

Copyright
by
Hamid Rahim Sheikh
2004

The Dissertation Committee for Hamid Rahim Sheikh
certifies that this is the approved version of the following dissertation:

**Image Quality Assessment Using Natural Scene
Statistics**

Committee:

Alan C. Bovik, Supervisor

Lawrence K. Cormack, Supervisor

J. K. Aggarwal

Chandrajit Bajaj

Gustavo de Veciana

Joydeep Ghosh

**Image Quality Assessment Using Natural Scene
Statistics**

by

Hamid Rahim Sheikh, B.Sc., M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2004

In the name of Allah, most Gracious, ever Merciful

This work is a humble offering to

The Grace that is my Lord

The love that are my parents

The joy that is my family

The wealth that is my health

The kindness that are my teachers

The care that are my friends

The bliss that would be my wife

The hope that would be my children

The wonder that is humanity

The promise that is posterity

Acknowledgments

I begin by penning down a testimony of gratitude to most Gracious and ever Merciful Allah, who is patient with His creation, covers and overlooks their faults, listens to their prayers, provides guidance and strength in times of need, and blesses their effort with fruition. Like a child cannot be thankful enough to his parents, so can a human not be thankful enough to God. I pray that may He give me the strength and the courage to live up to His expectations of me, and that may I never be ignorant or ungrateful of His Grace upon me.

I would also like to pay a tribute to the patience and sacrifice of my parents for my well-being. Without their lifelong support, I would not be standing where I do, and this dissertation would have been written by someone else. I would also like to thank my brother and sister for their love and support, and for the joy that they have always brought in my life. My friends too have always been very helpful in personal and professional spheres, and I would like to thank them all for putting up with me, and cheering me up when I feel down.

I am greatly indebted to my advisor, Dr. Alan Bovik, for putting up with me patiently for five years, and funding me even through years of no ‘output’ on my part. I am even more thankful to him for believing in me, and trusting my abilities to carry myself through the daunting task of

completing a dissertation. He provided me with support through rain and shine, and guidance on all and sundry, personal and professional. All of his students know him for being patient and forbearing with his students, never thrusting his opinion upon them, despite being watchful, always encouraging their creativity, and trusting their self-motivation. Personally, I could not have wished for a better advisor for myself. I found him to be completely compatible with my personality, and I consider myself extremely fortunate to have worked under his supervision.

I would also like to thank other members of my committee, Dr. Aggarwal, Dr. Bajaj, Dr. Cormack, Dr. de Veciana, and Dr. Ghosh for their help and guidance during my Ph.D., and Dr. Evans for encouraging me to do a Ph.D. in the first place.

I would like to thank current and past members of LIVE, Vidhya, James, Zhou, Shizhong, Umesh, Farooq, Mehul, Abtine, Yang, Hyohoon, Raghu, Brette, Kalpana, and Sumohana for their support, jokes, and friendship. Umesh deserves a special thanks for his help on academic and personal matters. The lab would be such a mess without him. He can be trusted to render selfless help to anyone, even if it comes at the cost of his personal time when he needs it himself. I would also like to thank Zhou for years of consistent guidance, arguments, and healthy criticism of my work, and for providing me with opportunities to devour his food supplies (see the acknowledgements in his dissertation!). I would also like to thank Melanie and Shirley for their help in administrative matters, and the folks at the international office and

the student health services for their attention and support.

On the professional side, I would like to thank Pierre Costa and Ahmad Ansari at SBC Research for funding a large part of my Ph.D., and to Raj Talluri, Youngjun Yoo, and Paul Fernandez at Texas Instruments for providing me with an excellent learning opportunity during my internship at TI.

I would also like to thank members of the Austin Ahmadiyya Muslim community for providing me with opportunities to keep my spiritual practices together, and a family environment with home culture and great food.

Lastly, I would like to thank the American people for providing me with a learning opportunity and a wonderful experience throughout my stay here in Austin, Texas. Their tolerance, hospitality, and flexibility towards people from other cultures and faiths has had a profound impact on me, and I have come to admire them greatly, and am thankful to have learned from them.

I am also grateful to God for lending me two entirely different perspectives on life: I have observed the world from the eyes of a third-world citizen, as well as from the perspective of the richest and the most powerful nation in the world. Coming to the USA has been an extremely rewarding experience for me, something that I would remember and cherish for the rest of my life. It has made me appreciative of the richness and the diversity, and yet the underlying sameness, of this mass of humanity, this mess of humanity, which is who we are . . . and who I am.

God bless humanity.

Preface

A young cup maker by the name of Hasan
In a new place, making his wine-cups, celebrating his love
has been strung with us by the strings of time
He has merged into us, as if our soul
For we are like raindrops pouring all night
(the night that has stretched over a thousand epochs)
falling against a windowpane, making snake-slithers
And now at this place at the dawn of eternity
We and this young cup maker
have been woven together into a dream!

—from *“Hasan the cup maker - IV”* by Noon Meem Rashid

Allow me to start this dissertation with a cliché: when it comes to communicating, a picture is worth a thousand words. It may be a cliché, but one that has scientific truth in it because the enormous amount of vision circuitry that we humans have, allows information to flow into us at a tremendous rate. Vision is the highest bandwidth sensory channel that we have for receiving input from, and gaining awareness of, our environment. A subtler side of the cliché is usually taken for granted, that being our ability to *store* and *reproduce* information in the form of a picture or a thousand words. Unlike other

animals, not only can we sense information, we can also express it through various forms, be it words, sounds, gestures, music, pictures, or dissertations.

Our intelligence is what sets us apart from animals. One product of this intelligence is that we are able to store and retrieve information *outside* our physical forms, in the form of writings, sketches, audio, images, videos, and what not. It is possible for us to sense or conceive information at some time, save it outside our physical forms, forget it, and then retrieve it later if need be. This dissertation could serve as an example, in which I have attempted to preserve what I have learned during years of gut-wrenching, hair-greying, and buttock-growing labor (please do not laugh). And while I am trying to forget the ‘information’ that I have saved in it, you, the reader, are trying to make sense out of it.

Time and again you will ask yourself ‘what is he saying?’ This unfortunately is a fundamental question that is inextricably tied with the whole notion of information reproduction: the question of *information fidelity*, whether or not the information retrieved from a reproduction is a true replica of the original source. There are many reasons why this may not be so, and while this could be out of deceit as well, we as engineers are more interested in the loss of information due to the limits imposed by devices and processes used during the creation of, and retrieval from, information storage. Unfortunately, due to the omnipresence of noise, there is only a faint asymptotic hope in being able to convey information exactly from a source to the receiver. So when I see this quote by L. Ron Hubbard outside the Church of Scientology that I

often pass by when I walk to school: “On the day we can learn to fully trust each other, there will be peace on Earth,” I cannot help but feel that only asymptotically this could happen, since trust comes from communication, and communication is thwarted by noise, and that the day that the honorable Ron Hubbard desires might be a tad late!

Thankfully, this dissertation concerns fidelity of information in images only. Images are a part of our daily lives now; we come across them a little too much in our everyday experience with the television, Internet, and print media. All these images (and videos) are reproductions of light stimuli that once existed in space-time, and the idea is for these frozen moments to be recaptured by those who view them. Being reproductions that they are, we have grown accustomed to not expecting too much out of them. We do not expect the richness of information that comes from ‘being there’ to come from looking at a picture or a video. Nevertheless, we have learned to have some use for images as sources of information, and our expectation of what an image should be is what could be called *image quality*. Thus, aesthetics aside, we consider an image or a video to be of high quality if it can evoke cognitive processes that are as close as possible to the ones caused by the original visual source within the limits of engineering and visual acuity.

Many things could happen to an image that could degrade its quality. These sources of *distortions* plague all practical systems right from the point where the light signal enters a camera, through different stages of processing, and finally to reproduction on paper or screen. People who design systems

dealing with pictures and videos are invariably interested in measuring the effect of their designs on the quality of these signals, since they try to maximize visual quality at a given cost. Humans can judge image quality almost as a reflex action, but researchers want computers to replace humans in this realm, not only due to vocational reasons, but also because there are too many images and videos out there, and it is just not economical, or even possible, to solicit human input at every moment of need.

Unfortunately, machine evaluation of image quality is a treacherously difficult problem that has yet to be solved satisfactorily. Researchers have been trying to develop quality assessment algorithms for the last three decades, and so far have been able to achieve only limited success. One of the biggest hurdles in their pursuit of a solution is that the problem involves understanding and modeling of not only human perception of visual stimuli, but also the behavior of humans when they interact with them. And to tell the truth, there is no consensus on what constitutes a ‘solution’ to the problem, or even what exactly the ‘problem’ is in the first place! Nevertheless, these algorithms improve each time a pertinent discovery is made by researchers in human vision and behavior, and inch closer to human ability by a few naïve, yet quantifiable, criteria of correspondence with human beings.

In this dissertation, I approach the quality assessment problem by asserting that distortions make images look unnatural, and that algorithms could be designed for quantifying the unnaturalness of images for machine evaluation of image quality.

Poets have always fascinated me with their ability to save information in mystical and mysterious ways. I have tried to adorn this dissertation with my translations of some pieces that have struck a soul-chord within me. Hopefully, my translation channel has enough capacity to allow readers to recover the messages with sufficient fidelity!

Hamid R. Sheikh

May, 2004.

Image Quality Assessment Using Natural Scene Statistics

Publication No. _____

Hamid Rahim Sheikh, Ph.D.
The University of Texas at Austin, 2004

Supervisors: Alan C. Bovik
Lawrence K. Cormack

Measurement of image quality is crucial for designing image processing systems that could potentially degrade visual quality. Such measurements allow developers to optimize designs to deliver maximum quality while minimizing system cost. This dissertation is about automatic algorithms for quality assessment of digital images.

Traditionally, researchers have equated image quality with *image fidelity*, or the closeness of a distorted image to a ‘reference’ image that is assumed to have perfect quality. This closeness is typically measured by modeling the human visual system, or by using different mathematical criteria for signal similarity.

In this dissertation, I approach the problem from a novel direction. I claim that quality assessment algorithms deal only with images and videos

that are meant for human consumption, and that these signals are almost exclusively images and videos of the visual environment. Image distortions make these so-called *natural scenes* look ‘unnatural’. I claim that this departure from ‘expected’ characteristics could be quantified for predicting visual quality.

I present a novel information-theoretic approach to image quality assessment using statistical models for natural scenes. I approach the quality assessment problem as an information fidelity problem, in which the distortion process is viewed as a channel that limits the flow of information from a source of natural images to the receiver (the brain). I show that quality of a test image is strongly related to the amount of statistical information about the reference image that is present in the test image.

I also explore image quality assessment in the absence of the reference, and present a novel method for blindly quantifying the quality of images compressed by wavelet based compression algorithms. I show that images are rendered unnatural by the quantization process during lossy compression, and that this unnaturalness could be quantified blindly for predicting visual quality.

I test and validate the performance of the algorithms proposed in this dissertation through an extensive study in which ground truth data was obtained from many human subjects. I show that the methods presented can accurately predict visual quality, and that they outperform current state-of-the-art methods in my simulations.

Table of Contents

Acknowledgments	v
Preface	viii
Abstract	xiii
Table of Contents	xv
List of Tables	xx
List of Figures	xxii
Chapter 1. Introduction	1
1.1 What Quality?	2
1.2 Why Measure Quality?	5
1.3 Measuring Quality	6
1.4 Objective Full-Reference Quality Assessment	7
1.4.1 The Mean Squared Error	7
1.4.2 Incorporating Human Visual System Modeling	8
1.4.3 Signal Fidelity and Structural Approaches	8
1.4.4 Natural Scene Statistics	9
1.4.5 Three Approaches to Quality Assessment	10
1.5 No-Reference Quality Assessment	11
1.6 Outline of the Dissertation	12
Chapter 2. Background	14
2.1 Image Quality Assessment	16
2.2 The Human Visual System	17
2.2.1 Anatomy of the HVS	18

2.2.2	Modeling HVS Function for QA Purposes	20
2.3	Full-Reference Image Quality Assessment Methods	28
2.4	No-Reference Image Quality Assessment methods	36
2.5	Natural Scene Statistics	41
Chapter 3.	Information-Theoretic Methods for Image Quality Assessment Using Natural Scene Statistics	47
3.1	Introduction	48
3.2	Limitations of Previous Full-Reference Methodologies	49
3.2.1	Proposed FR Solution	53
3.3	Limitations of Previous No-Reference Methodologies	54
3.3.1	Proposed NR Solution	55
Chapter 4.	An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics	56
4.1	Introduction	57
4.2	Information Fidelity Criterion for Image Quality Assessment	58
4.2.1	The Source Model	59
4.2.2	The Distortion Model	60
4.2.3	The Information Fidelity Criterion	62
4.3	Implementation Issues	67
4.3.1	Assumptions about the Source Model	67
4.3.2	Assumptions about the Distortion Model	67
4.3.3	Wavelet Bases and Inter-coefficient Correlations	68
4.3.3.1	Vector GSM	69
4.3.3.2	Downsampling	71
4.4	Subjective QA Study for Validation	73
4.5	Results	73
4.5.1	Simulation Details	74
4.5.2	Calibration of the Objective Score	75
4.5.3	Discussion	76
4.6	Similarities with HVS Error-Sensitivity Based QA Methods	82
4.7	Conclusions	89

