pixels (10 mega pixel) are now common. Unfor-

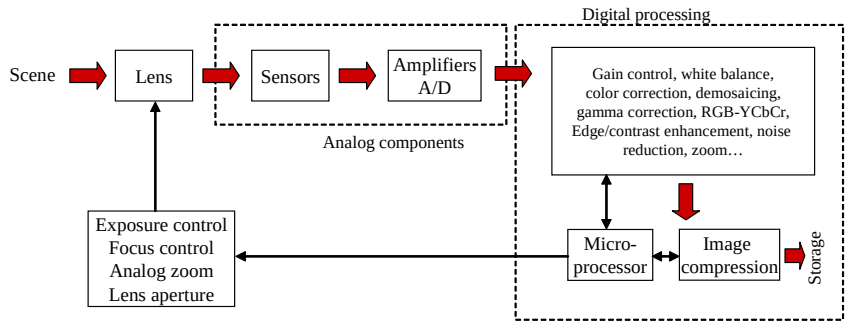


Figure 4.1: Pipeline of a digital still camera (DSC)

tunately, this inherently reduces the area of each sensor element and increases the amount of noise captured by each sensor.
Noise produced by DSCs has various origins [Nakamura, 2005]:

1. *Sensor noise (photon noise or shot noise)*: this type of noise is caused by the fluctuations of the detected photons, by the sensors of the DSC. By quantum-mechanical effects, sensor noise can not be avoided. Sensor noise is mostly noticable when taking photographs using short exposure time settings and in dark lighting conditions.

2. *Read-out noise*: noise generated by the electrical circuits (e.g. amplifiers) in the camera.

3. *Pattern noise*: refers to the correlated components of the read-out noise due to imbalances of the pixel readout. The human eye is very sensitive for patterns in the noise of a digital photographs. This kind of noise typically appears in the form of horizontal and vertical banding noise.

4. *Thermal (or dark) noise*: many CCD camers produce a noise signal even when there is no incident light present. The noise level increases gradually from the beginning to the end of the readout. Fortunately, this type of noise is non-random and the readout can be compensated by subtracting an offset that depends on the vertical position of the sensor element.

5. *Sensor cross-talk noise*: noise caused by interaction between neighboring sensor elements (by electron and photon leakage) [Hirakawa, 2008a].

Other imaging devices (such as medical scanners) also very often need to deal with noise in their reconstruction algorithms: for example in CT, measured x-ray intensities also inherently contain noise components that very often create disturbing streaking artifacts in the reconstructed images, especially for low radiation doses. In Magnetic Resonance Imaging (MRI), received radio frequency signals also contain statistical fluctuations, which can not be avoided.

Because the noise in digital images can have various origins, we will categorize different noise characteristics into different classes that can be studied individually:

- *The marginal distribution of the noise*: in many practical applications the marginal noise distribution is well approximated by a Gaussian distribution. Another distribution that is often used is the Poisson distribution (e.g. in medical imaging, microscopy).
- *The second order statistics of the noise*: noise can be spatially uncorrelated (*white noise*) or correlated (*colored noise*).
- *The stationarity of the noise process*: noise processes can be either *stationary* or *non-stationary*. In case of stationary noise, the noise statistics (such as the variance) are invariant to the position in the image. For non-stationary noise, the noise characteristics depend on the position in the image. An example is noise in Computed Tomography.
- *Signal dependency of the noise process*: the noise component can be either *additive*, or *signal-dependent*. In the former case, the observed image is simply the *sum* of the original image and a noise component. In the latter case, there exists dependencies between the noise samples and the original image.

In the remainder of this chapter, we will focus on the specification of parametric noise models and the corresponding estimation of the noise model parameters. In particular, we will present novel EM algorithms for the estimation of stationary correlated noise in Section 4.2.2 and non-stationary correlated noise in Section 4.3. We establish new approximate and exact analytical relationships between the camera response function and the noise level function, which is useful for modeling and estimating signal-dependent noise in images (Section 4.4).

The presented noise models will be further used in the chapter on image restoration (Chapter 5).

4.1 Probability density functions for modeling noise

Let y denote a pixel intensity at a given position of an observed image, and let x signify the corresponding pixel intensity of the original (*ideal*) image. For simplicity of the notations, we omit the position index and assume that all pixel intensities in the image are independent and identically distributed. The PDF that is most oftenly used for describing the noise in y, is the Gaussian distribution:

$$y|x \quad \sim \mathcal{N}\left(x, \sigma^2\right) \quad \Leftrightarrow f_{y|x}\left(y|x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y-x\right)^2}{2\sigma^2}\right).$$

Additive white Gaussian noise (AWGN) has generally been found to be a reasonable model for noise originating from electronic amplifiers (see e.g. [Sarpeshkar et al., 1993, Lim, 2006]).

When measuring light intensities from a single source, statistical fluctuations will be observed. According to quantum mechanics, the measurement of light intensity can be interpreted as a spatio-temporal integration, for which the "total number" of photons emitted by the source in the considered spatio-temporal interval is often assumed to be Poisson distributed:

$$y \sim \text{Poisson}(x) \Leftrightarrow f_{y|x}(y|x) = \frac{x^y \exp(-x)}{y!}.$$

Poisson noise is unbiased: it does not alter the intensity mean (i.e. $\text{E}[y] = \text{E}[x]$). Also, Poisson noise has a variance that increases linearly with the original intensity x ($\text{Var}[y] = x$), the surface area of the sensor elements and the integration (or exposure) time. Poisson noise arises in digital cameras (sensor noise), in medical and nuclear imaging (e.g. CT) and in microscopy [Rooms, 2005]. Consequently, for small sensor elements or short integration times, the Signal-to-noise ratio (SNR) will be low [Hirakawa, 2008a]. On the other hand, for sufficiently long integration times (i.e. for high SNRs), the Poisson distribution can be well approximated by a Gaussian distribution.

Another type of noise encountered in Magnetic Resonance Imaging (MRI), is Rician noise. In MRI, there is a trade-off between SNR and image resolution [Pižurica, 2002, p. 161]. In practice, the acquisition time is also limited for the comfort of the patient and to avoid patient motion. The main noise source in MRI images is thermal noise in the patient [Edelstein et al., 1986, Pižurica, 2002].

In general, the intensities y of the reconstructed image are assumed to follow an uncorrelated complex-valued Gaussian distribution, with mean x and variance σ^2. However, MRI magnitude images, which are obtained by taking the magnitude of y, are most commonly used. Consequently, $y' = |y|$ follows a Rice distribution:

$$y' \sim \text{Rice}(x') \Leftrightarrow f_{y'|x'}(y'|x') = \frac{y'}{\sigma^2} \exp\left(\frac{-\left(x'^2 + y'^2\right)}{2\sigma^2}\right) I_0\left(\frac{x'y'}{\sigma^2}\right),$$

where $x' = |x|$ and with $I_0(\cdot)$ the modified Bessel function of the first kind and zero order. In contrast to Poisson noise processes, Rice noise processes *do* alter the intensity mean, as:

$$\text{E}[y'|x'] = \sigma\sqrt{\frac{\pi}{2}} M\left(-\frac{1}{2}, 1, -\frac{x'}{2\sigma^2}\right) \neq x',$$

where $M(\cdot,\cdot,\cdot)$ is the confluent hypergeometric function [Abramowitz and Stegun, 1964]. The bias introduced by Rician noise is generally undesired, as the bias reduces constrast between bright and dark areas in the image [Aelterman et al., 2008]. For this reason, several bias removal techniques have been proposed for MRI, e.g. [Sled et al., 1998, Van Leemput et al., 1999].

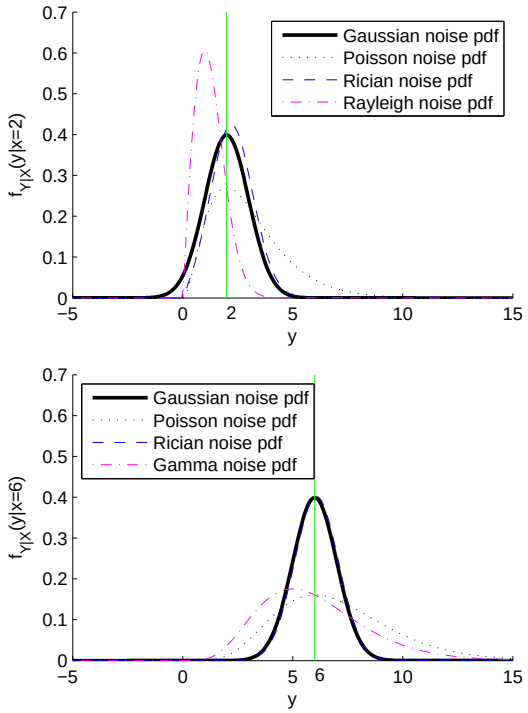


Figure 4.2: Marginal probability density functions (with parameter $\sigma = 1$).

The Rayleigh distribution is a special case of the Rice distribution and arises for low intensities in MRI images (i.e. when the mean $x' \approx 0$):

$$y' \sim \text{Rayleigh}\,(x') \Leftrightarrow f_{y'|x'}\,(y'|x') = \frac{y'}{\sigma^2} \exp\left(\frac{-y'^2}{2\sigma^2}\right).$$

Finally, multiplicative speckle noises, e.g. in Synthetic aperture radar (SAR), have successfully been modeled using the Gamma distribution [Baraldi and Panniggiani, 1995].

In Figure 4.2, the above probability density functions are depicted for two intensity levels ($x = 2$ and $x = 6$). It can be seen that the Poisson, Rician and Rayleigh distributions are asymmetric around $y = x$ (straight line). For higher intensity levels, the Poisson distribution has a higher variance than the Gaussian distribution (6 versus 1), while the Rician distribution approximately coincides with the Gaussian distribution.

4.2 Second-order statistics of noise

In most practical applications, neighboring noise samples are *not* statistically independent. Instead, spatial dependencies exist between these samples and the noise is called "colored noise". In Section 4.2.1, we will explain how these dependencies can be described by the second-order statistics of the noise. We discuss a few origins of colored noise in digital images. Next, in Section 4.2.2, we will present a new technique for estimating the noise correlations from an observed image.

4.2.1 From white to colored noise

We consider a stationary additive noise process, which can be described by:

$$y(\mathbf{p}) = x(\mathbf{p}) + w(\mathbf{p}) \tag{4.1}$$

where $x(\mathbf{p})$ is a pixel intensity of a noise-free image at position $\mathbf{p}$, $y(\mathbf{p})$ is the corresponding observed pixel intensity and $w(\mathbf{p})$ is a zero-mean additive noise component. We will further assume that the samples $w(\mathbf{p})$ are generated by a (wide-sense) spatial stationary process w, in which the correlation between two noise samples only depends on the position difference between the two noise samples, but not on their absolute position. Consequently, w can be completely described by mean and autocorrelation function.

A random process w obeying the above conditions is called *white* if its autocorrelation function is a Dirac delta function:

$$R_w(\mathbf{p}) = \mathrm{E}\left[w(\mathbf{p}')\overline{w(\mathbf{p}+\mathbf{p}')}\right] = \delta(\mathbf{p}). \tag{4.2}$$

According to the Wiener-Khinchin theorem, the *power spectral density* (PSD) is the (discrete time) Fourier transform of $R_w(\mathbf{p})$:

$$P(\boldsymbol{\omega}) = \sum_{\mathbf{p}\in\mathbb{Z}^2} R_w(\mathbf{p}) \exp\left(-j\boldsymbol{\omega}^T\mathbf{p}\right). \tag{4.3}$$

The PSD describes how the noise energy is distributed in frequency space. For white noise, the PSD is flat $P(\boldsymbol{\omega}) = 1$, hence the name *white*. Suppose a filter with frequency response $H(\boldsymbol{\omega}) \neq 1$ is applied to the *white* noise samples, then the resulting PSD $R_w^{'}(\mathbf{p})$ becomes [Baher, 2001]:

$$R_w^{'}(\mathbf{p}) = R_w(\mathbf{p})\left|H(\boldsymbol{\omega})\right|^2. \tag{4.4}$$

Clearly, the PSD $R_w^{'}(\mathbf{p})$ is subjected to the filter magnitude response $|H(\boldsymbol{\omega})|$. Hence one can think of correlated noise as white noise subjected to linear filtering. In analogy with the term "*white noise*" the resulting term is called "*colored noise*" (or *correlated* noise, because the filtering introduces correlations in the noise samples).

Next, we will discuss a number of origins of colored noise in images:

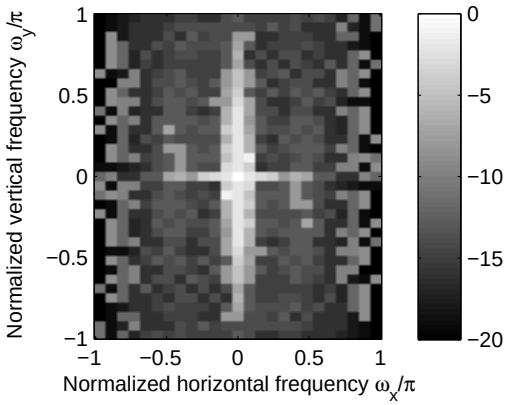
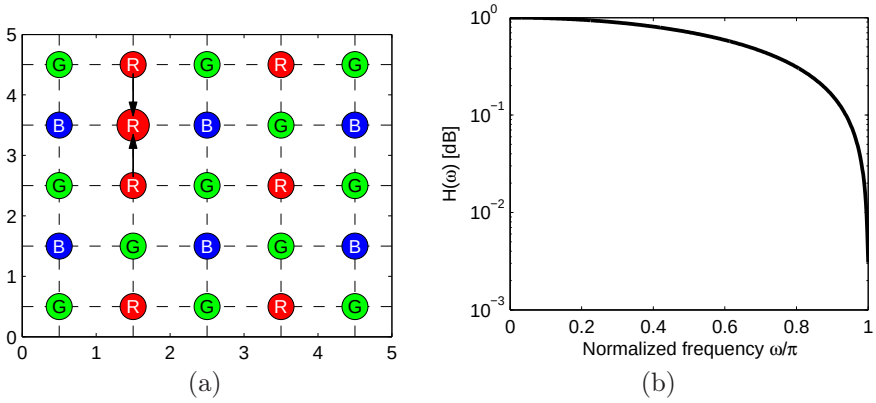


Figure 4.3: Power Spectral Density [dB] of noise in PAL broadcasting.

- *Phase Alternating Line (PAL) television*: the noise in PAL television images is a good example of colored noise. The correlations between the noise samples are caused by several mechanisms, such as deinterlacing [Kwon et al., 2003], demodulation and filter schemes. In Figure 4.3 the PSD of a noise patch from a PAL broadcast is shown. Here, there is a high concentration of energy in the lower horizontal frequencies, leading to horizontal stripes and artifacts.

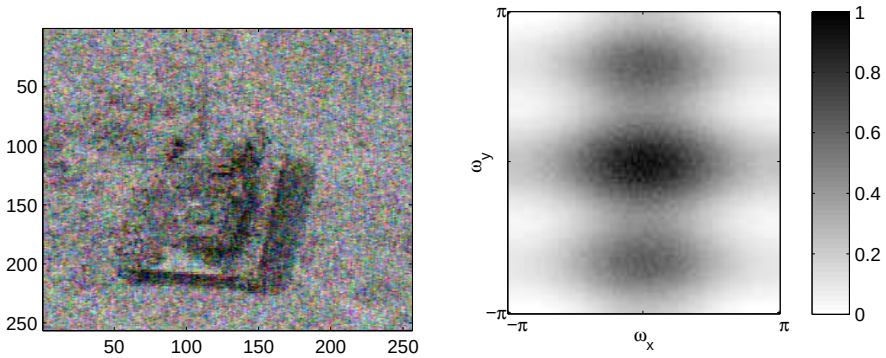
- *Color interpolation (demosaicing)*: modern digital cameras use a rectangular arrangement of photosensitive elements. In this matrix arrangement, photosensitive elements of different color sensitivity are placed in an interleaving way. This allows sampling of full color images without the use of three matrices of photosensitive elements. One popular example is the Bayer pattern [Bayer, 1976] (see Figure 4.4). Color interpolation (or demosaicing) is the process of estimating the values of missing photosensitive elements. The basic concept is illustrated in Figure 4.4(a): the interpolation is here a average of the neighboring red sensor element values: $R_2 = (R_1 + R_3)/2$. This is equivalent to a one-dimensional interpolation filter with frequency response $H(\omega) = \frac{1}{2}(1 + \exp(-j\omega))$ (see Figure 4.4(b)). This interpolation also inherently introduces correlations in the noise (see Figure 4.5). The image quality of this linear interpolation scheme is rather poor and for this reason most digital cameras use more sophisticated edge-adaptive interpolation schemes. However, the presence of sensor noise hampers the estimation of the edge direction. Generally, the presence of correlated noise in demosaiced images can not be avoided. In Chapter 5 we will describe a new demosaicing technique that is able to deal with sensor noise.

- *Post-processing techniques*: image noise often becomes correlated by the use of post-processing techniques, e.g., to enhance the quality of the im-



(a) (b)

Figure 4.4: (a) Bayer mosaic pattern and linear color interpolation for a missing red photosensitive element. (b) Frequency response of the linear filter $H(\omega) = \frac{1}{2}(1 + \exp(-j\omega))$.



(a) (b)

Figure 4.5: (a) Image corrupted with colored noise caused by demosaicing (b) PSD of the noise in the green color channel of (a).

age or to store the images. Examples are sharpening filters, digital zoom functions of cameras, JPEG compression... In [Luong, 2009] it was found that superresolution (SR) fusion techniques often create correlated noise with a very specific structure (see Figure 4.6(c)), mainly due to the particular alignment of the low resolution input images.

- *Thermal cameras*: images captured by thermal cameras of the push broom or whisk broom type often exhibit streaking noise artifacts, mainly caused by detector and sampling circuitry [Aelterman et al., 2010b]. This kind of noise can be approximated using a $1/f$ frequency characteristic (called *pink* noise) [Borel et al., 1996]. Pink noise also frequently arises in image sensors that acquire pixel data in time.

- *Computed Tomography (CT)*: in CT, noise correlations are often introduced by the specific reconstruction technique that is being used. Noise created by the backprojection algorithm (without reconstruction filter) is called *ramp-spectrum* noise, and has a f^1 frequency characteristic. Noise in CT will further be treated in Chapter 6.

- *Magnetic Resonance Imaging (MRI)*: noise in MRI images is traditionally considered *white* [Nowak, 1999b, Pižurica et al., 2003], although many MRI scanner manufacturers have included a wide range of techniques to allow for shorter scanning times (mainly to avoid patient motion artifacts in the images). To name a few: K-space subsampling, partial Fourier, elliptical filtering [Aelterman et al., 2010a]. The use of these techniques results in correlated noise in the reconstructed MRI images.

In Figure 4.6 some examples are shown of images corrupted with colored noise. The colored noise was artificially generated by subjecting white noise to a filter with magnitude response $\sqrt{P(\boldsymbol{\omega})}$ and subsequently by adding the filtered noise to the images.

For some applications that we will encounter in Chapter 5, the noise autocorrelation function is required in the wavelet domain (or another multiresolution transform domain). In case the noise PSD is *known* in advance (e.g. as in the above examples), the transform domain autocorrelation function for every subband can be computed using (4.4). However, this is only possible for a shift-invariant multiresolution transforms, such as the undecimated wavelet transform and steerable pyramid transform. For shift-variant transforms, the autocorrelation formula should also take the decimations of the discrete transform into account. A straightforward solution would then be to compute autocorrelation functions using (4.4) and subsequently to apply appropriate decimation operations to the obtained autocorrelation functions. However, this approach is not very practical in higher dimensions (e.g. 3D), as the spatial supports of the autocorrelation functions can become very large and many redundant computations need to be done.

In [Goossens et al., 2010a], we propose an alternative computation method that is similar to the fast DWT (see Section 2.1.3), but that processes autocorrelation functions instead of raw signals or images. Therefore, we derived an analytical formula that expresses the noise autocorrelation function for each invididual multiresolution subband as a function of the noise autocorrelation function of the subband coefficients in a finer multiresolution scale (taking the decimations of the transform into account). Next, by applying this formula recursively, the noise autocorrelation function can be exactly computed for every scale of the multiresolution transform.

For the DT-CWT, the calculation of the autocorrelation function is slightly more complicated than for, e.g., the DWT, because it also involves computing crosscorrelations, due to the linear transform needed to determine the real and imaginary parts of the complex wavelet coefficients (see Section 2.2.2). Nevertheless, devising a recursive computation scheme for the noise autocorrelation

function is still possible. For the exact details, we refer to [Goossens et al., 2010a].

In the next section, we investigate an even more challenging task: the estimation of the noise PSD from an observed image containing signal structures.

4.2.2 Estimation of colored Gaussian noise

Estimating the correlation properties of colored Gaussian noise from an observed image amounts to estimating the Power Spectral Density (PSD) of the noise in this image. However, this is a very challenging task, because one has to discriminate between the noise and the underlying image. For characterizing the noise, a good frequency resolution is desired, while for identifying the signal, good spatial resolution is needed. For this perspective, it is natural to estimate the noise correlations in a multiresolution transform domain, this has also the advantage that (noise-free) image models defined in this domain (see Chapter 3) can be used. In this section, we will present a technique that uses the Gaussian Scale Mixture as underlying prior model for the noise-free coefficients. A linear multiresolution transform retains the additivity of the noise, such that for one particular subband, we can write:

$$\mathbf{y}_j = \mathbf{x}_j + \mathbf{w}_j, \tag{4.5}$$

where $\mathbf{y}_j, \mathbf{x}_j, \mathbf{w}_j$ are vectors consisting of respectively the observed subband coefficients, the noise-free subband coefficients and the noise coefficients. The vectors are extracted in $\sqrt{d} \times \sqrt{d}$ overlapping local neighborhoods, centered at position $j = 1, ..., N$. Here the position index is again a one-dimensional index (like in raster scanning). The noise is Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{C}_w)$ and the noise-free coefficients are GSM distributed with covariance matrix $\mathbf{C}_x = \mathrm{E}[z]\,\mathbf{C}_u$:

$$\mathbf{x}|z \sim \mathcal{N}(\mathbf{0}, z\mathbf{C}_u). \tag{4.6}$$

Consequently, the density of $\mathbf{y}$ is a Gaussian mixture model with a specific constraint imposed to the covarianced matrices of the components:

$$\mathbf{y}|z \sim \mathcal{N}(\mathbf{0}, z\mathbf{C}_u + \mathbf{C}_w), \tag{4.7}$$

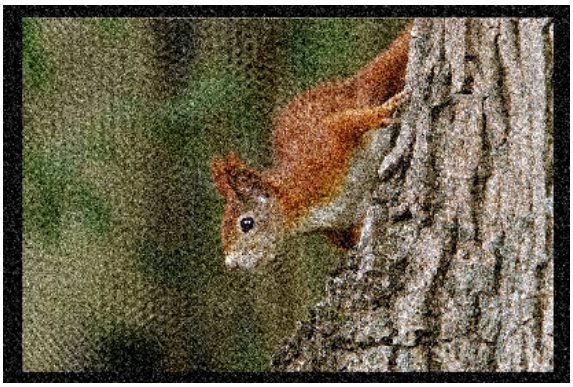
with covariance matrix $\mathbf{C}_y = \mathrm{E}[z]\,\mathbf{C}_u + \mathbf{C}_w$. Our goal is now to estimate the noise covariance matrix $\mathbf{C}_w$.

One of the main difficulties here is dealing with the hidden multiplier z, which has a continuous range of possible values. To allow for a simple numerical implementation, we will assume a discrete density for z: $\alpha_k = \mathrm{P}(z = z_k)$, for $k = 1, ..., K$ and with z_k fixed. The full set of model parameters is then given by $\mathbf{\Theta} = \{\mathbf{C}_x, \mathbf{C}_w, \mathbf{C}_k, \alpha_k\}$. To simplify the notations and to avoid scaling ambiguity (see Section 3.2.4), we will again assume that $\mathrm{E}[z] = 1$, such that $\mathbf{C}_x = \mathbf{C}_u$. As in Section 3.4.4, the parameter estimation can be done using an Expectation Maximization (EM) algorithm. However, estimating the GSM parameters jointly with the noise covariance matrix has been found to be a difficult

(a)

(b)

(c)



(a) PAL TV Noise $P(\boldsymbol{\omega}) \propto d + \sum_{m=-1,1} \exp\left(-a(\omega_y+mb)^2 - a(\omega_x+mc)^2\right)$

(b) Pink noise $P(\boldsymbol{\omega}) \propto 1/\omega_y$

(c) Noise after SR fusion $P(\boldsymbol{\omega}) \propto \sum_{m=-1}^{1} \sum_{\substack{n=-1 \\ m\neq n}}^{1} \exp\left(-a\sqrt{(\omega_y+mb)^2+(\omega_x+nc)^2}\right)$

Figure 4.6: Examples of images with *colored* noise with PSD $P(\boldsymbol{\omega})$ (use $a > 1$, $b \in [-\pi, \pi]$, $c \in [-\pi, \pi]$ and $0 < d < 1$).

task even using an EM algorithm [Portilla, 2004], in the sence that straightforward derivations do not lead to closed-form update formulas. In [Portilla, 2004] the update formula was therefore replaced by an easier, approximate equation and after every iteration it was checked whether the model likelihood was increased. If not, the update from the last iteration was un-done and replaced by a steepest ascent step, yielding a Generalized Expectation Maximization algorithm.

Here, we take a different approach: we rely on the fact that the density $f_{\mathbf{y}}(\mathbf{y})$ is a Gaussian Mixture model in which the components have covariance matrices $\mathbf{C}_k$ and subsequently we impose a Gaussian Scale Mixture constraint to the components. Compared to the GEM algorithm from [Portilla, 2004], our technique will have the advantage that the log-likelihood function does not need to be computed at every iteration to ensure convergence (which is generally a computationally intensive task). The Gaussian Scale Mixture constraint is as follows:

$$\mathbf{C}_k = z_k \mathbf{C}_x + \mathbf{C}_w. \tag{4.8}$$

Given the set of model parameters $\mathbf{\Theta}^{(i)}$ at iteration i, we want to optimize the new parameters $\mathbf{\Theta}$ in order to maximize the objective function:

$$\begin{aligned} &\text{maximize}\, \mathrm{E}\left[\log f_{\mathbf{y},k|\mathbf{\Theta}}\left(\mathbf{y}, k|\mathbf{\Theta}^{(i)}\right)|\mathbf{y}, \mathbf{\Theta}^{(i)}\right] \\ &\text{s.t. } \mathbf{C}_k = z_k \mathbf{C}_x + \mathbf{C}_w, \quad k = 1, ..., K \end{aligned} \tag{4.9}$$

This optimization problem can be converted into a constrained problem (which is also called a *"constrained"* EM algorithm):

$$Q(\mathbf{\Theta}, \mathbf{\Theta}^{(i)}) = \mathrm{E}\left[\log f_{\mathbf{y},k|\mathbf{\Theta}}\left(\mathbf{y}, k|\mathbf{\Theta}^{(i)}\right)|\mathbf{y}, \mathbf{\Theta}^{(i)}\right] - \sum_{k=1}^{K} \lambda_k \left\|\mathbf{C}_k - z_k \mathbf{C}_x - \mathbf{C}_w\right\|_F^2, \tag{4.10}$$

where $\lambda_k, k = 1, ..., K$ are Lagrange multipliers and where $\|\mathbf{A}\|_F^2 = \mathrm{tr}\left(\mathbf{A}\mathbf{A}^T\right)$ denotes the matrix Frobenius norm. It can be shown that the EM update equations are given by:

$$\text{mixture weights (E-step)} \quad \hat{\alpha}_k = \frac{1}{N}\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right), \; k = 1, ..., K \tag{4.11}$$

$$\text{Component covariances (M-step)} \quad \hat{\mathbf{C}}_k^{(1)} = \frac{\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right)\mathbf{y}_j\mathbf{y}_j^T}{\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right)}, \; k = 1, ..., K \tag{4.12}$$

$$\text{Signal covariance matrix (M-step)} \quad \hat{\mathbf{C}}_x = \sum_{k=1}^{K} \frac{\mu_1 - z_k}{\mu_1^2 - \mu_2}\lambda_k \hat{\mathbf{C}}_k^{(1)} \tag{4.13}$$

$$\text{Noise covariance matrix (M-step)} \quad \hat{\mathbf{C}}_w = \sum_{k=1}^{K} \frac{\mu_1 z_k - \mu_2}{\mu_1^2 - \mu_2}\lambda_k \hat{\mathbf{C}}_k^{(1)} \tag{4.14}$$

Algorithm 4.1 Algorithm for estimating the noise covariance matrix of a noisy wavelet subband.

repeat

$$\hat{\alpha}_k = \frac{1}{N}\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right), \quad \textbf{for } k = 1, ..., K$$

$$\hat{\mathbf{C}}_k^{(1)} = \frac{\sum_{j=1}^{N} \mathrm{P}(z=z_k|\mathbf{y}_j)\mathbf{y}_j\mathbf{y}_j^T}{\sum_{j=1}^{N} \mathrm{P}(z=z_k|\mathbf{y}_j)}, \quad \textbf{for } k = 1, ..., K$$

$$\hat{\mathbf{C}}_x = \sum_{k=1}^{K} \frac{\mu_1 - z_k}{\mu_1^2 - \mu_2}\hat{\alpha}_k\hat{\mathbf{C}}_k^{(1)}$$

$$\hat{\mathbf{C}}_w = \sum_{k=1}^{K} \frac{\mu_1 z_k - \mu_2}{\mu_1^2 - \mu_2}\hat{\alpha}_k\hat{\mathbf{C}}_k^{(1)}$$

$$\hat{\mathbf{C}}_k = z_k\mathbf{C}_x + \mathbf{C}_w$$

until convergence

where $\mu_1 = \sum_{k=1}^{K} \lambda_k z_k$ and $\mu_2 = \sum_{k=1}^{K} \lambda_k z_k^2$. For GSM mixtures of two components ($K = 2$), the GSM constraint (4.8) can be satisfied exactly. Further demanding that the weighted sum of the mixture component covariance matrices is equal to the sum of the signal covariance matrix and noise covariance matrix ($\sum_{k=1}^{K} \hat{\alpha}_k\hat{\mathbf{C}}_k = \hat{\mathbf{C}}_x + \hat{\mathbf{C}}_w$) allows us to calculate the Lagrange multipliers:

$$\lambda_k = \hat{\alpha}_k = \frac{1}{N}\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right). \tag{4.15}$$

However, if $K > 2$, the constraint (4.8) will no longer hold exactly: the M-step (4.12) does not satisfy the constraints of the GSM model. The reason is that (4.8) constitute a set of linear equations (one for every element of the covariance matrices), and if $K = 2$ the linear system has an exact solution. For larger K, minimizing $\|\mathbf{C}_k - z_k\mathbf{C}_x - \mathbf{C}_w\|_F^2$ leads to the least squares solution for $\mathbf{C}_x$ and $\mathbf{C}_w$ (which generally has a cost $\|\mathbf{C}_k - z_k\mathbf{C}_x - \mathbf{C}_w\|_F^2 > 0$), hence to satisfy (4.8), the Lagrange multipliers tend to infinity: $\lambda_k \to \infty$. As a solution to this problem, we modify our estimate (4.12) for the covariance matrix of component k, such that the constraint is satisfied:

$$\hat{\mathbf{C}}_k = z_k\hat{\mathbf{C}}_x + \hat{\mathbf{C}}_w. \tag{4.16}$$

The resulting algorithm is summarized in Algorithm 4.1. It can be shown that this approach is similar to a Bregman iteration [Bregman, 1967], in which the error (i.e. $\hat{\mathbf{C}}_k - \hat{\mathbf{C}}_k^{(1)}$) is added back to the right handed side of the constraint $\mathbf{C}_k = z_k\mathbf{C}_x + \mathbf{C}_w$. Bregman iterations will be discussed more into detail in Section 5.3. Unfortunately, because the function $\mathrm{E}\left[\log f_{\mathbf{y},k|\boldsymbol{\Theta}}\left(\mathbf{y}, k|\boldsymbol{\Theta}^{(i)}\right)|\mathbf{y}, \boldsymbol{\Theta}^{(i)}\right]$ is not convex in general, convergence results from Bregman optimization do not transfer to this constrained EM algorithm. In Appendix A we show that the constrained EM algorithm has the same convergence properties as the unconstrained EM algorithm for Gaussian mixtures, in the sence that every iteration

increases the likelihood function. One drawback of our method is that there is no guarantee to converge to a global *maximum*, because the likelihood function may exhibit multiple non-global maxima. This is the case for almost all EM algorithms [Dempster et al., 1977]. Consequently, the final solution can be improved by using good initial estimates of $\hat{\mathbf{C}}_x^{(0)}$ and $\hat{\mathbf{C}}_w^{(0)}$. In this work, we choose $\hat{\mathbf{C}}_x^{(0)} = q\mathbf{C}_y$ and $\hat{\mathbf{C}}_w^{(0)} = (1-q)\mathbf{C}_y$, with q close to 0 (e.g. $q = 0.1$). This choice is motivated by the fact that for sufficiently low SNRs, the subbands are dominated by noise, such that $\mathbf{C}_w \approx \mathbf{C}_y$. Alternatively, good initial estimates can be obtained using robust S-estimators of the (noise) covariance matrix [Campbell et al., 1998].

Important to remark is that the above algorithm *fails*, if the denominator in the update formulas (4.13)-(4.14) is zero, i.e. if $\mu_1^2 = \mu_2$. It is worthful to note that the kurtosis of the coefficient subband coefficients is given by $3\mu_2/\mu_1^2 - 3$, which becomes zero if $\mu_1^2 = \mu_2$. In this case, the probability density function $f_{\mathbf{y}}(\mathbf{y})$ is Gaussian, and every component of the GSM model will have the same hidden multiplier value $z_k = \mu_1$, such that also $f_{\mathbf{x}}(\mathbf{x})$ is Gaussian. Consequently, it becomes impossible to separate the signal from the noise: the highly kurtotic behavior of the noise-free coefficients $\mathbf{x}$ can not be exploited. Luckily, we can avoid this problem, by using a prior distribution for z that is well initialized (i.e. all z_k are different). A possible choice is to use a fixed initialization, e.g., $z_k = \exp(-3 + 7(k-1)/(K-1))$, $k = 1, ..., K$ and $\alpha_k = 1/K$, as in [Goossens et al., 2009d]. Nevertheless, it has been noted (e.g. in [Portilla, 2004]) that the kurtosis of noise-free coefficients decreases for lower frequency subbands, such that it becomes more and more difficult to estimate the noise covariance matrix in these subbands. A possible solution is then to use an appropriate interscale model (see Section 3.5) combined with this EM algorithm to detect significant coefficients in these subbands.

Another practical problem is that the eigenvalues of $\mathbf{C}_x$ and $\mathbf{C}_w$ can become negative during the iterative procedure (mainly due to estimation errors or numerical errors). In this case, we replace the negative eigenvalues of $\mathbf{C}_x$ and $\mathbf{C}_w$ by a small positive value (e.g. 10^{-4}), such that the positive definiteness of these covariance matrices is not lost. A similar approach is taken in [Portilla et al., 2003] for estimating $\mathbf{C}_x$, for the case that $\mathbf{C}_w$ is known.

A "blind" denoising technique can then be obtained by 1) estimating the signal and noise covariance matrix using the above EM algorithm, and by 2) using a general denoising technique which assumes that these covariance matrices are known in advance. Although general denoising techniques will be discussed in Chapter 5, Figure 4.8 already shows a denoising result for the *Baboon* and *Peppers* image. In particular, the *Baboon* image was chosen because it contains many fine details (e.g. the hairs) that could potentially be mistakenly recognized as components of white Gaussian noise. The denoising technique being used is BLS-GSM from [Portilla et al., 2003]. It can be seen that both the noise estimation and denoising techniques are quite effective in this example, as most fine details are well reconstructed, despite the high noise variance.

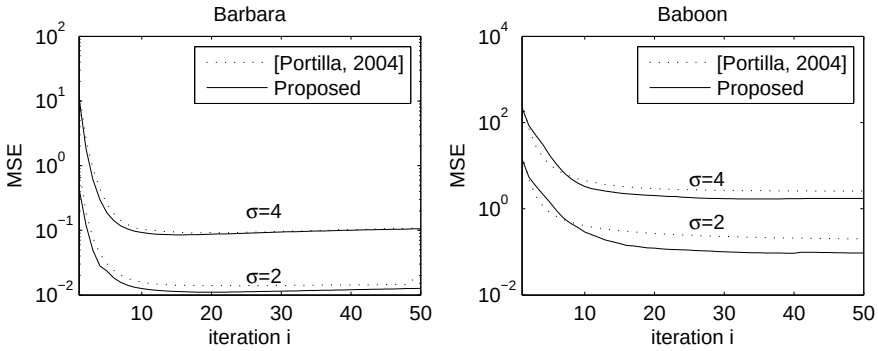


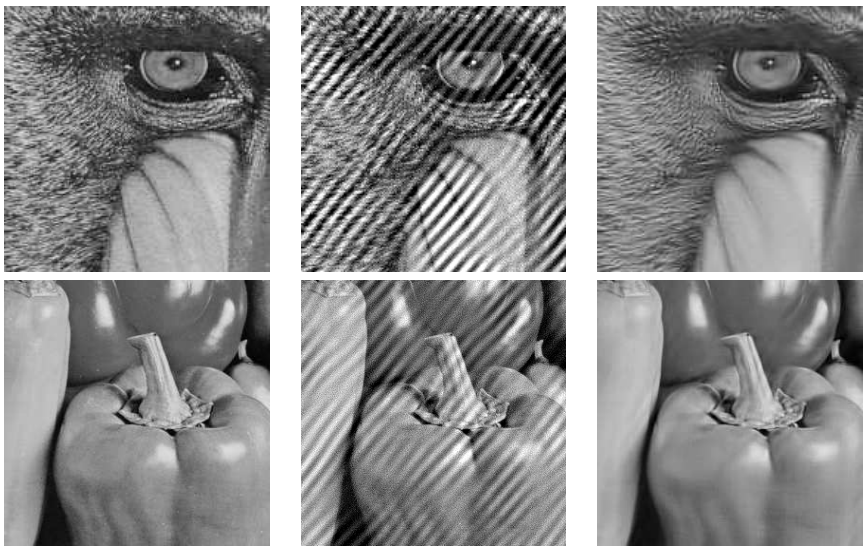
Figure 4.7: Evaluation of the noise covariance estimation error in the MSE sense for texture-rich images: (a) the *Barbara* image, (b) the *Baboon* image.

Finally, to evaluate the estimation performance of our estimation technique as a function of the EM iteration number, we added artificial white Gaussian noise to the *Barbara* and the *Baboon* image,[1] with standard deviation respectively $\sigma = 2$ and $\sigma = 4$. We then compare the noise estimation method from [Portilla, 2004] to the proposed estimation scheme in the DT-CWT domain, using the fixed initialization for z_k and α_k, with $K = 8$. Because the number of complex wavelet coefficients depends on the scale of the transform, we only used the finest scale for comparison in this experiment. Orthogonal Symlet wavelets with 16 vanishing moments are used for the first scale of the DT-CWT. Because the noise PSD is known, we can directly apply our exact computation method from [Goossens et al., 2010a] (which is briefly outlined at the end of Section 4.2.1) to obtain a ground truth for noise covariance matrix of every complex wavelet subband. The performance measure we use is the MSE between the estimated noise covariance matrix and the exact (ground truth) noise covariance matrix (MSE $= \left\| \hat{\mathbf{C}}_w - \mathbf{C}_w^{\text{ground truth}} \right\|_F^2$). The results are shown in Figure 4.7. It can be seen that, after a sufficient number of iterations, the proposed EM algorithm consistently obtains a lower MSE than the reference method from [Portilla et al., 2003], which indicates more accurate estimation of the noise covariance matrix.

4.3 Modeling and estimation of non-stationary noise

In practice, we encounter many situations where the noise energy and correlation structure depends on the position in the image (non-stationary noise).

[1] Two texture-rich images are used in this case, in order to have a non-trivial estimation task.



(a) (b) (c)

Figure 4.8: Blind denoising results using the EM algorithm from 4.1 for the DT-CWT. (a) Original noise-free image, (b) Image with correlated noise of variance 50^2, (c) Denoised image.

This can be both due to the acquisition device itself (e.g. interference from other devices), or by various post-processing steps (e.g. locally adaptive filtering). An example of an image with artificially generated non-stationary noise is given in Figure 4.9. In this example, the noise variance varies with the position in the image, but does not depend on the underlying image. Most often so, the noise is signal-dependent as well. In this section, we restrict ourselves to signal-*independent* (additive) noise; signal-dependent noise will be discussed later in Section 4.4.

In general, a model for non-stationary noise requires many parameters such that large ensembles of images are needed in order to reliably estimate these parameters. Therefore, we will make a few assumptions, which will allow us to work with one single image:

1. We assume that the statistics of the noise are Gaussian.

2. We consider locally stationary processes, which have properties that change slowly in space.

Because the noise properties vary slowly in space, it becomes possible to build a model for the noise from which the parameters can be estimated locally, using an EM algorithm that is similar as in Section 4.2.2.

Figure 4.9: Example of an image with artificially generated non-stationary noise.

4.3.1 Modeling of locally stationary Gaussian noise

In this section we will study *locally* stationary Gaussian noise processes, which have properties that change slowly in space. We say that a noise process is *locally* stationary, if in the neighborhood of any $\mathbf{q} \in \mathbb{Z}^2$, there exists a square window $\delta(\mathbf{q})$ of size $l(\mathbf{q})$, centered at position $\mathbf{q}$, where the process can be approximated by a stationary one : for $\mathbf{p} \in \delta(\mathbf{q})$ and for $|\mathbf{r}| \leq l(\mathbf{q})/2$, the autocorrelation is well approximated by [Mallat, 1998]:

$$\mathrm{E}\left[w(\mathbf{p})w(\mathbf{p}+\mathbf{r})\right] \approx \mathrm{E}\left[w(\mathbf{q})w(\mathbf{q}+\mathbf{r})\right] = R_w(\mathbf{q},\mathbf{r}). \tag{4.17}$$

We define the space-varying spectrum (SVS) of $w(\mathbf{p})$ as the Discrete Time Fourier transform (DTFT) of $R(\mathbf{q},\mathbf{r})$ with respect to $\mathbf{r}$:

$$S(\mathbf{q},\boldsymbol{\omega}) = \sum_{\mathbf{r}\in\mathbb{Z}^2} R(\mathbf{q},\mathbf{r}) \exp(-j\mathbf{r}^T\boldsymbol{\omega}) \tag{4.18}$$

For stationary processes, the SVS coincides with the Power Spectral Density (PSD), while for non-stationary processes, the PSD does not exist. We say that the SVS is *separable* if it can be factored as $S(\mathbf{q},\boldsymbol{\omega}) = S_0(\mathbf{q})S_1(\boldsymbol{\omega})$ with $\frac{1}{2\pi}\int_{-\pi}^{\pi} S_1(\boldsymbol{\omega})d\omega = 1$. The first component $S_0(\mathbf{q})$ represents the variance at position $\mathbf{q}$ while the second component $S_1(\boldsymbol{\omega})$ denotes the normalized Power Spectral Density (PSD). A specific class of locally stationary processes is obtained by the spatially variant filtering of white noise. Let $w(\mathbf{p})$ denote a white Gaussian noise process, then $Y(\mathbf{p})$ is obtained as:

$$Y(\mathbf{p}) = \sum_{\mathbf{q}\in\mathbb{Z}^2} w(\mathbf{q})H(\mathbf{p},\mathbf{p}-\mathbf{q}) \tag{4.19}$$

with $H(\mathbf{p},\mathbf{r})$ the impulse response of a linear spatially variant filter with

DTFT $\widehat{H}(\mathbf{p},\boldsymbol{\omega})$. The autocorrelation function of $Y(\mathbf{p})$ is then given by:

$$R(\mathbf{p},\mathbf{r}) = \mathrm{E}\left[Y(\mathbf{p})Y(\mathbf{p}+\mathbf{r})\right] \tag{4.20}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{+\pi} \overline{\widehat{H}(\mathbf{p},\boldsymbol{\omega})}\widehat{H}(\mathbf{p}+\mathbf{r},\boldsymbol{\omega})\exp(j\boldsymbol{\omega}^T\mathbf{r})d\boldsymbol{\omega}. \tag{4.21}$$

The local stationarity assumption (4.17) imposes that $\widehat{H}(\mathbf{p},\boldsymbol{\omega})$ has to satisfy some smoothness conditions (see [Mallat, 1998]). More specifically, if $|\mathbf{r}^T\frac{\partial\widehat{H}(\mathbf{p},\boldsymbol{\omega})}{\partial\mathbf{p}}| \ll |\widehat{H}(\mathbf{p},\boldsymbol{\omega})|$, for $|\mathbf{r}| \leq l(\mathbf{p})/2$, we have approximately:

$$\overline{\widehat{H}\left(\mathbf{p},\boldsymbol{\omega}\right)}\widehat{H}\left(\mathbf{p}+\mathbf{r},\boldsymbol{\omega}\right) \approx |\widehat{H}(\mathbf{p},\boldsymbol{\omega})|^2. \tag{4.22}$$

This condition implies that the SVS of the noise process varies slowly in time.

An example of such a noise process in 1D is depicted in Figure 4.10: Figure 4.10(a) shows a stationary white Gaussian noise process, Figure 4.10(b) displays the frequency response $\left|\widehat{H}(\mathbf{p},\boldsymbol{\omega})\right|$ of a time-variant linear filter. In particular, a high-resonant low-pass IIR filter with a transition band of 12dB/octave was used here; the cutoff frequency of the filter is varied slowly in time in order to obtain a locally stationary process. Next, Figure 4.10(c) shows the filtered noise signal and Figure 4.10(d) displays the multiscale undecimated wavelet analysis of the filter noise signal. The wavelet analysis is able to recover the time variant frequency response relatively well, despite the fact that only one examplar of the noise process is used.

Our approach to dealing with spatially variant correlated noise consists of 1) estimating the spatially variant autocorrelation function $R(\mathbf{p},\mathbf{r})$ in the wavelet domain, in presence of signal information and 2) denoising the degraded image in the wavelet domain using the estimated autocorrelation functions.

4.3.2 Estimation of locally stationary Gaussian noise

To estimate the noise covariance function $R^{(s,o)}(\mathbf{p},\mathbf{q})$ in the wavelet domain, in the presence of signal structures we again consider one wavelet subband (s,o). By the additivity of the noise, we have an equivalent additive relationship between the noisy wavelet coefficients $\mathbf{y}(\mathbf{p})$, the noise-free coefficients $\mathbf{x}(\mathbf{p})$ and the white noise $\mathbf{w}(\mathbf{p})$ at position $\mathbf{p} \in \mathcal{B}$:

$$\mathbf{y}(\mathbf{p}) = \mathbf{x}(\mathbf{p}) + \boldsymbol{\rho}(\mathbf{p})\mathbf{w}(\mathbf{p}). \tag{4.23}$$

The vectors $\mathbf{x}(\mathbf{p})$, $\boldsymbol{\epsilon}(\mathbf{p})$ and $\mathbf{y}(\mathbf{p})$ are formed by column-stacking the wavelet coefficients in local $\sqrt{d}\times\sqrt{d}$ overlapping windows centered at position $\mathbf{p}$. $\boldsymbol{\rho}(\mathbf{p})$ is a spatially variant $d\times d$ matrix that correlates the noise $\mathbf{w}(\mathbf{p}) \sim N(\mathbf{0},\mathbf{I})$. To distinguish noise from signal structures, we take again prior knowledge about the noise-free signal $\mathbf{x}(\mathbf{p})$ into account by modeling $\mathbf{x}(\mathbf{p})$ as a Gaussian Scale Mixture (GSM) with discrete hidden multiplier $z \in \{z_1, z_2, ..., z_K\}$. With this model, estimating $R^{(s,o)}(\mathbf{p},\mathbf{r})$ comes down to estimating $\boldsymbol{\rho}(\mathbf{p})\boldsymbol{\rho}^T(\mathbf{p})$, for which we can use a similar scheme as in Section 4.2.2. In the following, we will denote again $\alpha_k = \mathrm{P}\left(z = z_k\right), k = 1, ..., K$.

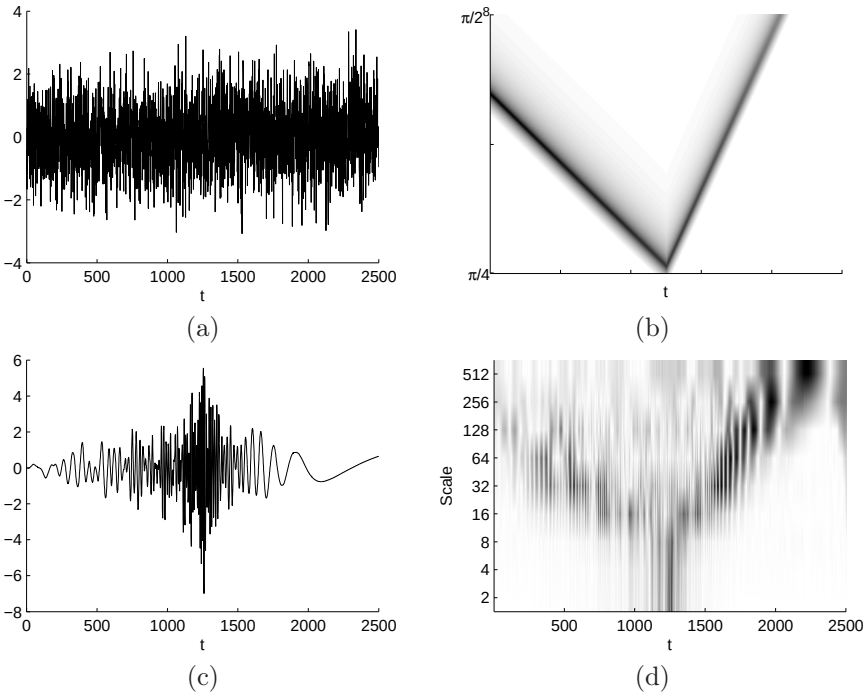


Figure 4.10: (a) White noise signal, (b) Frequency response of a time-variant filter as function of the time t (vertical frequency axis is in a logarithmic scale), (c) The white noise signal filtered by the time-variant filter from (b), (d). Undecimated discrete wavelet analysis of the noise signal from (c). *Black* corresponds to wavelet coefficients with significant magnitudes, *white* corresponds to nonsignificant wavelet coefficients.

Noise with separable space-varying spectrum

In a number of circumstances, the noise covariance matrix is constant for the whole image, up to a spatially varying scale factor $\sigma^2(\mathbf{p})$, representing the local noise variance. We have:

$$\boldsymbol{\rho}(\mathbf{p})\boldsymbol{\rho}^T(\mathbf{p}) = \sigma^2(\mathbf{p})\mathbf{C}_w. \tag{4.24}$$

It is clear that $\mathbf{C}_w$ can be estimated using information from the *whole* subband, while $\sigma^2(\mathbf{p})$ can only be obtained *locally*. Let $\boldsymbol{\theta}(\mathbf{p}) = \{\mathbf{C}_u, \mathbf{C}_w, \sigma^2(\mathbf{p})\} \cup \{\alpha_k, k = 1, ..., K\}$ denote the model parameters related to position $\mathbf{p}$. To estimate the total set of model parameters $\boldsymbol{\Theta} = \bigcup_{\mathbf{p}\in\mathcal{B}} \boldsymbol{\theta}(\mathbf{p})$ we again devise an EM algorithm for Gaussian mixtures with appropriate GSM constraint, as in Section 4.2.2:

$$\mathbf{C}_k(\mathbf{q}) = z_k\mathbf{C}_u + \sigma^2(\mathbf{q})\mathbf{C}_w \tag{4.25}$$

where $\mathbf{C}_k(\mathbf{q})$ denotes the covariance matrix of the kth Gaussian mixture component at position $\mathbf{q}$. Given a set of model parameters $\boldsymbol{\Theta}^{(i)}$ at iteration i, we

optimize the new parameters $\boldsymbol{\Theta}$ in order to increase the objective function:

$$Q(\boldsymbol{\Theta}^{(i)}, \boldsymbol{\Theta}) = \mathrm{E}\left[\log \prod_{\mathbf{q}\in\mathcal{B}} \prod_{\mathbf{p}\in\delta(\mathbf{q})} f(\mathbf{y}(\mathbf{p}), k|\boldsymbol{\theta}(\mathbf{q}))\big|\mathbf{y}, \boldsymbol{\Theta}^{(i)}\right] - \sum_{k=1}^{K}\sum_{\mathbf{q}\in\mathcal{B}} \lambda_k \left\|\mathbf{C}_k(\mathbf{q}) - z_k\mathbf{C}_u - \sigma^2(\mathbf{q})\mathbf{C}_w\right\|_F^2 \tag{4.26}$$

with the first term the expected complete-data log-likelihood function (where we use a simplifying assumption that coefficients in different overlapping local windows are statistically independent). The second term denotes the GSM constraint added to the problem using Lagrangian multipliers $\lambda_k, k = 1, ..., K$. It can be shown that the EM update equations are given by:

$$\hat{\alpha}_k = \frac{1}{N}\sum_{\mathbf{q}\in\mathcal{B}} \frac{1}{l^2(\mathbf{q})} \sum_{\mathbf{p}\in\delta(\mathbf{q})} \mathrm{P}\left(k|\mathbf{y}(\mathbf{p}), \boldsymbol{\theta}(\mathbf{q})\right), k = 1, ..., K \tag{4.27}$$

$$\hat{C}_k^{(1)}(\mathbf{q}) = \frac{\sum_{\mathbf{p}\in\delta(\mathbf{q})} \mathrm{P}\left(k|\mathbf{y}(\mathbf{p}), \boldsymbol{\theta}(\mathbf{q})\right)\mathbf{y}(\mathbf{p})\mathbf{y}^T(\mathbf{p})}{\sum_{\mathbf{p}\in\delta(\mathbf{q})} \mathrm{P}\left(k|\mathbf{y}(\mathbf{p}), \boldsymbol{\theta}(\mathbf{q})\right)}, k = 1, ..., K \tag{4.28}$$

$$\begin{pmatrix}\hat{\mathbf{C}}_u \\ \hat{\mathbf{C}}_w\end{pmatrix} = \begin{pmatrix} N\mu_2 & \mu_1\nu_1 \\ \mu_1\nu_1 & \nu_2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{k=1}^{K} \hat{\alpha}_k z_k \sum_{\mathbf{q}\in\mathcal{B}} \hat{\mathbf{C}}_k^{(1)}(\mathbf{q}) \\ \sum_{k=1}^{K} \hat{\alpha}_k \sum_{\mathbf{q}\in\mathcal{B}} \sigma^2(\mathbf{q})\hat{\mathbf{C}}_k^{(1)}(\mathbf{q}) \end{pmatrix} \tag{4.29}$$

$$\widehat{\sigma^2}(\mathbf{q}) = \frac{\mathrm{tr}\left(\sum_{k=1}^{K} \hat{\alpha}_k \mathbf{C}_w (\hat{\mathbf{C}}_k^{(1)}(\mathbf{q}) - \mathbf{C}_u)^T\right)}{\mathrm{tr}\left(\mathbf{C}_w\mathbf{C}_w^T\right)} \tag{4.30}$$

with μ_1, μ_2 as defined in Section 4.2.2 and with $\nu_b = \sum_{\mathbf{q}\in\mathcal{B}} \sigma^{2b}(\mathbf{q}), b = 1, 2$. The formulas above must be iterated until convergence of the likelihood. We note that update equations (4.29) and (4.30) depend on each other and must be used alternately in subsequent EM iterations in order to maximize the likelihood. In this iterative process, the *global* noise and signal covariance matrices $\mathbf{C}_u$, $\mathbf{C}_w$ as well as the *local* variance $\sigma^2(\mathbf{q})$ are estimated jointly. As for the EM algorithm from Section 4.2.2, the above formulas are the *exact* classical EM formulas for two mixture components (i.e. $K = 2$). For $K > 2$, we again modify the estimate (4.28) such that the GSM constraint is satisfied:

$$\hat{C}_k(\mathbf{q}) = z_k\hat{\mathbf{C}}_u + \sigma^2(\mathbf{q})\hat{\mathbf{C}}_w. \tag{4.31}$$

The computational time of this technique is significantly higher than the EM algorithm for stationary correlated noise from Section 4.2.2. This is because (4.28) and (4.29) require traversing the stationarity window $\delta(\mathbf{q})$ (in which local stationarity is assumed), for every position $\mathbf{q}$ in the subband. In practice, we use relatively large stationarity windows (e.g. 32×32) in order to obtain reliable estimates. Fortunately, the stationarity assumption can be further exploited to speed up the algorithm. For the details, see [Goossens et al., 2008c].

In Figure 4.11, some visual results are shown for this technique. First, the noise-free wavelet subband of Figure 4.11(g) is corrupted with additive noise, resulting from filtering white Gaussian noise by the space variant filter with

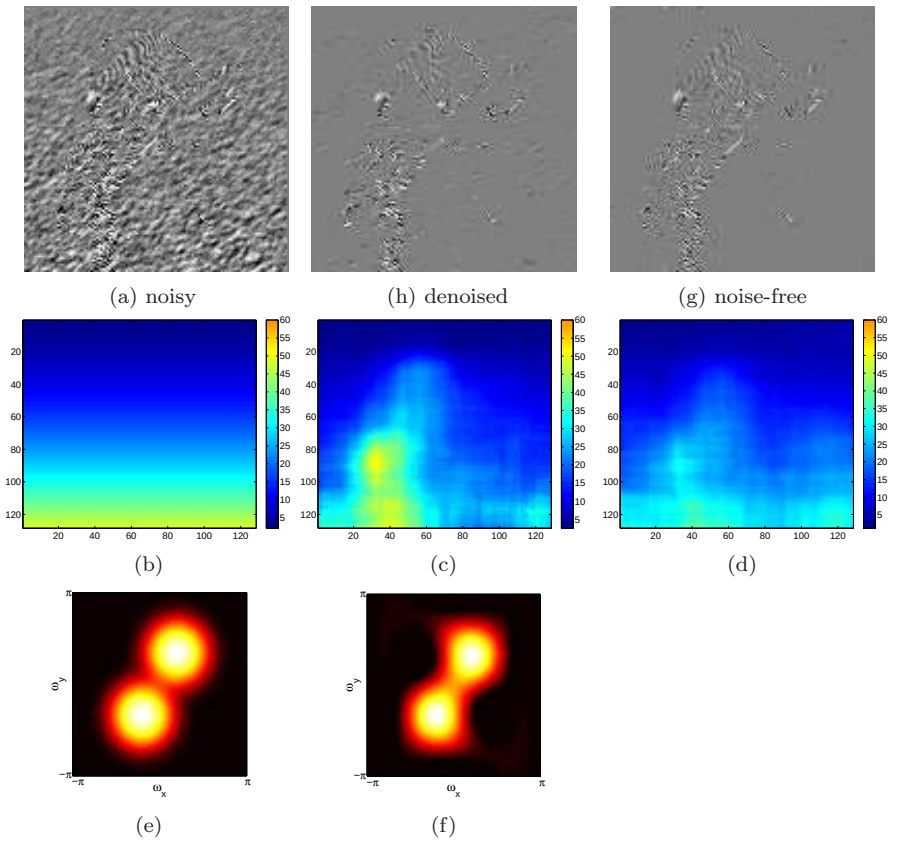


Figure 4.11: (a) Wavelet subband of Lena with added artificial noise with separable SVS (b) True local noise variance $\sigma^2(\mathbf{p})$ (c) Estimated local noise variance $\widehat{\sigma^2}(\mathbf{p})$, using the MAD-estimator (MSE=0.682) (d) Estimated local noise variance $\widehat{\sigma^2}(\mathbf{p})$, using the proposed method (MSE=0.451) (e) True noise PSD (f) Estimated noise PSD, using the proposed method (g) Original noise-free wavelet subband of Lena (h) Denoised wavelet subband of (a) using the estimated noise PSD (f) and local variance (d).

spectrum $|\widehat{H}(\mathbf{p},\boldsymbol{\omega})|^2 \sim [\mathbf{p}]_y^2 \exp(-60((\omega_x - 0.34\pi)^2 + (\omega_y - 0.20\pi)^2))$, see Figure 4.11(a). Here ω_x and ω_y denote respectively the x- and y-components of $\boldsymbol{\omega}$ and $[\mathbf{p}]_y$ is the y-component of $\mathbf{p}$. We use $l(\mathbf{p}) = 32$ and $d = 9$, corresponding to a 3×3 window for local correlations. The local noise variance $\sigma^2(\mathbf{p}) \propto [\mathbf{p}]_y^2$ is depicted in Figure 4.11(b). In Figure 4.11(c) the local noise variance is estimated locally using the robust Median of Absolute Deviations (MAD) estimator in a 32×32-window. Figure 4.11(d) shows the estimated $\widehat{\sigma^2}(\mathbf{p})$ using the proposed method with the same window size. The EM estimate is clearly much more robust to the presence of signal structures than the MAD estimate. This is mainly due to the fact that our method takes *signal correlations* into account whereas the MAD estimate does not. The estimated noise PSD in Figure 4.11(f) is

obtained by first converting the estimated noise covariance matrix $\hat{\mathbf{C}}_w$ into an autocorrelation function of size 128×128 by averaging over correlations that correspond to the same difference in position, setting correlations that can not be captured using a $\sqrt{d} \times \sqrt{d}$ window to zero and subsequently by computing the Discrete Fourier Transform. Despite the small window size 3×3 used for estimating local correlations, there is a very good resemblance to the original noise PSD in Figure 4.11(e). Next, the estimated noise parameters from Figure 4.11(d) and Figure 4.11(f) are used to denoise the wavelet subband, with an extension of the algorithm presented in [Goossens et al., 2009d] (such that it can deal with non-stationary noise, similar to the extension presented in [Portilla, 2005]. The result is shown in Figure 4.11(h). Due to the accurate noise estimation, the denoising algorithm reconstructs most of the signal structures present in Figure 4.11(a).

Noise with non-separable space-varying spectrum

In a more general scenario, the noise covariance matrix varies spatially and has to be estimated *locally*: $\boldsymbol{\rho}(\mathbf{p})\boldsymbol{\rho}^T(\mathbf{p}) = \mathbf{C}_w(\mathbf{p})$. To facilitate this, we will still estimate the signal covariance matrix $\mathbf{C}_u$ *globally*. The objective function now becomes:

$$Q(\boldsymbol{\Theta}^{(i)}, \boldsymbol{\Theta}) = \mathrm{E}\left[\log \prod_{\mathbf{q}\in\mathcal{B}} \prod_{\mathbf{p}\in\delta(\mathbf{q})} f(\mathbf{y}(\mathbf{p}), k|\boldsymbol{\theta}(\mathbf{q}))\big|\mathbf{y}, \boldsymbol{\Theta}^{(i)}\right] - \sum_{k=1}^{K} \sum_{\mathbf{q}\in\mathcal{B}} \lambda_k \left\|\mathbf{C}_k(\mathbf{q}) - z_k\mathbf{C}_u - \mathbf{C}_w(\mathbf{q})\right\|_F^2 . \tag{4.32}$$

Maximizing this function yields the same update equations as in the previous EM algorithm, except that (4.29) and (4.30) have to be replaced by:

$$\hat{\mathbf{C}}_u = \frac{1}{N} \sum_{k=1}^{K} \sum_{\mathbf{p}\in\mathcal{B}} \hat{\alpha}_k \left(\frac{z_k - \mu_1}{\mu_2 - \mu_1^2}\right) C_k^{(1)}(\mathbf{p}) \tag{4.33}$$

$$\hat{\mathbf{C}}_w(\mathbf{p}) = \sum_{k=1}^{K} \hat{\alpha}_k \hat{\mathbf{C}}_k^{(1)}(\mathbf{p}) - \mu_1 \hat{\mathbf{C}}_u, \quad \mathbf{p} \in \mathcal{B} \tag{4.34}$$

Finally, for $K > 2$, we again modify the estimate (4.28) such that the GSM constraint is satisfied:

$$\hat{C}_k(\mathbf{q}) = z_k \hat{\mathbf{C}}_u + \hat{\mathbf{C}}_w(\mathbf{p}). \tag{4.35}$$

One interesting point of this approach is that the convergence results from the estimation technique for *stationary* colored noise also apply to this algorithm.

To test the noise estimation method, we consider again a denoising experiment. Figure 4.12(a) shows a low-dose Pathological Thorax CT image from a 15-year old female girl that was captured on a Siemens Emotion 6 CT Scanner at Sophia Children's Hospital, EMC in Rotterdam (the Netherlands), with scan parameters (KVP=110, 127mAs and using a B30s convolution kernel).

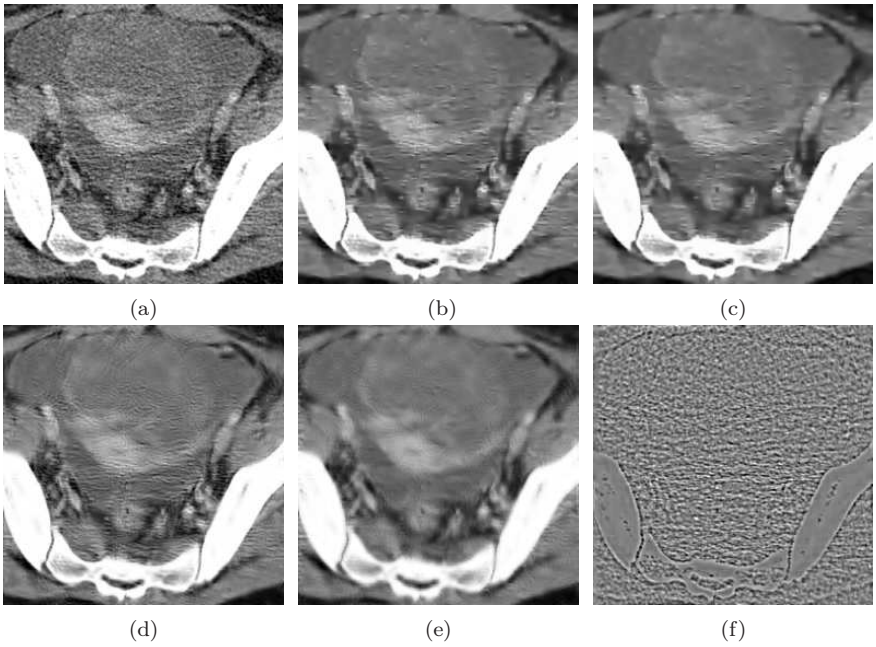


Figure 4.12: (a) Pathological Thorax Computed Tomography (CT) image of a 15-year old female. (b)-(c) Denoised versions of (a), using [Pižurica et al., 2003] with different threshold parameters. (d) Denoised version of (a), using [Portilla, 2004] (e) Denoised version of (a), using the proposed method. (f) Difference image between (d) and (a) (contrast enhanced, intensity 128 corresponds to difference zero).

The image suffers from noisy streak artifacts.[2] In Chapter 6, we will show that CT noise can be modeled as additive white Gaussian noise filtered by a space-variant filter. We compare the proposed noise estimation combined with the extension of the algorithm of [Goossens et al., 2009d], also discussed in Section 5.2.3 (Figure 4.12(d)) (see above), to the blind denoising methods of [Pižurica et al., 2003] (Figure 4.12(b)) and [Portilla, 2004] (Figure 4.12(c)). The method of [Pižurica et al., 2003] assumes white stationary noise and estimates the noise variance from the high-pass subband of the nondecimated spline wavelet transform. Due to the noise model mismatch, noise artifacts are left in the denoised image (Figure 4.12(b)) in areas where the local noise variance exceeds the estimated noise variance, whereas the proposed method does not. The method of [Portilla, 2004] assumes stationary correlated noise and also because of the non-stationarity, not all parts of the noise are removed. Our method uses the dual-tree complex wavelet transform from [Kingsbury, 2001], with 3 scales, $d = 9$ and $l(\mathbf{p}) = 16$. Figure 4.12(f) shows the difference image of Figure 4.12(d) and Figure 4.12(a). It can be noticed that some signal structures are present in the difference image, for example at the edges of the bright

[2]We will further discuss the topic of CT streak artifacts in Chapter 6.

areas in Figure 4.12(a). Here, due to the saturation in the scanner at intensity 255, there is a fast transition in the local noise variance. As a consequence, the local-stationarity assumption is violated and the local noise variance is slightly overestimated, causing oversmoothing of the edges. Nevertheless, the blind denoising method is able to remove the noise well in the organ regions, while preserving details better than the other methods. Alternatively, the saturation problem can be completely avoided by using the proper intensity windowing settings during CT acquisition. On Pentium IV 2 GHz processor, denoising a 256×256 image in an unoptimized implementation takes 143 s, from which 110 s are spent to noise estimation.

In the next section, we will drop the additivity assumption of the noise by considering signal-dependent noise.

4.4 Signal-dependent noise

From a theoretical point of view, dealing with signal-dependent noise is relatively easy: one only needs to model the joint density $f_{y,x}(y, x)$ of the noise-free pixel intensity x and a observed noisy pixel intensity y. Statistical methods, e.g. denoising, then directly follow. For example, for denoising the MMSE estimator is given by $\hat{x}_{\text{MMSE}} = \mathrm{E}[x|y] = \int_{-\infty}^{+\infty} x f_{x|y}(x|y)\, dx$ and the MAP estimator is simply $\hat{x}_{\text{MAP}} = \arg\max_x f_{x|y}(x|y)$. Unfortunately, there are a number of problems with this approach:

- Exact analytical expressions for the joint density $f_{y,x}(y, x)$ are often complicated in real-life situations and closed-form expressions do not always exist. If the input-output relationship between x and y are known in the noise-free case and if the noise sources can be simulated, a possible way to proceed is to use Monte Carlo (MC) techniques. However, it then becomes more difficult to study the influence of parameter changes, as this would require extra MC simulations. On the other hand, if the input-output relationship is not known, density estimation techniques need to be used. Again, studying the effects of parameter choices is not trivial.

- In many practical situations, the noise is also spatially correlated (on top of being signal-dependent), as we already explained. The noise modeling and estimation task then becomes significantly more difficult as we need to find the high-dimensional joint density of either the complete image or patches of this image ($f_{\mathbf{y},\mathbf{x}}(\mathbf{y}, \mathbf{x})$).

As a workaround to these issues, we will seek for approximative descriptions of signal-dependent noise. We will show that these kind of descriptions, despite being very simple, are often very accurate in practice and serve well for our needs (e.g. for image restoration, see further in Chapter 5). Moreover, the approximative descriptions also lead to much easier processing techniques, as we will show later.

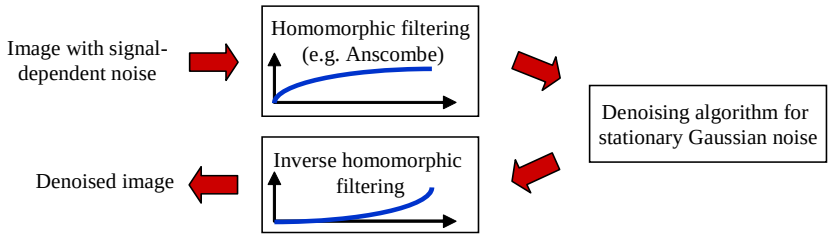


Figure 4.13: The principle of removal of signal-dependent noise using homomorphic filtering.

4.4.1 Variance stabilization

Variance stabilization (VS) aims at transforming signal-dependent noise into (approximately) signal-*in*dependent additive noise. VS techniques are very popular for dealing with signal-dependent noise because any denoising scheme designed for stationary AWGN can be used. Homomorphic filtering [Ding and Venetsanopoulos, 1987] is a VS technique that applies an invertible monotonically increasing non-linear function $\bar{\gamma}(\cdot)$ to every noise intensity y:

$$y|x \sim \mathcal{N}\left(\bar{\gamma}\left(x\right); \sigma_0^2\right)$$

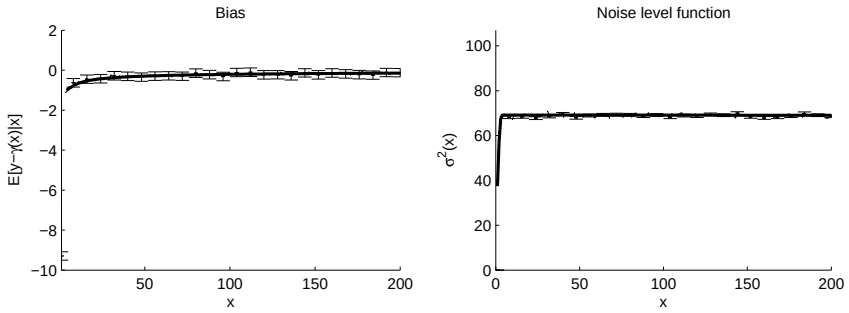
with σ_0^2 a constant. A well-known example of such a transform is the Anscombe transformation $\bar{\gamma}(x) = 2\sqrt{x + 3/8}$ for Poisson noise [Anscombe, 1948]. In general, if the standard deviation of x as function of the mean $\sigma(x)$ is known, the non-linear function can be computed using the indefinite integral [Foi, 2008]:

$$\bar{\gamma}\left(x\right) = \int \frac{\sigma_0}{\sigma(x')} \mathrm{d}x'.$$

Once the variance stabilizing transform is available, the removal of signal-dependent noise is performed through a three-step proceduce (see Figure 4.13): 1) normalize the noise variance by applying the homomorphic filtering, 2) use a denoising algorithm designed for stationary AWGN and 3) apply the inverse homomorfic filtering to obtain an estimate of the original, noise-free image.

The disadvantage of variance stabilization is that the signal model assumed for x may not hold for $\bar{\gamma}\left(x\right)$ and that very often the optimality of the estimator in signal space diminishes in homomorphic transform space [Hirakawa, 2008b].

As an experiment to assess the quality of the variance stabilization, we computed the bias $\mathrm{E}\left[y|x\right] - \bar{\gamma}\left(x\right)$ introduced by the Anscombe transform for Poissonian distributed data, using both Monte-Carlo simulations and an exact computation method (that will be explained in the following sections). Similarly, we computed the variance of the data after homomorphic filtering. The results are shown in Figure 4.14. Here, we used a modified version of the Anscombe transform in which the output is linearly mapped onto the range $[0, 255]$ (this does



(a) (b)

Figure 4.14: Experimental result for the Anscombe transform of Poissonian distributed data (a) Signal bias after homomorphic filtering, (b) Noise variance after homomorphic filtering.

not affect the bias and variance results, upon scaling). It can be noted that the bias and variance are near-constant, but for large input intensities. For small input intensities the variance of the output decreases and the bias increases in magnitude. We can conclude that the variance is stabilized, but not perfectly. Moreover, a bias error is introduced when the variance stabilizing transform is being inverted (e.g. through the algebraic inverse $\bar{\gamma}^{-1}(x)$). The solution is then to derive an unbiased inverse transform [Anscombe, 1948, Mäkitalo and Foi, 2009]. For example, for the Anscombe transform, an asymptotically unbiased inverse transformation is given by: $\bar{\gamma}^{-1}(x) = (x/2 - 1/8)^2$.