Chapter 5. Image Information and Visual Quality	90
5.1 Introduction	92
5.2 Visual Information Fidelity for Image Quality Assessment . . .	94
5.2.1 The Source Model	94
5.2.2 The Distortion Model	95
5.2.3 The Human Visual System Model	96
5.2.4 The Visual Information Fidelity Criterion	97
5.3 Implementation Issues	104
5.3.1 Assumptions About the Source Model	104
5.3.2 Assumptions About the Distortion Model	106
5.3.3 Assumptions About the HVS Model	106
5.4 Subjective QA Study for Validation	107
5.5 Results	107
5.5.1 Simulation Details	108
5.5.2 Calibration of the Objective Score	108
5.5.3 Discussion	111
5.5.3.1 Overall performance	111
5.5.3.2 Cross-distortion performance	112
5.5.3.3 Dependence on the HVS parameter	115
5.5.3.4 Computational Complexity	116
5.6 Conclusions	116
Chapter 6. No-Reference Quality Assessment Using Natural-Scene Statistics: JPEG2000	118
6.1 Introduction	119
6.2 The JPEG2000 Image Compression Standard	121
6.3 Blind Measurement of Image Naturalness	126
6.3.1 Statistical Model for Natural Images in the Wavelet Domain	126
6.3.2 Compressed Natural Images	129
6.3.3 Features for Blind Quality Assessment	131
6.3.4 Image-dependent Threshold Calculations	133
6.3.5 Simplified Marginal Model	135
6.4 Results	136

6.4.1	Subjective QA Study for Training and Testing	136
6.4.2	Simulation Details	136
6.4.3	Quality Calibration	137
6.4.4	Quality Prediction Results	137
6.4.5	Discussion	138
6.4.6	Implementation Complexity	139
6.5	Conclusions	140
Chapter 7.	Conclusions and Future Work	148
7.1	Dissertation Summary	149
7.2	Contributions	150
7.3	Future Work	153
7.3.1	Full-Reference Image Quality Assessment	153
7.3.2	No-Reference Image Quality Assessment	155
7.3.2.1	Future Improvements for NR QA of JPEG2000	157
7.3.3	Reduced-Reference Image Quality Assessment	157
7.3.4	Video Quality Assessment	158
7.3.5	Image Information Metrics	159
Appendices		161
Appendix A.	Subjective Quality Assessment Study	162
A.1	Scope of the Subjective QA Study	162
A.2	Subjective Quality Assessment Methodologies	163
A.3	Designing the Image Database	165
A.3.1	Choice of Input Images	165
A.3.2	Distortion Types	166
A.4	Details of the Experiments	167
A.4.1	Test Methodology	167
A.4.2	Equipment and Display Configuration	168
A.4.3	Human Subjects, Training, and Testing	168
A.5	Processing of Raw Scores	171
A.5.1	Outlier Detection and Subject Rejection	171

A.5.2	MOS Scores	172
A.5.3	DMOS Scores	172
A.6	Results	173
A.6.1	Sample Images	173
A.6.2	Confidence Interval and RMS Error	173
A.6.3	Dependence of Quality on Distortion Parameters	175
Appendix B.	Similarities between IFC and HVS based FR QA	
	Methods	178
Bibliography		181
Vita		204

List of Tables

4.1	Validation scores for different quality assessment methods. The methods tested were PSNR, Sarnoff JND-Metrix 8.0 [84], MSSIM [131], IFC for scalar GSM without downsampling, IFC for scalar GSM with downsampling by 3 along orientation and 5 across, IFC for vector GSM, and IFC for vector GSM using horizontal and vertical orientations only, IFC for vector GSM and horizontal/vertical orientations with only the smallest eigenvalue in (4.22). The methods were tested against MOS from the subjective study after a non-linear mapping. The validation criteria are: correlation coefficient (CC), mean absolute error (MAE), root mean squared error (RMS), outlier ratio (OR) and spearman rank-order correlation coefficient (SROCC).	76
4.2	Validation scores for different quality assessment methods. The methods tested were PSNR, Sarnoff JND-Metrix 8.0 [84], MSSIM [131], IFC for scalar GSM without downsampling, IFC for scalar GSM with downsampling by 3 along orientation and 5 across, IFC for vector GSM, and IFC for vector GSM using horizontal and vertical orientations only, IFC for vector GSM and horizontal/vertical orientations with only the smallest eigenvalue in (4.22). The methods were tested against DMOS from the subjective study after a non-linear mapping. The validation criteria are: correlation coefficient (CC), mean absolute error (MAE), root mean squared error (RMS), outlier ratio (OR) and spearman rank-order correlation coefficient (SROCC).	77
4.3	Validation scores for the vector GSM IFC using all orientations versus using: only the horizontal and vertical orientations, and the subbands oriented at ± 60 deg. Only the smallest eigenvalue has been used in (4.22) for generating this table.	82
5.1	Validation scores for different quality assessment methods. The methods tested were PSNR, Sarnoff JND-Metrix 8.0 [84], MSSIM [131], VIF, and VIF using horizontal and vertical orientations only. The methods were tested against DMOS from the subjective study after a non-linear mapping. The validation criteria are: correlation coefficient (CC), mean absolute error (MAE), root mean squared error (RMS), outlier ratio (OR) and spearman rank-order correlation coefficient (SROCC).	111

5.2	RMSE performance of the QA methods on individual distortion types.	112
A.1	Distortion types and parameter used to control the degree of distortion. A JPEG2000 bitstream at a rate of 2.5 bits per pixel was transmitted over a simulated channel for the wireless channel distortion.	167
A.2	Subjective evaluation sessions: number of images in each session and the number of subjects participating in each session. The reference images were included in each session.	169
A.3	Average size of the 95% confidence interval about the mean μ , $[\mu - \delta, \mu + \delta]$ and RMSE for MOS and DMOS on a 1-100 quality scale.	174

List of Figures

2.1	Schematic diagram of the eye and the human visual system. . .	18
2.2	Normalized Contrast Sensitivity Function [4]	22
2.3	Frequency decompositions of various models.	24
2.4	(a) Implementation of masking effect for channel based HVS models. (b) Visibility threshold model (simplified): threshold elevation versus mask contrast	26
2.5	Non-linear contrast response saturation effects in neurons. . .	27
4.1	The quality assessment problem could be analyzed using an information theoretic framework in which a source transmits information through a channel to a receiver. The mutual information between the input of the channel (the reference image) and the output of the channel (the test image) quantifies the amount of information that could ideally be extracted by the receiver (the human observer) from the test image.	58
4.2	Downsampling a subband for reducing correlation between coefficients. The orientation selective filters in the steerable pyramid decomposition select image frequencies along the principal orientation (left). Downsampling along and across the principal orientation is a good way of reducing correlations in the filtered signal (right). In my simulations, I downsample by 3 along orientation and by 5 across orientation.	72
4.3	Scatter plots for the four objective quality criteria: PSNR, Sarnoff's JND-Metrix, MSSIM, and log(IFC) for vector GSM using horizontal/vertical orientations. The IFC shown here uses only the horizontal and vertical subbands at the finest scale, and only the smallest eigenvalue in (4.22). The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).	78

4.4	Scatter plots for the quality predictions by the four methods after compensating for quality calibration: PSNR, Sarnoff's JND-Metrix, MSSIM, and IFC for vector GSM using horizontal/vertical orientations. The IFC shown here uses only the horizontal and vertical subbands at the finest scale, and only the smallest eigenvalue in (4.22). The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).	79
4.5	An HVS error-sensitivity based quality measurement system. I show that this HVS model is the dual of the IFC under the assumptions regarding source and distortion models.	83
5.1	Mutual information between \mathcal{C} and \mathcal{E} quantifies the information that the brain could ideally extract from the reference image, whereas the mutual information between \mathcal{C} and \mathcal{F} quantifies the corresponding information that could be extracted from the test image.	93
5.2	The VIF can capture the quality improvement from linear contrast enhancement. A VIF value greater than unity indicates this improvement, while a VIF value less than unity signifies a loss of visual quality. The MSE between the reference image and the test images is approximately the same in this figure. .	103
5.3	Spatial maps showing how VIF captures spatial information loss.	105
5.4	Scatter plots for the four objective quality criteria: PSNR, Sarnoff's JND-Metrix, MSSIM, and $\log(\text{VIF})$ for VIF using horizontal/vertical orientations. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).	109
5.5	Scatter plots for the quality predictions by the four methods after compensating for quality calibration: PSNR, Sarnoff's JND-Metrix, MSSIM, and VIF using horizontal/vertical orientations. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).	110

5.6	Calibration curves for the four quality assessment methods for individual distortion types. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan). Note that VIF can be stably calibrated for predicting quality for a wider range of distortion types. The mapping used for MSSIM in this figure is $\log_{10}(1 - \text{MSSIM})$ for illustrative purposes.	114
5.7	Dependence of VIF performance on the σ_n^2 parameter. Note that VIF performs better than other methods against which it is compared in this chapter for all range of values of σ_n^2 shown in this figure: VIF (solid), PSNR (dashed), Sarnoff JND-Metrix 8.0 (dash-dot), and MSSIM (dotted)	115
6.1	Uncompressed image.	122
6.2	Four-level DWT coefficients (magnitude) of the uncompressed image using the 9/7 biorthogonal Wavelet. The subband coefficients are contrast stretched and non-linearly enhanced for better display.	123
6.3	DWT coefficients (magnitude) of the image after compression by JPEG2000. The subband coefficients are contrast stretched and non-linearly enhanced for better display.	124
6.4	Image compressed by JPEG2000	125
6.5	Joint histograms of $(\log_2 P, \log_2 C)$ for an uncompressed natural image at different scales and orientations of its wavelet decomposition. Top left: Diagonal subband at the finest scale. Top right: Horizontal subband at the finest scale. Bottom left: Vertical subband at the second-finest scale. Bottom right: Diagonal subband at the third-finest scale.	128
6.6	Joint histograms of $(\log_2 P, \log_2 C)$ for one subband of an image when it is compressed at different bit rates using the JPEG2000 compression algorithm. Top left: No compression. Top right: 2.44 bits/pixel. Bottom left: 0.78 bits/pixel. Bottom right: 0.19 bits/pixel.	130
6.7	Partition of the (P, C) space into quadrants. Also, the set of coefficients from which P is calculated in my simulations. . . .	131

6.8	$Mean(\log_2(C))$ versus subband enumeration index. The means of horizontal and vertical subbands at a given scale are averaged. (a) The subband enumeration n used in (b) & (c). (b) Uncompressed natural images. $Mean(\log_2(C))$ falls off approximately linearly with n . (c) $Mean(\log_2(C))$ for an image at different bit rates (solid) and the corresponding linear fits (dotted). The fits are computed from $n = 1 \dots 4$ only. Note that the estimated fits are quite close to the uncompressed image represented by the top most solid line. These linear fits are used for computing the image-dependent thresholds for the corresponding images.	142
6.9	Histograms of the logarithm (to the base 2) of the horizontal subband the finest scale for one image, before and after compression (at 0.75 bits per pixel). The dotted line denotes the threshold at -6.76.	143
6.10	Results: (a) Quality predictions versus MOS for one run of the algorithm. Normalized histograms of the RMSE for several runs of the algorithm using the joint statistics of P and C (b) and using the marginal statistics of C only (c). The dotted line in the histograms show the mean value, while the dashed line shows the standard deviation of human scores.	144
6.11	Results: Normalized histograms of the linear correlation coefficient for several runs of the algorithm using the joint statistics of P and C (b) and using the marginal statistics of C only (c). The dotted lines in the histograms show the mean value.	145
6.12	Variation of the RMSE with threshold offset in (6.5) for one simulation involving the entire database.	145
6.13	Content-dependence of the performance of the proposed NR QA algorithm over the quality range: (a) Algorithm performs consistently well over the range of quality. (b) The algorithm consistently over-predicts the quality. The MOS-Prediction point 'x' and the ± 2 standard deviation intervals are shown for each figure. The variance in the MOS is due to inter-human variability and the variation in the prediction is due to changes in the training set.	146
6.14	Content-dependence of the performance of the proposed NR QA algorithm over the quality range: The algorithm consistently under-predicts the quality. The MOS-Prediction point 'x' and the ± 2 standard deviation intervals are shown for each figure. The variance in the MOS is due to inter-human variability and the variation in the prediction is due to changes in the training set.	147
A.1	Interface for the subjective study.	169

A.2	Sample pictures from the set used to derive the database. . . .	174
A.3	Dependence of quality on distortion parameters for different distortion types: JPEG2000 and JPEG.	175
A.4	Dependence of quality on distortion parameters for different distortion types: White Noise and Gaussian Blur.	176
A.5	Dependence of quality on distortion parameters: bit errors in the JPEG2000 bit stream in a fast-fading Rayleigh channel. .	177

Chapter 1

Introduction

O Born of Time, tonight, down in your street,
in the icy darkness of the night,
I stand in front of your door, my hair in a mess
From the window above those enchanting eyes gaze upon me again tonight
Time, O Born of Time, is a potter's wheel on which
like decanters, goblets, wine-cups, candlesticks and vases,
men are made and unmade
I too am a human,
but the years that I passed in the incarceration of suffering!
Hasan the cup maker is a mound of dust today
without a trace of dampness
O Born of Time, in the morning, at Yusuf's perfumery
your eyes said something to me again
By the sparkling radiance of your eyes
has quivered in this dust-mound a soft throb of moisture
This, perhaps, will turn this dust into pliant clay!

Who knows the infiniteness of desire, O Born of Time, but
if you wish, I could be once again,
the same cup maker whose goblets
were the pride of every tavern of every city and town
Which lit up the abodes of the rich and poor.
Who knows the infiniteness of desire, O Born of Time, but
if you wish, I could turn once again to my long forsaken wine-cups
to the dry pits of clay and dust
to the means of sustenance and expression of my art.
That from this very clay, from this paint and polish
I shall conjure up the sparks with which
the ruins of hearts could be enlightened!

—from *“Hasan the cup maker - I”* by Noon Meem Rashid

1.1 What Quality?

Measurement of image quality is crucial to many image and video processing systems. Due to inherent physical limitations and economic reasons, the quality of images and videos could visibly degrade right from the point when they are captured to the point when they are viewed by a human observer. Examples of such distortions are numerous.

Examples of Distortions During Acquisition: Sources of distortions such as improper focusing, exposure, or camera movement (motion blur) are well

known even among nontechnical people. Other distortions that are equally prevalent, although less known to the general public, include light diffraction at the aperture, optical blur due to the lens, internal reflections inside the camera, chromatic aberrations in the lens, photosensor noise, loss of chromatic calibration, sampling noise, quantization noise etc. Most of these distortions could be minimized to acceptable levels for everyday use by investing time, effort, and money into the acquisition process. Yet some sources of distortions existing in other scenarios plague images to a much greater degree than what we are used to in our everyday experience, such as non-optical imaging devices used for medical image processing applications, astronomy, seismology, radar, sonar, etc.

Examples of Distortions During Processing: While there is an infinite number of ways in which image quality could degrade during processing, examples of a few practical distortion types include compression artifacts due to lossy compression, distortion resulting from transmission errors and packet losses during communication, and loss of quality due to changes in resolution, format, color conversion, and enhancement operations.

Examples of Distortions During Reproduction: These distortions occur when an image is displayed to a human observer. They include display nonlinearities, resolution limitations of the display devices, loss of chrominance and luminance calibration, dithering and halftoning for printing, etc.

Since image and video quality could degrade in almost all systems of practical importance, it is crucial for designers and developers to keep the

tradeoffs between visual quality and system cost in mind, and to optimize systems for providing maximum visual quality at a minimum cost. For example, it is common knowledge that the quality of DVD videos is much superior to the old analog VHS videos. This is because distortions have been significantly reduced due to advances in technology, and the resulting improvement in visual quality of DVD videos easily offsets the additional cost that consumers have to pay over VHS.

So it is all about the delicate balance between quality and cost. Obviously this cost could be in monetary terms: there is only X amount of money that could be spent on a device, and the consumer who pays those X dollars requires the best visual quality for his money. As members of a conscientious capitalist society, the goal of a designer and engineer is to present possibilities of profits to the entrepreneur who sells these image and video systems and the associated content to consumers. Such profits could only come from efficient design of these systems by utilizing the explosion in knowledge and technology to drive the cost down and the quality up.

As another example, the cost and benefit equation could be far more complex in systems used to capture and process medical images for diagnosis. In such applications, it is not a matter of profits by reducing the cost and increasing quality. Rather it is a matter of saving lives by enabling doctors to elicit greater diagnostic information from images by keeping the distortions to a minimum. For such systems, the aim of the designer is to maximize image quality at whatever cost that is humanly possible to incur.

1.2 Why Measure Quality?

Now that I have explained the importance of quality of images and videos and the associated cost-quality balance, the obvious question that arises is why do we need to *measure* quality. The answer is simple and could be illustrated by a few examples. If a designer is designing this high-end television, and wants to know what does the quality-cost curve look like, he obviously needs a mechanism for *measuring* the quality of the output video when his design is running at certain configuration costing a certain resource. In another scenario, a designer of a medical imaging device may want to decide which of the two alternative X-ray devices gives better results. He too needs a way of scientifically comparing the quality of the two systems. Another situation could be that of an entrepreneur, who sells entertainment content over cable, satellite, or the Internet, and wants to monitor the quality of the content that is delivered to his customers.

Thus, quality assessment algorithms are needed for three types of applications:

1. For optimization purposes, where one maximizes quality at a given cost.
2. For comparative analysis between different alternatives.
3. For quality monitoring in real-time applications.

1.3 Measuring Quality

The obvious way of measuring image or video quality is to solicit human opinion. This is known as *subjective* quality assessment method and the average opinion about quality of a group of human subjects is sought. However, such evaluations are time-consuming, cumbersome, and expensive to conduct, and the methodology is difficult to embed into real-world applications. For example, a researcher developing a state-of-the-art television system could not request two-dozen subjects for quality evaluation each time he tweaks a parameter in his design. Neither could a service provider place a group of employees at each point where the quality of the video service is to be monitored. More importantly, such subjective evaluation methods can not be embedded into image and video processing algorithms that trade-off resources for quality in real-time, for example, a video streaming server providing content over the Internet may want to adjust itself based on the real-time resource load, and may want to do so optimally.

What we really need are automatic algorithms for *objective* quality assessment that could analyze images and videos and report their quality without human involvement. The goal is to design methods whose evaluations are *in close agreement with human judgements*. Such methods could eliminate the need for expensive subjective studies. Unfortunately, the problem of machine evaluation of quality has turned out to be surprisingly difficult!

1.4 Objective Full-Reference Quality Assessment

Researchers in the field of image quality assessment (QA) have attempted to measure quality using the so-called full-reference (FR) framework. This framework is a consequence of our limited understanding of human perceptions of quality. It involves the following hypothesis: *The quality of an image could be evaluated by comparing it against a reference signal of perfect quality. A measure of the similarity between the reference image and the image being evaluated could be calibrated to serve as a measure of perceptual quality.* A full-reference algorithm therefore computes the similarity between the image or video whose quality is to be evaluated (called the *test* signal) and the associated reference signal.

1.4.1 The Mean Squared Error

One obvious way of measuring this similarity is to compute an error signal by subtracting the test signal from the reference, and then computing the average energy of the error signal. The mean-squared-error (MSE) is the simplest, and the most widely used, FR QA method. Unfortunately, MSE correlates poorly with subjective image quality, and the goal of researchers over the past three decades has been to design algorithms that can predict quality much more accurately than MSE.

1.4.2 Incorporating Human Visual System Modeling


Researchers assume that incorporating knowledge of the human visual system (HVS) and human perception into objective quality assessment algorithms could increase their accuracy. This HVS-based FR paradigm has been the dominant paradigm for the last three decades. The underlying premise is that humans do not perceive images as signals in a high-dimensional space, but are interested in various *attributes* of those images, such as brightness, contrast, shape and texture of objects, orientations, smoothness, etc. Since the sensitivity of the HVS is different for different aspects of images, it makes sense to account for these sensitivities while making a comparison between the test and the reference signals. Thus, HVS-based methods involve extensive modeling of the human visual system and the perceptual behavior of humans. Such methods, while somewhat successful, are plagued by the complexity of HVS models and the need for robust and accurate calibration, which may render practical implementations difficult.

1.4.3 Fidelity and Structural Approaches

Other researchers try to circumvent the challenges of modeling the HVS by exploring mathematical distance measures other than the MSE that would correlate better with subjective quality. Such methods have also achieved limited success. Recently, a novel similarity measure was proposed that aims to follow a new approach to the quality assessment problem [131]. This method claims that the purpose of the HVS is to extract cognitive information from

images, which almost exclusively comes from the *structure* of objects in images. Thus, one should quantify *structural distortion* in images instead of quantifying error strength. This approach has also achieved reasonable success.

1.4.4 Natural Scene Statistics

In this dissertation, I follow a different approach to the quality assessment problem. I assert that QA algorithms are involved with signals that are meant for human consumption, and such images and videos almost exclusively  come from the so-called set of natural signals. Researchers have long known that natural images¹, that is images and videos of the three dimensional visual environment, are extremely rare in the set of all possible images, just as works of English literature are extremely rare in the set of all possible combinations of the English alphabet. I claim that it is necessary to exploit the sparseness of this space of natural images, whose statistical characteristics could be disturbed by the distortions occurring in image processing systems. Thus, I claim that distortion processes make distorted images and videos look ‘unnatural’ to human observers, and this departure from expected natural behavior could be quantified for measurement of image quality.

I further propose to approach the quality assessment problem from a novel perspective: an information-theoretic perspective based on natural scene

¹In my treatment of natural images for quality assessment, I define natural images as ‘high quality’ images of the visual environment, including images of outdoor scenes, manmade objects and indoor environments, although the scope of the algorithms presented in this dissertation could be specialized to subclasses of natural images, or to other scenarios such as medical images, Radar, Sonar, computer generated graphics etc.

statistics (NSS). In this approach, the quality assessment problem is viewed as an information fidelity problem rather than a signal fidelity problem, in which a source of natural images tries to communicate to a receiver (the human brain) through a channel that imposes limitations (by introducing distortions) on how much information could flow through it. I formalize this perspective using well-known information theoretic techniques, that is, by quantifying the mutual information between the input and the output of the channel. Thus I claim that this mutual information, which quantifies the information about the reference image (the source) that could ideally be extracted by the receiver (the brain) from the test image (the output of the channel), is one aspect of statistical information fidelity that should relate well with visual quality. Moreover, I also view the HVS as a channel limiting the flow of information to the brain due to inherent physiological reasons. Using a channel model for the HVS as well, I quantify the statistical information content of an image, that is, the information that could ideally be extracted from the reference image by the brain. Thus, one fundamental contribution of this dissertation is the exploration of the connections between image information and visual quality.

1.4.5 Three Approaches to Quality Assessment

The approaches discussed above describe three ways in which one could look at the visual image quality assessment problem. One viewpoint is *structural*, from the image-content perspective, in which images are considered to be projections of objects in the three dimensional environment, which could





come from a wide variety of lighting conditions. Such variations constitute *non-structural* distortion that should be treated differently from structural ones, e.g., blurring or blocking, that could hamper identification of objects in the image. The second viewpoint is *psychophysical*, from the human visual receiver perspective, in which researchers simulate the processing of images by the human visual system, and predict visibility and perceptual significance of errors. The third viewpoint, the one that I take in this dissertation, is the *statistical* viewpoint that considers natural images to be signals with certain statistical properties. These three views are fundamentally connected with each other by the following hypothesis: *the physics of image formation of the natural three dimensional visual environment leads to certain statistical properties of the visual stimuli, in response to which the human visual system has evolved over eons*. However, different aspects of each of these perspectives may have different complexities when it comes to analysis and modeling. In this dissertation, I show that the statistical approach to image quality assessment requires few assumptions, is simple and methodical to derive, and yet it outperforms the other two approaches. Also, it is reassuring to discover that the statistical approach to quality assessment is indeed a *dual* of the psychophysical approach to the same problem.

1.5 No-Reference Quality Assessment

Providing the reference signal to the quality assessment algorithm renders such algorithms infeasible for most applications. FR QA methods are

basically useful only as a ‘lab tool’ for designing systems since the resource challenges in providing the reference signal in, say, a quality monitoring application, are virtually insurmountable. Thus, a different framework is needed for objective quality assessment, in which the algorithm does not rely on the availability of the reference signal to evaluate the quality of the test signal. This is the so-called no-reference (NR) quality assessment problem. It is obvious that human observers can easily evaluate the quality of images or videos without needing to view the reference signal. This inspires me into believing that the design of NR QA should theoretically be possible. Unfortunately, the NR problem is, as yet, an unsolved problem with no known generic NR QA algorithm that exists today. Only limited success has been achieved by limiting the scope of NR methods to specific distortion types.

In this dissertation, I claim that statistical models for natural images could provide the ‘reference’ against which test images could be compared for evaluation of their quality. Thus, I claim again that the unnaturalness of the test images could be quantified without the reference image for the evaluation of image quality by using NSS models.

1.6 Outline of the Dissertation

In Chapter 2, I present the background of the QA problem by discussing the HVS, FR, and NR methods for image quality assessment. I also discuss natural scene statistics models in Chapter 2. In Chapter 3, I talk about the limitations of the previous approaches and present my proposed solution. In

Chapters 4 and 5, I present two methods based on an information-theoretic setup for FR QA using NSS modeling, while in Chapter 6 I present a new NR QA method for images compressed by the JPEG2000 distortion. I conclude the dissertation in Chapter 7.

Chapter 2

Background

May the Lord on the Day of Reckoning be my redeemer
For I have seen Mrs. Salamanca's eyes
Mrs. Salamanca's eyes
whose horizons stretch beyond the blue vastness of the southern seas
The blue vastness of the southern seas
The islands of which sparkle in the abundance of morning light
In the sparkling islands are playgrounds of golden, maroon, and red birds
Stretched as if the plains of paradise
Birds flexing their wings in the endlessness of eternity!

May the Lord on the Day of Reckoning be my redeemer
For I have planted kisses on Mrs. Salamanca's lips
The kisses the springs of whose sweetness
Lie beyond the drunken gardens of golden, maroon, and red trees of the
northern lands
Where, from the breasts of life's fragrant blossoms
Terrified dream-bees suck nectar, and sip from that

with whose ecstatic vividness,
beneath the two ends of time's unseen arch,
resound the symphonies of the wetlands of matter and space
The songs of the wetlands with each other entwine
like the lips of Mrs. Salamanca with mine!

May the Lord on the Day of Reckoning be my redeemer
For I have seen
Mrs. Salamanca naked in the bed all night
That neck, those arms, those legs, those breasts
Which heave with the tempests of the southern seas
And effuse the fragrance of the gardens of the northern lands
Where every moment perfumes and tempests embrace and refrain
The uncovered body of Mrs. Salamanca
From horizon to horizon as if a grapevine
whose nourishment is the heavenly light, and the fruit:
Ecstasy that is truly boundless
Who, except the Lord, is flawless!

— *"Mrs. Salamanca" by Noon Meem Rashid*

2.1 Image Quality Assessment

There are two dominant types of quality assessment algorithms in the literature: the full-reference methods and the no-reference methods. As explained in Chapter 1, full-reference methods rely on the availability of a reference image, assumed to have perfect quality, against which a test (distorted) image is compared for evaluation of its quality. Researchers have primarily focused on FR methods that rely on modeling the human visual system to achieve accuracy in quality prediction, or by using various signal fidelity criteria to measure the ‘closeness’ of the test image to the reference. In this chapter, I will present an overview of some of the FR literature that belongs to both of these classes, as well as other application-dependent methods.