4.4.2 Gaussian modeling of signal-dependent noise

To overcome the limitations of variance stabilization, we will investigate explicit models for signal-dependent noise. Therefore, we start from a simple model that describes the functional relation between a noisy pixel intensity y, the noise-free pixel intensity $x \in [x_{\min}, x_{\max}]$ and a noise sample w:

$$y = \zeta(x, w), \tag{4.36}$$

where w is Gaussian distributed $w \sim \mathcal{N}(0, 1)$ and $\zeta(\cdot)$ is a *nonlinear* function that "mixes" the signal and the noise. Next, certain restrictions apply to $\zeta(x, w)$: firstly, let $\gamma(x) = \zeta(x, 0)$ denote the intensity mapping function (i.e. in absence of noise, $y = \gamma(x)$) and let us assume that $\gamma(x)$ is continuous and monotonic on $x \in [x_{\min}, x_{\max}]$. Consequently, the inverse function $\gamma^{-1}(y)$ exists for $y \in [\gamma(x_{\min}), \gamma(x_{\max})]$. Secondly, $\zeta(x, w)$ is analytic in a small interval around $w = 0$, such that the Maclaurin series

$$\zeta(x, w) = \sum_{n=0}^{+\infty} \left. \frac{\partial \zeta}{\partial w} \right|_{w=0} w^n \tag{4.37}$$

converges on this interval. This allows for a first-order MacLaurin series approximation:

$$y \approx \gamma(x) + \left.\frac{\partial \zeta}{\partial w}\right|_{w=0} w, \tag{4.38}$$

from which the conditional mean $\mathrm{E}[y|x]$ and variance $\mathrm{Var}[y|x]$ can be easily obtained as:

$$\mathrm{E}[y|x] \approx \gamma(x), \tag{4.39}$$

$$\mathrm{Var}[y|x] \approx \left(\left.\frac{\partial \zeta}{\partial w}\right|_{w=0}\right)^2. \tag{4.40}$$

More specifically, the linearization from (4.37) causes $y|x$ to be Gaussian distributed with mean and variance as above. We will call this approach the *Gaussian modeling of signal-dependent noise* (as in [Foi et al., 2008]). Remark that both $(\mathrm{E}[y|x] - x)$ and $\mathrm{Var}[y|x]$ are generally a function of x, hence the signal-dependency of the noise translates into 1) a bias $\gamma(x) - x$ (as $\mathrm{E}[y|x] \neq 0$) and 2) a signal-dependent variance $\mathrm{Var}[y|x]$.

Equation (4.36) generalizes a number of noise models used in the literature. For example:

- An additive Gaussian noise model can be obtained as:

$$\zeta(x, w) = x + \sigma w, \tag{4.41}$$

  where $\sigma = \sqrt{\mathrm{Var}[y|x]}$ is the noise standard deviation.

- A multiplicative noise model, which has been used, e.g., to model speckle noise in ultrasound images (see [Achim et al., 2001a], [Pižurica, 2002, p. 158]) is given by:

$$\zeta(x, w) = (x + 1)w. \tag{4.42}$$

- A mixed first order additive-multiplicate noise model [Hirakawa and Parks, 2005b] is found by choosing:

$$\zeta(x, w) = x + (a_1 + a_2 x)\, w. \tag{4.43}$$

  with a_1 the standard deviation of the signal-independent noise component and with a_2 the multiplicative noise component gain factor.

- A related mixed additive-multiplicate noise model [Lim, 2006] is defined by:

$$\zeta(x, w) = x + \sqrt{a_1 + a_2 x + a_3 x^2}\, w, \tag{4.44}$$

  where the constant a_1 is the variance of *read-out noise* generated by the sensors in the camera, a_2 is a *photon noise* gain factor and a_3 is the variance of *pattern noise*.

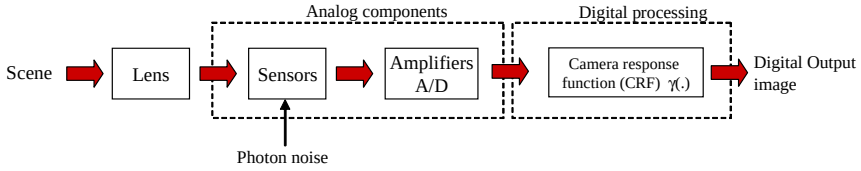


Figure 4.15: Simplified pipeline for a digital still camera (compare to Figure 4.1). Also see text.

- The signal-dependent CCD noise model from [Faraji and MacLean, 2004] is given by:
$$\zeta(x, w) = x + \sigma x^a w \tag{4.45}$$
where σ and a are constants that depend on the CCD photon transfer curve.

We remark that the above noise models rely on an identity intensity response mapping function $\zeta(x, 0) = \gamma(x) = x$. To obtain noise models with non-identity response mapping functions, x can be replaced by $\gamma(x)$ in the right-handed sides of the above equations (4.41)-(4.45).

Modeling the Digital Still Camera pipeline

Consider the simplified model for noise in a digital camera from Figure 4.15: a Poisson distributed intensity, captured by a sensor element in a DSC, is amplified and subjected to a nonlinear camera response function (CRF) $\gamma(x)$. Camera response functions [Liu et al., 2006a], which are also widely used in high dynamic range (HDR) imaging [Debevec and Malik, 1997,De Neve et al., 2009], account for different (pointwise) post-processing steps in the DSC[3], such as contrast enhancement, gamma correction, ... For noise-free images, the role of the CRF function is the same as the role of the intensity mapping function (i.e. $\zeta(x, w) = \gamma(x, 0)$). In Figure 4.15 we assume that the DSC has sufficiently been well-engineered, such that the contributions of read-out noise and pattern noise can be ignored. However, what follows can be easily extended to include these additional noise terms. By using Gaussian modeling, we obtain the following "mixing" function:
$$\zeta(x, w) = \gamma\left(x + \sqrt{x}w\right), \tag{4.46}$$
for which the first and second order conditional statistical moments are given by:
$$\mathrm{E}\left[y|x\right] \approx \gamma(x), \tag{4.47}$$
$$\mathrm{Var}\left[y|x\right] \approx |x| \left(\frac{\mathrm{d}\gamma}{\mathrm{d}x}\right)^2. \tag{4.48}$$

[3]We will later explain how to deal with non-pointwise operations.

Table 4.1: Various camera response function (CRF)-noise level function (NLF) pairs.

Type	CRF $\gamma(x)$	NLF $\sigma(x)$
Gamma correction	x^α	$\frac{\alpha x^\alpha}{\sqrt{x}}$
Contrast enhancement	ax	$a\sqrt{x}$
Logarithm	$\log x$	$\frac{1}{\sqrt{x}}$
"Gamma" NLF	$-\frac{\sqrt{x}e^{-\alpha x}}{\alpha} + \frac{1}{2}\frac{\sqrt{\pi}\mathrm{erf}(\sqrt{\alpha x})}{\alpha^{3/2}}$	$x \cdot \exp(-\alpha x)$
Composite CRF	$\gamma_2\left(\gamma_1(x)\right)$	$\sqrt{x}\frac{\partial \gamma_2}{\partial \gamma_1}\left(\gamma_1(x)\right)\frac{\partial \gamma_1}{\partial x}$

Let $\sigma(x) = \sqrt{\mathrm{Var}\left[y|x\right]}$ denote the standard deviation of the noise as function of the "ideal" intensity x (which we will refer to as the noise level function, NLF[4]), then based on (4.48) we find the relationship between the NLF and the CRF:

$$\sigma(x) \approx \sqrt{x}\frac{\mathrm{d}\gamma}{\mathrm{d}x} \text{ and } \gamma(x) \approx \int \frac{\sigma(x)}{\sqrt{x}}\mathrm{d}x + C. \tag{4.49}$$

Hence, if we know the CRF (e.g. because we have information about the exact algorithms being used in the DSC), we can compute the NLF. On the other hand, if we can somehow estimate the NLF (see Section 5.3.5), the CRF is determined up to a constant C. The latter case corresponds to *reverse-engineering*: the DSC is considered to be a black-box; based on the noise characteristics of acquired images, information about the internal processing is obtained (through the CRF). In Table 4.1, a number of CRF-NLF pairs are listed, for a number of components of the DSC pipeline (Figure 4.1).

Equation (4.49) is also useful when considering composite CRFs $\gamma(x) = \gamma_2\left(\gamma_1(x)\right)$ (see bottom row of Table 4.1): by the chain rule for derivatives, the resulting NLF is basically the product of the NLFs corresponding to respectively γ_1 and γ_2, with appropriate scalings and warpings. This easily allows to find the NLF for a combination of processing algorithms. We illustrate this through a simple example: suppose we want to compute the NLF for gamma correction followed by a contrast enhancement. Following Table 4.1, the respective NLFs are:

$$\sigma_1(x) = \frac{\alpha x^\alpha}{\sqrt{x}} \quad \text{and} \quad \sigma_2(x) = a\sqrt{x} \tag{4.50}$$

The composite NLF can then be expressed in terms of the NLFs of the individual operations:

$$\sigma(x) = \frac{\sigma_2(\gamma_1(x))}{\sqrt{\gamma_1(x)}}\sigma_1(x) \tag{4.51}$$

$$= a\frac{\sqrt{x^\alpha}}{\sqrt{x^\alpha}}\frac{\alpha x^\alpha}{\sqrt{x}} = \frac{a\alpha x^\alpha}{\sqrt{x}} \tag{4.52}$$

[4]We remark that the NLF is sometimes defined as a function of the *processed* intensity $x' = \gamma(x)$, while here it is a function of the "RAW" pixel intensity x. Caution is advised not to mix both definitions.

If we would change the order of the NLFs, the composite NLF becomes:

$$\sigma(x) = \frac{\alpha\,(ax)^{\alpha}}{\sqrt{x}} \tag{4.53}$$

which is a different expression than (4.52). Hence the NLF is highly dependent on the order of the invididual CRFs.

We emphasize that the equations (4.47)-(4.48) are only *approximate* and only take the photon noise component into account. To assess the accuracy of the approximation, we performed a number of Monte-Carlo experiments. Therefore, we generated 2000×256 Poisson distributed random variables with gradually increasing mean in the range $[0, 255]$. Next, we applied several CRFs to these variables and we estimated the variance. The results are shown in Figure 4.16. The first and second CRFs simulate the dynamic range compression present in most cameras [Mann, 2000], to pre-compensate the intensity response curve of Cathode Ray Tube (CRT) monitors. It can be noted that in these examples, the approximation (4.48) is quite accurate. However, this is not always the case, as we will explain next.

Assessing the inaccuracy of the model

Our initial requirement that $\zeta\,(x, w)$ must analytic in a small interval around $w = 0$, is too strong in a number of cases. The requirement implies that the CRF $\gamma(x)$ must be analytic for $x \in [x_{\min}, x_{\max}]$, which is not always the case in practice. Therefore, consider the CRF-NLF pair (for $x \in [0, 255]$):

$$\begin{aligned} \gamma(x) &= \int \frac{2\mathrm{d}x}{(1+\exp\left(\alpha\,(x_1-x)\right))\,(1+\exp\left(-\alpha\,(x_2-x)\right))}, \\ \sigma(x) &= \frac{2\sqrt{x}}{(1+\exp\left(\alpha\,(x_1-x)\right))\,(1+\exp\left(-\alpha\,(x_2-x)\right))}. \end{aligned} \tag{4.54}$$

with $0 < x_1 < x_2 < 255$ and $\gamma(0) = 0$. The above CRF simulates "soft" saturation between $[x_1, x_2]$. As the parameter $\alpha \to \infty$, the following saturation curve is obtained:

$$\gamma_{\mathrm{sat}}(x) = \begin{cases} 0 & x < x_1 \\ (x - x_1)/(x_2 - x_1) & x_1 \le x < x_2 \\ 1 & x_2 \le x \end{cases}$$

It can be shown that $\gamma(x)$ is analytic for $x \in [0, 255]$, although it approaches $\gamma_{\mathrm{sat}}(x)$ for $\alpha \to \infty$, which is not analytic for $x = x_1$ and $x = x_2$. In Figure 4.17, the Monte-Carlo experiment is repeated for the CRF-NLF from (4.54) for $x_1 = 40$ and $x_2 = 216$ and for two different values of α. As the parameter α increases, the CRF becomes less smooth in $x \in \{x_1, x_2\}$. More specifically, it can be checked that the magnitude of the second derivative of $\gamma(x)$ in $x \in \{x_1, x_2\}$ increases approximately linearly. Consequently, the approximations (4.47) and (4.48) become less accurate, as Figure 4.17(b) illustrates.

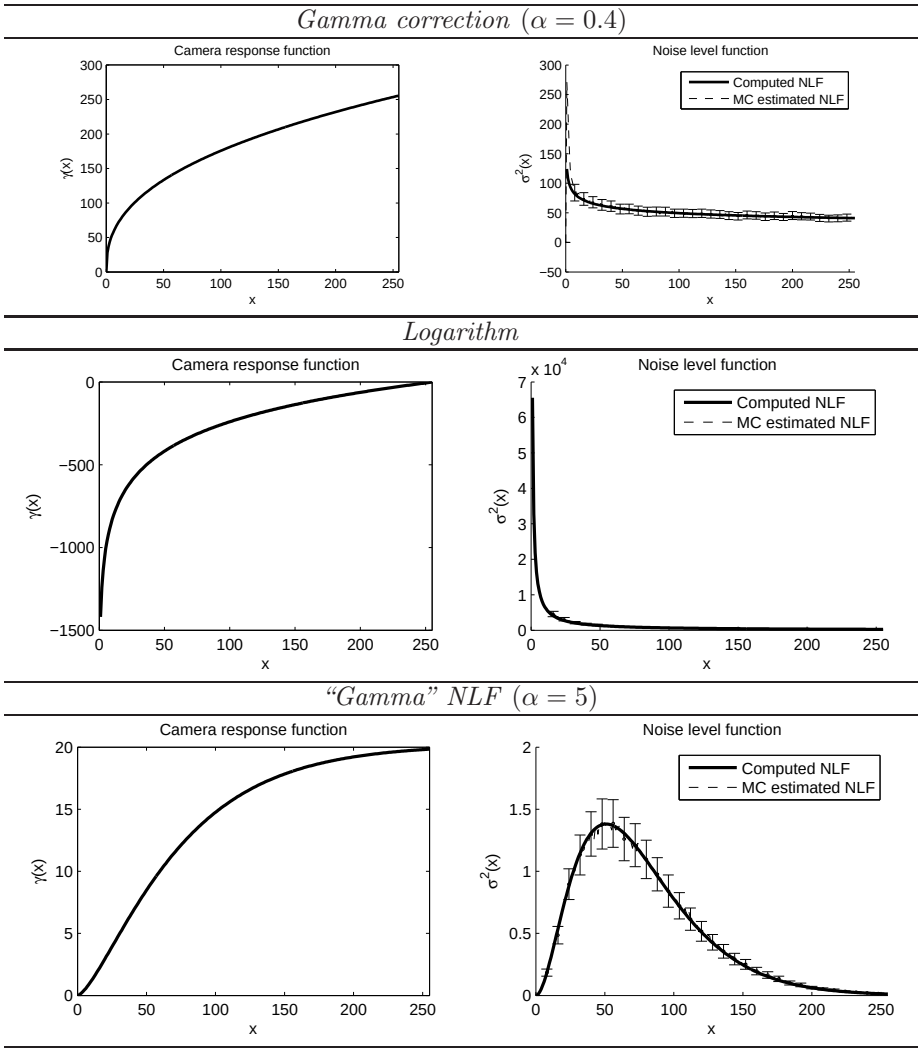


Figure 4.16: Illustration of the accuracy of the approximation $\sigma^2(x) \approx |x| \left(\frac{\partial \gamma}{\partial x}\right)^2$, for different camera response functions listed in Table 4.1. *Note: results are obtained in the absence of saturation.*

However, the most severe limitation is in the bias estimation: according to (4.55), the bias predicted by the model $\mathrm{E}[y|x] - \gamma(x)$ is always 0. In case of saturation, this is certainly not valid: e.g. by the clipping of the high intensity values that are corrupted with noise, the conditional mean $\mathrm{E}[y|x]$ is shifted to the left: $\mathrm{E}[y|x] < \gamma(x)$. A possible solution is to compute more terms of the

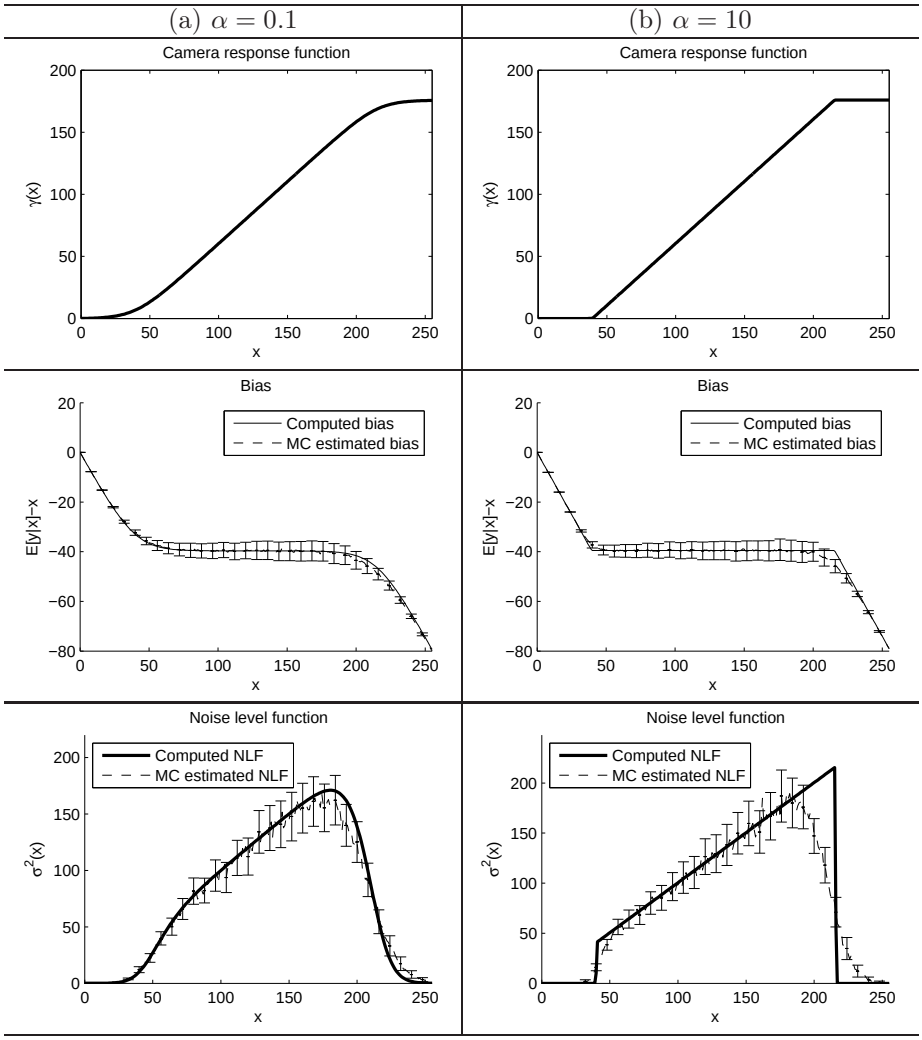


Figure 4.17: Illustration of the *in*accuracy of the approximation $\sigma^2(x) \approx |x| \left(\frac{\mathrm{d}\gamma}{\mathrm{d}x}\right)^2$, for non-analytic CRFs.

McLaurin series:

$$y \quad \approx \quad \gamma(x) + \sum_{n=1}^{N} \left.\frac{\partial^n \zeta}{\partial w^n}\right|_{w=0} w^n$$

The first order conditional statistical moment becomes:

$$\begin{aligned} \mathrm{E}[y|x] &\approx \gamma(x) + \sum_{n=1}^{N} \left.\frac{\partial^n \zeta}{\partial w^n}\right|_{w=0} \frac{\mathrm{E}[w^n]}{n!} \\ &= \gamma(x) + \sum_{n=1}^{N} \left.\frac{\partial^{2n} \zeta}{\partial w^{2n}}\right|_{w=0} \frac{2^{-n}}{n!} \end{aligned} \tag{4.55}$$

where we used statistical moments of the Gaussian distribution:

$$\mathrm{E}[w^n] = \begin{cases} \frac{n!}{2^{n/2}(n/2)!} & n \text{ even} \\ 0 & n \text{ odd} \end{cases}$$

In (4.55), the bias $\mathrm{E}[y|x] - \gamma(x)$ will generally be non-zero. As predicted by the "soft saturation" experiment, the absolute value of the bias $|\mathrm{E}[y|x] - \gamma(x)|$ increases with the magnitude of the higher order derivatives $\left|\partial^{2n}\zeta/\partial w^{2n}\right|_{w=0}$. Given the fact that $2^{-n}/n!$ drops quickly to zero as $n \to \infty$, the most significant term will be the second derivative.

We can conclude that the Gaussian modeling technique using an N-term McLaurin series-based computation of the conditional moments gives accurate results only if the higher-order derivatives $\left|\partial^{2n}\zeta/\partial w^{2n}\right|_{w=0}$ are sufficiently bounded - which is not the case near the saturation points of $\gamma(x)$.

4.4.3 Exact conditional moments $\mathrm{E}[y^n|x]$

In this section, we will present a computation technique that does not require that the mixing function $\zeta(x, w)$ is analytic around $w = 0$, and that can be used in the presence of saturation. Note that equation (4.36) can be seen as a nonlinear coordinate transform:

$$\begin{cases} x' &= x \\ y &= \zeta(x, w) \end{cases}.$$

Now, assuming that this coordinate transform is *invertible* on $(x, w) \in \Omega$, we can write an inverse relationship:

$$\begin{cases} x' &= x \\ w &= \zeta'(x', y) \end{cases},$$

where $\zeta'(x', w)$ satisfies $\zeta'(x, \zeta(x, w)) = w$ for $(x, w) \in \Omega$. The conditional density $f_{y|x}(y|x)$ can be written in this new coordinate system:

$$\begin{aligned} f_{y|x}(y|x) &= \frac{f_{x,y}(x, y)}{f_x(x)} \\ &= \left|\frac{\partial \zeta'}{\partial y}\right| \frac{f_{x,w}(x, \zeta'(x, y))}{f_x(x)} \\ &= \left|\frac{\partial \zeta'}{\partial y}\right| f_{w|x}(\zeta'(x, y)|x) \end{aligned}$$

where $|\partial\zeta'/\partial y|$ is the magnitude of the determinant of the Jacobian matrix for the coordinate transform. Based on the above definitions, we can exactly compute the conditional noncentral moments $M_n(x) = \mathrm{E}\left[(y-\gamma(x))^n|x\right]$ as:

$$M_n(x) = \int_{-\infty}^{+\infty} \left|\frac{\partial\zeta'}{\partial y}\right| (y-\gamma(x))^n f_w\left(\frac{\gamma^{-1}(y)-x}{\sqrt{x}}\right) \mathrm{d}y. \quad (4.56)$$

We derive the *non*central moments here (remark that in general $\mathrm{E}[y] \neq \gamma(x)$), because we found that this yields simpler analytical expressions for $M_n(x)$.[5] From these moments, the conditional mean and variance are readily obtained as:

$$\mu(x) = \mathrm{E}[y|x] = M_1(x) + \gamma(x) \quad (4.57)$$

$$\sigma^2(x) = M_2(x) - 2(\mu(x) - \gamma(x))\mu(x) + \mu^2(x) - \gamma^2(x) \quad (4.58)$$

The major advantage of these equations is that their *exactness*; only required is that the partial derivative $\partial\zeta'/\partial y$ exists. However, based on (4.58) it is not possible to derive a bidirectional relationship between the NLF $\sigma^2(x)$ and the $\gamma(x)$ (note in this respect that $M_n(x)$ even depends on the inverse $\gamma^{-1}(y)$).

Gaussian-Poisson modeling

Now, we will consider again a Poisson distributed random variable $x$that is sent through the CRF $\gamma(x)$, where the response of the CRF is saturated at $x = 0$ and $x = 255$ ($\gamma(x) = 0$ for $x \leq 0$ and $x \geq 255$). We further assume that the CRF is a continuous monotonic (invertible) function on the interval $x \in [0, 255]$. To simplify the equations, we use the Gaussian-Poisson approximation, which here leads to:

$$f_{w|x}(w|x) = f_w(w) = (2\pi)^{-1/2}\exp\left(-w^2/2\right), \quad (4.59)$$

Consequently, we can write:

$$\zeta(x,w) = \begin{cases} \gamma(x+\sqrt{x}w) & 0 \leq x+\sqrt{x}w \leq 255 \\ 0 & x+\sqrt{x}w < 0 \\ 255 & 255 < x+\sqrt{x}w \end{cases} \quad (4.60)$$

The function $\zeta(x,w)$ is invertible in w for $0 \leq x+\sqrt{x}w \leq 255$, see Figure 4.18.

Next, we define $\zeta'(x,y)$ such $\zeta'(x,\zeta(x,w)) = w$, but in a way that $\zeta'(x,y)$ is invertible for $y \in \mathbb{R}$:

$$\zeta'(x,y) = \begin{cases} \frac{\gamma^{-1}(y)-x}{\sqrt{x}} & 0 \leq y \leq 255 \\ y & \text{else} \end{cases} \quad (4.61)$$

[5] In case central moments are needed, conversion formulas can be used.

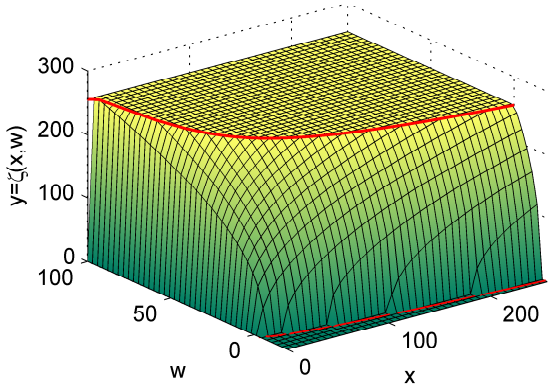


Figure 4.18: Signal and noise mixing function $\zeta(x, w)$ according to (4.60), for the gamma-correction $\gamma(x) = 255\,(x/255)^{0.4}$ with $x \in [0, 255]$ and including saturation to the range $[0, 255]$. The thick lines delineate the regions $x + \sqrt{x}w = 0$ and $x + \sqrt{x}w = 255$.

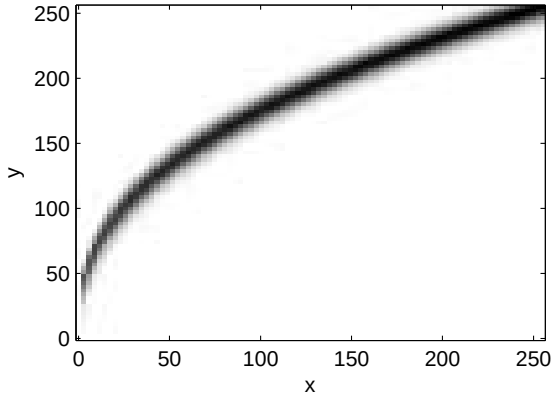


Figure 4.19: Conditional probability density function $f_{y|x}\,(y|x)$ for the mixing function from Figure 4.18. Black corresponds to high probabilities, white to low probabilities.

The determinant of the Jacobian matrix for the coordinate transform (see above) is given by:

$$\frac{\partial \zeta'}{\partial y} = \begin{cases} \frac{1}{\sqrt{x}} \frac{\mathrm{d}\gamma^{-1}}{\mathrm{d}y} & 0 \le y \le 255 \\ 1 & \text{else} \end{cases} \tag{4.62}$$

Based on (4.59)-(4.62), the conditional noncentral moments can be expressed

as:

$$
\begin{aligned}
M_n(x) &= \int_0^{255} \frac{(y-\gamma(x))^n}{\sqrt{x}} \left|\frac{\mathrm{d}\gamma^{-1}}{\mathrm{d}y}\right| f_w\left(\frac{\gamma^{-1}(y)-x}{\sqrt{x}}\right)\mathrm{d}y \\
&\quad + \int_{-\infty}^{0} \frac{(0-\gamma(x))^n}{\sqrt{x}} f_w\left(\frac{\gamma^{-1}(y)-x}{\sqrt{x}}\right)\mathrm{d}y \qquad (4.63) \\
&\quad + \int_{255}^{+\infty} \frac{(255-\gamma(x))^n}{\sqrt{x}} f_w\left(\frac{\gamma^{-1}(y)-x}{\sqrt{x}}\right)\mathrm{d}y. \qquad (4.64)
\end{aligned}
$$

Subsequently, the bias $\mathrm{E}[y-\gamma(x)|x]$ and NLF can be computed through (4.57)-(4.58). Hence, for a given CRF, we can compute the bias $\mathrm{E}[y|x]-\gamma(x)$, the variance $\mathrm{Var}[y|x]$ and even higher order statistical moments, as a function of the signal intensity in the presence of Poisson noise We will now discuss a number of special cases.

Example I: saturation of a signal in Poisson noise with an identity CRF

In case the CRF is the identity function on $[0,255]$ ($\gamma(x)=x$ for $x\in[0,255]$), closed-form analytical expressions can be found for $M_n(x)$:

$$
\begin{aligned}
M_1(x) &= \sqrt{\frac{x}{2\pi}}\left(e^{-\frac{1}{2}|x|} - e^{-\frac{(x-255)^2}{2|x|}}\right) + \left(\frac{255}{2}-x\right) \\
&\quad + \frac{255-x}{2}\mathrm{erf}\left(\frac{x-255}{\sqrt{2x}}\right) + \frac{x}{2}\mathrm{erf}\left(\sqrt{\frac{x}{2}}\right), \\
M_2(x) &= -\sqrt{\frac{x}{2\pi}}\left(xe^{-\frac{1}{2}|x|} + (255-x)e^{-\frac{(x-255)^2}{2|x|}}\right) + \frac{(255-x)^2}{2} + \frac{x^2}{2} \\
&\quad + \frac{|x|-x^2}{2}\mathrm{erf}\left(\sqrt{\frac{x}{2}}\right) - \frac{|x|-(x-1)^2}{2}\mathrm{erf}\left(\frac{x-1}{\sqrt{2x}}\right) \qquad (4.65)
\end{aligned}
$$

where $\mathrm{erf}(x) = 2\pi^{-1/2}\int_0^x e^{-t^2}\mathrm{d}t$ is the error function. In Figure 4.20(a), the bias function and NLF corresponding to these equations are shown and compared to MC simulations (in the same way as in Section 4.4.2). Even tough the MC simulations make use of "true" Poisson random variables while equation (4.65) relies on the Gaussian approximation of the Poisson distribution, the results are very accurate. The NLF increases linearly in the interval $[0,200]$ (as expected, as the variance of a poisson random variable is equal to its mean), but decreases around $x=230$ due to the saturation. A similar remark can be made for the bias: the bias is zero for $x<200$, but becomes negative for larger x due to the saturation. Saturation does not have any influence for small x-values, this is because Poisson random variables are always positive.

Example II: saturation of a signal in Poisson noise with gamma correction

As a second example, we consider a combination of saturation and gamma correction. The CRF is given by:

$$\gamma(x) = \begin{cases} 255\,(x/255)^{\alpha} & 0 \leq x \leq 255 \\ 0 & x < 0 \\ 255 & 255 < x \end{cases}$$

Unfortunately, for this problem, no closed-form analytical expressions exist for $M_n(x)$ (as far as the authors are aware of). Nevertheless, the integrals in (4.64) can still be computed through numerical integration techniques. The bias function and NLF are depicted in Figure 4.20(b). The nonlinearity of the CRF causes a bias for both small and large x values. The NLF is decreasing, although for $x < 30$ it is not predicted accurately by (4.65). Here, the discrepancy between the computed NLF and the estimated NLF by MC can be attributed to the Gaussian approximation of the Poisson distribution.

The results can be further improved by using the true Poisson density for $f_{y|x}\,(y|x)$ instead of the Gaussian density (i.e. by dropping the Gaussian-Poisson approximation). Because the Poisson density deals with discrete random numbers (and consequently y is a discrete number), the integrals in (4.64) will need to be replaced by a sum (over possibly an infinite number of terms). This leads to expressions that are even less analytically tractable, hence we need turn to numerical evaluation. Doing this, we found that the technique suffers from overflow errors. Fortunately, these errors can be partially avoided by working with the *logarithm* of the Poisson density (using numerical techniques to compute the log-gamma function), thereby translating products into sums and applying the exponential function in the final stage.

4.4.4 Possible extensions to the signal-dependent noise models

As mentioned in Section 4.4.2, the Gaussian-Poisson modeling does not take noise sources other than photon noise into account. To incorporate pattern noise and read-out noise (as described in [Lim, 2006]) into the model, we can extend the Gaussian-Poisson modeling as follows (see (4.44)):

$$\zeta(x, w) = \gamma\left(x + \sqrt{a_1 + a_2 x + a_3 x^2}\, w\right) \tag{4.66}$$

Following the same reasoning as in Section 4.4.2, we find the following approximation for the NLF:

$$\sigma(x) = \sqrt{a_1 + a_2 x + a_3 x^2}\,\frac{\mathrm{d}\gamma}{\mathrm{dx}}.$$

An extra complication in practice is that the constants a_1, a_2, a_3 depend on various camera settings, such as the exposure time, ISO setting, automatic gain

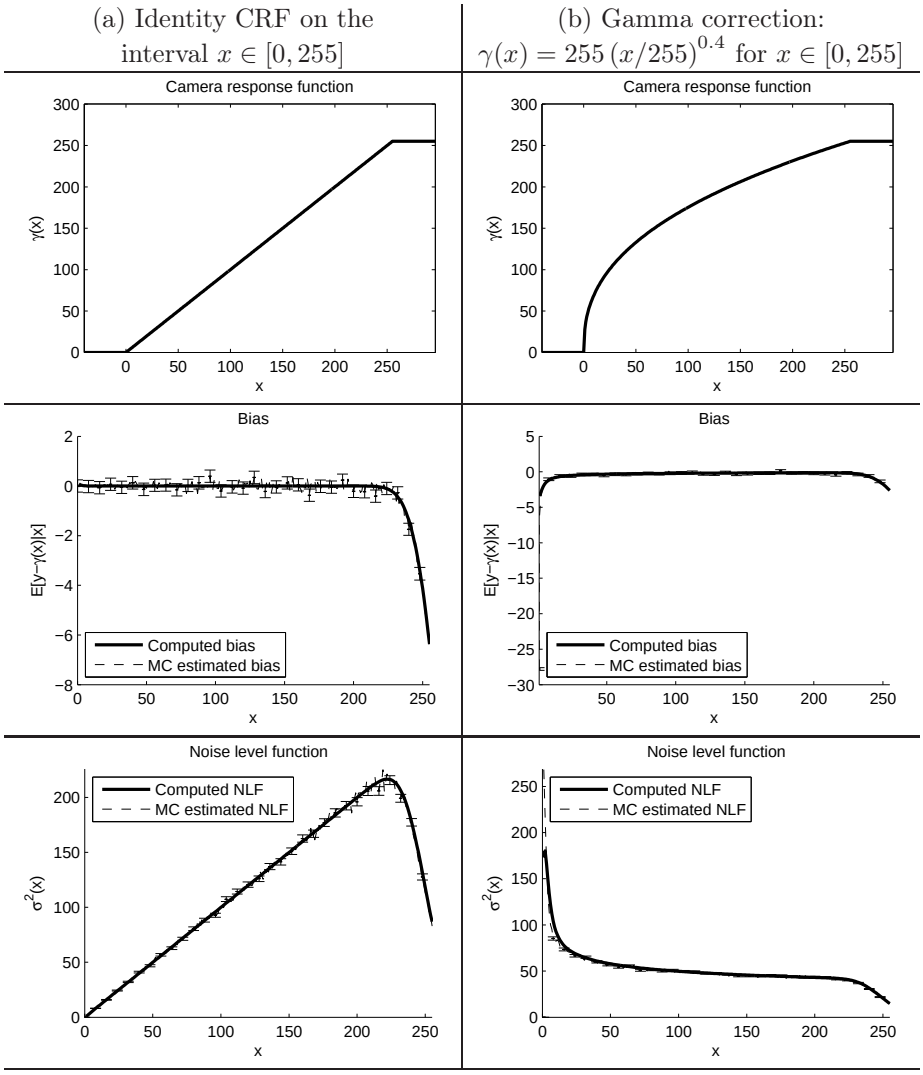


Figure 4.20: Results for the exact computation of the bias $\mathrm{E}\left[y - \gamma\left(x\right)|x\right]$ and NLF $\mathrm{Var}\left[y|x\right]$

control, etc. In case the CRF $\gamma(x)$ is known exactly and with an appropriate estimation technique for $\sigma(x)$ and x (see Section 5.3.5), these parameters can be estimated from the image through a linear regression.

Furthermore, the CRF is assumed to be a pixelwise nonlinear operation, which holds for a number of post-processing steps such as contrast enhancement and gamma correction, but not for steps involving color operations:

- *Color correction*: is used to compensate cross-color bleeding in the im-

age sensor CFA [Nakamura, 2005], e.g., caused by cross-talk [Hirakawa, 2008a]. Color correction typically applies a fixed pixel-wise linear transform to the R,G,B intensities in the image.

- *White balance correction*: the goal of white balancing techniques is to give the DSC a reference to white in the captured image. Once the white point is estimated from the image, the colors are converted to the correct ones. Simple techniques correct the average pixel luminance values, while more sophisticated techniques rely on chromatic adaptation and color constancy techniques [Nakamura, 2005]. The latter involves a pixel-wise nonlinear transform to the R,G,B intensities.

To deal with these situations, a possible solution is to extend the signal-dependent noise model from (4.36) to vectors (in which each component denotes color component R/G/B):

$$\mathbf{y} = \boldsymbol{\zeta}\left(\mathbf{x}, \mathbf{w}\right) \tag{4.67}$$

where the noise $\mathbf{w} \sim \mathcal{N}\left(0, \mathbf{I}_3\right)$ is spherically Gaussian distributed. A vector-valued extension of the CRF can be defined as:

$$\boldsymbol{\gamma}(\mathbf{x}) = \boldsymbol{\zeta}\left(\mathbf{x}, \mathbf{0}\right). \tag{4.68}$$

Next, a Maclaurin series approximation can be applied (under the same conditions as mentioned in Section 4.4.2):

$$\mathbf{y} \approx \boldsymbol{\gamma}(\mathbf{x}) + \sum_{c=1}^{3} \left.\frac{\partial \boldsymbol{\zeta}}{\partial w_c}\right|_{w_c=0} w_c \tag{4.69}$$

where w_c is the cth component of $\mathbf{w}$. The conditional mean $\mathrm{E}\left[\mathbf{y}|\mathbf{x}\right]$ and covariance $\mathrm{Var}\left[\mathbf{y}|\mathbf{x}\right]$ are readily obtained as:

$$\mathrm{E}\left[\mathbf{y}|\mathbf{x}\right] \approx \boldsymbol{\gamma}(\mathbf{x}) \tag{4.70}$$

$$\mathrm{Var}\left[\mathbf{y}|\mathbf{x}\right] \approx \sum_{c,c'=1}^{3} \left.\frac{\partial \boldsymbol{\zeta}}{\partial w_c}\right|_{w_c=0} \left(\left.\frac{\partial \boldsymbol{\zeta}}{\partial w_{c'}}\right|_{w_{c'}=0}\right)^T \tag{4.71}$$

It is clear that the resulting noise models are considerably more complicated, as now signal-dependent cross-channel correlations are taken into account. Nevertheless, the vector-extension relies on the same principles as those we presented above.

4.4.5 Estimation of signal-dependent noise

Before closing this chapter, we give a few notes on how signal-dependent noise in images can be estimated, without knowledge of the CRF. In case both an original image and a noisy image are available, we can:

- estimate the joint density $f_{x,y}\left(x, y\right)$ as the joint histogram of x and y. Subsequently, the conditional moments $\mathrm{E}\left[\left(y - \mathrm{E}\left[y\right]\right)^n |x\right]$ can be computed numerically.

- fit a noise level function $\sigma(x)$ to the data, e.g., by linear regression techniques (see e.g. [Portilla, 2005, Liu et al., 2006b]).

In Section 5.3.5, we will discuss an alternative technique that is able to estimate the noise level function $\sigma(x)$ directly from an observed image.

4.5 Conclusion

Due to physical limitations of the acquisition, noise in digital images can generally not be avoided. Accurate modeling of noise in images is therefore very important for image processing applications. In this chapter, we first discussed a number of noise sources for digital images and the corresponding marginal statistics. Next, we focused on stationary noise processes. We explained that in practice individual noise samples are correlated. The noise Power Spectral Density (PSD) describes the noise power distribution in frequency and is suited to describe the noise correlations. We investigated the noise PSD in a number of practical situations such as PAL television, color interpolation, thermal cameras, MRI. Thereafter, we presented a novel constrained EM algorithm for estimating the noise correlations from an observed image in a multiresolution transform domain. This EM algorithm was also extended to deal with locally stationary correlated image noise, in which the correlation structure of the noise samples varies slowly with the position in the image. We also considered signal-dependent noise and we explained the advantages and disadvantages of variance-stabilization transforms. To overcome some of the limitations, we provided an alternative approach that seeks for an approximative description of signal-dependent noise through conditional first and second order moments of the joint distribution of a noise-free pixel intensity and the observed noisy pixel intensity, which is called "Gaussian modeling" of signal-dependent noise. We discussed how the signal-dependency characteristic of the noise is altered by subsequent post-processing steps, such as a logarithmic transform, gamma correction and clipping. The resulting models are of great importance for image restoration techniques, that will be treated in Chapter 5. Finally, a few extensions to signal-dependent noise models for color images are presented.

The work presented in this chapter has so far resulted in one book chapter on the topic of colored noise [Aelterman et al., 2010b], one journal article on the computation of autocorrelation functions from complex wavelet subbands [Goossens et al., 2010a] and one conference publication on the estimation of locally stationary noise in images [Goossens et al., 2008c]. One journal publication on the presented constrained EM algorithm is in preparation [Goossens et al., 2010c].

5

Digital image restoration

Digital image restoration is the recovery of "original images" from observed degraded images. Many of the techniques developed for image restoration have origins in the fields of applied mathematics, estimation theory, linear algebra and numerical analysis [Banham and Katsaggelos, 1997]. In general, the recovery process comes down to solving a numerically complex *ill-posed* inverse problem [Kirsch, 1996]. According to [Hadamard, 1923], a problem is classified as ill-posed if either 1) a solution *does not exist* for certain input images, 2) the solution is *not unique*, or 3) the solution *differs significantly* when small perturbations are brought to the image. Image restoration techniques are ill-posed, due to for example,

- *Image noise*, which may result in images that are inconsistent with any natural scene [Luong, 2009]. In this case the original image cannot be obtained directly from its observation.

- *Blurring processes*, which may significantly attenuate Fourier components with higher spatial frequencies in the image, in a way that these components can not be recovered uniquely. By directly "undoing" the attenuation of these components, the noise will also amplified tremendously, causing the solution to become *instable*.

To overcome these problems, most restoration techniques apply a form of regularization [Kirsch, 1996]. A possible way to achieve regularization is to encode "general" prior knowledge about images and/or the degradation, in the form of an image model (see Chapter 3) and a degradation model (e.g. noise models from Chapter 4). Restoration techniques attempt to exploit all available information (i.e. the observed image itself and the prior knowledge) in order to maximize the quality of the restored image. This is in contrast to image enhancement methods which attempt to produce images that are pleasing to the observer, without use of an explicit degradation model. Another clan of related techniques are image reconstruction techniques. These are generally treated separately from restoration techniques, since reconstruction techniques typically operate on a subset of the data (e.g. demosaicing), despite solving

the same mathematical problem, i.e. solving systems of linear or nonlinear equations [Banham and Katsaggelos, 1997].

While image restoration techniques were initially developed for astronomical imaging or for improving aging and deteriorated films, today, the application area is much broader:

- *Digital photography*: as already explained in Chapter 4, digital still cameras and video camcorders inevitably produce noise. Noise removal is desirable not only to improve the visual quality of the images, but also to improve the performance of digital compression.

- *Video communication/transmission* (e.g. over the internet), where compression techniques are used to reduce the communication bandwidth. For low bandwidths, the compressed images often contain disturbing coding artifacts, such as quantization artifacts and block artifacts (e.g. JPEG, MPEG). Deblocking is the process of suppressing block artifacts, by smoothing over the block boundaries. Other examples are motion compensation (digital image stabilization) and motion blur compensation.

- *Medical imaging*: noise and aliasing often cause artifacts in reconstructed medical images (e.g. CT, MRI) which may potentially cause misdiagnosis (e.g. when the physician misses certain lesions in the images).

- *Biomedical imaging* [Rooms, 2005]: images obtained with optical systems, such as microscopes contain distortions, such as blurring and noise, which causes loss of actual information. The goal here is to recover some of this information, to facilitate post-processing, analysis...

- Other application areas include: the printing industry (inverse halftoning, [Hein and Zakhor, 1995]), assembly line manufacturing [Banham and Katsaggelos, 1997], industrial inspection and defense-oriented applications, such as the detection of land mines [Pižurica, 2002].

Many different application areas share the same problems. We briefly categorize several restoration problems that will be treated in this chapter:

- *Noise reduction (denoising)* is the process of suppressing or entirely removing noise from images. The noise is either introduced by the acquisition device or the transmission process. The goal of image denoising is therefore to reduce the noise while preserving original image details.

- *Deconvolution (deblurring)*: imaging systems are non-ideal, i.e. they images an ideal mathematical point as a smeared-out version of this point. The image of this ideal point is called the Point Spread Function (PSF) [Rooms, 2005]. Image deblurring is the process of "undoing" this PSF, either by exploiting prior knowledge with respect to the blur kernel or by jointly estimating the PSF and the deblurred image [Luong, 2009].

- *Demosaicing* (see Section 4.2): is the process of filling in missing pixel intensities in a color filter array (CFA) of a digital still camera. Strictly speaking, demosaicing is a reconstruction technique because it starts with a subset of the data. However, because demosaicing creates several artifacts (such as color and zippering artifacts), we will consider joint demosaicing and denoising as a restoration technique.

In this chapter, we will consider three different classes of image restoration techniques, as shown in Figure 5.1. In *image domain* methods, the image restoration filter directly operates in the image domain, by manipulation of the pixel intensities (without the use of multiresolution transforms or other transforms, such as the FFT). Examples of such methods are Total Variation classes [Rudin and Osher, 1994] (for image denoising), Bilateral filtering [Tomasi and Manduchi, 1998], anisotropic diffusion schemes (e.g. [Rudin et al., 1992,Chan et al., 2001]), Non-local means denoising [Buades. et al., 2005], etc. In *transform domain* methods, the restoration technique recovers the image by first performing a forward transform, then by processing of the transform coefficients and finally by applying the inverse transform. Examples are the Fourier domain Wiener filter, various Richardson-Lucy based restoration techniques [Rooms, 2005] and wavelet-based techniques. Finally, in *joint image domain/transform domain* methods, manipulate both image intensities and transform coefficients, typically in an iterative scheme.

We remark that from a mathematical point of view, a distinction between image domain restoration and transform domain restoration does not make much sense, since many (if not all) image domain restoration techniques have an equivalent implementation in transform domain and vice versa. However, the difference is *in the way the image model and degradation model are defined and are being used*: as we already discussed in Chapter 3, it is much easier to model prior information with respect to the undegraded image in a sparse (multiresolution) transform domain rather than directly in the image domain. On the other hand, the degradation model is generally easier to express directly in the image domain (see e.g. Chapter 4). Some techniques exploit the fact that the degradation model in the image domain easily translates into an equivalent wavelet domain degradation model (e.g. denoising methods for additive white Gaussian noise), and consequently the transform domain technique can optimally use all available information about the degradation.

However, this is not always trivial: for example, consider images containing missing pixels with known locations. An image domain degradation model simply keeps track of the locations of the missing pixels, while an equivalent wavelet domain degradation model would need to take into account that wavelet coefficients at different scales and orientations may be missing or are incorrect. Interpolation of these missing pixel intensities from e.g. neighboring pixel intensities is generally not as straightforward in the wavelet domain as in the image domain. A second example is blurring, which is defined as a convolution in the image domain. A translation of this degradation model to the wavelet domain inevitably leads to a multi-subband degradation model involving semi-

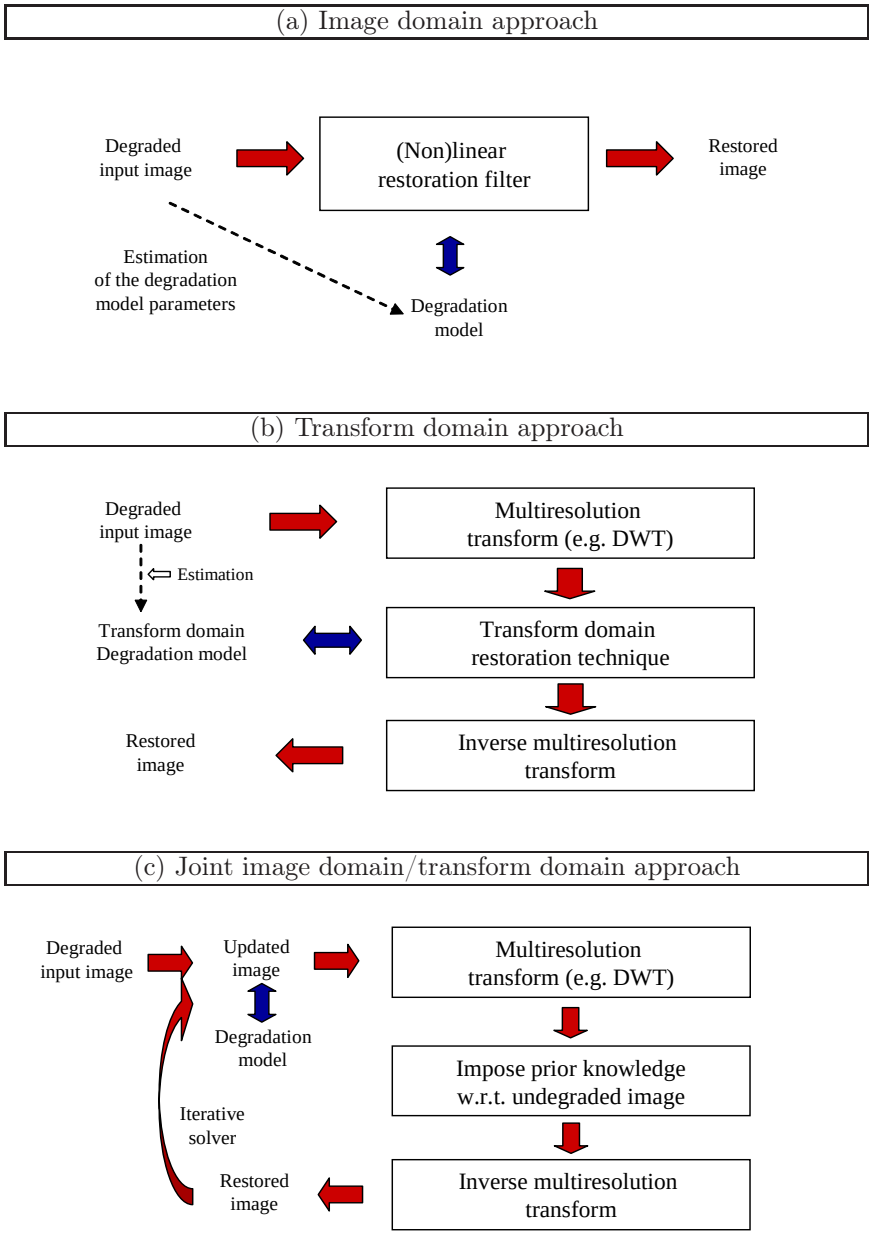


Figure 5.1: Three different designs of image restoration algorithms (a) Image domain techniques directly apply a restoration filter based on a degradation model in the image domain, (b) Transform domain techniques process the transform coefficients in order to estimate the original image, a transform domain degradation model needs to be used, (c) Joint domain techniques combine the advantages of the (a) and (b) designs.

block circulant matrices (see [Banham et al., 1994]), which is more complicated than the corresponding image domain model (a convolution kernel).

Although (multiresolution) transform domain techniques are generally better able to exploit sparseness properties of the underlying undegraded image, it is not necessarily true that transform domain techniques offer a better restoration performance (e.g. in SNR or visually) than image domain techniques. Quite often, unrealistic assumptions with respect to the degradation model (or image model) need to be made in order to obtain a reasonably practical transform domain algorithm. For example, optimally removing Poisson noise in the wavelet domain is considered to be a difficult task, due to the signal-dependency of the noise (there are some recent successes, see e.g. [Hirakawa, 2007]). As already mentioned in Section 4.4.1, many denoising techniques therefore rely on a variance stabilizing (VS) transform, such that the signal-dependent noise is translated into additive noise in the transform domain. The nonlinearity of the VS transform is then being completely ignored in the image model to keep the algorithms simple.

To obtain the best of both worlds, we will also investigate joint image domain/transform domain designs, in which the degradation is naturally modeled in the image domain, while the image model is defined in a multiresolution transform domain. This decoupling has the very important consequence that a restoration technique can be build by 1) selecting an appropriate multiresolution transform from Chapter 2, 2) defining an image model from Chapter 3 for the selected representation and 3) by using e.g. a degradation model from Chapter 4. The mechanism behind the joint transform domain design is also very intuitive, as illustrated in Figure 5.1: conceptually, first the degraded image is "checked" against the image model in transform domain and corrections are made in places where the image model does not match. Subsequently, the obtained image is "tested" against the degradation model, pixel intensities that are "inconsistent" with the degradation model (e.g. due to oversmoothing) are restored by adding back information from the degraded image. Then this process is repeated, until no modifications are being performed to the restored image. Mathematically, this is achieved by iteratively solving an appropriately defined optimization problem. Later, we will see that a proper formulation of this optimization problem can lead to guaranteed convergence.

In the next sections, we will discuss a number of restoration techniques for all three of the above designs. Starting from "more simple" restoration problems we will gradually increase the complexity of the problems and illustrate through examples how these problems can be solved efficiently. A list of novel restoration algorithms is given in Table 5.1. For each of the methods, the table lists the transform domain (Chapter 2), image model (Chapter 3) and degradation model (Chapter 4) being targeted.

Section	Application	Transform dom.	Image model	Degradation model
5.1.2	Image Denoising	Image Domain	Nonlocal	Correlated Gauss. noise
5.2.2	Image Denoising	Multiresolution	BKF	White Gauss. noise
5.2.3	Image Denoising	Multiresolution	Inter+intra	Correlated Gauss. noise
5.2.4	Image Denoising	Multiresolution	MPGSM	Correlated Gauss. noise
5.2.5	Demosaicing	DT-CWT	-	Bayer pattern
5.3.3	Image Denoising	Joint	Laplace/BKF	Correlated Gauss. noise
5.3.4	Denoising+ deblurring	Joint	Laplace/BKF	Additive noise + blur
5.3.5	Signal-dependent noise removal	Joint	Laplace/BKF	Uncorrelated signal-dependent noise

Table 5.1: Overview of the new restoration algorithms in this chapter.

5.1 Image domain restoration

5.1.1 Overview of existing techniques

For the restoration of images in the image domain, many techniques have been proposed in the past. The Wiener-filter is theoretically optimal in the MMSE sense for restoring images with Gaussian statistics corrupted with *Gaussian noise* [Rooms et al., 2010]. Let $\mathbf{x}$ and $\mathbf{y}$ denote vectors of pixel intensities of respectively the original image and the observed image. The Wiener filter estimates the original image as follows:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} H(\mathbf{x};\mathbf{y}) + J(\mathbf{x}) \tag{5.1}$$

with $H(\mathbf{x};\mathbf{y})$ a so-called data fitting (or fidelity) function, which is a quadratic function in case of the Wiener filter and with $J(\mathbf{x}) \propto \|\mathbf{x}\|^2$ a regularization term. Unfortunately, as we explained in Chapter 3, image statistics are rarely Gaussian and consequently the Wiener-filter generally performs poorly.

The Tikhonov-Miller [Tikhonov et al., 1990] restoration method is an improvement to the Wiener filter in the sense that it imposes a smoothness constraint as a regularization functional to suppress oscillations and restoration errors due to noise amplification. Tikhonov-Miller uses quadratic data fitting and regularization terms:

$$H(\mathbf{x};\mathbf{y}) = \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \quad \text{and} \quad J(\mathbf{x}) = \|\mathbf{S}\mathbf{x}\|^2, \tag{5.2}$$

with $\mathbf{S}$ a linear "sparsifying" transform (generally a high-pass filter, although generally any multiresolution transform from Chapter 2 can be used as well). Solving the l_2-regularized problem (5.2) leads to a system of linear equations, from which the solution can efficiently be computed using the DFT, or Gauss-Seidel methods [Kahan, 1958]. Alternatively, the solution can be computed iteratively using the method of successive approximations [Banham and Katsaggelos, 1997]; the resulting algorithm is then known as iterative constrained least squares (ICLS).

Both the Wiener filter and the Tikhonov-Miller restoration method are linear methods that have the *dis*advantage that they can not recover frequency

components of the image that are fully removed by the degradation process (for example high frequencies in the image that are completely attenuated due to blur).

To overcome these problems, l_1-regularized problems have been studied, where

$$H(\mathbf{x};\mathbf{y}) = \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \quad \text{and} \quad J(\mathbf{x}) = |\mathbf{S}\mathbf{x}|_1 . \tag{5.3}$$

The l_1-regularized problem is difficult to solve numerically, because $J(\mathbf{x})$ is non-differentiable. A well-known example is Total Variation (TV) [Rudin and Osher, 1994], where $H(\mathbf{x};\mathbf{y}) = \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ and $J(\mathbf{x}) = \|\mathbf{x}\|_{BV}$ is the TV norm (also known as the bounded variation norm):

$$\|\mathbf{x}\|_{BV} = |\boldsymbol{\nabla}\mathbf{x}|_1 \tag{5.4}$$

which is the magnitude of a discrete version of the spatial gradient of $\mathbf{x}$. Common techniques to solve l_1-regularized problems are steepest descent (which is generally considered to be very slow by taking too many iterations before convergence), various improvements on steepest descent, employing better optimized time steps, nonlinear conjugate gradient methods and (approximative) Newton-based methods.

Other filters that are successful are Bilateral filtering [Tomasi and Manduchi, 1998], nonlinear diffusion schemes [Perona and Malik, 1990, Weickert, 1998, Black et al., 1998]. Nonlinear diffusion schemes compute solutions of a set of coupled partial differential equations (PDEs) that are inspired by heat-diffusion equations. The conductance coefficient that is used by [Perona and Malik, 1990] is a function of the gradient magnitude (edge-stopping function) and ensures that edges are preserved during the simulated diffusion process. Here, the TV norm is sometimes also used as edge-stopping function, which is called TV diffusion. In [Pižurica et al., 2006], a Bayesian formulation for the edge-stopping functions is used, based on a mixture of truncated Laplace distributions (see Section 3.2.2). In [Steidl et al., 2004], is shown that TV diffusion, TV regularization and Haar-wavelet soft shrinkage (see further) are equivalent under certain (but quite general) conditions.

For the restoration of blurred images with *Poisson noise*, the Richardson-Lucy (RL) method has been proposed in [Richardson, 1972, Lucy, 1974]. An iterative EM algorithm for RL has been presented in [Shepp and Vardi, 1982]. Because RL does not include regularization, extensions have been proposed, such as RL with Tikhonov-Miller regularization [Dey et al., 2004], RL Conchello [Conchello and McNally, 1996] and RL with TV regularization [Dey et al., 2004].

More recently, the NLMeans filter has been introduced in [Buades. et al., 2005]. By its success, many improvements have been proposed (e.g. [Mahmoudi and Sapiro, 2005, Kervrann and Boulanger, 2006, Kervrann et al., 2007, Bilcu and Vehvilainen, 2007, Dauwe et al., 2008]). Most of the authors focus on the removal of additive white Gaussian noise. In the next section, we will explain our own improvements to the NLMeans filter and how the filter can be used for the removal of correlated Gaussian noise.

5.1.2 Improved Non-Local Means filter

As already mentioned in Section 3.7, the NLMeans filter estimates a noise-free pixel intensity as a weighted average of *all* pixels in the image, where the weights are proportional to the similarity between the local neighborhood of the pixel being processed and local neighborhoods of the surrounding pixels:

$$\widehat{[\mathbf{x}_j]}_c = \frac{\sum_{j'=1}^{N} g\left(\mathbf{y}_{j'} - \mathbf{y}_j\right) \mathbf{y}_{j'}}{\sum_{j'=1}^{N} g\left(\mathbf{y}_{j'} - \mathbf{y}_j\right)} \tag{5.5}$$

where $g(\cdot)$ is a weighting function that depends on the Euclidean distance between the two local neighborhood vectors:

$$g\left(\mathbf{y}_{j'} - \mathbf{y}_j\right) = \exp\left(-\frac{\left\|\mathbf{y}_{j'} - \mathbf{y}_j\right\|^2}{h^2}\right) \tag{5.6}$$

with h a constant that is proportional to the noise variance. Since only the central pixel intensity of the local neighborhood is estimated, we will refer to this filter as the pixel-based NLMeans. A *block*-based NLMeans filter does also exist [Buades et al., 2005, Goossens et al., 2008a], in this filter the weights of different neighborhoods are accumulated before calculating the final weighted average. The NLMeans filter in this form is intuitive and potentially very powerful, however, there are two limitations: 1) the objective and visual quality are somewhat inferior to other recent non-local techniques (e.g. [Elad and Aharon, 2006b, Dabov et al., 2007]) and 2) the NLMeans filter has a computational complexity that is quadratic in the number of pixels in the image, which makes the algorithm impractical in real applications. Therefore, several authors have investigated better similarity measures [Azzabou et al., 2007, Kervrann et al., 2007], use adaptive local neighborhoods [Kervrann and Boulanger, 2006], or refine the similarity estimates in different iterations [Brox and Cremers, 2007]. In [Zimmer et al., 2008], rotationally invariant distance measures are used. Other authors propose algorithmic acceleration techniques [Mahmoudi and Sapiro, 2005, Wang et al., 2006, Bilcu and Vehvilainen, 2007, Dauwe et al., 2008], based for example on neighborhood preclassification [Mahmoudi and Sapiro, 2005, Dauwe et al., 2008] and FFT-based computation of the neighborhood similarities [Wang et al., 2006].

In earlier work [Dauwe et al., 2008] we noted that the exponential form of $g(\mathbf{x})$ assigns positive non-zero weights to *dissimilar* neighborhoods. Even though these weights are very small, the estimated pixel intensities can be severely biased due to many small positive contributions. We therefore proposed a preclassification based on the first three statistical moment to exclude dissimilar blocks [Dauwe et al., 2008]. In [Goossens et al., 2008a] we have shown that the NLMeans algorithm is equivalent to the first iteration of optimizing the robust Leclerc loss function (for the Jacobi optimization method). This makes it possible to replace the Leclerc loss function with other loss functions (see Table 3.2) that are commonly used for robust estimation. We will now look at the characteristics of these robust weighting functions in more detail.

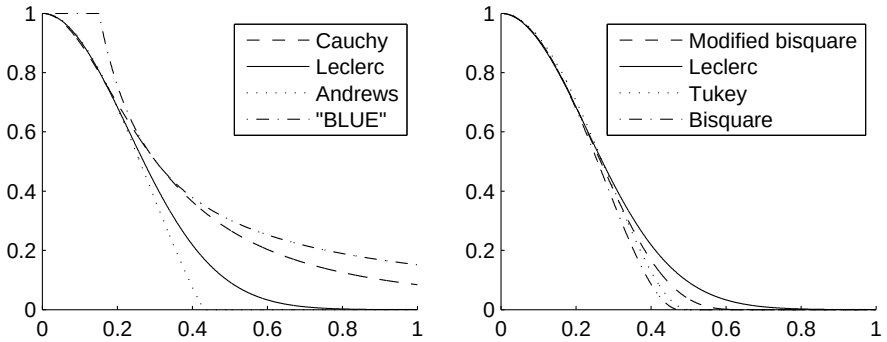


Figure 5.2: Comparison of the weighting functions $g(y)$ for commonly used robust functions. Here, $y = \|\mathbf{y}_i - \mathbf{y}_j\|$ is the Euclidean distances between the vectors $\mathbf{y}_i$ and $\mathbf{y}_j$.

- The *Cauchy* weighting function has a very slow decay (see Figure 5.2(a)). Larger weights are assigned to dissimilar blocks than the Leclerc robust function, which will eventually lead to oversmoothing.
- The *Leclerc* weighting function has a faster decay, but still assigns positive non-zero weights to dissimilar blocks.
- The *Andrews* weighting function imposes a hard threshold while comparing neighborhoods (the weight is zero as soon as a given threshold is exceeded), while the *Tukey* and *Bisquare* weighting functions rather use a soft threshold (Figure 5.2(b)). Experimentally we found that applying a soft threshold often improves the visual quality.
- To improve upon the Tukey and Bisquare weighting functions, we also modified the Bisquare robust function in order to have a steeper slope (see Table 3.2 and Figure 5.2(b)).

Next, one iteration of the NLMeans filter may not remove all of the noise (e.g. in case no other similar candidate neighborhoods are present in the image). Theoretically, because the NLMeans estimate is an average, the filter requires an infinite number of neighborhoods in order to completely suppress the noise, which is not possible for finite image dimensions. Therefore, we successfully applied a post-filter in [Goossens et al., 2008a] to remove the remaining noise. The effect of this post-filter is shown in Figure 5.3. It can be seen that the post-filter gives a significant improvement both in PSNR as in visual quality.

Further, we extended the NLMeans filter to images with correlated noise, by considering the Mahalanobis distance instead of the Euclidean distance. For

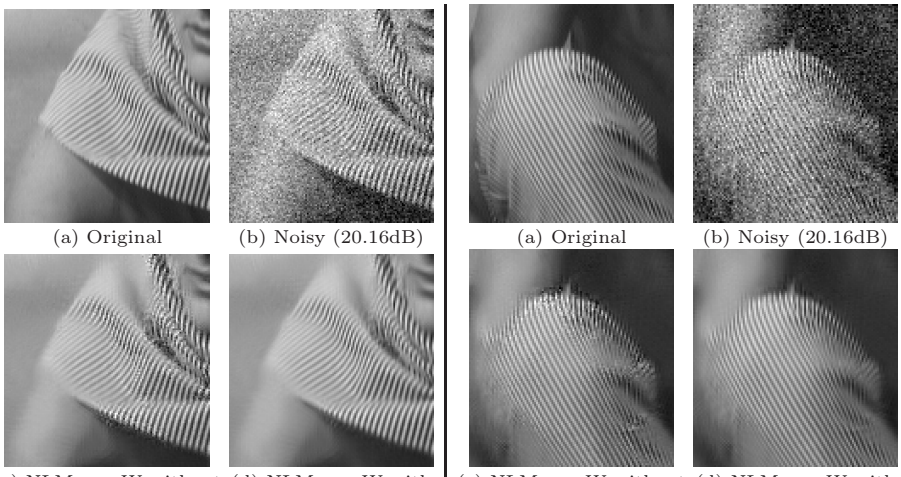


Figure 5.3: Denoising example of Barbara (two magnified portions of the image are shown): the effect of using the proposed post-processing filter (a) Crop out of the original image (b) Image corrupted by white Gaussian noise ($\sigma = 25$). (c) The result of the NLMeans filter *without* post-processing filter (d) The result of the NLMeans filter *with* post-processing filter. PSNR values (for the full image) are between parentheses.

the Leclerc weighting function[1], this gives:

$$g\left(\mathbf{y}_{j'} - \mathbf{y}_j\right) = \exp\left(-\frac{\left(\mathbf{y}_{j'} - \mathbf{y}_j\right)^T \mathbf{C}_w^{-1} \left(\mathbf{y}_{j'} - \mathbf{y}_j\right)}{h^2}\right), \tag{5.7}$$

where $\mathbf{C}_w$ is the noise covariance matrix in the image domain. The distance function (5.7) can be efficiently computed by first applying a prewhitening operation to the observed neighborhood vectors (i.e. $\mathbf{y}_j^{'} = \mathbf{C}_w^{-1/2}\mathbf{y}_j$) and subsequently by using the Euclidean distance function on the prewhitened neighborhood vectors. In practice, we "prewhiten" the whole image based on the noise PSD, which is computationally more efficient than prewhitening all neighborhood vectors individually. The PSD may not be invertible, in this case we regularize the inversion using a small positive constant.

Equivalently, the NLMeans filter can be used to suppress non-stationary Gaussian noise, by using the following weighing function:

$$g\left(\mathbf{y}_{j'} - \mathbf{y}_j\right) = \exp\left(-\frac{\left\|\mathbf{y}_{j'} - \mathbf{y}_j\right\|^2}{h^2\left(\sigma_{j'}^2 + \sigma_j^2\right)}\right) \tag{5.8}$$

where σ_j^2 is the noise variance at position j in the image. Here the noise variance of positions j and j' is summed in the weight computation. This

[1] Other weighting functions can be extended analogously.

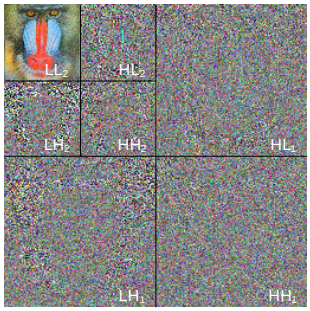


Figure 5.4: 2D DWT of an image with white Gaussian noise.

distance function can be motivated by the Gaussianity of the noise: if $\mathbf{y}_{j'}$ and $\mathbf{y}_j$ are Gaussian distributed with variance $\sigma^2_{j'}$ and σ^2_j, respectively, then their difference $\mathbf{y}_{j'} - \mathbf{y}_j$ will be Gaussian with variance $\sigma^2_{j'} + \sigma^2_j$. A similar extension of the NLMeans filter has been made for dealing with Poisson noise [Coupé et al., 2008].

Finally, we proposed a modified implementation of the NLMeans filter, yielding a significant speed up of the algorithm without sacrificing quality. Our modification exploits the symmetry of the weighting function (i.e. $g(\mathbf{y}_{j'} - \mathbf{y}_j) = g(\mathbf{y}_j - \mathbf{y}_{j'})$ and uses a 2D moving average filter for computing the Euclidean distances between neighborhood vectors. For 11×11-neighborhoods, the speed-up factor is approximately 121, which brings down the computation time for processing a 512×512 image from several minutes to approximately 10 seconds on a recent PC.

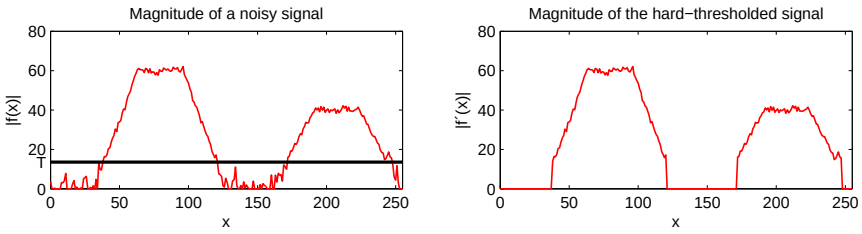
5.2 Multiresolution image restoration

In this section, we will discuss a number of image restoration techniques that are implemented in a multiresolution transform domain, following the design of Figure 5.1(b).

5.2.1 Existing multiresolution techniques for noise reduction

Wavelet shrinkage by thresholding

Image denoising by wavelet shrinkage has attracted the interest of many researchers in the field, due to its simplicity and effectiveness. Figure 5.4 shows a 2D DWT decomposition of an image with artificial noise: the LL_2 subband clearly contains the most signal information, while the subbands LH_i, HL_i and HH_i are almost completely corrupted by the noise. Some image structures (mostly from the textures and edges) are still visible through the noise. These structures correspond to wavelet coefficients with significant magnitudes



(a) (b)

Figure 5.5: Hard-thresholding of a signal. (a) Noisy signal, (b) Thresholded signal.

in the noiseless image. Noise reduction is then typically performed by *shrinkage*-based approaches: wavelet coefficients with non-significant magnitudes are mostly dominated by noise; by decreasing the coefficients in magnitude, the noise can noticably be suppressed. Unfortunately, large coefficients also contain noise, which is not removed by this method.

Hard-thresholding imposes a fixed threshold per subband in which all coefficients with magnitude smaller than this threshold are set to zero. *Soft-thresholding* also reduces the magnitude of the significant coefficients but to a lesser degree. In this respect, wavelet-domain thresholding is very similar to *noise gating*, a much older technique that is commonly used to suppress background noise in audio signals: the noise gate allows the signal to pass only when its magnitude is above a given threshold. Below this threshold, the "gate" is closed and nothing is passed. This principle is illustrated in Figure 5.5: in this example the noise can be almost completely removed by simply discarding all signal magnitudes that are below the threshold T.

To address the question of threshold selection, a number of techniques have been specifically developed to optimize the threshold for hard and/or soft shrinkage in terms of a given optimization criterion [Donoho, 1995, Vidakovic, 1998a, Vidakovic, 1998b, Chang et al., 2000b, Abramovich et al., 1998].

Bayesian techniques

Bayesian estimation techniques incorporate a prior distribution of noise-free wavelet coefficients [Vidakovic, 1998a, Vidakovic, 1998b, Leporini et al., 1999] and also often lead to shrinkage estimators. These methods minimize a specific optimization criterion, called the *Bayesian risk* [Van Trees, 1968]. Let x and y respectively denote a noise-free wavelet coefficient and the corresponding observed noisy wavelet coefficient at position j, then the Maximum a posteriori (MAP) estimate calculates the mode of the posterior distribution $f_{x|y}(x_j|y_j)$:

$$\hat{x}_{\mathrm{MAP},j} = \arg\max_{x_j} f_{x|y}(x_j|y_j)\,, \tag{5.9}$$

where the posterior distribution is obtained by applying Bayes' rule (hence the name Bayesian estimation). For example, if x follows a *Laplace* distribution

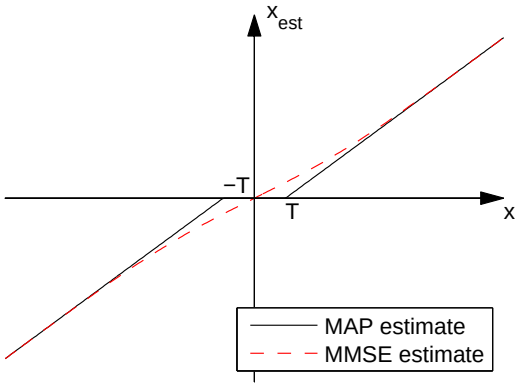


Figure 5.6: Bayesian MAP and MMSE shrinkage for estimating a Laplacian distributed signal.

with parameter s (as defined in Section 3.2.1) and if $y|x \sim \mathcal{N}\left(x, \sigma^2\right)$ is Gaussian distributed, then the MAP estimate for x is given by the *soft-shrinkage* rule:

$$\begin{aligned} \hat{x}_{\mathrm{MAP},j} &= \mathrm{softshrink}\left(y_j; \sigma^2/s\right) \\ &= \mathrm{sign}\left(y_j\right) \max\left(0, |y_j| - \sigma^2/s\right). \end{aligned}$$

A second Bayesian estimate is the Minimum Mean Square (MMSE) error estimate, which minimizes the quadratical cost function $\mathrm{E}\left[(\hat{x} - x)^2\right]$:

$$\hat{x}_{\mathrm{MMSE},j} = \mathrm{E}\left[x_j | y_j\right], \tag{5.10}$$

which is the conditional expectation of the noise-free wavelet coefficient, given its noisy observation. The MAP and MMSE[2] estimates for a Laplacian distributed signal in Gaussian noise are illustrated in Figure 5.6. For large $|x|$ both estimates are approximately equal, while for small $|x|$ the MMSE estimate tends to shrink "less aggressively".

For more complicated conditional PDFs $f_{x|y}\left(x_j|y_j\right)$, the MAP estimate (5.9) may not be available in closed-form. In those cases, general optimization techniques can be used, often leading to iterative techniques that are easy to implement. On the other hand, the computation of the MMSE estimate requires an integration, for which the use of numerical integration techniques may be necessary, especially when considering vectors of wavelet coefficients. Nevertheless, the MMSE estimate is often preferred over the MAP estimate because it naturally maximizes the Peak Signal To Noise Ratio (PSNR), defined by:

$$\mathrm{PSNR} = 10 \log_{10} \frac{255^2}{\mathrm{MSE}} \quad \text{with} \quad \mathrm{MSE} = \frac{1}{N} \sum_{j=1}^{N} \left(\hat{x}_j - x_j\right)^2.$$

[2] A "simple" closed-form expression for the MMSE estimate is in this case not available (see e.g. [Simoncelli and Adelson, 1996]), although we will go deeper into this in Section 5.2.2.

Table 5.2: Overview of a few recent wavelet-based denoising techniques from literature. (For specific details, see text.)

"Classical" Thresholding	*Bayesian multivariate*
[Donoho, 1995]	[Strela et al., 2000]
[Vidakovic, 1998a]	[Figueiredo and Nowak, 2001]
[Vidakovic, 1998b]	[Şendur and Selesnick, 2002a]
[Chang et al., 2000b]	[Portilla et al., 2003]
[Abramovich et al., 1998]	[Portilla, 2005]
[Johnstone and Silverman, 1997]	[Malfait and Roose, 1997]
Bayesian univariate	*Bayesian multivariate (cont.)*
[Simoncelli and Adelson, 1996]	[Pižurica et al., 2002]
[Chipman et al., 1997]	[Scheunders, 2004]
[Abramovich et al., 1998]	[Pižurica and Philips, 2006]
[Clyde et al., 1998]	[Fan and Xia, 2001a]
[Crouse et al., 1998]	[Benazza-Benyahia and Pesquet, 2004]
[Mihçak, 1999]	[Benazza-Benyahia and Pesquet, 2005]
[Moulin and Liu, 1999]	[Goossens et al., 2009d]
[Mallat, 1999]	[Goossens et al., 2009c]
[Achim et al., 2001b]	
[Fadili and Boubchir, 2005]	

MSE and PSNR are extensively used in the literature to compare the performance of restoration and compression algorithms. It can be shown that the MSE, when computed as the sum of the squared differences between wavelet and scaling coefficients $\hat{x}_j$ and x_j *in all subbands*, is invariant under orthonormal transforms. Hence, for orthonormal multiresolution representations such as the DWT, maximizing the PSNR in the wavelet domain as explained above, automatically maximizes the PSNR after wavelet reconstruction. For redundant transforms this is not the case: maximizing the PSNR in transform domain is strictly speaking *sub-optimal* in terms of image domain PSNR [Raphan and Simoncelli, 2008].

Because the prior distribution encodes prior knowledge about noise-free images, the choice of the prior distribution is *crucial* in the design of Bayesian denoising techniques. In fact, Bayesian estimators can be derived for any of the parametric densities presented in Section 3.2. This fact, together with the good time-frequency localization properties of the wavelet transform, recently led to the birth of many denoising techniques (see the Bayesian section of Table 5.2): [Simoncelli and Adelson, 1996] studied the MMSE estimates for a generalized Laplace prior. [Moulin and Liu, 1999] investigate the Bayesian MAP estimate for the same prior distribution. It was found that the MAP estimate for very heavy tailed generalized Laplace distributions is equivalent to *hard thresholding*. [Abramovich et al., 1998, Clyde et al., 1998] derive Bayesian estimates for weighted mixtures of a Gaussian and a point mass at zero. [Chipman et al., 1997, Crouse et al., 1998] use the MMSE estimate for a mixture of two univariate Gaussian distributions, in particular [Crouse et al., 1998] use this model in the context of a HMT tree (see Section 3.5.1).

In [Tan and Jiao, 2007], a great similary between the methods of [Mihçak,

1999, Figueiredo and Nowak, 2001, Şendur and Selesnick, 2002a, Portilla et al., 2003] was noted. These algorithms all employ elliptically symmetric prior distributions (see Section 3.2.3). It was found that the GSM model from [Portilla et al., 2003], which aims to characterize the joint statistics of wavelet coefficients in a local neighborhood through a GSM model, and the bivariate shrinkage method of [Şendur and Selesnick, 2002a] offers a denoising performance (in terms of PSNR) that is significantly better than the methods from [Figueiredo and Nowak, 2001, Mihçak, 1999]. Many authors (e.g. [Chang et al., 2000a, Pižurica et al., 2002, Tan and Jiao, 2007]) conclude that great importance need to be attached to the joint dependence among wavelet coefficients. While [Şendur and Selesnick, 2002a] use a bivariate probability density for modeling inter-scale dependencies, [Portilla et al., 2003] use higher-dimensional (typically 9 or 10 dimensional) multivariate densities for characterizing the wavelet coefficients in a local neighborhood and (optionally) a parent coefficient.

The MMSE estimate for the GSM used in [Portilla et al., 2003] is particularly interesting because the expression is a posterior weighted average of local Wiener estimates for different values of the GSM-multiplier z:

$$\hat{\mathbf{x}}_j = \mathrm{E}\left[\mathbf{x}_j|\mathbf{y}_j\right] = \int_0^{+\infty} f_{z|\mathbf{y}}\left(z|\mathbf{y}_j\right) z\mathbf{C}_u \left(z\mathbf{C}_u + \mathbf{C}_w\right)^{-1} \mathbf{y}_j \mathrm{d}z, \tag{5.11}$$

where $\mathbf{C}_w$ is the noise covariance matrix for the considered wavelet subband. In practice, the integral in (5.11) must be evaluated using numerical techniques. We will show in Section 5.2.2 that for a specific case of GSM model, the Bessel-K Form density, closed form expressions can be found as well.

[Pižurica and Philips, 2006] propose to shrink a wavelet coefficient by multiplying it *with the probability that it contains a significant noise-free component* (called hypothesis H_1):

$$\begin{aligned} \hat{x}_j &= \mathrm{P}\left(H_1|y_j\right) y_j \\ &= \left(1 + \frac{\mathrm{P}\left(H_0\right)}{\mathrm{P}\left(H_1\right)} \frac{f_{y|H}\left(y_j|H_0\right)}{f_{y|H}\left(y_j|H_1\right)}\right)^{-1} y_j. \end{aligned} \tag{5.12}$$

A locally adaptive version of this approach was also introduced in [Pižurica and Philips, 2006]. This technique attempts to exploit spatial correlations between the wavelet coefficients within the same subband (also see Figure 5.7):

$$\begin{aligned} \hat{x}_j &= \mathrm{P}\left(H_1|y_j, v_j\right) y_j \\ &= \left(1 + \frac{\mathrm{P}\left(H_0\right)}{\mathrm{P}\left(H_1\right)} \frac{f_{v|H}\left(v_j|H_0\right)}{f_{v|H}\left(v_j|H_1\right)} \frac{f_{y|H}\left(y_j|H_0\right)}{f_{y|H}\left(y_j|H_1\right)}\right)^{-1} y_j \end{aligned} \tag{5.13}$$

where v_j is the local spatial activity indicator (see Section 3.4.2). In this case, the probability of signal presence is conditioned not only on the coefficient value but also on the LSAI computed from the surrounding coefficients, to improve the denoising performance. This locally adaptive version offers a denoising

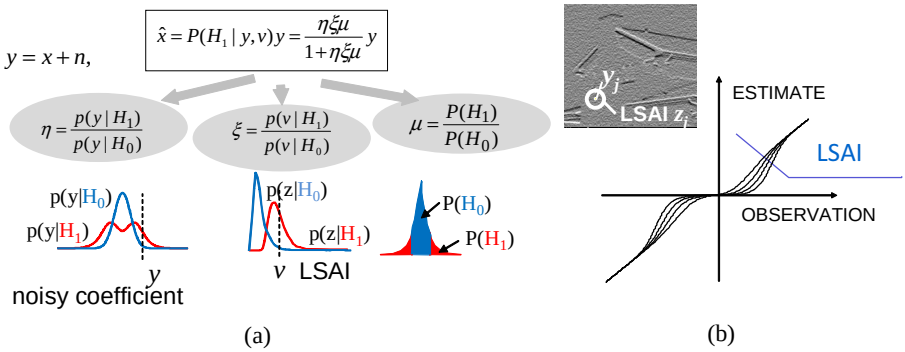


Figure 5.7: Illustration of the probabilistic shrinkage (ProbShrink) from [Pižurica and Philips, 2006]. (a) Shrinkage rule formula and the different conditional densities, (b) Shrinkage rule for different values of the Local Spatial Activity Indicator (LSAI).

performance close to the performance of [Portilla et al., 2003], but at a *lower* computational complexity.

Similarly, a number of denoising methods have been devised for Markov Random Field[3] priors [Malfait and Roose, 1997, Jansen and Bultheel, 1999, Jansen and Bultheel, 2001, Pižurica et al., 2002]. These techniques estimate a significance map of the wavelet coefficients, typically using Metropolis stochastic sampling [Malfait and Roose, 1997, Jansen and Bultheel, 1999, Jansen and Bultheel, 2001] or using *iterated conditional modes (ICM)* [Pižurica et al., 2002].

More recently, further improvements have been brought to the "traditional" GSM method of [Portilla et al., 2003]: the orientation-adaptive GSM method from [Hammond and Simoncelli, 2008] makes use of the steerability properties of the steerable pyramid transform to improve the denoising performance. In [Guerrero-Colón et al., 2008a] the GSM covariance matrix is estimated *locally* in non-overlapping regions. Mixtures of GSMs [Guerrero-Colón et al., 2008b] cluster the local covariance matrices globally, in order to exploit non-local redundancy (or repetitivity) in images. Fields of Gaussian Scale Mixture models (FoGSM) [Lyu and Simoncelli, 2008] combine the GSM model with a Markov Random Field model and yield a slightly better denoising performance than MGSM, on average.

Finally, denoising techniques increasingly employ "newer" multiresolution transforms specifically designed to overcome some of the shortcomings of the wavelet transform (see Chapter 2). These recent transforms analyze images in a multidirectional fashion, which allows then to deal with line and curve singularities [Tan and Jiao, 2007]. For example, in [Starck et al., 2000] it was found that simple hard-thresholding rules in the curvelet domain yielded results that were comparable to state-of-the-art methods, employing more complicated

[3] See Section 3.4.1.

rules in the wavelet domain. A similar result was obtained for *Ridgelet-biframes* in [Tan and Jiao, 2006]. In [Tessens et al., 2008], an extension of ProbShrink from [Pižurica and Philips, 2006] is proposed for the *curvelet* transform, reaching the same conclusion. In [Easley et al., 2009], the *shearlet* system is adopted to this end.

5.2.2 Estimators for the univariate Bessel K distribution

As already mentioned in Section 3.2.5, the marginals of the BKF distribution fit well with the observed histograms of wavelet coefficients. Because the BKF distribution is a specific case of a Gaussian Scale Mixture, estimators for scalar wavelet coefficients under the BKF prior can be more easily extended to wavelet coefficient vectors than for e.g. non-GSM distributions.

In this section, we will derive exact expressions for the MMSE and MAP estimates of a Bessel K Form (BKF) distributed signal, corrupted with additive white Gaussian noise. In [Fadili and Boubchir, 2005], an expression for the MMSE estimator for the Bessel K distribution was derived, however, this derivation is based on an asymptotic approximation of the Bessel K function. For completeness, we give the exact results here. As far as we know, these expressions have not been reported so far. We follow the same reasoning as in [Selesnick, 2008].[4]

MMSE estimate

To simplify the notations, we will consider only one coefficient in a subband of a given multiresolution transform and therefore we will omit the position subscript j in the remainder of this section. Using the GSM representation of the Bessel K Form, the MMSE estimate (5.10) yields:

$$\hat{x}_{\mathrm{MMSE}} = \mathrm{E}\left[x|y\right] = \int_0^{+\infty} f_{z|y}\left(z|y\right) \mathrm{E}\left[x|y,z\right] \mathrm{d}z$$

where we can use Bayes' rule $f_{z|y}\left(z|y\right) = f_{y|z}\left(y|z\right) f_z\left(z\right) / f_y\left(y\right)$. To proceed, we need to compute the likelihood function $f_y\left(y\right)$:

$$\begin{aligned} f_y\left(y\right) &= \int_0^{+\infty} f_{y|z}\left(y|z\right) f_z\left(z\right) \mathrm{d}z \\ &= \frac{1}{\Gamma\left(\tau\right)} \int_0^{+\infty} \frac{z^{\tau-1} e^{-z}}{\sqrt{2\pi\left(z\sigma_u^2 + \sigma^2\right)}} \exp\left(-\frac{1}{2}\frac{y^2}{z\sigma_u^2 + \sigma^2}\right) \mathrm{d}z \\ &= \frac{\exp\left(\sigma^2/\sigma_u^2\right)}{\sqrt{2\pi\sigma_u}\Gamma\left(\tau\right)} \int_{\sigma^2/\sigma_u^2}^{+\infty} \left(t - \frac{\sigma^2}{\sigma_u^2}\right)^{\tau-1} t^{-\frac{1}{2}} \exp\left(-t - \frac{y^2}{2t\sigma^2}\right) \mathrm{d}t, \end{aligned}$$

[4] in [Selesnick, 2008], exact MMSE and MAP estimates are given for a special case of the BKF, i.e. for $\tau = 1$, which corresponds to a Laplace distribution and which yields simpler expressions.

with σ^2 and σ_u^2 respectively the noise variance and variance of the Gaussian (hidden) random variable u. Next, we apply the generalized Binomial theorem (allowing non-integer values for τ):

$$(t+a)^{\tau-1} = \sum_{n=0}^{+\infty} \begin{pmatrix} \tau-1 \\ n \end{pmatrix} t^{\tau-1-n} a^n,$$

with $a = -\sigma^2/\sigma_u^2$ and with $\begin{pmatrix} \tau-1 \\ n \end{pmatrix} = (\tau-1)(\tau-2)\cdots(\tau-n)/n!$. This leads to:

$$\begin{aligned} f_y(y) &= \frac{\exp\left(\sigma^2/\sigma_u^2\right)}{\sqrt{2\pi\sigma_u}\Gamma(\tau)} \sum_{n=0}^{+\infty} \begin{pmatrix} \tau-1 \\ n \end{pmatrix} a^n \int_{\sigma^2/\sigma_u^2}^{+\infty} t^{\tau-n-\frac{3}{2}} \exp\left(-t - \frac{y^2}{2t\sigma^2}\right) \mathrm{d}t \\ &= \frac{\exp\left(\sigma^2/\sigma_u^2\right)}{\sqrt{2\pi\sigma_u}\Gamma(\tau)} \sum_{n=0}^{+\infty} \begin{pmatrix} \tau-1 \\ n \end{pmatrix} \left(-\frac{\sigma^2}{\sigma_u^2}\right)^n \Gamma\left(\tau-n-\frac{1}{2}, \frac{\sigma^2}{\sigma_u^2}; \frac{y^2}{2\sigma^2}\right), \end{aligned} \tag{5.14}$$

where $\Gamma(\alpha, x; \beta) = \int_x^{+\infty} t^{\alpha-1} \exp\left(-t - \frac{\beta}{t}\right) dt$ is the generalized incomplete Gamma function [Selesnick, 2008]. Based on (5.14), the MMSE estimate directly follows:

$$\begin{aligned} &\hat{x}_{\mathrm{MMSE}} = y - \alpha(y) y \quad \text{where} \\ &\alpha(y) = \frac{\sigma^2}{\sigma_u^2} \frac{\sum_{n=0}^{+\infty} \begin{pmatrix} \tau-1 \\ n \end{pmatrix} \left(-\frac{\sigma^2}{\sigma_u^2}\right)^n \Gamma\left(\tau-n-\frac{3}{2}, \frac{\sigma^2}{\sigma_u^2}; \frac{y^2}{2\sigma^2}\right)}{\sum_{n=0}^{+\infty} \begin{pmatrix} \tau-1 \\ n \end{pmatrix} \left(-\frac{\sigma^2}{\sigma_u^2}\right)^n \Gamma\left(\tau-n-\frac{1}{2}, \frac{\sigma^2}{\sigma_u^2}; \frac{y^2}{2\sigma^2}\right)}. \end{aligned} \tag{5.15}$$

Because the factor $\alpha(y)$ in the second term in y is between 0 and 1, the MMSE estimate also corresponds to shrinkage. Interesting is the special case $\tau = 1$ (Laplace distribution), for which (5.15) amounts to:

$$\hat{x}_{\mathrm{MMSE}} = y - \frac{\sigma^2}{\sigma_u^2} \frac{\Gamma\left(-\frac{1}{2}, \frac{\sigma^2}{\sigma_u^2}; \frac{y^2}{2\sigma^2}\right)}{\Gamma\left(\frac{1}{2}, \frac{\sigma^2}{\sigma_u^2}; \frac{y^2}{2\sigma^2}\right)} y, \tag{5.16}$$

This expression is identical to the result for the Laplace distribution given in [Selesnick, 2008].

MAP estimate

An implicit form for the MAP estimate can be directly derived from the results from [Selesnick, 2008]:

$$|\hat{x}| = |y| \left(1 - \sqrt{2}\frac{\sigma^2}{\sigma_u} \frac{K_{\frac{3}{2}-\tau}\left(\sqrt{2}\,|\hat{x}|/\sigma_u\right)}{K_{\frac{1}{2}-\tau}\left(\sqrt{2}\,|\hat{x}|/\sigma_u\right)}\right)_+ \tag{5.17}$$

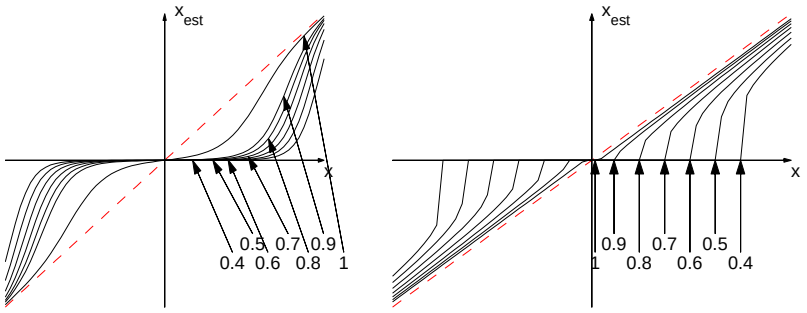


Figure 5.8: (a) MMSE estimates for the Bessel K Form density according to (5.15), (b) MAP estimates for the Bessel K Form density, using (5.17).

where $(\cdot)_+ = \max(0, \cdot)$. The rule (5.17) can be practically implemented by successive substitution [Selesnick, 2008], i.e. by starting from $\hat{x} = y$ and by subsequently applying the right handed side of (5.17) iteratively. Note that for $\tau = 1$ we correctly get the *soft-thresholding* rule, as $K_{1/2}(u) = K_{-1/2}(u)$.

In Figure 5.8, both the MAP and MMSE shrinkage rules are depicted, for different values of the shape parameter τ. MMSE shrinkage rules are much smoother, while the MAP estimate approximates hard-thresholding for small τ. Recall that the kurtosis of the BKF distribution is given by $3+3/\tau$, hence small τ correspond to highly kurtotic distributions. In this respect, we have reached the same conclusion as for the generalized Laplace distribution in [Moulin and Liu, 1999]. The advantage of using the BKF distribution over the generalized Laplace distribution is that the estimates (5.17)-(5.15) can be easily generalized to vectors when the distribution of the noise is spherically symmetric. More specifically, if the covariance matrix $\mathbf{C}_u = \sigma_u^2 \mathbf{I}$ and the noise covariance matrix $\mathbf{C}_w = \sigma^2 \mathbf{I}$, then the multivariate MAP estimate is given by:

$$\|\hat{\mathbf{x}}\| = \|\mathbf{y}\| \left(1 - \sqrt{2} \frac{\sigma^2}{\sigma_u} \frac{K_{\frac{3}{2}-\tau}\left(\sqrt{2}\,\|\hat{\mathbf{x}}\|/\sigma_u\right)}{K_{\frac{1}{2}-\tau}\left(\sqrt{2}\,\|\hat{\mathbf{x}}\|/\sigma_u\right)} \right)_+ . \tag{5.18}$$

and a similar expression can be given for the MMSE estimate. In (5.18), the shrinkage is applied to the coefficient magnitude, leaving the orientation of the vector $\mathbf{y}$ untouched. This is an interesting generalization of shrinkage to vectors. Unfortunately, in practice, the signal and noise pdfs are rarely spherically symmetric. The reason is that most multiresolution transforms do not yield uncorrelated coefficients, as already explained in Section 3.3. In the next section, we will develop a vector-based shrinkage approach that works for non-spherically symmetric signal and/or noise probability densities.

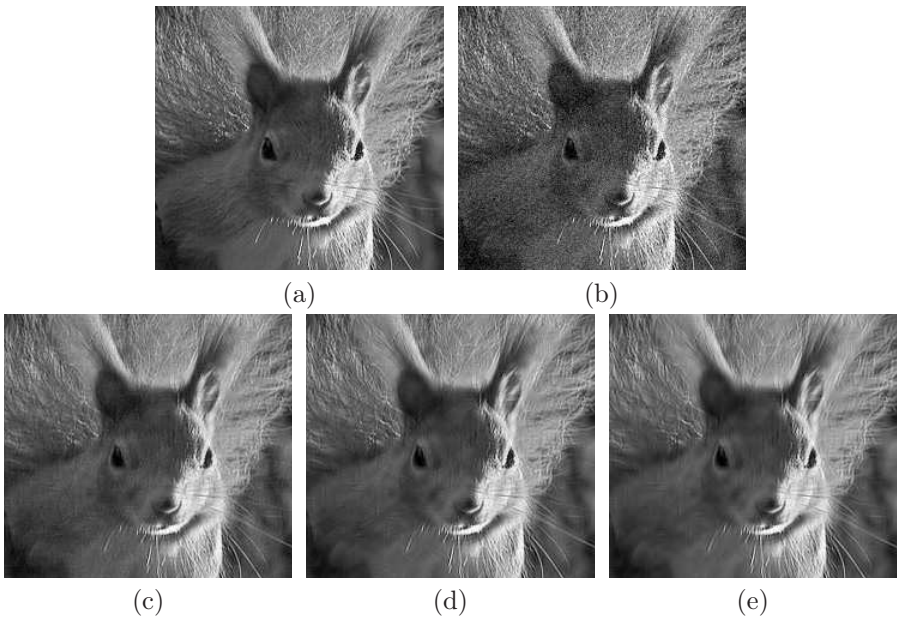


Figure 5.9: Denoising example in shearlet domain (a) Original image, (b) Noise image (PSNR=24.61dB), (c) Denoised using soft-thresholding (PSNR=30.74dB), (d) MAP estimate (5.17) (PSNR=31.18dB), (e) MMSE estimate (5.15) (PSNR=31.06dB). The parameter τ was estimated empirically for every shearlet subband using a dedicated EM algorithm (see Chapter 3).

5.2.3 Vector-ProbShrink

In this section, we will derive a vector-based shrinkage rule for the joint intra-/interscale probabilistic model from Section 3.6, called Vector-ProbShrink. Similar as in [Pižurica and Philips, 2006], we will estimate *the probability that an observed noisy neighborhood vector* $\mathbf{y}$ *contains a significant noise-free component (signal of interest present).* For the additive signal-plus-noise model,

$$\mathbf{y}_j = \mathbf{x}_j + \mathbf{w}_j, \tag{5.19}$$

the Vector-ProbShrink estimator is a generalization to vectors of ProbShrink (5.12):

$$\hat{\mathbf{x}}_j = \mathrm{P}\left(H_1|\mathbf{y}_j\right)\mathbf{y}_j. \tag{5.20}$$

where $\mathrm{P}\left(H_1|\mathbf{y}_j\right)$ is a so-called shrinkage factor. To ease the computation of this shrinkage factor, we will write (5.20) in the following form:

$$\hat{\mathbf{x}}_j = \left(1 - \frac{f_{\mathbf{y}|H}\left(\mathbf{y}_j|H_0\right)\mathrm{P}\left(H_0\right)}{f_{\mathbf{y}}\left(\mathbf{y}_j\right)}\right)\mathbf{y}_j. \tag{5.21}$$

The remainder of the derivation is simply to compute the different parts of this equation: $f_{\mathbf{y}|H}\left(\mathbf{y}_j|H_0\right)$, $f_{\mathbf{y}}\left(\mathbf{y}_j\right)$ and $\mathrm{P}\left(H_0\right)$. First we note that the (con-

ditional) pdfs of the observation can be expressed in terms of the (conditional) pdfs of the noise-free signal vectors, as follows:

$$f_{\mathbf{y}|H}\left(\mathbf{y}_j|H_0\right) = \int_{\mathbb{R}^d} f_{\mathbf{y}|\mathbf{x},H}\left(\mathbf{y}_j|\mathbf{x},H_0\right) f_{\mathbf{x}|H}\left(\mathbf{x}|H_0\right) \mathrm{d}\mathbf{x}, \tag{5.22}$$

$$f_{\mathbf{y}}\left(\mathbf{y}_j\right) = \int_{\mathbb{R}^d} f_{\mathbf{y}|\mathbf{x}}\left(\mathbf{y}_j|\mathbf{x}\right) f_{\mathbf{x}}\left(\mathbf{x}\right) \mathrm{d}\mathbf{x}. \tag{5.23}$$

Due to the additivity of the signal-noise model and the assumption of Gaussian noise, the conditional densities $f_{\mathbf{y}|\mathbf{x},H}\left(\mathbf{y}_j|\mathbf{x},H_0\right)$ and $f_{\mathbf{y}|\mathbf{x}}\left(\mathbf{y}_j|\mathbf{x}\right)$ are Gaussian pdfs:

$$f_{\mathbf{y}|\mathbf{x},H}\left(\mathbf{y}_j|\mathbf{x},H_0\right) = f_{\mathbf{y}|\mathbf{x}}\left(\mathbf{y}_j|\mathbf{x}\right) = N\left(\mathbf{y};\mathbf{x},\mathbf{C}_w\right). \tag{5.24}$$

Consequently, equations (5.22)-(5.23) are d-dimensional convolutions. The prior distribution $f_{\mathbf{x}}\left(\mathbf{x}\right)$ that we consider here, is the Bessel K Form density (Section 3.2.5), as in the previous section. The conditional prior distribution $f_{\mathbf{x}|H}\left(\mathbf{x}|H_0\right)$, under the hypothesis H_0 that the signal of interest is absent, is given by (3.43). Unfortunately, closed analytical forms for (5.22)-(5.23) are not trivial to find, which hampers the practical implementation. Instead, we marginalize the density $f_{\mathbf{X}}(\mathbf{x})$ based on the GSM representation (see Section 3.2.5) as $f_{\mathbf{x}}\left(\mathbf{x}\right) = \int_{z=0}^{+\infty} f_{\mathbf{X}|Z}(\mathbf{x}|z) f_Z(z) dz$. We further remark that if $f_{\mathbf{x}|H}\left(\mathbf{x}|H_0\right)$ is the density of a Gaussian Mixture, the above convolution involves adding the noise covariance matrix $\mathbf{C}_w$ to each component of the mixture. Therefore, we approximate the indicator function in (3.43) using a Gaussian function:

$$f_{\mathbf{x}|H}\left(\mathbf{x}|H_0\right) \approx C_0 f_{\mathbf{x}}\left(\mathbf{x}\right) \exp\left(-\frac{\mathbf{x}^T\mathbf{C}_w^{-1}\mathbf{x}}{2T^2}\right) \tag{5.25}$$

where C_0 is a density normalization factor. This results in a Gaussian conditional prior density on $\mathbf{x}$:

$$f_{\mathbf{y}|z,H}\left(\mathbf{y}|z,H_0\right) = N\left(\mathbf{x};\mathbf{0},\left((z\mathbf{C}_x)^{-1} + (T^2\mathbf{C}_w)^{-1}\right)^{-1}\right) \tag{5.26}$$

where $N\left(\mathbf{x};\mathbf{0},\mathbf{C}\right)$ denotes the Gaussian density evaluated in $\mathbf{x}$. Next, to simplify the dependency on z in (5.26), we use a trick used by [Portilla et al., 2003], to reduce the computation time of the BLS-GSM method: we express $f_{\mathbf{x}|z}\left(\mathbf{x}|z\right)$ and $f_{\mathbf{y}|z}\left(\mathbf{y}|z\right)$ in a new basis where $\mathbf{C}_x$ and $\mathbf{C}_w$ are both diagonal, using:

$$z\mathbf{C}_x + \mathbf{C}_w = \mathbf{U}\mathbf{Q}(z\mathbf{\Lambda} + \mathbf{I})\mathbf{Q}^T\mathbf{U}^T \tag{5.27}$$

where $\mathbf{U}\mathbf{U}^T = \mathbf{C}_w$. $\mathbf{Q}$ and the diagonal matrix $\mathbf{\Lambda}$ are obtained by the diagonalisation $\mathbf{U}^{-1}\mathbf{C}_x\mathbf{U}^{-T} = \mathbf{Q}^T\mathbf{\Lambda}\mathbf{Q}$. By applying the linear transform to the observation vectors $\mathbf{y}_j$, i.e. $\mathbf{v}_j = (\mathbf{U}\mathbf{Q})^{-1}\mathbf{y}_j$, the conditional density of $\mathbf{v}_j$ is given:

$$f_{\mathbf{v}|z}\left(\mathbf{v}_j|z\right) = N\left(\mathbf{y};\mathbf{0},z\mathbf{\Lambda} + \mathbf{I}\right) \tag{5.28}$$

In [Goossens et al., 2009d], we show that the conditional density $f_{\mathbf{y}|z,H}\left(\mathbf{y}|z,H_0\right)$ can also be expressed in this basis as:

$$f_{\mathbf{y}|z,H}\left(\mathbf{y}|z,H_0\right) = N\left(\mathbf{y};\mathbf{0},(z^{-1}\mathbf{\Lambda}^{-1} + T^{-2}\mathbf{I})^{-1} + \mathbf{I}\right) \tag{5.29}$$

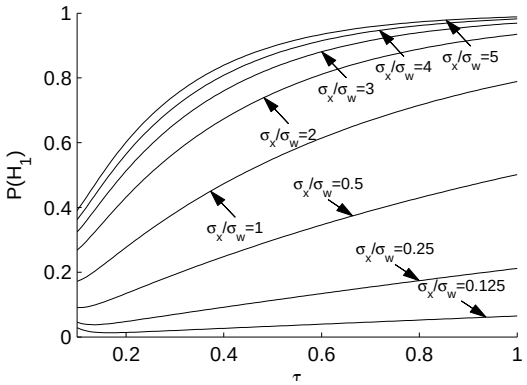


Figure 5.10: Probability of signal presence $\mathrm{P}(H_1)$ as function of τ, according to (5.31) for different ratios σ_x/σ_w and for $d = 9$.

Since the linear transform matrix $(\mathbf{UQ})^{-1}$ only has to be computed once per subband, independent of z, this greatly reduces the computational complexity of the proposed method, since the estimation rule (5.20) using only requires the evaluation of Gaussian densities with diagonal covariance matrix in $\mathbf{v}_j$.

Finally, the probability $\mathrm{P}(H_0) = 1 - \mathrm{P}(H_1)$ globally estimates the absence of the signal of interest on the *whole* subband:

$$\mathrm{P}(H_0) = \int_0^{+\infty} f_z(z) \left(\prod_{i=1}^{d} \frac{T^2}{T^2 + z\mathbf{\Lambda}_{ii}} \right)^{1/2} \mathrm{d}z. \tag{5.30}$$

In case of diagonal covariance matrices ($\mathbf{C}_x = \sigma_x^2\mathbf{I}$, $\mathbf{C}_w = \sigma_w^2\mathbf{I}$) and for the threshold $T = 1$, we find $\mathbf{\Lambda}_{ii} = \sigma_x^2/\sigma_w^2$ such that:

$$\mathrm{P}(H_0) = 1 - \mathrm{P}(H_1) = \int_0^{+\infty} f_z(z) \left(\frac{\sigma_w^2}{z\sigma_x^2 + \sigma_w^2} \right)^{d/2} dz. \tag{5.31}$$

This is weighted average of the ratios of the volumes of the hyperspheres with radiuses σ_w and $\sqrt{z\sigma_x^2 + \sigma_w^2}$. It is interesting to note that the weighting function $f_z(z)$, which is the density of a Gamma distribution, inherently relates the probability $\mathrm{P}(H_0)$ to kurtosis of the coefficients in the considered subband: if the kurtosis is low ($\tau \to 1$), many coefficients will contain a significant noise-free component ($\mathrm{P}(H_1) \to 1$). This is illustrated in Figure 5.10.

In Figure 5.11, the conditional densities $f_{\mathbf{x}|H}(\mathbf{x}|H_0)$ and $f_{\mathbf{x}|H}(\mathbf{x}|H_1)$ are shown for a two-dimensional random vector $\mathbf{x}$, corrupted with positively (in 2D) correlated Gaussian noise. In this case, the positive correlation between the noise components and negative correlation between the noise-free signal components cause a diagonal cut in $f_{\mathbf{x}|H}(\mathbf{x}|H_1)$. In the absence of correlations between the noise components and between the noise-free signal components, this cut is ring-shaped. The shrinkage function and its contours are depicted in

Figure 5.11(c)-(d). The contours are *not* elliptical, despite the signal and noise pdfs having elliptically symmetric distributions. This is because the axes of these ellipses are not aligned to each other in this example. Hence in contrast to the MMSE/MAP estimates from Section 5.2.2, our probabilistic shrinkage technique can deal with situations in which both the signal vector $\mathbf{x}$ and the observation vector $\mathbf{y}$ are correlated.

In Figure 5.12, the experiment is repeated for spherical symmetric signal and noise densities. As expected, the shrinkage factor $\mathrm{P}(H_1|\mathbf{y}_j)$ has also spherical isoprobability contours.

Furthermore, *Vector-ProbShrink* can be combined with the HMT interscale model, simply by replacing the posterior probability $\mathrm{P}(H_1|\mathbf{y}_j)$ with the one predicted by the HMT model $\mathrm{P}\left(H_1|\mathbf{y}^{(i)},...,\mathbf{y}^{(I)}\right)$:

$$\hat{\mathbf{x}}_j = \mathrm{P}\left(H_1|\mathbf{y}^{(i)},...,\mathbf{y}^{(I)}\right)\mathbf{y}_j. \tag{5.32}$$

Further details with respect to the HMT model initialization and other implementation specific information are given in [Goossens et al., 2009d].

Finally, we remark that equation (5.20) can also be interpreted as an approximation of MMSE estimator for our model from (3.6) with $\mathrm{E}[\mathbf{x}_j|\mathbf{y}_j,H_1] \approx \mathbf{y}_j$ and $\mathrm{E}[\mathbf{x}_j|\mathbf{y}_j,H_0] \approx \mathbf{0}$:

$$\hat{\mathbf{x}}_j = \mathrm{E}[\mathbf{x}_j|\mathbf{y}_j] = \mathrm{P}(H_1|\mathbf{y}_j)\,\mathrm{E}[\mathbf{x}_j|\mathbf{y}_j,H_1] + \mathrm{P}(H_0|\mathbf{y}_j)\,\mathrm{E}[\mathbf{x}_j|\mathbf{y}_j,H_0]. \tag{5.33}$$

In case we are (almost) certain that a given wavelet coefficient vector is purely noise we select $\mathbf{0}$ as the estimate for the noise-free coefficient vector, hence $\mathrm{E}[\mathbf{x}_j|\mathbf{y}_j,H_0] \approx \mathbf{0}$. On the other hand, using this approximation, significant structures like edges are preserved and no noise is suppressed: $\mathrm{E}[\mathbf{x}_j|\mathbf{y}_j,H_1] \approx \mathbf{y}_j$. This results in the shrinkage rule (5.20).

5.2.4 MMSE estimation for MPGSM

Because the MPGSM model from Section 3.4.4 can account for the local variability of the covariance matrix of the noise-free coefficients, while the GSM model from the previous section assumes that the covariance matrix is constant, MPGSM is generally more powerful than GSM. In this section, we derive the MMSE estimator for estimating local neighborhoods of noise-free coefficient vectors from the observed noisy coefficient vectors using the MPGSM prior model. In MPGSM, each noise-free signal vector $\mathbf{x}_j$ is decomposed into two parts (3.30):

$$\mathbf{x}_j = \mathbf{V}_k\mathbf{t}_j + \bar{\mathbf{V}}_k\mathbf{r}_j.$$

where we assume that $\mathbf{V}_k, k = 1,...,K$ are orthogonal projection matrices and that $\bar{\mathbf{V}}_k, k = 1,...,K$ are the complementary orthogonal projection matrices (see Section 3.4.4). Consequently, $\mathbf{t}_j$ and $\mathbf{r}_j$ are orthogonal projections of $\mathbf{x}_j$ onto the basis vector from $\mathbf{V}_k$ and $\bar{\mathbf{V}}_k$, respectively ($\mathbf{t}_j = \mathbf{V}_k^T\mathbf{x}_j$ and $\mathbf{r}_j = \bar{\mathbf{V}}_k^T\mathbf{x}_j$). In the following, we will consider one position in the subband and

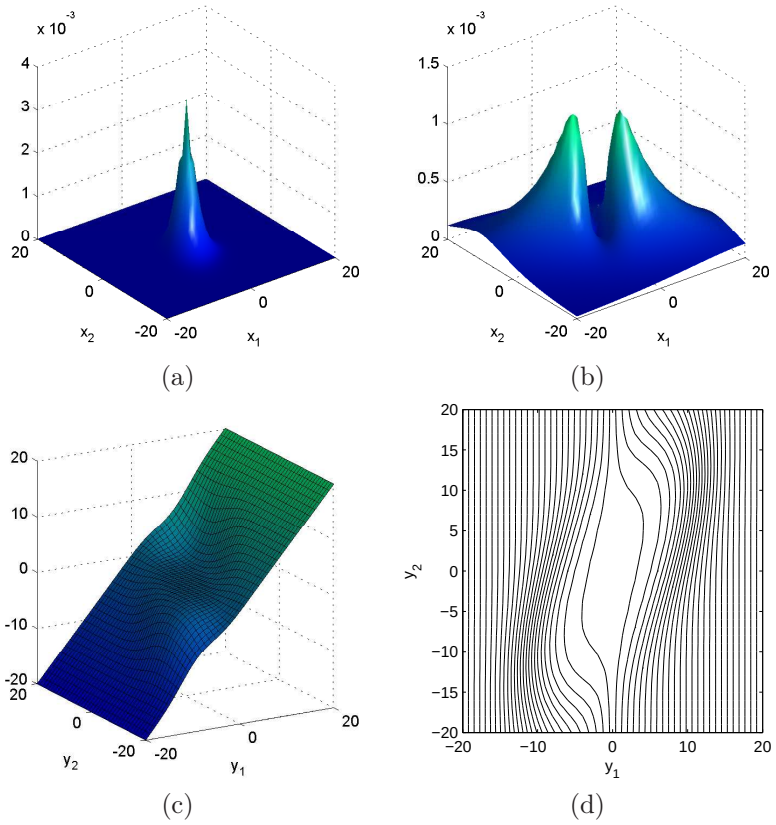


Figure 5.11: Illustration of the densities (*non-spherically symmetric* case), modeling a wavelet coefficient x_1 and its right neighbor x_2. (a) Conditional density $f_{\mathbf{x}|H}(\mathbf{x}|H_0)$, (b) Conditional density $f_{\mathbf{x}|H}(\mathbf{x}|H_1)$, (c) The shrinkage function $\mathrm{P}(H_1|\mathbf{y})\, y_1$, (d) Isocontours of (c).

drop the position subscripts j, to simplify the notations. We consider the estimation of this noise-free signal vector, in an additive noise model:

$$\mathbf{y} = \mathbf{x} + \mathbf{w}.$$

By relying on this relationship and the prior distribution (3.36), the observation model likehood function can be computed through several marginalizations over

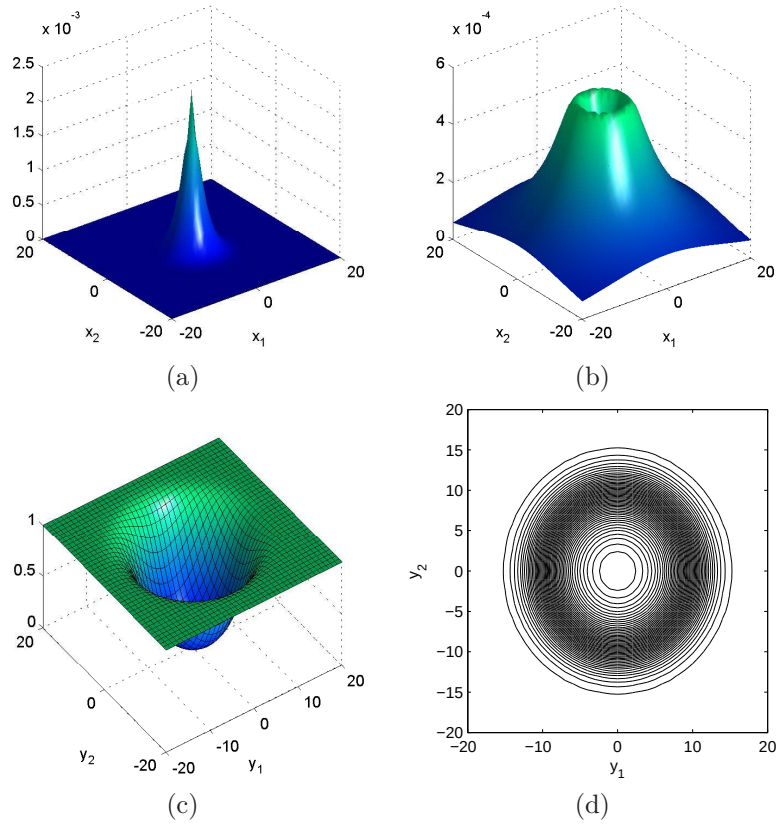


Figure 5.12: Illustration of the densities (*spherically symmetric* case), modeling a wavelet coefficient x_1 and its right neighbor x_2. (a) Conditional density $f_{\mathbf{x}|H}(\mathbf{x}|H_0)$, (b) Conditional density $f_{\mathbf{x}|H}(\mathbf{x}|H_1)$, (c) The shrinkage factor $\mathrm{P}(H_1|\mathbf{y})$, (d) Isocontours of (c).

its hidden variables (H_k and z):

$$
\begin{aligned}
f_{\mathbf{y}}(\mathbf{y}) &= \sum_{k=1}^{K} \mathrm{P}(H_k)\, f_{\mathbf{y}|H}(\mathbf{y}|H_k) \\
&= \sum_{k=1}^{K} \mathrm{P}(H_k) \int_{\mathbb{R}^d} f_{\mathbf{x}|H}(\mathbf{x}|H)\, f_{\mathbf{y}|\mathbf{x},H}(\mathbf{y}|\mathbf{x},H)\, \mathrm{d}\mathbf{x} \qquad (5.34) \\
&= \sum_{k=1}^{K} \mathrm{P}(H_k) \int_{-\infty}^{+\infty} \mathrm{d}z f_z(z) \int_{\mathbb{R}^d} \mathrm{d}\mathbf{x} f_{\mathbf{x}|H}(\mathbf{x}|z,H_k)\, f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \qquad (5.35)
\end{aligned}
$$

with $f_{\mathbf{x}|H}(\mathbf{x}|z,H_k) = f_{\mathbf{r}|H}(\bar{\mathbf{V}}_k^T\mathbf{x}|H_k) f_{\mathbf{t}|z,H}(\mathbf{V}_k^T\mathbf{x}|z,H_k)$ (see (3.36)). To significantly reduce the number of model parameters for MPGSM compared to

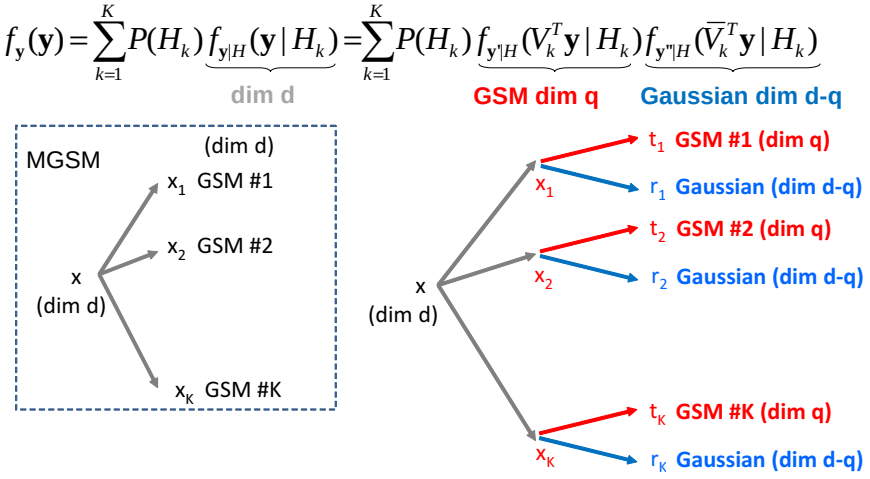


Figure 5.13: Illustration of MPGSM denoising.

MGSM, we applied several restrictions to the signal covariance matrix $\mathbf{C}_x$ in Section 3.4.4 (equation (3.31)). To fully enjoy the benefits of this parameter reduction and to arrive at a denoising estimator with a lower computational complexity than MGSM, we will apply the restrictions made to $\mathbf{C}_x$ to the noise covariance matrix $\mathbf{C}_w$ as well. Therefore, we will assume that under hypothesis H_k, the projected and complementary projected noise components, respectively $\mathbf{V}_k^T\mathbf{w}$ and $\bar{\mathbf{V}}_k^T\mathbf{w}$ are uncorrelated, such that the noise covariance matrix $\mathbf{C}_w$ satisfies a specific property:

$$\mathbf{C}_w = \mathbf{V}_k \left(\mathbf{V}_k^T \mathbf{C}_w \mathbf{V}_k\right) \mathbf{V}_k^T + \bar{\mathbf{V}}_k \left(\bar{\mathbf{V}}_k^T \mathbf{C}_w \bar{\mathbf{V}}_k\right) \bar{\mathbf{V}}_k^T \quad (\text{under } H_k)$$

where additionally $\bar{\mathbf{V}}_k^T \mathbf{C}_w \bar{\mathbf{V}}_k$ is assumed to be a diagonal matrix. The integration in (5.35) is then fairly simple, as both conditional densities $f_{\mathbf{x}|H}(\mathbf{x}|z, H_k)$ and $f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ are Gaussian. Furthermore, every *projected* component $\mathbf{V}_k^T\mathbf{y}|H_k$ is *Gaussian Scale Mixture plus Gaussian noise* distributed, while every *complementary projected* component $\bar{\mathbf{V}}_k^T\mathbf{y}|H_k$ is *Gaussian* distributed. This allows us to estimate the noise-free coefficients for each hypothesis H_k in both projection domains (as $\hat{\mathbf{x}} = \mathbf{V}_k\hat{\mathbf{t}}_k + \bar{\mathbf{V}}_k\hat{\mathbf{r}}_k$), followed by aggregation of the resulting estimates according to the posterior probability of each hypothesis. Conditioned on the hypothesis H_k, the MMSE estimator for the noise-free coefficients is equivalent to that for the observation model in [Portilla et al., 2003]:

$$\begin{aligned}\hat{\mathbf{t}}_k &= \mathrm{E}\left[\mathbf{t}|\mathbf{V}_k^T\mathbf{y}, H_k\right] \\ &= \int_0^{+\infty} f_{z|\mathbf{t},H}\left(z|\mathbf{t}, H_k\right) z\mathbf{C}_{u,k} \left(z\mathbf{C}_{u,k} + \mathbf{V}_k^T\mathbf{C}_w\mathbf{V}_k\right)^{-1} \mathbf{V}_k^T\mathbf{y}\mathrm{d}z, \end{aligned} \quad (5.36)$$

which is a weighted average of local Wiener solutions for different z. If we denote the covariance matrix of the projected noise component $\bar{\mathbf{V}}_k^T\mathbf{w}$ as $\mathbf{\Omega}_k$,

we estimate $\hat{\mathbf{r}}_k$ in the complementary space as follows:

$$\hat{\mathbf{r}}_k = \mathrm{E}\left[\mathbf{r}_j|\bar{\mathbf{V}}_k^T\mathbf{y}, H_k\right] = \boldsymbol{\Psi}_k\left(\boldsymbol{\Psi}_k + \bar{\mathbf{V}}_k^T\mathbf{C}_w\bar{\mathbf{V}}_k\right)^{-1}\bar{\mathbf{V}}_k^T\mathbf{y} \tag{5.37}$$

By the assumed diagonality of the covariance matrices in (5.37) each component can be estimated *independently*, which offers computational advantages especially when $q \ll d$. Finally, $\hat{\mathbf{x}}$ is obtained by an overall optimization over K models $H_1, ..., H_K$, by averaging over the solutions of all K MPGSM components:

$$\mathbf{x} = \mathrm{E}\left[\mathbf{x}|\mathbf{y}\right] = \sum_{k=1}^{K}\mathrm{P}\left(H_k|\mathbf{y}\right)\left(\mathbf{V}_k\hat{\mathbf{t}}_k + \bar{\mathbf{V}}_k\hat{\mathbf{r}}_k\right) \tag{5.38}$$

In (5.38), the presence of the posterior probability $\mathrm{P}\left(H_k|\mathbf{y}\right)$ reveals the adaptability of the model: the final estimate is a weighted mean of estimates according to different projection spaces and covariance matrices.

Figure 5.13 illustrates the MPGSM-MMSE estimator: while MGSM is simply a mixture of GSM distributions (*left*), MPGSM components decompose into GSM part $\mathbf{t}$ and a Gaussian part $\mathbf{r}$ (*right*). Denoising is performed by estimating the different sub-components using (5.36) and (5.37) and finally by averaging the results by applying (5.38). In Figure 5.14(c)-(d), the MMSE estimate (5.38) is depicted for the densities $f_{\mathbf{x}}\left(\mathbf{x}\right)$, $f_{\mathbf{y}}\left(\mathbf{y}\right)$, which are shown in respectively Figure 5.14(a)-(b). The prior density $f_{\mathbf{x}}\left(\mathbf{x}\right)$ clearly has a non-symmetric shape, this manifests in asymmetric and inequal multivariate shrinkage functions for $\left[\hat{\mathbf{x}}\right]_1$ and $\left[\hat{\mathbf{x}}\right]_2$.

5.2.5 Complex-wavelet based demosaicing

In the past, many techniques have been proposed for solving the demosaicing problem. Some authors focus on the *frequency domain* interpretation of the problem [Dubois, 2005, Alleysson and Susstrunk, 2005, Alleysson and de Lavarene, 2008]: by noting the similarities with luminance and chrominance demultiplexing in NTSC/PAL television [Alleysson and de Lavarene, 2008], analogous schemes can be devised for demosaicing. These schemes consist of linear filters that demultiplex luminance and chrominance signals and are efficiently implemented in the Fourier domain. Other techniques directly operate in the *image domain* because this domain allows for spatially adaptive directional filtering (e.g. [Kakarala and Baharav, 2002, Hirakawa and Parks, 2005a, Lukac and Plataniotis, 2005b, Muresan and Parks, 2005, Li, 2005, Kimmel, 1999, Lee et al., 2006, Menon et al., 2007]).

The problem of demosaicing is not only a matter of interpolation of missing color intensities: as we already explained in Chapter 4, the data captured by the sensors are also subject to a serie of image post-processing techniques, which modifies the noise signal-dependency characteristics. Existing demosaicing schemes attempt to preserve edge sharpness and textures, this is typically done by adapting the interpolation to the edge direction. Despite the fact that these techniques often give excellent results for *noise-free* images, their performance in the presence of noise is often very poor. This is because noise may

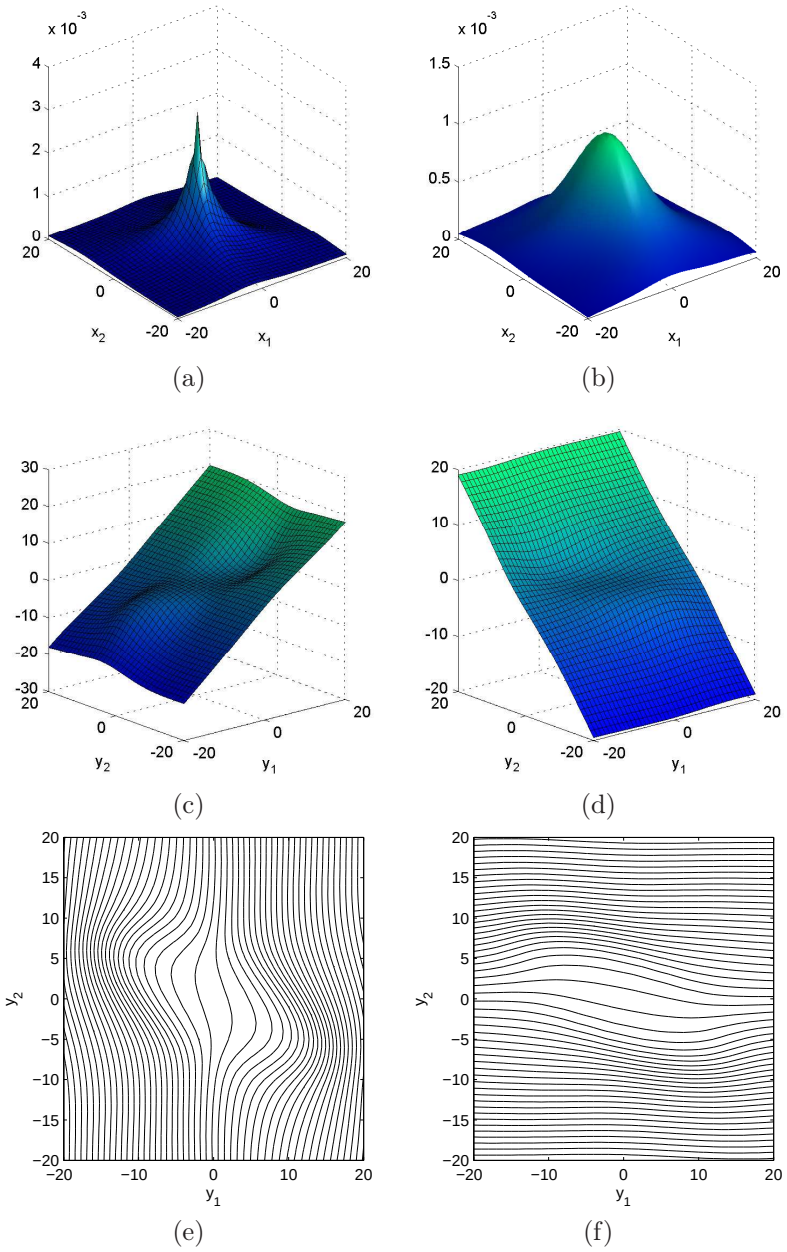


Figure 5.14: Probability densities and resulting shrinkage functions for MPGSM (a) prior pdf $f_{\mathbf{x}}(\mathbf{x})$, (b) likelihood function $f_{\mathbf{y}}(\mathbf{y})$, (c) first component of the estimate $[\hat{\mathbf{x}}(\mathbf{y})]_1$, (d) second component of the estimate $[\hat{\mathbf{x}}(\mathbf{y})]_2$, (e) iso-contours of (c), (f) iso-contours of (d).

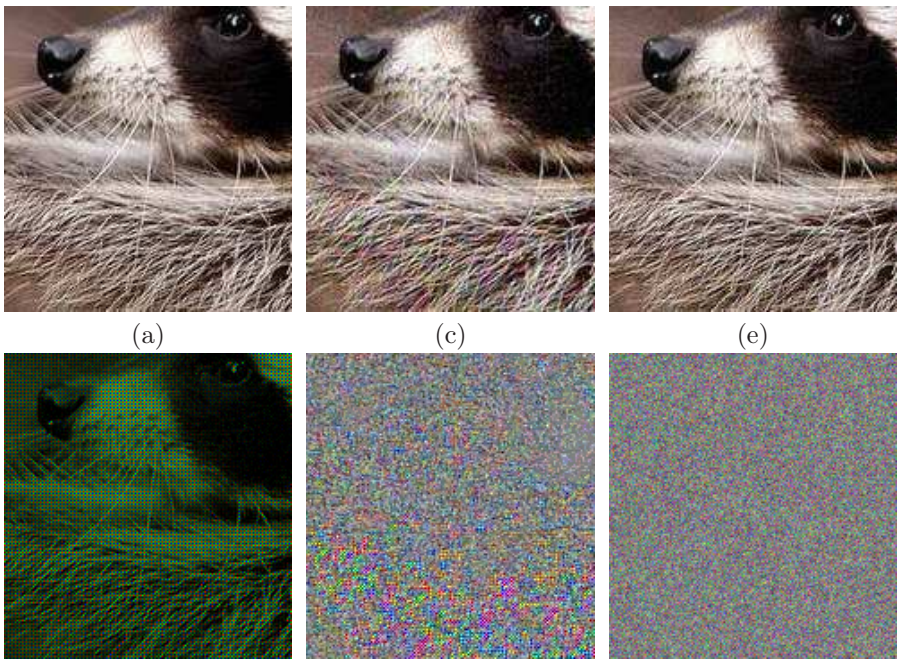
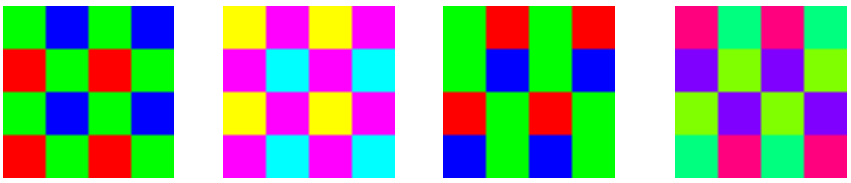


Figure 5.15: Illustration of the influence of noise in demosaicing (a) Original, noise-free image (b) CFA image corrupted with artificial white Gaussian noise ($\sigma = 10$), (c) Demosaicing result using DL-MMSE [Zhang and Wu, 2005], (d) 3× enhanced difference image between (c) and (a), (e) Image with white Gaussian noise ($\sigma = 10$) added after demosaicing (for comparison to (c)), (f) 3× enhanced difference image between (e) and (a).

lead to false edge structures which are visually very disturbing. Furthermore, noise patterns may be correlated in such a complicated way that it becomes very difficult to restore the image afterwards, even using state-of-the-art denoising techniques [Hirakawa, 2008b]. We illustrate this in Figure 5.15: due to the presence of noise, the demosaicing technique is not able to correctly interpolate along the hairs of the badger, causing color artifacts in Figure 5.15(c). Removal of these noise artifacts would require a sophisticated noise reduction algorithm. However, noise reduction as post-processing is generally considered to be impractical, as the underlying image models for demosaicing and denoising are often quite different. Furthermore, the image model mismatch causes many complicated interactions that are intractable to solve mathematically [Hirakawa, 2008b].

More recently, Hirakawa extended the frequency-domain point of view to the wavelet domain [Hirakawa, 2008b], to combine noise reduction with demosaicing [Hirakawa and Parks, 2006, Hirakawa et al., 2007, Hirakawa, 2008b]. This way, the demosaicing procedure can take noise properties into account



(a) Bayer (b) Cyan-Magenta-Yellow (c) Lukac (d) Hirakawa - Pat. I

Figure 5.16: Color Filter Array patterns used for demosaicing. (a) [Bayer, 1976], (b) [Hirakawa and Wolfe, 2008], (c) [Lukac and Plataniotis, 2005a], (d) [Hirakawa and Wolfe, 2008].

and the denoising can be optimized for the specific interpolation formulas being used, while still relying on the same image and noise model. Furthermore, many efficient prior models and noise reduction techniques are available in the wavelet domain, hence it is attractive to design the demosaicing method in this domain as well.

Spectral analysis of Color Filter Arrays

The Bayer pattern CFA [Bayer, 1976], shown in Figure 5.16(a), uses a quincunx sampling of the red, green and blue color components, in which the green component are sampled twice as dense as the other color components. Although the Bayer pattern is the de facto standard in industry, various alternative CFAs have been proposed and/or studied. For example, the CMY pattern (Figure 5.16(b)) is often used in video-cameras because C/M/Y photosensitive elements are less sensitive to noise and because video frame rates do not permit long integration times [Hirakawa, 2008b]. Other patterns are designed to improve the characteristics of the Bayer pattern: the patterns in Figure 5.16(c) specifically maximizes the demosaicing performance while keeping the computational complexity low [Lukac and Plataniotis, 2005a]. The pattern in Figure 5.16(c) is designed to jointly maximize the bandwidth of the luminance and chrominance channels while maintaining perfect reconstruction [Hirakawa and Wolfe, 2008]. In the following, we will only consider the Bayer CFA because this pattern is dominantly used in practice, although what follows is not restricted to Bayer patterns.

Let $\mathbf{c}(\mathbf{p}) \in \left\{ \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T, \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T, \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T \right\}$ denote the vector of RGB-color intensities of the CFA at position $\mathbf{p}$. The mosaic image acquired by a DSC, $y_{\text{mosaic}}(\mathbf{p})$, can be expressed in terms of the original "ideal" image $\mathbf{x}(\mathbf{p})$ by:

$$y_{\text{mosaic}}(\mathbf{p}) = \mathbf{c}^T(\mathbf{p})\mathbf{x}(\mathbf{p}). \tag{5.39}$$

The goal of demosaicing is to recover the hypothetical "original" image $\mathbf{x}(\mathbf{p})$ from the mosaic image $y_{\text{mosaic}}(\mathbf{p})$.[5] Note that in the mosaic image $y_{\text{mosaic}}(\mathbf{p})$

[5] In practice, $y_{\text{mosaic}}(\mathbf{p})$ is also corrupted by noise, but to keep the discussion simple, we

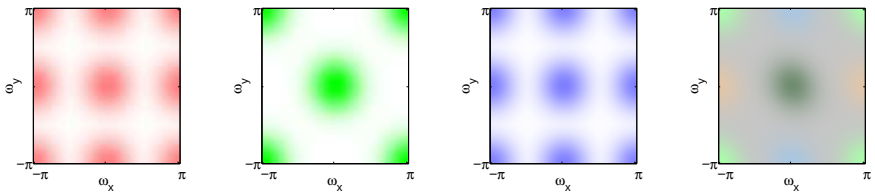


Figure 5.17: Frequency spectra of the Red (a), Green (b) and Blue (c) channels in a CFA. (d) Frequency spectrum of the mosaic image, which is a basically superposition of (a), (b), (c).

neighboring pixel intensities are measured using different color filters, hence the mosaic image should not be treated as a "photographic" image. Nevertheless, investigating the spectrum of the mosaic image $y_{\text{mosaic}}(\mathbf{p})$ allows us to gain some insight in the demosaicing problem.

First, we note that the spectrum of the mosaic image is a superposition of the individual spectra of the images $[\mathbf{c}(\mathbf{p})]_k [\mathbf{x}(\mathbf{p})]_k$, $k = 1, ..., 3$. In Figure 5.17, these spectra are depicted for an artificial image $\mathbf{x}(\mathbf{p})$ (for which we used a monochrome Gaussian function for illustrative purposes). The subsampling of the color components of the CFA causes the creation of frequency aliasing in the frequency domain. By the aliasing, the spectrum of the center-band signal is replicated at spatial frequencies $(\pm\pi, \pm\pi)$ (*green*) or at $(\pm\pi, \pm\pi)$, $(0, \pm\pi)$ and $(\pm\pi, 0)$ (*red* and *blue*).[6] Because the spectrum of the mosaic image is the superposition of the spectra of the individual color channels, replications of the center-band signals will also occur in the mosaic spectrum: if we take a closer look at Figure 5.17(d) we note that the center-band is in fact the luminance channel, while the red, green and blue channels can be found in the side-bands. The demosaicing problem can then effectively be solved by demultiplexing the spectra of the individual color channels from the spectrum of the mosaic image. This is possible under appropriate bandwidth assumptions for the individual color components.

Wavelet domain demosaicing scheme

To define these bandwidth assumptions, [Hirakawa, 2008b] considers the green (G) channel as luminance channel and the difference between red and green (R-G) and blue and green (B-G) as a crude approximation to the chrominance

will consider noise-free images for the moment.

[6] Due to the shifts of the blue and red color elements in the CFA, the angle of the Fourier responses is rotated by 180° (i.e. the sign is switched) at either $(\pm\pi, \pm\pi)$, $(0, \pm\pi)$ or $(\pm\pi, 0)$. The sign is not shown in Figure 5.17, as only the magnitude is depicted.

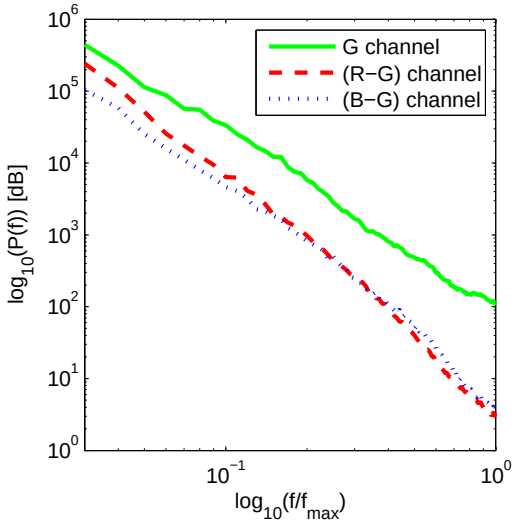


Figure 5.18: Normalized radially averaged PSDs over a set of six color photographs. The photographs were first low-pass filtered and subsampled by a factor 2, to eliminate effects of the demosaicing used during the acquisition of the images themselves.

channel. This defines the color transform:

$$\begin{cases} Y &= G \\ \alpha &= R - G \\ \beta &= B - G \end{cases} \tag{5.40}$$

The reason for choosing this transform over better suited *YCbCr* or *La*b** transforms is merely because it will make the demultiplexing reconstruction formulas simpler. For the new channels Y, α, β, the first assumption is that the green channel signal (Y) is bandlimited to 3/4 of the Nyquist bandwidth. The second assumption is that the red-green and blue-green differences α, β are bandlimited to 1/4 of the Nyquist bandwidth. Both assumptions reflect the fact that the human eye is less sensitive to details in the chrominance information than in the luminance information (this assumption is also used in e.g. the JPEG compression standard). Because the green channel is a crude approximation to the luminance channel, we checked these assumptions by computing normalized radially averaged PSDs of the Y, α, β-channels over six color photographs. The result is shown in Figure 5.18. In general, the approximation is quite good: almost over the whole frequency range the power of the α- and β-channels is about 10dB lower than the power of the Y-channel. Based on these assumptions, the mosaic spectrum (from Figure 5.17(d)) can be schematically interpreted as in Figure 5.19(a). The relative phase shift of the spectra of the α_1- and α_2-channels is visualized by the horizontal and respectively vertical stripes pattern in Figure 5.19(a).

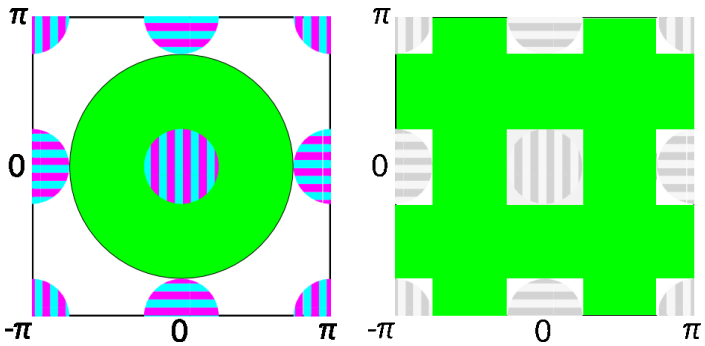


Figure 5.19: (a) Schematic interpretation of the mosaic image spectrum. The Y-channel (green) is symbolized by the green disk. The α_1- and α_2-channels are indicated by respectively the magenta disks and cyan disks. (b) Frequency bands of the mosaic image spectrum that are not assumed to contain chromacity information.

Now, to separate the green from the red-green and blue-green difference images under the bandwidth assumptions, Hirakawa proposes to use a two-level wavelet packet transform (WPT). This is because each wavelet subband in a two-level WPT has a frequency support of approximately 1/4 of the original image bandwidth both horizontally and vertically, which easily allows to demultiplex the different parts of the frequency spectrum of the mosaic image.

Let $A_j^{(LL,LL)}$ denote the scaling coefficient at position j of the (LL,LL)-subband of the WPT of the mosaic image, then given the bandwidth assumptions, $A_j^{(LL,LL)}$ can be expressed in terms of the scaling coefficients of the luminance channels channels, $Y_j^{(LL,LL)}$, and the chrominance channels, $\alpha_j^{(LL,LL)}$ and $\beta_j^{(LL,LL)}$:

$$A_j^{(LL,LL)} = Y_j^{(LL,LL)} + \alpha_j^{(LL,LL)} + \beta_j^{(LL,LL)}. \tag{5.41}$$

Our goal is to recover the individual luminance and chrominance channels $Y_j^{(LL,LL)}$, $\alpha_j^{(LL,LL)}$ and $\beta_j^{(LL,LL)}$ from $A_j^{(LL,LL)}$. For the (HL,LL) subband (see Figure 5.19(a)), we can write a similar expression:

$$A_j^{(H^\star L,LL)} = \rho_{1,1,0}\alpha_j^{(LL,LL)} + \rho_{2,1,0}\beta_j^{(LL,LL)} \tag{5.42}$$

where the superscript '$\star$' in $(H^\star L,LL)$ signifies time reversal of the row wavelet filter for the first scale and where $\rho_{\cdot,\cdot,\cdot}$ are modulation factors that depend on the position of the invidual color sensor elements in the CFA. More specifically, for the Bayer pattern, these modulation factors are defined as:

$$\rho_{k,m,n} = (-1)^{mP_k(1,0)+nP_k(0,1)+(m+n)P_k(1,1)}, \tag{5.43}$$

where $P_1(m,n)$ and $P_2(m,n)$ are 1 if the Bayer Pattern has a red, respectively blue value at location (m,n), and 0 otherwise. Finally, applying the same tricks

to the HL, LL subbands gives:

$$A_j^{(H^\star H^\star, LL)} = \rho_{1,1,1}\alpha_j^{(LL,LL)} + \rho_{2,1,1}\beta_j^{(LL,LL)}, \tag{5.44}$$

with $A_j^{(H^\star H^\star, LL)}$ a WPT coefficient of subband (HH, LL) in which both the wavelet row and column filters have been time-reversed. Now, equations (5.41)-(5.44) constitute a system of linear equations with a unique solution:

$$\begin{aligned}
\widehat{\alpha_j^{(LL,LL)}} &= \frac{\rho_{2,1,0}A_j^{(H^\star H^\star, LL)} - \rho_{2,1,1}A_j^{(H^\star L, LL)}}{\rho_{1,1,1}\rho_{2,1,0} - \rho_{1,1,0}\rho_{2,1,1}} \\
\widehat{\beta_j^{(LL,LL)}} &= \frac{\rho_{1,1,1}A_j^{(H^\star L, LL)} - \rho_{1,1,0}A_j^{(H^\star H^\star, LL)}}{\rho_{1,1,1}\rho_{2,1,0} - \rho_{1,1,0}\rho_{2,1,1}} \\
\widehat{Y_j^{(LL,LL)}} &= A_j^{(LL,LL)} - \widehat{\alpha_j^{(LL,LL)}} - \widehat{\beta_j^{(LL,LL)}}.
\end{aligned} \tag{5.45}$$

These formulas effectively demultiplex the *scaling coefficients* of the luminance and chrominance channels of the image. Because the bandwidth of the luminance channel Y is assumed to be 3/4 of the Nyquist bandwidth while the bandwidth of the chrominance channels α, β is 1/2 of the Nyquist bandwidth, the high-pass and band-pass frequencies of the luminance channel Y can be partially reconstructed by "filling in" the corresponding wavelet coefficients (see Figure 5.19(b)):

$$\widehat{Y_j^{(mm,nn)}} = A_j^{(mm,nn)} \text{ if } (mm, nn) \neq (LL, LL)\,. \tag{5.46}$$

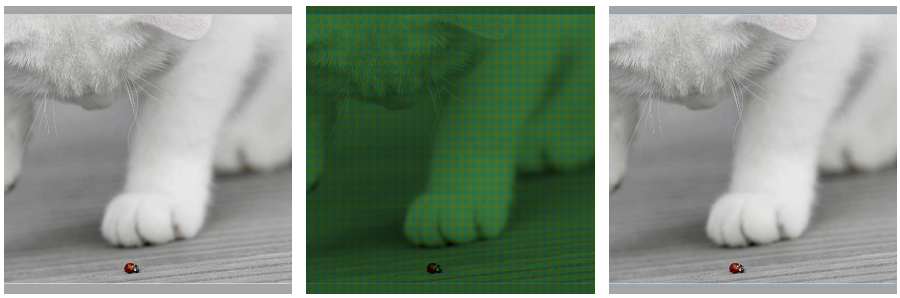
Consequently, this demosaicing algorithm can be summarized as follows:

1. Compute a two-level wavelet packet transform of the mosaic image (the $(\cdot\cdot, LL)$ are computed using time-reversal of the wavelet filters).
2. Use equations (5.45) and (5.46) to demultiplex the luminance Y and chrominance α, β channels.
3. Apply the inverse color transform for (5.40) to obtain R, G, B channels.
4. Perform an inverse wavelet packet transform for each of the channels.

In Figure 5.20, a demosaicing result for the above image is shown. Enlargements with more details are given in Figure 5.21. It can be noted that despite the fact that the technique overall gives a very high visual quality (for example, the fine cat hairs are reconstructed well), the method suffers from 1) discoloration artifacts and 2) zippering artifacts.