In no-reference quality assessment, the algorithm does not have access to the reference image, and only the test image could be processed to assess its quality. From an application point of view, NR methods are more desirable than FR methods, which are primarily meant for testing and validation applications. FR methods cannot be used for, say, monitoring video quality over a lossy network. Unfortunately, the NR quality assessment problem is largely unsolved, with limited success achieved by restricting the scope of the algorithms to specific distortion types, such blocking due to block-based compression algorithms, or blurring etc. Only a handful of NR methods have been proposed in the literature, and they too are mostly meant for the blocking artifact in images and videos. I will also review some NR QA methods in this chapter.

As mentioned in Chapter 1, I claim that image and video processing algorithms almost exclusively deal with signals that are meant to convey reproductions of visual information for human consumption. These signals belong to the class of the so-called *natural images*. The distortions present in real-world image and video processing systems make these signals look unnatural to human beings, and this unnaturalness could be quantified using statistical models for natural images. I will present some statistical properties of natural images later in this Chapter.

Section 2.2 provides a brief introduction to the human visual system, and its modeling for quality assessment purposes. This will lay a foundation for better understanding of the material in Section 2.3, where I present some full-reference quality assessment methods that have been proposed in the literature. Section 2.4 provides an overview of some no-reference methods. In Section 2.5 I will present an overview of natural scene statistics models.

2.2 The Human Visual System

Figure 2.1 schematically shows the early stages of the HVS. The functioning of higher levels of HVS is increasingly complicated, and is rarely modeled for image processing systems, primarily due to the lack of a complete understanding and the complexity of such models. However, the components of HVS depicted in Figure 2.1 are fairly well understood and accepted by the vision science community, and frequently used for image processing applications. A more detailed description of HVS may be found in [13, 126].

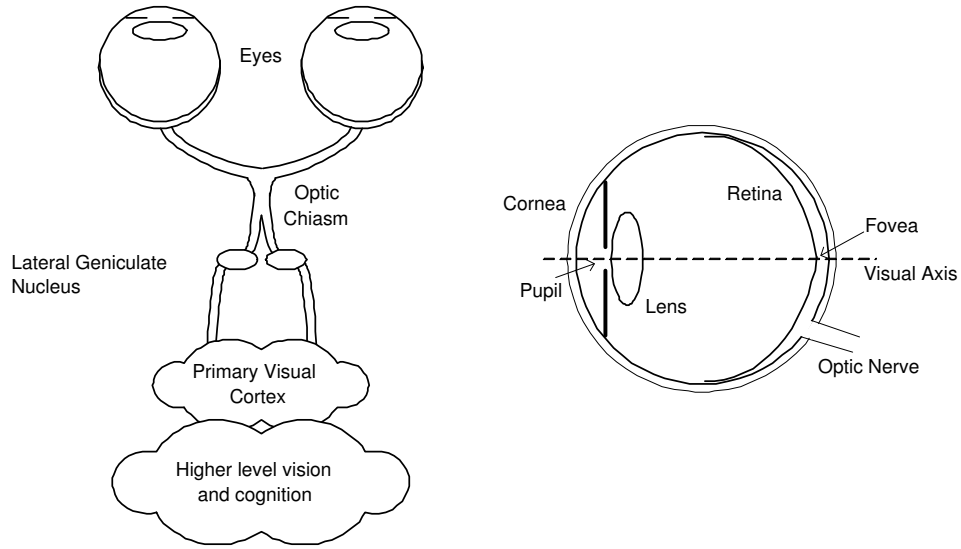


Figure 2.1: Schematic diagram of the eye and the human visual system.

2.2.1 Anatomy of the HVS

The visual stimulus in the form of light coming from objects in the environment is focussed by the optical components of the eye onto the retina, a membrane at the back of the eye that contains several layers of neurons, including photoreceptor cells. The optics consist of the cornea, the pupil (the aperture that controls the amount of light entering the eye), the lens and the fluids that fill the eye. The optical system focusses the visual stimulus onto the retina, but in doing so blurs the image slightly due to the inherent limitations and imperfections. The blur is low-pass, typically modeled as a linear space-invariant system characterized by a Point Spread Function (PSF). Photoreceptor cells sample the image that is projected onto the retina.

There are two types of photoreceptor cells in the retina: the cone cells

and the rod cells. The cones are responsible for vision in normal light conditions, while the rods are responsible for vision in very low light conditions, and hence are generally ignored in the modeling. There are three different types of cones, corresponding to three different light wavelength bands to which they are most sensitive. The L-cones, M-cones and the S-cones (corresponding to the Long, Medium and Short wavelengths at which their respective sensitivities peak) split the image projected onto the retina into three visual streams. The signals from the cone cells pass through several layers of interconnecting neurons in the retina before being carried off to the brain by the optic nerve.

The photoreceptor cells are non-uniformly distributed over the surface of the retina. The region of the retina that lies on the visual axis has the highest density of cone cells and is called the fovea (Figure 2.1). This density falls off rapidly with distance from the fovea. The distribution of the ganglion cells, the neurons that carry the electrical signal from the eye to the brain through the optic nerve, is also highly non-uniform, and drops off even faster than the density of the cone receptors. The net effect is that the HVS does not process the entire visual stimulus at uniform resolution.

The visual streams originating from the eye are reorganized in the Optic Chiasm and the Lateral Geniculate Nucleus (LGN) in the brain, before being relayed to the Primary Visual Cortex. The neurons in the visual cortex are known to be tuned to various aspects of the incoming streams, such as spatial and temporal frequency, orientation, and direction of motion. Typically, only spatial frequency and orientation selectivity is modeled by QA metrics. The

neurons in the cortex have receptive fields that are well approximated by two-dimensional Gabor functions. The ensemble of these neurons may be modeled as an octave-band Gabor filter bank [13, 126], where the spatial frequency spectrum (in polar representation) is sampled at octave intervals in the radial frequency dimension and uniform intervals in the orientation dimension [6]. Another aspect of the neurons in the visual cortex is their saturating response to stimulus contrast, where the output of a neuron saturates as the input contrast increases.

Many aspects of the neurons in the primary visual cortex are not modeled for quality assessment applications. The visual streams generated in the cortex project to other parts of the brain for further specialized processing of such attributes as motion and color. The functionality of the higher layers of the HVS is currently an active research topic in vision science.

2.2.2 Modeling HVS Function for QA Purposes

Optics and the PSF. The optics of the HVS are typically modeled as a linear space-invariant filter. Prior to applying the PSF, however, the pixel values may need to be converted into luminance values to reflect the intensity of light coming off from the display screen, which is typically accomplished by performing a point-wise non-linear operation on the digital pixel values.

Foveal and Peripheral Vision. As stated above, the densities of the cone cells and the ganglion cells in the retina are not uniform, peaking at the fovea and decreasing rapidly with distance from the fovea. A natural result

is that whenever a human observer fixates at a point in his environment, the region around the fixation point is resolved with the highest spatial resolution, while the resolution decreases with distance from fixation point. The high-resolution vision due to fixation by the observer onto a region is called *foveal* vision, while the progressively lower resolution vision is called *peripheral* vision. Most image quality assessment models work with foveal vision only, while some may incorporate peripheral vision as well [48, 53, 54, 127]. Models may also re-sample the image with the sampling density of the receptors in the fovea in order to provide a better approximation of the HVS as well as to provide more robust calibration of the model [53, 54].

Light Adaptation. The HVS operates over a wide range of light intensity values, spanning several orders of magnitude from a moonlit night to a bright sunny day. It copes with such a large range by a phenomenon known as *light adaptation*, which operates by controlling the amount of light entering the eye through the pupil, as well as adaptation mechanisms in the retinal cells that adjust the gain of post-receptor neurons in the retina. The result is that the retina primarily encodes the contrast of the visual stimulus instead of coding absolute light intensities. Many models work with band-limited contrast for complex natural scenes [68], which is tied with the channel (subband) decomposition (see below).

Contrast Sensitivity Functions. The Contrast Sensitivity Function models the variation in the sensitivity of HVS to different spatial and temporal frequencies that are present in the visual stimulus. This variation may be

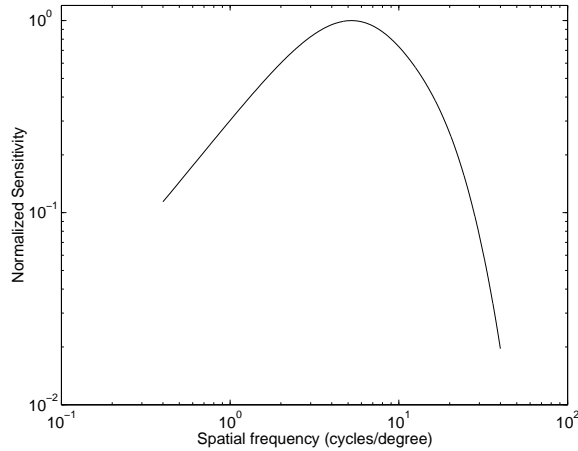


Figure 2.2: Normalized Contrast Sensitivity Function [4]

explained by the characteristics of the ganglion cells and the cells in the Lateral Geniculate Nucleus, or as internal noise characteristics of HVS neurons. Consequently, some models of HVS choose to implement CSF as a linear filtering operation, while others implement CSF through weighting factors for subbands after a frequency decomposition. The CSF varies with distance from the fovea as well, but for foveal vision, the spatial CSF is typically modeled as a space-invariant band-pass function (Figure 2.2).

Channel (Subband) Decomposition. While the cortical receptive fields are well represented by 2D Gabor functions, the Gabor decomposition is difficult to compute and lacks some of the mathematical conveniences that are desired for simple implementation, such as invertibility, reconstruction by addition etc. Watson constructed the Cortex Transform [142] to model the frequency and orientation selective channels (or subbands), which have similar profiles as 2D Gabor functions but are more convenient to implement. Chan-

nel decomposition models used by Watson [142], Daly [16, 17], Lubin [53, 54] and Heeger and Teo [32], Teo and Heeger [112, 113] attempt to model the HVS as closely as possible without incurring prohibitive implementation difficulties. Simpler transforms such as the Wavelet transform, or even the Discrete Cosine Transform have been reported in the literature primarily due to their suitability for certain applications rather than their accuracy in modeling the cortical neurons. The subband configuration for some of the models described in this chapter is given in Figure 2.3.

Masking and Facilitation. Masking and facilitation are important aspects of the HVS in modeling the interactions between different image components present at the same spatial location. Masking/facilitation refers to the fact that the presence of one image component (called the *mask*) will decrease/increase the visibility of another image component (called the *test* signal). The mask generally reduces the visibility of the test signal in comparison with the case when the mask is absent. However, the mask may sometimes facilitate detection as well. Usually, the masking effect is strongest when the mask and the test signal have similar frequency and orientation.

Masking is typically implemented within each channel. Most models implement masking in the form of a gain-control mechanism that weights the error signal (between the distorted and the reference images) in a channel by a space-varying *visibility threshold* for that channel [144]. The visibility threshold adjustment at a point is calculated based on the energy of the reference signal (or the reference and the distorted signals) in the neighborhood

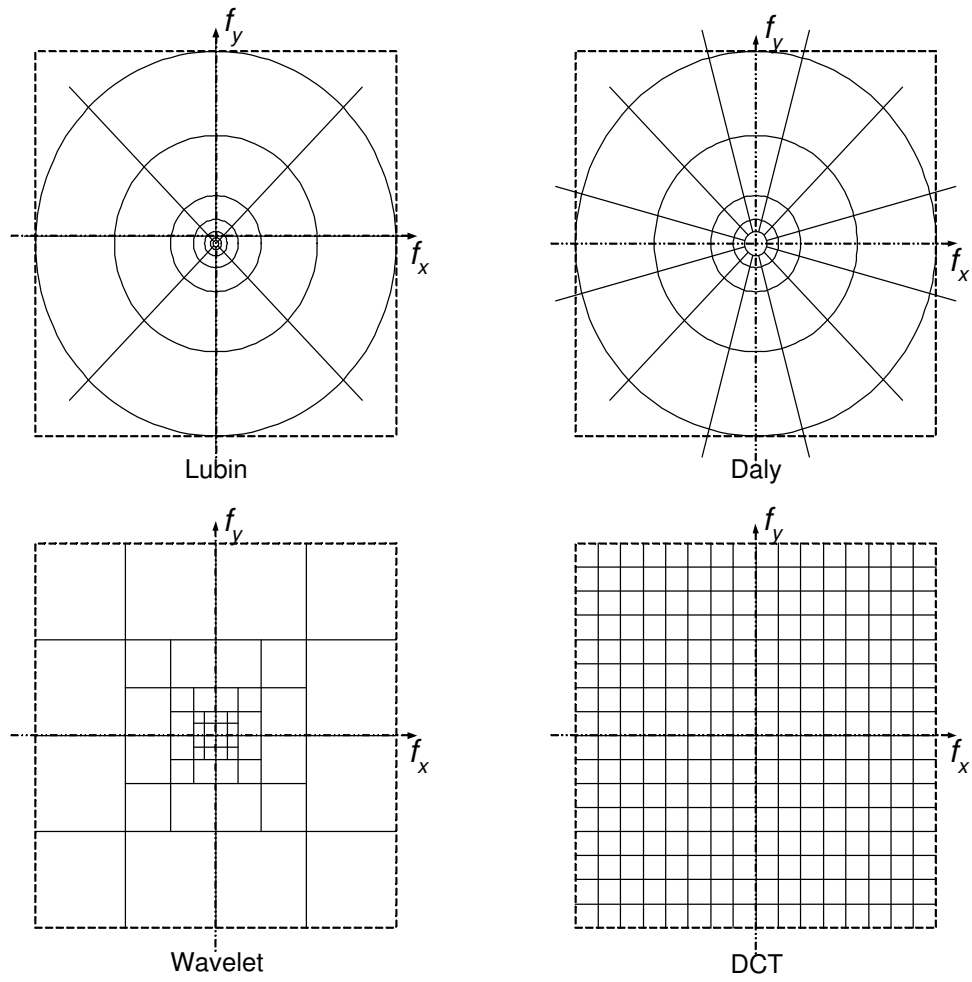


Figure 2.3: Frequency decompositions of various models.

of that point, as well as the HVS sensitivity for that channel in the absence of masking effects (also called the *base-sensitivity*). Figure 2.4(a) shows how masking is typically implemented in a channel. For every channel the base error threshold (the minimum visible contrast of the error) is modified (typically by elevating it) to account for the presence of the masking signal. The threshold elevation is related to the contrast of the reference (or the distorted) signal in that channel through a relationship that is depicted in Figure 2.4(b). The elevated visibility threshold is then used to normalize the error signal. This normalization typically converts the error into units of Just Noticeable Difference (JND), where a JND of 1.0 means that the distortion at that point in that channel is just at the threshold of visibility.