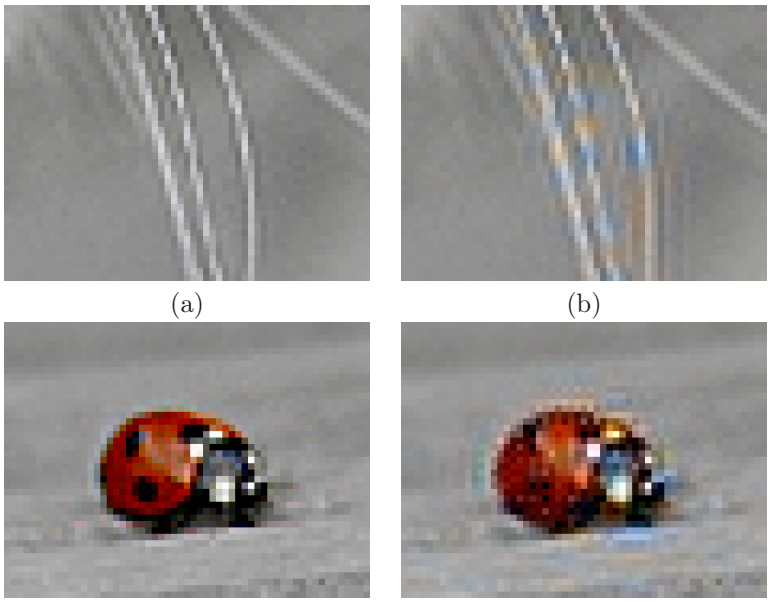
Novel complex-wavelet based approach

Because of the potential of the wavelet-based demosaicing approach, we improved this technique in our recent work [Aelterman et al., 2009]. First of all,



(a) (b) (c)

Figure 5.20: A wavelet-based demosaicing result (a) original image, (b) mosaic image, (c) wavelet-based demosaiced image.



(a) (b)

(c) (d)

Figure 5.21: Crop-out of the wavelet-based demosaicing result revealing color artifacts and zipperings: (a),(c) original image, (b),(d) demosaiced image.

the issues with the DWT discussed in Section 2.1.4, specifically shift variance and aliasing, also negatively impact the above algorithm. The use of dual-tree complex wavelet packets not only allows us to tackle these problems, but also makes it possible to deal with discoloration artifacts and even to further improve the quality of the reconstructed images, as we will show further.

We recall that the primary advantage of using wavelet-based processing in-

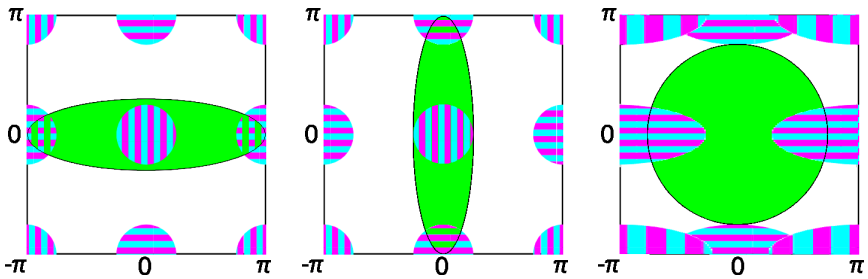


Figure 5.22: Schematic representation of two different violations of the bandwidth assumptions: (a)-(b) mosaic spectrum of an image with a sharp vertical (a) and horizontal edge (b); (c) The chrominance bandwidth exceeds the maximal allowed bandwidth (less common).

stead of Fourier demosaicing lies in the joint spatial and frequency localization of the wavelet coefficients. By looking at the mosaicing spectra in Figure 5.19, we note that the reconstruction formulas (5.46) are *not unique*! Under the bandwidth assumptions, we have for example that $A_j^{(H^\star L,LL)} = A_j^{(LH^\star,LL)}$. This ambiguity stems from the fact that several aliasing copies of the chrominance channels are present in the spectrum of the mosaic image. If the bandwidth assumptions are perfectly fulfilled, it does not matter which aliasing copy is being used to reconstruct the luminance channel. However, the discoloration artifacts in Figure 5.21 occur when these assumptions are not correct: either when the bandwidth of the luminance channel exceeds 3/4 of the Nyquist bandwidth, or when the bandwidth of the chrominance channel exceeds 1/4 of the Nyquist bandwidth. The former case occurs for example for *sharp horizontal and vertical edges* (see Figure 5.21(b)). A schematical illustration is given in Figure 5.22(a)-(b). In the latter case, chrominance information is incorrectly transferred to the *high-pass luminance frequencies* of the demosaiced image (see Figure 5.22(c)), causing the zipper artifacts in Figure 5.21(d).

The solution to this problem lies in 1) detecting these two artifacts, 2) correcting for them. The use of wavelets gives a lot of advantages here: in contrast to Fourier-based demosaicing techniques, we can easily make the demosaicing formulas (5.45) and (5.46) adaptive to the local context.

Solving discoloration artifacts and reconstruction of luminance high-pass frequencies

As already said, discoloration artifacts occur when the luminance channel violates its bandwidth assumption. This will most likely result in an increase in the energy of either the $A_j^{(H^\star L,LL)}$ or $A_j^{(LH^\star,LL)}$-coefficients. The complex wavelet framework allows to detect high frequencies by using the magnitude of the complex wavelet coefficients. For the DT-CWT, the magnitude is (approximately) alias-free and shift-invariant. This leads to the following edge

detector[7]:

$$\begin{cases} \text{horizontal edge,} & \text{if } \left|A_j^{(H^\star L,LL)}\right|^2 > \left|A_j^{(LH^\star,LL)}\right|^2 + T \\ \text{vertical edge,} & \text{if } \left|A_j^{(H^\star L,LL)}\right|^2 < \left|A_j^{(LH^\star,LL)}\right|^2 + T \\ \text{no edge,} & \text{else} \end{cases} \tag{5.47}$$

with T a predefined threshold. Depending on the local outcome of the edge detection, we have the situation of Figure 5.19(a), Figure 5.22(b) or Figure 5.22(b). In the first case, we proceed with the standard reconstruction formulas (5.45). In the last two cases, we can recover the chrominance channels *exactly* from the chrominance aliasing copies that do not have a spectral overlap with the luminance channel. Next, if we have the chrominance channels, we can again reconstruct the luminance high-pass frequencies for the frequency plane region in which the luminance and chrominance channels overlap. This is achieved by a simple processing operations on the DT-CWT packet coefficients of the mosaic image, thereby relying on the frequency localizing properties of the complex wavelets.

In the results section (Section 5.4), we will show that these modifications lead to a vast improvement in demosaicing performance, compared to the (non-complex) wavelet-based demosaicing algorithm.

Because the demosaicing is completely performed in the DT-CWT domain, the method can easily be combined with denoising. A straightforward approach would be to first apply the reconstruction formulas, with our novel extensions. Next, the wavelet coefficients can be processed using any of our denoising methods from the previous sections. The primary advantage of this solution is that the DT-CWT only needs to be performed *once*, resulting in an efficient processing algorithm that could be implemented on a DSC in the future. However, this approach consists of two sequential steps and one may expect the performance to increase further by performing joint denoising and demosaicing. [Hirakawa, 2008b] propose to solve the joint problem by reformulating the reconstruction formulas in a Bayesian framework. Even though Hirakawa used a wavelet packet transform with cycle spinning, his joint technique can be seemlessly be combined with our complex-wavelet based technique. Further exploring this topic is one of the future research directions of our work.

[7]For mosaic images, traditional edge detection methods (e.g. Sobel, Canny) cannot be used directly, because the CFA causes discontinuities everywhere in the mosaic image; hence these methods would detect edges at every position in the image.

5.3 Bregman framework for image restoration

In this section, we present a general framework that can be used for solving a wide variety of image restoration problems. This framework emerged from a technique presented in [Bregman, 1967] that searches for the common points in two convex sets. Later, this technique has been applied to image restoration in [Osher et al., 2008, Goldstein and Osher, 2008]. Bregman considered the following generalized constrained optimization problem:

$$\min_{\mathbf{x}} J(\mathbf{x}), \quad \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}_0 \tag{5.48}$$

where $J(\mathbf{x})$ is a convex function. In general, this problem is very difficult to solve directly if $J(\mathbf{x})$ is non-differentiable. Therefore, continuation methods approximately solve (5.48), by translating the problem into an unconstrained optimization problem:

$$\min_{\mathbf{x}} J(\mathbf{x}) + \frac{\lambda_i}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}_0\|_2^2, \tag{5.49}$$

where $\lambda_i, i = 1, ..., I$ is an increasing sequence of penalty weights $\lambda_1 < \lambda_2 < \lambda_3 < \cdots < \lambda_I$. To fully enforce the constraint $\mathbf{A}\mathbf{x} = \mathbf{b}_0$, it is necessary to choose λ_I very large, such that the relative contribution of $J(\mathbf{x})$ in (5.49) becomes very small. Unfortunately, this causes computational instabilities in many applications [Goldstein and Osher, 2008].

Bregman proposed another approach to solve the optimization problem (5.48): by introducing the Bregman divergence, which is associated to the convex functional J:

$$D_J^{\mathbf{p}}(\mathbf{x}, \mathbf{x}_i) = J(\mathbf{x}) - J(\mathbf{x}_i) - \mathbf{p}^T(\mathbf{x} - \mathbf{x}_i) \tag{5.50}$$

where $\mathbf{p}$ is a subgradient[8] of J in $\mathbf{x}_i$. The Bregman divergence is always positive ($D_J^{\mathbf{p}}(\mathbf{x}, \mathbf{x}_i) \geq 0$) and convex in its first argument ($D_J^{\mathbf{p}}(\mathbf{x}, \mathbf{x}_{i+1}) \geq D_J^{\mathbf{p}}(\mathbf{x}_i, \mathbf{x}_{i+1})$ for $\mathbf{x}_i$ on the line segment between $\mathbf{x}$ and $\mathbf{x}_{i+1}$). Further, the Bregman divergence is not a distance measure because it is generally not symmetric.

An illustration of the Bregman divergence is given in Figure 5.23. From the figure, it can be seen that the Bregman divergence is a measure for the closeness to the optimum of the function $J(\mathbf{x})$. Hence, by minimizing the Bregman divergence, the optimum of the function can be found. Rather than choosing λ_i very large, Bregman suggested to solve (5.48) iteratively:

$$\begin{aligned} \mathbf{x}_{i+1} &= \arg\min_{\mathbf{x}} D_J^{\mathbf{p}}(\mathbf{x}, \mathbf{x}_i) + \lambda H(\mathbf{x}; \mathbf{b}_0) \\ &= \arg\min_{\mathbf{x}} J(\mathbf{x}) - \mathbf{p}_i^T(\mathbf{x} - \mathbf{x}_i) + \lambda H(\mathbf{x}; \mathbf{b}_0) \text{ and} \\ \mathbf{p}_{i+1} &= \mathbf{p}_i - \boldsymbol{\nabla} H\left(\mathbf{x}^{i+1}; \mathbf{b}_0\right) \end{aligned} \tag{5.51}$$

with $\mathbf{p}_i$ a subgradient of J in $\mathbf{x}_i$ and with λ now a constant. Such update from $\mathbf{x}_i \rightarrow \mathbf{x}_{i+1}$ is called a *Bregman iteration*. It can be shown that under weak

[8] A vector $\mathbf{p}$ is a subgradient of the convex function J in $\mathbf{x}_i$ if $J(\mathbf{x}) - J(\mathbf{x}_i) \geq \mathbf{p}^T(\mathbf{x} - \mathbf{x}_i)$.

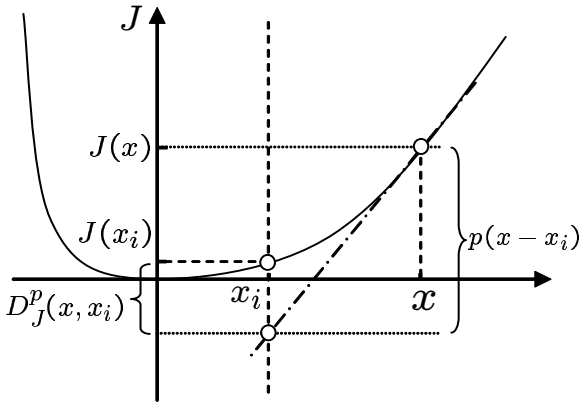


Figure 5.23: Illustration of the Bregman divergence associated to a convex functional $J(x)$.

assumptions for J and H, the value of $H(\mathbf{x}_i; \mathbf{b}_0)$ decreases in every iteration. Because λ is kept constant, the numerical problems of the continuation methods are avoided. Moreover, the algorithm will converge:: $H(\mathbf{x}_i; \mathbf{b}_0)$ can be made arbitrarily small by using more iterations: $H(\mathbf{x}_i; \mathbf{b}_0) \to 0$ as $i \to \infty$.

For $H(\mathbf{x}; \mathbf{b}_0) = \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}_0\|_2^2$, the Bregman iteration (5.51) can be computed more efficiently using [Goldstein and Osher, 2008]:

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} J(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}_i\|_2^2 \tag{5.52}$$

$$\mathbf{b}_{i+1} = \mathbf{b}_i + (\mathbf{b}_0 - \mathbf{A}\mathbf{x}_{i+1}) . \tag{5.53}$$

where the vector $\mathbf{b}_i$ in the constraint $\|\mathbf{A}\mathbf{x} - \mathbf{b}_i\|_2^2$ is now updated iteratively. This has an intuitive interpretation: in every iteration, the error term $\mathbf{b}_0 - \mathbf{A}\mathbf{x}_{i+1}$ is added back to the right handed side of the constraint $\mathbf{A}\mathbf{x} = \mathbf{b}_i$. By the convergence results, we have that $\mathbf{A}\mathbf{x}_i = \mathbf{b}_0$ for $i \to \infty$. When we compare (5.52) again with the original constrained problem (5.49), we note that the primary advantage that we have over the continuation methods is that λ is no longer increasing but *constant*. Instead, the right handed side of the constrained is modified in every iteration and the stability problems are completely avoided.

5.3.1 Splitting the Bregman iteration

In the following, we will consider l_1-regularization from (5.1) with $J(\mathbf{x}) = |\mathbf{S}\mathbf{x}|_1$ and with general data fitting function $H(\mathbf{x}; \mathbf{y})$. Here, $\mathbf{S}$ is a linear "sparsifying" transform (for which multiresolution transforms from Chapter 2 can be used). For this functional, the first part of the Bregman iteration (5.51) is a complex optimization problem. Because we are dealing with images here, the dimension

of $\mathbf{x}$ is very high. To alleviate this problem, Split Bregman [Goldstein and Osher, 2008] converts (5.52) again in a second constrained optimization problem by introducing a *"split"* variable $\mathbf{d}$:

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} |\mathbf{d}|_1 + H(\mathbf{x};\mathbf{y}) + \frac{\lambda}{2}\|\mathbf{S}\mathbf{x} - \mathbf{d}\|_2^2 . \tag{5.54}$$

This problem can be solved elegantly using the Bregman iteration from (5.52)-(5.53), which gives:

$$(\mathbf{x}_{i+1}, \mathbf{d}_{i+1}) = \min_{(\mathbf{x},\mathbf{d})} |\mathbf{d}|_1 + H(\mathbf{x};\mathbf{y}) + \frac{\lambda}{2}\|\mathbf{S}\mathbf{x} - \mathbf{d}_i - \mathbf{b}_i\|_2^2 , \tag{5.55}$$

$$\mathbf{b}_{i+1} = \mathbf{b}_i + (\mathbf{S}\mathbf{x}_{i+1} - \mathbf{d}_{i+1}) . \tag{5.56}$$

Next, the l_1 and l_2-components of (5.55) are "split" by performing the minimization with respect to $\mathbf{x}$ and $\mathbf{d}$ *separately:*

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} H(\mathbf{x};\mathbf{y}) + \frac{\lambda}{2}\|\mathbf{S}\mathbf{x} - \mathbf{d}_i - \mathbf{b}_i\|_2^2 , \tag{5.57}$$

$$\mathbf{d}_{i+1} = \arg\min_{\mathbf{d}} |\mathbf{d}|_1 + \frac{\lambda}{2}\|\mathbf{S}\mathbf{x} - \mathbf{d}_i - \mathbf{b}_i\|_2^2 . \tag{5.58}$$

The minimization problems in (5.57)-(5.58) are now considerably easier: (5.58) has a closed form solution, given by *soft-shrinkage*:

$$[\mathbf{d}_{i+1}]_k = \text{softshrink}\left([\mathbf{S}\mathbf{x}_i]_k + [\mathbf{b}_i]_k , \frac{1}{\lambda}\right) , \tag{5.59}$$

where $[\cdot]_k$ denotes the kth component of a vector. Also, (5.57) can be solved efficiently. For example, let $H(\mathbf{x};\mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, then (5.57) becomes:

$$\mathbf{x}_{i+1} = \left(\mathbf{I} + \lambda\mathbf{S}^H\mathbf{S}\right)^{-1}\left(\mathbf{y} + \mathbf{S}^H(\mathbf{d}_i + \mathbf{b}_i)\right) . \tag{5.60}$$

with $(\cdot)^H$ the Hermitian transpose of a matrix. In some applications, $\mathbf{S}$ is a diagonal matrix, such that the updated image $\mathbf{x}_{i+1}$ can be found using simple point-wise operations. In case the sparsifying transform $\mathbf{S}$ has a block-circulant structure, this system of equations can be solved using the 2D FFT, since a 2D DFT matrix will diagonalize $\mathbf{S}$. If the matrix $\mathbf{I} + \lambda\mathbf{S}^H\mathbf{S}$ is *diagonally dominant*, which means that the magnitude of a diagonal entry is larger than the sum of the magnitudes of all other entries in that row:

$$\left|1 + \lambda\left[\mathbf{S}^H\mathbf{S}\right]_{mm}\right| \gg \sum_{\substack{n=1,\\ n\neq m}}^{N} \left|\lambda\left[\mathbf{S}^H\mathbf{S}\right]_{mn}\right| , \tag{5.61}$$

then the Gauss-Seidel method [Kahan, 1958] can be used. In all other cases it is adviced to use a few steps of the conjugate gradient method to reach an approximate solution of the problem [Goldstein and Osher, 2008].

In theory, (5.57) and (5.58) need to be applied alternately until convergence, before applying (5.56) iteratively. In [Goldstein and Osher, 2008] it was found that for many applications the algorithm still converges when only one iteration of the inner iteration is being performed. The resulting algorithm is summarized in Algorithm 5.1.

The Split-Bregman algorithm is in fact similar to various iterative thresholding schemes (e.g. [Figueiredo and Nowak, 2003,Daubechies et al., 2004,Bioucas-Dias and Figueiredo, 2007, Combettes and Pesquet, 2007]). Split Bregman has several advantages compared to other approaches: 1) a relatively low memory footprint [Goldstein and Osher, 2008], 2) simple and fast iteration steps and 3) the technique is generally easy to implement, even for complex problems.

Solving the regularization problem (5.54) has one subtle issue: it requires an initial choice of the regularization parameter λ. Although there exists extensive literature on selecting the regularization parameter (see e.g. [Titterington, 1991, Kang and Katsaggelos, 1995]), for most applications, we are interested in solving the following constrained problem, which does not contain such a regularization parameter:

$$\min_{\mathbf{x}} J(\mathbf{x}), \quad \text{s.t. } H(\mathbf{x};\mathbf{y}) < \epsilon \tag{5.62}$$

with ϵ an upper bound for data fitting error of the solution (usually a small positive number). Practically, starting from an initial solution with data fitting cost $H(\mathbf{x};\mathbf{y}) \geq \epsilon$, the search for the optimal solution will be stopped whenever a candidate solution is found with $H(\mathbf{x})$ smaller than a predefined threshold ϵ. The idea from [Goldstein and Osher, 2008] is to solve (5.62) again using Bregman iterations. For example, for quadratic data functions of the form $H(\mathbf{x};\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, the problem (5.62) can be reduced to a sequence of unconstrained problems of the form

$$\mathbf{x}_{i+1} = \min_{\mathbf{x}} J(\mathbf{x}) + \mu \|\mathbf{y}_i - \mathbf{A}\mathbf{x}\|_2^2, \tag{5.63}$$

$$\mathbf{y}_{i+1} = \mathbf{y}_i + \mathbf{y} - \mathbf{A}\mathbf{x}_i. \tag{5.64}$$

If we further specify $J(\mathbf{x})$ in a sparsifying transform domain, we can add the second constraint $\mathbf{S}\mathbf{x} = \mathbf{d}$, next to the constraint $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 < \epsilon$. Then the first Bregman step (5.63) can be replaced by the constrained problem (5.54), which gives the Split Bregman algorithm from Algorithm 5.2. Now, as the constraints $\mathbf{S}\mathbf{x} = \mathbf{d}$ and $\mathbf{y}_i = \mathbf{A}\mathbf{x}$ will be perfectly fulfilled after each Bregman iteration, the final result of the algorithm becomes independent of the parameter choice for λ and μ. However, the computation time largely depends on the choice of these parameters (and of course also the parameter ϵ). It is found in [Goldstein and Osher, 2008] that when these parameters are properly chosen, the outer loop of Algorithm 5.2 only needs to be performed a small number of times.

Algorithm 5.1 Split Bregman algorithm minimizing $|\mathbf{d}|_1 + H(\mathbf{x})$ s.t. $\mathbf{Sx} = \mathbf{d}$.

initialize $\mathbf{x}_1 = \mathbf{y}$, $\mathbf{d}_1 = \mathbf{0}$, $\mathbf{b}_1 = \mathbf{0}$

while $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2 >$ tolerance

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} H(\mathbf{x};\mathbf{y}) + \tfrac{\lambda}{2}\|\mathbf{Sx} - \mathbf{d}_i - \mathbf{b}_i\|_2^2, \quad \text{(step I)}$$
$$\mathbf{d}_{i+1} = \arg\min_{\mathbf{d}} |\mathbf{d}|_1 + \tfrac{\lambda}{2}\|\mathbf{Sx} - \mathbf{d}_i - \mathbf{b}_i\|_2^2, \quad \text{(step II)}$$
$$\mathbf{b}_{i+1} = \mathbf{b}_i + (\mathbf{Sx}_{i+1} - \mathbf{d}_{i+1}). \quad \text{(step III)}$$

end

Algorithm 5.2 Split Bregman algorithm minimizing $|\mathbf{d}|_1$ s.t. $\mathbf{Sx} = \mathbf{d}$ and $\|\mathbf{y} - \mathbf{Ax}\|_2^2 < \sigma$.

initialize $\mathbf{x}_1 = \mathbf{y}$, $\mathbf{d}_1 = \mathbf{0}$, $\mathbf{b}_1 = \mathbf{0}$

while $\|\mathbf{y} - \mathbf{Ax}_i\|_2 \geq \sigma$

while $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2 >$ tolerance

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} \mu\|\mathbf{y}_i - \mathbf{Ax}\|_2^2 + \tfrac{\lambda}{2}\|\mathbf{Sx} - \mathbf{d}_i - \mathbf{b}_i\|_2^2, \quad \text{(step I)}$$
$$\mathbf{d}_{i+1} = \arg\min_{\mathbf{d}} |\mathbf{d}|_1 + \tfrac{\lambda}{2}\|\mathbf{Sx} - \mathbf{d}_i - \mathbf{b}_i\|_2^2, \quad \text{(step II)}$$
$$\mathbf{b}_{i+1} = \mathbf{b}_i + (\mathbf{Sx}_{i+1} - \mathbf{d}_{i+1}). \quad \text{(step III)}$$

end

$$\mathbf{y}_{i+1} = \mathbf{y}_i + \mathbf{y} - \mathbf{Ax}_i$$

end

5.3.2 Bayesian MAP estimation through Bregman optimization

So far, we considered the Bregman framework as a generic technique for solving l_1-regularized problems where the regularization takes place in a sparsifying transform domain. There is a lot of flexibility here:

- The function $H(\mathbf{x})$ needs to be convex and differentiable, but other than that there are no restrictions. It turns out that $H(\mathbf{x})$ is naturally linked with the degradation model (or noise model), as we will see further on.
- For the matrix $\mathbf{S}$, any multiresolution transform presented in Chapter 2 can be used. Moreover, $\mathbf{S}$ does not even have to be an *invertible* matrix (or even a *square* matrix): the update step (5.60) depends on the Hermitian transpose $\mathbf{S}^H$, which is the adjoint transform, but not the inverse $\mathbf{S}^{-1}$. This is beneficial when the inverse transform is more time-consuming than

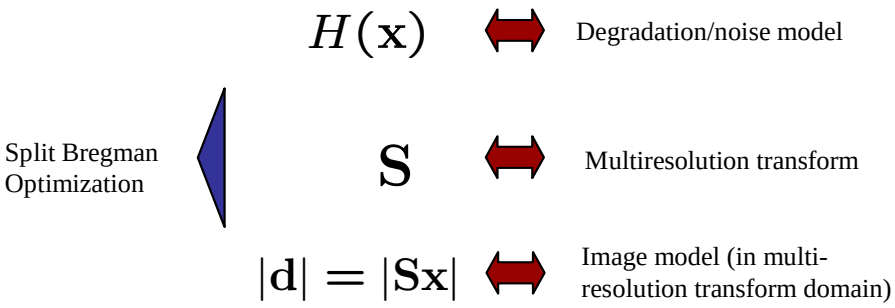


Figure 5.24: Different components of the Bregman optimization framework.

the adjoint transforms (e.g. the Radon transform, non-uniform DFT...).

- Previously we used the l_1-norm $|\mathbf{d}|_1$ in the formulation of (5.54). The framework allows other norms to be used as well, as long as these functionals are *convex*. $|\mathbf{d}|_1$ is naturally linked to the image model, as we will see next.

Bregman optimization can be used to find the maximum of Bayesian posterior distributions. The multivariate extension of the Bayesian MAP estimate (5.9) from Section 5.2.1, is defined as:

$$\begin{aligned} \hat{\mathbf{x}}_{\text{MAP}} &= \arg\max_{\mathbf{x}} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \\ &= \arg\max_{\mathbf{x}} \log f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) + \log f_{\mathbf{x}}(\mathbf{x}) . \end{aligned} \tag{5.65}$$

Our goal is here to estimate a *"clean"* image $\mathbf{x}$ from an observed degraded image $\mathbf{y}$, using prior information with respect to $\mathbf{x}$ in the form of a probability distribution. As we explained in Chapter 3, the prior distribution is usually modeled in the transform domain. First, let $\mathbf{d} = \mathbf{S}\mathbf{x}$ with $\mathbf{S}$ an invertible matrix, then we have:

$$f_{\mathbf{x}}(\mathbf{x})\,\mathrm{d}\mathbf{x} = |\det \mathbf{S}|\, f_{\mathbf{d}}(\mathbf{S}\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{5.66}$$

where we expressed the transform domain PDF $f_{\mathbf{d}}(\mathbf{d})$ in terms of the image domain PDF of $\mathbf{x}$ through a change of variables. The determinant of the Jacobian matrix, $|\det \mathbf{S}|$, will not have any further influence in the optimization problem, since this matrix is constant. In various intra-scale statistical models from Section 3.4, the subband coefficients from different subbands were assumed to be statistically independent. Let $\mathbf{D}_l$ denote a projection matrix that projects the transform coefficients $\mathbf{d}$ onto the lth subband ($l = 1, ..., L$), such that the coefficients for subband l are given by $\mathbf{d}^{(l)} = \mathbf{D}_l\mathbf{S}\mathbf{x}$, then we can write:

$$f_{\mathbf{x}}(\mathbf{x}) \propto \prod_{l=1}^{L} f_{\mathbf{d}^{(l)}}(\mathbf{D}_l\mathbf{S}\mathbf{x}) \tag{5.67}$$

and consequently, the MAP estimate becomes:

$$\hat{\mathbf{x}}_{\text{MAP}} = \arg\max_{\mathbf{x}} -H(\mathbf{x};\mathbf{y}) + \sum_{l=1}^{L} \log f_{\mathbf{d}^{(l)}}\left(\mathbf{D}_l\mathbf{S}\mathbf{x}\right) \tag{5.68}$$

$$= \arg\min_{\mathbf{x}} H(\mathbf{x};\mathbf{y}) - \sum_{l=1}^{L} \log f_{\mathbf{d}^{(l)}}\left(\mathbf{D}_l\mathbf{S}\mathbf{x}\right) \tag{5.69}$$

with $H(\mathbf{x};\mathbf{y}) = -\log f_{\mathbf{y}|\mathbf{x}}\left(\mathbf{y}|\mathbf{x}\right)$ the data fitting function.

Now, the Split Bregman algorithm from Algorithm 5.1 can be used to compute the MAP estimates. This has the advantage that the computation is greatly simplified, independent of the multiresolution transform or image model being used. In particular, for quadratic data fitting functions, the step I of the Bregman iteration is an l_2-regularized problem which is equivalent to the (linear) Tikhonov-Miller restoration problem (see Section 5.1). Step II of the Bregman iteration amounts to *soft-thresholding* if the distributions $f_{\mathbf{d}^{(l)}}\left(\mathbf{D}_l\mathbf{S}\mathbf{x}\right), l = 1, ..., L$ are Laplace distributions. Alternatively we can also consider more heavy tailed distributions such as the Bessel K Form distribution (Section 3.2.5). In this case, step II can be solved using the MAP estimation presented in Section 5.2.2.

Hence, given an (arbitrary) image model, a sparsifying multiresolution transform and a degradation/noise model, the Split Bregman technique offers a generic approach for solving image restoration problems (see Figure 5.24). In principle, any image model from Chapter 3, multiresolution transform from Chapter 2 and noise model from Chapter 4 can be used for this. Even though the global restoration problem can be quite complicated (as we will see further on), the intermediate individual Bregman update steps are usually quite simple and computationally efficient.

5.3.3 Split Bregman based removal of correlated noise

To illustrate the Bregman iteration technique in practical applications, we will work out a simple denoising example before tackling the more difficult problems in the next sections. We start from an observed image corrupted with additive correlated Gaussian noise:

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}\left(\mathbf{x}, \mathbf{C}_w\right). \tag{5.70}$$

The coefficients for the band-pass and high-pass subbands $l = 2, ..., L$ are assumed to be Laplacian distributed with parameter s_l. We further assume that the noise is spatially stationary. This means that $\mathbf{C}_w$ is a Toeplitz matrix. To facilitate implementation in the 2D Fourier domain, we approximate this matrix by a block-circulant matrix. This means that we (artificially) assume that noise on the left image border is correlated with noise on the right image border. This simplification has only effect near the image border.

Our goal is to estimate the original noise-free image $\mathbf{x}$. The data fitting term for (5.70) is given by:

$$H(\mathbf{x};\mathbf{y}) = -\log f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{y})^T \mathbf{C}_w^{-1}(\mathbf{x}-\mathbf{y}).$$

For Split Bregman, the unconstrained MAP optimization problem can be formulated as follows:

$$\hat{\mathbf{x}}_{\text{MAP}} = \arg\min_{\mathbf{x}} H(\mathbf{x};\mathbf{y}) + \sum_{l=1}^{L} \left| \frac{\mathbf{d}^{(l)}}{s_l} \right|_1 + \frac{\lambda}{2} \sum_{l=1}^{L} \left\| \mathbf{D}_l \mathbf{S}\mathbf{x} - \mathbf{d}^{(l)} \right\|_2^2, \tag{5.71}$$

where the first term is the data fitting term, the second term imposes prior knowledge to the transform coefficients $\mathbf{d}^{(l)}$ and the third term expresses that $\mathbf{d}^{(l)}, j = 1, ..., L$ are transform coefficients for subband l of the original noise-free image $\mathbf{x}$ (recall that $\mathbf{S}$ is a sparsifying transform matrix and that $\mathbf{D}_l$ projects the transform coefficients onto the lth subband). The optimization problem can be efficiently solved using (5.55)-(5.58), which leads to:

$$\mathbf{x}_{i+1} = \left(\mathbf{C}_w^{-1} + \lambda \sum_{l=2}^{L} \mathbf{D}_l^T \mathbf{S}^H \mathbf{S} \mathbf{D}_l \right)^{-1} \left(\mathbf{C}_w^{-1}\mathbf{y} + \lambda \sum_{l=2}^{L} \mathbf{D}_l^T \mathbf{S}^H \left(\mathbf{d}_i^{(l)} + \mathbf{b}_i^{(l)} \right) \right) \tag{5.72}$$

$$\left[\mathbf{d}_{i+1}^{(l)} \right]_k = \text{softshrink}\left([\mathbf{D}_l \mathbf{S}\mathbf{x}]_k + \left[\mathbf{b}_i^{(l)} \right]_k, \frac{1}{\lambda s_l} \right), \quad j = 1, ..., L \tag{5.73}$$

$$\mathbf{b}_{i+1}^{(l)} = \mathbf{b}_i^{(l)} + \left(\mathbf{D}_l \mathbf{S}\mathbf{x}_{i+1} - \mathbf{d}_{i+1}^{(l)} \right), \quad j = 1, ..., L. \tag{5.74}$$

with $\mathbf{b}_i^{(l)}, j = 1, ..., L$ a sequence of Bregman splitting variables. The resulting algorithm is in fact an iterative thresholding algorithm, in which the coefficients of every subband are thresholded according to the given threshold for that subband, i.e.: $1/s_l$. For a self-inverting transform with perfect reconstruction (i.e. $\sum_{l=1}^{L} \mathbf{D}_l^T \mathbf{S}^T \mathbf{S} \mathbf{D}_l = \mathbf{I}$) and for white noise (i.e. $\mathbf{C}_w = \sigma^2 \mathbf{I}$) the algorithm can be further simplified to the algorithm presented in Algorithm 5.3. Bregman step I (5.72) then corresponds to solving a diagonal system of linear equations, which is computationally very efficient.

We remark that for orthonormal transforms $\mathbf{S}$ (i.e. $\mathbf{S}^T\mathbf{S} = \mathbf{S}\mathbf{S}^T = \mathbf{I}$), the whole optimization problem can be expressed in transform domain, as

$$\begin{aligned} \|\mathbf{S}\mathbf{x} - \mathbf{S}\mathbf{y}\|_2^2 &= (\mathbf{S}\mathbf{x} - \mathbf{S}\mathbf{y})^T (\mathbf{S}\mathbf{x} - \mathbf{S}\mathbf{y}) \\ &= (\mathbf{x}-\mathbf{y})^T \mathbf{S}^T \mathbf{S} (\mathbf{x}-\mathbf{y}) \\ &= \|\mathbf{x}-\mathbf{y}\|_2^2. \end{aligned}$$

It can easily be checked that in this case, applying solely one Bregman iteration yields the optimal MAP solution. For non-orthonormal or redundant multiresolution transforms, multiple iteration will be needed for convergence and the

Algorithm 5.3 Split Bregman algorithm for multiresolution transform domain denoising for *white* noise (for *correlated* noise, simply use equation (5.72) instead of the first equation).

initialize $\mathbf{x}_1 = \mathbf{y}$, $\mathbf{d}_1 = \mathbf{0}$, $\mathbf{b}_1 = \mathbf{0}$

while $H(\mathbf{x};\mathbf{y}) \geq \epsilon$

  while $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2 >$ tolerance

$$\mathbf{x}_{i+1} = \frac{1}{1+\lambda\sigma^2}\left(\mathbf{y} + \lambda\sigma^2 \sum_{l=2}^{L} \mathbf{D}_l^T \mathbf{S}^H \left(\mathbf{d}_i^{(l)} + \mathbf{b}_i^{(l)}\right)\right), \quad \text{(step I)}$$

$$\left[\mathbf{d}_{i+1}^{(l)}\right]_k = \text{softshrink}\left(\left[\mathbf{D}_l \mathbf{S}\mathbf{x}\right]_k + \left[\mathbf{b}_i^{(l)}\right]_k, \tfrac{1}{\lambda s_l}\right), \quad j = 1, ..., L, \quad \text{(step II)}$$

$$\mathbf{b}_{i+1}^{(l)} = \mathbf{b}_i^{(l)} + \left(\mathbf{D}_l \mathbf{S}\mathbf{x}_{i+1} - \mathbf{d}_{i+1}^{(l)}\right), \quad j = 1, ..., L. \quad \text{(step III)}$$

  end

  $\mathbf{y}_{i+1} = \mathbf{y}_i + \mathbf{y} - \mathbf{x}_i$

end

solution will be (slightly) different than for the equivalent transform domain MAP estimation. The main difference then is that the data data fitting is performed directly in the *image domain* instead of in transform domain. This permits to easily include a wide variety of degradations (that are defined in the image domain) into the model, as we will explain further.

5.3.4 Multiresolution joint denoising and deblurring

As a second example of Split-Bregman, we will now extend the degradation model by including blur, which will lead to a novel restoration algorithm. A classical restoration problem is the recovery of an original image after blurring it and adding Gaussian noise:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w} \quad \text{with } \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_w) \tag{5.75}$$

with $\mathbf{w}$ additive white Gaussian noise. The matrix $\mathbf{A}$ is a generally non-invertible matrix that models the blurring operation. For simplicity, we will assume that $\mathbf{A}$ is *block-circulant* (such that the blurring operation is expressed as a circular convolution, which again facilitates implementation in the 2D Fourier domain) and *known in advance*. The observations are conditionally Gaussian distributed $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{C}_w)$, such that the data fitting term is given by:

$$H(\mathbf{x};\mathbf{y}) = \frac{1}{2}(\mathbf{A}\mathbf{x} - \mathbf{y})^T \mathbf{C}_w^{-1}(\mathbf{A}\mathbf{x} - \mathbf{y}). \tag{5.76}$$

Solving the joint deblurring and denoising problem only requires modifying step I of the Bregman iteration (5.72), as now the matrix $\mathbf{A}$ is involved in the optimization. Because $\mathbf{A}$ is block-circulant, $\mathbf{A}^T\mathbf{A}$ will also be block-circulant, hence computing Bregman step I is best performed using the 2D DFT. The resulting algorithm is shown in Algorithm 5.4.

Compared to Section 5.3.3, the main modification we made is a change of the *degradation* model which resulted in a different formula of step I of the Bregman iteration, leaving step II unaffected.

To illustrate that Split-Bregman allows for a lot of flexibility, we will now consider the case that the blur operation $\mathbf{A}$ and/or the noise covariance matrix $\mathbf{C}_w^{-1}$ are *unknown*. Let us denote by $X(m,n)$ and $Y(m,n)$ the DFT of respectively $\mathbf{x}$ and $\mathbf{y}$, then the data fitting function (5.76) can be written in the DFT domain:

$$H(\mathbf{x};\mathbf{y}) = \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|A(m,n)X(m,n) - Y(m,n)\right|^2 \left(B(m,n)\right)^2 \tag{5.77}$$

with M and N respectively the number of rows and columns in the image, with $A(m,n)$ the frequency response of the unknown blurring filter and with $B(m,n)$ the square root of the reciprocal of the noise PSD. We will now consider the estimation of $B(m,n)$ from a degraded image (the estimation of $A(m,n)$ is entirely analogous). Because for some (m,n), the data fitting cost $\left|A(m,n)X(m,n) - Y(m,n)\right|^2$ can become 0, this is again an ill-posed problem. Therefore we add an extra regularization term to the problem:

$$\begin{aligned}\min_{\mathbf{x},\{B(m,n)\}} \quad & J(\mathbf{x}) + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|A(m,n)X(m,n) - Y(m,n)\right|^2 \left(B(m,n)\right)^2 \\ & + \frac{\mu}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} G(m,n)\left(B(m,n) - B_0\right)^2 \end{aligned} \tag{5.78}$$

with B_0 a constant parameter, $G(m,n) > 0$ a frequency dependent weight and with μ a regularization constant. $G(m,n)$ can be chosen to penalize certain spatial frequencies of the noise PSD. Selecting $G(m,n) = 1$ would enforce the noise PSD to be flat, i.e. $B(m,n) = B_0$. In the following, we will assume that the constant B_0 (which is proportional to the noise standard deviation) is known in advance, such that we can concentrate on estimating the noise correlation structure instead of the noise variance.[9] In many applications the noise PSD has a low-pass or high-pass characteristic rather than a flat characteristic (see Chapter 4). To incorporate this form of prior knowledge in the noise estimation method, we can for example select the frequency response of a high-pass filter, such as the spatial gradient:

$$G(m,n) = \max\left(\epsilon_0, \sqrt{m^2+n^2}\right) \tag{5.79}$$

[9] If necessary, B_0 can be estimated using separate techniques such as the techniques presented in Chapter 4.

with ϵ_0 a small positive constant. Next, $B(m,n)$ can be estimated as the minimum of (5.78), which gives:

$$\widehat{B(m,n)} = \frac{\mu G(m,n)}{|A(m,n)X(m,n) - Y(m,n)|^2 + \mu G(m,n)}, \tag{5.80}$$

which solely requires a point-wise division in the DFT domain. A blind denoising technique, assuming that $A(m,n)$ is known in advance, can readily be obtained as follows:

1. Start from an initial estimate of the noise PSD, e.g. $B(m,n) = 1$, which corresponds to white noise.
2. Solve $\min_{\mathbf{x}} J(\mathbf{x}) + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} |A(m,n)X(m,n) - Y(m,n)|^2 \left(B(m,n)\right)^2$, for which Algorithm 5.4 can be used.
3. Update the (reciprocal of the) noise PSD estimate using (5.80).
4. Repeat step 2 until convergence (e.g. $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2 \leq$ tolerance, with i the iteration index).

Recall that step 2 can be implemented using the Split Bregman algorithm from Algorithm 5.4. Because this algorithm already contains two loops, there seems to be additional computational overhead due to the necessity of a *third* loop. Fortunately, in practice this problem is not very significant, because the Bregman iterative optimization can start from the solution $\mathbf{x}_i$ from the previous outer iteration (hence after every outer iteration, fewer inner iterations will be necessary). Furthermore, the Bregman iterations are relatively simple and can be efficiently implemented.

5.3.5 Joint signal-dependent noise and bias removal

The removal of signal-dependent noise is a nontrivial image processing problem that is very significant for many applications. However, compared to the design of algorithms for signal-independent (additive) noise, only limited effort has been spent to tackle the problem of signal-dependent noise. As we already mentioned, most techniques rely on *variance stabilization* (see Section 4.4.1). Better is to approximate the signal-dependent noise into mixed additive and multiplicative noise, as in [Hirakawa and Parks, 2005b]. The noise level function (NLF) is then defined as:

$$\sigma(x) \approx \sigma(x_0) + \left.\frac{\partial \sigma}{\partial x}\right|_{x=x_0} x. \tag{5.81}$$

Equation (5.81) comprises a first order Taylor approximation around the working point x_0, where the working point is often chosen to be a dark intensity, favoring the dark regions over the brighter regions. Another way to deal with

Algorithm 5.4 Split Bregman algorithm for multiresolution transform domain deblurring and denoising.

initialize $\mathbf{x}_1 = \mathbf{y}$, $\mathbf{d}_1 = \mathbf{0}$, $\mathbf{b}_1 = \mathbf{0}$

while $H(\mathbf{x};\mathbf{y}) \geq \epsilon$

while $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2 >$ tolerance

$$\mathbf{x}_{i+1} = \left(\mathbf{A}^T\mathbf{C}_w^{-1}\mathbf{A}+\lambda\right)^{-1}\left(\mathbf{C}_w^{-1}\mathbf{A}^T\mathbf{y}+\lambda\sum_{l=2}^{L}\mathbf{D}_l^T\mathbf{S}^H\left(\mathbf{d}_i^{(l)}+\mathbf{b}_i^{(l)}\right)\right), \quad \text{(step I)}$$

$$\left[\mathbf{d}_{i+1}^{(l)}\right]_k = \text{softshrink}\left(\left[\mathbf{D}_l\mathbf{S}\mathbf{x}\right]_k + \left[\mathbf{b}_i^{(l)}\right]_k, \tfrac{1}{\lambda s_l}\right), \quad j = 1, ..., L, \quad \text{(step II)}$$

$$\mathbf{b}_{i+1}^{(l)} = \mathbf{b}_i^{(l)} + \left(\mathbf{D}_l\mathbf{S}\mathbf{x}_{i+1} - \mathbf{d}_{i+1}^{(l)}\right), \quad j = 1, ..., L. \quad \text{(step III)}$$

end

$\mathbf{y}_{i+1} = \mathbf{y}_i + \mathbf{y} - \mathbf{A}\mathbf{x}_i$

end

signal-dependent noise is to estimate the noise variance locally at position j as in e.g. [Johnstone and Silverman, 1997, Argenti et al., 2002, Goossens et al., 2006] and to denoise the images using an algorithm for *non-stationary* noise. The disadvantage here is that the noise variance estimate is often not very reliable, especially the presence of image discontinuities such as edges, textures... Other approaches directly deal with signal-dependent noise (in particular, Poisson noise) in the wavelet domain [Hirakawa, 2007, Hirakawa and Wolfe, 2009], unfortunately this turns out to be a very complicated task. In [Foi et al., 2008, Foi, 2008], the NLF is first estimated in the wavelet domain, by maximum likelihood fitting of a parametric model for the NLF. Next, the estimated NLF is used for denoising in a second processing step.

In this section, we will present a denoising technique for the signal-dependent noise models defined in Section 4.4 within the Bregman optimization framework. Again, Bregman optimization has the major advantage in this application that the noise model and image prior models are decoupled and can be defined in different transform domains. Compared to the method of [Foi, 2008], our approach estimates the noise variance of each pixel in the image jointly with the "denoised" image.

Let $x_j = [\mathbf{x}]_j$, $w_j = [\mathbf{w}]_j$ and $y_j = [\mathbf{y}]_j$ denote the jth component of respectively $\mathbf{x}$, $\mathbf{w}$ and $\mathbf{y}$. For the signal-dependent noise models from Section 4.4, the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is Gaussian:

$$y_j|x_j \sim \mathcal{N}\left(\mu(x_j), \sigma^2(x_j)\right), \tag{5.82}$$

where the bias function $\mu(x)$ and the noise level function $\sigma(x)$ are obtained

using the techniques discussed in Section 4.4. Our goal is then to estimate an (approximately) "unbiased" version of the original image x_j. Based on (5.82), the data fitting term readily follows:

$$H(\mathbf{x};\mathbf{y}) = \frac{1}{2}\sum_{j=1}^{N}\left(\frac{y_j - \mu(x_j)}{\sigma^2(x_j)}\right)^2,$$

with N the number of pixels in the image. Step I of the Bregman iteration becomes:

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{x}} \frac{1}{2}\sum_{j=1}^{N}\left(\frac{y_j - \mu(x_j)}{\sigma^2(x_j)}\right)^2 + \frac{\lambda}{2}\|\mathbf{S}\mathbf{x} - \mathbf{d}_i - \mathbf{b}_i\|_2^2. \tag{5.83}$$

Unfortunately, because $\mu(x)$ and $\sigma(x)$ are nonlinear functions in general, (5.83) does no longer correspond to a quadratic problem. One way to proceed is to use nonlinear optimization techniques, such as nonlinear conjugate gradients, however, this may dramatically increase the computation time. An alternative is to estimate a "nonlinearized" version of the signal directly, i.e. to estimate $x' = \mu(x)$ instead of x. Therefore we assume that $\mu(x)$ is monotonically increasing on the range space of x (which is the case in all examples we encountered in our work), such that the inverse $\mu^{-1}(x)$ exists. Additionally, we will impose the prior distribution on the signal in this "nonlinearized" domain as well. Applying the variable substitution $x' = \mu(x)$ to (5.83) then yields:

$$\mathbf{x}'_{i+1} = \arg\min_{\mathbf{x}} \frac{1}{2}\sum_{j=1}^{N}\left(\frac{y_j - x'_j}{\sigma'^2(x'_j)}\right)^2 + \frac{\lambda}{2}\left\|\mathbf{S}\mathbf{x}' - \mathbf{d}_i - \mathbf{b}_i\right\|_2^2, \tag{5.84}$$

where $\mathbf{x}' = \left[x'_1 \cdots x'_N\right]$ and with the NLF warped to the nonlinear domain of x' using the de-biasing function $\mu^{-1}(x)$:

$$\sigma'^2\left(x'_j\right) = \sigma^2\left(\mu^{-1}\left(x'_j\right)\right).$$

Solving (5.84) is then straightforward:

$$\mathbf{x}'_{i+1} = \left(1 + \lambda\boldsymbol{\sigma}'^2(\mathbf{x})\right)^{-1}\left(\mathbf{y} + \lambda\boldsymbol{\sigma}'^2\left(\mathbf{x}'\right)\mathbf{S}\left(\mathbf{d}_i - \mathbf{b}_i\right)\right) \tag{5.85}$$

with

$$\boldsymbol{\sigma}'^2\left(\mathbf{x}'\right) = \begin{pmatrix} \sigma'^2\left(x'_1\right) & & & \\ & \sigma'^2\left(x'_2\right) & & \\ & & \ddots & \\ & & & \sigma'^2\left(x'_N\right) \end{pmatrix}. \tag{5.86}$$

Algorithm 5.5 Split Bregman algorithm for multiresolution transform domain removal of signal-dependent noise and debiasing.

initialize $\mathbf{x}_1^{'} = \mathbf{y}$, $\mathbf{d}_1 = \mathbf{0}$, $\mathbf{b}_1 = \mathbf{0}$

while $\left\|\mathbf{x}_{i+1}^{'} - \mathbf{x}_i^{'}\right\|_2 >$ tolerance

$$\mathbf{x}_{i+1}^{'} = \left(1 + \lambda \boldsymbol{\sigma}^{'2}\left(\mathbf{x}_i^{'}\right)\right)^{-1}\left(\mathbf{y} + \lambda \boldsymbol{\sigma}^{'2}\left(\mathbf{x}_i^{'}\right)\mathbf{S}\left(\mathbf{d}_i - \mathbf{b}_i\right)\right), \quad \text{(step I)}$$

$$\left[\mathbf{d}_{i+1}^{(l)}\right]_k = \text{softshrink}\left(\left[\mathbf{D}_l \mathbf{S}\mathbf{x}_i^{'}\right]_k + \left[\mathbf{b}_i^{(l)}\right]_k, \tfrac{1}{\lambda s_l}\right), \quad j = 1, ..., L, \quad \text{(step II)}$$

$$\mathbf{b}_{i+1}^{(l)} = \mathbf{b}_i^{(l)} + \left(\mathbf{D}_l \mathbf{S}\mathbf{x}_{i+1}^{'} - \mathbf{d}_{i+1}^{(l)}\right), \quad j = 1, ..., L. \quad \text{(step III)}$$

end

$$\mathbf{x}_{i+1} = \left[\ \mu^{-1}\left([\mathbf{x}_i]_1\right) \quad \cdots \quad \mu^{-1}\left([\mathbf{x}_i]_N\right)\ \right]$$

Equation (5.85) then constitutes a linear diagonal system of equations which is trivial to solve. Finally, an estimate of the unbiased image is obtained as:

$$\hat{x}_j = \mu^{-1}(\hat{x}_j^{'}) \tag{5.87}$$

An example of such an de-biasing function $\mu^{-1}(\cdot)$ is depicted in Figure 5.25. Here the de-biasing function expands the intensity range of the denoised image from $[0, 255]$ to $[-64, 320]$. This approach is elegant and effective in dealing with signal-dependent noise, but the main drawback is that there is a potential risk that estimation errors in the low and high intensity ranges are being amplified by the de-biasing function. This is particularly the case for low SNR. Nevertheless, combined with an effective image prior, this effect is not very pronounced as we will see in Section 5.4. The complete algorithm is summarized in Algorithm 5.5.

Finally, we remark that it is also possible to estimate $\sigma^{'2}\left(x_j^{'}\right)$ from the observed image, in a similar manner as we did for the noise PSD in Section 5.3.4, yielding a generic denoising algorithm for signal-dependent noise. Unfortunately, it is not evident to find the de-biasing function $\mu^{-1}(\cdot)$ simultaneously. This will be explored in our future work.

5.4 Experimental results

In this section, we will present experimental results for the above image restoration methods. Because of the large number of restoration methods in this chapter, we will give extended quantitative results solely for the image denoising methods, in order to make space for the more interesting visual results and also

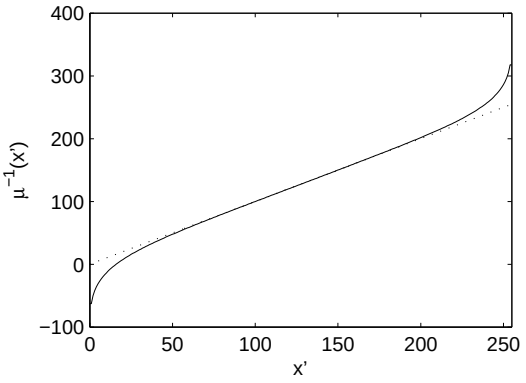


Figure 5.25: Example of a de-biasing function that expands the intensity range from $[0, 255]$ to $[-64, 320]$.

because the denoising technique are the underlying foundation of the more sophisticated restoration methods. As we discussed generic techniques for image restoration in this chapter, the number of possible combinations (signal model, noise model, multiresolution transform) is very high. Therefore, we will restrict ourselves to a few interesting practical restoration applications.

5.4.1 Denoising results for white noise

First, we investigate the influence of different interscale and intrascale image models on the denoising performance. Therefore, we compare the following methods on the test set of 8 images shown in Figure 5.26:

- The method from [Crouse et al., 1998], which uses a HMT with a mixture of Gaussians prior model in the decimated wavelet domain.
- *ProbShrink* from [Pižurica and Philips, 2006], which makes use of an undecimated wavelet transform combined with an intrascale model based on local spatial activity indicators (see Section 3.4.2).
- *BLS-GSM* from [Portilla et al., 2003], which is implemented using the full STP transform with 8 orientations. The local neighborhood used consists of a 3×3 local neighborhood and 1 parent coefficient.
- The method from [Romberg et al., 2001a], which is an extension of [Crouse et al., 1998] to the DT-CWT.
- *Bivariate Shrinkage* from [Şendur and Selesnick, 2002a], in which the DT-CWT parent-child dependencies are modeled using a bivariate distribution (see Section 3.5.2).
- *Vector-ProbShrink*, presented in Section 5.2.3.

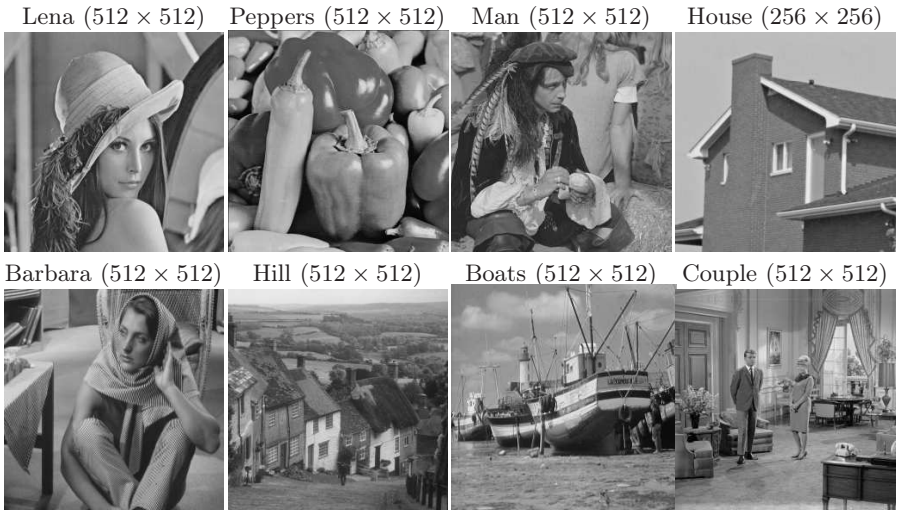


Figure 5.26: Test images from the USC-SIPI database (`http://sipi.usc.edu/database/`), used for the validation of the denoising algorithms.

For all of the DT-CWT based methods we use Farras nearly symmetric filters [Abdelnour and Selesnick, 2001] for the first scale and 10-tap Q-shift filters [Kingsbury, 2003] starting from the second scale. We use overlapping 3×3 local neighborhoods to keep the computional overhead low. The noise variance is assumed to be known to all the algorithms. Quantitative PSNR results for different test images and different input noise levels are reported in Figure 5.27. To increase the reliability of the results, the PSNR results are averaged over 50 runs for each of the denoising methods.

Averaged over all images of the test set, our combined inter/intrascale method performs equally well as the BLS-GSM method of Portilla (see Figure 5.27(f)), but at a lower computational cost, since the redundancy factor of the DT-CWT is 4, while the full STP transform with 8 orientations has redundancy factor $56/3 \approx 18.67$. A comparison for the computation times of Vector-ProbShrink and BLS-GSM is given in Table 5.3. By incorporating interscale dependencies using a HMT model in the DT-CWT domain, Vector-ProbShrink is approximately 3-4 times faster than BLS-GSM, while maintaining the same PSNR performance.

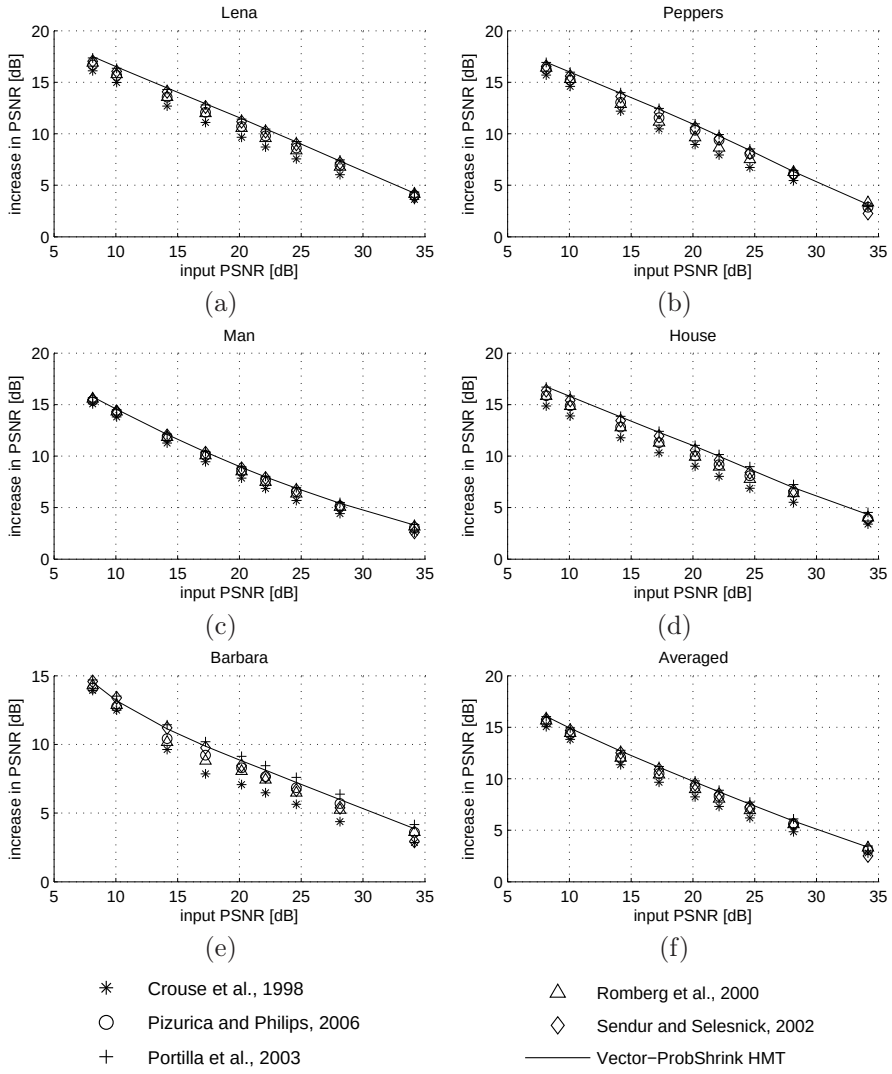


Figure 5.27: PSNR denoising results for different wavelet-based denoising techniques. Reported is the increase in PSNR (i.e. the output PSNR after denoising *minus* the input PSNR).

Table 5.3: Comparison of the execution times of the BLS-GSM method and the proposed method. To allow for a fair comparison, both methods are implemented in C++ with the same level of optimization. Reported values are the execution times averaged over 10 runs and their standard deviations (between parentheses).

	Input image size			
Method	256×256		512×512	
BLS-GSM / Full STP	6.41s	(0.03s)	25.89s	(0.05s)
Vector-ProbShrink / DT-CWT	2.02s	(0.01s)	6.99s	(0.07s)

In Figure 5.28, a second set of results is generated for the same test set of images, but now mainly for techniques that exploit *non-local* dependencies in images. Some of the new techniques that are included in these results are:

- *K-SVD* Denoising [Elad and Aharon, 2006b], which makes use of a locally trained dictionary of image patches.
- *BLS-SVGSM* [Guerrero-Colón et al., 2008a], which uses the SVGSM prior model combined with an MMSE estimator (implemented with block sizes of 32×32).
- *BM-3D* [Dabov et al., 2007], which first performs block-matching to *group* similar patches. Then, 3D transform domain filtering is applied to the resulting stacks. Finally, the denoised image is synthesized by aggregating the 3D filtered patches.
- *BLS-MPGSM* (Section 5.2.4), using a 5×5 local neighborhood.
- The improved *NLMeans filter* from Section 5.1.2.

It can be seen that both *BLS-MPGSM* and the improved *NLMeans filter* are competitive to BM-3D, but slightly under-performing in general. For BLS-MPGSM, we believe the main reason is that non-local information is only exploited partially (i.e. in the EM training phase and not the denoising phase). On the other hand, the *NLMeans* filter tends to destroy *"weak"* edges, which degrades the PSNR performance. However, we found that in terms of the visual quality of the images, our methods are sometimes *better*, because BM-3D often creates paint brush artifacts due to incorrect block-matching caused by the presence of noise.

Visual examples for the different denoising techniques are shown in Figure 5.29, for the *Peppers* image with AWGN ($\sigma = 80$). BLS-MPGSM is omitted in this comparison, as for a high noise level, this method only performs *marginally* better than BLS-GSM. Even for this high noise level, the proposed methods are able to reconstruct well most of the details of the image, well being competitive to BM-3D in terms of PSNR.

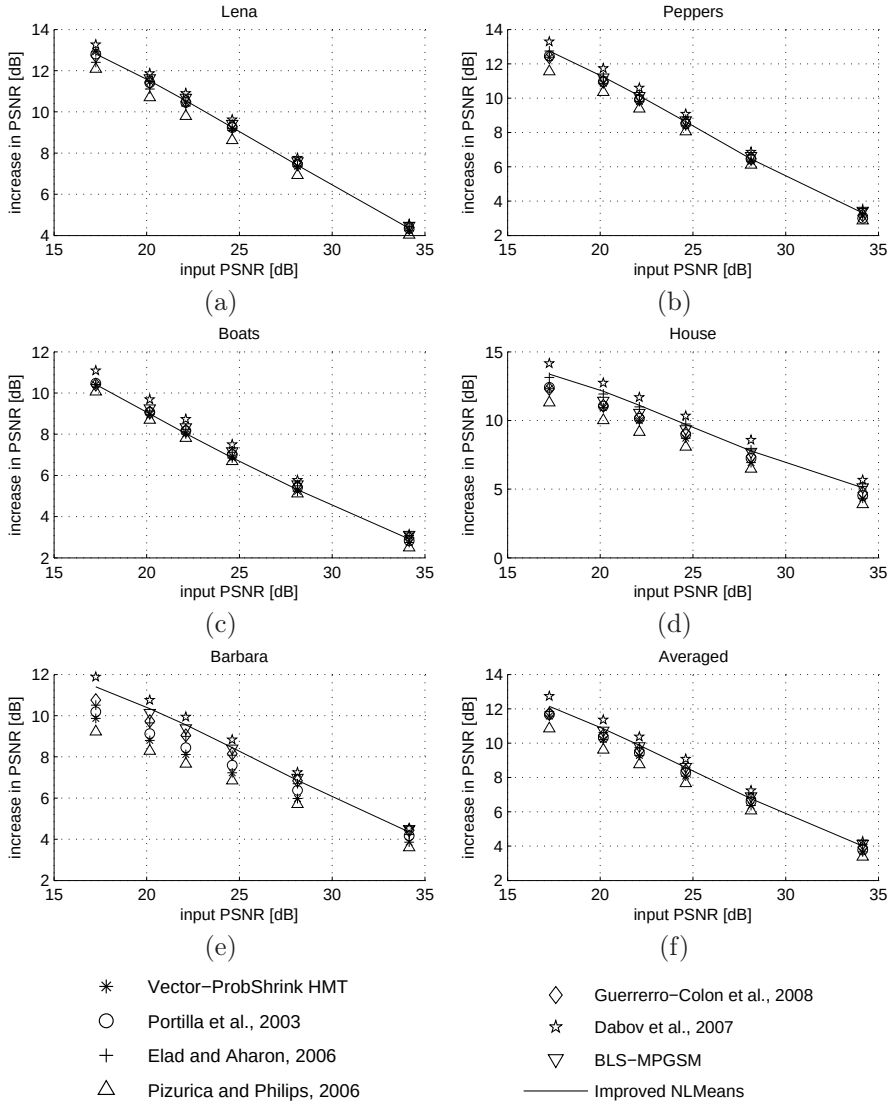
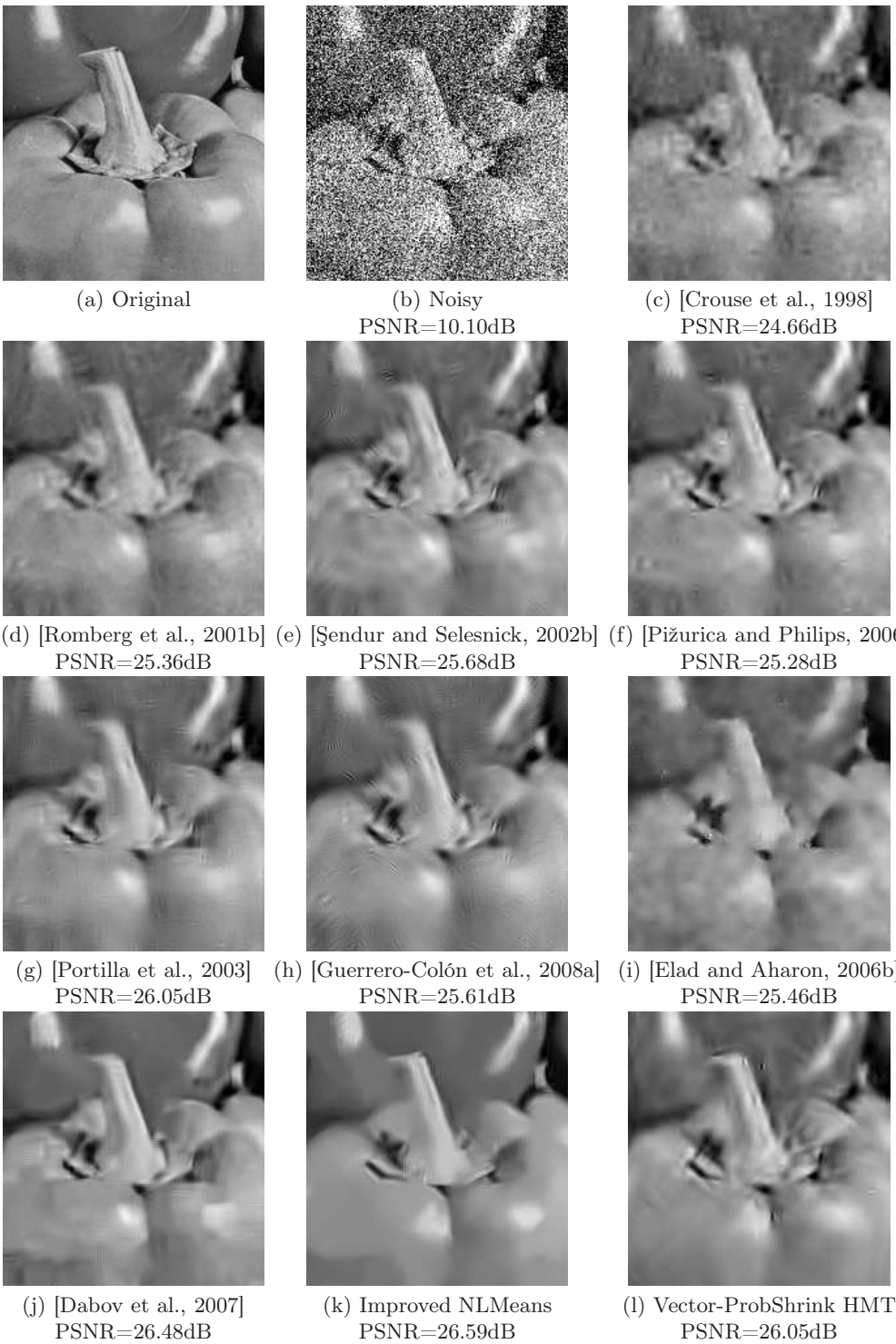


Figure 5.28: PSNR denoising results for different wavelet-based and non-wavelet-based denoising techniques. Reported is the increase in PSNR (i.e. the output PSNR after denoising *minus* the input PSNR).



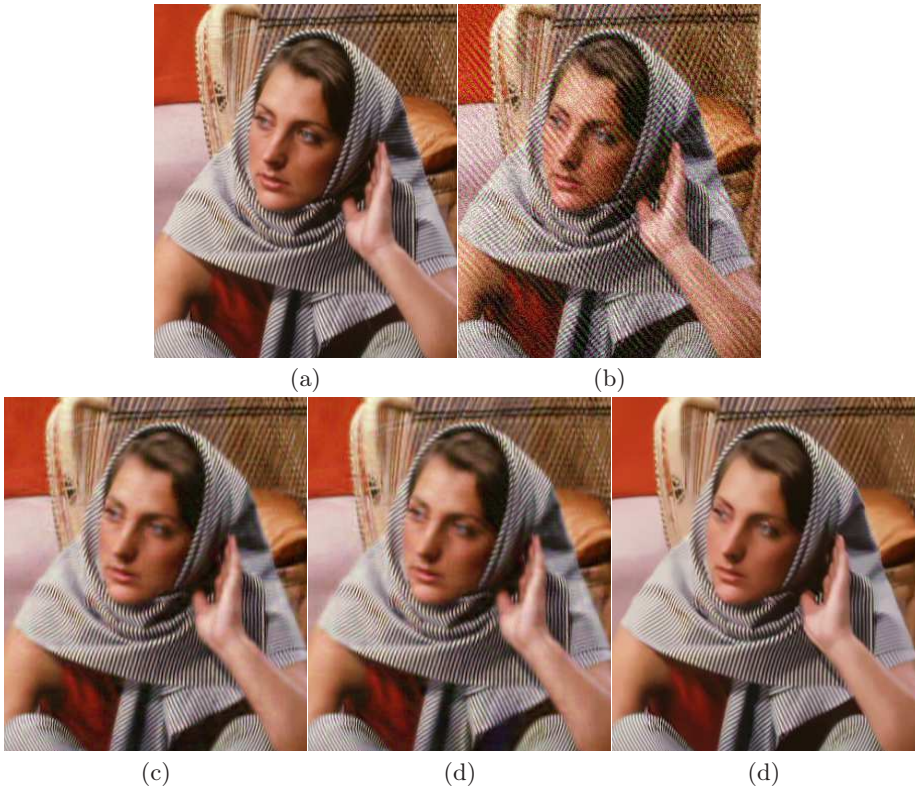
(a) Original
(b) Noisy PSNR=10.10dB
(c) [Crouse et al., 1998] PSNR=24.66dB
(d) [Romberg et al., 2001b] PSNR=25.36dB
(e) [Şendur and Selesnick, 2002b] PSNR=25.68dB
(f) [Pižurica and Philips, 2006] PSNR=25.28dB
(g) [Portilla et al., 2003] PSNR=26.05dB
(h) [Guerrero-Colón et al., 2008a] PSNR=25.61dB
(i) [Elad and Aharon, 2006b] PSNR=25.46dB
(j) [Dabov et al., 2007] PSNR=26.48dB
(k) Improved NLMeans PSNR=26.59dB
(l) Vector-ProbShrink HMT PSNR=26.05dB

Figure 5.29: Visual denoising results for different wavelet-based and non-wavelet-based denoising techniques.

5.4.2 Denoising results for colored noise

In Figure 5.30 and Figure 5.31, visual results are given for color images corrupted with artificial colored noise. The noise was added independently to the three RGB-color channels. To allow an easy comparison of the denoising methods, no noise correlations between different color channels are being considered in these results. The Vector-ProbShrink and BLS-MPGSM algorithms are applied in the RGB-color space to each colour channel individually. For the improved NLMeans filter, the block similarity is computed as a mean square difference in RGB-space. Although for all three methods, the PSNR increases by more than 8dB after processing, the visual performance performance of the methods is quite different. For example, BLS-MPGSM generally tends to generate more *ringing* artifacts than the other methods. We found that this behavior can be improved by using 3×3 neighborhoods instead of 5×5 neighborhoods. However doing so, the PSNR performance may decrease because we may be ignoring some significant correlations between coefficients. On the other hand, the *NLMeans* filter overall produces *sharper* images. However, in flat regions (e.g. the skin) the *NLMeans* filter causes oversmoothing. Furthermore, this filter fails to reconstruct the details around the eye of the parrot in Figure 5.31. We believe this is caused by the high noise level and by the lack of candidate similar patches for these image details.

Hence we can conclude that the presented image models and denoising techniques already give a high PSNR and visual quality, in spite of the high input noise levels used in these experiments. Apparently each of the methods have their own advantages or disadvantages, which may suggest that there is even more room for improvement.



(a) (b)

(c) (d) (d)

Figure 5.30: Visual results for the removal of stationary correlated noise from images. (a) Original image, (b) Noisy image (PSNR=20.60dB), (c) Vector-ProbShrink (PSNR=29.44dB), (d) BLS-MPGSM (PSNR=31.01dB), (e) Improved NLMeans (PSNR=30.53dB).

(a) (b)

(c) (d) (d)

Figure 5.31: Visual results for the removal of stationary correlated noise from images. (a) Original image, (b) Noisy image (PSNR=14.55dB), (c) Vector-ProbShrink (PSNR=27.22dB), (d) BLS-MPGSM (PSNR=26.90dB), (e) Improved NLMeans (PSNR=25.68dB).

5.4.3 Demosaicing

In Figure 5.32, demosaicing results are shown for both the cycle-spinning method and the proposed method. Here, the directional selectivity of the complex wavelets, combined with the suppression of discoloration artifacts vastly improves the visual quality of the reconstructed images. For the result in Figure 5.33 we reach similar conclusions. The difference images in Figure 5.33 reveal that the reconstruction errors are much smaller than DL-MMSE, which is one of the state-of-the-art demosaicing methods.

The proposed demosaicing method is quite fast: demosaicing one 512×512 image on a recent PC takes 1.5 s in an unoptimized Matlab implementation, while DL-MMSE requires 24.0 s for the same task. This improvement in computation time is achieved here by 1) the simplicity of the reconstruction formulas and 2) by the excellent space-frequency localizing properties of the complex wavelets.

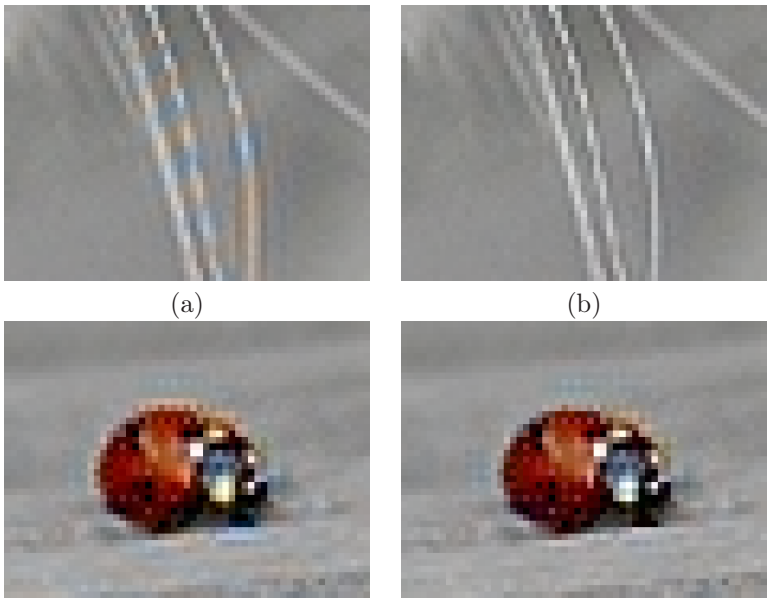


Figure 5.32: Demosaicing results (compare to Figure 5.21) (a),(c) wavelet-based demosaicing with cycle-spinning, (b),(d) our complex-wavelet based demosaicing method.

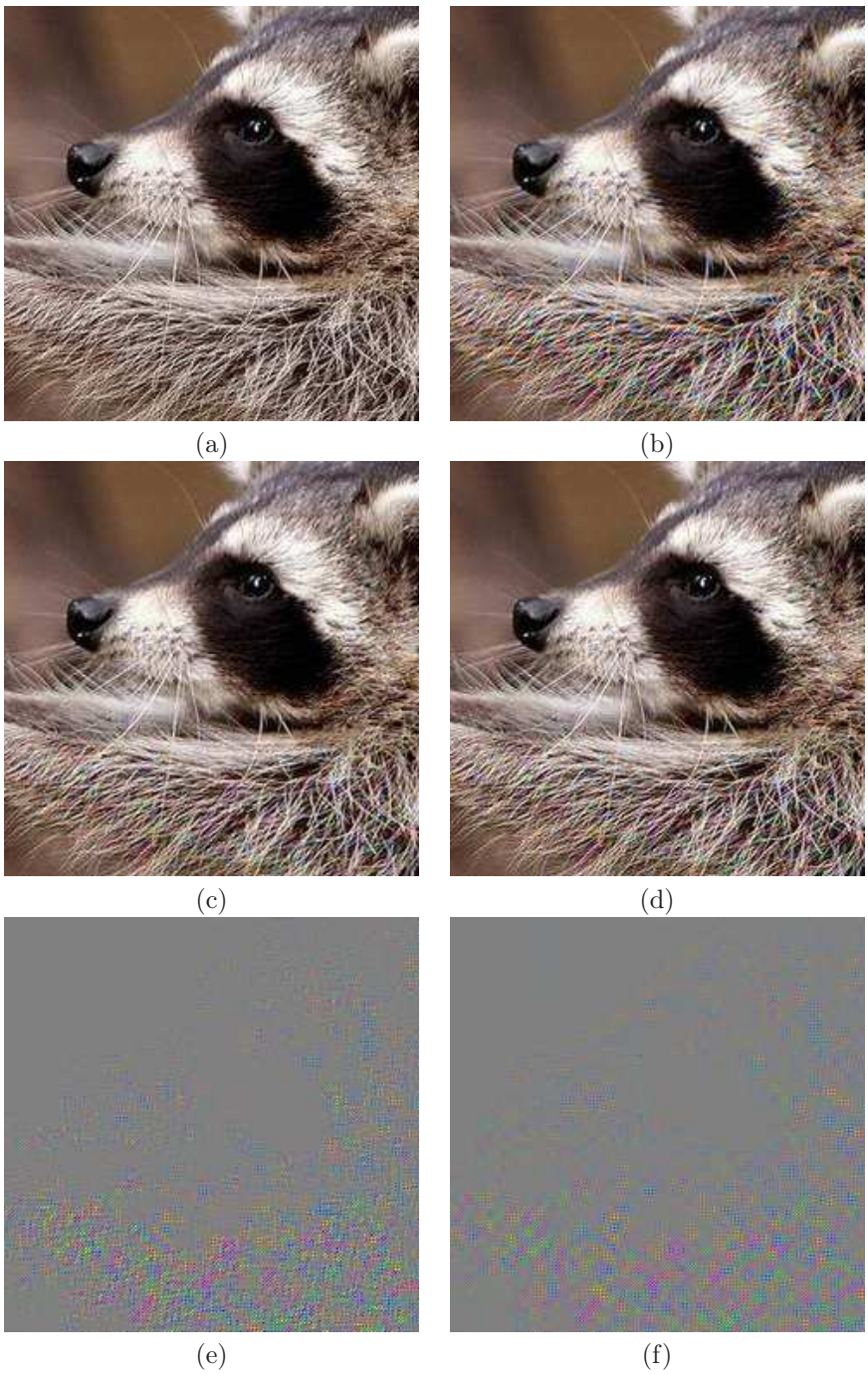
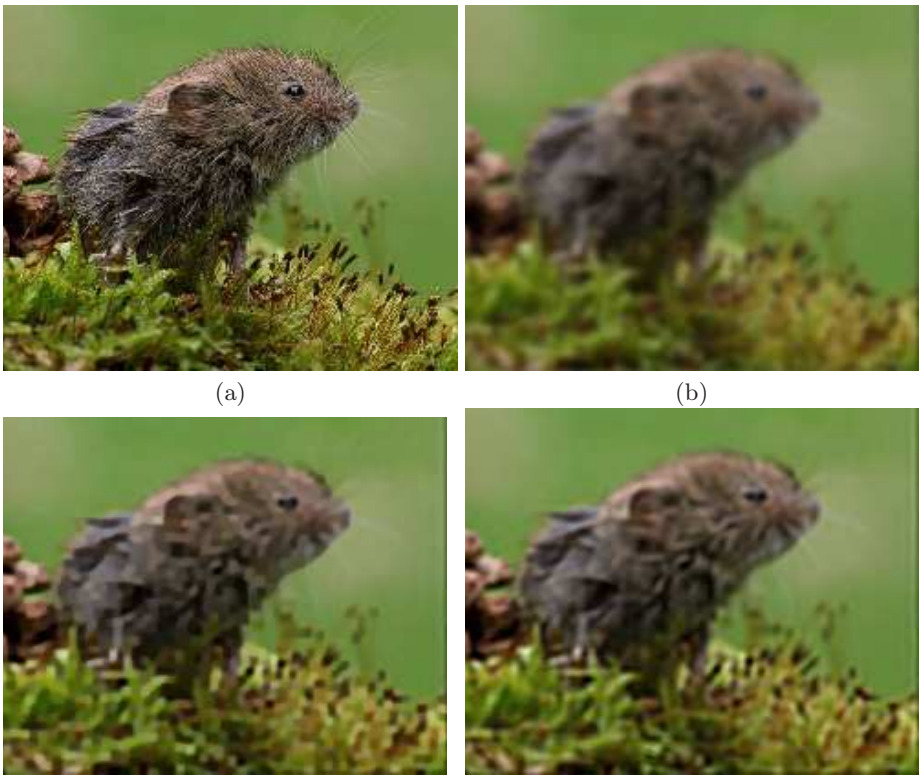


Figure 5.33: Demosaicing results (a) Original image, (b) wavelet-based demosaicing (PSNR=26.76dB) (c) DL-MMSE (PSNR=30.74dB) [Zhang and Wu, 2005], (d) our complex-wavelet based demosaicing method (PSNR=32.36dB), (e) difference image for (c), (f) difference image for (d).

5.4.4 Image Restoration using Split Bregman techniques

Results for joint denoising and deblurring

In Figure 5.34 and Figure 5.35, we applied the Split Bregman based restoration algorithm from Section 5.3.4 to a color image corrupted with artificial isotropic Gaussian blur and white Gaussian noise. The degradation model parameters are assumed to be known. We use two types of regularization: Total Variation (TV) and shearlets. In spite of the fact that TV regularization is computationally much simpler than shearlet regularization, we see that TV creates cartoon-like artifacts in the restored images. The images restored using shearlets have a more natural appearance and fine edge-like structures are better reconstructed.



(a) (b)

(c) (d)

Figure 5.34: Joint denoising and deblurring results. (a) Original image, (b) degraded image with isotropic Gaussian blur ($\sigma_{blur} = 2$) and white Gaussian noise with variance $\sigma^2_{noise} = 9$, (c) Bregman-based restoration using the TV regularization, (d) Bregman-based restoration using shearlet regularization.

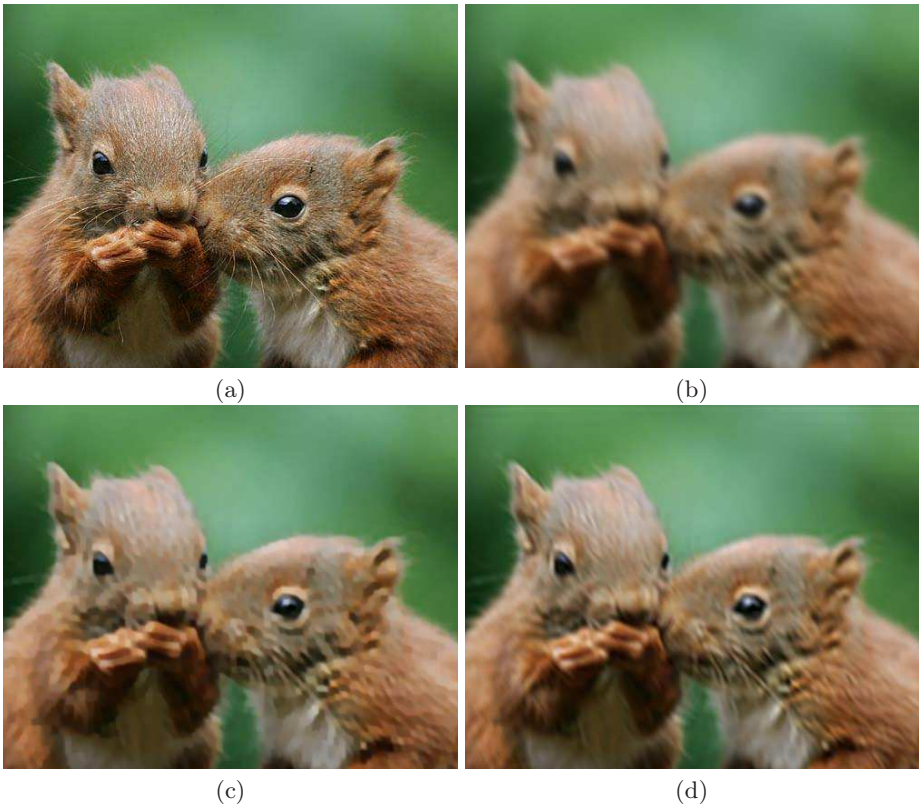


Figure 5.35: Joint denoising and deblurring results. (a) Original image, (b) degraded image with anisotropic Gaussian blur ($\sigma_{blur} = 3$) white Gaussian noise with variance $\sigma^2_{noise} = 9$, (c) Bregman-based restoration using the TV regularization, (d) Bregman-based restoration using shearlet regularization.

Results for blind denoising of correlated Gaussian noise

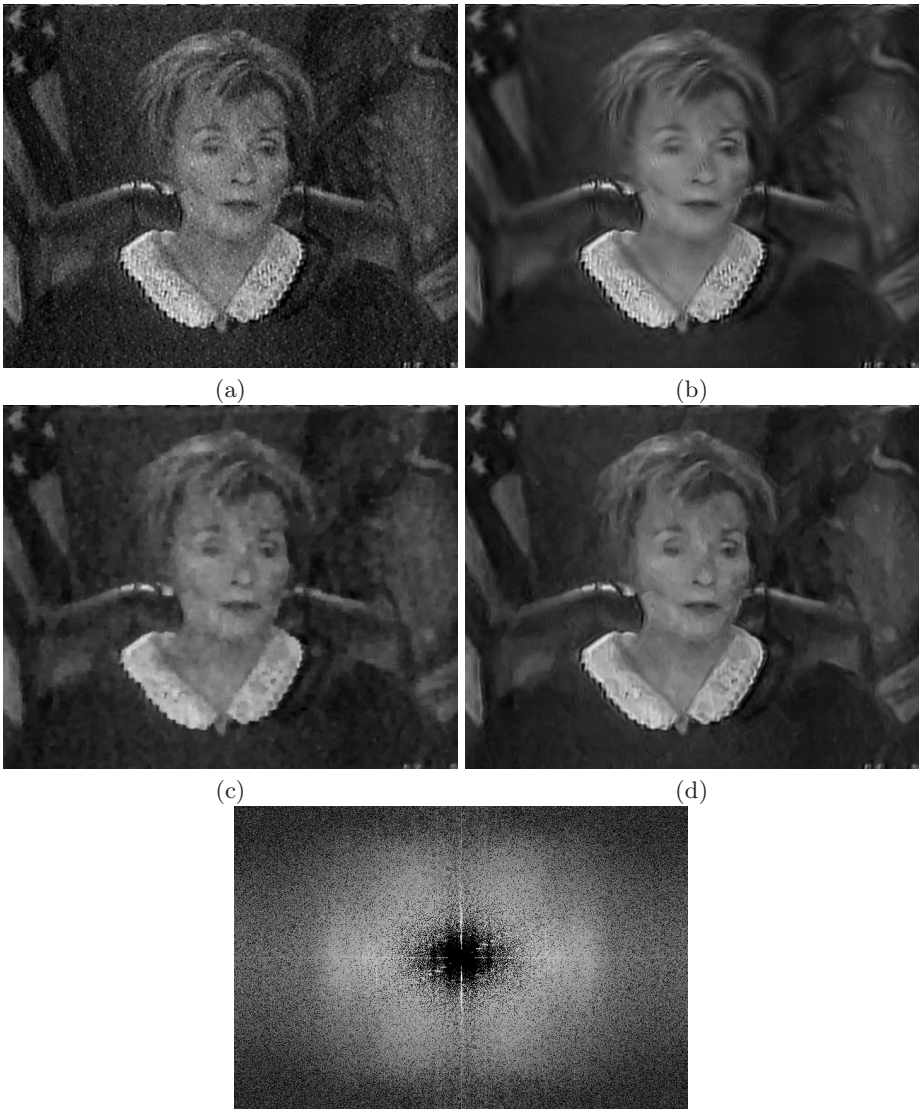
In Figure 5.36 and Figure 5.37, we processed two noisy grayscale PAL TV images. As explained in Chapter 4, we can assume that these images are degraded by additive stationary colored Gaussian noise. Because we do not known the noise PSD in advance, we use the blind denoising algorithm for correlated noise presented in Section 5.3.3 and Section 5.3.4. We do not include an extra blurring operation in this experiment and set $\mathbf{A} = \mathbf{I}$. This allows us to compare the following two restoration methods:

- The blind DT-CWT based denoising algorithm from Section 4.2.2.
- The Bregman based restoration algorithm from Section 5.3.4.

Recall that the blind DT-CWT based technique completely estimates the noise covariance matrices in every subband of the DT-CWT individually. As such, the technique can not directly provide an estimate of the noise PSD, since this would also require estimating cross-correlations between complex wavelet subbands. Because the Bregman-based algorithm can deal with blurring degradations of the input image and can directly estimate the noise PSD and convolution kernel, this algorithm is inherently "more powerful" than the blind DT-CWT denoising method.

For the *Judy* image (Figure 5.36), the estimated noise PSD is shown in Figure 5.36(e). It can be noted that the estimated noise PSD has an isotropic band-pass characteristic. The reason that the low-pass frequencies of the PSD are attenuated is a combined effect of the regularization in estimating the PSD (see (5.78)) and the TV regularization used for estimating the underlying noise-free image. The Bregman optimization algorithm jointly estimates the noise PSD and the noise-free image. The final denoising result is shown in Figure 5.36(c). Again, the image reveals many cartoon-like artifacts due to the use of TV regularization. Because the algorithm gives an accurate estimate of the noise PSD, we can in fact use other denoising algorithms for correlated noise (that assume the noise PSD to be known) as well. In Figure 5.36(d), the result of the NLMeans filter for *correlated* noise from Section 5.1.2 is depicted, yielding slightly better results.

This experiment is repeated for the *Boy* image in Figure 5.37, reaching similar conclusions. As the difference images (i.e. the difference between the restored image and the noisy input image) are relatively free of signal structures, this indicates that the restoration methods correctly identify and remove the correlated noise in the images.



(a) (b) (c) (d) (e)

Figure 5.36: Blind denoising of a PAL TV image. (a) Recorded noisy image, (b) Complex-wavelet based noise estimation and denoising (from Section 4.2.2), (c) Bregman-based noise estimation and restoration using the TV norm, (d) Bregman-based noise estimation using the TV norm and restoration using the *NLMeans* filter, (e) The estimated PSD (*black* corresponds to low noise power, *gray* to high noise power).

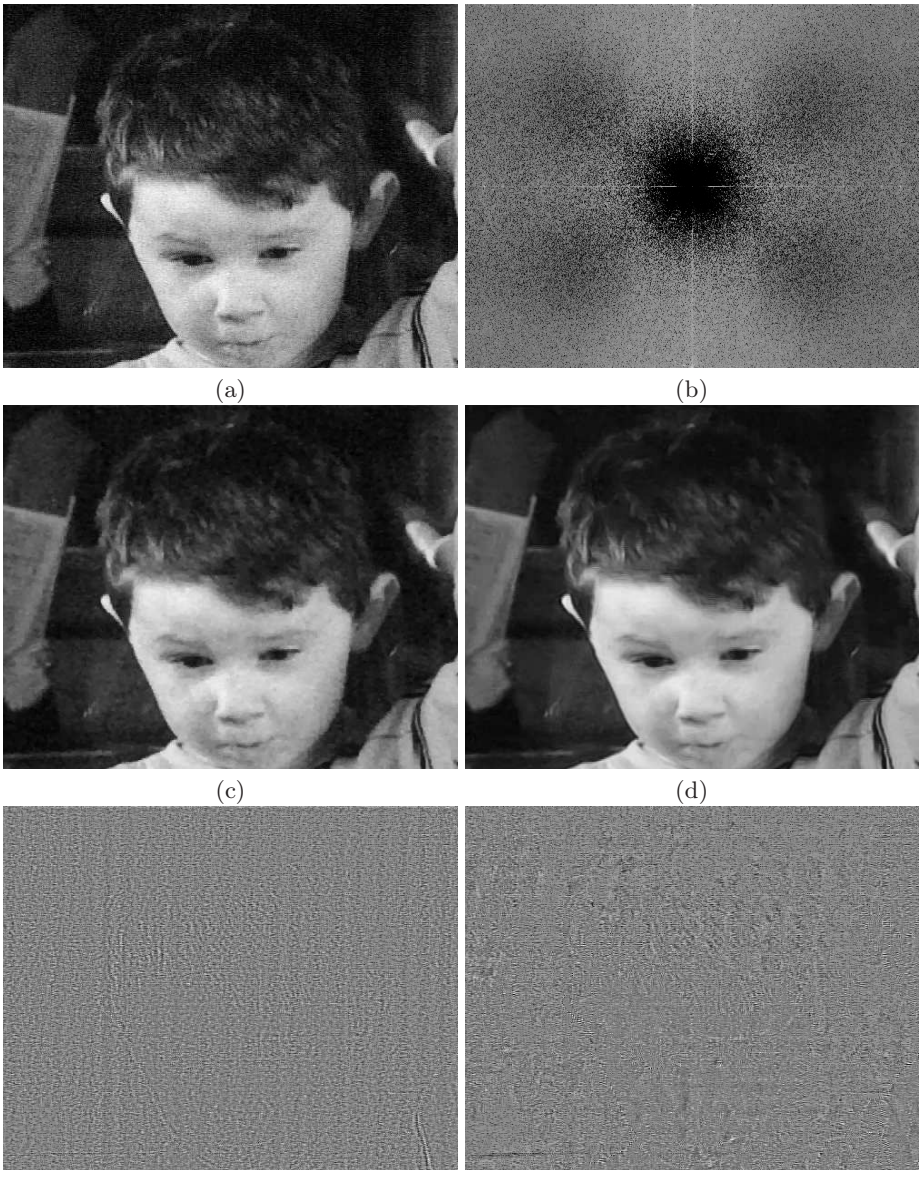


Figure 5.37: Blind denoising of a PAL TV image. (a) Recorded noisy image, (b) The estimated PSD (*black* corresponds to low noise power, *gray* to high noise power), (c) Bregman-based noise estimation and restoration using the TV norm, (d) Bregman-based noise estimation using the TV norm and restoration using the *NLMeans* filter. (e) Contrast enhanced difference image between (a) and (c). (f) Contrast enhanced difference image between (a) and (d).

As a final experiment we consider the removal of signal-dependent noise and de-biasing. In Figure 5.38(a), an original noise-free image is shown in which the intensity range has been stretched to $[-64, 320]$. In Figure 5.38(b), white Gaussian noise is added with standard deviation $\sigma = 40$. Next, the intensity range of the noisy image is clipped to the range $[0, 255]$. This results in a reduced contrast, as can be seen in Figure 5.38(b). The goal is now to recover the image in Figure 5.38(a) from the image in Figure 5.38(b). For this task, we use the Bregman algorithm from Section 5.3.5. In in Figure 5.38(c), the result is shown for TV regularization. We see that the algorithm is able to reconstruct many details of the see waves and the cliffs in the background from the noise information in Figure 5.38(b). Because the de-biasing function amplifies estimation errors in the low and high intensity ranges of the image, most artifacts are visible in these ranges in Figure 5.38(b). Because the Bregman algorithm gives an estimate of the noise variance for each position in the image ($\sigma'^2\left(x_1'\right)$), we can again plug these parameter values into an alternative denoising method, such as the NLMeans filter from Section 5.1.2, by employing the weighting function (5.8). Finally, we remove the bias from the denoised image using (5.87). The result is shown in Figure 5.38(d). It can be seen that the NLMeans result suffers much less from the amplification of estimation errors, giving a qualitatively better result. Despite the high variance of the noise in the input image and the significant loss of information due to clipping, both restoration methods are well able to recover the original image. In Figure 5.38(e)-Figure 5.38(f) the intensity histograms of the recovered image and the original image are being compared. These histograms overlap well for a large part of the intensity range.

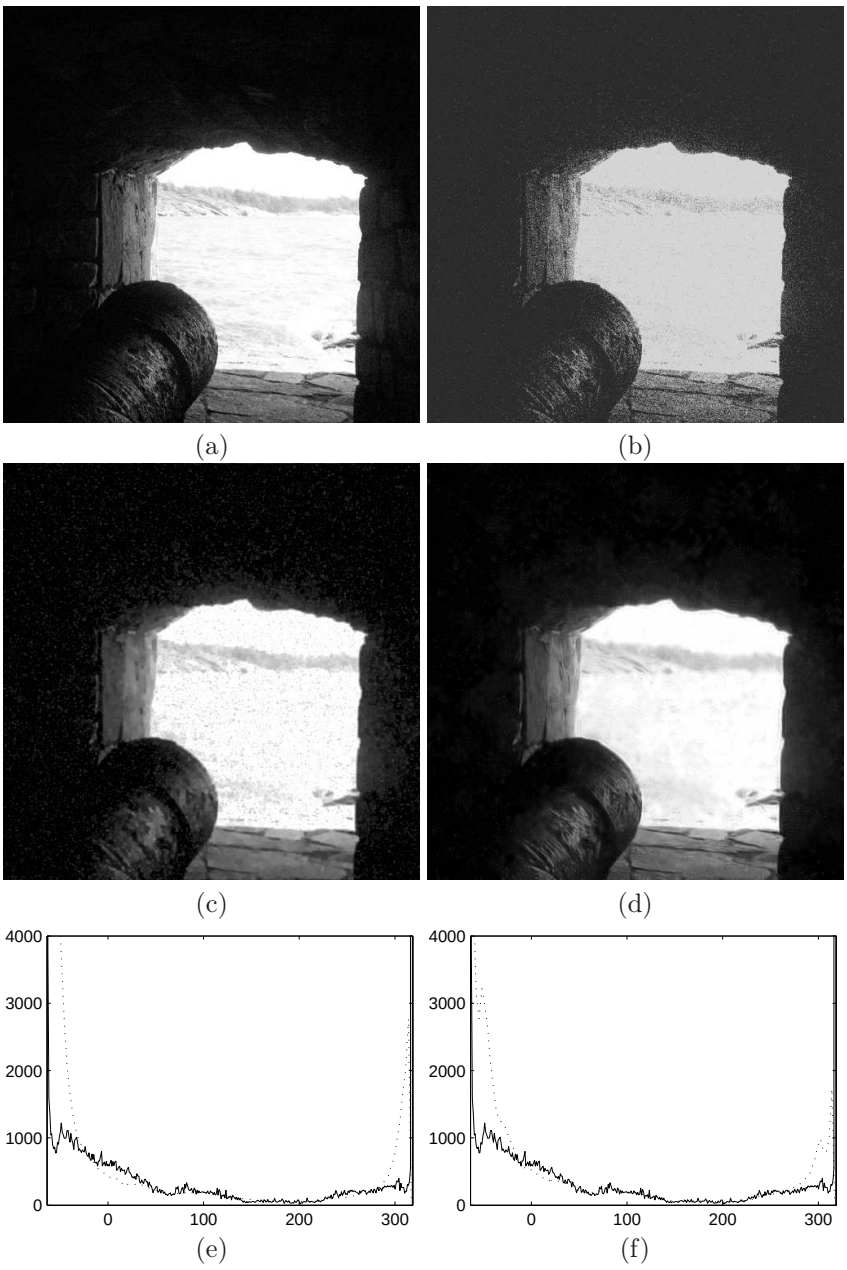


Figure 5.38: Removal of signal-dependent noise and de-biasing. (a) Original image with intensity range $[-64, 320]$, (b) Degraded image with Gaussian noise added and clipped to $[0, 255]$ (PSNR=13.87dB), (c) Restoration using the TV norm (PSNR=22.69dB), (d) Restoration using NLMeans algorithm (PSNR=25.36dB), (e) Intensity histogram of the denoised image (c), (f) Intensity histogram of the denoised image (d) (the dotted line is the histogram of the original image).

5.5 Conclusion

In this chapter, we have presented a number of novel image restoration algorithms according to three different designs. First, we have further improved the NLMeans filter, which exploits the self-similarity in images and is able to attain a high denoising performance, both in PSNR as visually. Unfortunately, the filter is sometimes not able to reconstruct well very fine structures, especially when these structures are not pronounced in the image. Therefore, multiresolution-based restoration methods can bring a solution. We have developed MAP and MMSE estimation rules for the Bessel K Form density. This brought us some extra insights in how thresholding rules are related to the kurtosis of the subband coefficients. To deal with correlated noise, we presented the Vector-ProbShrink denoising method, which is based on our novel joint inter/intrascale statistical model, yielding a denoising performance that is comparable to current state-of-the-art multiresolution denoising techniques.

We also derived the MMSE estimator for the MPGSM model and we showed that the MPGSM denoising method offers a vast improvement in PSNR especially for texture-rich images. Hence, both Vector-ProbShrink and MPGSM can be considered to be improved versions of the ProbShrink and BLS-GSM methods. Both methods achieve this in a different manner: while Vector-ProbShrink relies on a joint inter/intrascale model, MPGSM uses a more sophisticated intrascale model (but ignores interscale dependencies).

We remark that these model improvements are orthogonal, i.e. it is possible to combine MPGSM with an interscale model as well. However, for future research, we believe that most improvement can be gained by further incorporating non-local concepts in the MPGSM model estimation. As we explained in Section 3.4.4, by the nonlocal training, the model can capture similarities present in the multiresolution subbands. However, the denoising process itself is still local and cannot fully take advantage of the nonlocal information present in images.

Next, we discussed a novel complex-wavelet packet based demosaicing algorithm, that is particularly intriguing because it fully exploits the properties of the complex wavelets in order to reconstruct fine details in the image and at the same time it has a low computational complexity. The algorithm can be easily extended to perform joint denoising and demosaicing.

Finally, we presented the Bregman optimization framework for image restoration, which solves the restoration problem in an iterative way but allows to tackle more complicated restoration problems. The great benefit of this approach is that it easily allows to combine one (or multiple) multiresolution transforms from Chapter 2, a statistical image model from Chapter 3 and a noise model from Chapter 4 to this end. We illustrated this through a few restoration examples, such as joint denoising and deconvolution, blind restoration of images in correlated noise and joint denoising and debiasing. The Bregman framework has however a much wider applicability, especially in image reconstruction applications. We are currently investigating Bregman-based demosaicing schemes and MRI compressed sensing reconstruction for arbitrary

non-uniformly sampled K-space trajectories [Aelterman et al., 2010c].

The contributions of this chapter have already resulted in the following publications: [Goossens et al., 2008a] (on the improved non-local means filter), [Goossens et al., 2009d] (Vector-ProbShrink), [Goossens et al., 2009c] (the MPGSM denoising method) and [Aelterman et al., 2009] (the complex wavelet-based demosaicing algorithm).

6

Noise models for CT images

In Computed Tomography, cross-sectional images (also called slices) are made from an object by illuminating the object from many different directions. This technique enables physicians to look inside the human body and allows easy visualisation of various abnormalities. After the enormous success by the introduction x-ray Computed Tomography in the 70's, this image formation method has been extended to magnetic resonance imaging, ultrasound and microwaves, but also to nuclear medicine, e.g., positron emission tomography (PET) and single photon emission computed tomography (SPECT). These new imaging modalities are now also used on a daily basis since the early 90's. Nevertheless, by its greatly increased availability, x-ray Computed Tomography is still very popular nowadays.

One important issue in Computed Tomography is that statistical random noise can not be avoided because of the involved health risks of radiation exposure. Typically, the radiation dose is kept as low as possible while keeping an acceptable image quality. Nevertheless, noise artifacts can be very distracting and may even cause wrong diagnosis. Because of the increasing number of clinical applications of low-dose CT, the precise characterization of noise in CT becomes even more important. This is not only for the analysis of the CT images, but also for the restoration (increasing the SNR) and for devising better reconstruction techniques. As already mentioned, the CT image quality highly depends on the artifacts that can be found in these images.

Statistical models are also useful for defining image quality measures, to study e.g. which reconstruction algorithm performs better and to optimize reconstruction parameters. Another emerging field is Computer-assisted diagnosis (CAD), where computer-based techniques are used for identifying visually subtle features. Knowledge of and understanding the noise properties in CT is crucial and statistical noise characterization can help to further improve CAD techniques.

In this chapter, we start from the traditional filtered backprojection (FBP) reconstruction algorithm (Section 6.1) and we review a number of existing statistical models for CT noise (Section 6.4.1). Next, we present a novel, improved statistical model for noise in CT images reconstructed with the FBP algorithm

in Sections 6.4.2-6.6. Finally, experimental results are given in Section 6.7.

6.1 The Filtered Backprojection Algorithm

6.1.1 Parallel-beam CT

In this section, we briefly review the filtered backprojection algorithm, that is still viable for todays commercial CT scanners. We only discuss main background information of importance for our further analysis of CT noise. For comprehensive treatment of CT reconstruction algorithms, we refer to the many textbooks on this topic, e.g. [Kak and Slaney, 2001, Hsieh, 2003].

In tomography, an image is formed by radiating an object from different directions. Any objects present in the scanning plane attenuate the transmitted x-rays. An array of detectors measures the x-ray photon energy in a particular direction. Different projections of the image are obtained by measuring in different directions. The computational aspect is then the reconstruction of the image from its projections.

For monoenergetic x-ray photons (i.e. in the absence of beam hardening), the x-ray intensities measured by the detectors for a single uniform material follow the Lambert-Beers law [Kak and Slaney, 2001, Hsieh, 2003]:

$$X^m(\vartheta, t) = X_{\text{ref}} \exp\left(-\mu \Delta u\right) \tag{6.1}$$

where X_{ref} is the transmitted x-ray intensity, $X^m(\vartheta, t)$ is the measured data by the detectors, ϑ is the projection angle, t is the sampling position along the detector, μ is the linear attenuation coefficient of the material (in cm^2/g) and Δu is the thickness of the material. Typically, bones have higher attenuation coefficients than tissues, which means that the measured data will have a lower magnitude.

Of course, in practice materials being scanned are not uniform and the attenuation coefficient varies with the position. To deal with this situation, (6.1) can be extended as follows:

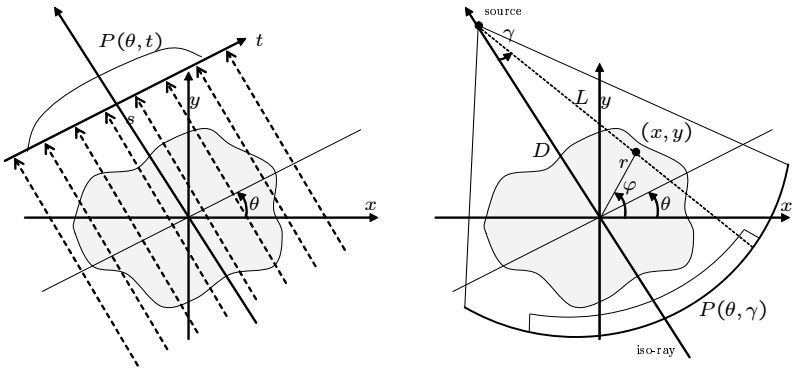
$$X^m(\vartheta, t) = X_{\text{ref}} \exp\left(-\int_L \mu(u \sin\vartheta + t\cos\vartheta, -u\cos\vartheta + t\sin\vartheta)\mathrm{d}u\right) \tag{6.2}$$

where L is a line over which the integration takes place. In (6.2), the attenuation distribution $\mu(x, y)$ depends on the position in the image. An illustration is given in Figure 6.1.

Next, the projection data is obtained by taking the negative logarithm of the measured data (which we will further call "logarithmic transform"):

$$P^m(\vartheta, t) = -\log\left(\frac{X^m(\vartheta, t)}{X_{\text{ref}}}\right) \tag{6.3}$$

$$= \int_L \mu(u\cos\vartheta - t\sin\vartheta, u\sin\vartheta + t\cos\vartheta)\mathrm{d}u \tag{6.4}$$



(a) (b)

Figure 6.1: (a) Acquisition of a parallel-beam CT image. The t axis denotes the sampling position along the detector, ϑ is the projection angle. (b) Acquisition of a fan-beam CT image. γ is the angle between the iso-ray and the line that connects the source to the reconstructed pixel. L is the distance from the x-ray source to the reconstructed pixel at position (x, y). D is the distance between the x-ray source and the iso-center.

The transform that maps $\mu(x, y)$ onto $P^m(\vartheta, t)$ is called the Radon transform. The goal of CT reconstruction is to recover the attenuation distribution $\mu(x, y)$ from the projection data $P^m(\vartheta, t)$. We remark that the relationship from (6.3) is only valid in ideal circumstances. In practice, the relationship generally does not hold because of the polyenergetic nature of the x-ray beam, scattered radiation and patient motion. These problems will be discussed in more detail in Section 6.2. To deal with these acquisition imperfections, extra pre-/post-processing steps are usually implemented in the scanner.

A widely used reconstruction technique is the FBP algorithm, because this technique has proven to be highly accurate and is amenable to fast implementation. Here, a point (x, y) of the image is reconstructed by integrating over all lines $x \cos\vartheta + y \sin\vartheta = t$ that go through this point [Kak and Slaney, 2001, Hsieh, 2003]:

$$\mu(x,y) = \int_0^{\pi} \mathrm{d}\vartheta \int_{-\infty}^{+\infty} \mathrm{d}t P^m(\vartheta,t) q\left(t - (x\cos\vartheta + y\sin\vartheta)\right), \tag{6.5}$$

where

$$q(t) = \int_{-\infty}^{+\infty} |\omega|\, G(\omega) \exp\left(j\omega t\right) \mathrm{d}\omega \tag{6.6}$$

is the impulse response of the FBP filter. Here, $G(\omega)$ is a smoothing filter. In this technique, the projection data is filtered in the t-direction with the FBP kernel $q(t)$ and subsequently backprojected. The factor $|\omega|$, also called "ramp filter", represents the Jacobian for a change of variables between polar and Cartesian coordinates. Because the "ramp filter" amplifies high frequencies,

Table 6.1: Smoothing filters and their corresponding impulse responses.

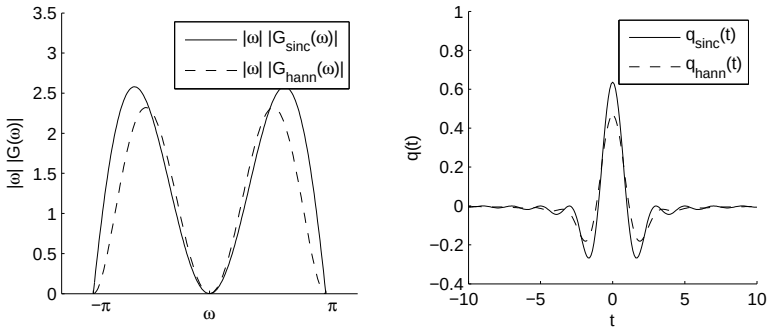
Smoothing filter	Frequency and impulse response
Sinc	$G_{\mathrm{sinc}}(\omega) = \frac{\sin\omega}{\omega}\mathrm{I}\left(\lvert\omega\rvert \le \pi\right)$ $q_{\mathrm{sinc}}(t) = \begin{cases} \frac{1+\cos\pi t}{\pi\left(1-t^2\right)} & \lvert t\rvert^2 \neq 1 \\ 0 & \lvert t\rvert^2 = 1 \end{cases}$
Hann	$G_{\mathrm{hann}}(\omega) = \frac{1}{2}\left(1+\cos\omega\right)\mathrm{I}\left(\lvert\omega\rvert \le \pi\right)$ $q_{\mathrm{hann}}(t) = \begin{cases} \left(\frac{1}{2}\pi^2 - 2\right)/2\pi & t = 0 \\ \left(\frac{1}{4}\pi^2 - 2\right)/2\pi & \lvert t\rvert = 1 \\ \frac{(1-3t^2)\cos\pi t + (1-t^2)\pi t \sin\pi t + t^2 - 2t^4 - 1}{2\pi t^2 (1-t^2)^2} & \text{else} \end{cases}$

noise at those frequencies is amplified. Therefore, a (low frequency) smoothing filter with frequency response $G(\omega)$ is added for regularization purposes, to balance noise in the reconstructed image and the spatial blur introduced by filtering. Without loss of generality, we normalize the FBP filter as follows:

$$\int_{-\infty}^{+\infty} |\omega|^2 \left|G(\omega)\right|^2 \mathrm{d}\omega = 1. \tag{6.7}$$

This means that the FBP filter has noise gain 1, i.e. the variance of white noise not affected by filtering. In Table 6.1, typical smoothing filters are given together with their impulse responses. Later, we will see that the choice of the smoothing filter mainly determines the noise characteristics. Consequently, the noise characteristics can be improved by a proper design of $G(\omega)$. In Figure 6.2, the magnitude response of the FBP filters from Table 6.1, together with their impulse response is shown.

The backprojection formula (6.5) is a technique to invert the Radon transform, but is by no means exact and thus should not be called "inverse Radon transform"! For example, the mean of the attentuation distribution can not be recovered. To see this, consider adding a constant to the measured projection



(a) (b)

Figure 6.2: (a) Magnitude response of FBP filters: the sinc-filter and Hann-filter. (b) The corresponding impulse responses.

data $P^m(\vartheta, t)$:

$$\begin{aligned}
\mu'(x,y) &= \int_0^{\pi} \mathrm{d}\vartheta \int_{-\infty}^{+\infty} \mathrm{d}t \, (P^m(\vartheta,t) + \mu_0) \, q\,(t - (x\cos\vartheta + y\sin\vartheta)) \\
&= \mu(x,y) + \mu_0 \int_0^{\pi} \mathrm{d}\vartheta \int_{-\infty}^{+\infty} q\,(t - (x\cos\vartheta + y\sin\vartheta))\,\mathrm{d}t \\
&= \mu(x,y) + \mu_0 \int_0^{\pi} \mathrm{d}\vartheta \int_{-\infty}^{+\infty} q\,(t)\,\mathrm{d}t \\
&= \mu(x,y)
\end{aligned}$$

where we relied on $\int_{-\infty}^{+\infty} q\,(t)\,\mathrm{d}t = |0|\,G(0) = 0$. In practice, the mean is "recovered" by calibration with reference to water at a standard temperature and air pressure. The intensities of CT images are usually rescaled to match the Hounsfields scale [Hsieh, 2003]:

$$\mathrm{HU} = \frac{\mu(x,y) - \mu_{H_2O}}{\mu_{H_2O}} \times 1000 \tag{6.8}$$

By this definition, water has HU = 0. Air has the lowest attenuation value, with definition (6.8), this corresponds to HU = −1000. Bones have HU ≈ 400 or more.

6.1.2 Discrete implementation

The reconstruction formulas that we have studied so far, are meant for projection data with a continuous domain. In practice, data is acquired for a finite number of projection angles and detector positions. Now we will briefly discuss the discrete implementation of the FBP reconstruction algorithm for parallel-beam CT reconstruction. Let ϑ_k, $k = 1, ..., K$ and t_l, $l = 1, ..., L$ respectively denote the projection angles and detector positions. For the FBP

algorithm, it is required that a sufficient number of projection angles ϑ_k are available and that the angles span the range $[0, \pi]$. In particular, the number of projection angles depends on the desired spatial resolution or voxel size of the reconstructed image. After acquisition, the measured data $X^m(\vartheta_k, t_l)$ is obtained. The outline of the algorithm is then as follows:

1. *Logarithmic transform:* apply the logarithmic transform from (6.3) to the measured data $X^m(\vartheta_k, t_l)$ to obtain $P^m(\vartheta_k, t_l)$.

2. *Fourier transform*: transform the mearured projection data $P^m(\vartheta_k, t_l)$ to the DFT domain (the DFT filtering takes place in the t-direction).

3. *Ramp/smoothing filtering*: filter the projection data in the DFT domain with $|\omega|\, G(\omega)$. The FBP filter $|\omega|\, G(\omega)$ is evaluated in the frequency points $\omega_l = l/\pi, l = -L/2, ..., L/2-1$. Let $P^f(\vartheta_k, t_l)$ denote the obtained filtered projection data (after applying the inverse DFT).

4. *Zero-padding and backprojection*: evaluate the backprojection formula (6.5), where the integral over ϑ is replaced by a finite sum (for $m = 1, ..., M$ and $n = 1, ..., N$):

$$\mu(m, n) = \sum_{k=1}^{K} P^f \left(\vartheta_k, \text{round}\left(m \cos\vartheta_k + n \sin\vartheta_k\right)\right), \tag{6.9}$$

   where round $(\cdot)$ signifies rounding to the nearest integer. This regridding is necessary because the sampling coordinates $(\vartheta_k, m\cos\vartheta_k + n\sin\vartheta_k)$ do not coincide with the sampling grid for $P^f(\vartheta_k, t_l)$. Rounding to nearest integer (or nearest neighbor interpolation) is computationally very efficient, although this technique causes small aliasing artifacts to be visible in the reconstructed images. Other interpolation methods, such as linear or cubic spline interpolation, are used as well. In general, let $h(t)$ denote the interpolation kernel, then:

$$\mu(m, n) = \sum_{k=1}^{K} \sum_{l=1}^{L} P^f\left(\vartheta_k, t_l\right) h\left(t_l - m\cos\vartheta_k + n\sin\vartheta_k\right). \tag{6.10}$$

   Ideally, assuming that the projection data is bandlimited, a sinc-interpolation kernel with appropriate bandwidth needs to be used. Because this requires some time-costly computations, a more efficient alternative is to ideally upsample $P^f\left(\vartheta_k, t_l\right)$ by padding with zeros in the DFT domain in step 2 [Seppä, 2007]. Subsequently, a simpler interpolation method (or even nearest neighbor interpolation) can be used in this step.

This version of the discrete FBP algorithm is widely used, not only because of the simplicity of the implementation (which allows the reconstruction algorithm

to be easily ported to dedicated DSP chips), but also because the good quality of the reconstructed images.

In most commercially available CT scanners, fan-beam based acquisition (see Figure 6.1(b)) is used instead of parallel-beam acquisition. For fan-beam CT, the x-ray tube, no longer needs to be translated. This significantly reduces the scan time. The x-ray source is a single point source and all x-rays are transmitted from this single point. To reconstruct the image, the fan-beam data can either be reformatted into parallel-beam, such that the above FBP algorithm can be used, or dedicated reconstruction formulas can be used [Hsieh, 2003, p. 76-86].

6.2 Sources of noise and imperfection in the measurement data

In real-life situations, the measurements obtained by clinical CT scanners rarely satisfy the conditions posed in Section 6.1 (i.e. monoenergetic x-rays, noise-free measurements...). In practice, the measured data must be preprocessed before any reconstruction technique can be applied [Hsieh, 2003]. To give a better view on the complexity of the entire CT reconstruction chain, we mention a number of sources of imperfection:

- Because of *quantum mechanical effects* during the measurement of photon energy, the detector response varies from one measurement to another. This statistical noise is usually called quantum or photon noise and can not be avoided in the acquisition. As we will see further on, the noise energy can be reduced by increasing the tube current but this also increases the radiation exposure.

- A second noise source is *electronic noise* introduced by electronic circuitry (e.g. amplifiers) in the CT scanner. In normal circumstances, this kind of noise only contributes to a small fraction of the statistical noise. Because of the ramp-filtering in the FBP algorithm, both the photon noise and the electronic noise are magnified after reconstruction. More concretely, a deviation on one projection data sample causes a bright or dark straight line in the reconstructed image (called streaking artifact).

- A third problem that affects the accuracy of the measurements is *scattered radiation*. X-ray photons are partially diffracting and do not completely travel along straight lines. Consequently, the Lambert-Beers law (equation (6.2)) does not hold exactly. Typically, scattered radiation produces shading and streaking image artifacts [Hsieh, 2003] or results in loss of resolution. This problem is even more pronounced in ultrasound tomography (e.g. [Maleki et al., 1992]) or more recently microwave tomography [Van den Bulcke and Franchois, 2009], where the ultrasound waves are scattered in all possible directions. In this case, completely

different reconstruction techniques need to be used. For an overview of these techniques, we refer to [Kak and Slaney, 2001].

- A fourth problem is the assumption that x-rays are *mono*energetic, so that x-ray photons emitted from the x-ray source all have the same energy, which is rarely satisfied in practice. For polyenergetic x-rays, there is no linear relationship between the measured projection data and the thickness of the object; this causes so-called beam-hardening artifacts in the reconstructed image.

- A fifth problem is *patient motion*: the acquisition time is in the order of tens of seconds and involuntary patient motions, such as breathing, are inevitable. Because the basic assumptions for CT (equation (6.2)) are violated, this inherently leads to image artifacts.

- Other sources are [Hsieh, 2003]: nonlinearity of the detector elements, off-focal radiation of the x-ray source, the presence of metal objects, x-ray photon starvation, CT gantry misalignment, x-ray tube arcing, deficiency in the projection sampling, partial volume effect, focal spot drift. For a detailed overview of these problems, we refer to [Hsieh, 2003].

In this work, we will mainly concentrate on the first two problems, i.e. electronic and photon noise. The noise modeling consists of two steps:

1. Modeling of noise in the projection data, i.e. before reconstruction (Section 6.3).

2. Investigation of the noise properties *after* FBP reconstruction (Section 6.4).

These steps are discussed in more detail in the following Sections.

6.3 Signal-dependency characteristics of the projection data noise

In this section, we will put forward a model for the signal-dependency of the noise in the projection data. For this task, we build further on the signal-dependent noise modeling techniques presented in Section 4.4.2. In Figure 6.3, a simplified processing chain that converts an ideal (noise-free) input x-ray intensities $X(\vartheta, t)$ into projection data $P^m(\vartheta, t)$ is shown. The input signal, measured by the photon detectors $X^d(\vartheta, t)$, is first amplified electronically and then converted to a digital signal by the analog to digital (A/D) converter. Next, the logarithmic transform (6.3) is applied, which results in the measured projection data $P^m(\vartheta, t)$. We will now discuss the noise sources in the individual processing blocks of the chain. First, we assume that the measured response of the x-ray detector $X^d(\vartheta, t)$ is directly proportional to the x-ray

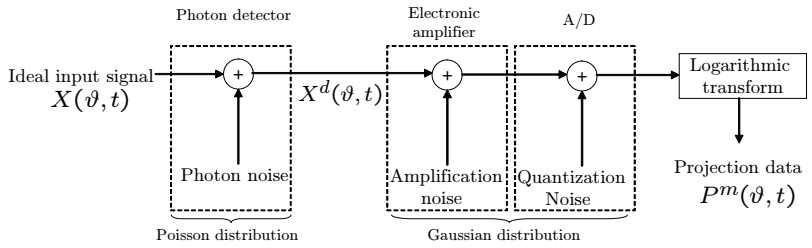


Figure 6.3: The processing chain for CT projection data and the introduction of noise. Figure is based on the description in [Kak and Slaney, 2001, p. 190].

photon energy. This allows us to model the detected x-ray intensities using a Poisson distribution:

$$X^d(\vartheta,t) \sim \text{Poisson}\left(X(\vartheta,t)\right). \tag{6.11}$$

We further assume that the noise measured by one detector is statistically independent of noise measured by other detectors.[1] The response $X^d(\vartheta,t)$ is amplified electronically, which introduces amplification noise. Also the A/D conversion, which is subsequently applied, can be considered as a uniform noise source. Because the amplification noise has a white Gaussian distribution with variance that is typically larger than the variance of the quantization noise, the joint contribution can be modeled by a white Gaussian noise source with mean $X(\vartheta,t)$ and variance σ_ϵ^2:

$$X^m(\vartheta,t) \sim \mathcal{N}\left(X(\vartheta,t),\sigma_\epsilon^2\right). \tag{6.12}$$

Recall that the measured projection data is obtained by taking the negative logarithm of $X^m(\vartheta,t)$ (see equation (6.3)), hence a precise characterization of the distribution of $P^m(\vartheta,t)$ is even more complicated. To simplify the problem and to enable us to model the CT noise after backprojection (see further), we will approximate the distribution of the projection data $P^m(\vartheta,t)$ using its first and second order moments, i.e. with a Gaussian distribution. The Gaussian approximation turns out to be very accurate for radiation doses used in practice (which means that X_{ref} is sufficiently large), as we will show further on. We will now briefly outline the procedure to compute these moments.

First, note that the measured data can be expressed in terms of the noise-free data, as follows:

$$\begin{aligned} X^m(\vartheta,t) &= X^d(\vartheta,t) + \epsilon(\vartheta,t) && (6.13)\\ &= X(\vartheta,t)\left(\frac{X^d(\vartheta,t)}{X(\vartheta,t)} + \frac{\epsilon(\vartheta,t)}{X(\vartheta,t)}\right) && (6.14)\end{aligned}$$

[1] This assumption implies that cross-talk between detector elements is ignored. In practice, cross-talk can not be completely avoided. However, by proper design specifications, the cross-talk can be kept below an acceptable level. [Hsieh, 2003, p. 160]

where we rely on the inequality $0 < X(\vartheta,t) \leq X_{\text{ref}}$. Next, a logarithmic transform is applied, which converts the product into a sum:

$$P^m(\vartheta,t) = P(\vartheta,t) + \log\left(\frac{X^d(\vartheta,t)}{X(\vartheta,t)} + \frac{\epsilon(\vartheta,t)}{X(\vartheta,t)}\right) \tag{6.15}$$

In essence, we can now adopt the strategy from Section 4.4.3 which consists of computing the statistical moments of P^m conditioned on P. To arrive at analytically tractable expressions, we apply a Taylor series expansion $\log(1+x) = \sum_{k=1}^{+\infty} \frac{(-1)^{k+1}}{k} x^k$ in $x = 0$:

$$P^m = P + \sum_{k=1}^{+\infty} \frac{(-1)^{k+1}}{k} \left(\frac{X^d}{X} + \frac{\epsilon}{N} - 1\right)^k \tag{6.16}$$

where we dropped the coordinates (ϑ,t), to simplify the notations. In (6.14)-(6.16), the working point is chosen suitably, relying on the fact that $\mathrm{E}\left[X^d|P\right] = X$ and $\mathrm{E}\left[\epsilon\right] = 0$. The moments of $P^m(\vartheta,t)$ can be computed based on the moments of the Poisson distribution $\mathrm{E}\left[\left(X^d\right)^k\right]$ and the moments of the Gaussian distribution $\mathrm{E}\left[\epsilon^k\right]$, up to a certain number of terms of the Taylor series. For example, for the first moment, we find:

$$\begin{aligned}
\mathrm{E}\left[P^m|P\right] &= P + \sum_{k=1}^{+\infty} \frac{(-1)^{k+1}}{k} \mathrm{E}\left[\left(\frac{X^d}{X} + \frac{\epsilon}{N} - 1\right)^k\right] \\
&= P + \sum_{k=1}^{+\infty} \frac{(-1)^{k+1}}{k} \sum_{l=0}^{k} \binom{k}{l} \mathrm{E}\left[\left(\frac{X^d}{X}\right)^l \left(\frac{\epsilon}{N} - 1\right)^{k-l}\right] \\
&= P + \frac{1}{2X} + \left(\frac{5}{12} + \frac{1}{2}\sigma_\epsilon^2\right) \frac{1}{X^2} + \mathcal{O}\left(\frac{1}{X^3}\right).
\end{aligned} \tag{6.17}$$

Normally, we have $X \gg 1$,[2] such that first and second order terms in N^{-1} are neglible and such that the projection data is in good approximation unbiased ($\mathrm{E}\left[P^m|P\right] \approx P$). The variance of the projection data can be computed similarly:

$$\mathrm{Var}\left[P^m|P\right] = \frac{1}{X} + \left(\frac{3}{2} + \sigma_\epsilon^2\right) \frac{1}{X^2} + \mathcal{O}\left(\frac{1}{X^3}\right) \tag{6.18}$$

With the substitution $X = X_{\text{ref}} \exp(-P)$, we can write the projection data noise variance in terms of the ideal noise-free projection data:

$$\sigma^2(P) \approx \frac{e^P}{X_{\text{ref}}} + \left(\frac{3}{2} + \sigma_\epsilon^2\right) \left(\frac{e^P}{X_{\text{ref}}}\right)^2. \tag{6.19}$$

This relationship is particularly interesting, because it tells us a number of facts about the noise variance:

[2]Note that the x-ray intensity X is the expected number of photons detected per unit time.

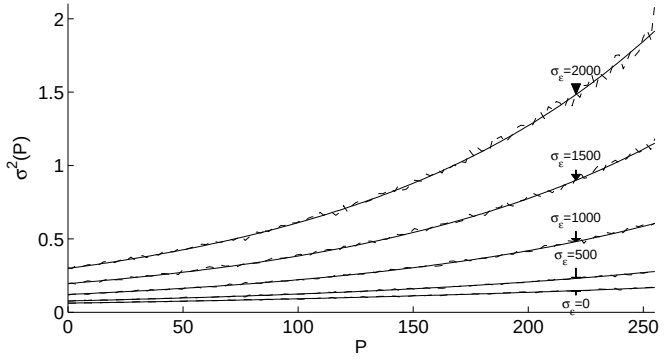


Figure 6.4: Comparison of the approximation of the noise variance $\sigma^2(P)$ to the estimated variance using numerical simulations ($X_{\text{ref}} = 255$).

- The noise variance increases with the signal P, or equivalently decreases with X. The minimum noise variance (called noise floor), is $\left(X_{\text{ref}} + \sigma_\epsilon^2 + \frac{3}{2}\right) / X_{\text{ref}}^2$. Hence, the variance can be kept low, by either reducing the amount of electronic noise or by increasing X_{ref}. The reference signal X_{ref} (which can be measured through calibration), is proportional to the tube current, but further also depends on the exposure time of the detector elements and the size of the detector elements. Hence to increase X_{ref}, one can either 1) increase the tube current, i.e. the radiation dose, 2) increase the exposure time (which also leads to higher radiation dose), 3) increase the size of the detector elements (which will typically reduce the resolution, i.e. fewer detector elements can be placed in the same area).

- The maximal noise variance is reached for *maximal* P. This is caused by a *lower*[3] incidence of x-ray photons at the receiver for maximal P (e.g. when scanning through bones).

- Because the electronic noise variance only appears starting from the second term in $1/X_{\text{ref}}$, the contribution of the electronic noise to the total noise variance is small compared to the photon noise variance.

In Figure 6.4, the approximated noise variance (6.18) is compared to the estimated noise variance obtained using numerical Monte-Carlo simulations, i.e., by generating random noise samples according to (6.11)-(6.12), for different electronic noise levels σ_ϵ and for a fixed reference signal $X_{\text{ref}} = 255$. In the experiment, the reference signal is kept very low because the approximation error increases for low X_{ref}, simulating the effect of low dose CT. In more realistic scenarios, the order of X_{ref} is 10^5 to 10^6 or higher.

[3]Note that by the negative logarithm transform (6.3), P is logarithmically proportional to the *reciprocal* of the x-ray intensity X.

By modeling the measured projection data using a Gaussian distribution, $P^m(\vartheta, t)$ is considered to be a Gaussian Random Field (GRF) with mean and variance respectively given by equation (6.17) and (6.18). This simplification will easily allow us to obtain analytical results with respect to the PSD of the noise in a reconstructed CT image (see next section).

In the previous analysis, we silently assumed that the measured output at the detector is equal to the detected number of x-ray photons. In practice, this is not always true [Hsieh, 2003]: A/D converters have a limited dynamic range which may cause the measurements to be clipped. Moreover, perfect linearity over the whole dynamic range is difficult to accomplish in practice. Generally speaking, we have $X^m(\vartheta, t) \sim \mathcal{N}\left(\gamma\left(X(\vartheta, t)\right), \sigma_\epsilon^2\right)$, where $\gamma(x)$ is a nonlinear detector response function (DRF).[4] The DRF can be obtained from theoretical analysis and phantom experiments [Hsieh, 2003]. For our work, we assume that $g(x)$ is linear over a large part of its range:

$$\gamma(x) = \gamma(x_0) + (x - x_0)\gamma_0 \tag{6.20}$$

where $\gamma_0 = \frac{\partial \gamma}{\partial x}(x_0)$ is the system gain factor. Taking (6.20) into account, the variance of the projection data now becomes:

$$\mathrm{Var}\left[P^m\right] = \frac{\gamma_0^2}{X^2}\left[X + \frac{3}{2} + \frac{\sigma_\epsilon^2}{\gamma_0^2}\right] + \mathcal{O}\left(\frac{1}{X^3}\right). \tag{6.21}$$

Here, the system gain factor causes the noise floor to change. In case the linearity assumption is not adequate enough but $\gamma(x)$ can be precisely characterized, the derivation of the first and second order moments can be repeated based on a higher order Taylor series expansion of $g(x)$. To avoid artifacts after FBP reconstruction, the inverse DRF generally needs to be applied to the projection measurements [Hsieh, 2003].

6.4 Noise modeling after FBP reconstruction

In this section, we present a noise model for CT images reconstructed with the FBP algorithm. The model builds further on the first and second order statistical moments of the projection data which were derived in Section 6.3. We assume parallel-beam acquisition, however, the results can be extended to fan-beam or cone-beam geometries as well (which will be published in later work).

6.4.1 Existing models

In [Riederer et al., 1978, Hanson, 1981, Kak and Slaney, 2001], the noise power spectral density (NSD) was derived for projection data corrupted with stationary AWGN with variance S_0, using the continuous FBP reconstruction

[4] This function is in fact a CT analogue of the camera response function (CRF) used in digital still cameras (see Chapter 4).

algorithm. It was shown that the spectral density of the noise in this case only depends on the radial frequency:

$$S(\omega) = S_0 |\omega| |G(\omega)|^2 . \tag{6.22}$$

Consequently, the shape of the NSD mainly depends on the smoothing filter $G(\omega)$. The NSD is depicted in Figure 6.5(a) for the case of a ramp filter (i.e. $G(\omega) = 1$). Because the NSD is rotationally symmetric, the corresponding noise is *isotropic* and does not show directional structures. However, we remark that noise in CT images often contains directional streaks - this can not be explained by an isotropic NSD model.

In [Hanson, 1981] the low frequency behavior of the NSD is characterized by the density of noise-equivalent quanta (NEQ) detected in projection measurements. The NEQ is defined as the total effective number of quanta detected per unit distance along the projections. [Faulkner and Moores, 1984] derived the NSD for the *discrete* backprojection algorithm. [Kijewski and Judy, 1987] further improved the description of the NSD for the discrete backprojection algorithm of [Faulkner and Moores, 1984], taking both angular sampling and sampling within each projection into account. They showed that, because of the aliasing that arises due to undersampling, the zero-frequency (DC) component of the NSD is non-zero in general and that the aliasing destroys the rotational symmetry of the NSD. [Wang and Vannier, 1993] derived analytical expressions for the noise autocorrelation function and the noise variance for helical CT under the assumption that the projection data noise is stationary, white and Gaussian. [Hsieh, 1997] investigates the non-stationary characteristics of noise in helical CT. The non-stationarity is there the combined effect of the weights used for helical reconstruction and the scaling factors used in fan-beam (instead of parallel-beam) backprojection, again assuming stationary Gaussian projection data noise.

As we discussed in the previous section, an x-ray signal follows a Poisson distribution, and consequently the x-ray noise is not additive but signal-dependent. This leads to noise streaking artifacts with a non-symmetrical PSD. Recently, a number of authors have studied the signal-dependency of projection data noise. In [Hsieh, 1998], a relationship between the projection data mean and the noise variance is determined. This relationship is similar to (6.21), up to the term $3\gamma_0^2/2X^2$, which is missing in [Hsieh, 1998] due to the approximations under which the relationship is obtained. A filter operation on the projection data is adapted to the local noise characteristics, in order to suppress noise. [Lu et al., 2001] experimentally found an alternative relationship between the projection data mean and the noise variance. They propose a scale transformation (i.e. variance stabilization) to undo the signal-dependency of the noise.

More recently, a number of authors [Pan and Yu, 2003, Zhu and Starlack, 2007, Wunderlich and Noo, 2008, O'Connor and Fessler, 2007] have studied the prediction of the noise variance and covariance in reconstructed CT images, by taking the signal-dependency of the noise into account. In [Pan and Yu,

2003], variance images are used to optimize the SNR of the reconstructed images. In [O'Connor and Fessler, 2007], expressions for the image covariance are derived for fan-beam CT. [O'Connor and Fessler, 2007] provide an efficient computation technique for predicting the noise variance based on the assumption of local stationarity. In [Wunderlich and Noo, 2008], an alternative procedure for computing the noise variance and covariance is proposed and used to optimize the lesion detectability performance of a Channelized Hotelling observer. [Zhu and Starlack, 2007] derive a straightforward technique for predicting the noise variance that is a simple adaptation of the FBP algorithm with modified convolution kernels and weighting factors.

Despite the fact that many of the recent noise (co)variance prediction methods give accurate results on phantom data, there are a few problems when trying to apply these methods in practice:

- The prediction formulas are often complicated, which does not only potentially result in error-prone implementations but also incurs an associated large computational cost (often several times the CT reconstruction time of one image). This makes these methods less practical.
- Except for the method proposed by [Zhu and Starlack, 2007], many methods assume that the variance of the projection data $\mathrm{Var}\left[P^m\right]$ is known exactly in advance. For phantom data, this is the case and the techniques can be used to optimize e.g. reconstruction parameters. However, because $\sigma^2\left(P\right)$ is a function of the noise-free data P, for non-phantom data this assumption is not very realistic. Indeed, if P is available in advance, there is no reason to take a CT scan at all! Hence techniques are needed to estimate $\sigma^2\left(P\right)$ from observed projection data.

In the next section, we will bring a potential solution to these problems.

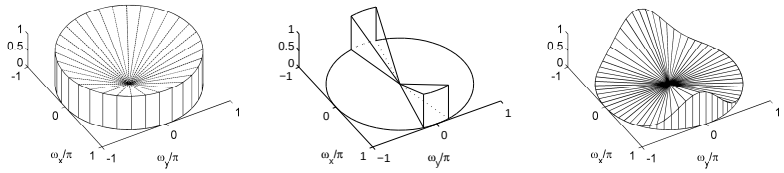
6.4.2 Analytical formulation of the local NSD

The continuous FBP algorithm

To provide more insight in the structure and properties of CT noise in reconstructed CT images, we will study the *local* NSD in polar frequency coordinates. We use the term *local* NSD to refer to the space-varying spectrum (see Section 4.3) of the additive noise component of the observed signal. To this end, we will derive a formula for the NSD as a function of the position in the image, for the FBP algorithm. The main result is that for parallel-beam CT, the local NSD is *separable* in the image domain. This leads to accurate (but simple) techniques to estimate the noise properties.

First, consider the measurement of a single spike with intensity I_0 captured by a detector at position t_0 under an angle ϑ_0. According to the signal-plus-noise model from Section 6.3, this gives the following projection measurements:

$$P^m(\vartheta,t) \approx \delta\left(t-t_0\right)\delta\left(\vartheta-\vartheta_0\right)\left(I_0-\sigma\left(I_0\right)\nu(\vartheta,t)\right), \tag{6.23}$$



(a) (b) (c)

Figure 6.5: (a) Isotropic NSD. (b) The NSD for a single projection measurement. (c) Anisotropic NSD model obtained for many projection measurements.

where $\nu(\vartheta, t)$ is white Gaussian noise with mean 0 and variance 1. The approximation in (6.23) is due to the Gaussianity assumption of the measured projection data, but the advantage of this will be clear soon. When applying the FBP formula (6.5), the reconstructed CT image of this spike is given by:

$$\mu(x, y) = q(x\cos\vartheta_0 + y\sin\vartheta_0 - t_0)\left(I_0 - \sigma\left(I_0\right)\nu(\vartheta_0, t_0)\right). \tag{6.24}$$

This is a line impulse with equation $x\cos\vartheta_0 + y\sin\vartheta_0 = t_0$ and with intensity $(I_0 - \sigma\left(I_0\right)\nu(\vartheta_0, t_0))$, which is filtered orthogonally by the FBP filter $q(t)$. Here we immediately see the streaking behavior of CT noise: for a small deviation on I_0 caused by the noise $\nu(\vartheta_0, t_0)$, either a bright or dark streak is created in the image. If we denote by

$$Q(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\omega|^2 \left|G(\omega)\right|^2 \exp\left(j\omega t\right) \mathrm{d}\omega \tag{6.25}$$

the autocorrelation function of the FBP filter, we can compute the noise autocovariance function for a point (x', y') on the same line by ensemble averaging:

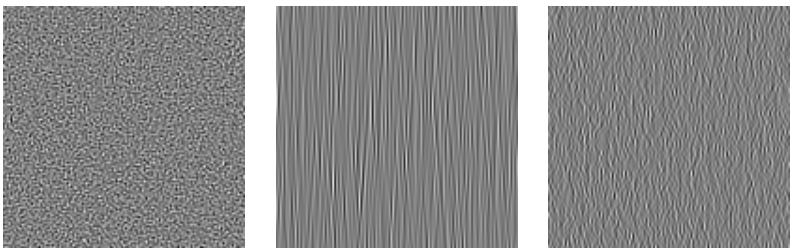
$$\begin{aligned} R(x, y) &= \mathrm{E}\left[\left(\mu\left(x', y'\right) - \mathrm{E}\left[\mu\left(x', y'\right)\right]\right)\left(\mu\left(x+x', y+y'\right) - \mathrm{E}\left[\mu\left(x+x', y+y'\right)\right]\right)\right] \\ &= \sigma^2\left(I_0\right) Q(x\cos\vartheta_0 + y\sin\vartheta_0 - t_0), \end{aligned} \tag{6.26}$$

or, in polar coordinates, $R(\vartheta, t) = \sigma^2(I_0)Q\left(t\cos\left(\vartheta - \vartheta_0\right) - t_0\right)$. The NSD can be found by taking the Polar Fourier transform of $R(\vartheta, t)$:

$$S\left(\vartheta, \omega\right) = \sigma^2\left(I_0\right)\delta\left(\vartheta - \vartheta_0\right)|\omega|\left|G(\omega)\right|^2. \tag{6.27}$$

We see that for a single spike, the NSD on each point of the line is *separable* in polar frequency coordinates, with angular component $\delta\left(\vartheta - \vartheta_0\right)$, radial component $|\omega|\left|G(\omega)\right|^2$ and scalar factor $\sigma^2\left(I_0\right)$. Figure 6.5(b) shows an illustration of the NSD corresponding to (6.27) for a particular choice of the smoothing, i.e. $G(\omega) = 1$.

Now that we have the separability result for one projection measurement, the question is if this result still holds for more measurements. In general, for any reconstructed CT image, the reconstructed intensity at position $(x, y) =$



(a) (b) (c)

Figure 6.6: Stationary synthetic noise generated for the NSDs from Figure 6.5. (a) Isotropic noise. (b) The noise introduced by projection measurement for a fixed angle of 90°. (c) Anisotropic noise for many projection measurements for different angles.

$(r\cos\varphi, r\sin\varphi)$ is a sum of all lines that pass through this point. The line bundle of all these lines, with parameter ϑ is given by:

$$r\cos(\varphi - \vartheta) = t \tag{6.28}$$

with $r \in \mathbb{R}$ and $\vartheta \in [0, \pi]$. In Appendix B, it is shown that the local NSD created by one line of the bundle (i.e. for one particular ϑ) at position (r, φ) is:

$$\begin{aligned} S^{(\vartheta)}_{(r,\varphi)}(\alpha,\omega) &= \delta(\vartheta-\alpha)\,|\omega|\,|G(\omega)|^2 \cdot \\ &\qquad \int_{-\infty}^{+\infty} \sigma^2(P(\alpha,t))\,q^2(t - r\cos(\varphi-\alpha))\,\mathrm{d}t \end{aligned} \tag{6.29}$$

where the positional dependency of the NSD is made explicit by the subscript (r, φ). The overall NSD is the sum of the NSD contributions of all lines of the line bundle, which can be obtained by integrating over the line bundle parameter ϑ:

$$\begin{aligned} S_{(r,\varphi)}(\alpha,\omega) &= \int_0^{\pi} S^{(\vartheta)}_{(r,\varphi)}(\alpha,\omega)\,\mathrm{d}\vartheta \\ &= |\omega|\,|G(\omega)|^2 \int_{-\infty}^{+\infty} \mathrm{d}t\,\sigma^2(P(\alpha,t))\,q^2(t - r\cos(\varphi-\alpha)) \\ &= |\omega|\,|G(\omega)|^2\,\psi_{(r,\varphi)}(\alpha), \end{aligned} \tag{6.30}$$

with

$$\psi_{(r,\varphi)}(\alpha) = \int_{-\infty}^{+\infty} \sigma^2(P(\alpha,t))\,q^2(t - r\cos(\varphi-\alpha))\,\mathrm{d}t. \tag{6.31}$$

We see that the local NSD in (6.30) is again separable in polar frequency coordinates. The angular component $\psi_{(r,\varphi)}(\alpha)$ is the noise power at position

(r, φ) in direction $\pm\alpha$. To see this, we integrate $S_{(r,\varphi)}(\alpha, \omega)$ over its radial frequency range:

$$\begin{aligned} \int_{-\pi}^{\pi} S_{(r,\varphi)}(\alpha, \omega) |\omega| \, \mathrm{d}\omega &= \int_{-\pi}^{\pi} |\omega|^2 |G(\omega)|^2 \psi_{(r,\varphi)}(\alpha) \, \mathrm{d}\omega \\ &= \psi_{(r,\varphi)}(\alpha) \end{aligned} \tag{6.32}$$

where we relied on proper normalization of the FBP filter (equation (6.7)). By the signal dependency of the noise (see equation (6.31)), $\psi_{(r,\varphi)}(\alpha)$ is typically non-constant and consequently, the corresponding NSD $S_{(r,\varphi)}(\alpha, \omega)$ depends on α. This leads to an anisotropic NSD, as illustrated in Figure 6.5(c). In some directions, the noise power is higher than in other directions.

Finally, we calculate the local noise variance at position (r, φ) as the surface integral of the local NSD over its domain:

$$\begin{aligned} v(r, \varphi) &= \int_0^{2\pi} \int_0^{\pi} S_{(r,\varphi)}(\alpha, \omega) |\omega| \, \mathrm{d}\omega \mathrm{d}\alpha \\ &= \int_0^{\pi} \psi_{(r,\varphi)}(\alpha) \, \mathrm{d}\alpha \end{aligned} \tag{6.33}$$

which equals the local noise power in direction $\pm\alpha$, integrated over all possible directions $\alpha \in [0, \pi]$.

The discrete FBP algorithm

The above derivation can be extended to the discrete FBP algorithm as well, using the same reasoning as in [Kijewski and Judy, 1987]. An extension to the discrete FBP algorithm involves including sampling within the projections and angular sampling. For the continuous FBP algorithm, the DC-component of the NSD is always zero. By undersampling within the projections (i.e. along the detector array), this DC-component is generally non-zero. On the other hand, undersampling in the angular direction causes the NSD to become anisotropic even if it was predicted to be isotropic by the continuous FBP noise model. Hence the noise description for the discrete FBP algorithm which we will now explain, is more accurate especially when undersampling comes in to play. For our analysis, we will model every step of the discrete FBP algorithm discussed in Section 6.1.2. As a first step, we consider the effect of sampling along the detector array. According to the Nyquist-Shannon sampling theorem, the spectrum (6.30) is replicated at multiples of the sampling frequency:

$$S^{(1)}_{(r,\varphi)}(\alpha, \omega) = \psi_{(r,\varphi)}(\alpha) \sum_{l=-\infty}^{+\infty} |G(\omega - \omega_l)|^2 |\omega - \omega_l| \tag{6.34}$$

with $\omega_l = 2\pi l/(2a) = l\pi/a, l = 1, ..., L$, a is the sampling period along the detector array. As explained in Section 6.1.2, the backprojection algorithm does not use the projection samples directly because the required sampling

coordinates do not coincide with the usual sampling grid. Therefore, the first step is regridding using interpolation. Let $H(\omega)$ denote the Fourier transform of the interpolation kernel $h(t)$ (see Section 6.1.2), then the NSD is modified as follows:

$$S^{(2)}_{(r,\varphi)}(\alpha,\omega) = |H(\omega)|^2 S^{(1)}_{(r,\varphi)}(\alpha,\omega). \tag{6.35}$$

By the Fourier Slice Theorem, backprojecting from a discrete number of angles K replicates the power spectrum along slices through the frequency space origin at these angles. This only affects the angular component of the NSD:

$$\begin{aligned} S^{(3)}_{(r,\varphi)}(\alpha,\omega) &= |H(\omega)|^2 \left(\sum_{k=1}^{K} \psi_{(r,\varphi)}(\alpha)\, \delta(\alpha - \vartheta_k) \right) \\ &\quad \cdot \sum_{l=-\infty}^{+\infty} |G(\omega - \omega_l)|^2 |\omega - \omega_l| \end{aligned} \tag{6.36}$$

with ϑ_k the projection angle. In particular, we note that the aliasing effect causes the angular component of $S^{(3)}_{(r,\varphi)}(\alpha,\omega)$ to have a degraded orientation selectivity.