Some methods implement masking and facilitation as a manifestation of contrast response saturation. Figure 2.5 shows a set of curves each of which may represent the saturation characteristics of neurons in the HVS. Masking can be explained by the need for greater input stimulus (target contrast) needed for the detection of the target in the presence of the mask since the neuron operates closer to the saturation region in the presence of the mask than in its absence. Metrics may model masking with one or more of these curves.

Pooling. Pooling refers to the task of arriving at a single measurement of quality, or a decision regarding the visibility of the artifacts, from the visual streams. Most quality assessment metrics use Minkowski pooling to pool the error signal from the different frequency and orientation selective streams, as

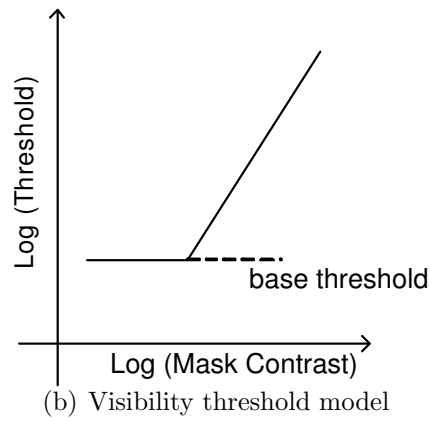
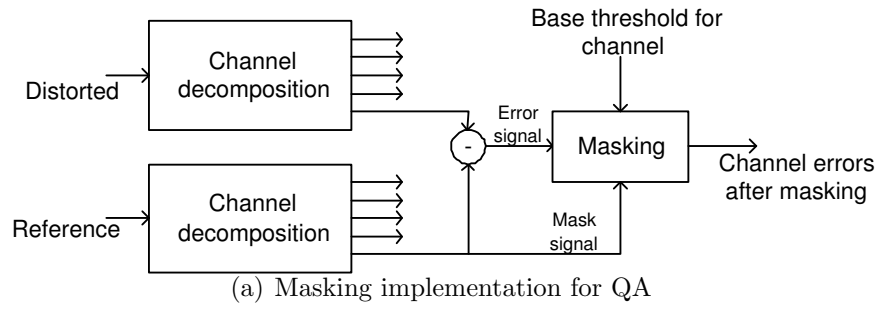


Figure 2.4: (a) Implementation of masking effect for channel based HVS models. (b) Visibility threshold model (simplified): threshold elevation versus mask contrast

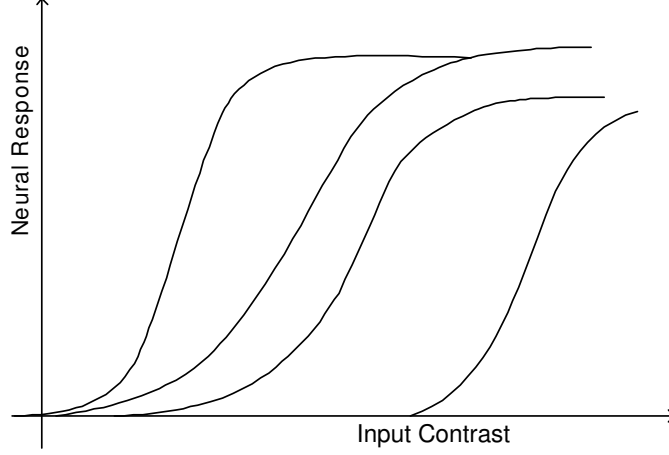


Figure 2.5: Non-linear contrast response saturation effects in neurons.

well as across spatial coordinates, to arrive at a fidelity measurement. For most quality assessment methods, pooling takes the form:

$$E = \left(\sum_l \sum_k |e_{l,k}|^\beta \right)^{\frac{1}{\beta}} \quad (2.1)$$

where $e_{l,k}$ is the normalized and masked error of the k -th coefficient in the l -th channel, and β is a constant typically between 1 and 4. This form of error pooling is commonly called Minkowski error pooling. Minkowski pooling may be performed over space (index k) and then over frequency (index l), or vice-versa, with some non-linearity between them, or possibly with different exponents β . A spatial map indicating the relative importance of different regions may also be used to provide spatially variant weighting to different $e_{l,k}$ [66, 151].

Summarizing the above discussion, an elaborate quality assessment algorithm may implement the following HVS features (apart from modeling

display device characteristics):

1. Eye optics modeled by a low-pass PSF.
2. Non-uniform retinal sampling.
3. Color processing.
4. Light adaptation (luminance masking).
5. Contrast sensitivity functions.
6. Spatial frequency, temporal frequency (for video QA) and orientation selective signal analysis.
7. Masking and facilitation.
8. Contrast response saturation.
9. Pooling.

2.3 Full-Reference Image Quality Assessment Methods

The standard paradigm for quality assessment (QA) of images has been to assume the availability of a reference image of perfect quality. An image whose quality is to be evaluated is considered to be a sum of the perfect reference signal and an error signal. The loss of quality is assumed to be directly related to the strength of the error signal, and the error strength is quantified in a meaningful way. The simplest implementation of the concept

is the MSE and the corresponding fidelity metric, the Peak Signal to Noise Ratio (PSNR):

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2 \quad (2.2)$$

$$\text{PSNR} = 10 \log_{10} \frac{L^2}{\text{MSE}} \quad (2.3)$$

where N is the number of pixels in the image, and x_i and y_i are the i -th pixels in the original and the distorted images respectively. L is the dynamic range of the pixel values. For 8bits/pixel monochromatic signal, L is equal to 255. MSE and PSNR are widely used for quality assessment because they are simple to calculate, and are mathematically easy to deal with for optimization purposes (MSE is differentiable, for example). However, they have been widely criticized as well for not correlating well with perceived visual quality [21, 67, 130]. There are a number of reasons why PSNR may not correlate well with the human perception of quality:

1. Digital pixel values on which the MSE is typically computed, may not exactly represent the light stimulus entering the eye.
2. The sensitivity of the HVS to the errors may be different for different aspects of the error signal, such as its spatial frequency, orientation, spatial position and extent, and may also vary with visual context. Such aspects may not be captured adequately by the global difference operator in the definition of MSE.

3. Simple error summation, like the one implemented in the MSE formulation, may be markedly different from the way the HVS and the brain arrives at an assessment of the perceived distortion.
4. Two distorted image signals with the same amount of error energy may have very different *structure* of errors, and hence different perceptual quality.

Image and video quality metrics in the last three decades have tried to improve upon the MSE by addressing the above issues. Perhaps the earliest such attempt was made by Mannos and Sakrison [56], which has been extended by other researchers over the years by incorporating more knowledge about the HVS.

In this section, I present a review of some FR image QA metrics that are based on quantifying the error between the reference and the distorted images in a perceptually meaningful way. A more detailed review of image QA metrics may be found in [21, 67, 134].

Perhaps the earliest attempt to improve upon the MSE by incorporating HVS modeling was made by Mannos and Sakrison [56]. The HVS modeling employed was a simple point non-linearity and CSF weighting. Some of the more recent HVS based FR quality metrics employ more sophisticated models for improving quality predictions.

The Visible Differences Predictor (VDP) by Daly [16, 17] aims to compute a probability-of-detection map between the reference and the distorted

signals. The value at each point in the map is the probability that a human observer will perceive a difference between the reference and the distorted images at that point. The reference and the distorted images (expressed as luminance values instead of pixels) are passed through a series of processes: point non-linearity, CSF filtering, channel decomposition, contrast calculation, masking effect modeling, and probability-of-detection calculation. A modified Cortex Transform [142] is used for channel decompositions, which transforms the image signal into five spatial levels followed by six orientation levels, leading to a total of 31 independent channels (including the baseband). For each channel, a threshold elevation map is computed from the contrast in that channel. A psychometric function is used to convert error strengths (weighted by the threshold elevations) into a probability-of-detection map for each channel. Pooling is then carried out across the channels to obtain an overall detection map.

Lubin [53, 54] presents an algorithm that too attempts to estimate a detection probability of the differences between the original and the distorted images. A blur is applied to model the PSF of the eye optics. The signals are then re-sampled to reflect the photoreceptor sampling in the retina. A Laplacian pyramid [9] is used to decompose the images into seven resolutions (each resolution is one-half of the immediately higher one), followed by band-limited contrast calculations [68]. A set of orientation filters implemented through steerable filters of Freeman and Adelson [26] is then applied for orientation selectivity in four orientations. The CSF is modeled by normalizing

the output of each frequency-selective channel by the base-sensitivity for that channel. Masking is implemented through a sigmoid non-linearity, after which the errors are convolved with disk-shaped kernels at each levels before being pooled into a distortion map using the Minkowski pooling across frequency. An additional pooling stage may be applied to obtain a single number for the entire image. The work in [53, 54] forms the basis for the famous Sarnoff JND Vision Model [83].

Teo and Heeger’s metric [32, 112, 113] uses PSF modeling, luminance masking, channel decomposition, and contrast normalizations. The channel decomposition process uses the Quadrature Steerable Filters [99] with six orientation levels and four spatial resolutions. A detection mechanism is implemented based on squared error. Masking is modeled through contrast normalization and response saturation. The contrast normalization is different from Daly’s or Lubin’s method in that they take the outputs of channels at all orientations at a particular resolution to perform the normalizations. Thus, this model does not assume that the channels at the same resolution are independent. Only channels at different resolutions are considered to be independent. The output of the channel decomposition after contrast normalization is decomposed four-fold by passing through four non-linearities of shapes as illustrated in Figure 2.5, the parameters for which are optimized to fit the data from psychophysical experiments.

Watson’s DCT metric [138] is based on 8×8 DCT transform commonly used in image and video compression. Unlike the models above, this method

partitions the spectrum into 64 uniform subbands (8 in each Cartesian dimension). After the block-based DCT and the associated subband contrasts are computed, a visibility threshold is calculated for each subband coefficient within each block based on the base-sensitivity for that subband. The base sensitivities are derived empirically. The thresholds are corrected for luminance and texture masking. The error in each subband is weighted by the corresponding visibility threshold and pooled using Minkowski pooling spatially. Pooling across subbands is then performed using Minkowski formulation with a different exponent. Another HVS based QA method in the DCT domain is presented in [55], and generalized to multidimensional contrast perception models in [70].

Safranek-Johnston’s perceptual image coder [81] incorporates a quality metric using a similar strategy as in Watson’s DCT metric. The channel decomposition uses a Generalized Quadrature Mirror Filter (GQMF) [149] for analysis and synthesis. This transform splits the spectrum into 16 uniform subbands (four in each Cartesian dimension). Masking and pooling takes similar approaches as in Watson’s DCT metric.

Bradley [7] reports a Wavelet Visible Difference Predictor (WVDP), which is a simplification of Daly’s VDP described above. He uses Watson et al. [141] derivation of 9/7 Wavelet quantization-noise detection thresholds for a 9/7 biorthogonal wavelet and combines it with a threshold elevation and psychometric detection probability scheme similar to Daly’s. Another Wavelet based metric has been proposed by Lai and Kuo [45]. Their metric is based on

the Haar Wavelet and their masking model incorporates channel interactions as well as suprathreshold effects (modeling the HVS when the distortion is much larger than the threshold of distortion visibility).

The quality metrics proposed above are scalar valued metrics. Damera-Venkata et al. [18] proposed a metric for quantifying the performance of image restoration systems, in which the degradation is modeled as a linear frequency distortion and additive noise injection. They develop two complementary metrics that separately quantify these distortions. They observed that if the additive noise is uncorrelated with the reference image, then an error measure from an HVS based metric will correlate well with the subjective judgement. Using a spatially adaptive restoration algorithm [43] (which was originally designed for inverse-half-toning), they isolate the effects of noise and linear frequency distortion. The noise is quantified using a multichannel HVS based metric. A Distortion Measure quantifies the spectral distortion between the reference and the model restored image.

Some researchers have attempted to measure image quality using single-channel models with the masking-effect models specifically targeting certain types of distortions, such as the blocking artifact. Blocking is recognized as one of the most annoying artifacts in block-DCT based image/video compression such as JPEG, especially at high compression ratios. Karunasekera and Kingsbury [38, 39] proposed a quality metric for blocking artifact. Edge detection is performed first on the error image. An activity map is calculated from the reference image in the neighborhood of the edges, and an activity-masked

edge image is computed such that edges that occur in high activity areas are de-emphasized. The activity-masked edge image is adjusted for luminance masking. A non-linear transformation is applied before pooling. The parameters for the model are obtained from experiments that measure the sensitivity of human observers to edge artifacts embedded in narrow-band test patterns.