The last step incorporates the discrete representation of the final reconstructed image. This corresponds to convolving the spectrum in rectangular frequency coordinates with a two-dimensional Dirac pulse train (also called comb filter):

$$\mathrm{comb}(b\omega_x, b\omega_y) = \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} \delta\left(\omega_x - \frac{2\pi n}{b}, \omega_y - \frac{2\pi m}{b}\right) \tag{6.37}$$

with ω_x and ω_y respectively horizontal and vertical frequencies (i.e. $\omega_x = \omega\cos\alpha$, $\omega_y = \omega\sin\alpha$) and with b the sampling distance of the discrete representation (usually $b = a$).

6.5 The discrete NSD model and its relation to directional multiresolution representations

We have shown that for both the continuous and discrete FBP algorithms, the local NSD can be written in a polar-separable form (6.30). This expression has a radial component that consists of the FBP filter and an angular component $\psi_{(r,\varphi)}(\vartheta)$ that describes the noise power in the direction ϑ. Our goal is now to predict the NSD $S_{(r,\varphi)}(\vartheta,\omega)$ for each position in a CT image, reconstructed using the FBP algorithm. This involves noise estimation.

First, we remark that, although the radial component depends on the backprojection filter being used (which is considered to be known for our purposes), the angular component $\psi_{(r,\varphi)}(\vartheta)$ is an *unknown* continuous function of $\vartheta \in [0,\pi]$. Noise estimation then amounts to estimating this unknown function.

To arrive at a computational technique to estimate $\psi_{(r,\varphi)}(\vartheta)$ as a function of ϑ, we sample $\psi_{(r,\varphi)}(\vartheta)$ in the projection angles ϑ_k, i.e.:

$$\psi_{(r,\varphi)}(\vartheta_k) = \int_{-\infty}^{+\infty} \sigma^2(P(\vartheta_k,t))\, q^2(t - r\cos(\varphi - \vartheta_k))\,\mathrm{d}t. \quad (6.38)$$

This results in a high number of parameters to estimate: the number of projection angles K times the number of pixels in the image. For typical medical images, the dimensions are (at least) 512×512, and the number of projection angles is roughly $512\sqrt{2} \approx 724$. Fortunately, for most practical applications (e.g. noise analysis, signal detection, ...), a discrete NSD model with an *undersampled* number of projection angles is sufficient to obtain a good description of the local noise statistics. Therefore, we put forward the following NSD model:

$$S_{(r,\varphi)}(\vartheta,\omega) = \left(\sum_{k=1}^{K'} \psi_{(r,\varphi)}(\vartheta_k)\, f(\vartheta - \vartheta_k)\right) |\omega|\,|G(\omega)|^2 \quad (6.39)$$

where $f(\vartheta - \vartheta_k) \geq 0$ is a steerable function (see Section 2.3.1). This choice not only permits a variable number of analysis angles K', but also allows to accurately compute the NSD response in other angles $\vartheta \notin [\vartheta_1, ..., \vartheta_{K'}]$:

$$\psi_{(r,\varphi)}(\vartheta) = \sum_{k=1}^{K'} \psi_{(r,\varphi)}(\vartheta_k)\, b_k(\vartheta) \quad (6.40)$$

with $b_k(\vartheta)$ interpolation functions as explained in Section 2.3.1. For uniformly spaced projection angles ϑ_k, these interpolation functions are Dirichlet functions (2.37), which means that for $K' = K$, the NSD model is exploiting the bandlimitedness assumption of the projection data. As said before, this accuracy is not required, and we take $K' \ll K$. We will illustrate this on an example in Section 6.7. For the discrete NSD model, the local variance at position (r,φ) can be expressed as:

$$\begin{aligned} v(r,\varphi) &= \int_0^{\pi} \psi_{(r,\varphi)}(\vartheta)\,\mathrm{d}\vartheta \\ &= \sum_{k=1}^{K'} \psi_{(r,\varphi)}(\vartheta_k) \int_0^{\pi} b_k(\vartheta)\,\mathrm{d}\vartheta \\ &= \frac{1}{K'} \sum_{k=1}^{K'} \psi_{(r,\varphi)}(\vartheta_k) \end{aligned} \quad (6.41)$$

which is the average local noise power over all orientations. Once in a given point (r,φ) the local noise power in each orientation ϑ_k is estimated, one can obtain an estimate of the local variance in that point by averaging over contributions of $\psi_{(r,\varphi)}(\vartheta_k)$ in each direction. This is analogous to the continuous formula (6.33).

Now we turn to the actual parameter estimation of this statistical model. Suppose we are given a variance image $\sigma^2\left(P(\vartheta,t)\right)$, e.g. estimated from the projection data $P^m(\vartheta,t)$ (we will explain later how to obtain such a variance image), then we want to estimate the noise model parameters $\psi_{(r,\varphi)}\left(\vartheta_k\right), k = 1,...K'$. To do so, we start from (6.31) for continuous ϑ. By multiplying both sides of (6.31) with $f(\vartheta-\vartheta_k)$ and integrating over ϑ, we find:

$$\int_0^\pi \psi_{(r,\varphi)}\left(\vartheta\right) f(\vartheta-\vartheta_k)\mathrm{d}\vartheta \quad = \quad R_k(r,\varphi) \quad \text{with}$$

$$R_k(r,\varphi) = \int_{-\infty}^{+\infty} \mathrm{d}t \int_0^\pi \mathrm{d}\vartheta \sigma^2\left(P\left(\vartheta,t\right)\right) f\left(\vartheta-\vartheta_k\right) q^2\left(t-r\cos\left(\varphi-\vartheta\right)\right) \quad (6.42)$$

Next, we can substitute (6.40) into this equation, which results in a linear system of equations in K' unknowns $\psi_{(r,\varphi)}\left(\vartheta_l\right), l = 1,...,K'$:

$$\sum_{l=1}^{K'} \left(\mathbf{A}\right)_{k,l} \psi_{(r,\varphi)}\left(\vartheta_l\right) = R_k(r,\varphi) \quad \text{with}$$

$$\left(\mathbf{A}\right)_{k,l} = \int_0^\pi f\left(\vartheta-\vartheta_k\right) b_l\left(\vartheta\right) \mathrm{d}\vartheta.$$

Solving this system leads directly to an estimator for the noise power in the orientation ϑ_k:

$$\hat{\psi}_{(r,\varphi)}\left(\vartheta_k\right) = \sum_{l=1}^{K'} \left(\mathbf{A}^{-1}\right)_{k,l} R_l(r,\varphi) \quad (6.43)$$

The elements of the matrix $\mathbf{A}$ (and hence $\mathbf{A}^{-1}$) solely depend on the choice of steering function and interpolation function, and can be precomputed. For example, for uniformly spaced projection angles, $K' = 3$ and $f\left(\vartheta\right) = \cos^2\left(\vartheta\right)$, we find:

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{10\pi}{3} & -\frac{2\pi}{3} & -\frac{2\pi}{3} \\ -\frac{2\pi}{3} & \frac{10\pi}{3} & -\frac{2\pi}{3} \\ -\frac{2\pi}{3} & -\frac{2\pi}{3} & \frac{10\pi}{3} \end{pmatrix}. \quad (6.44)$$

Hence, to estimate the noise model parameters, it suffices to compute $R_k(r,\varphi)$ according to (6.42). Then, a simple linear combination of $R_k(r,\varphi)$ gives the noise power in all orientations ϑ_k. The implementation of this estimation method is fairly simple: first we recognize that (6.42) is in the same form as the FBP reconstruction formula (6.5). Consequently, $R_k(r,\varphi)$ is found by *backprojecting the weighted variance images* $\sigma^2\left(P\left(\vartheta,t\right)\right) f\left(\vartheta-\vartheta_k\right)$ *with a modified reconstruction filter* $q^2(t)$. As in [Zhu and Starlack, 2007], the same result was found for variance prediction (i.e. for estimating the *diagonal* elements of the local noise covariance matrix), our result can be considered to be a generalization of the variance prediction method from [Zhu and Starlack, 2007] to off-diagonal elements of the local noise covariance matrix.

Furthermore, in our approach, the factor $f(\vartheta-\vartheta_k)$ in (6.39) acts as a directional filter, which suggests the use of an directional multiresolution transform

for the practical computation of the noise model parameters. More specifically, if we analyze the reconstructed image using a set of filters defined in polar frequency coordinates:

$$H_{i,k}(\omega,\vartheta) = R_i(\omega)\sqrt{f(\vartheta-\vartheta_k)}, \tag{6.45}$$

where i is the multiresolution scale, then for a FBP reconstructed image, the NSD after filtering is given by:

$$S^{(i,k)}_{(r,\varphi)}(\vartheta,\omega) = \psi_{(r,\varphi)}(\vartheta_k)\left(f(\vartheta-\vartheta_k)\right)|\omega|\,|R_i(\omega)|^2\,|G(\omega)|^2,$$

which is also of the form (6.39). Then, $\psi_{(r,\varphi)}(\vartheta_k)$ represents the noise power at position (r,φ) for orientation k, which can alternatively be estimated from the multiresolution subband coefficients at scale i and orientation k. Hence, multiresolution transforms lend themselves well for estimating the noise model parameters.

6.6 CT noise characteristics

After estimating the parameters of the CT noise model (according to (6.43)), we can study some derived noise characteristics such as the degree of noise anisotropy and the noise streak orientation based on the estimates of the local noise power $\hat{\psi}_{(r,\varphi)}(\vartheta_k)$ in the orientation ϑ_k. We define the (local) degree of noise anisotropy as the normalized coefficient of variation of $\hat{\psi}_{(r,\varphi)}(\vartheta_k)$:

$$\xi(r,\varphi) = \frac{\sqrt{\frac{1}{K'-1}\sum_{k=1}^{K'}\left(\hat{\psi}_{(r,\varphi)}(\vartheta_k) - \frac{1}{K'}\sum_{l=1}^{K'}\hat{\psi}_{(r,\varphi)}(\vartheta_l)\right)^2}}{\sqrt{K'}\cdot\frac{1}{K'}\sum_{k=1}^{K'}\hat{\psi}_{(r,\varphi)}(\vartheta_k)} \tag{6.46}$$

which is the standard deviation of the samples $\left\{\hat{\psi}_{(r,\varphi)}(\vartheta_k), k=1,...,K'\right\}$ divided by their mean. The normalization constant $\frac{1}{\sqrt{K'}}$ is chosen such that the maximal anisotropy 1 is reached if for exactly one angle ϑ_l the local noise power is positive ($\hat{\psi}_{(r,\varphi)}(\vartheta_l) > 0$) and for all other angles, the local noise power is zero ($\hat{\psi}_{(r,\varphi)}(\vartheta_k) = 0,\ k \neq l$). The noise anisotropy is 0, if all $\hat{\psi}_{(r,\varphi)}(\vartheta_k) = \psi_0$ are equal, with ψ_0 a given constant. If the noise anisotropy is large, e.g. $\xi(r,\varphi) \approx 1$, then some orientations θ_k will contribute more to the NSD than other orientations. In this case, it is useful to define the noise streak orientation as:

$$\theta_0(r,\varphi) = \frac{\pi}{2} + \arg\max_{\vartheta_k}\hat{\psi}_{(r,\varphi)}(\vartheta_k), \tag{6.47}$$

which is the direction contributing most to the variance. The constant $\pi/2$ is added to compensate the rotation added by the Fourier transform (because the Fourier transform of a line impulse is a line impulse that is rotated 90°). Figure 6.7 illustrates the derived noise characteristics. The noise spectra are

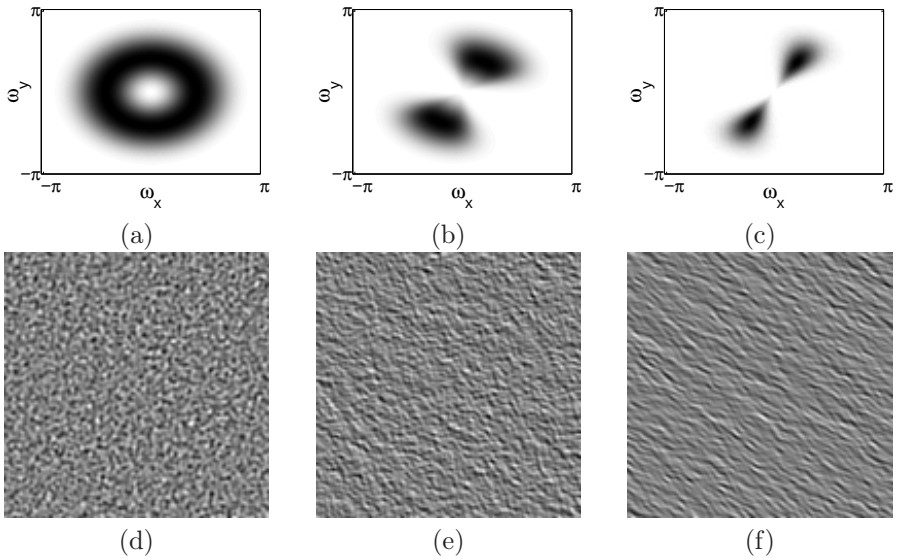


Figure 6.7: Illustration of the derived noise characteristics ($K = 11$) (a) Isotropic NSD: $\psi_{(r,\varphi)} = [1, 1, ..., 1]$, $\xi(r,\varphi) = 0$, $\theta_0(r,\varphi) = \textit{undetermined}$, (b) Anisotropic NSD: $\psi_{(r,\varphi)} = [0, 0, 1, 0, 1, 0, 1, 0, 0, ..., 0]$, $\xi(r,\varphi) \approx 0.52$, $\theta_0(r,\varphi) = \frac{17\pi}{22}$ (c) Anisotropic NSD: $\psi_{(r,\varphi)} = [0, 0, 0, 0, 1, 0, 0, ..., 0]$, $\xi(r,\varphi) = 1$, $\theta_0(r,\varphi) = \frac{17\pi}{22}$, (d)-(f): artificial stationary Gaussian noise generated for the NSDs in (a)-(c) (see text).

synthesized according to the discrete NSD model (Section 6.5), with parameters $\psi_{(r,\varphi)}$ as given in the figure caption. To visually show the anisotropy we also generated artificial stationary correlated Gaussian noise, for each of the NSD models. Figure 6.7(f) clearly shows that the noise contains line structures (called streaking artifacts), while the isotropic noise in Figure 6.7(d) does not show any orientation. In the next section, we will illustrate the estimation of the noise characteristics for a synthetic image.

6.7 Experimental results

Noise estimation

As a practical example of the presented noise model for FBP reconstructed images, we test the discrete NSD model on an artificial software phantom image. Figure 6.8(a) shows an image of nine small disks with a high constant intensity, corrupted with Poisson noise in the projection space as described in Section 6.3. The disks have increasing radii and are located at equal distances from each other. Clearly, the visibility of the small disks is significantly reduced by the noise, in practice this could mean that small lesions in the image are

missed in the diagnostic process. To fully extract the available information in the image, studying the noise characteristics is very important.

The CT image Figure 6.8(a) contains many noise *streaking artifacts*. These artifacts are mostly concentrated near the disks. This is because the projection data noise variance is maximal for projection data with high intensities (in fact, the streaking artifacts are superpositions of lines that intersect with the disks). As we explained in the previous sections, streaking artifacts locally have an asymmetric NSD. The degree of "asymmetry" can be measured by the anisotropy measure (6.46). Figure 6.8(b) shows this anisotropy measure applied to the "ideal" standard deviation of the projection data $\sigma\left(P(\vartheta,t)\right)$, obtained through (6.19), taking all projection angles into account ($K'=K$). The figure reveals that the noise anisotropy is maximal for lines that intersect with the disks, as predicted. On the other hand, inside the nine disks, the noise anisotropy is very low.

Next, we computed the noise dominant streaking orientation (6.47) for each position in the image (Figure 6.8(c)) and the local variance according to equation (6.41) (Figure 6.8(d),(f)). The predicted noise streaking orientations correspond well to the visual observations in Figure 6.8(a). Of course, the noise streaking orientation has no meaning in regions where the noise is isotropic. This result predicts that the local noise variance (Figure 6.8(d)) is maximal inside the disks and significantly lower outside.

To validate the correctness of the variance estimation, we also estimated the local variance using Monte Carlo simulations (averaged over 1000 runs of the FBP reconstruction algorithm). The local variance estimated using the Monte Carlo simulations is shown in Figure 6.8(f). The predicted local variance from Figure 6.8(d) agrees very well with the local variance estimated using Monte Carlo simulations from Figure 6.8(f). We conclude that the presented NSD model accurately predicts the local noise PSD.

We also tested the local variance estimation on the phantom image by using an undersampled number of projection angles ($K'<K$) for the local NSD characterization. Visual results are given in Figure 6.9 for $K'=61$ and $K'=31$. In particular, the visual difference between Figure 6.9(a) and Figure 6.8(d) is very small, compared to a huge reduction of a factor $362/61\approx 5.93$ of the number of noise model parameters. We computed the estimation error for an increasing number of projection angles from $K'=31$ to $K'=362$. The MSE of the local variance estimation for the phantom image is depicted in Figure 6.10. The MSE drops significantly: for $K'=61$, the MSE is approximately 10% of the MSE for $K'=31$. This result suggests that the NSD can be well described by local noise powers $\psi_{(r,\varphi)}\left(\vartheta_k\right)$ for relatively small number of orientations K'.

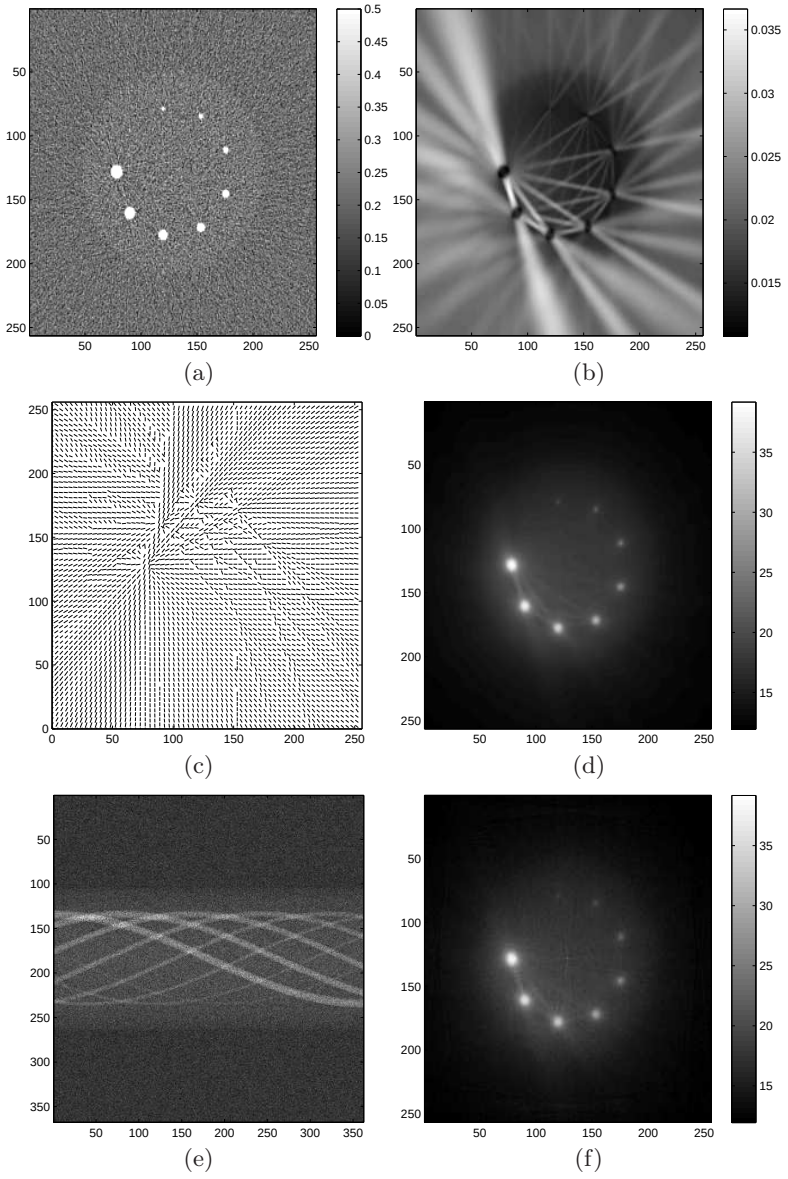


Figure 6.8: Estimation of the CT noise characteristics. (a) Artificial CT image with known $\sigma^2(P(\alpha,t))$, reconstructed from $K = 362$ projections, (b) estimated noise anisotropy (6.46), (c) estimated dominant noise streak orientation, (d) local noise variance, directly estimated from Figure 6.8(a) for $K' = K$, (e) noisy projection data $P^m(\theta,t)$; (f) local noise variance estimated using Monte Carlo simulations (averaged over 1000 runs of the FBP reconstruction algorithm).

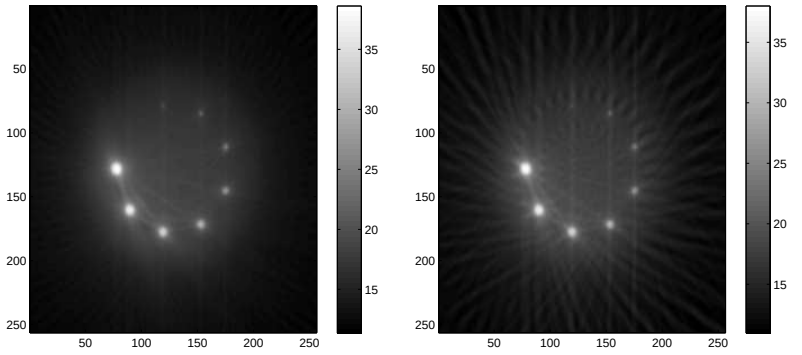


Figure 6.9: Estimation of the local noise variance using an undersampled number of projection angles. (a) for $K' = 61$, (b) for $K' = 31$.

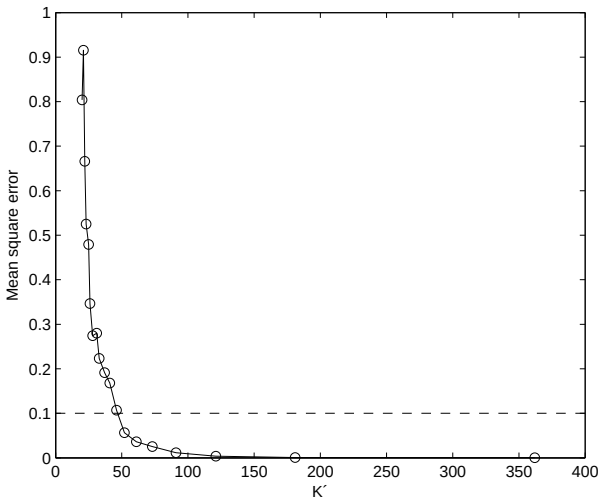
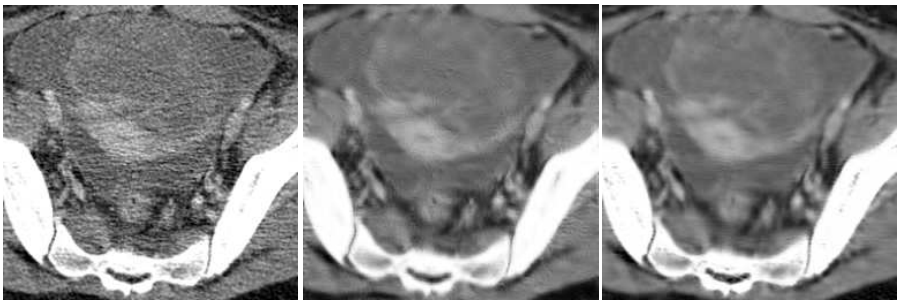


Figure 6.10: Estimation error (MSE) as function of the number of projection angles K'.

Application to image denoising

As a second example, we apply the presented noise model to image denoising. Exact details of this technique will be covered by a European patent application, which is currently in preparation. Therefore, the details will be published in our later work. Preliminary visual results are given in Figure 6.11. Using the CT noise model from this chapter, the denoising technique can even better be adapted to the spatially variant noise. In particular, there is a less stringent local stationarity requirement, which leads to less remaining noise artifacts in the processed image. Our denoising technique using the specialized



(a) (b) (c)

Figure 6.11: Denoising example using two different noise models for CT images: (a) the CT image from Figure 4.12, (b) denoised using the spatially variant noise model from 4.3.2, (c) denoised using the CT noise model presented in this chapter.

CT noise model is also much faster in computation time: while the technique from Chapter 4 takes 143 s to denoise a 256×256 image, this technique only requires 300 ms, which permits processing of large data sets (full CT image volumes) in a relatively short time period.

6.8 Conclusion

In this chapter we have shown that signal-dependent measurement noise leads to non-stationary and anisotropic noise after filtered backprojection, both for the continuous and discrete FBP algorithm. We presented a novel spatially variant noise model to describe the position-dependent and orientation-dependent properties of CT noise obtained after reconstruction. Next, we proposed a discrete NSD model that allows efficient estimation of the noise model parameters from a given image. We defined a number of local measures that characterize the non-stationary noise properties, including the local noise anisotropy, the dominant streaking orientation. Our preliminary result for an artificial image indicates that the model describes the spatially variant properties of the noise very well. A direct application of the presented theory is to increase the SNR of low-dose CT images through image restoration, but the wider range of applicability is in the field of computer aided diagnosis.

Contributions from this chapter have been presented in [Goossens et al., 2007a, Goossens et al., 2008b]. One journal manuscript has been submitted for publication [Goossens et al., 2010d]. On the application of the proposed model to CT noise reduction (which is not presented here), one European patent application is pending.

7

Models for measuring medical image quality

In this chapter, we will discuss a number of models that can be used for assessing medical image quality. In order to explain what "quality" actually means, we first remark that medical image quality is *entirely* different from (general) image quality:

- In (non-medical) image restoration, the goal is to produce images that are "aestetically pleasing," which often relies on a subjective impression: one algorithm is better than the other algorithm if it creates images that "look better". Looking better is related to the visibility of errors (or artifacts) in the image and is strongly influenced by the properties of the human visual system (HVS). To deal with the subjectiveness, some quality measures such as Mean Squared Error (MSE), Peak-Signal-To-Noise-Ratio are often used. Because it has been found that these measures not always correlate well with the human subjective perception of image quality (because certain types or artifacts are not taken into account), in the last decades, many researchers have incorporated properties of the HVS into the image quality metric. An example of such a metric is the Structural-Similarity-Index Metric [Wang et al., 2004].

- In medical image processing, the goal is *not* to create visually pleasing images: a medical image has a specific purpose (e.g. to allow the physician to diagnose a disease). To objectively assess medical image quality, first we must specify the task and next we must determine quantitatively how well the task is performed [Barrett, 1990]. The tasks being considered for quality assessment are detection tasks, in which abnormalities in images (e.g. tumors, vein calcifications, lesions...) are being detected. Image quality can then be expressed objectively in terms of the detectability of abnormalities.

Image quality assessment is not only useful to determine how well image processing algorithms perform the task for which they are designed (e.g. image restoration or enhancement techniques), the metrics can also be used to optimize various settings and parameters of these algorithms. For example, in denoising, often a trade-off between detail preservation and noise removal needs to be made. In Section 5.2.3 this resulted in a threshold parameter that defines the "signal of interest." Here, a quality metric can be used to determine which value to select for this parameter. From this perspective, image quality assessment is essential for designing good image processing and reconstruction algorithms. This also holds on a non-algorithmic level: for example, good image quality metrics allow display development engineers to evaluate alternative technology choices, e.g. concerning the type of backlights, glass and the LCD panels being used, before building and testing the devices.

Traditionally, determining the image quality of medical images has often been done by a panel of *human observers*: a small group of experienced physicians is asked to do a small clinical study where the physicians are asked to make a binary decision: an abnormality is either present in the diagnostic image, or not. Because it is a very time-consuming and expensive process, mathematical model observers [Barrett et al., 1998, Barrett and Myers, 2004, Gallas and Barrett, 2003] have been developed. These models predict the performance of human observers doing the same task and may eventually replace humans in quality assessment tasks.

At present, several studies indicate that there is in general a gap between the performances of model observers and human observers. In our opinion, this gap mainly stems from 1) the complexity of the problem and our incomplete knowledge of the processing in the HVS and 2) the simplifying assumptions made to arrive at practical processing algorithms.

In many studies in the literature, model observers have exact knowledge of the abnormality shape and background *statistics*. For example, abnormalities in medical images can range from very subtle to obvious. Furthermore, the size of the abnormalities can vary. While many model observers assume that the exact characteristics of the abnormality are known in advance (a training phase is often used for this), in clinical practice, physicians do not have this knowledge. This inherently leads to discrepancies between model observers and human observers, in the sense that model observers generally *outperform* human observers.

The main contribution of this chapter is the extension of two types of model observers, the ideal observer (IO) and the channelized Hotelling observer (CHO), *to signal-known-statistically (SKS) tasks*, in which the *uncertainty* with respect to the abnormality (further called signal) is modeled. This way, the new model observers are "handicapped" compared to model observers that have full information about the signal, and potentially behave more similar to humans. We study four causes of signal uncertainty: uncertainty with respect to signal strength, orientation, size and location. The cases will show how model observers can be defined that can deal with these causes of uncertainty. The

most striking result is that model observers that perform the signal detection task in an optimal sense (which will be defined later) through linear dimension reduction, need to make use of *steerable* multiresolution representations. This also makes the connection of this topic with the other Chapters of this dissertation. This work has already resulted into one conference paper [Goossens et al., 2010f] and one journal paper is in preparation [Goossens et al., 2010e].

A second research direction to bridge the performance gap is to include temporal information into the CHO model, because nowadays medical images are often viewed in stack-browsing mode (i.e. slices of a medical volume are shown sequentially). In collaboration with ir. Ljiljana Platiša (Ghent University) and with the American Food and Drug Administration (FDA), this research track resulted in the development of multi-slice CHO models in [Platiša et al., 2009a, Platiša et al., 2009b, Platiša et al., 2010a, Platiša et al., 2010b].

In the first part of this chapter (Section 7.1), we briefly review existing model observers. We present our novel extension to model observers for SKS detection tasks in Section 7.2 and Section 7.3; this part of the work have been contributed by the author. Experimental results are given in Section 7.4. Finally, we briefly present the multi-slice channelized Hotelling observer models in Section 7.5. These models have been worked out by ir. Ljiljana Platiša, in cooperation with dr. Ewout Vansteenkiste, prof. W. Philips, dr. Subok Park (FDA), dr. Aldo Badano (FDA), dr. Brandon Gallas (FDA) and the author.

The work presented here has been performed within the context of the IBBT-MEVIC[1] project. The aim of "Medical Virtual Imaging Chain" (MEVIC) was to simulate the complete imaging chain, from image acquisition (e.g. CT or MRI) up to image processing, medical displays and eventually the human observer.

7.1 Existing model observers for medical image quality assessment

As said before, for medical purposes image quality is defined in terms of how well a certain detection task can be performed. Clinically relevant tasks are, e.g., tumor detection, bone metastatis and vein calcification detection in digital medical images. For the detection tasks, a trade-off between the probability of true positive detections (i.e. the true positive rate) and the probability of false positive detections (i.e. the false positive rate) must be made. This trade-off is quantified in the receiver operating characteristic (ROC), and useful measures are the *area under the ROC curve* (AUC) and *detection Signal-To-Noise-Ratio* (SNR) [Barrett, 1990].

One important "theoretical" numerical observer is the Bayesian Ideal Observer (IO) [Barrett, 1990, Barrett et al., 1995, Barrett et al., 1998, Barrett and Myers, 2004], which provides an upper bound of the detection performance in terms of the ROC, because it makes use of all available information present in

[1] `http://www.ibbt.be/en/project/mevic`.

the images or known about the images. The IO even theoretically outperforms humans because it is not affected by imperfections in the HVS. Therefore it can be argumented that the IO objectively measures the quality of medical images [Barrett, 1990]. However, we call this observer a "theoretical" observer because the IO is impractical and often impossible to implement, because the observer needs the specification of probability models for both healthy and diseased images [Gallas and Barrett, 2003]. For medical images, this is a very challenging statistical modeling task. Hence practically, these models are not available.

To alleviate this problem, the channelized Hotelling observer (CHO) was introduced. It was inspired by psychophysical evidence of the HVS making use of frequency selective channels [Gallas and Barrett, 2003]. Namely, studies from [Hubel and Wiesel, 1959] in the 1950s and 1960s demonstrated that from the responses of simple neural cells in the primary visual cortex (also known as V1) of animals, local orientation and contrast of image features could be distinguished. Later, it was found that other features, such as color, spatial frequency, direction and motion are detected similarly, using single or complex neural cells.

Based on these studies, a fairly well-agreed-on model for the V1 has emerged [Olshausen and Field, 2005]. This model computes a linearly weighted sum of input signals over space and time (typically using Gabor-like functions). Subsequently, the weighted sum is either normalized by neighboring neuron responses or passed through a pointwise nonlinear function (see Figure 7.1(a)). The CHO applies a similar principle, in which the input image is first linearly projected onto a set of basis functions (called *channels*), thereby performing a dimensionality reduction. Next, a decision of signal presence or absence is made based on the obtained projection coefficients (further called *channel responses*).[2]

The CHO model is illustrated in Figure 7.1(b). Some authors have attempted to incorporate nonlinearities or other characteristics of the HVS in the CHO model as well: in [Zhang et al., 2006] it is found that a CHO model with a number of nonlinear components predicts the human performance slightly better than a linear CHO, although they found that the impact on medical image quality evaluation in general is minimal. In general, mostly *linear* CHO models are used.

Even though the CHO model has brought new interesting insights as predictor of human performance (see e.g. [Gilland et al., 2004, Gifford et al., 2005, Shidahara et al., 2006]), several studies have shown that human observers are *inefficient* compared to the IO or CHO [Park et al., 2005]. As already mentioned, one important source of the inefficiency is the *intrinsinc uncertainty* about the signal characteristics of the human observer, such as the contrast, size, shape, orientation and location of the signal [Park et al., 2005]. In many CHO studies,

[2]We remark that the goal of the CHO is not in the first place to *mimick* the HVS, as this task is currently too complex. Instead, the goal is to approximate the performance of a human observer performing the same task. Therefore, knowledge of the HVS helps to explain several discrepancies in the results of model and human observer experiments.

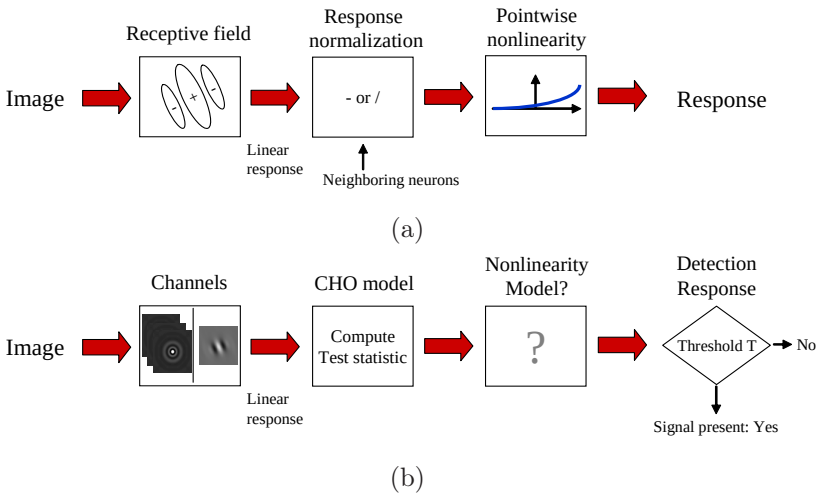


Figure 7.1: (a) Model for the V1 [Olshausen and Field, 2005], (b) The channelized Hotelling observer [Gallas and Barrett, 2003].

it is assumed that exact information about the background images and signals is available, the so-called background-known-exactly/signal-known-exactly or (BKE/SKE) detection task. The model observer then exactly knowns "what to look for". Although this permits a straightforward design of equivalent human observer studies to validate the model observers [Eckstein and Abbey, 2001], this assumption is truly not valid for humans, as physicians do not know precisely in advance the location, shape and contrast of the abnormality they are looking for. A possible solution would be to additionally reveal this information to the physicians (in the form of markers), however, the risk is that the model observers would eventually be tuned for a clinically non-relevant task. There are also other sources of inefficiency of human observers, e.g., neural receptor sampling errors, randomness of neural responses and loss of information during neural transmission [Lu and Dosher, 1999]. Sometimes the inefficiency is just due to visual fatigue of the human observer [Krupinski and Berbaum, 2009].

Hence image quality does not only depend on the images and on the task, but also on the observer that performs the task [Barrett, 1990]. To design mathematical model observers that serve as good predictors of the human observer, it is crucial to learn about the discrepancies between these two types of observers. Knowledge of these discrepancies allows us to better fine-tune medical imaging systems, and may serve in the future as an computer assisted diagnosis tool to aid physicians. As a starting point, in Section 7.2 we will design CHOs for SKS tasks, which incorporate intrinsic uncertainty with respect to the signal characteristics.

In the remainder of this section, we will first review a number of background and signal models that are used for the observer studies. Next, we will briefly explain the ideal observer and the channelized Hotelling observer.

7.1.1 Signal detection theory and the ideal observer

Given an image, the task of the model observer is to decide whether a certain signal is present in the image or not. The presence of the signal could indicate an abnormality. According to classical decision theory, this task is a binary classification task with two hypotheses: the signal is absent (H_0) or the signal is present (H_1). Let $\mathbf{b}$ denote a vector of intensities of the random-background. In this notation, the images are column-stacked into vectors (such as in raster scanning). $\mathbf{x}$ and $\mathbf{y}$ respectively denote the *known* signal and the image. In the following, we will assume an additive relationship between the background image $\mathbf{b}$ and the signal $\mathbf{x}$. Although the signal detection theory also holds for non-additive relationships between the background and signal, this assumption is commonly used in literature because it facilitates analytical tractability of the model observers' performance. The hypotheses are formulated as follows:

$$\begin{cases} \mathbf{y} = \mathbf{b} & (H_0) \\ \mathbf{y} = \mathbf{b} + \mathbf{x} & (H_1) \end{cases}$$

Under H_0, the observed image is only a background image, while under H_1, the observed image also contains a signal $\mathbf{x}$. To perform the decision task, we must also specify the observer and a figure of merit for measuring the observer performance (such as AUC and SNR). One candidate observer is the Bayesian ideal observer, which uses the following test statistic:

$$\lambda(\mathbf{y}) = \frac{f_{\mathbf{Y}|H}(\mathbf{y}|H_1)}{f_{\mathbf{Y}|H}(\mathbf{y}|H_0)} \tag{7.1}$$

where $f_{\mathbf{Y}}(\mathbf{y}|H_0)$ and $f_{\mathbf{Y}}(\mathbf{y}|H_1)$ are conditional probability distributions of the data $\mathbf{y}$, under the hypotheses of respectively signal absence and signal presence. Based on this test statistic, the ideal observer decides whether a signal is present:

$$\begin{cases} H_0 & \text{if } \lambda(\mathbf{y}) < T \\ H_1 & \text{if } \lambda(\mathbf{y}) \geq T \end{cases} \tag{7.2}$$

with T a predefined threshold. The performance of an observer can be quantified through the true positive rate (TPR) and false positive rate (FPR), which are defined as follows:

$$\text{TPR} = \text{P}\left(\hat{H}_1|H_1\right), \tag{7.3}$$

$$\text{FPR} = \text{P}\left(\hat{H}_1|H_0\right), \tag{7.4}$$

where a true positive detection is made when the signal is correctly identified, while a false positive detection denotes an errorneous detection of the signal. By changing the parameter T, a trade-off between both probabilities can be made. The ROC is a graphical plot of TPR as a function of FPR. Optimizing the observer performance is done by minimizing FPR for a given TPR, or vice

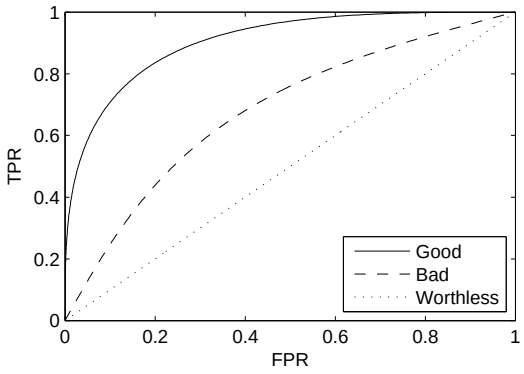


Figure 7.2: Three Receiver Operating Characteristics (ROC) corresponding to different detector performances. The higher the area under the ROC (AUC), the better the performance.

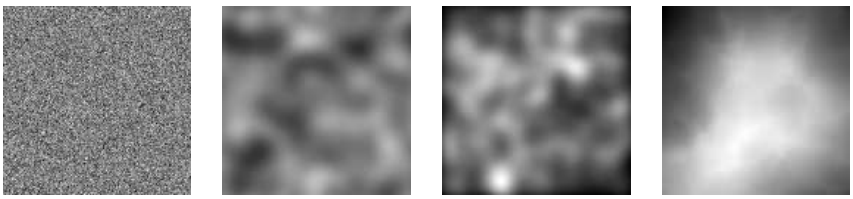
versa. Consequently, the observer performance can be expressed by means of the area under the ROC curve (AUC), see Figure 7.2.

For objective quality assessment, it is advocated to use the Bayesian ideal observer whenever possible [Park et al., 2009]. Unfortunately, the performance of this observer can not be computed easily, because exact probability density functions of the background and the signal are required. For complex medical images as encountered in clinical practice, this is a very difficult task. To compute the test statistic of the ideal observer anyway, several restrictions are imposed to the background and signal pdfs.

7.1.2 Background models

The backgrounds are designed to mimick clinically realistic images in a practical manner, such that different observer strategies can easily be computed. In the literature, both uniform and spatially inhomogeneous backgrounds are being considered (see Figure 7.3):

- *White Gaussian background (WGB)*: being one of the most simple statistical models for background images. Obviously this background model can account for measurement noise up to certain extent, but is certainly not clinically realistic.
- *Correlated Gaussian background (CGB)*: this type of background is equivalent to the correlated stationary Gaussian noise from Chapter 4. This model is again not encountered in clinical practice, yet it has the advantage that the ideal observer is calculable.
- *Lumpy background (LB)* [Rolland and Barrett, 1992]: is produced by placing a random number of Gaussian functions (called lumps) at random locations in the image. The locations of the lumps are often uniformly



(a) (b) (c) (d)

Figure 7.3: Examples of background images. (a) White Gaussian noise background, (b) Correlated Gaussian noise background, (c) Lumpy background [Rolland and Barrett, 1992], (d) Clustered lumpy background [Bochud et al., 1999].

distributed over the image, while the number of lumps is drawn from a Poisson probability distribution.

- *Clustered lumpy background (CLB)* [Bochud et al., 1999, Castella et al., 2009]: is a spatially inhomogeous model that has been shown to mimmic mammographical anatomical structures. The creation of the background is similar as for LB, the main difference is that *clusters* of lumps are formed, which are located close to each other. Moreover, the lump profile in this background model is a generalized Gaussian functions instead of a Gaussian function.

In this chapter, to easily compare the performance of different observers, we will mainly consider Gaussian lumpy backgrounds,

$$\mathbf{b} \sim N(\mu_b, \mathbf{C}_b), \tag{7.5}$$

with mean μ_b and covariance matrix $\mathbf{C}_b$. Realistic values for the parameters μ_b and $\mathbf{C}_b$ can be firstly estimated from a set of images, typically using an assumption of spatial stationarity.[3] This comes down to modeling the background through its second order statistics (see Chapter 3). In [Barrett et al., 1995], it is found that for the Bayesian ideal observer, the AUC estimated *under the assumption that the background is Gaussian* is the first order approximation for the AUC of the *true* ideal observer. Hence the approximation gives us *an idea* of the performance of the ideal observer for more complex backgrounds.

Nevertheless, for more realistic quality assessment tasks, more complex background images, such as actual clinical medical images are desired. We will see in Section 7.1.4 that the CHO will bring a possible solution here.

7.1.3 Signal models

We model the signal by a fixed profile with known or unknown location and shape parameters. One specific example of such a signal is an elliptical Gaussian

[3]Most medical images are *not* spatially stationary. However, the assumption imposes a certain structure to $\mathbf{C}_b$ with a much smaller number of free parameters, which can be advantageous in most cases.

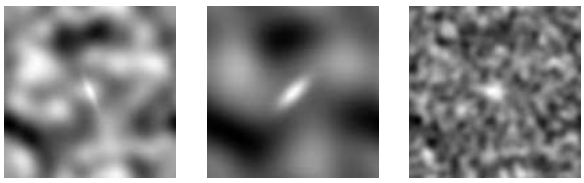


Figure 7.4: Sample background images with asymmetric signals inserted in the center of the image, for Gaussian lumpy backgrounds of different parameter setups. Possible sources of signal uncertainty are: size and shape of the signal, the location and the orientation.

profile:

$$[\mathbf{x}]_{\mathbf{p}} = a \exp\left(- \left(\mathbf{A}_\vartheta \mathbf{p} - \mathbf{q}\right)^T \mathbf{D}^{-1} \left(\mathbf{A}_\vartheta \mathbf{p} - \mathbf{q}\right)\right) \tag{7.6}$$

where a is the signal amplitude, $\mathbf{p}$ is a 2D vector denoting the spatial position, $\mathbf{q}$ is the position of the center of the profile and $\mathbf{D} = \begin{pmatrix} 2b\sigma^2 & 0 \\ 0 & 2\sigma^2 \end{pmatrix}$ is a diagonal matrix with b a fixed constant and σ a scale parameter. $\mathbf{A}_\vartheta$ is a 2D rotation matrix:

$$\mathbf{A}_\vartheta = \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix}. \tag{7.7}$$

In case the diagonal matrix $\mathbf{D}$ has equal diagonal elements, the signal is rotationally symmetric (or simply, symmetric) and invariant under rotations $\mathbf{A}_\vartheta$. Otherwise, the orientation of the main axis is ϑ.

We say that the detection task is signal-known-statistically (SKS) [Park et al., 2005], if at least one of the parameters $b, \sigma, \mathbf{q}, \vartheta$ is not known in advance but specified by a probability density function. Otherwise, all parameters are known which corresponds to a signal-known-exactly (SKE) task.

For notational convenience, we will denote the vector of unknown signal parameters by $\alpha = \{b, \sigma, \mathbf{q}, \vartheta\}$ in the remainder of this chapter, the set of all possible parameter values as Ω (i.e. $\alpha \in \Omega$) and we will explicitly show the dependency of the signal $\mathbf{x}$ on the unknown parameters as $\mathbf{x}_\alpha$.

In Figure 7.4, some sample background images and signals are shown, where the signal is given by (7.6). It is clear that such signals are not very clinically realistic. However, what follows is not restricted to this choice of signals: most important is the parametrization in amplitude, scale and orientation. Therefore, the elliptical Gaussian profile serves as a study object.

7.1.4 Channelized Hotelling observers

Based on the signal and background model above, it becomes possible to compute the likelihood ratio (LRT) of the ideal observer (7.1). However, for more complex backgrounds or signals, the ideal LRT is much more complicated or even unknown. A practical solution is brought by the CHO, which makes use

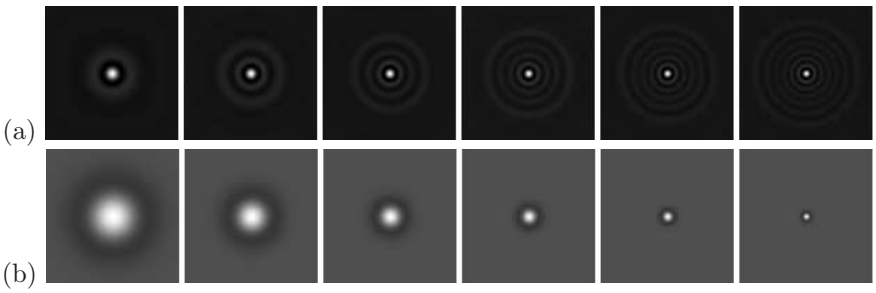


Figure 7.5: Images of rotationally symmetric channels. (a) Laguerre-Gauss channels [Gallas and Barrett, 2003], (b) DDOG channels [Abbey and Barrett, 2001].

of linear dimensionality reduction [Gallas and Barrett, 2003, Myers and Barrett, 1987]. The images are linearly projected onto a small set of channels, and an ideal linear observer[4] (also called Hotelling observer) is applied to the dimension reduced vector:

$$\mathbf{y}' = \mathbf{U}^T \mathbf{y} \tag{7.8}$$

where $\mathbf{y}'$ is the dimension reduced observation vector and where $\mathbf{U}$ is an $N \times S$ projection matrix. Here, N is the number of pixels in the image, while S is the number of channels being used.

Crucial here is the selection of the channels (and channel parameters), such that the detection performance of the CHO does not deviate too much from the performance of either the IO or the human observer. A second consideration is that the same set of channels should be applicable in many situations [Gallas and Barrett, 2003]. For these reasons, rotationally symmetric Laguerre-Gauss channels [Gallas and Barrett, 2003] (Figure 7.5(a)) are often used because of their efficiency, and Dense Difference-of-Gaussian (DDOG) channels [Abbey and Barrett, 2001] as a model for the spatial-frequency selectivity in the human visual system (Figure 7.5b)).

In the following, we will consider a number of signal-absent/signal-present images $\mathbf{y}$. As we only work with ensemble statistics of these images, we will denote the sample ensemble mean under the hypothesis of signal-absent and signal-present respectively as $\langle \mathbf{y}|H_0\rangle$ and $\langle \mathbf{y}|H_1\rangle$.

An implementation of the CHO consists of two phases [Gallas and Barrett, 2003]:

1. *Training phase*: in this phase the model parameters of the CHO are trained for the specific task that one has in mind (e.g. lesion detection), under 'ideal' circumstances (no distortion applied to the images). The training method is Linear Discriminant Analysis (LDA). For the training set, the CHO processes a relatively large number of signal-absent/signal-present pairs of images. More specifically, the signal is estimated as:

$$\hat{\mathbf{x}} = \langle \mathbf{y}'|H_1\rangle - \langle \mathbf{y}'|H_0\rangle \tag{7.9}$$

[4] An ideal linear observer is an ideal observer restricted to linear test statistics.

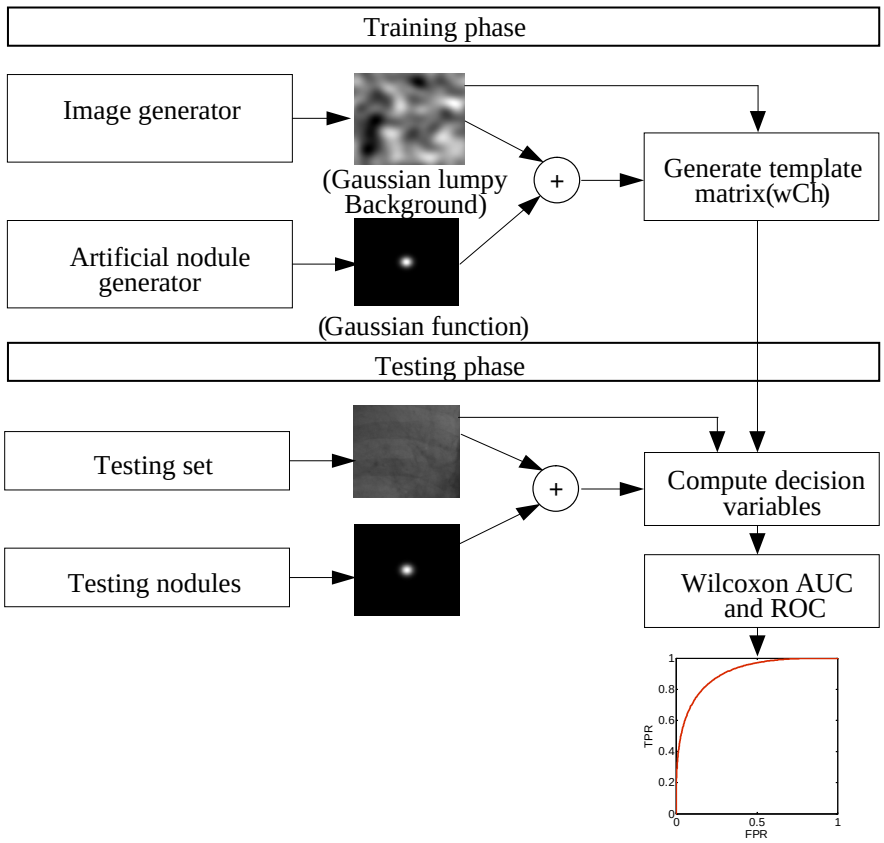


Figure 7.6: The channelized Hotelling observer in action.

which is the difference between two sample averages (under both hypotheses) over all images from the training phase. According to LDA, the intra-class covariance matrices are assumed to be equal, and are estimated as:

$$\begin{aligned}\hat{\mathbf{C}}_b &= \frac{1}{2}\left\langle \left(\mathbf{y}' - \langle \mathbf{y}'|H_0\rangle\right)\left(\mathbf{y}' - \langle \mathbf{y}'|H_0\rangle\right)^T |H_0 \right\rangle + \\ &\quad \frac{1}{2}\left\langle \left(\mathbf{y}' - \langle \mathbf{y}'|H_1\rangle\right)\left(\mathbf{y}' - \langle \mathbf{y}'|H_1\rangle\right)^T |H_1 \right\rangle . \end{aligned} \tag{7.10}$$

For linear discriminant analysis, the test statistic (see (7.2)) is linear:

$$\lambda(\mathbf{y}') = \mathbf{w}_{\mathrm{CHO}}^T \mathbf{y}' \tag{7.11}$$

where $\mathbf{w}_{\mathrm{CHO}}$ is the channel template matrix, which can be estimated as follows:

$$\widehat{\mathbf{w}_{\mathrm{CHO}}} = \left(\hat{\mathbf{C}}_b\right)^{-1} \hat{\mathbf{x}}. \tag{7.12}$$

The channel template matrix is used in the testing phase, in order to draw a conclusion about signal presence/absence. Remark that, for computing the channel template matrix, the model observer needs to know which images contain (H_1) or do not contain (H_0) the signal. In case real medical images are used, this information can be extracted from annotations made by a physician.

2. *Testing phase*: here, the trained CHO model is applied to a test set of data, which again consists of signal-absent/signal-present pairs of images. The most important difference with the training phase is that the model observer now does not have the preknowledge about whether a signal is present in the image. Hence the model observer effectively needs to perform a detection. A signal is detected (H_1) if the test statistic (7.11) is larger than a predefined threshold, otherwise the signal is assumed to be absent (H_0). The ROC is estimated using the Mann–Whitney–Wilcoxon test [Wilcoxon, 1945, Mann and Whitney, 1947] and subsequently the AUC detection performance is computed from the ROC.

The workflow of the CHO is summarized in Figure 7.6. Perhaps one of the most interesting features of the CHO is that detection performance can be computed for arbitrary images with arbitrary signals. For example, in a previous study [Platiša et al., 2009c], we used real radiographic images of the chest with simulated lung nodules (see Figure 7.7). For these images, a good agreement was found between the CHO and the human observers.
However, there are also a number of shortcomings related to the CHO:

- First, the linear discriminant analysis assumes that the conditional probability density functions $f_{\mathbf{y}'}(\mathbf{y}'|H_0)$ and $f_{\mathbf{y}'}(\mathbf{y}'|H_1)$ are Gaussian. As some of the channels are in fact filters with band-pass characteristics (see Figure 7.8(b)), the Gaussianity assumption is *inconsistent* with the highly kurtotic behavior of the filter responses that we studied in Chapter 3. An interesting future research topic is to derive test statistics for the probability density distributions from Chapter 3 and to compare them with human observer experiments. In the remainder of this chapter we will stick to the Gaussian distributions, as these distributions are commonly used within the context of medical image quality assessment.

- Second, the estimated AUC-values exhibit statistical fluctuations due to the limited testing and training sets being used. To have a sufficiently low AUC variance (which is necessary for comparing the image quality between two display systems), thousands of images are needed. Gathering image data is also a time-consuming and often expensive task. Moreover, study designs used in practice often involve multiple physicians (readers), and the variability between the different observers needs to be taken into account. A common solution that permits to use a limited number of image data, is to use multiple reader/multiple case (MRMC) studies in which multiple CHO models are trained on subsets of the complete

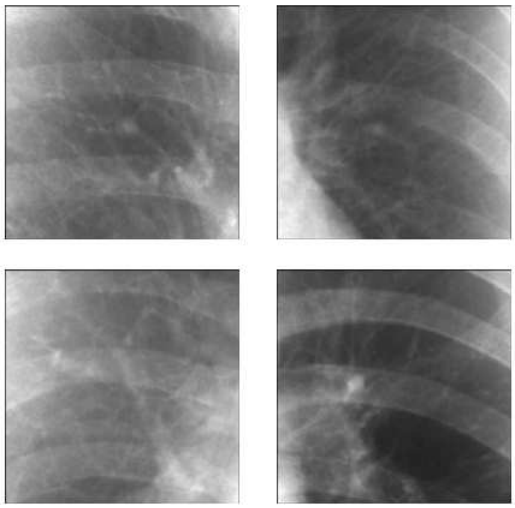


Figure 7.7: Chest lung nodules from the database [Shiraishi et al., 2000]. (*left*) images with nodule inserted in the center of the image (the nodule is very subtle in the image at the bottom), (*right*) images without nodule.

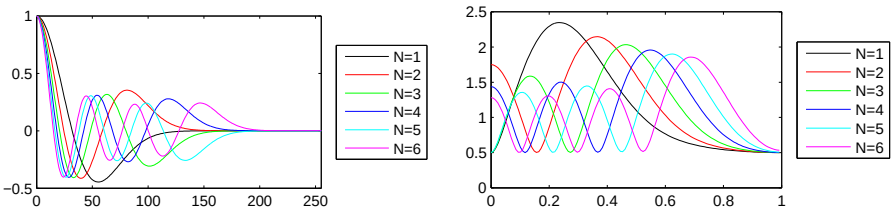


Figure 7.8: (a) Laguerre-Gauss (LG) functions for different orders N, (b) Frequency responses of the LG functions.

data set, in analogy to multiple physicians inspecting multiple cases. An example is the one shot method (see e.g. [Gallas et al., 2007]), which enables to analyze the variance on the AUC estimates, due to within-reader variability and between-reader variability.

- Third, a CHO model using rotationally symmetric channels is not well suited for detecting *rotationally asymmetric signals*, as the detector can not distinguish directional features of the signal. An example is given in Figure 7.9: here two artificial lesions are shown that all generate the same channel responses for a set of LG channels. In other words: the CHO can not distinguish the lesions from one another. A trivial solution would then be to use asymmetric channels as in [LaCroix et al., 1999], however,

rotationally asymmetric signals typically have an unknown dominant orientation in most practical circumstances and a CHO model for SKE tasks can not take this uncertainty into account. We will show in Section 7.2 that dealing with this form of uncertainty necessarily causes an optimal observer (e.g. in the maximum likelihood sense) to use a nonlinear test statistic.

In the next section, we will address the third issue, by extending the CHO model to deal with asymmetrical signals with unknown (random) orientation. However, the problem can be stated more generally, as the size or amplitude of the signal can also be unknown. The detection task then becomes a signal-known-statistically (SKS) task.

7.2 New model observers for SKS tasks

Signal-known-statistically (SKS) detection tasks are more complicated than SKE tasks, simply because less information is available about the signal that is to be detected.

The performance of human and model observers in SKS tasks has been studied in [Eckstein and Abbey, 2001, Manjeshwar and Wilson, 2001, Castella et al., 2009, Park et al., 2005]. [Eckstein and Abbey, 2001] compute the SKS task perfomance by using different templates for different combinations of signal parameter values, and subsequently the authors optimally combine (in maximum likelihood sense) the template outputs. Their results suggest that the performance of SKE tasks could be used as a first order approximation to performance in SKS tasks. [Manjeshwar and Wilson, 2001] conclude that the human observer performance is severly limited in location uncertainty but the detection performance can be improved by 77% by adding a marker around the signal. [Park et al., 2005] estimate the human observer efficiency relative to the IO. The authors find that the relative human efficiency for SKE tasks is much lower than for SKS tasks, which indicates a great potential for SKS model observers that behave closer to humans. [Castella et al., 2009] investigate the influence of signal variations and uncertainty on human detection performance. It is reported that the human observer is mostly sensitive to *signal size uncertainty*, but not significantly to *signal shape uncertainty*.

Table 7.1 gives a summary of existing model observers. It shows whether signal or background knowledge is required, and if the model is related to the HVS. We remark that, opposed to SKE detection tasks, channelized Hotelling observers *for SKS tasks* have not been studied extensively before.

In this section, we will develop new nonlinear model observers for detecting signals with unknown (random) parameters, based on the theory of joint detection and estimation (JDE) [Olmo et al., 2000]. These classes of observers jointly estimate the unknown signal parameters and the signal presence, which means that depending on the task for which they are designed, these observers are also able to determine the *location*, *orientation* and *size* of signals on an

Table 7.1: Brief summary of prior knowledge used by common model observers.

Model	Knowledge about signal	Background knowledge	Human Visual System
Ideal observer (IO) [Barrett et al., 1998]	Full knowledge	Full knowledge	Not related
Ideal linear observer (ILO) [Gallas and Barrett, 2003]	Full knowledge	Full knowledge	Not related
Channelized ideal linear observer [Park et al., 2007]	Full knowledge	Full knowledge	Frequency or orientation selective channels
Channelized Hotelling observer (CHO) for SKE tasks [Gallas and Barrett, 2003]	Learned during training phase	Learned during training phase	Frequency or orientation selective channels
CHO for SKS tasks (Section 7.2.2)	Learned during training phase	Learned during training phase	Frequency, orientation and scale selective channels

image. To keep the ability to work with relatively small training sets of images, we also explain the general framework in which channelized Hotelling implementations of these observers can be derived, while retaining the theoretical performance of their non-channelized equivalents as much as possible.

One important result is that a number of specific detection tasks in which some signal parameters are known and others are unknown, pose restrictions on the choice of the channels. For example, if we want to optimally detect signals with unknown orientations, then orientation-selective channels should be used that are angularly *steerable* (see Chapter 2). On the other hand, if the size (scale) of the signal is unknown, the channels should be frequency-selective and shiftable in scale. For more complex tasks, e.g. the location, orientation and size are unknown, the channel design constraints can seamlessly be combined, resulting in channels that are both orientation and frequency selective and that are at the same time localized in space.
The purpose of the developed JDE model observers is two-fold:

1. We show that the detection performance of the JDE model observers is generally closer to the performance of the IO than the ideal *linear* observer (ILO), which is constrained to linear decision boundaries. Hence, the JDE model observers, or in particular the channelized implementations of these observers, are very practical in use while closely matching the performance of the IO.

2. Our hope is that the new observers provide more insights into the discrepancies between human observers and the IO, with respect to the intrinsic signal uncertainty. The need for frequency- and orientation-selective channels for optimal detection also agrees with the presence of frequency- and orientation-responsive neurons in the human V1.

7.2.1 A Variational approximation of the IO for SKS tasks

First, we will derive the IO test statistic for SKS tasks. As this will not give a closed form expression for the test statistic (which makes it difficult to directly

compute the IO performance), we will apply a variational approximation. This will provide us some insights in the detection performance of the new observers later.

For an SKS detection task, the observer has to deal with the presence of hidden signal parameters α. According to (7.1), the test statistic of the ideal observer can be computed by marginalizing the conditional densities over the hidden parameters [Park et al., 2005]:

$$\lambda(\mathbf{y}) = \frac{f_{\mathbf{Y}|H}(\mathbf{y}|H_1)}{f_{\mathbf{Y}|H}(\mathbf{y}|H_0)} = \frac{\int_\Omega f_{\mathbf{Y}|\alpha,H}(\mathbf{y}|\alpha,H_1) f_\alpha(\alpha)\mathrm{d}\alpha}{f_{\mathbf{Y}|H}(\mathbf{y}|H_0)} \tag{7.13}$$

The ideal observer requires the specification of the conditional probability density functions of the data: $f_{\mathbf{Y}|H}(\mathbf{y}|H_0)$ and $f_{\mathbf{Y}|H}(\mathbf{y}|H_1)$. The conditional probability density function $f_{\mathbf{Y}|H}(\mathbf{y}|H_1)$ in (7.13) contains high-dimensional integrals and computation of $\lambda(\mathbf{y})$ is usually done through MCMC methods [Park et al., 2003, Park et al., 2005]. Unfortunately, for the MCMC simulations it is much more difficult to examine the influence of the unknown parameters on e.g. the detection performance, unless many simulations are performed for different combinations of parameters.

An alternative approach that is analytically tractable makes use of Bayesian Variational approximation [Beal, 2003]. Using Jensens' inequality, a lower bound of the log-likelihood function $\log\lambda(\mathbf{y})$ is maximized in order to reach a decision of signal presence:

$$\begin{aligned}
\log\lambda(\mathbf{y}) &= \log\int_\Omega f_\alpha(\alpha)\frac{f_{\mathbf{Y},\alpha|H}(\mathbf{y},\alpha|H_1)}{f_\alpha(\alpha)}\mathrm{d}\alpha - \log f_{\mathbf{Y}|H}(\mathbf{y}|H_0) \\
&\geq \int_\Omega f_\alpha(\alpha)\log\frac{f_{\mathbf{Y},\alpha|H}(\mathbf{y},\alpha|H_1)}{f_\alpha(\alpha)}\mathrm{d}\alpha - \log f_{\mathbf{Y}|H}(\mathbf{y}|H_0) \\
&= \mathrm{KL}\left[f_\alpha(\alpha)\|\, f_{\alpha|\mathbf{y},H}(\alpha|\mathbf{y},H_1)\right] - \log f_{\mathbf{Y}|H}(\mathbf{y}|H_0)
\end{aligned} \tag{7.14}$$

with $\mathrm{KL}\left[\cdot\|\,\cdot\right]$ the Kullback-Leibler divergence between two probability density functions. Although this approach still faces high-dimensional integrals, under the Gaussianity assumption of the background (Section 7.1.2), the log-likelihood function has a simple form and the test statistic can be reduced to:

$$\begin{aligned}
t_{\mathrm{var}} &= \log\lambda(\mathbf{y}) \\
&= \int_\Omega f_\alpha(\alpha)\left(\mathbf{x}_\alpha^T\mathbf{C}_b^{-1}\mathbf{y}\right)\mathrm{d}\alpha \\
&= \mathrm{E}_\alpha\left[\mathbf{x}_\alpha^T\right]\mathbf{C}_b^{-1}\mathbf{y}
\end{aligned} \tag{7.15}$$

where the subscript signifies that the mathematical expectation is taken with respect to α, i.e. $\mathrm{E}_\alpha\left[g(\mathbf{x})\right] = \int_\Omega f_\alpha(\alpha)g(\mathbf{x})\mathrm{d}\alpha$. It can be shown that the SNR[5]

[5]For the definition of SNR for detection tasks, we refer to [Barrett et al., 1995]. We remark that in the literature about model observers, the SNR is defined as the *square root* of the power ratio between the signal and noise (in contrast to most electrical engineering literature, where it is simply the power ratio). For consistency with the other Chapters in this book, we stick to the EE definition.

performance of the observer with test statistic (7.15) is given by:

$$\mathrm{SNR}_{t,\mathrm{var}} = \frac{\left(\mathrm{E}_\alpha \left[\mathbf{x}_\alpha\right]^T \mathbf{C}_b^{-1} \mathrm{E}_\alpha \left[\mathbf{x}_\alpha\right]\right)^2}{\mathrm{E}_\alpha \left[\mathbf{x}_\alpha^T \mathbf{C}_b^{-1} \mathbf{x}_\alpha\right]}. \tag{7.16}$$

Furthermore, because the t_{var} is Gaussian distributed, the AUC is directly related to the SNR [Barrett et al., 1998]:

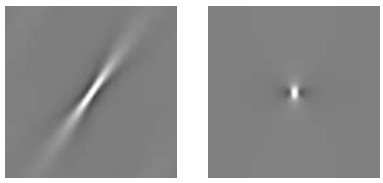
$$\mathrm{AUC}_{\mathrm{var}} = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\sqrt{\mathrm{SNR}_{t,\mathrm{var}}}}{2}\right). \tag{7.17}$$

Because the Variational approximation (7.14) amounts to approximating the probability density $f_\Lambda(\lambda)$ with a Gaussian distribution, the *linear* test statistic t_{var} is also the discriminant function of the ILO. By comparing (7.15) to the discriminant function for signal known exactly (SKE) tasks (see e.g. [Barrett, 1990]), we note that the only difference is that the signal profile is averaged over all possible parameter values. Hence the ILO detects the *expected* signal, under all signal uncertainty. For example, in case the orientation of an asymmetric signal (see (7.1.3)) is unknown but uniformly distributed on $[0, \pi]$, the ideal linear observer will average the ensemble of signals for all possible orientations and will attempt to detect this averaged symmetric signal. However, there are two issues with the ILO to take into account:

1. The ILO does not provide estimates of the unknown signal parameters α (which would be useful as an indication why a signal was detected in a given image).

2. The log-likelihood function (7.14) is a (non-tight) lower bound for the true log-likelihood function, due to the use of Jensens' inequality. Maximization of the lower-bound does not necessarily give the global optimum of the true log-likelihood function. For example, in [Park et al., 2005] it has been experimentally found that ensemble averaging of the signal over all possible signal locations (in a location-unaware task) yields a poor signal detection performance. In general, the ILO can not distinguish between two significantly different realizations of a signal with a given ensemble average (over all possible values of the unknown parameters). The same goes for different signals with the same ensemble average, but with significantly different realizations at the time. Figure 7.9 shows an example for orientation-unaware tasks. The ILO is often not adequate in practice, and the above observations suggest that nonlinear decision rules may perform closer to optimal since the IO in SKS tasks is generally nonlinear.

Next, we will investigate an alternative solution that allows us to estimate of the unknown signal parameters jointly with the signal presence. This will also give a performance that is often closer to the ideal observer.



(a) (b)

Figure 7.9: Example of two asymmetric signals with the same ensemble average when averaged over all possible orientations ($0°-180°$), e.g. in an orientation-unaware task. Even though the of appearance is quite different, the ILO can not distinguish these two signals from each other.

7.2.2 Model Observer based on Joint Detection and Estimation

In the theory of joint detection and estimation (JDE) [Olmo et al., 2000], the composite parameter estimation and hypothesis testing is seen as the joint estimation of a mixed set of discrete and continuous parameters. The joint approach is obviously independent of the order in which the detection and estimation are being performed, and generally yields better overall performance (in the MAP sense) than when both detection and estimation are done invidivually. Let us denote the a posterior distribution of the parameters as $\varphi(\alpha, H_k) = f_{\alpha,H|\mathbf{Y}}(\alpha, H_k|\mathbf{y})$. For the background model (see Section 7.1.2), the joint MAP estimation of the unknown parameters and the hypothesis is performed as follows:

$$\begin{aligned}
\widehat{(\alpha, H_k)} &= \arg \max_{(\alpha,H_k)} \varphi(\alpha, H_k) \\
&= \arg \max_{(\alpha,H_k)} -\frac{1}{2} \left(\mathbf{y} - k\mathbf{x}_\alpha\right)^T \mathbf{C}_b^{-1} \left(\mathbf{y} - k\mathbf{x}_\alpha\right) + \log\, f_\alpha(\alpha) + \log\, P(H_k) \\
&= \arg \max_{(\alpha,H_k)} k\mathbf{x}_\alpha^T \mathbf{C}_b^{-1} \left(\mathbf{y} - \frac{1}{2}\mathbf{x}_\alpha\right) + \log\, f_\alpha(\alpha) + \log\, P(H_k) \qquad (7.18)
\end{aligned}$$

where the statistical independence of α and H_k has been exploited. The decision rule that provides the solution to (7.18), can be written as follows (see Figure 7.10(a)):

$$\hat{H}_k = \begin{cases} H_1 & \text{if } \exists \alpha^\star : \mathbf{x}_{\alpha^\star}^T \mathbf{C}_b^{-1} \left(\mathbf{y} - \frac{1}{2}\mathbf{x}_{\alpha^\star}\right) \geq \log \frac{P(H_0)}{P(H_1)} \\ H_0 & \text{else} \end{cases} \qquad (7.19)$$

Important to note is that if α is fixed (as in SKE tasks) or if $\mathbf{x}_\alpha$ is a linear function of α, the decision boundary (i.e. $\mathbf{y} = \mathbf{x}_\alpha/2$) becomes linear in $\mathbf{x}_\alpha$. Moreover, for the test statistic

$$t_{\text{JDE}} = \mathbf{x}_\alpha^T \mathbf{C}_b^{-1} \mathbf{y}, \qquad (7.20)$$

Table 7.2: MAP estimators for unknown signal parameters. In this table, $(\cdot)_+ = \max(0, \cdot)$. For the channelized observers (third and fourth row), see explanation in Section 7.3. Here $\mathbf{U}_\alpha$ is the channel matrix, $\mathbf{C}_b^{'}$ is the background covariance matrix in the channel domain.

Observer / distribution of α	MAP-optimal estimate
non-channelized / $f_\alpha(\alpha)$ uniform on Ω	$\hat{\alpha} = \arg\max_{\alpha\in\Omega} \mathbf{x}_\alpha^T \mathbf{C}_b^{-1} \left(\mathbf{y} - \frac{1}{2}\mathbf{x}_\alpha\right)$
non-channelized / $f_\alpha(\alpha)$ non-uniform on Ω	$\hat{\alpha} = \arg\max_{\alpha\in\Omega} \log f_\alpha(\alpha) + \left(\mathbf{x}_\alpha^T \mathbf{C}_b^{-1} \left(\mathbf{y} - \frac{1}{2}\mathbf{x}_\alpha\right) - \log\frac{P(H_0)}{P(H_1)}\right)_+$
channelized / $f_\alpha(\alpha)$ uniform on Ω	$\hat{\alpha} = \arg\max_{\alpha\in\Omega} \left(\frac{\mathbf{x}_\alpha^T}{\|\mathbf{U}_{\hat{\alpha}}\|_F^2} \left(\mathbf{U}_\alpha \mathbf{C}_b^{'-1} \mathbf{U}_\alpha^T\right) \left(\mathbf{y} - \frac{1}{2}\mathbf{x}_\alpha\right)\right)$
channelized / $f_\alpha(\alpha)$ non-uniform on Ω	$\hat{\alpha} = \arg\max_{\alpha\in\Omega} \log f_\alpha(\alpha) + \left(\frac{\mathbf{x}_\alpha^T}{\|\mathbf{U}_{\hat{\alpha}}\|_F^2} \left(\mathbf{U}_\alpha \mathbf{C}_b^{'-1} \mathbf{U}_\alpha^T\right) \left(\mathbf{y} - \frac{1}{2}\mathbf{x}_\alpha\right) - \log\frac{P(H_0)}{P(H_1)}\right)_+$

the decision rule has a clear interpretation: if t_{JDE} is close to 0, the decision will be H_0, on the other hand, if t_{JDE} is close to $\mathbf{x}_\alpha^T \mathbf{C}_b^{-1} \mathbf{x}_\alpha$, the signal will be detected (H_1). The exact boundary depends on the prior probabilities $P(H_0)$ and $P(H_1)$. According to (7.18), the estimation of α can also be done independently from the signal detection. Because the found $\hat{\alpha}$ will automatically be the best candidate for making a decision of signal presence (i.e. $\alpha^\star = \hat{\alpha}$ in (7.19)), the JDE can be implemented using a simple *sequential* scheme (see Figure 7.10b):

1. Estimate the set of unknown signal parameters α, in order to maximize the conditional likelihood function:

$$\hat{\alpha} = \arg\max_{\alpha} \max\left(\varphi(\alpha, H_0), \varphi(\alpha, H_1)\right) \tag{7.21}$$

   Table 7.2 lists exact expressions for the estimation of α.

2. Use this estimate in order to detect signal presence, using the test statistic $t_{\mathrm{JDE}} = \mathbf{x}_{\hat{\alpha}}^T \mathbf{C}_b^{-1} \mathbf{y}$ evaluated in $\hat{\alpha}$. The final decision is H_1 if and only if $t_{\mathrm{JDE}} > \frac{1}{2}\mathbf{x}_{\hat{\alpha}}^T \mathbf{C}_b^{-1} \mathbf{x}_{\hat{\alpha}} + \log\frac{P(H_0)}{P(H_1)}$, otherwise the decision is H_0.

These findings have a number of practical consequences: 1) the detector is clearly *nonlinear*, however, the nonlinearity is mainly in the estimation part of the scheme (and in particular in the dependence of $\mathbf{x}_\alpha$ on α). Once the signal parameters are estimated, the test statistic is linear in $\mathbf{x}_{\hat{\alpha}}$ (decision part). This finding will allow use to derive a channelized Hotelling implementation of this scheme in Section 7.3.

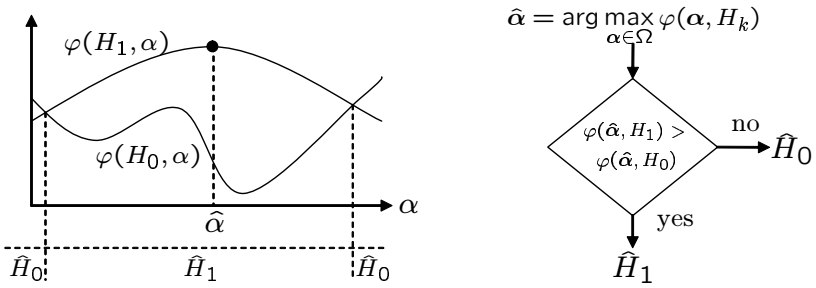


Figure 7.10: (a) Illustration of the joint MAP detection and estimation of one single parameter α. Shown are the objective functions $\varphi(H_0,\alpha)$ and $\varphi(H_1,\alpha)$ for hypotheses H_0 and H_1, respectively. The MAP estimate is $\widehat{(\alpha, H_k)} = \arg\max \varphi(H_k,\alpha)$. The decision is H_1 if and only if there exists $\alpha^\star : \varphi(H_1,\alpha^\star) > \varphi(H_0,\alpha^\star)$ (the decision boundaries are indicated using dashed lines). If so, and if $\alpha^\star$ is abscis of the maximum of $\varphi(H_1,\alpha^\star)$, the estimate of the parameter $\hat{\alpha}$ is given by $\alpha^\star$. (b) Sequential scheme that gives the optimal JDE estimation.

Detection performance

Next, we will investigate the detection performance of the JDE scheme. First of all, conditioned on $\hat{\alpha}$, the test statistic is the same as for SKE tasks. For SKS, the performance is highly influenced by the quality of the estimate $\hat{\alpha}$, or more particularly, the joint density function $f_{\alpha,\hat{\alpha}}(\alpha,\hat{\alpha})$. For example, one may expect that the lower the estimation error on $\hat{\alpha}$, the better the detection performance will be. To derive the SNR, the first and second order conditional moments of the test statistic are required:

$$\begin{aligned} \mathrm{E}\left[t_{\mathrm{JDE}}\,|\alpha,\hat{\alpha},H_0\right] &= 0, \\ \mathrm{E}\left[t_{\mathrm{JDE}}\,|\alpha,\hat{\alpha},H_1\right] &= \mathbf{x}_{\hat{\alpha}}^T\mathbf{C}_b^{-1}\mathbf{x}_\alpha, \text{ and} \\ \mathrm{Var}\left[t_{\mathrm{JDE}}\,|\alpha,\hat{\alpha},H_k\right] &= \mathbf{x}_{\hat{\alpha}}^T\mathbf{C}_b^{-1}\mathbf{x}_{\hat{\alpha}}. \end{aligned}$$

From these expressions, the detection SNR can be directly computed [Barrett, 1990]:

$$\begin{aligned} \mathrm{SNR}_{t,\mathrm{JDE}}^2 &= \frac{\left(\mathrm{E}\left[t_{\mathrm{JDE}}\,|H_1\right] - \mathrm{E}\left[t_{\mathrm{JDE}}\,|H_0\right]\right)^2}{\left(\mathrm{Var}\left[t_{\mathrm{JDE}}\,|H_1\right] + \mathrm{Var}\left[t_{\mathrm{JDE}}\,|H_0\right]\right)/2} \\ &= \frac{\left(\mathrm{E}_{\hat{\alpha},\alpha}\left[\mathbf{x}_{\hat{\alpha}}^T\mathbf{C}_b^{-1}\mathbf{x}_\alpha\right]\right)^2}{\mathrm{E}_{\hat{\alpha}}\left[\mathbf{x}_{\hat{\alpha}}^T\mathbf{C}_b^{-1}\mathbf{x}_{\hat{\alpha}}\right]}. \end{aligned} \tag{7.22}$$

Now, it is interesting to note that if α and $\hat{\alpha}$ are statistically independent, but identically distributed, (7.22) is the same expression as (7.16). This suggests

that if the estimator would select $\hat{\alpha}$ randomly from the probability density function $f_\alpha(\alpha)$, the SNR performance would be equal to the SNR performance of the ideal linear observer. Moreover, if one does a *better* job than guessing $\hat{\alpha}$ randomly (e.g. by using (7.21)), the SNR performance of the JDE scheme will be higher than that of the ideal linear observer.

Summary

For SKS tasks, the JDE scheme has the following advantages compared to the ideal observer:

1. The scheme provides an explicit estimate of the unknown parameters $\hat{\alpha}$.
2. High dimensional integrals are generally avoided. The integration task is replaced by solving an optimization problem (7.21), for which more efficient techniques exist.
3. The test statistic is still linear, which has the most practical consequence that, depending of the parameters that are known or unknown, it is often possible to devise an efficient channelized Hotelling scheme for the considered problem, as we will explain in the next section.
4. The JDE scheme generally outperforms the ideal linear observer, as we will show later, and can attain a performance that is close (or equal) to the IO.

The JDE observer can be considered a practical approximation to the IO for SKS tasks. However, in practice, complete probability density functions of the images are not available. Therefore in the next section, we will extend the JDE observer to use linear channels and this will allow some interesting perspectives with respect to the channel choice and design.

7.3 Channelized Hotelling observers for SKS detection tasks

As we explained in Section 7.1.4, for more realistic scenarios (like more complex non-Gaussian backgrounds), we would like to train the observer from a limited set of images. Unfortunately, the JDE detection scheme from the previous section can not directly be used for this. Therefore, similar to the CHO for SKE tasks (Section 7.1.4), we constrain the observer to a small set of linear channels. Recall that the test statistic of the JDE observer is given by:

$$t_{\text{JDE}} = \mathbf{x}_{\hat{\alpha}}^T \mathbf{C}_b^{-1} \mathbf{y}. \tag{7.23}$$

Next, we introduce linear projections onto a set of K channels (i.e. $\mathbf{y}$ is mapped to $\mathbf{U}_{\hat{\alpha}}^T \mathbf{y}$, with $\mathbf{U}_{\hat{\alpha}}$ a $N \times K$ matrix). This is a projection from the N-dimensional

space $\mathbb{R}^N$ (spatial domain) to the K-dimensional space $\mathbb{R}^K$ (channel domain). In the channel domain, the test statistic can be written as:

$$t'_{\mathrm{JDE}} = \frac{\mathbf{x}_{\hat{\alpha}}^T}{\|\mathbf{U}_{\hat{\alpha}}\|_F^2} \left(\mathbf{U}_{\hat{\alpha}} \mathbf{C}_b'^{-1} \mathbf{U}_{\hat{\alpha}}^T \right) \mathbf{y} \tag{7.24}$$

with $\|\mathbf{U}_{\hat{\alpha}}\|_F^2$ a channel matrix energy normalization factor, which will prove useful later. The notation indicates that the channels are *adaptive* to the signal parameters. For example, if one uses Laguerre-Gauss channels for detecting rotationally symmetric Gaussian signals with known scale parameter σ (see (7.6) with $b = c = 1$), then it is common to tune the channels so that the first order Laguerre-Gauss function matches the signal ($a_{\mathrm{LG}} = \sqrt{2\pi}\sigma$, see [Gallas and Barrett, 2003]). If the scale parameter σ is unknown, we also keep the Laguerre-Gauss parameter a_{LG} *variable*. Once the estimate of σ is available (through (7.21)), we can use this value to tune the channel parameter $\hat{a}_{\mathrm{LG}} = \sqrt{2\pi}\hat{\sigma}$.

Ideally, we would like that $t'_{\mathrm{JDE}} \approx t_{\mathrm{JDE}}$, such that t'_{JDE} is a approximately a sufficient statistic, and such that the influence of the dimension reduction by channel projection is minimal. For this, the covariance matrix $\mathbf{C}_b'^{-1}$ and the channel matrix $\mathbf{U}_{\hat{\alpha}}$ need to be chosen suitably, such that $\mathbf{C}_b^{-1} \approx \mathbf{U}_{\hat{\alpha}} \mathbf{C}_b'^{-1} \mathbf{U}_{\hat{\alpha}}^T / \|\mathbf{U}_{\hat{\alpha}}\|_F$. Now, we would like to estimate the unknown signal parameters in the channel space as well, for reasons of elegance and computational efficiency. In analogy to (7.24), we replace the MAP estimates $\hat{\alpha}$ by their channelized equivalents, which are listed in Table 7.2. For example, if α is uniformly distributed in Ω, we have

$$\hat{\alpha} = \arg\max_{\alpha \in \Omega} \left(\frac{\mathbf{x}_\alpha^T}{\|\mathbf{U}_{\hat{\alpha}}\|_F^2} \left(\mathbf{U}_\alpha \mathbf{C}_b'^{-1} \mathbf{U}_\alpha^T \right) \left(\mathbf{y} - \frac{1}{2} \mathbf{x}_\alpha \right) \right). \tag{7.25}$$

In general, the objective function in (7.25) is highly nonconvex and may contain many local but non-global maxima, which practically means that projections need to be performed using a whole set of channel matrices $\mathbf{U}_\alpha$ with varying parameters α. Because the search space and the dimensionality of the data are still huge, this technique would not be very practical. The key to the solution for this problem is to choose the channels *suitably*, such that a transform on the signal in the spatial domain can be expressed as an equivalent transform on the signal in channel domain. Instead of searching for parameter estimates in the spatial domain, the optimization (7.25) can then completely take place in channel domain, without loss of accuracy. Therefore we will add an extra requirement for designing the channel matrix:

$$\mathbf{U}_\alpha = \mathbf{A}_\alpha^T \mathbf{U}_0 = \mathbf{U}_0 \mathbf{A}_\alpha'^T \tag{7.26}$$

where $\mathbf{U}_0$ is a *fixed* channel matrix, which does not depend on the unknown parameters α. One can think of $\mathbf{A}_\alpha$ as a linear transform matrix that maps $\mathbf{x}_\alpha$ onto the reference signal $\mathbf{x}_0$ in the image space (i.e. a signal constructed

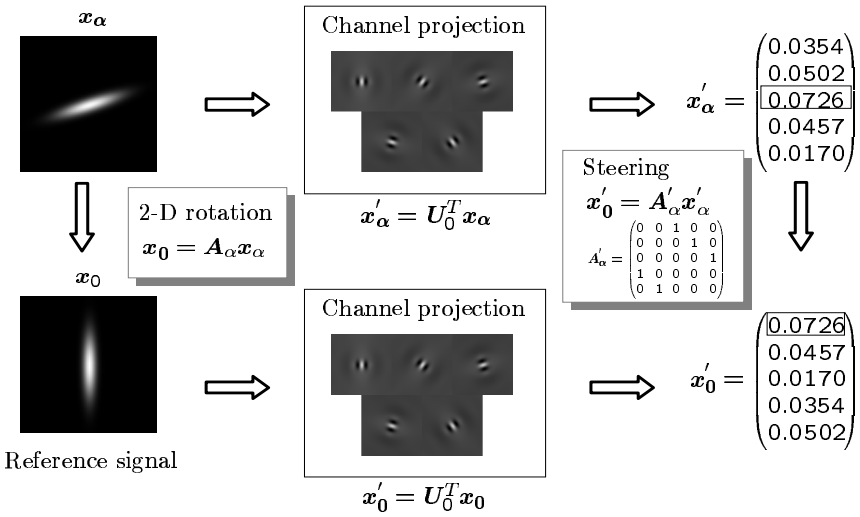


Figure 7.11: Illustration of the channel design condition (equation (7.26)): a rotation by 72° in the image domain corresponds to a matrix multiplication in the channel domain. The channel matrix $\mathbf{U}_0$ and the matrices $\mathbf{A}_\alpha$, $\mathbf{A}'_\alpha$ need to be chosen suitably, such that in the figure the end result for transiting from $\mathbf{x}_\alpha$ to $\mathbf{x}'_0$ is the same, independent of the path being followed. The problem of orientation-unaware detection is discussed in more detail in Section 7.3.2.

from *known* reference parameters): $\mathbf{A}_\alpha \mathbf{x}_\alpha = \mathbf{x}_0$. Similarly, $\mathbf{A}'_\alpha$ does the same in the channel space by mapping $\mathbf{U}_0^T \mathbf{x}_\alpha$ onto $\mathbf{U}_0^T \mathbf{x}_0$. Figure 7.11 illustrates the case of the orientation-unaware detection of a signal. Before explaining the design of fixed channel matrices satisfying (7.26), we will first show that the estimation part now completely takes place in channel space. Therefore, we substitute (7.26) into the estimator (7.25):

$$\hat{\alpha} = \arg\max_\alpha \frac{1}{\|\mathbf{A}_\alpha^T \mathbf{U}_0\|_F^2} \mathbf{x}_0'^T \mathbf{C}_b'^{-1} \left(\mathbf{A}_\alpha'^T \mathbf{y}' - \frac{1}{2} \mathbf{x}'_0 \right) \tag{7.27}$$

with $\mathbf{y}' = \mathbf{U}_0^T \mathbf{y}$ and $\mathbf{x}'_0 = \mathbf{U}_0^T \mathbf{x}_0$ the projected observed image and the projected reference signal, respectively. Now we see that both the reference signal $\mathbf{x}_0$ and the observed image $\mathbf{y}$ are projected only *once*, using the fixed channel matrix. Exactly the same applies to test statistic from equation (7.24), which becomes:

$$\begin{aligned} t'_{\text{JDE}} &= \frac{1}{\|\mathbf{A}_\alpha^T \mathbf{U}_0\|_F^2} \mathbf{x}_{\hat{\alpha}}^T \mathbf{A}_{\hat{\alpha}}^T \mathbf{U}_0 \mathbf{C}_b'^{-1} \mathbf{A}'_{\hat{\alpha}} \mathbf{U}_0^T \mathbf{y} \\ &= \frac{1}{\|\mathbf{A}_\alpha^T \mathbf{U}_0\|_F^2} \mathbf{x}_0'^T \mathbf{C}_b'^{-1} \mathbf{A}'_{\hat{\alpha}} \mathbf{y}'. \end{aligned} \tag{7.28}$$

Because $\mathbf{y}'$ and $\mathbf{x}'_0$ can be precomputed prior to the optimization in (7.27),

both the estimation of α and the detection task are done in channel space. The decision of signal presence is given by:

$$\hat{H}_k = \begin{cases} H_1 & \text{if } t'_{\text{JDE}} \geq \frac{\mathbf{x}_0'^T \mathbf{C}_b'^{-1} \mathbf{x}_0'}{\|\mathbf{A}_\alpha^T \mathbf{U}_0\|_F^2} + \log \frac{P(H_0)}{P(H_1)} \\ H_0 & \text{else} \end{cases} \tag{7.29}$$

where the right hand of the inequality in (7.29) is constant and independent of the observed image. The consequences are huge:

- As already mentioned, it is advocated to use channels that are applicable in a wide range of situations. Through Figure 7.11 we end up with a constant channel matrix $\mathbf{U}_0$, which is independent of α (but which depends on the type of SKS task, as we will see). Hence, this would allow us to use a fixed channel matrix for a given SKS task (also see further). From a computational point of view, the channelization ($\mathbf{U}_0^T \mathbf{y}$) can immediately take place as a first step in processing the data.
- Because the JDE can be split up in a *nonlinear estimation part* and a *linear detection part* (see Section 7.2.2), the decision boundaries for the mapped observation vector $\mathbf{A}'_{\hat{\alpha}} \mathbf{U}_0^T \mathbf{y}$ are *linear*! Hence we can immediately obtain a CHO implementation for SKS tasks, which can be used for arbitrary signals and backgrounds. The details will be explained in Section 7.4.

The only remaining question is: how to design channel matrices and mapping transforms that satisfy the channel design constraint from equation (7.26)? Briefly, the design procedure can be summarized as follows:

- First fix the reference signal $\mathbf{x}_0$: i.e. choose reference parameter values, perhaps arbitrarily.
- Define the transform $\mathbf{A}_\alpha$ that maps any signal $\mathbf{x}_\alpha$ onto the reference signal.
- Choose the channel matrix suitably, such that there exists a transform matrix $\mathbf{A}'_\alpha$ in channel space satisfying (7.26).

Because the third step is a difficult task in general, we will treat each unknown parameter (signal amplitude, rotation, scale and location) at the time in the following subsections.

7.3.1 Detection of signals with random amplitude

First we consider the case where only the signal amplitude $a > 0$ is unknown and uniformly distributed on $[a_{\min}, a_{\max}]$, and $\alpha = [a]$. For clinical applications, this problem with variable signal amplitude (or contrast-to-noise ratio of the signal) is a very relevant. To proceed, we apply the design procedure outlined before. For the reference signal, it is convenient to choose $a_0 = 1$, such that

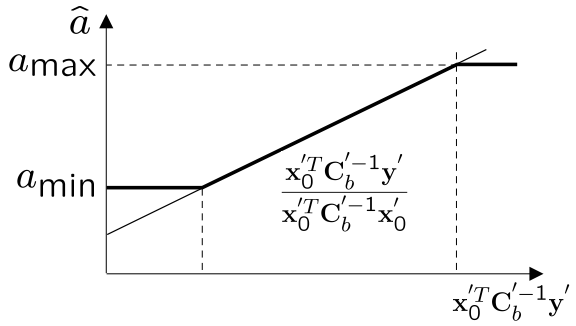


Figure 7.12: Illustration of the JDE amplitude estimation.

$\mathbf{A}_a = a^{-1}\mathbf{I}$ is simply a diagonal matrix. To satisfy (7.26) it is sufficient to choose $\mathbf{A}_a^{'} = a^{-1}\mathbf{I}$ so there are no further restrictions necessary on the channel matrix. Using (7.27), the JDE amplitude estimate is as follows:

$$\hat{a} = \min\left(a_{\max}, \max\left(a_{\min}, \frac{\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{y}^{'}}{\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{x}_0^{'}}\right)\right). \tag{7.30}$$

This estimator is linear in the domain $[a_{\min}, a_{\max}]$, and saturates outside this domain. An illustration of (7.30) is given in Figure 7.12. Next, (7.30) can be put into (7.38) to yield the final test statistic:

$$t_{\text{JDE}}^{'} = \begin{cases} \frac{\left(\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{y}^{'}\right)^2}{\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{x}_0^{'}}, & \text{if } a_{\min} < \hat{a} < a_{\max} \\ a_{\min}\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{y}^{'}, & \text{if } \hat{a} = a_{\min} \\ a_{\max}\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{y}^{'} & \text{if } \hat{a} = a_{\max} \end{cases} \tag{7.31}$$

Hence the final test statistic saturates at the points $a_{\min}\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{y}^{'}$ and $a_{\max}\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{y}^{'}$, consistent with the fact that the signal amplitude can not exceed the interval $[a_{\min}, a_{\max}]$.

Similar estimators can be derived for other prior distributions. Suppose a has an exponential distribution with mean $1/\lambda$, according to JDE theory the amplitude is estimated as follows:

$$\hat{a} = \begin{cases} \left(\frac{\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{y}^{'}-\lambda}{\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{x}_0^{'}}\right)_+ & \text{if } t_{\text{JDE}}^{'} = \frac{\left(\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{y}^{'}-\lambda\right)^2}{2\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{x}_0^{'}} \geq \log\frac{P(H_0)}{P(H_1)} \quad (H_1) \\ \text{undefined} & \text{else } (H_0) \end{cases} \tag{7.32}$$

which is again linear on $\mathbf{x}_0^{'T}\mathbf{C}_b^{'-1}\mathbf{y}^{'} \in [\lambda, +\infty[$. In case the result is undefined, the observed signal is too weak to obtain a reliable estimate of the amplitude. This case corresponds to *signal absence* (H_0).

As can be seen, the estimation of the signal amplitude is a task that does not rely on a particular choice of the channels. This allows us the freedom to design specific channels for more difficult SKS tasks, some of which we will discuss in the next subsections.

7.3.2 Orientation-unaware detection

Now we assume that the signal orientation angle ϑ is unknown and uniformly distributed on $[0, \pi]$, and $\alpha \equiv [\vartheta]$. Extensions to non-uniform prior distributions for ϑ are also possible. As a reference orientation angle, we choose $\vartheta = 0$. The matrix $\mathbf{A}_\vartheta$ should rotate the spatial plane and transform $\mathbf{x}_\vartheta$ into $\mathbf{x}_0$. Because of the discrete sampling, the construction of the matrix $\mathbf{A}_\vartheta$ could be done based on bandlimited resampling, in order to obtain an "ideal" rotation matrix. Fortunately, because we completely work in the channel space, the practical problem of the computation of $\mathbf{A}_\vartheta$ is avoided. The main problem becomes the design of the channels, according to (7.26). In the context of orientation-unaware detection, equation (7.26) states that rotating the image in the image space should correspond in a linear operation ("steering") in channel space. Hence we should use channels that can be "steered" to any orientation in the image space, based on a linear sum of a fixed set of channels. The solution is provided by the *steerable* filters from Section 2.3.2.

We therefore design channels that are both sensitive to specific scales as to specific orientations, as a product of steerable functions $f_{\text{steer}}^{(\vartheta_k)}(r, \varphi)$ and a rotationally symmetric (RS) function $f_{\text{RS}}^{(s)}(r, \varphi)$ (e.g. LG functions or DDOG functions):

$$f^{(\vartheta_k, s)}(r, \varphi) = f_{\text{steer}}^{(\vartheta_k)}(r, \varphi) f_{\text{RS}}^{(s)}(r, \varphi), \quad k = 1, ..., K \text{ and } s = 1, ..., S. \tag{7.33}$$

To construct the channel matrix $\mathbf{U}$, it suffices to sample the channel functions (7.33) and store the samples in an $N \times (KS)$-matrix where every column contains one sampled channel function. Some examples of steerable LG functions are shown in Figure 7.13(a), for $K = 5$ orientations. For simplicity, we will use only one scale ($S = 1$) and $f_{\text{RS}}^{(s)}(r, \varphi) = 1$ in the remainder of this text. The theory can be easily extended to multiple scales.

Estimation of the unknown orientation angle

By opting for steerable channels, the transform matrix $\mathbf{A}_\vartheta^{'}$, which we will call "steering" matrix from now on, is also automatically determined:

$$\left[\mathbf{A}_\vartheta^{'}\right]_{mn} = b_{m-n}(\vartheta) = \frac{1}{K} \frac{\sin\left(\pi(m-n) - \vartheta K\right)}{\sin\left(\pi(m-n)/K - \vartheta\right)}.$$

For an odd numbers of orientations K, the steering matrix has the interesting property of being unitary ($\mathbf{A}_\vartheta^{'T} \mathbf{A}_\vartheta^{'} = \mathbf{I}$), such that $\left\| \mathbf{A}_\vartheta^{'T} \mathbf{U}_0 \right\|_F^2 = \|\mathbf{U}_0\|_F^2$.

Moreover, $\mathbf{A}_{\vartheta}^{'}$ is a *circulant* matrix, which means that it can be diagonalized by the DFT matrix $\mathbf{F}_K$:

$$\mathbf{A}_{\vartheta}^{'} = \mathbf{F}_K \mathbf{D}_{\vartheta} \mathbf{F}_K^H \tag{7.34}$$

with $\mathbf{D}_{\vartheta}$ a diagonal matrix depending on ϑ. We will exploit this property for estimating $\hat{\vartheta}$ in a computationally efficient manner. It can be shown that the diagonal elements of $\mathbf{D}_{\vartheta}$ are given by:

$$[\mathbf{D}_{\vartheta}]_{kk} = \begin{cases} \exp\left(2j\vartheta\left(k-1\right)\right) & 1 \leq k \leq \left(K+1\right)/2 \\ \exp\left(2j\vartheta\left(k-1-K\right)\right) & \left(K+1\right)/2 < k \leq K \end{cases} \tag{7.35}$$

Following equation (7.27), ϑ can be estimated as follows:

$$\hat{\vartheta} \quad = \quad \arg\max_{\vartheta \in [0,\pi]} g(\vartheta), \tag{7.36}$$

with $g(\vartheta) = \mathbf{x}_0^{'T} \mathbf{C}_b^{'-1} \mathbf{A}_{\vartheta}^{'T} \mathbf{y}^{'}$. Based on the diagonalization (7.34), we can alternatively write $g(\vartheta)$ as:

$$g(\vartheta) = \sum_k \tilde{b}_k \left[\mathbf{D}_{\vartheta}\right]_{kk}, \tag{7.37}$$

with $\tilde{b}_k = \mathbf{x}_0^{'T} \mathbf{C}_b^{'-1} \left[\mathbf{F}_K\right]_{:,m} \left[\mathbf{F}_K^H\right]_{m,:} \mathbf{y}^{'}$. Consequently, $g(\vartheta)$ is a trigonometric polynomial in ϑ of degree K (and period π), with at most $2K$ extrema in the interval $[0, \pi[$ [Powell, 1981, p. 150]. Finally, the test statistic is given by:

$$t_{\mathrm{JDE}}^{'} \quad = \quad g(\hat{\vartheta}) / \left\|\mathbf{U}_0\right\|_F^2 \tag{7.38}$$

We remind the reader that the range of possible values $\hat{\vartheta}$ is continuous and does not depend on the choice of the number of orientations K. In fact, even if the number of orientations K is very low (e.g. $K = 2$), it is possible to detect signals with arbitrary orientations.

7.3.3 Scale-unaware detection

In the case of scale-unaware detection, we consider signals with unknown size σ, but for example uniformly distributed on $[\sigma_{\min}, \sigma_{\max}]$, and $\alpha \equiv [\sigma]$. We prefer to express σ on a logarithmic scale because a relative change of the size of the signal, $a\sigma$, then results in a translation of the scaling variable in a logarithmic scale ($\log a + \log \sigma$). In the following, we will refer to this as shiftability in scale. In [Freeman and Adelson, 1991, Simoncelli and Freeman, 1995], shiftability in scale is developed in the context of periodic signals. Applied to frequency scales, this means that upshifting the high-frequency selective channels results in low-frequency selective channels (hence causing aliasing and hampering the scale estimation). The solution is to design scale-shiftable channels in continuous radial frequency, and to apply ideal low-pass filtering prior to sampling, to suppress aliasing. In polar-frequency coordinates we have:

$$f_{\mathrm{scale-shiftable}}^{(\sigma)}(\omega, \varphi) = \mathrm{sinc}\left(\mathrm{sign}\left(\omega\right) \log_2\left(\frac{|\omega|}{\pi} + \epsilon\right) - \sigma\right) I(|\omega| < \omega_{max}) \tag{7.39}$$

with ω the radial frequency, φ the angular frequency, ω_{max} the maximal radial frequency, ϵ a small positive number to make sure the result is defined for $\omega = 0$ (e.g. $\epsilon = 10^{-6}$) and $\mathrm{sinc}\,(\cdot)$ the sinc-function. The function $\log_2\left(\frac{|\omega|}{\pi} + \epsilon\right)$ defines a logarithmic warping of the radial frequencies to dyadic scales. A plot of this function is shown in Figure 7.14a. and the 1-d radial magnitude responses of the scale-shiftable channels (equation (7.39)) are given in Figure 7.14b. The elements of the transform matrix $\mathbf{A}_\sigma^{'}$ are given by:

$$\left[\mathbf{A}_\sigma^{'}\right]_{mn} = \mathrm{sinc}\left(((m-n)-\sigma)\right).$$

The scale estimate is given by (see equation (7.27)):

$$\hat{\sigma} \quad = \quad \arg \max_{\sigma \in [\sigma_{\min}, \sigma_{\max}]} \mathbf{x}_0^{'T} \mathbf{C}_b^{'-1} \mathbf{A}_\sigma^{'T} \mathbf{y}^{'}. \tag{7.40}$$

In contrast to Section 7.3.2, $\mathbf{A}'$ now has a Toeplitz structure. To find the maximum in (7.40), we first apply a brute-force search for a fixed number of uniformly spaced scales in $[\sigma_{\min}, \sigma_{\max}]$. Next, the estimate is refined using iterative Newton-Raphson optimization. Because the number of channels is usually relatively low (less than 10), the impact on the computation time is minimal.

7.3.4 Location-unaware detection using a scanning CHO

For the location unaware detection (where only $\mathbf{q}$ is unknown, $\alpha \equiv \{\mathbf{q}\}$), we will restrict ourselves to integer translations ($\mathbf{q}$ is an integer vector), although extensions to non-integer translations are also possible, based on e.g. bandlimited interpolation. As a reference position, we choose the center of the image: $\mathbf{q} = \mathbf{q}_c$. The matrix $\mathbf{A}_\vartheta$ should translate the spatial plane and transform $\mathbf{x}_p$ into $\mathbf{x}_0$ (i.e. move the signal from position $\mathbf{q}$ to position $\mathbf{q}_c$). In essence, it is possible to use the same trick as we did for rotation and scale unaware detection, based on steerable channels (in this context these channels are called *time/space-shiftable*). However, if we would like to have a fine spatial resolution, the required number of channels increases rapidly: for example, to build a grid of 20×20 spatially selective channels, 400 channels are needed. For the CHO this also requires at least 400 training images in order for $\mathbf{C}_b^{'}$ to be even invertible, while the actual number of images needed for correct operation may be 10 to 100 times larger. Moreover, channels that are shiftable in space can not be shiftable in scale at the same time (or vice versa) [Simoncelli and Freeman, 1995], as this would require channels that are simultaneously bandlimited and compactly supported, which is not possible [Daubechies, 1992]. So we conclude that constructing channels based on the same technique as in previous subsections is impractical.

Nevertheless, with a simple workaround it is possible to approximate the JDE observer. The first step of the sequential scheme is again the estimation of the missing parameters, in this case the unknown center position $\mathbf{q}$ of the

signal. Equation (7.25) amounts to:

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}} \mathbf{x}_0^T \mathbf{C}_b^{'-1} \mathbf{U}_{\mathbf{p}}^T (\mathbf{y}) \,. \tag{7.41}$$

Because $\mathbf{U}_{\mathbf{p}}$ represents shifted versions of the same channel matrix $\mathbf{U}_0$, we can efficiently compute channel responses for different positions in the image, as the projection operation $\mathbf{U}_{\mathbf{p}}^T (\mathbf{y})$ for all $\mathbf{q}$ can be performed using Fast Fourier Transforms. The location estimator then exhaustively scans the whole image for the highest test statistic, using the same set of channels. For this reason we will call this the *scanning technique*. The highest test statistic is subsequently used for making the decision of signal presence. Because the CHO basically scans the whole image to perform its detection task, this type of CHO is known as "scanning CHO" [Park et al., 2005].

We remark that compared to the JDE estimator, the position estimator (7.41) is based on a constant covariance matrix $\mathbf{C}_b^{'}$ in channel space. This is only efficient in case the image is spatially (wide-sense) stationary, while most medical image are not. An alternative would be to make the matrix $\mathbf{C}_b^{'}$ position-dependent, thus would again require a large number of training images.

7.3.5 More complex detection tasks

Now that we have described the channel selection in detail, we can have a look at more complicated SKS problems, in which multiple signal parameters are unknown. First of all, a brief summary of the channel choices is given in Table 7.3. Listed are various detection tasks, the type and number of channels to use, and the profile that needs to be changed for every situation. In general, the channels can be designed in polar coordinates as follows:

$$f^{(k,s)}(r,\varphi) = f^{(k)}_{\text{angular}}(r,\varphi) f^{(s)}_{\text{radial}}(r,\varphi),$$

with $f^{(k)}_{\text{angular}}(r,\varphi)$ the angular profile (impulse response) of the channels and with $f^{(s)}_{\text{radial}}(r,\varphi)$ the radial profile. For every detection task in Table 7.3, it is specified which profile needs to be changed. For example, in case the signal size is unknown, we should use the scale-shiftable radial profile, or $f^{(s)}_{\text{radial}}(r,\varphi) = f^{(s)}_{\text{scale-shiftable}}(\omega,\varphi)$. If, on top of that, the signal orientation is unknown, a steerable angular profile should be used: $f^{(k)}_{\text{angular}}(r,\varphi) = f^{(\vartheta_k)}_{\text{steer}}(r,\varphi)$. This way we obtain channels that are jointly shiftable in scale and in orientation (also see the discussion in [Simoncelli et al., 1992] on this topic). For the tasks listed, there never occur conflicts when several tasks are being combined.

Table 7.3: Overview of the proposed channels to use for specific detection tasks.

Detection task	Proposed channels	Profile	Number of channels
SKE	Laguerre-Gauss, DOG	Radial	K
SKS, signal amplitude unknown	Laguerre-Gauss, DOG	Radial	K
SKS, signal orientation unknown	steerable Laguerre-Gauss, steerable DOG	Angular	K K
SKS, signal size unknown	rotationally symmetric scale-shiftable channels	Angular	S
SKS, signal location unknown	Any from the above	N/A	Any
SKS, all parameters unknown	steerable scale-shiftable channels	Radial+angular	KS

7.4 Results

7.4.1 Detection performance experiment

As a first experiment, we generated two sets of 10 000 simulated 2D Gaussian correlated background (CGB) images with variance 1 and with power spectral density $P(\omega) \propto \exp\left(-\|\omega\|^2/(2\sigma_{BK}^2)\right)$ where the parameter is given by $\sigma_{BK} = 10$. One set contains signal-free BK images, a second set contains BK images with an asymmetric elliptical signal inserted in the center (see (7.6)), with parameters $D_{11} = 1$, $D_{22} = 10$ and with a uniformly distributed orientation $\vartheta \in [0, \pi]$. We trained different CHO models to this training set, and generated a third and fourth sets with the same parameters, for the testing phase. The different CHO models used in this experiment are:

1. CHO with the first 10 rotationally symmetric LG channels, with parameter[6] $a_{LG} = 1.9$ (*LG-symmetric*),
2. The CHO for SKS tasks (Section 7.3), with $10 \cdot K$ steerable LG channels and with $a_{LG} = 1.9$, for different number of orientations K (*LG-steerable*).

Here, *LG-symmetric* is trained on a signal that is averaged over all possible orientations (it can be shown that this is an approximation to the ideal linear observer for the considered SKS task). We give results for the test sets in terms of Area Under the ROC curve (AUC) (Figure 7.15(a)) and estimation performance (MSE in estimating ϑ, Figure 7.15b). We also compare the detection performance to the performance of the ideal observer which has full knowledge of ϑ in order to have a theoretical upper bound of the detection performance.

We note that for signal amplitudes > 0.0075, the *LG-steerable* significantly outperforms *LG-symmetric* both in terms of detection performance and estimation performance. The detection performance even improves when using more orientations. In this case, the orientation selectivity is increased, how-

[6]This constant has been chosen experimentally to maximize the detection performance in terms of AUC.

ever, one should take into account that the number of required training images also becomes larger.

For signal amplitudes < 0.0075, *LG-symmetric* has a slightly better detection performance. We remark that in this case the signal amplitude is extremely low and the signal is completely invisible to the human eye, hence this amplitude range is not very suited for human observer experiments. We conclude that a significant improvement in detection performance is obtained when these models are extended to deal properly with SKS tasks.

7.4.2 Artificial asymmetric lung nodule detection experiment

As a second experiment, we train the new CHO model on real radiographic images of the chest [Shiraishi et al., 2000], with simulated lung nodules. We use 5 scales of LG channels ($S = 5$, $a_{LG} = 1.9$) and 5 orientations of steerable channels. This gives a total of 25 channels. The CHO is trained on 200 non-overlapping 256×256 patches extracted from the radiographic images, to obtain pairs of signal-absent and signal-present images. For the signal-present images, simulated lung nodules with parameters $D_{11} = 2, D_{22} = 10$ (see Section 7.1.2) are added to the center of the image. The anisotropy of the nodule is a bit exaggerated compared to clinically realistic cases, for illustrative purposes. For training, we selected the signal amplitude such that the AUC value is in the range 0.90-0.95. This resulted in an amplitude $a = -0.1$ for a background intensity range of $[0, 1]$. Note that the amplitude is chosen to be negative (i.e. nodules are darker than the background), as this is the case for real lung nodules in radiographic images.

Next, we selected an image for testing and we added 10 simulated nodules to the image (see Figure 7.16(a)), with the same parameters as in the training phase, but now added at random positions and with uniformly distributed orientation $\vartheta \in [0, \pi]$. A scanning CHO is then applied to the image. The resulting decision variables are shown in Figure 7.16(b). Finally, based on the decision variables, a decision of signal presence is drawn. The detected nodules, together with their estimated orientations are shown in Figure 7.16(c). By looking at Figure 7.16(c) it seems that there is one false detection in a darker area of the image, however, this is not of importance here because this detected nodule is located in an irrelevant part of the image and could alternatively be omitted by simple thresholding techniques.

In the bottom row of Figure 7.16, the experiment is repeated for another image and with larger nodules. Again the CHO model is able to detect the nodules well, with only one missed detection. Figure 7.16(e) confirms that the nodule is correctly detected, it is simply not shown in Figure 7.16(f) because it overlaps with another nodule. Hence, the CHO gives accurate positions and orientation angles of the nodules in the image.

7.5 Multi-slice Observer Models

In the previous sections, we focused on the processing of 2D images, i.e. the detection of planar signals in projection images. In clinical practice there is a trend toward volumetric imaging in CT, MRI, PET/SPECT, 3D breast imaging. Consequently, it becomes more and more important to assess and optimize the image quality for volumetric images as well. Volumetric images are often presented to a physician as a stack of slices in which each slice is viewed for a fixed amount of time. The browsing speed and browsing position can be adjusted by the physician; in this sense this is similar to a video imaging application. To assess volumetric images, one can think of CHO model designs for mimicking humans, e.g. by including temporal information in the detection process. For example, several studies (see e.g. [Dan et al., 1996]) have shown that the lateral geniculate nucleus (LGN) in the visual system of cats has specific temporal responses which tend to decorrelate the visual signals. Hence we could develop a CHO model that acts similarly. However, it is not clear yet how the HVS processes these visual signals when the images are presented in a stack browsing mode.

Given the complexity of this problem and the limited amount of information available on this topic in the literature, we use a bottom-up approach in our research: motivated by psychophysical evidence, we apply incremental extensions to the existing CHO model. Subsequently these extensions will be validated using psychovisual experiments in the near future. Several factors need to be taken into account here, such as anatomical properties of the information in the images, image acquisition parameters (e.g. slice thickness, spatial resolution), browsing speed, display technology being used (since LCD displays often have a slow temporal response), etc.

The architectures of two multi-slice CHO models, msCHO$_a$ from [Chen et al., 2001] and our proposed models msCHO$_b$ and msCHO$_c$ [Platiša et al., 2009b, Platiša et al., 2010b], are shown in Figure 7.17-Figure 7.18. To incorporate temporal information into the multi-slice CHOs, a sequential design is used: in the first stage, a 2D CHO is applied to the individual slices of the medical volume (Figure 7.17), yielding decision variables which indicate whether there is a signal present in each of the slices. In practice, we often know in advance that the support of the signal in the temporal direction is limited. Therefore, we define a temporal region of interest (ROI) containing the support of the signal. In a time-aware detection task, the exact time when the signal reaches a maximum amplitude is also known. In that case, the ROI is centered around that point in time and only the decision variables for the ROI are further given consideration. In the second stage, the decision variables of the 2D CHO in the ROI are passed to a 1D HO.

The three multi-slice CHO models differ in their use of the channel template matrix $\mathbf{w}_{\mathrm{CHO}}$. For msCHO$_a$, a channel template matrix is computed *for each slice* and applied to the same slice. On the other hand, for msCHO$_b$, only one channel template matrix is calculated (typically for the slice where the signal amplitude is maximal). Subsequently this channel template matrix is applied to

all slices in the ROI. msCHO$_a$ is motivated by the common assumption in the literature that humans tune signal-matched filters to the varying background information and consequently use a *separate* filter for each individual slice in the stack. This is in contrast to msCHO$_b$, which assumes that humans are more likely to examine (a number of) multiple consecutive slices of the stack with a *unique* signal matched filter. For msCHO$_c$, the channel template matrix is calculated for all channels and slices.

Our simulation results in [Platiša et al., 2010b] indicate that msCHO$_a$ and msCHO$_b$ perform equally well, while msCHO$_c$ generally outperforms msCHO$_a$ and msCHO$_b$ in the detection signals with exactly known parameters. However, by its design msCHO$_a$ and msCHO$_b$ are less susceptible to dimensionality problems and less sensitive to the number of training samples than msCHO$_c$. Because the number of training volumes available is often very limited in practice, robustness to small training sets also plays an important role in the model. To investigate which multi-slice model best corresponds to the human observer, psychovisual experiments are required. This is currently the topic of ongoing research.

Finally, in the multi-slice models, there is also uncertainty associated with the signal properties (e.g. the spread in time) and we can again consider SKS detection tasks. Because of the sequential design of our multi-slice CHOs, these detection tasks can be efficiently solved using the techniques presented in Section 7.3.

7.6 Conclusion

As an objective approach to medical image quality assessment, studies with mathematical model observers promise to eventually replace time-consuming human observer studies. The quality of images is usually assessed for a given task of interest, such as the detection of abnormalities in an image. The ideal observer (IO) and the channelized Hotelling observer (CHO) are good candidates for this, the IO for providing an upper bound for the detection performance and the CHO as a practical approximation to the IO. One major drawback of these model observers is that they do not take intrinsic uncertainty with respect to the signal properties into account.

In this chapter, we explained a new theoretical framework for deriving CHOs in signal-known-statistically (SKS) tasks, i.e. when the signals have unknown parameters. We showed that the ideal linear observer is often inadequate for this task and that the optimal observer is nonlinear. Joint estimation and detection theory has proven to be very useful for this, as this allowed to split up the task in a *nonlinear* estimation part, and a *linear* detection part. We have derived CHOs for these tasks, and we have explained that the SKS task poses some additional requirements on the channel design. For example, for scale-unaware detection, the channels need to be *shiftable in scale*, while for rotationally-unaware detection, the channels should be *steerable*. The combination of location- and orientation unaware detection leads to optimal estimation

in the steerable pyramid domain (Chapter 2). This opens the way for more general multiresolution representations, as presented in Chapter 2 to be used for medical image quality assessment. Furthermore, it could be interesting to introduce the image models from Chapter 3 as well, even though this requires some further investigation.

Coincidentally or not, the findings of scale- and orientation-selective channels also agree with the presence of frequency- and orientation-responsive neurons in the human primary visual cortex. Hence this may suggest that the new model observers behave similar to humans in detection tasks, as the intrinsic uncertainty of the signal is taken into account.

Finally, we also explained two designs of two multi-slice CHO models for assessing the quality of volumetric images. These designs are again driven by properties of the HVS. Although psychophysical validation studies with human observers are outside the scope of this dissertation, these studies will be the topic of our future work.

Our contributions to CHOs for SKS tasks have already resulted into one conference paper [Goossens et al., 2010f], one journal paper is currently in preparation [Goossens et al., 2010e]. The work on multi-slice CHO models has resulted in five publications as co-author [Platiša et al., 2009c, Platiša et al., 2009b, Platiša et al., 2009a, Platiša et al., 2010a, Platiša et al., 2010b].

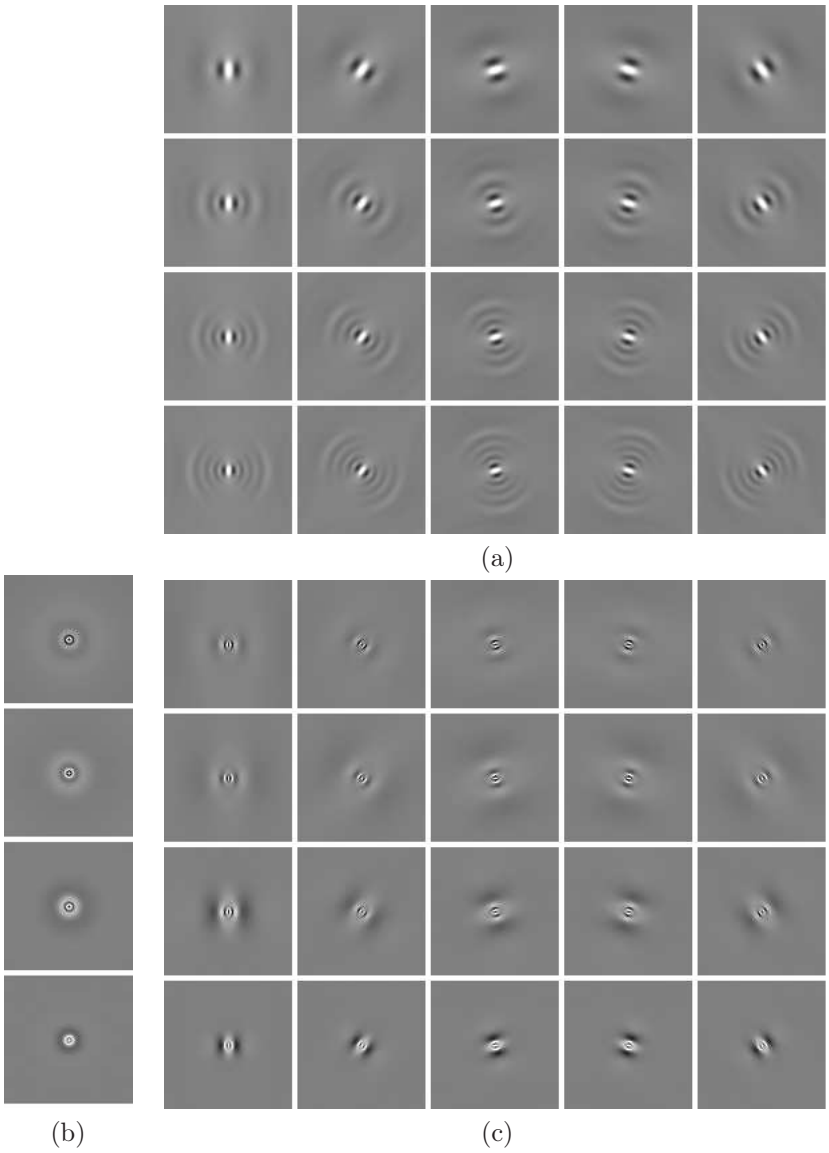
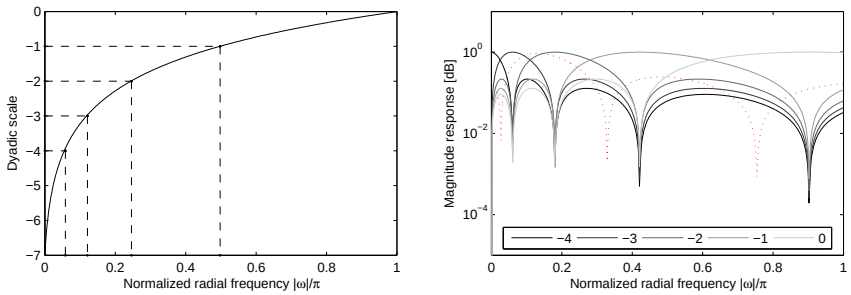
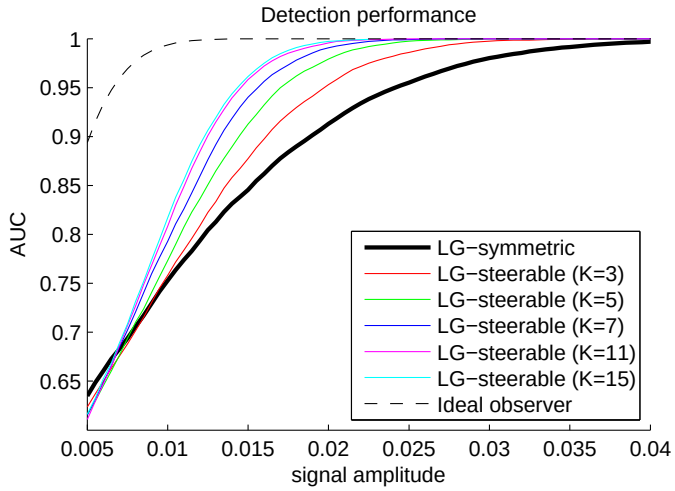


Figure 7.13: Examples of channels used for projections. (a) Orientation-steerable LG channels - for SKS tasks with unknown signal orientation (b) Rotationally symmetric scale-shiftable channels - for SKS tasks with unknown signal size (c) Orientation-steerable and scale-shiftable channels - for SKS tasks with unknown signal orientation and size. Gray corresponds to intensity 0, white with positive intensities and black with negative intensities.

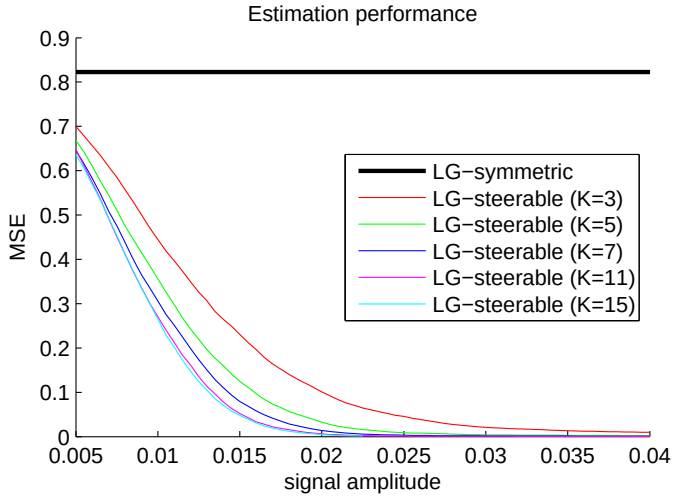


(a) (b)

Figure 7.14: (a) The function $\log_2\left(\frac{|\omega|}{\pi}+\epsilon\right)$ which maps radial frequencies to dyadic scales (b) Magnitude responses of the filters $f^{(\sigma)}_{\mathrm{scale-shiftable}}(\omega,\varphi)$ for different scales σ listed in the legend. The dotted line is the interpolated magnitude response for dyadic scale $\sigma = 0.2$.



(a)



(b)

Figure 7.15: Comparison of the detection and estimation performance of different CHO models: (a) AUC detection performance, (b) Mean square error (MSE) of the estimated angle $\hat{\vartheta}$.

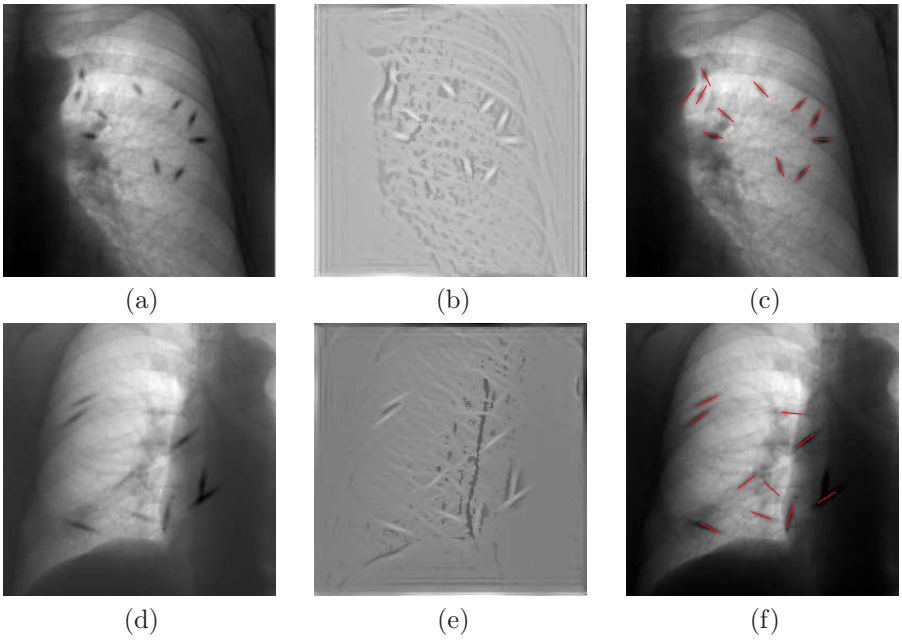


Figure 7.16: (a),(d) Two chest radiographs from the database from [Shiraishi et al., 2000] with simulated lung nodules. (b),(e) Decision variables t'_{JDE} of a scanning CHO applied to the chest images (white corresponds to high values, gray to low values). (c),(f) Detected lung nodules with their estimated orientation.

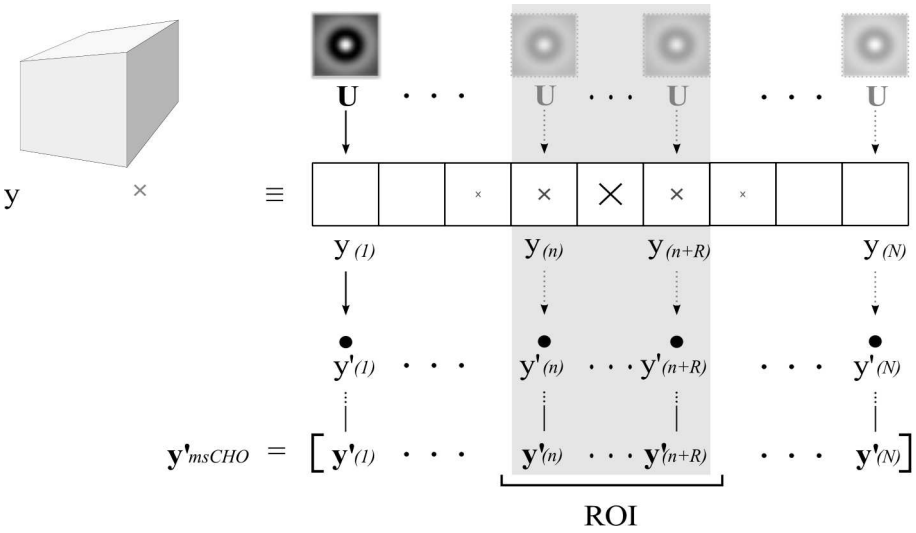


Figure 7.17: Multi-slice CHO - part I. The slices $\mathbf{y}_{(1)}, ..., \mathbf{y}_{(N)}$ are first processed using 2D-LG channels, to obtain the channel outputs $\mathbf{y}'_{msCHO} = \left\{\mathbf{y}'_{(1)}, ..., \mathbf{y}'_{n}\right\}$. Figure taken from [Platiša et al., 2010b].

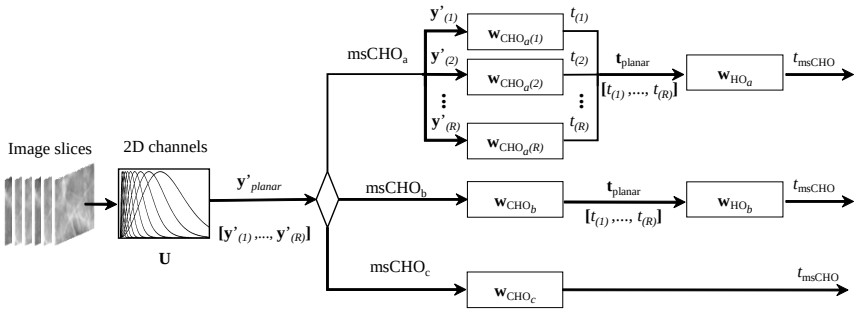


Figure 7.18: Multi-slice CHO - part II. msCHO_a applies a CHO to each slice in the ROI, with a different channel template matrix $\mathbf{w}_{\text{CHO}\,a}(r)$ per slice. Next, a Hotelling observer with channel template matrix $\mathbf{w}_{\text{HO}\,a}$ computes the final channel responses. msCHO_b is identical to msCHO_a except that the same channel template matrix$\mathbf{w}_{\text{HO}\,b}$ is used for all slices of the stack. Figure taken from [Platiša et al., 2010b].

8

Concluding remarks

Because of physical limitations and cost restrictions, noise, originating from sensors or detectors and electronic amplifiers in the acquisition devices and other image degradations can not be avoided in digital imaging applications. Fortunately, by the advent of powerful computers and availability of computational resources in the last decades, more sophisticated and intelligent image processing techniques, prove to be useful to significantly improve the quality of digital images.

The key solution for designing efficient image restoration algorithms lies in 1) accurate modeling of noise in digital images and 2) incorporating prior knowledge with respect to the ideal, undegraded image. The more prior knowledge that can be utilized, the better the performance of the image processing algorithms will be. Multiresolution geometrical transforms play an important role here, since these transforms enable representing images using a small number of significant coefficients. Consequently, the statistical image modeling task for *images* is much easier in a multiresolution transform domain. Also, the frequency and orientation analyzing properties of a multiresolution transform can be exploited in order to describe and estimate the characteristics of *noise* in images. We illustrated this for the estimation of (non)stationary colored noise in images and also for the more complicated modeling of noise in CT images.

To demonstrate how statistical models for images and noise can be utilized in image processing applications, we considered various image restoration applications. We saw that many restoration tasks, such as the estimation and suppression of colored noise in images, can directly be performed in a multiresolution transform domain. For other, slightly more complicated restoration tasks, we discussed an iterative optimization strategy that is based on Bregman iterations. The Bregman framework can be seen as a generic restoration methodology, in which an arbitrary image model can be combined with an arbitrary noise model for a given multiresolution transform. This approach is very flexible and powerful in the sense that complicated restoration problems can be solved with limited effort with only a relatively small number of lines of programming code.

In some situations, a clever choice of a multiresolution transform and process-

ing strategy enables devising sophisticated but computationally efficient techniques. As an example, we discussed the complex-wavelet based demosaicing algorithm that can seamlessly be combined with denoising. To arrive at such a solution, we judiciously exploited spatial, directional and frequency localization properties of the complex wavelets.

We found another interesting example of this phenomenon in the quality assessment of medical images, where again multiresolution concepts come into play. In the literature, one specific model observer, the Channelized Hotelling observer has proven to be very useful for the automatic task of detection of abnormalities in images. However, recent studies have shown that this mathematical model observer generally outperforms human observers. By incorporating uncertainty with respect to the abnormality (e.g. tumor size, contrast...) we developed a more complex model observer that can be efficiently implemented in a directional multiresolution transform domain while still performing optimally in MAP sense.

Many models that we encountered in this dissertation are inspired by (or related to) properties of the human visual system (HVS). For example, the presence of orientation and frequency selective responses of simple neurons is a motivation for the use of directionally selective representations. The knowledge of the HVS may provide extra insights for developing efficient image processing algorithms and conversely we "may" learn more about the human visual system.

Brief summary of the main contributions

The main novelties and contributions presented in this dissertation are the following. In Chapter 2 we presented a novel design technique for complex wavelet filters for the first scale of the DT-CWT in order to improve the directional selectivity of the transform. Next, we proposed a novel design of a discrete shearlet transform that combines low redundancy, shift invariance and directional analysis properties. In Chapter 3, we developed two new statistical models for image multiresolution transform coefficients: an improved intra-scale model (MPGSM), which captures the variability of the spatial covariance matrix, and a joint intra/inter-scale model, which also incorporates dependencies between transform coefficients within different scales.

In Chapter 4 we presented novel EM algorithms for the estimation of both stationary and non-stationary correlated noise in images. We established an approximate analytical relationship between the camera response function and the noise level function, which is useful for modeling and estimating signal-dependent noise in images.

In Chapter 5 we presented an improved non-local means (NLMeans) algorithm. A different robust weighting function combined with a post-processing filter resulted in a significant improvement both visually and in PSNR. We also discussed how the NLMeans filter can be used for denoising images corrupted with non-stationary and correlated noise. We derived exact MAP and MMSE estimation rules for the Bessel K Form density; the Bessel K Form MAP

estimator is particularly suited for use within the Bregman-based restoration framework.

Next, we introduced the Vector-ProbShrink denoising method, which is based on the joint inter/intra-scale statistical model from Chapter 3. The proposed MMSE estimator for the MPGSM model offers a vast improvement in PSNR compared to existing wavelet-based denoising methods, especially for texture-rich images. Both the Vector-ProbShrink and MPGSM denoising are currently among the state-of-the-art wavelet-based denoising techniques, where MPGSM generally outperforms Vector-ProbShrink, but at a much higher computational cost.

We also presented a complex wavelet-packet based demosaicing method, which is an improved version of the wavelet-packed based method of Hirakawa. In particular, we provided a solution for dealing with the discoloration artifacts and the loss of high-pass luminance frequencies of the method of Hirakawa. We have further shown how the Bregman optimization framework can be used for solving more "difficult" image restoration problems. This resulted in a new technique for joint denoising and deblurring, a new technique for estimating the Power Spectral Density of stationary Gaussian noise from an image jointly with denoising, and a technique to remove signal-dependent noise from images.

In Chapter 6 we introduced a new non-stationary model for noise in computed tomography images (CT) reconstructed by the filtered backprojection algorithm. In Chapter 7 we derived new Channelized Hotelling observers for detecting signals with unknown parameters in medical images.

Current progress of the research and directions for future work

Several directions are open for future research or are currently under further investigation. For image restoration, the most promising approach seems to be the use of the Bregman optimization framework, because 1) the framework is very generic, powerful and can be adapted to many applications, 2) the resulting Bregman-based algorithms are relatively simple to implement. One disadvantage is that this approach is computationally very intensive on a Central Processing Unit (CPU). It can take several minutes on a PC to process one single grayscale image. Fortunately, the Bregman class of algorithms tend to transfer well to parallel computing architectures, such as Graphical Processing Units (GPU). As a proof of concept, we already implemented one of the Bregman algorithms, in combination with the DT-CWT, on a recent GPU, for which the total processing time is in the order of a few seconds for a high-definition television color image.

In this dissertation, we presented several image models, noise models and multiresolution transforms. In fact, any combination of an image model, noise model and transform can be used to solve certain practical problems. In Chapter 5 we focused on the most interesting applications to illustrate the proposed solution methodology. During the writing of this dissertation, in our research

group, we have also been investigating Bregman-based demosaicing schemes, the use of the non-local prior from Chapter 3 for Bregman-based image restoration and MRI compressed sensing reconstruction techniques for arbitrary non-uniformly sampled K-space trajectories [Aelterman et al., 2010c]. This last technique allows to significantly reduce the acquisition time of an MRI image while maintaining image quality.

There is also a lot of room to improvement to the statistical image models. We believe that most gain can be achieved by efficiently incorporating non-local dependencies in the multiresolution transform models (see our discussion in Chapter 3). Promising in this respect are Bayesian network, Markov Random Field approaches with message passing schemes.

Next, another interesting application is the use of the proposed CT noise model to improve the reconstruction quality of low-dose CT images, for which we have shown one visual result at the end of Chapter 6. Within the context of the IBBT-ICA4DT project, we initiated a small preliminary study involving physicians. The study indicated that the prototype performed better than other existing techniques and even resulted in images of better quality than the unprocessed image in some cases. By this encouraging success, the method has been subject to further development and a European patent application is now in preparation.

Finally, the mathematical model observers deserve some extra attention. In the near future, we will perform a validation study of the presented CHO models by comparing to human observer experiments. We are also planning a study on the application of the proposed CHO models to MRI quality evaluation for the detection of multiple-sclerosis, in collaboration with the Ghent University Hospital and Université d'Angers in France. Within the context of the IBBT-CIMI project, our research group is also extending the CHO models to color data, as the use of color information for visualization becomes more and more important in medical imaging.

A

Appendix A: convergence of the EM algorithm for GSMs

In Appendix A, we will proof that for the constrained EM algorithm from Section 4.2.2 with $\sum_{k=1}^{K} \lambda_k = 1$ (which we can accomplish by the degrees of freedom we have), the update direction of the EM update equations (4.11)-(4.14) has a positive projection onto the gradient of the likelihood function, such that the likelihood function increases in every iteration. Convergence properties are then entirely the same as for the (unconstrained) EM algorithm for Gaussian Mixtures [Dempster et al., 1977]. The proof is similar as the proof for the EM algorithm for Gaussian Mixtures [Xu and Jordan, 1996], with a few modifications that we will describe more into detail below. In the following, we will simplify the notations by denoting by $\mathbf{\Theta}^i = \{\mathbf{C}_x, \mathbf{C}_w, \mathbf{C}_k, \alpha_k\}$ the set of parameters at iteration i, and by denoting by $\mathbf{\Theta}^{i+1} = \left\{\hat{\mathbf{C}}_x, \hat{\mathbf{C}}_w, \hat{\mathbf{C}}_k, \hat{\alpha}_k\right\}$ the set of parameters at iteration $i+1$. We will consider the EM maximization step from iteration i to $i+1$. First, we note that according to (4.16) we have the following relationship:

$$\hat{\mathbf{C}}_k = z_k \hat{\mathbf{C}}_x + \hat{\mathbf{C}}_w = \sum_{l=1}^{K} \hat{\mathbf{C}}_l^{(1)} \beta_{k,l} \tag{A.1}$$

with

$$\beta_{k,l} = \frac{z_l \left(\mu_1 - z_k\right) + \mu_1 z_k - \mu_2}{\mu_1^2 - \mu_2} \lambda_k \tag{A.2}$$

where we assume that $\mu_1^2 \neq \mu_2$ (see Section 4.2.2). The matrix $\boldsymbol{\beta} = [\beta_{k,l}]$ has a number of interesting properties that are fairly easy to show: 1) the matrix is positive semidefinite (SD) matrix with eigenvalues either 1 or 0, 2) the matrix is idempotent ($\boldsymbol{\beta}^2 = \boldsymbol{\beta}$) with rank 2, 3) $\sum_{l=1}^{K} \beta_{k,l} = 1$ and 4) for $K = 2$, $\boldsymbol{\beta}$ is the identity matrix ($\boldsymbol{\beta} = \mathbf{I}$).

Equation (4.10) can then be equivalently written as:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i) = \mathrm{E}\left[\log f_{\mathbf{y},k|\boldsymbol{\Theta}}\left(\mathbf{y}, k|\boldsymbol{\Theta}\right)|\mathbf{y}, \boldsymbol{\Theta}\right] - \sum_{k=1}^{K} \lambda_k \left\| \sum_{l=1}^{K} \beta_{k,l}\mathbf{C}_k - z_k\mathbf{C}_x - \mathbf{C}_w \right\|_F^2, \tag{A.3}$$

such that the solution $(\mathbf{C}_x, \mathbf{C}_w)$ that maximizes $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i)$ automatically satisfies the constraint $z_k\mathbf{C}_x + \mathbf{C}_w = \sum_{l=1}^{K} \beta_{k,l}\mathbf{C}_k$. Next, the gradient of $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i)$ with respect to $\mathbf{C}_k$ is given by:

$$\frac{\partial}{\partial \mathbf{C}_k} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i) = -\frac{1}{2}\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right) \mathbf{C}_k^{-1}\left(\mathbf{C}_k - \mathbf{y}_j\mathbf{y}_j^T\right)\mathbf{C}_k^{-1}. \tag{A.4}$$

where the partial derivative of the second term in (A.3) with respect to $\mathbf{C}_k$ is zero because the constraint holds at any time. Consequently, the EM update equation for $\mathbf{C}_k$ (equation (4.12)) can be written as:

$$\hat{\mathbf{C}}_k = \mathbf{C}_k + \frac{\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right)\mathbf{y}_j\mathbf{y}_j^T}{\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right)} - \mathbf{C}_k \tag{A.5}$$

$$= \mathbf{C}_k + \left(\frac{2\mathbf{C}_k}{\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right)} \frac{\partial}{\partial \mathbf{C}_k} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i)\right)\mathbf{C}_k \tag{A.6}$$

By denoting $\mathrm{vec}\left(\cdot\right)$ the operation that stacks the columns of the matrix into a vector, (A.6) can also be written as (see [Xu and Jordan, 1996]):

$$\mathrm{vec}\left(\hat{\mathbf{C}}_k\right) = \mathbf{B}_k \mathrm{vec}\left(\frac{\partial}{\partial \mathbf{C}_k} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i)\right) \quad \text{with} \quad \mathbf{B}_k = \frac{2\mathbf{C}_k \otimes \mathbf{C}_k}{\sum_{j=1}^{N} \mathrm{P}\left(z = z_k|\mathbf{y}_j\right)} \tag{A.7}$$

where $'\otimes'$ denotes the Kronecker product. Furthermore, it has been shown in [Xu and Jordan, 1996] that $\mathbf{B}_k$ is positive definite with probability 1, if N is sufficiently large. Due to the idempotency of $\boldsymbol{\beta}$ and the constraint $\sum_{l=1}^{K} \beta_{k,l}\mathbf{C}_k = z_k\mathbf{C}_x + \mathbf{C}_w$, it follows that:

$$\mathrm{vec}\left(\hat{\mathbf{C}}_k\right) = \sum_{l=1}^{K} \beta_{k,l} \mathrm{vec}\left(\hat{\mathbf{C}}_l\right) \tag{A.8}$$

$$= \mathrm{vec}\left(\mathbf{C}_k\right) + \left(\sum_{l=1}^{K} \beta_{k,l}\mathbf{B}_l\right) \mathrm{vec}\left(\frac{\partial}{\partial \mathbf{C}_l} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i)\right). \tag{A.9}$$

If we group all parameters $\hat{\mathbf{C}}_k, k = 1, ...K$ into a column vector

$$\mathbf{C}^\star = \left[\mathrm{vec}\left(\hat{\mathbf{C}}_1\right)^T \cdots \mathrm{vec}\left(\hat{\mathbf{C}}_K\right)^T\right]^T,$$

we will have that:

$$\hat{\mathbf{C}}^\star = \mathbf{C}^\star + \mathbf{B}^\star \frac{\partial}{\partial \mathbf{C}^\star} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i) \tag{A.10}$$

where $\mathbf{B}^{\star} = \sum_{k,l=1}^{K} \beta_{k,l} \left(\mathbf{E}_k \otimes \mathbf{B}_l \right)$. Here, $\mathbf{E}_k$ is a diagonal matrix with $[\mathbf{E}_k]_{ll} = \delta\left(k-l\right)$. Since the convex combination of positive SD matrices $\sum_{l=1}^{K} \beta_{k,l} \mathbf{B}_l$ is positive SD, and because the Kronecker product and the sum of two positive SD matrices are positive SD, $\mathbf{B}^{\star}$ is positive SD as well. Similarly, for the mixture weights $\boldsymbol{\alpha} = [\alpha_k]$ it holds that [Xu and Jordan, 1996]:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \mathbf{A} \frac{\partial}{\partial \boldsymbol{\alpha}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i) \tag{A.11}$$

with $\mathbf{A}$ a positive definite matrix (with probability 1). Now, if we combine (A.10) and (A.11), we can write for the complete parameter set:

$$\boldsymbol{\Theta}^{i+1} = \boldsymbol{\Theta}^i + \mathbf{P} \left. \frac{\partial}{\partial \boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i) \right|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^i} \tag{A.12}$$

where $\mathbf{P}$ is a positive definite matrix (with probability 1). Next, by projecting the update direction $\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$ onto the gradient direction and by using the positive definiteness of $\mathbf{P}$ we find that:

$$\left(\left. \frac{\partial}{\partial \boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^i) \right|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^i} \right)^T \left(\boldsymbol{\Theta}^{i+1} - \boldsymbol{\Theta}^i \right) > 0, \tag{A.13}$$

such that every update step $\boldsymbol{\Theta}^{i+1} - \boldsymbol{\Theta}^i$ has a positive projection onto the gradient of the likelihood function. Consequently, every EM update step increases the likelihood function of the data. □

B

Appendix B: Derivation of the local NSD for parallel beam CT

In Appendix B, we derive the local NSD created by one line of the line bundle $r\cos(\varphi-\vartheta)=t$; i.e. for one particular $\vartheta=\vartheta_0$. First, we consider the FBP algorithm for parallel beam CT (see Section 6.1.1). Therefore, we introduce:

$$\begin{aligned} P^m_{\vartheta_0(\vartheta,t)} &= \delta(\vartheta-\vartheta_0)\,P^m(\vartheta,t) \\ &= \delta(\vartheta-\vartheta_0)\left[P(\vartheta,t)-\sigma\left(P(\vartheta,t)\right)\nu(\vartheta,t)\right] \end{aligned} \tag{B.1}$$

In (6.27), we found that for one spike, the local NSD in orientation ϑ, is proportional to $\delta(\vartheta-\vartheta_0)\,|\omega|\,|G(\omega)|^2$. Under the condition that the projection data $P^m_{\vartheta_0(\vartheta,t)}$ is locally stationary in the t-direction, the local NSD at position (r,φ) has a radial component that only depends on the FBP filter and is given by:

$$S^{(\vartheta_0)}_{(r,\varphi)}(\vartheta,\omega)=U(r,\varphi,\vartheta)\,\delta(\vartheta-\vartheta_0)\,|\omega|\,|G(\omega)|^2 \tag{B.2}$$

with $U(r,\varphi,\vartheta)$ an unknown function that we still need to determine. Now, if we denote $\mu_{\vartheta_0}(r,\varphi)$ as the FBP reconstruction of $P^m_{\vartheta_0(\vartheta,t)}$, the variance of $\mu_{\vartheta_0}(r,\varphi)$ is given by the volume under to NSD $S^{(\vartheta_0)}_{(r,\varphi)}(\vartheta,\omega)$:

$$\begin{aligned} \mathrm{Var}\left[\mu_{\vartheta_0}(r,\varphi)\right] &= U(r,\varphi,\vartheta)\int_0^{\pi}\mathrm{d}\vartheta\,\delta(\vartheta-\vartheta_0)\int_{-\infty}^{+\infty}\mathrm{d}\omega\,|\omega|\,|G(\omega)|^2 \\ &= U(r,\varphi,\vartheta) \end{aligned} \tag{B.3}$$

where we used the normalization imposed to the FBP filter (6.7). The variance of $\mu_{\vartheta_0}(r,\varphi)$ can also be computed alternatively using ensemble statistics:

$$\begin{aligned} \mathrm{Var}\left[\mu_{\vartheta_0}(r,\varphi)\right] &= \mathrm{Var}\left[\int_{-\infty}^{+\infty}\mathrm{d}t P^m(\vartheta,t)q\left(t-r\cos(\varphi-\vartheta)\right)\right] \\ &= \int_{-\infty}^{+\infty}\sigma^2\left(P(\vartheta,t)\right)q^2\left(t-r\cos(\varphi-\vartheta)\right)\mathrm{d}t \end{aligned} \tag{B.4}$$

By combining equations (B.2), (B.3), (B.4), we arrive at the following expression for the local NSD:

$$S^{(\vartheta_0)}_{(r,\varphi)}(\vartheta,\omega) = \delta(\vartheta-\vartheta_0)\,|\omega|\,|G(\omega)|^2 \cdot \int_{-\infty}^{+\infty} \sigma^2(P(\vartheta,t))\,q^2(t-r\cos(\varphi-\vartheta))\,\mathrm{d}t \qquad \text{(B.5)}$$

which is separable in the polar frequency coordinates (ϑ,ω).