In [11] a Peak Signal to Perceptible Noise Ratio (PSPNR) is defined, which is also a single-channel metric. They model luminance masking and activity masking to obtain a JND profile. The PSPNR has the same definition as given in (2.3) except that the MSE expression is adjusted for the JND profile. Another single-channel metric is the Objective Picture Quality Scale (PQS) by Miyahara [62], a number of features that can capture various distortions are combined into one score.

A novel single channel FR quality assessment metric is proposed in [128], where the metric attempts to measure the structural correlation between the reference and the distorted images. The metric performs quite well across image types and distortions. An interesting feature of this metric is that its output is between 0 and 1, where 1.0 implying perfect matching with the reference and 0 implies no structural match with the reference. The metric was improved later into the Structural Similarity (SSIM) metric by resolving some numerical instability issues [131].

A number of comparative studies have been undertaken in the literature that compare the relative performance of different FR QA methods [3, 23, 27, 49, 57, 59, 80, 137]. In my view, all except the recent study by

VQEG, involved a validation study that was not comprehensive enough to draw solid conclusions. This is evidenced by the fact that researchers claimed good performance of their methods on small scale studies, but many such methods did not perform as well as previously reported under more stringent testing conditions. The Phase-I of VQEG study concluded, after extensive psychometric testing, that all of the ten video quality assessment algorithms that were tested, which included some of the most sophisticated video quality assessment algorithms of the time¹, were ‘statistically indistinguishable’ from one another [22, 120]. Unfortunately, this group included the PSNR! In the Phase-II of testing [121], two (four) of the six proponents were declared to be statistically better than the rest in the 525 (625)² test cases. The Phase-II also concluded that none of the methods tested was statistically the same as the ‘null model’, or an ideal quality assessment algorithm that predicts the mean subjective quality exactly. This means that there is still a lot of improvement to be made.

2.4 No-Reference Image Quality Assessment methods

No-reference quality assessment for images and videos has recently received a great deal of attention since providing the reference signal for FR

¹Some description of the proponent methods has been provided in the final reports, as well as at the VQEG web site. Since almost all of the proponents were from the industry, the technical details provided in these documents are scanty.

²This refers to the number of horizontal lines in the display devices used for the testing.

quality metrics is simply impossible in almost all applications³. The main philosophy of previous NR quality assessment has been that of blind distortion measurement. In this section I will review some of the background work on NR quality assessment.

In [150] a blocking artifact measurement method is described for video signals. The algorithm assumes that the blocking edge occurs every eight pixels, and the measure of blocking artifact is the luminance-weighted norm of the pixel difference across block boundaries. This is then adjusted by activity based normalization. Liu et al. [52] describe another NR blocking artifact measurement algorithm, which is also based on measuring the edge strength at the expected block location. Once again, their method assumes the presence of the block boundary every eight pixels, and the blocking metric measures the average gray-level slope at the block boundary adjusted by the average gray-level slope at other locations.

Frequency domain blocking artifact metrics have also been investigated by Wang et al. [129] and Tan and Ghanbari [108]. The algorithm described in [129] assumes that the presence of a blocky-signal will introduce spikes in the Power Spectral Density (PSD) of the edge-enhanced image at known frequencies. The strength of these spikes relative to a median-smoothed version of the PSD (median smoothing will remove these spikes) quantifies the strength of

³If one has the resources to provide a reference image or video for FR QA, why use the distorted version at all? FR QA is therefore mostly used for in-lab testing and design, and not for tasks such as quality assessment at the user end, or quality monitoring over video broadcasting networks etc.

the blocking artifact. Masking effects are also incorporated by using the JND profile as described by Chou and Li [11]. In [108], a similar technique for measuring the strengths of frequencies present in the image gradient that can be attributed to the blocking artifact, relative to the strengths of other frequencies, is described. Both these methods assume periodicity (with a period of eight pixels) of the blocking artifact but unlike the spatial domain techniques described above, they are robust to shifts in the block-boundary grid.

A DCT-domain approach for measuring the blocking artifact blindly in images is described by Bovik and Liu [5]. The algorithm estimates the amplitude of a discontinuity between two adjacent 8×8 blocks using the DCT coefficients of those blocks. This is then normalized by the local activity. An efficient implementation is also described.

Although the methods described above measure the blocking distortion, no attempt is made to relate the measurement to perceptual quality or annoyance for human observers. In [133] activity and blocking artifact measures are calibrated against results from subjective quality experiments to yield a NR perceptual quality assessment algorithm. Meesters and Martens [60] also propose a blockiness measurement algorithm whose output is designed to conform with subjective judgements of blockiness in images.

Kayargadde and Martens [41] present an NR QA scheme that measures noise and blur in images blindly. They first degrade images using white noise and Gaussian blur and then construct a psychometric space using Multidimensional Scaling (MDS). MDS technique constructs alternate spaces from

the noise and blur measurement data, in which the positions of the images are related to their noisiness and blur measures (subjective or objective), with respect to those of other images in that set. Based on MDS analysis on subjective noise and blur measures [40], they derive quality predictions from the MDS analysis of the objective noise and blur measures. However, as pointed out by the authors as well, the distortions targeted by their method are contrived, independent of the images and stationary, and hence are easier to quantify, whereas the distortion occurring in most practical systems is image dependent and spatially varying.

Oguz et al. [64] propose a Visible Ringing Measure (VRM) that captures the ringing artifact around strong edges. The algorithm is based on constructing an image mask that exposes only those parts of the image that are in the vicinity of strong edges, and the ringing measure is considered to be the pixel intensity variance around the edges in the masked image. However, the results are only reported as part of a ringing artifact removal algorithm on one graphics image (a cartoon), no results are reported on natural images, and VRM has not been compared to (or calibrated against) human judgements of quality. The VRM was used as an alternative to MSE for quantifying the performance of the post-processing algorithm in [153].

Marziliano et al. [58] present a NR blur metric that is based on measuring average edge transition-widths as an indication of blur in an image. Subjective experiments using JPEG2000 distortion on five images (total of 25 distorted images) and ten subjects were used to test the metric. The metric

performs much better on the Gaussian blur since it is a stationary, image-independent distortion. However when tested on the LIVE image quality assessment database [93], its performance is much poorer [58]. This is because JPEG2000 causes blur in a nonuniform manner in addition to ‘ringing’ artifacts around strong edges, and the diversity of image content and quality range in the LIVE database highlights the limitations of [58]. The authors also tested their algorithm on their own subjective study with JPEG2000 compressed images, and the performance of their algorithm on JPEG2000 images has room for improvement.

Li [50] presents an algorithm that aims to measure several distortions present in an image blindly: global blur (based on assumed Gaussian blurs of step edges), additive white and impulse noise (based on local smoothness violation), blocking artifact (based on simple block boundary detection), and ringing artifact (based on anisotropic diffusion). The paper reports results of individual distortion measurements on *one* image only, with no subjective validations of the metrics reported.

All the techniques mentioned above have the following theme: *distortion statistics are known*. No explicit modeling of the natural scene statistics are done, rather an attempt is made to quantify distortion based on the statistics of the *distortion* alone. Jannsen and Blommaert [36] present a novel way of interpreting image quality in terms of the ‘naturalness’ (the degree to which an image conforms to the ‘expected’) and ‘usefulness’ (the degree to which objects in an image can be discriminated). In a nutshell, their predictions for

image quality, for images distorted by chroma scaling and variations in the color temperatures of the display devices, are based on the statistical distributions of the brightness and chrominance components of natural images⁴, and the degree to which a given image matches the set in distribution. While the NR reported results still have much room for improvement in terms of prediction accuracy and the target distortion types (the authors only worked on luminance and chrominance distortion), and the authors have not proposed extensions of using naturalness measures to quantify FR image quality, the concept of using statistics of natural images for quality assessment, together with knowledge about the distortion process, is the main theme of the research presented in this dissertation.

2.5 Natural Scene Statistics

The key emphasis of this dissertation is to exploit the fact that image and video quality assessment algorithms almost exclusively operate upon so-called natural scenes. One could easily imagine an experiment in which a human subject is asked to judge the quality of random noise, or to decide whether one patch of random noise looks better than another patch. Such quality assessment tasks have no meaning. Yet these questions are valid in the previous FR framework since no modeling of the image source is employed. The crucial fact is that human viewers are primarily interested in the quality of images that have cognitive relevance, and such images are, almost exclusively,

⁴Assumed to be given by the normalized histograms of a set of natural images.

pictures and videos of the natural three-dimensional visual environment. It should be obvious that the subspace of natural signals is an exceedingly tiny subset in the set of all possible images. It is easy to imagine the sparseness of this subspace by hypothetically considering how long it would take for a random image generator to generate anything that looks like a natural scene (more traditionally, the time required for a monkey on a typewriter to reproduce Shakespeare!). For researchers pursuing evermore efficient systems, it makes sense to exploit the sparseness of this subspace if it leads to algorithmic advantages.

More specifically, images and videos of the visual environment captured using high quality capture devices operating in the visual spectrum are broadly classified as natural scenes. This differentiates them from text, computer generated graphics, cartoons and animations, paintings and drawings, random noise, or images and videos captured from non-visual stimuli such as Radar and Sonar, X-Rays, ultra-sounds etc. Natural scenes form an extremely tiny subset of the set of all possible images [25, 78]. Many researchers have attempted to understand the structure of this subspace of natural images by studying their statistics [8, 25, 33, 47, 78, 97, 101, 104, 114, 122–125]. Researchers believe that the visual stimulus emanating from the natural environment drove the evolution of the HVS, and that modeling natural scenes and the HVS are essentially dual problems [102]. While many aspects of the HVS have been studied and incorporated into quality assessment algorithms, a usefully comprehensive (and feasible) understanding is still lacking. NSS

modeling may serve to fill this gap.

An important property of natural scenes is *scale invariance*, which means that the statistics of natural scenes are approximately constant with respect to changes in scale (zooming in or zooming out) up to a multiplicative constant [25, 78]. This property manifests itself in simple image models, such as the approximately $1/f$ fall-off of the amplitude spectral density, and has been hypothesized to occur due to the distribution of the sizes of objects in natural images [79]. Other interesting invariances have also been reported, such as the invariance of the image intensity histograms' shape, and the histograms of image gradient with scale. Non-linear multiresolution models [8, 96, 125]. capture those non-linear statistical dependencies that cannot be analyzed using linear methods. It has been claimed that these nonlinear dependencies could be explained solely by the presence of edges in natural images [20]. Furthermore, researchers have claimed that scaling of many statistics of natural images could be explained by the so-called dead leaves model, which is an image formation model where objects occlude one another [47, 114]. Localized models using Principal Components Analysis (PCA) and Independent Components Analysis (ICA), including local statistics that attract human gaze, have revealed insightful similarities between the statistics of natural scenes and the physiology of the HVS [31, 75, 76, 119]. Researchers have also proposed that the HVS has evolved to efficiently code the natural stimulus. This 'efficient coding hypothesis' has been analyzed using mathematical as well as physiological methods [65].

The FR methods that I propose in this dissertation model natural images in the wavelet domain using Gaussian Scale Mixtures (GSM) [122]. Scale-space-orientation analysis (loosely referred to as a wavelet decomposition) of images has been found to be useful for many applications, and researchers have found it useful for natural image modeling as well. It is well known that the coefficients of a subband in a wavelet decomposition are neither independent nor identically distributed, though they may be approximately second-order uncorrelated [97], and a coefficient is likely to have a large variance if its neighborhood has a large variance. Also the marginal densities of the wavelet coefficients are sharply peaked around zero with heavy tails, which are typically modeled as Laplacian density functions, while the localized statistics of wavelet coefficients, such as local mean and variance, are highly space-varying. Researchers have characterized this behavior of natural images in the wavelet domain by using GSMs [122, 125], which could be thought of as Gaussian random fields (RF) whose space-varying variance is also a random field [2]. A more detailed introduction to GSM's will be given in the following chapters.

Natural scene statistics have been explicitly incorporated into a number of image processing algorithms. Image compression algorithms, such as EZW [88], SPIHT [82] or JPEG2000 [111] make use of multiscale coefficient trees to capture non-linear redundancies. Kivanç et al. use a doubly stochastic statistical model of wavelet coefficients of natural images to design denoising algorithms [61]. Lam and Goodman present doubly stochastic models for the Discrete Cosine Transform (DCT) coefficients of natural images using Gener-

alized Gaussian Distributions [46]. Romberg et al. present a two-state hidden Markov model for the wavelet coefficients of images and use this model for image denoising [77]. The model was also used for segmentation [10]. Wainwright et al. use a multiscale Gaussian scale mixture model for denoising [124, 125]. The model has also been used for denoising in [71, 74]. In [72], an iterative projection technique is used to project noisy images onto the space of images having autocorrelations expected from noise-free natural images. Buccigrossi and Simoncelli use a multiscale NSS model for designing optimal bit allocation schemes for a wavelet based image compression algorithm [8], while similar models using complex wavelets have been used for image texture analysis and synthesis [73].

While the characteristics of the human visual systems and properties of certain distortion types have been explicitly incorporated into quality assessment algorithms, assumptions made about the statistics of images are usually quite simplistic. Specifically, most algorithms suffice to assume that images are smooth and low-pass in nature. In this dissertation, I claim that distortions present in real-world systems make the images look unnatural, and that this departure from an expected natural behavior could be quantified in various ways for a variety of QA applications.

Using this philosophy, I present an information-theoretic setup in which I quantify the unnaturalness of images using a well-known statistical distance criterion: the mutual information. I use the information-theoretic framework to design two full-reference quality assessment algorithms that are based on

various aspects of *image information* and *information fidelity*, that not only constitute a novel approach to the quality assessment problem, but are also highly competitive with the state-of-the-art methods, and outperform them in my simulations (Chapters 4 and 5). I also adapt an NSS model using prior information about the distortion process (and hence incorporating distortion modeling as well) to construct a simplified model that can characterize images compressed by wavelet based compression algorithms, such as JPEG2000, as well as uncompressed natural images. I use this model to quantify the departure of an image from natural behavior, and make no-reference quality predictions (Chapter 6).

The success of these methods show that human perceptions of image quality in the presence of perceptible distortions are indeed related to the naturalness of images. This observation has similarly been reported by Jannsen and Blommaert [36] for images distorted by chroma scaling and color temperatures of the display devices, where the researchers blindly measured the departure from naturalness by using chroma statistics of natural images.

Chapter 3

Information-Theoretic Methods for Image Quality Assessment Using Natural Scene Statistics

You would laugh, O Born of Time, isn't it strange?
That I am a treasure trove of passion
and a worshipper of worldliness as well
and the greatness I don't have, I seek that as well!
You who laughed that night at my hesitations
laugh again at my duality!
But what has anyone ever gained from love except for himself?
O Born of Time
Every love is question to which, except the lover,
there is no answer
It is enough if the call of the soul reverberates!

O Born of Time
It was indeed the call of my soul
that reverberated against the icy rims of centuries of my art

And it was indeed the shore of the ocean of your eyes
engulfing the centuries gone by
This ocean, which, to my soul, is a mirror
This ocean, which, to my ever forming,
ever un-forming wine-cups, is a mirror
This ocean, which, to every art,
every lover of art,
is a mirror.

—from “*Hasan the cup maker - II*” by Noon Meem Rashid

3.1 Introduction

The dominant paradigm in the field of full-reference quality assessment has been that of modeling the human visual system and taking into account its different sensitivities to different characteristics of images and distortions when a measure of ‘signal error’ is computed. Some methods also utilize signal fidelity criteria not based on HVS modeling to measure the closeness of a test image to the corresponding reference. Recently, a structural approach to the quality assessment problem has been proposed that claims to quantify the loss of image structure instead of measuring just the error strength. Moreover, certain application dependent methods exploit special distortion characteristics to design quality assessment algorithms. Many such methods were reviewed in Chapter 2.

No-Reference quality assessment methods are more desirable from an application perspective, since they do not rely on the availability of the reference image. Most of the NR methods proposed in the literature have focused on measuring the blocking artifact. A number of NR methods proposed in the literature were reviewed in Chapter 2.

3.2 Limitations of Previous Full-Reference Methodologies

I explained in Chapter 2 that the full-reference quality assessment algorithms proposed in the literature lack any modeling of image sources, and the sparseness of the subspace of natural images is not exploited. This means that under the existing framework, a question like ‘what is the quality of a randomly generated image corrupted by JPEG distortion?’ or a statement like ‘the quality of a random image with reference to another random image is really bad’ have valid and quantifiable meanings, but subjectively these two statements are irrelevant, since human beings are not interested in the quality of random images, or the loss of fidelity in them. Moreover, most of the previous methods view the QA problem as a signal fidelity problem, where images are deterministic points in a multidimensional space and image differences are measured using distance metrics, ignoring the inherently statistical nature of the quality assessment problem. For example, not only are images themselves less yielding to deterministic modeling, but many physiological processes in the brain, human behavior, and unpredictable environmental factors, call for

a statistical approach.

Many limitations of HVS based full reference methods are well known, and some of them have been discussed in [131]:

The Quality Definition Problem. The assumption that image fidelity is the same as image quality is only loosely true. In [95], this assumption is questioned in the light of empirical evidence. As will become apparent later, one FR method that I propose in this dissertation has the ability to discriminate quality *improvements* over the reference image! Thus, the very assumption that fidelity or ‘closeness’ to the reference image relates with image quality is not always true, and ‘distance’ may sometimes imply improvement in visual quality.

The Suprathreshold Problem Most HVS models are based on psychophysical experiments that deal with signals at the threshold of visual acuity. The generalizations of these models to the case where the distortions and degradations are typically much greater than the visibility threshold is hard to justify.

The Natural Image Complexity Problem The psychophysical experiments from which almost all HVS models are derived use simple patterns such as sinusoids, Gabors, noise patches, lines, boxes, etc. and hence the generalizations of such models for use in QA algorithms that deal with complex natural scenes is again hard to justify.

The Cognitive Interaction Problem QA algorithms assume homogene-

ity of visual information and are blind to the fact that human observers may base their judgements of quality on the *content* of the image. Also, perceptual quality is dependent on the task that the human subject needs to perform. Such *contextual* effects are also very hard to model in any QA algorithm.

Apart from the ‘philosophical’ limitations of the HVS-based quality assessment framework, there are some practical limitations as well. The strong dependence of these methods on data obtained from careful psychophysical experiments, the need for knowledge of the viewing conditions such as the viewing distance, ambient light levels, screen resolution, screen type and illumination settings, as well as the need to calibrate HVS models, hampers practical implementation of these methods.

Signal fidelity methods that rely solely on mathematical similarity measures also have their limitations. Their main disadvantage is that it is hard to analyze the reasons of failure or success of such methods, and their design is basically a trial-and-error exercise. In [23], fourteen of such methods were evaluated in a single study in hopes of discovering a metric that predicts quality well. In [3] also, twenty-two signal fidelity criteria are tested, along with four HVS based methods. However, such measures are generally attractive since they do not require any modeling of the HVS, calibration, or psychophysical data. They are typically faster in processing as well, and may have nice mathematical properties, such as differentiability, that could make them amenable to analysis and optimizations.

In [131], it was claimed that structural QA methods avoid some of the

limitations of HVS based methods since they are not based on threshold psychophysics or the HVS models derived thereof. However they have some limitations of their own. Specifically, although the structural paradigm for QA is an ambitious paradigm, there is no widely accepted definition of structure and structural distortion that is perceptually meaningful and quantifiable. In [131], the SSIM was constructed by *conjecturing* the functional forms of structural and non-structural distortions and the interaction between them. Although these forms are intuitive, and the resulting SSIM performs remarkably well for its simplicity, yet the *precise formulations* of structural and non-structural distortions leaves one wondering if it is the structural *framework* leads to the good performance of SSIM, or is it because of the *specific equations* used in it, and whether or not other formulations for ‘structure’ and ‘non-structure’ would also yield good quality assessment metrics. Furthermore, SSIM also requires optimization of some parameters for accuracy and numerical stability, and the exact ‘physical’ interpretation of these parameters is a tad unclear.

In this dissertation, I take a new approach to the quality assessment problem. As mentioned in Chapter 1, the third alternative to QA, apart from HVS based, signal fidelity and structural approaches, is the statistical approach, which I use in an information-theoretic setting. Needless to say, even my approach will make certain assumptions, but once assumptions regarding the source and distortion models and the suitability of the Shannon information as a valid measure of perceptual information fidelity are made, the components of the proposed algorithms and their interactions fall through

with minimal reliance on arbitrary formulations. Moreover, I will show that this new approach is a dual of the HVS based QA approach, with key similarities and contrasts between the two. This duality is a manifestation of the duality between NSS and HVS models.

3.2.1 Proposed FR Solution

In this dissertation, I propose a novel approach to the full-reference quality assessment problem. Instead of a signal fidelity framework for quantifying image quality, I propose to use an information-theoretic framework based on natural scene statistics. Specifically, I model the problem as a source of natural images communicating to a receiver (the human brain) through a channel (the distortion operation) that imposes limits on how much information could flow through it. I claim that the amount of information shared between the input and the output of the channel (the information that actually makes it through the channel) should relate well with visual quality. Moreover, I will also quantify the amount of information that the brain could extract from the reference image given the limitations of the human visual system. I will present two methods based on this new framework, each having its own merit and scope of application. They are presented in the Chapters 4 and 5.

3.3 Limitations of Previous No-Reference Methodologies

The field of NR quality assessment is far from being a mature topic, with the current research leaving much to be desired. One obvious shortcoming of the existing methodologies is their inability to cope with anything other than the blocking artifact¹. With the emerging compression technologies that avoid the blocking artifact, such as the JPEG2000 image compression standard and the H.264 video compression standard [1, 85], as well as certain types of distortion not arising from lossy compression becoming more common, such as channel errors in wireless communications, there is a need to conduct research into expanding the scope of current NR QA algorithms.

Current algorithms for doing NR QA are specific to the blocking distortion. The explicit use of natural scene statistics models is absent from almost all current methodologies. Only rudimentary image statistics models are employed in the design of the QA algorithms: natural images are smoothly varying with most of their spectral energy concentrated in the lower spatial frequencies, or that natural images do not have blocking artifacts or white noise. More explicit models for natural images could be used for designing current NR QA metrics.

¹Stationary, image independent distortions such as Gaussian blurs and white noise are contrived problems and do not arise in most applications, and hence doing NR QA for such types of distortions is not very useful.

3.3.1 Proposed NR Solution

In this dissertation, I propose to use explicit natural scene statistics model for NR quality assessment as well. I propose to adhere to the philosophy of doing NR QA by doing blind distortion measurement with respect to a statistical model for natural images. Thus I assume that the image whose quality is being evaluated is a distorted reproduction of a natural scene, and that *a priori* information about the distortion process is known as well. Using source and distortion models, I quantify the unnaturalness of images compressed by wavelet based compression methods, and relate this unnaturalness to image quality. In Chapter 6, I will present some research that I have completed for NR QA for JPEG2000 compressed images.

Chapter 4

An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics

O Born of Time

That pool at the caravanserai of Aleppo, the night and its stillness

In which we swam in harmony together

Like a circumference circumscribes a circle

We swam together all night

our souls and bodies in unison

We swam away from a self-fulfilling apprehension

Like water moves inside a tear-drop

Content with each other we swam

against the currents of declining youth

And you had exclaimed: "Hasan, it has brought you here as well,
the thirst in your soul!"

(And as I recalled the thirst in my soul

my throat was blessed by the munificence of my tears!)

—from "*Hasan the cup maker - III*" by Noon Meem Rashid

4.1 Introduction

In this chapter, I explore a novel information theoretic criterion for image fidelity using Natural Scene Statistics (NSS)¹. Images and videos of the three dimensional visual environment come from a common class: the class of natural scenes. Natural scenes form a tiny subspace in the space of all possible signals, and researchers have developed sophisticated models to characterize the statistics of this subspace. Most real-world distortion processes disturb these statistics and make the image or video signals unnatural. I propose to use natural scene models in conjunction with distortion models to quantify the Shannon information shared between the test and the reference images, and posit that this shared information is an aspect of fidelity that relates well with visual quality [90].

Section 4.2 presents the development of the information fidelity criterion. Implementation and subjective validation details are provided in Sections 4.3 and 4.4, while the results are discussed in Section 4.5. In Section 4.6 I compare and contrast the proposed method with HVS based methods, and conclude the chapter in Section 4.7.

¹Copyright 2004 IEEE. Some of the material in this chapter has been reproduced, with permission, from: H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics," IEEE Trans. Image Processing, accepted March 2004.

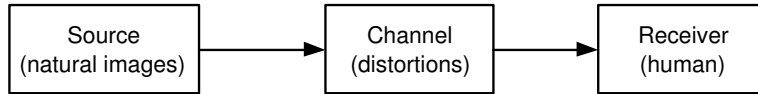


Figure 4.1: The quality assessment problem could be analyzed using an information theoretic framework in which a source transmits information through a channel to a receiver. The mutual information between the input of the channel (the reference image) and the output of the channel (the test image) quantifies the amount of information that could ideally be extracted by the receiver (the human observer) from the test image.

4.2 Information Fidelity Criterion for Image Quality Assessment

In this chapter, I propose to approach the quality assessment problem as an information fidelity problem, where a natural image source communicates with a receiver through a channel. The channel imposes fundamental limits on how much information could flow from the source (the reference image), through the channel (the image distortion process) to the receiver (the human observer). Figure 4.1 shows the scenario graphically. A standard way of dealing with such problems is to analyze them in an information-theoretic framework, in which the mutual information between the input and the output of the channel (the reference and the test images) is quantified using a model for the source and a distortion model. Thus, my assertion in proposing this framework is that the *information* that a test image has of the reference is a good way of quantifying fidelity that could relate well with visual quality.

4.2.1 The Source Model

The NSS model that I use for the development of the proposed FR method is the Gaussian Scale Mixture (GSM) model in the wavelet domain. It is convenient to deal with one subband of the wavelet decomposition at this point and later generalize this for multiple subbands. I model one subband of the wavelet decomposition of an image as a GSM random field (RF), $\mathcal{C} = \{C_i : i \in \mathcal{I}\}$, where \mathcal{I} denotes the set of spatial indices for the RF. \mathcal{C} is a product of two stationary RF's that are independent of each other [122, 125]:

$$\mathcal{C} = \mathcal{S} \cdot \mathcal{U} = \{S_i \cdot U_i : i \in \mathcal{I}\} \quad (4.1)$$

where $\mathcal{S} = \{S_i : i \in \mathcal{I}\}$ is an RF of positive scalars and $\mathcal{U} = \{U_i : i \in \mathcal{I}\}$ is a Gaussian scalar RF with mean zero and variance σ_U^2 . Note that for the GSM defined in (4.1), while the marginal distribution of C_i may be sharply-peaked and heavy-tailed, such as those of natural scenes in the wavelet domain, conditioned on S_i , C_i are normally distributed, that is,

$$p_{C_i|S_i}(c_i|s_i) \sim \mathcal{N}(0, s_i^2 \sigma_U^2) \quad (4.2)$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian density with mean μ and variance σ^2 . Another observation is that given S_i , C_i are independent of $S_j \forall j \neq i$, meaning that the variance of the coefficient C_i specifies its distribution completely. Additionally, if the RF \mathcal{U} is white, then the elements of \mathcal{C} are conditionally independent given \mathcal{S} . The GSM framework can model the marginal statistics of the wavelet coefficients of natural images, the non-linear dependencies that

are present between the coefficients, as well as the space-varying localized statistics, by introducing correlations in the RF \mathcal{S} [51].