Bibliography

[Abbey and Barrett, 2001] Abbey, C. and Barrett, H. (2001). Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. *J. Opt. Soc. Am. A*, 18(3):473–488.

[Abdelnour and Selesnick, 2001] Abdelnour, A. F. and Selesnick, I. W. (2001). Design of 2-band orthogonal near-symmetric CQF. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 3693–3696.

[Abramovich et al., 1998] Abramovich, F., Sapatinas, T., and Silverman, B. (1998). Wavelet thresholding via a Bayesian approach. *J. of the Royal Statist. Society B*, 60:725–749.

[Abramowitz and Stegun, 1964] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions*. New York: Dover. ninth dover printing, tenth GPO printing ed.

[Achim et al., 2001a] Achim, A., Bezerianos, A., and Tsakalides, P. (2001a). Novel Bayesian multiscale method for speckle removal in medical ultrasound images. *IEEE Trans. Medical Imaging*, 20(8):772–783.

[Achim et al., 2001b] Achim, A., Bezerianos, A., and Tsakalides, P. (2001b). Wavelet-based ultrasound image denoising using an alpha-stable prior probability model. In *Proc. International Conference on Image Processing*, volume 2, pages 221–224.

[Aelterman et al., 2010a] Aelterman, J., Deblaere, K., Goossens, B., Pižurica, A., and Philips, W. (2010a). Dual Tree Complex Wavelet-Based Denoising of correlated noise in 3D Magnetic Resonance Imaging. Under revision.

[Aelterman et al., 2009] Aelterman, J., Goossens, B., Luong, H., Pižurica, A., and Philips, W. (2009). Locally Adaptive Complex Wavelet-Based Demoisaicing for Color Filter Array Images. In *SPIE Electronic Imaging 2009*, pages 72480J–72480J–12, San José, CA, USA.

[Aelterman et al., 2008] Aelterman, J., Goossens, B., Pižurica, A., and Philips, W. (2008). Removal of Correlated Rician Noise in Magnetic Resonance Imaging. In *Proc. 16th European Signal Processing Conference (EUSIPCO)*.

[Aelterman et al., 2010b] Aelterman, J., Goossens, B., Pižurica, A., and Philips, W. (2010b). *Recent Advances in Signal Processing*, chapter Suppression of Correlated Noise. IN-TECH.

[Aelterman et al., 2010c] Aelterman, J., Luong, H., Goossens, B., Pižurica, A., and Philips, W. (2010c). Split Bregman based Reconstruction of non-uniformly sub-Nyquist sampled MRI data. submitted to Elsevier Signal Processing Journal.

[Aharon et al., 2006] Aharon, M., Elad, M., and Bruckstein, A. (2006). The K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on signal processing*, 54(11):4311–4322.

[Alleysson and de Lavarene, 2008] Alleysson, D. and de Lavarene, C. (2008). Frequency selection demosacking: A review and a look ahead. In *Proc. SPIE Visual Communications and Image Processing (VCIP-08)*, volume 6822, San José, CA, USA.

[Alleysson and Susstrunk, 2005] Alleysson, D. and Susstrunk, S. (2005). Linear demosaicing inspired by the human visual system. *IEEE Transactions on Image Processing*, 14(4):1–11.

[Anderson, 1963] Anderson, T. (1963). Asymptotic theory for Principal Component Analysis. *Annals of Mathematical Statistics*, 34(1):112–148.

[Andrews and Mallows, 1974] Andrews, D. and Mallows, C. (1974). Scale mixtures of normal distributions. *J. Royal Stat. Stoc.*, 36:99–102.

[Anscombe, 1948] Anscombe, F. J. (1948). The Transformation of Poisson, Binomial and Negative-Binomial Data. *Biometrika*, 35:245–254.

[Antonini et al., 1992] Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I. (1992). Image coding using wavelet transform. *IEEE Trans. Image Process.*, 1(2):205–220.

[Argenti et al., 2002] Argenti, F., Torricelli, G., and Alparone, L. (2002). Signal-dependent noise removal in the undecimated wavelet domain. In *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, volume 4, pages IV–3293–IV–3296.

[Averbuch et al., 2006] Averbuch, A., Coifman, R., Donoho, D., Elad, M., and Israeli, M. (2006). Fast and Acurrate Polar Fourier Transform. *Journal on Appl. and Comp. Harm. Analysis*, 21:145–167.

[Azzabou et al., 2007] Azzabou, N., Paragias, N., and F., G. (2007). Image Denoising Based on Adapted Dictionary Computation. In *Proc. of IEEE International Conference on Image Processing (ICIP)*, pages 109–112, San Antonio, Texas, USA.

[Baher, 2001] Baher, H. (2001). *Analog and Digital Signal Processing*. Wiley, Chichester.

[Bamberger and Smith, 1992] Bamberger, R. H. and Smith, M. J. T. (1992). A filter bank for the directional decomposition of images: theory and design. *IEEE Trans. Signal Process.*, 40(4):882–893.

[Banham et al., 1994] Banham, M., Galatsanos, N., Gonzalez, H., and Katsaggelos, A. (1994). Multichannel restoration of single channel images using a wavelet-based subband decomposition. *IEEE Trans. Image processing*, 3(6):821–833.

[Banham and Katsaggelos, 1996] Banham, M. R. and Katsaggelos, A. K. (1996). Spatially adaptive wavelet-based multiscale image restoration. *IEEE Trans. Image Process.*, 5(4):619–634.

[Banham and Katsaggelos, 1997] Banham, M. R. and Katsaggelos, A. K. (1997). Digital image restoration. *IEEE Signal Processing Magazine*, 14(2):24–41.

[Baraldi and Panniggiani, 1995] Baraldi, A. and Panniggiani, F. (1995). A refined gamma map sar speckle filter with improved geometrical adaptivity. *IEEE Trans. on Geosci. and Remote Sensing*, 33(5):1245–1257.

[Barrett, 1990] Barrett, H. H. (1990). Objective assessment of image quality: effects of quantum noise and object variability. *J. Opt. Soc. Am. A*, 7(7):1266–1278.

[Barrett et al., 1998] Barrett, H. H., Abbey, C. K., and Clarkson, E. (1998). Objective assessment of image quality. iii. roc metrics, ideal observers and likelihood-generating functions. *J. Opt. Soc. Am. A*, 15:1520–1535.

[Barrett et al., 1995] Barrett, H. H., Denny, J. L., Wagner, R. F., and Myers, K. J. (1995). Objective assessment of image quality. ii. fisher information, fourier crosstalk, and figures of merit for task performance. *J. Opt. Soc. Am. A*, 12(5):834–852.

[Barrett and Myers, 2004] Barrett, H. H. and Myers, K. J. (2004). *Foundations of Image Science*. John Wiley and Sons, New York.

[Bartholomew, 1987] Bartholomew, D. (1987). *Latent Variable Models and Factor Analysis*. London: Charles Griffin & Co. Ltd.

[Basseville et al., 1992] Basseville, M., Benveniste, A., Chou, K., Golden, S., Nikoukhah, R., and Willsky, A. (1992). Modeling and estimation of multiresolution stochastic processes. *IEEE Trans. Inform. Theory*, 38(2):766–784.

[Bayer, 1976] Bayer, B. (1976). Color imaging array. United States Patent 3971065.

[Bayram and Selesnick, 2008] Bayram, I. and Selesnick, I. W. (2008). On the Dual-Tree Complex Wavelet Packet and M-Bandtransforms. *IEEE Trans. Signal Process.*, 56(6):2298–2310.

[Beal, 2003] Beal, M. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London.

[Bell and Sejnowsky, 1997] Bell, A. and Sejnowsky, T. (1997). The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3201–3439.

[Bellman, 1961] Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

[Belzer et al., 1995] Belzer, B., Lina, J.-M., and Villasenor, J. (1995). Complex, linear-phase filters for efficient image coding. *IEEE Trans. Signal Process.*, 43(10):2425–2427.

[Benazza-Benyahia and Pesquet, 2004] Benazza-Benyahia, A. and Pesquet, J.-C. (2004). An extended sure approach for multicomponent image denoising. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, 2:ii – 945–8.

[Benazza-Benyahia and Pesquet, 2005] Benazza-Benyahia, A. and Pesquet, J.-C. (2005). Building robust wavelet estimators for multicomponent images using Steins' principle. *IEEE Trans. Image Proc.*, 14(11):1814–1830.

[Bilcu and Vehvilainen, 2007] Bilcu, R. C. and Vehvilainen, M. (2007). Fast nonlocal means for image denoising. In Martin, R. A., DiCarlo, J. M., and Sampat, N., editors, *Proc. SPIE Digital Photography III*, volume 6502. SPIE.

[Bioucas-Dias and Figueiredo, 2007] Bioucas-Dias, J. and Figueiredo, M. (2007). A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans. Image Processing*, 16(12):2992–3004.

[Bishop et al., 1998] Bishop, C. M., Svensen, M., and Williams, C. (1998). Gtm: The Generative Topographic Mapping. *Neural Computation*, 1(1):215–234.

[Black et al., 1998] Black, M., Sapiro, G., Marimont, D., and Heeger, D. (1998). Robust anisotropic diffusion. *IEEE Transactions on Image Processing*, 7(3):421–432.

[Bochud et al., 1999] Bochud, F. O., Abbey, C. K., and Eckstein, M. P. (1999). Statistical texture synthesis of mamographic images with clustered lumpy backgrounds. *Opt. Express*, 4(11):33–43.

[Borel et al., 1996] Borel, C., Cooke, B., and Laubscher, B. (1996). Partial Removal of Correlated noise in Thermal Imagery. In *Proceedings of SPIE*, volume 2759, pages 131–138.

[Boubchir, 2007] Boubchir, L. (2007). *Bayesian approaches for image denoising in oriented and non-oriented multiscale sparse transforms*. PhD thesis, University of Caen, France.

[Bregman, 1967] Bregman, L. (1967). The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex optimization. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217.

[Brox and Cremers, 2007] Brox, T. and Cremers, D. (2007). Iterated Nonlocal Means for Texture Restoration. In *Proc. Int. Conf. on Scale Space and Variational Methods in Computer Visions (SSVM'07)*, volume 4485, Ischia, Italy. Springer, LNCS.

[Buades. et al., 2005] Buades., A., Coll., B., and Morel, J. (2005). A non local algorithm for image denoising. *SIAM interdisciplinary journal: Multiscale Modeling and Simulation*, 2(2):60–65.

[Buades et al., 2005] Buades, A., Coll, B., and Morel, J. (2005). A review of image denoising algorithms, with a new one. *SIAM interdisciplinary journal: Multiscale Modeling and Simulation*, 4(2):290–530.

[Buades et al., 2008] Buades, A., Coll, B., and Morel, J.-M. (2008). Nonlocal Image and Movie Denoising. *Int J. Comput. Vis.*, 76:123–139.

[Burt and Adelson, 1983] Burt, P. J. and Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.*, COM-31(4):532–540.

[Burton and Moorhead, 1987] Burton, G. J. and Moorhead, I. R. (1987). Color and Spatial structures in natural scenes. *Applied Optics*, 26(1):157–170.

[Campbell et al., 1998] Campbell, N. A., Lopuhaä, H. P., and Rousseeuw, P. J. (1998). On the calculation of a robust s-estimator of a covariance matrix. *Stat Med*, 17(23):2685–2695.

[Candès, 1998] Candès, E. (1998). *Ridgelets: Theory and Applications*. PhD thesis, Departement of Statistics, Stanford University.

[Candès et al., 2006] Candès, E., Demanet, L., Donoho, D., and Ying, L. (2006). Fast Discrete Curvelet Transforms. *Multiscale modeling and simulation*, 5(3):861–899.

[Castella et al., 2009] Castella, C., Eckstein, M. P., Abbey, C. K., and Kinkel, K. O. (2009). Mass detection on mammograms: influence of signal shape uncertainty on human and model observers. *J. Opt. Soc. Am. A*, 26(2):425–436.

[Castleman et al., 1998] Castleman, K., Schulze, M., and Wu, Q. (1998). Simplified design of steerable pyramid filters. In *Proceedings of the 1998 IEEE Int. Symposium on Circuits and Systems (ISCAS)*, volume 5, pages 329–332.

[Chan et al., 2001] Chan, T. F., Osher, S., and Shen, J. (2001). The digital tv filter and nonlinear denoising. *IEEE Trans. Image Processing*, 10(2):231–241.

[Chang et al., 2000a] Chang, S., Yu, B., and Vetterli, M. (2000a). Spatially adaptive wavelet thresholding with context modeling for image denoising. *IEEE Trans. Image Process.*, 9(9):1522–1531.

[Chang et al., 1998] Chang, S. G., Yu, B., and Vetterli, M. (1998). Spatially adaptive wavelet thresholding with context modeling for image denoising. In *Proc. IEEE Internat. Conf. on Image Proc.*, Chicago, IL, USA.

[Chang et al., 2000b] Chang, S. G., Yu, B., and Vetterli, M. (2000b). Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. Image Processing*, 9(9):1532–1546.

[Chaux et al., 2006] Chaux, C., Duval, L., and Pesquet, J.-C. (2006). Image analysis using a dual-tree m-band wavelet transform. *IEEE Trans. Image Processing*, 15(8):2397–2412.

[Chellappa, 1985] Chellappa, R. (1985). *Two-dimensional discrete gaussian markov random field models for image processing.* North-Holland.

[Chen et al., 2001] Chen, M., Bowsher, J., Baydush, A., Gilland, K., DeLong, D., and Jaszczak (2001). Using the hotelling observer on multi-slice and multi-view simulated SPECT myocardial images. In *Conf. Rec. IEEE Nucl. Sci. Symp.*, pages 2258–2262.

[Chen et al., 1998] Chen, S. S., Donoho, D., and Saunders, M. (1998). Atomic Decomposition by Basis Pursuis. *SIAM J. Sci. Comput.*, 20(1):33–61.

[Chipman et al., 1997] Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. of the Amer. Statist. Assoc*, 92:1413–1421.

[Choi et al., 2000a] Choi, H., Romberg, J., Baraniuk, R., and Kingsbury, N. G. (2000a). Hidden Markov tree modeling of complex wavelet transforms. In *Proc. IEEE Conf. Acoustics, Speech and Sig. Processing (ICASSP)*, pages 136–136.

[Choi et al., 2000b] Choi, H., Romberg, J., Baraniuk, R., and Kingsbury, N. G. (2000b). Hidden Markov tree modeling of complex wavelet transforms. In *Proc. IEEE Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, pages 136–136.

[Clyde et al., 1998] Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85(2):391–401.

[Coifman and Wickerhauser, 1992] Coifman, R. and Wickerhauser, M. (1992). Entropy-based algorithms for best basis selection. *IEEE Trans. Information Theory*, 38(2):713–718.

[Coifman and Donoho, 1995] Coifman, R. R. and Donoho, D. L. (1995). Translation-invariant denoising. In Antoniadis, A. and Oppenheim, G., editors, *Wavelets and Statistics*, pages 125–150, New York. Springer Verlag.

[Coifman and Meyer, 1997] Coifman, R. R. and Meyer, F. G. (1997). Brushlets: a tool for directional image analysis and image compression. *Appl. Comp. Harmonic Anal.*, 5:147–187.

[Comaniciu and Meer, 2002] Comaniciu, D. and Meer, P. (2002). Mean Shift: A Robust Approach toward Feature Space Analysis. *IEEE Trans. Pattern Analysis Machine Intell.*, 24(5):603–619.

[Combettes and Pesquet, 2007] Combettes, P. and Pesquet, J.-C. (2007). A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):1–12.

[Conchello and McNally, 1996] Conchello, J.-A. and McNally, J. G. (1996). Fast regularization technique for expectation maximization algorithm for optical sectioning microscopy. In *Proc. SPIE Vol. 2655, Three-Dimensional Microscopy: Image Acquisition and Processing III*, pages 199–208. SPIE.

[Condat et al., 2008] Condat, L., Van De Ville, D., and Forster-Heinlein, B. (2008). Reversible, Fast, and High-Quality Grid Conversions. *IEEE Trans. Image Processing*, 17(5):679–693.

[Coupé et al., 2008] Coupé, P., Hellier, P., Kervrann, C., and Barillot, C. (2008). Bayesian non local means-based speckle filtering. In *Proc. of the 2008 IEEE Int. Symposium on Biomed. Imaging: From Nano to Macro*, pages 1291–1294.

[Crouse et al., 1998] Crouse, M. S., Nowak, R. D., and Baranuik, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Proc.*, 46(4):886–902.

[Şendur and Selesnick, 2002a] Şendur, L. and Selesnick, I. (2002a). Bivariate shrinkage with local variance estimation. *IEEE Signal Processing Letters*, 9(12):438–441.

[Şendur and Selesnick, 2002b] Şendur, L. and Selesnick, I. W. (2002b). Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Trans. On Signal Processing*, 50:2744–2756.

[da Cunha et al., 2006] da Cunha, A., Zhou, J., and Do, M. (2006). The nonsubsampled contourlet transform: theory, design, and applications. *IEEE Trans. Image Processing*, 15(10):3089–3101.

[Dabov et al., 2007] Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. (2007). Image denoising by sparse 3d transform-domain collaborative filtering. *IEEE Trans. Image Processing*, 16(8):2080–2095.

[Dan et al., 1996] Dan, Y., Atick, J. J., and Reid, R. C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J Neurosci*, 16(10):3351–3362.

[Daubechies, 1992] Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia.

[Daubechies et al., 2004] Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math*, 57(11):1413–1457.

[Daubechies et al., 2003] Daubechies, I., Han, B., Ron, A., and Shen, Z. (2003). Framelets: MRA based construction of wavelet frames. *Appl. Comp. Harm. Analysis*, 14:1–46.

[Daubechies and Sweldens, 1996] Daubechies, I. and Sweldens, W. (1996). Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, 4(3):245–267.

[Dauwe et al., 2008] Dauwe, A., Goossens, B., Luong, H., and Philips, W. (2008). A Fast Non-Local Image Denoising Algorithm. In *Proc. SPIE Electronic Imaging*, volume 6812, San José, USA.

[De Neve et al., 2009] De Neve, S., Goossens, B., Luong, H., and Philips, W. (2009). An Improved HDR Image Synthesis Algorithm. In *IEEE Int. Conf. Image Processing (ICIP2009)*, pages 1545–1548, Cairo, Egypt.

[Debevec and Malik, 1997] Debevec, P. E. and Malik, J. (1997). Recovering High Dynamic Range Radiance Maps from Photographs. In *SIGGRAPH*, pages 369–378.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 19(1):1–38.

[Dey et al., 2004] Dey, N., Blanc-Féraud, L., Zimmer, C., Roux, P., Kam, Z., Olivo-Marin, J., and Zerubia, J. (2004). 3D microscopy deconvolution using richardson-lucy algorithm with total variation regularization. `http://www-sop.inria.fr/ariana/en/publications.php?name=Dey`. Research Report 5272, INRIA.

[Ding and Venetsanopoulos, 1987] Ding, R. and Venetsanopoulos, A. N. (1987). Generalized homomorphic and adaptive order statistic filters for the removal of impulsive and signal-dependent noise. *IEEE Trans. Circuits Syst.*, 34(8):948–955.

[Do and Vetterli, 2001] Do, M. N. and Vetterli, M. (2001). Frame reconstruction of the Laplacian pyramid. In *IEEE Conf. on Acoustics, Speech and Signal Processing*, volume 6, pages 3641–3644.

[Do and Vetterli, 2003a] Do, M. N. and Vetterli, M. (2003a). *Beyond Wavelets, chapter Contourlets.* Academic Press.

[Do and Vetterli, 2003b] Do, M. N. and Vetterli, M. (2003b). The finite ridgelet transform for image representation. *IEEE Trans. Image Processing*, 12(1):16–28.

[Do and Vetterli, 2005] Do, M. N. and Vetterli, M. (2005). The contourlet transform: An efficient directional multiresolution image representation. *IEEE Trans. Image Process.*, 14(12):2091–2106.

[Donoho, 1995] Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*, 41(3):613–627.

[Donoho, 1999] Donoho, D. L. (1999). Wedgelets: Nearly minimax estimation of edges. *Annals of Statistics*, 27(3):859–897.

[Donoho and Flesia, 2001] Donoho, D. L. and Flesia, A. G. (2001). Can recent innovations in harmonic analysis 'explain' key findings in natural image statistics? *Network*, 12(3):371–393.

[Dubois, 2005] Dubois, E. (2005). Frequency-domain methods for demosaicking of bayer-sampled color images. *IEEE Signal Processing Letters*, 12(12):847–850.

[Dumitrescu, 2009] Dumitrescu, B. (2009). SDP Approximation of a Fractional Delay and the Design of Dual-Tree Complex Wavelet Transform. *IEEE Trans. Sig. Process.*, 56(9):4255.

[Easley et al., 2009] Easley, G., Labate, D., and Colonna, F. (2009). Shearlet Based Total Variation for Denoising. *IEEE Trans. Image Process.*, 18(2):260–268.

[Easley et al., 2008] Easley, G., Labate, D., and Lim, W. (2008). Sparse Directional Image Representations using the Discrete Shearlet Transform. *Applied and Computational Harmonic Analysis*, 25:25–46.

[Eckstein and Abbey, 2001] Eckstein, M. P. and Abbey, C. K. (2001). Model observers for signal-known-statistically tasks. In *Proc. SPIE Medical Imaging*, volume 4324.

[Edelstein et al., 1986] Edelstein, W. A., Glover, G. H., Hardy, C. J., and Redington, R. W. (1986). The intrinsic signal-to-noise ratio in NMR imaging. *Magn Reson Med*, 3(4):604–618.

[Elad and Aharon, 2006a] Elad, M. and Aharon, M. (2006a). Image denoising via learned dictionaries and sparse representation. In *IEEE Conf. on Comp. Vis. and Pat. Recog. (CVPR)*, volume 1, pages 895–900.

[Elad and Aharon, 2006b] Elad, M. and Aharon, M. (2006b). Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries. *IEEE Trans. Image Processing*, 15(12):3736–3745.

[Fadili and Boubchir, 2005] Fadili, J. M. and Boubchir, L. (2005). Analytical form for a Bayesian wavelet estimator of images using the Bessel K form densities. *IEEE Trans. on Image Process.*, 14(2):231–240.

[Fan and Xia, 2001a] Fan, G. and Xia, X. (2001a). Image denoising using local contextual hidden Markov model in the wavelet domain. *IEEE Signal Processing Letters*, 8(5):125–128.

[Fan and Xia, 2001b] Fan, G. and Xia, X.-G. (2001b). Improved Hidden Markov Models in the Wavelet-Domain. *IEEE Trans. Signal Process.*, 49(1):115–120.

[Faraji and MacLean, 2004] Faraji, H. and MacLean, J. (2004). Adaptive suppression of ccd signal-dependent noise in light space. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, volume 3, pages iii–217–20.

[Faulkner and Moores, 1984] Faulkner, K. and Moores, B. M. (1984). Analysis of x-ray computed tomography images using the noise power spectrum and autocorrelation function. *Phys. Med. Biol.*, 29(11):1343–1352.

[Field, 1987] Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394.

[Figueiredo and Nowak, 2001] Figueiredo, M. T. and Nowak, R. D. (2001). Wavelet-based image estimation: an empirical bayes approach using jeffrey's noninformative prior. *IEEE Trans. Image Process.*, 10(9):1322–1331.

[Figueiredo and Nowak, 2003] Figueiredo, M. T. and Nowak, R. D. (2003). An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.*, 12(8):906–916.

[Fischer et al., 2007] Fischer, S., Šroubek, F., Perrinet, L., Redondo, R., and Cristobal, G. (2007). Self-Invertible 2D Log-Gabor Wavelets. *International Journal of Computer Vision*, 75(2):231–246.

[Foi, 2008] Foi, A. (2008). Practical denoising of clipped or overexposed noisy images. In *Proc. 16th European Signal Process. Conf. (EUSIPCO)*, Lausanne, Switzerland.

[Foi et al., 2008] Foi, A., Trimeche, M., Katkovnik, V., and Egiazarian, K. (2008). Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Trans. image Process.*, 17(10):1737–1754.

[Freeman and Adelson, 1991] Freeman, W. and Adelson, E. (1991). Design and use of steerable filters. *IEEE Transactions on Pattern analysis and Machine Intelligence*, 13(9):891–906.

[Friedman, 1987] Friedman, J. H. (1987). Exporatory Projection Pursuit. *J. Amer. Statist. Assoc.*, 82:249–266.

[Gallas et al., 2007] Gallas, B., Pennello, G., and Myers, K. (2007). Multi-reader multicase variance analysis for binary data. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, 24(12):B70–B80.

[Gallas and Barrett, 2003] Gallas, B. D. and Barrett, H. H. (2003). Validating the use of channels to estimate the ideal linear observer. *J. Opt. Soc. Am. A*, 20(9):1725–1738.

[Gamerman, 1997] Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. and Machine Intel.*, 6:721–741.

[Gifford et al., 2005] Gifford, H. C., King, M. A., Pretorius, P. H., and Wells, R. G. (2005). A Comparison of Human and Model Observers in Multislice LROC Studies. *IEEE Trans. Medical Imaging*, 24(2):160–169.

[Gilland et al., 2004] Gilland, K. L., Tsui, B. M. W., Qi, Y., and Gullberg, G. T. (2004). Comparison of Rotationally Symmetric and Oriented Channels for the Hotelling Observer for Myocardial SPECT Images. In *Nuclear Science Symposium*, volume 6, pages 3620–3623.

[Goldstein and Osher, 2008] Goldstein, T. and Osher, S. (2008). The split Bregman method for L1 regularized problems. Technical Report 08-29, Computational and Applied Math, University of California.

[Gómez et al., 1998] Gómez, E., Gómez-Villegas, M. A., and Marín, J. M. (1998). A Multivariate Generalization of the Power Exponential Family of Distributions. *Communications in Statistics - Theory and Methods*, 27(3):589–600.

[Gómez et al., 2008] Gómez, E., Gómez-Villegas, M. A., and Marín, J. M. (2008). Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications. *Communications in Statistics - Theory and Methods*, 37(6):972–985.

[Goossens et al., 2009a] Goossens, B., Aelterman, J., Luong, H. Q., Pižurica, A., and Philips, W. (2009a). Efficient Design of a Low Redundant Discrete Shearlet Transform. In *2009 International Workshop on Local and Non-Local Approximation in Image Processing (LNLA2009)*, pages 112–124, Tuusula, Finland. IEEE.

[Goossens et al., 2010a] Goossens, B., Aelterman, J., Pižurica, A., and Philips, W. (2010a). A Recursive Scheme for Computing Autocovariance functions of complex wavelet subbands. *IEEE Trans. Signal Processing*. In press.

[Goossens et al., 2007a] Goossens, B., De Bock, J., Pižurica, A., and Philips, W. (2007a). Low-Dose CT Image Denoising By Locally Adaptive Wavelet Domain Estimation. In *Proc. IEEE Benelux Chapter on Engineering in Medicine and Biology (EMBS)*, Heeze, the Netherlands.

[Goossens et al., 2008a] Goossens, B., Luong, H., Pižurica, A., and Philips, W. (2008a). An improved Non-Local Means Algorithm for Image Denoising. In *2008 International Workshop on Local and Non-Local Approximation in Image Processing*, Ghent, Belgium. (invited paper).

[Goossens et al., 2008b] Goossens, B., Luong, H., Pižurica, A., and Philips, W. (2008b). Space-variant Characterization of Image Noise in Computed Tomography. In *Liége Imaging Days*, Liége, Belgium.

[Goossens et al., 2006] Goossens, B., Pižurica, A., and Philips, W. (2006). Wavelet domain image denoising for non-stationary and signal-dependent noise. In *IEEE International Conference on Image Processing (ICIP)*, pages 1425–1428, Atlanta, GA, USA.

[Goossens et al., 2007b] Goossens, B., Pižurica, A., and Philips, W. (2007b). Noise Removal from Images by Projecting onto Bases of Principle Components. In *Proc. Int. Conf. Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pages 190–199, Delft, The Netherlands.

[Goossens et al., 2007c] Goossens, B., Pižurica, A., and Philips, W. (2007c). Removal of Correlated Noise by Modeling Spatial Correlations and Interscale Dependencies in the Complex Wavelet Domain. In *IEEE Int. Conf. on Image Processing (ICIP)*, pages 317–320, San Antonio, Texas, USA.

[Goossens et al., 2008c] Goossens, B., Pižurica, A., and Philips, W. (2008c). EM-Based Estimation of Spatially Variant Correlated Image Noise. In *IEEE Int. Conf. on Image Processing (ICIP)*, pages 1744–1747, San Diego, CA, USA.

[Goossens et al., 2009b] Goossens, B., Pižurica, A., and Philips, W. (2009b). A Filter Design Technique for Improving the Directional Wavelet Transform. In *IEEE Int. Conf. Image Processing (ICIP2009)*, pages 3805–3808, Cairo, Egypt. IEEE. (to appear).

[Goossens et al., 2009c] Goossens, B., Pižurica, A., and Philips, W. (2009c). Image Denoising Using Mixtures of Projected Gaussian Scale Mixtures. *IEEE Trans. Image Processing*, 18(8):1689–1702.

[Goossens et al., 2009d] Goossens, B., Pižurica, A., and Philips, W. (2009d). Removal of correlated noise by modeling the signal of interest in the wavelet domain. *IEEE Trans. Image Processing*, 18(6):1153–1165.

[Goossens et al., 2010b] Goossens, B., Pižurica, A., and Philips, W. (2010b). Design of First-Scale Wavelet Filters with Improved Directional Sensitivity for the Dual-Tree Complex Wavelet Transform. In preparation.

[Goossens et al., 2010c] Goossens, B., Pižurica, A., and Philips, W. (2010c). An em algorithm for estimating the noise covariance matrix from a noisy wavelet subband under a gaussian scale mixture prior model. In preparation.

[Goossens et al., 2010d] Goossens, B., Pižurica, A., and Philips, W. (2010d). A Statistical Model for Non-Stationary Noise in Computed Tomography Imaging. Submitted to IEEE Trans. Medical Imaging, under revision.

[Goossens et al., 2010e] Goossens, B., Platiša, L., Vansteenkiste, E., and Philips, W. (2010e). Channelized Hotelling Observers for Detecting Signals with Random Parameters. In preparation.

[Goossens et al., 2010f] Goossens, B., Platiša, L., Vansteenkiste, E., and Philips, W. (2010f). The Use of Steerable Channels for Detecting Asymmetrical Signals with Random Orientations. In *SPIE Medical Imaging 2010*, San Diego, CA, USA.

[Gopinath, 2003] Gopinath, R. A. (2003). The Phaselet Transform - An Integral Redundancy Nearly Shift-Invariant Wavelet Transform. *IEEE Trans. Signal Processing*, 51(7):1792–1805.

[Guerrero-Colón, 2008] Guerrero-Colón, J. A. (2008). *Bayesian methods for the restoration of digital camera images in overcomplete pyramids*. PhD thesis, Universidad de Granada, Escuela Técnica Superior de Ingenierías Informática.

[Guerrero-Colón et al., 2007] Guerrero-Colón, J. A., Mancera, L., and Portilla, J. (2007). Image restoration using space-variant gaussian scale mixtures in overcomplete pyramids. *IEEE Trans. Image Proc.*, 6701:27–41. (accepted on 26th Sept 2007).

[Guerrero-Colón et al., 2008a] Guerrero-Colón, J. A., Mancera, L., and Portilla, J. (2008a). Image restoration using space-variant gaussian scale mixtures in overcomplete pyramids. *IEEE Trans. Image Processing*, 17(1):27–41. (accepted on 26th Sept 2007).

[Guerrero-Colón et al., 2008b] Guerrero-Colón, J. A., Simoncelli, E. P., and J., P. (2008b). Image Denoising using Mixtures of Gaussian Scale mixtures. In *IEEE Int. Conf on Image Processing (ICIP2008)*, pages 565–568, San Diego, CA, USA.

[Guo and Labate, 2007] Guo, K. and Labate, D. (2007). Optimally Sparse Multidimensional Representation using Shearlets. *SIAM J Math. Anal.*, 39:298–318.

[Guo et al., 2009] Guo, K., Labate, D., and Lim, W. (2009). Edge Analysis and Identification using the Continuous Shearlet Transform. *Applied and Computational Harmonic Analysis*, 27(1):24–46.

[Hadamard, 1923] Hadamard, J. (1923). *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. H. Milford, Oxford University Press.

[Hammond and Simoncelli, 2008] Hammond, D. and Simoncelli, E. (2008). Image Modeling and Denoising with Orientation-Adapted Gaussian Scale Mixtures. *IEEE Trans. Image Process.*, 17(11):2089–2101.

[Hammond et al., 2009] Hammond, D., Vandergheynst, P., and Gribonval, R. (2009). Wavelets on Graphs via Spectral Graph Theory. Submitted.

[Hanson, 1981] Hanson, K. (1981). Noise and contrast discrimination in computed tomography. *Radiology of the Skull and Brain*, 5: Technical Aspects of Computed Tomography:3941–3954.

[Hastie and Stuetzle, 1989] Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84:502–516.

[Hein and Zakhor, 1995] Hein, S. and Zakhor, A. (1995). Halftone to Continuous-Tone Conversion of Error-Diffusion Coded Images. *IEEE Trans. Image Processing*, 4:204–216.

[Hinton, 1999] Hinton, G. E. (1999). Products of experts. In *Int. Conf. on Art. Neur. Netw. (ICANN)*, volume 1, pages 1–6.

[Hirakawa, 2007] Hirakawa, K. (2007). Signal-dependent noise characterization in Haar filterbank representation. In *Proc. SPIE Optics & Photonics*, pages 67011I.1–67011I.12.

[Hirakawa, 2008a] Hirakawa, K. (2008a). Cross-talk explained. In *Proc. 15th IEEE International Conference on Image Processing ICIP 2008*, pages 677–680.

[Hirakawa, 2008b] Hirakawa, K. (2008b). *Single-Sensor Imaging: Methods and Applications for Digital Cameras*, chapter Color Filter Array Image Analysis for Joint Denoising and Demosaicking. CRC Press.

[Hirakawa et al., 2007] Hirakawa, K., Meng, X.-L., and J. Wolfe, P. (2007). A framework for wavelet-based analysis and processing of color filter array images with applications to denoising and demosaicing. *Proc. IEEE Int. Conf. on Acoust. Speech Signal Process. (ICASSP 2007)*, pages 597–600.

[Hirakawa and Parks, 2005a] Hirakawa, K. and Parks, T. W. (2005a). Adaptive homogeneity-directed demosaicing algorithm. 14(3):360–369.

[Hirakawa and Parks, 2005b] Hirakawa, K. and Parks, T. W. (2005b). Image denoising for signal-dependent noise. *IEEE Proc. Acoust., Speech, Signal Processing*, 2:29–32.

[Hirakawa and Parks, 2006] Hirakawa, K. and Parks, T. W. (2006). Joint demosaicing and denoising. *IEEE Trans. Image Process.*, 15(8):2146–2157.

[Hirakawa and Wolfe, 2008] Hirakawa, K. and Wolfe, P. J. (2008). Spatio-spectral color filter array design for optimal image recovery. 17(10):1876–1890.

[Hirakawa and Wolfe, 2009] Hirakawa, K. and Wolfe, P. J. (2009). Skellamshrink: Poisson intensity estimation for vector-valued data. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2009*, pages 3441–3444.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of Complex of Statistical variables into Principal Components. *Journal of Educational Psychology*, 24:417–441.

[Hsieh, 1997] Hsieh, J. (1997). Nonstationary noise characteristics of the helical scan and its impact on image quality and artifacts. *Med. Phys.*, 24:1375–1384.

[Hsieh, 1998] Hsieh, J. (1998). Adaptive streak artifact reduction in computed tomography resulting from excessive x-ray photon noise. *Med. Phys.*, 25(11):2139–2147.

[Hsieh, 2003] Hsieh, J. (2003). *Computed Tomography: Principles, Design, Artifacts and Recent Advances.* SPIE Press.

[Hubel and Wiesel, 1959] Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.*, 148:574–591.

[Hyvärinen, 1999] Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3):626–634.

[Jacquin, 1992] Jacquin, A. (1992). Image coding based on a fractal theory of iterated contractive image transformations. *IEEE Trans. Image Process.*, 1(1):18–30.

[Jansen and Bultheel, 1999] Jansen, M. and Bultheel, A. (1999). Geometrical priors for noisefree wavelet coefficients in image denoising. In Müller, P. and Vidakovic, B., editors, *Bayesian inference in wavelet based models*, volume 141 of *Lecture Notes in Statistics*, pages 223–242. Springer Verlag.

[Jansen and Bultheel, 2001] Jansen, M. and Bultheel, A. (2001). Empirical bayes approach to improve wavelet thresholding for image noise reduction. *J. of the Amer. Statist. Assoc. (JASA)*, 96(454):629–639.

[Johnstone and Silverman, 1997] Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society B*, 59(2):319–351.

[Jolliffe, 1986] Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag.

[Kahan, 1958] Kahan, W. (1958). *Gauss-Seidel Methods of Solving Large Systems of Linear Equations*. PhD thesis, University of Toronto, Canada.

[Kak and Slaney, 2001] Kak, A. C. and Slaney, M. (2001). *Principles of Computerized Tomographic Imaging*. Society of Industrial and Applied Mathematics.

[Kakarala and Baharav, 2002] Kakarala, R. and Baharav, Z. (2002). Adaptive demosaicing with the principal vector method. *IEEE Trans. Consumer Electronics*, 48(4):932–937.

[Kang and Katsaggelos, 1995] Kang, M. G. and Katsaggelos, A. K. (1995). General Choise of the Regularization Functional in Regularized Image Restoration. *IEEE Trans. Image Processing*, 4(5):594–602.

[Kervrann and Boulanger, 2006] Kervrann, C. and Boulanger, J. (2006). Optimal spatial adaptation for patch-based image denoising. *IEEE Trans. Image Processing*, 15(10):2866–2878.

[Kervrann et al., 2007] Kervrann, C., Boulanger, J., and Coupé, P. (2007). Bayesian Non-Local Means Filter, Image Redundancy and Adaptive Dictionaries for Noise Removal. In *Proc. Int. Conf. on Scale Space and Variational Methods in Computer Visions (SSVM'07)*, pages 520–532, Ischia, Italy.

[Kijewski and Judy, 1987] Kijewski, M. and Judy, P. (1987). The noise power spectrum of CT images. *Phys. Med. Biol.*, 32(5):565–575.

[Kimmel, 1999] Kimmel, R. (1999). Demosaicing: image reconstruction from color ccd samples. *IEEE Trans. Image Processing*, 8(9):1221–1228.

[Kingsbury, 2006] Kingsbury, N. (2006). Rotation-invariant Local Feature Matching with Complex Wavelets. In *Proc. European Signal Processing Conference (EUSIPCO)*.

[Kingsbury, 2001] Kingsbury, N. G. (2001). Complex wavelets for shift invariant analysis and filtering of signals. *Journal of Applied and Computational Harmonic Analysis*, 10(3):234–253.

[Kingsbury, 2003] Kingsbury, N. G. (2003). Design of q-shift complex wavelets for image processing using frequency domain energy minimalisation. In *Proc. IEEE Conf. on Image Process.*, Barcelona.

[Kirsch, 1996] Kirsch, A. (1996). *An Introduction To The Mathematical Theory of Inverse Problems*. Springer.

[Kivinen et al., 2007] Kivinen, J. J., Sudderth, E. B., and Jordan, M. I. (2007). Image denoising with nonparametric Hidden Markov Trees. In *IEEE Int. Conf. on Image Processing (ICIP)*, San Antonio, Texas, USA.

[Kotz and Kozubowski, 2001] Kotz, S. and Kozubowski, T.and Podgorski, K. (2001). *The Laplace Distributions And Generalizations: A Revisit with Applications to Communications, Economics, Engineering, Finance*. Birkhäuser, Boston.

[Kotz et al., 2000] Kotz, S., Kozubowski, T. J., and Podgorski, K. (2000). An asymmetric multivariate laplace distribution. *Computational Statistics*, 4:531–540.

[Krupinski and Berbaum, 2009] Krupinski, E. and Berbaum, K. (2009). Does Reader Visual Fatigue Impact Performance? In *Medical Image Perception Society (MIPS) XIII Conference*.

[Kwon et al., 2003] Kwon, O., Sohn, K., and Lee, C. (2003). Deinterlacing using Directional Interpolation and Motion Compensation. *IEEE Trans. Consumer Electronics*, 49(1):198–203.

[Laakso et al., 1996] Laakso, T. I., Välimäki, V., Karjalainen, M., and Laine, U. K. (1996). Splitting the unit delay – tools for fractional delay filter design. *IEEE Sig. Process. Mag.*, 13:30–60.

[LaCroix et al., 1999] LaCroix, K. J., Tsui, B. M. W., and Frey, E. C. (1999). Rotationally Symmetric vs. Oriented Frequency Channels for the Hotelling Observer: A Comparison with Human Observers. In *Nuclear Science Symposium*, volume 3, pages 1402–1406.

[Lee and Mumford, 1999] Lee, A. and Mumford, D. (1999). An occlusion model generating scale-invariant images. In *IEEE Workshop on Statistical and Computational Theories of Vision*, Fort Collins, CO, USA.

[Lee, 1996] Lee, T. (1996). Image Representation Using 2D Gabor Wavelets. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(10):1.

[Lee et al., 2006] Lee, W., Lee, S., and Kim, J. (2006). Cost-efffective color filter array demosaicing using spatial correlation. *IEEE Trans. Consumer Electronics*, 52(2):547–554.

[LePennec and Mallat, 2005] LePennec, E. and Mallat, S. (2005). Sparse geometric image representation with bandelets. *IEEE Trans. Image Processing*, 14(4):423–438.

[Leporini et al., 1999] Leporini, D., Pasquet, J.-C., and Krim, H. (1999). Best basis representation with prior statistical models. In Müller, P. and Vidakovic, B., editors, *Lecture Notes in Statistics*, pages 155–172, Springer Verlag.

[Li, 1995] Li, S. Z. (1995). *Markov Random Field Modeling in Computer Vision*. Springer-Verlag.

[Li, 2005] Li, X. (2005). Demosaicing by successive approximation. *IEEE Trans. Image Processing*, 14(3):370–379.

[Lim, 2006] Lim, S. (2006). Characterization of noise in digital photographs for image processing. In *Proc. SPIE Digital Photography II*, volume 6069, page 60690O, San José, CA, USA.

[Lina and Mayrand, 1995] Lina, J.-M. and Mayrand, M. (1995). Complex Daubechies wavelets. *Appl. Comp. Harm. Analysis*, 2(3):219–229.

[Liu et al., 2006a] Liu, C., Freeman, W. T., Szeliski, R., and Kang, S. B. (2006a). Noise estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 901–908.

[Liu et al., 2006b] Liu, C., Freeman, W. T., Szeliski, R., and Kang, S. B. (2006b). Noise Estimation from a Single Image. In *Proc. of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pages 901–908, Washington, DC, USA. IEEE Computer Society.

[Lu et al., 2001] Lu, H., Hsiao, I.-T., Li, X., and Liang, Z. (2001). Noise properties of low-dose CT projections and noise treatment by scale transformations. In *Proc. IEEE Nuclear Science Symposium Conference*, volume 3, pages 1662–1666.

[Lu and Do, 2007] Lu, Y. M. and Do, M. N. (2007). Multidimensional Directional Filter Banks and Surfacelets. *IEEE Trans. Image Processing*, 16(4):918–931.

[Lu and Dosher, 1999] Lu, Z. L. and Dosher, B. A. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *J Opt Soc Am A Opt Image Sci Vis*, 16(3):764–778.

[Lucy, 1974] Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79(6):745–754.

[Lukac and Plataniotis, 2005a] Lukac, R. and Plataniotis, K. N. (2005a). Color filter arrays: design and performance analysis. 51(4):1260–1267.

[Lukac and Plataniotis, 2005b] Lukac, R. and Plataniotis, K. N. (2005b). Universal demosaicking for imaging pipelines with an RGB color filter array. *Pattern Recognition*, 38:2208–2212.

[Luong, 2009] Luong, H. (2009). *Advanced Image and Video Resolution Enhancement Techniques*. PhD thesis, Ghent University, Department of Telecommunications and Information Processing (TELIN-IPI-IBBT).

[Luong et al., 2006] Luong, H., Ledda, A., and Philips, W. (2006). Non-local image interpolation. In *Proc. of IEEE International Conference on Image Processing (ICIP)*, pages 693–696.

[Lyu and Simoncelli, 2008] Lyu, S. and Simoncelli, E. (2008). Modeling Multiscale Subbands of Photographic Images with Fields of Gaussian Scale Mixtures. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(4):693–706.

[Mahmoudi and Sapiro, 2005] Mahmoudi, M. and Sapiro, G. (2005). Fast image and video denoising via nonlocal means of similar neighborhoods. *IEEE Signal Processing Letters*, 12(12):839–842.

[Mäkitalo and Foi, 2009] Mäkitalo, M. and Foi, A. (2009). On the inversion of the Anscombe transformation in low-count poisson image denoising. In *Proc. Int. Workshop on Local and Non-Local Approx. in Image Process. (LNLA2009)*, pages 26–32, Tuusula, Finland.

[Maleki et al., 1992] Maleki, M. H., Devaney, J., and Schatzberg, A. (1992). Tomographic reconstruction from optical scattered intensities. *JOSA A.*, 9(8):1536–1363.

[Malfait and Roose, 1997] Malfait, M. and Roose, D. (1997). Wavelet-based image denoising using a Markov Random Field a priori model. *IEEE Trans. Image Process.*, 6(4):549–565.

[Mallat, 1989a] Mallat, S. (1989a). Multifrequency channel decomposition of images and wavelet models. *IEEE Trans. Acoust., Speech, Signal Proc.*, 37(12):2091–2110.

[Mallat, 1989b] Mallat, S. (1989b). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11(7):674–693.

[Mallat, 1996] Mallat, S. (1996). Wavelets for Vision. *Proc. IEEE*, 84(4):604–614.

[Mallat, 1998] Mallat, S. (1998). A wavelet tour of signal processing. *Ann. Stat.*, 26(1):1–47.

[Mallat, 1999] Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Academic Press.

[Mallat, 2009] Mallat, S. (2009). Geometrical grouplets. *Applied and Computational Harmonic Analysis*, 26(2):161–180.

[Mallat and Gabriel, 2007] Mallat, S. and Gabriel, P. (2007). A review of Bandlet methods for geometrical image representation. *Numerical Algorithms*, 44(3):205–234.

[Mallat and Zhang, 1993] Mallat, S. and Zhang, Z. (1993). Matching Pursuit in a time-frequency dictionary. *IEEE Trans. Signal Process.*, 41:3397–3415.

[Manjeshwar and Wilson, 2001] Manjeshwar, R. M. and Wilson, D. L. (2001). Effect of inherent location uncertainty on detection of stationary targets in noisy image sequences. *J. Opt. Soc. Am. A*, 18(1):78–85.

[Mann and Whitney, 1947] Mann, H. and Whitney, D. (1947). On a test whether one or two random variables is stochastically larger than the orther. *Ann. Math. Statisics*, 18:50–60.

[Mann, 2000] Mann, S. (2000). Comparametric equations with practical applications in quantigraphic image processing. *IEEE Trans. Image Process.*, 9(8):1389–1406.

[Menon et al., 2007] Menon, D., Andriani, S., and Calvagno, G. (2007). Demosaicing with directional filtering and a posteriori decision. *IEEE Trans. Image Processing*, 16(1):132–141.

[Mihçak, 1999] Mihçak, M. K. (1999). Low-complexity Image Denoising based on Statistical Modeling of Wavelet Coefficients. *IEEE Signal Processing Letters*, 6(12):300–303.

[Moulin and Liu, 1999] Moulin, P. and Liu, J. (1999). Analysis of multiresolution image denoising schemes using generalized-gaussian and complexity priors. *IEEE Trans. Info. Theory, Special Issue on Multiscale Analysis*, 3(3):909–919.

[Moussouris, 1974] Moussouris, J. (1974). Gibbs and Markov random systems with constraints. *J. Stat. Phys.*, 10(1):11–33.

[Mumford and Gidas, 2001] Mumford, D. and Gidas, B. (2001). Stochastic Models for Generic Images. *Quaterly of Applied Mathematics*, 59(1):85–111.

[Muresan and Parks, 2005] Muresan, D. D. and Parks, T. W. (2005). Demosaicing using optimal recovery. 14(2):267–278.

[Myers and Barrett, 1987] Myers, K. J. and Barrett, H. H. (1987). Addition of a channel mechanism to the ideal-observer model. *J. Opt. Soc. Am. A*, 4(12):2447–2457.

[Nabney, 2001] Nabney, I. T. (2001). *NETLAB: Algorithms for Pattern Recognition, Advances in Pattern Recognition.* Springer, Berlin.

[Nakamura, 2005] Nakamura, J. (2005). *Image Sensors and Signal Processing for Digital Still Cameras*. CRC Press, Inc., Boca Raton, FL, USA.

[Neal and Hinton, 1998] Neal, R. and Hinton, G. (1998). *A View of the EM Algorithm that justifies Incremental, Sparse, and other variants*. Dordrecht: Kluwer Academic Publishers.

[Nikias and Shao, 1995] Nikias, C. L. and Shao, M. (1995). *Signal Processing with Alpha-Stable Distributions and Applications*. Wiley-Interscience.

[Nowak, 1999a] Nowak, R. (1999a). Multiscale hidden markov models for Bayesian image analysis. In Müller, P. and Vidakovic, B., editors, *Bayesian inference in wavelet based models*, New York: Springer Verlag.

[Nowak, 1999b] Nowak, R. (1999b). Wavelet-based rician noise removal for magnetic resonance imaging. *IEEE Trans Image Process*, 8(10):1408–1419.

[O'Connor and Fessler, 2007] O'Connor, Y. Z. and Fessler, J. A. (2007). Fast Predictions of Variance Images for Fan-Beam Transmission Tomography With Quadratic Regularization. *IEEE Trans. Medical Imaging*, 26(3):335–346.

[Olmo et al., 2000] Olmo, G., Magli, E., and Lo Presti, L. (2000). Joint statistical signal detection and estimation. Part I: theoretical aspects of the problem. *Signal Processing*, 80(1):57–73.

[Olshausen and Field, 1996] Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

[Olshausen and Field, 1997] Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325.

[Olshausen and Field, 2005] Olshausen, B. A. and Field, D. J. (2005). How close are we to understanding v1? *Neural Computation*, 17(8):1665–1699.

[Osher et al., 2008] Osher, S., Mao, O., Dong, B., and Yin, W. (2008). Fast linearized bregman iterations for compressed sensing and sparse denoising. CAM Report 08-37, UCLA.

[Pan and Yu, 2003] Pan, X. and Yu, L. (2003). Image reconstruction with shift-variant filtration and its implication for noise and resolution properties in fan-beam computed tomography. *Med. Phys.*, 30(4):590–600.

[Park et al., 2007] Park, S., Barrett, H. H., Clarkson, E., Kupinski, M. A., and Myers, K. J. (2007). Channelized-ideal observer using Laguerre-Gauss channels in detection a Gaussian signal. *J. Opt. Soc. Am. A*, 24(12):B136–B149.

[Park et al., 2005] Park, S., Clarkson, E., Kupinski, M. A., and Barrett, H. H. (2005). Efficiency of the human observer detecting random signals in random backgrounds. *J. Opt. Soc. Am. A*, 22(1):3–16.

[Park et al., 2003] Park, S., Kupinski, M. A., Clarkson, E., and Barrett, H. H. (2003). Ideal-observer performance under signal and background uncertainty. In Taylor, C. J. and Noble, J. A., editors, *Information Processing in Medical Imaging*, volume 2732 in Lecture Notes in Computer Science, pages 342–353. Springer-Verlag, New York.

[Park et al., 2009] Park, S., Witten, J. M., and Myers, K. J. (2009). Singular Vectors of a Linear Imaging System as Efficient Channels for the Bayesian Ideal Observer. *IEEE Trans. Medical Imaging*, 28(5):657–668.

[Perona and Malik, 1990] Perona, P. and Malik, J. (1990). Scale-Space and Edge Detection Using Anisotropic Diffusion. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(7):629–639.

[Philips, 1996] Philips, W. (1996). Warped polynomials and their applications in signal and image processing. Technical Report WP 96-02, ELIS, RUG, Gent.

[Pižurica, 2002] Pižurica, A. (2001-2002). *Image Denoising Using Wavelets and Spatial Context Modeling*. PhD thesis, University of Ghent.

[Pižurica and Philips, 2003] Pižurica, A. and Philips, W. (2003). Multiscale statistical image models and Bayesian methods. In *SPIE Conference Wavelet Applications in Industrial Processing*, Providence, Rhode Island, USA.

[Pižurica and Philips, 2006] Pižurica, A. and Philips, W. (2006). Estimating the probability of the presence of a signal of interest in multiresolution single- and multiband image denoising. *IEEE Trans. Image Process.*, 15(3):654–665.

[Pižurica and Philips, 2007] Pižurica, A. and Philips, W. (2007). Analysis of least squares estimators under Bernoulli-Laplacian priors. In *Twenty eighth Symposium on Information Theory in the Benelux*.

[Pižurica et al., 2002] Pižurica, A., Philips, W., Lemahieu, I., and Acheroy, M. (2002). A joint inter- and intrascale statistical model for wavelet based Bayesian image denoising. *IEEE Trans. Image Process.*, 11(5):545–557.

[Pižurica et al., 2003] Pižurica, A., Philips, W., Lemahieu, I., and Acheroy, M. (2003). A versatile wavelet domain noise filtration technique for medical imaging. *IEEE Trans. Medical Imaging*, 22(3):323–331.

[Pižurica et al., 2006] Pižurica, A., Vanhamel, I., Sahli, H., Philips, W., and Katartzis, A. (2006). A Bayesian formulation of edge-stopping functions in nonlinear diffusion. *IEEE Signal Process. Letters*, 8(13):501–504.

[Platiša et al., 2009a] Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2009a). Channelized Hotelling Observers for Detection Tasks in Multi-Slice Images. In *Medical Image Perception XIII (MIPS2009)*, Santa Barbara, CA, USA.

[Platiša et al., 2009b] Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2009b). Channelized Hotelling Observers for the Detection of 2D Signals in 3D Simulated Images. In *IEEE Int. Conf. Image Processing (ICIP2009)*, pages 1781–1784, Cairo, Egypt.

[Platiša et al., 2010a] Platiša, L., Goossens, B., Vansteenkiste, E., Badano, A., and Philips, W. (2010a). Using channelized Hotelling Observers to quantify the temporal effect of medical liquid crystal displays on the detection performance. In *Proc of SPIE Medical Imaging 2010*, San Diego, CA, USA.

[Platiša et al., 2010b] Platiša, L., Goossens, B., Vansteenkiste, E., Park, S., Gallas, B. D., Badano, A., and Philips, W. (2010b). Channelized Hotelling Observers for the Assessment of Volumetric Imaging data sets. submitted.

[Platiša et al., 2009c] Platiša, L., Vansteenkiste, E., Goossens, B., Marchessoux, C., Kimpe, T., and Philips, W. (2009c). Optimization of Medical Imaging Display Systems: using the - experimental study. In *SPIE Medical Imaging*, Orlando, Florida, USA. Submitted.

[Portilla, 2004] Portilla, J. (2004). Full blind denoising through noise covariance estimation using Gaussian Scale Mixtures in the wavelet domain. *IEEE Int. Conf. on Image Process. (ICIP)*, 2:1217–1220.

[Portilla, 2005] Portilla, J. (2005). Image restoration using Gaussian Scale Mixtures in Overcomplete Oriented Pyramids (a review). In *Proc. of the SPIE's 50th Annual Meeting: Wavelets XI*, volume 5914, pages 468–482, San Diego, CA.

[Portilla and Guerrero-Colón, 2007] Portilla, J. and Guerrero-Colón, J. (2007). Image restoration using adaptive gaussian scale mixtures in overcomplete pyramids. In Van De Ville, D., Goyal, V., and Papadakis, M., editors, *Proceedings of SPIE - Wavelets XII*, volume 6701.

[Portilla and Simoncelli, 2001] Portilla, J. and Simoncelli, E. (2001). Adaptive Wiener Denoising using a Gaussian Scale Mixture Model in the Wavelet Domain. *IEEE Int. Conf. on Image Process. (ICIP)*, 2:37–40.

[Portilla et al., 2003] Portilla, J., Strela, V., Wainwright, M., and Simoncelli, E. (2003). Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on image processing*, 12(11):1338–1351.

[Powell, 1981] Powell, M. (1981). *Approximation Theory and Methods*. Cambridge University Press.

[Rabbani et al., 2006] Rabbani, H., Vafadust, M., Gazor, S., and Selesnick, I. W. (2006). Image Denoising Employing a Bivariate Cauchy Distribution with Local Variance in Complex Wavelet Domain. In *12th Digital Signal Processing Workshop - 4th Signal Processing Education Workshop*, pages 203–208.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Raphan and Simoncelli, 2008] Raphan, M. and Simoncelli, E. (2008). Optimal denoising in redundant representations. *IEEE Trans. Image Process.*, 17(8):1342–1352.

[Richardson, 1972] Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59.

[Riederer et al., 1978] Riederer, S., N.J., P., and Chesler, D. (1978). The noise power spectrum in computed x-ray tomography. *Phys. Med. Biol.*, 23(3):446–454.

[Rolland and Barrett, 1992] Rolland, J. P. and Barrett, H. H. (1992). Effect of random background inhomogeneity on observer detection performance. *J. Opt. Soc. Am. A*, 9(5):649–658.

[Romberg et al., 2001a] Romberg, J., Choi, H., Baraniuk, R., and Kingsbury, N. G. (2001a). Hidden Markov tree modeling of complex wavelet transforms. In *Proc. IEEE Conf. on Image Process.*

[Romberg et al., 2001b] Romberg, J., Choi, H., and Baraniuk, R. G. (2001b). Bayesian tree structured image modeling using wavelet-domain Hidden Markov Models. *IEEE Trans. Image Process.*, 10(7):1056–1068. Enter text here.

[Romberg et al., 2002] Romberg, J., Wakin, M., and Baraniuk, R. (2002). Multiscale wedgelet image analysis: fast decompositions and modeling. In *Proc. International Conference on Image Processing 2002*, volume 3, pages 585–588.

[Rooms, 2005] Rooms, F. (2005). *Nonlinear Methods in Image Restoration Applied to Confocal Microscopy*. PhD thesis, Universiteit Gent (UGent).

[Rooms et al., 2010] Rooms, F., Goossens, B., Pižurica, A., and Philips, W. (2010). *Optical and Digital Image Processing*, chapter Image restoration and application in biomedical processing. Wiley-VCH Verlag GmbH & Co. KGaA.

[Roth and Black, 2009] Roth, S. and Black, M. J. (2009). Fields of Experts. *Int J. Comput. Vis.*, 82(2):205–229.

[Roweis et al., 2002a] Roweis, S. T., Saul, L. K., and Hinton, G. E. (2002a). *Global Coordination of Local Linear Models*, volume 14. MIT Press, Cambridge, MA, USA.

[Roweis et al., 2002b] Roweis, S. T., Saul, L. K., and Hinton, G. E. (2002b). Global Coordination of Local Linear Models. *Advances in Neural Information Processing Systems*, 14:889–896.

[Ruderman, 1997] Ruderman, D. L. (1997). Origins of Scaling in Natural Images. *Vision Research*, 37(23):3385–3398.

[Rudin and Osher, 1994] Rudin, L. and Osher, S. (1994). Total variation based image restoration with free local constraints. In *Proc. of IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 31–35.

[Rudin et al., 1992] Rudin, L. I., Osher, S., and Fatami, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268.

[Sallee and Olshausen, 2003] Sallee, P. and Olshausen, B. A. (2003). Learning sparse multiscale image representations. *Adv. in Neur. Inf. Proc. Sys (NIPS)*, 15:1327–1334.

[Sarpeshkar et al., 1993] Sarpeshkar, R., Delbruck, T., and Mead, C. (1993). White noise in mos transistors and resistors. *IEEE Circuits and Devices Magazine*, 9(6):23–29.

[Scheunders, 2004] Scheunders, P. (2004). Wavelet thresholding of multivalued images. *IEEE Trans. Image Proc.*, 13(4):"475–483".

[Selesnick, 2008] Selesnick, I. (2008). The Estimation of Laplace Random Vectors in Additive White Gaussian Noise. *IEEE Trans. Signal Process.*, 56(8):3482–3496.

[Selesnick, 2001] Selesnick, I. W. (2001). Hilbert Transform Pairs of Wavelet Bases. *IEEE Signal Processing Letters*, 8(6):170–173.

[Selesnick, 2002] Selesnick, I. W. (2002). The design of approximate Hilbert transform pairs of wavelet bases. *IEEE Trans. on Signal Processing*, 50(5):1144–1152.

[Selesnick, 2006] Selesnick, I. W. (2006). Laplace random vectors, Gaussian noise, and the generalized incomplete Gamma function. In *Proc. IEEE Int. Conf. on Image Process.*, pages 2097–2100.

[Selesnick et al., 2005a] Selesnick, I. W., Baraniuk, R. G., and Kingsbury, N. G. (2005a). The Dual-Tree Complex Wavelet Transform. *IEEE Signal Processing Magazine*, 22(6):123–151.

[Selesnick et al., 2005b] Selesnick, I. W., Baraniuk, R. G., and Kingsbury, N. G. (2005b). The Dual-Tree Complex Wavelet Transform. *IEEE Signal Processing Magazine*, 22(6):123–151.

[Seppä, 2007] Seppä, M. (2007). High-quality two-stage resampling for 3-D volumes in medical imaging. *Medical Image Analysis*, 11:346–360.

[Shapiro, 1993] Shapiro, J. M. (1993). Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Proc.*, 41(12):3445–3462.

[Shepp and Vardi, 1982] Shepp, L. A. and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Medical Imaging*, 1(2):113–122.

[Shi and Selesnick, 2006] Shi, F. and Selesnick, I. W. (2006). Multivariate Quasi-Laplacian Mixture Models for Wavelet-based Image Denoising. In *Proc. Int. Conf. on Image Processing (ICIP)*, pages 2097–2100.

[Shidahara et al., 2006] Shidahara, M., Inoue, K., Maruyama, M., Watabe, H., Taki, Onishi, Y., Ito, H., Arai, H., and Fukuda, H. (2006). Predicting human performance by channelized Hotelling observer statistically processed brain perfusion spect. *Ann. Nucl. Med.*, 20(9):605–613.

[Shiraishi et al., 2000] Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K., Matsui, M., Fujita, H., Kodera, Y., and Doi, K. (2000). Development of a Digital Image Database for Chest Radiographs with and without a Lung Nodule: Receiver Operating Characteristic Analysis of Radiologists. *Detection of Pulmonary Nodules, AJR*, 174:71–74.

[Shukla et al., 2005] Shukla, R., Dragotti, P., Do, M., and Vetterli, M. (2005). Rate-distortion optimized tree-structured compression algorithms for piecewise polynomial images. *IEEE Trans. Image Processing*, 14(3):343–359.

[Simoncelli, 1996] Simoncelli, E. (1996). A Rotation-Invariant Pattern Signature. In *IEEE Int. Conf. Image Processing*.

[Simoncelli et al., 1992] Simoncelli, E., Freeman, W. T., Adelson, E. H., and Heeger, D. J. (1992). Shiftable Multi-scale Transforms. *IEEE Trans. Information Theory*, 38(2):587–607.

[Simoncelli and Olshausen, 2001] Simoncelli, E. and Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, 24:1193–1216.

[Simoncelli, 1999] Simoncelli, E. P. (1999). Modeling the joint statistics of image in the wavelet domain. In *Proc. SPIE Conf. on Wavelet Applications in Signal and Image Processing VII*, volume 3813, pages 188–195, Denver, CO.

[Simoncelli and Adelson, 1996] Simoncelli, E. P. and Adelson, E. H. (1996). Noise removal via Bayesian wavelet coring. In *Proc. IEEE Internat. Conf. Image Proc. ICIP*, Lausanne, Switzerland.

[Simoncelli and Freeman, 1995] Simoncelli, E. P. and Freeman, W. T. (1995). The Steerable Pyramid: A flexible architecture for Multi-scale Derivative Computation. In *Proc IEEE Int.Conf. Image Processing*, Washington, DC.

[Sled et al., 1998] Sled, J., Zijdenbos, A., and Evans, A. (1998). A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Trans. Medical Imaging*, 17(1):87–97.

[Srivastava et al., 2003] Srivastava, A., Lee, A. B., Simoncelli, E., and Zhu, S.-C. (2003). On Advances in Statistical Modeling of Natural Images. *Journal of Mathematical Imaging and Vision*, 18:17–33.

[Srivastava et al., 2002] Srivastava, A., Liu, X., and Grenander, U. (2002). Universal Analytical Forms for Modeling Image Probabilities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(9):1200–1214.

[Starck et al., 2000] Starck, J. L., Candès, E. J., and Donoho, D. L. (2000). The curvelet transform for image denoising. *IEEE Trans. on Image Process.*, 11(6):670–684.

[Steidl et al., 2004] Steidl, G., Weickert, J., Brox, T., Mrazek, P., and Welk, M. (2004). On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs. *SIAM J. Numer. Anal.*, 42(2):686–658.

[Stiller, 1997] Stiller, C. (1997). Object-based estimation of dense motion fields. *IEEE Trans. Image Process.*, 2:234–250.

[Strela et al., 2000] Strela, V., Portilla, J., and Simoncelli, E. P. (2000). Image denoising using a local gaussian scale mixture model in the wavelet domain. In *Proc. SPIE, 45th Annual Meeting*, San Diego.

[Tan and Jiao, 2006] Tan, S. and Jiao, L. (2006). Image denoising using the ridgelet biframe. *J. Opt. Soc. Amer. A.*, 23:2449–2461.

[Tan and Jiao, 2007] Tan, S. and Jiao, L. (2007). Multivariate statistical models for image denoising in the wavelet domain. *Int. J. Comput. Vision*, 75(2):209–230.

[Tay et al., 2006] Tay, D., Kingsbury, N. G., and Palaniswami, M. (2006). Orthonormal Hilbert-pair of wavelets with (almost) maximum vanishing moments. *IEEE Signal Processing Letters*, 13(9):553–536.

[Teh et al., 2003] Teh, Y. W., Welling, M., Osindero, S., and Hinton, G. E. (2003). Energy-based models for sparse overcomplete representations. *J. Mach. Learn. Res.*, 4:1235–1260.

[Tenenbaum et al., 2000] Tenenbaum, J., de Silva, V., and Langford, J. (2000). A global geometric framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323.

[Tessens et al., 2008] Tessens, L., Pižurica, A., Alecu, A., Munteanu, A., and Philips, W. (2008). Context adaptive image denoising through modeling of curvelet domain statistics. *Journal of Electronic Imaging*, 17(3):033021–1 – 033021–17.

[Tikhonov et al., 1990] Tikhonov, A., Goncharsky, A., Stepanov, V., and Yagola, A. (1990). *Numerical Methods For The Solution of Ill-Posed Problems*. Kluwer Academic Publishers.

[Tipping and Bishop, 1999] Tipping, M. E. and Bishop, C. M. (1999). Mixtures of Probabilistic Principal Component Analysers. *Neural Computation*, 11(2):443–482.

[Titterington, 1991] Titterington, D. M. (1991). Choosing the regularization parameter in image restoration. *Lecture Notes-Monograph Series*, 20:392–402.

[Tomasi and Manduchi, 1998] Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. *International Conference on Computer Vision (ICCV)*, pages 839–846.

[Tzikas et al., 2007] Tzikas, D., Likas, A., and Galatsanos, N. (2007). Variational bayesian blind image deconvolution with student-t priors. In *Proc. IEEE International Conference on Image Processing ICIP 2007*, volume 1, pages I–109–I–112.

[Unser and Van De Ville, 2009] Unser, M. and Van De Ville, D. (2009). Higher-order Riesz Transforms and Steerable Wavelet Frames. In *IEEE Int. Conf. Image Processing*, pages 3801–3804, Cairo, Egypt.

[Van De Ville et al., 2005] Van De Ville, D., Blu, T., and Unser, M. (2005). Isotropic polyharmonic b-splines: scaling functions and wavelets. *IEEE Trans. Image Processing*, 14(11):1798–1813.

[Van De Ville and Unser, 2008] Van De Ville, D. and Unser, M. (2008). Complex Wavelet Bases, Steerability, and the Marr-like pyramid. *IEEE Trans. Image Processing*, 17(11):2063–2080.

[Van den Bulcke and Franchois, 2009] Van den Bulcke, S. and Franchois, A. (2009). A Full-Wave 2.5D Volume Integral Equation Solver for 3D Millimeter-Wave Scattering by Large Inhomogeneous 2D Objects. *IEEE Trans. Antennas and Propagation*, 57(2):535–545.

[Van Leemput et al., 1999] Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999). Automated model-based bias field correction of mr images of the brain. *IEEE Trans. Medical Imaging*, 18(10):885–896.

[Van Trees, 1968] Van Trees, H. L. (1968). *Detection, Estimation and Modulation Theory*. Wiley, New York.

[Velisavljevic et al., 2006] Velisavljevic, V., Beferull-Lozano, B., Vetterli, M., and Dragotti, P. (2006). Directionlets: anisotropic multidirectional representation with separable filtering. *IEEE Trans. Image Processing*, 15(7):1916–1933.

[Verbeek, 2006] Verbeek, J. (2006). Learning Non-linear Image Manifolds by Global Alignment of Local Linear Models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(8):1236–1250.

[Verbeek et al., 2003] Verbeek, J., Vlassis, N., and Krï¿œse, B. (2003). Efficient greedy learning of Gaussian mixture models. *Neural Computations*, 15(2):469–485.

[Verveer and Duin, 1995] Verveer, P. and Duin, R. (1995). An evaluation of Intrinsic Dimensionality Estimators. *IEEE Transactions on Pattern Analysis and Machinal Intelligence*, 17(1):81–86.

[Vidakovic, 1998a] Vidakovic, B. (1998a). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. of the American Statistical Association*, 93:173–179.

[Vidakovic, 1998b] Vidakovic, B. (1998b). Wavelet-based nonparametric Bayes methods. In Dey, D. D., Müller, P., and Sinha, D., editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, volume 133 of *Lecture Notes in Statistics*, pages 133–155. Springer Verlag, New York.

[Vlassis and A., 2002] Vlassis, N. and A., L. (2002). A greedy EM algorithm for Gaussian mixture learning. *Neur. Proc. Lett*, 15(1):77–87.

[Vo et al., 2007] Vo, A., Nguyen, T., and Oraintara, S. (2007). Image denoising using shiftable directional pyramid and scale mixtures of complex gaussians. In *Proc. IEEE International Symposium on Circuits and Systems ISCAS 2007*, pages 4000–4003.

[Wainwright and Simoncelli, 2000] Wainwright, M. J. and Simoncelli, E. P. (2000). Scale mixtures of Gaussians and the statistics of natural images. *Adv. Neural Information Processing Systems (NIPS 1999)*, 12:855–861.

[Wainwright et al., 2001] Wainwright, M. J., Simoncelli, E. P., and Willsky, A. S. (2001). Random Cascades on Wavelet Trees and their use in modeling and analyzing natural images. *Applied Computational and Harmonic Analysis*, 11(1):89–123.

[Wang and Vannier, 1993] Wang, G. and Vannier, M. (1993). Helical CT image noise - analytical results. *Med. Physics*, 20(6):1635–1640.

[Wang et al., 2006] Wang, J., Guo, Y., Ying, Y., Liu, Y., and Peng, Q. (2006). Fast non-local algorithm for image denoising. In *Proc. of IEEE International Conference on Image Processing (ICIP)*, pages 1429–1432.

[Wang and Ostermann, 1998] Wang, Y. and Ostermann, J. (1998). Evaluation of mesh-based motion estimation in h.263-like coders. *IEEE Transactions on Circuits and Systems for Video Technology*, 8:243 – 252.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612.

[Weickert, 1998] Weickert, J. (1998). *Anisotropic Diffusion in Image Processing*. ECMI Series. Teubner-Verlag.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80–83.

[Wunderlich and Noo, 2008] Wunderlich, A. and Noo, F. (2008). Image covariance and lesion detectability in direct fan-beam x-ray computed tomography. *Phys. Med. Biol.*, 53(10):2471–2493.

[Xu and Jordan, 1996] Xu, L. and Jordan, M. I. (1996). On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation*, 8:129–151.

[Yi et al., 2009] Yi, S., Labate, D., Easley, G., and Krim, H. (2009). A Shearlet Approach to Edge Analysis and Detection. *IEEE Trans. Image Processing*, 18(5):929–941.

[Zhang and Wu, 2005] Zhang, L. and Wu, X. (2005). Color demosaicking via directional linear minimum mean square-error estimation. *IEEE Transactions on Image Processing*, 14(12):2167–2178.

[Zhang et al., 2006] Zhang, Y., Pham, B. T., and Eckstein, M. P. (2006). The Effect of Nonlinear Human Visual System Components Backgrounds. *IEEE Trans. Medical Imaging*, 25(10):1348–1362.

[Zhang and Zafar, 1992] Zhang, Y. Q. and Zafar, S. (1992). Motion-compensated wavelet transform coding for color video compression. *IEEE Trans. Circuits and Systems for Video Technology*, 2(3):285–296.

[Zhu and Starlack, 2007] Zhu, L. and Starlack, J. (2007). A practical reconstruction algorithm for CT noise variance maps using FBP reconstruction. In *Proc. SPIE Medical Imaging*, volume 6510, pages 651023.1–651023.8.

[Zhu et al., 1997] Zhu, S. C., Wu, Y. N., and Mumford, D. (1997). Minimax entropy principles and its application to texture modeling. *Neural Computation*, 9(8):1627–1660.

[Zimmer et al., 2008] Zimmer, S., Didas, S., and Weickert, J. (2008). A rotationally invariant block matching strategy improving image denoising with non-local means. In *Proceedings of International Workshop on Local and Non-Local Approximation in Image Processing*, pages 135–142.