4.2.2 The Distortion Model

The distortion model that I use in my work is also described in the wavelet domain. It is a simple signal attenuation and additive Gaussian noise in each subband:

$$\mathcal{D} = \mathcal{G}\mathcal{C} + \mathcal{V} = \{g_i C_i + V_i : i \in \mathcal{I}\} \quad (4.3)$$

where \mathcal{C} denotes the RF from a subband in the reference signal, $\mathcal{D} = \{D_i : i \in \mathcal{I}\}$ denotes the RF from the corresponding subband from the test (distorted) signal, $\mathcal{G} = \{g_i : i \in \mathcal{I}\}$ is a deterministic scalar attenuation field and $\mathcal{V} = \{V_i : i \in \mathcal{I}\}$ is a stationary additive zero-mean Gaussian noise RF with variance σ_V^2 . The RF \mathcal{V} is white, and is independent of \mathcal{S} and \mathcal{U} . This model captures two important, and complementary, distortion types: blur and additive noise. I assume that most distortion types that are prevalent in real world systems could roughly be described *locally* by a combination of these two. In this model, the attenuation factors g_i can capture the loss of signal energy in a subband to the blur distortion, while the process \mathcal{V} can capture additive noise components separately. Additionally, changes in image contrast that result from variations in ambient light are not modeled as noise since they too can be incorporated into the attenuation field \mathcal{G} .

The choice of a proper distortion model is crucial for image fidelity assessments that are expected to reflect perceptual quality. In essence we want

the distortion model to characterize what the HVS perceives as distortion. Based on my experience with different distortion models, I am inclined to hypothesize that the visual system has evolved over time to optimally estimate natural signals embedded in *natural distortions*: blur, white noise, and brightness and contrast variations due to changes in ambient lighting. The visual stimulus that is encoded by the human eyes is blurred by the optics of the eye as well as the spatially-varying sampling in the retina. It is therefore natural to expect evolution to have worked towards near-optimal processing of blurry signals, say for controlling the focus of the lens, or guiding visual fixations. Similarly, white noise arising due to photon noise or internal neuron noise (especially in low light conditions) affects all visual signals. Adaptation in the HVS to changes in ambient lighting has been known to exist for a long time [126]. Thus the HVS signal estimators would have evolved in response to natural signals corrupted by natural distortions, and would be near-optimal for them, but sub-optimal for other distortion types (such as blocking and colored noise) or signal sources. Hence ‘over-modeling’ the signal source or the distortion process is likely to fail for QA purposes, since it imposes assumptions on the existence of near-optimal estimators in the HVS for the chosen signal and distortion models, which may *not* be true. In essence distortion modeling combined with NSS source modeling is the *dual* of HVS *signal estimator* modeling.

Another hypothesis is that the field \mathcal{G} could account for the case when the additive noise \mathcal{V} is linearly correlated with \mathcal{C} . Thus if $D_k = C_k + V_k$ and

$V_k = aC_k + N_k$, where a is some constant and N_k is another noise RF uncorrelated with C_k , then $g_k = 1 + a$. Previously researchers have noted that as the correlation of the noise with the reference signal increases, the MSE becomes poorer in predicting perceptual quality [18]. While the second hypothesis could be a corollary to the first, I feel that both of these hypotheses (and perhaps more) need to be investigated further with psychophysical experiments so that the exact contribution of the distortion model in the quality prediction problem could be understood properly. For the purpose of image quality assessment presented in this dissertation, the distortion model of (4.3) is adequate, and works well in my tests.

4.2.3 The Information Fidelity Criterion

Given a statistical model for the source and the distortion (channel), the obvious information fidelity criterion is the mutual information between the source and the distorted image. Mutual information captures the information (in bits) that the source (the reference image) and the distorted (the test image) signals have in common, and it is one way of characterizing image fidelity. In a medical imaging application, Garcia et al. [28] show that the mutual information between a reference and a distorted mammogram image is strongly correlated with the ability of a human radiologist to correctly identify tumors. Thus, a statistical measure of information has been shown to affect semantic, task dependent, understanding of images. This evidence, together with the results in this dissertation, give faith to the hypothesis that statis-

tical measures of information (such as the mutual information) could form a measure of the semantic information in an image.

In this section, I derive the mutual information between the test and the reference signals in the wavelet domain. I first derive the mutual information for one subband and later generalize for multiple subbands.

Let $C^N = (C_1, C_2, \dots, C_N)$ denote N elements from \mathcal{C} . In this section I will assume that the underlying RF \mathcal{U} is uncorrelated (and hence \mathcal{C} is an RF with conditionally independent elements given \mathcal{S}), and that the distortion model parameters \mathcal{G} and σ_V^2 are known *a priori*. Let $D^N = (D_1, D_2, \dots, D_N)$ denote the *corresponding* N elements from \mathcal{D} . The mutual information between these is denoted $I(C^N; D^N)$, and it quantifies the information in bits that is shared between the corresponding N coefficients in the two RFs.

Due to the availability of the reference image, it makes sense to tune the source model specifically for the reference image. I propose to do this by considering *conditional* mutual information. Incidentally, due the non-linear dependance among the C^N by way of \mathcal{S} , it is much easier to analyze the mutual information assuming \mathcal{S} is known. This is also in line with some divisive normalization based HVS models, which process the input signals after ‘conditioning’ them with division by local standard deviation estimates [103, 125]. Thus the information criterion that I propose for FR QA in this chapter is the conditional mutual information $I(C^N; D^N | S^N = s^N)$, where $S^N = (S_1, S_2, \dots, S_N)$ are the corresponding N elements of \mathcal{S} , and s^N denotes a *realization* of S^N . I will denote $I(C^N; D^N | S^N = s^N)$ as $I(C^N; D^N | s^N)$

throughout the rest of this dissertation. With the stated assumptions on \mathcal{C} and the distortion model (4.3), one can show:

$$I(C^N; D^N | s^N) = \sum_{i=1}^N I(C_i; D^N | C^{i-1}, s^N) \quad (4.4)$$

$$= \sum_{j=1}^N \sum_{i=1}^N I(C_i; D_j | C^{i-1}, D^{j-1}, s^N) \quad (4.5)$$

$$= \sum_{i=1}^N I(C_i; D_i | C^{i-1}, D^{i-1}, s^N) \quad (4.6)$$

$$= \sum_{i=1}^N I(C_i; D_i | s_i) \quad (4.7)$$

where I get (4.4) and (4.5) by the chain rule [15], and (4.6) and (4.7) by conditional independence of \mathcal{C} given \mathcal{S} , independence of the noise \mathcal{V} , the fact that the distortion model keeps D_i independent of C_j , $\forall j \neq i$, and that given S_i , C_i and D_i are independent of S_j $\forall j \neq i$. Using the fact that C_i are Gaussian given $S_i = s_i$ with variance $s_i^2 \sigma_U^2$, and V_i are also Gaussian with variance σ_V^2 , we get:

$$I(C^N; D^N | s^N) = \sum_{i=1}^N I(C_i; D_i | s_i) \quad (4.8)$$

$$= \sum_{i=1}^N (h(D_i | s_i) - h(D_i | C_i, s_i)) \quad (4.9)$$

$$= \sum_{i=1}^N (h(g_i C_i + V_i | s_i) - h(V_i)) \quad (4.10)$$

$$= \frac{1}{2} \sum_{i=1}^N \log_2 \left(1 + \frac{g_i^2 s_i^2 \sigma_U^2}{\sigma_V^2} \right) \quad (4.11)$$

where $h(X)$ denotes the differential entropy of a continuous random variable X , and for X distributed as $\mathcal{N}(\mu, \sigma^2)$, $h(X) = 1/2 \log_2 2\pi e \sigma^2$ [15].

Equation (4.11) was derived for one subband. It is straightforward to use separate GSM RF's for modeling each subband of interest in the image. I will denote the RF modeling the wavelet coefficients of the reference image in the k -th subband as \mathcal{C}^k , and in test (distorted) image as \mathcal{D}^k , and assume that \mathcal{C}^k are independent of each other. I will further assume that each subband is distorted independently. Thus, the RF's \mathcal{V}^k are also independent of each other. The information fidelity criterion (IFC) is then obtained by summing over all subbands:

$$\text{IFC} = \sum_{k \in \text{subbands}} I(C^{N_k, k}, D^{N_k, k} | s^{N_k, k}) \quad (4.12)$$

where $C^{N_k, k}$ denotes a vector of N_k coefficients from the RF \mathcal{C}^k of the k -th subband, and similarly for $D^{N_k, k}$ and $s^{N_k, k}$.

Equation (4.12) is the proposed information fidelity criterion that quantifies the Shannon information that is shared between the source and the distorted images. An attractive feature of this criterion is that like MSE and some other mathematical fidelity metrics, it does not involve parameters associated with display device physics, data from visual psychology experiments, viewing configuration information or stabilizing constants that dictate the accuracy of HVS based FR QA methods (and some structural ones too). The IFC does not require training data either. However some implementation parameters

will obviously arise once (4.12) is implemented. I will discuss implementation in the next section.

The IFC is not a distortion metric, but a fidelity criterion. It theoretically ranges from zero (no fidelity) to infinity (perfect fidelity within a non-zero multiplicative constant in the absence of noise²). Perfect fidelity within a multiplicative constant is something that is in contrast with the approach in SSIM [131], in which contrast distortion (multiplicative constant) was one of the three attributes of distortion that was regarded as a visual degradation, albeit one that has a different (and ‘orthogonal’) contribution towards perceptual fidelity than noise and local-luminance distortions. The IFC views multiplicative constants (contrast stretches) as signal gains or attenuations *interacting* with additive noise. Thus, with this approach, the same noise variance would be perceptually less annoying if it were added to a contrast stretched image than if it were added to a contrast attenuated image. Since each subband has its own multiplicative constant, blur distortion could also be captured by this model as the finer scale subbands would be attenuated more than coarser scale subbands.

²Differential entropy is invariant to translation, and so the IFC is infinite for perfect fidelity within additive constant in the absence of noise as well. However since we are applying the IFC in the wavelet domain, the zero-mean assumptions on \mathcal{U} and \mathcal{V} imply that this case will not happen.

4.3 Implementation Issues

In order to implement the fidelity criterion in (4.12) a number of assumptions are required about the source and the distortion models. I outline them in this section.

4.3.1 Assumptions about the Source Model

Note that mutual information (and hence the IFC) can only be calculated between RF's and not their *realizations*, that is, a particular reference and the test image under consideration. I will assume ergodicity of the RF's and that reasonable estimates for the statistics of the RF's can be obtained from their realizations. I then quantify the mutual information between the RF's having statistics obtained from particular realizations.

For the scalar GSM model, estimates of s_i^2 can be obtained by localized sample variance estimation, since for natural images \mathcal{S} is known to be a highly spatially correlated field, and σ_U^2 can be assumed to be unity without loss of generality.

4.3.2 Assumptions about the Distortion Model

The IFC assumes that the distortion model parameters \mathcal{G} and σ_V^2 are known *a priori*, but these would need to be estimated in practice. I propose to partition the subbands into blocks and assume that the field \mathcal{G} is constant over such blocks, as are the noise statistics σ_V^2 . Localized treatment (by 'blocking') of image distortions is necessitated by their non-stationary nature. The value

of the field \mathcal{G} over block l , which I denote as g_l , and the variance of the RF \mathcal{V} over block l , which I denote as $\sigma_{V,l}^2$, are fairly easy to estimate (by linear regression) since both the input (the reference signal) as well as the output (the test signal) of the system (4.3) are available [37]:

$$\hat{g}_l = \widehat{\text{Cov}}(C, D) \widehat{\text{Cov}}(C, C)^{-1} \quad (4.13)$$

$$\hat{\sigma}_{V,l}^2 = \widehat{\text{Cov}}(D, D) - g_l \widehat{\text{Cov}}(C, D) \quad (4.14)$$

where the covariances are approximated by sample estimates using sample points from the corresponding blocks in the reference and test signals.

4.3.3 Wavelet Bases and Inter-coefficient Correlations

The derivation leading to (4.11) assumes that \mathcal{U} is uncorrelated, and hence \mathcal{C} is independent given \mathcal{S} . In practice, if the wavelet decomposition is orthogonal, the underlying \mathcal{U} could be approximately uncorrelated. In such cases, one can use (4.11) for computing the IFC. However real cartesian-separable orthogonal wavelets are not good for image analysis since they have poor orientation selectivity. In my implementation, I chose the steerable pyramid decomposition with six orientations [98]. This gives better orientation selectivity than possible with real cartesian separable wavelets. However the steerable pyramid decomposition is over-complete, and the neighboring coefficients \mathcal{C} from the same subband are linearly correlated. In order to deal with such correlated coefficients, I propose two simple approximations that work well for quality assessment purposes.

4.3.3.1 Vector GSM

The first approximation is to partition the RF into non-overlapping block-neighborhoods and then assume that the neighborhoods are uncorrelated with each other. One could then use a vector form of the IFC by modeling each neighborhood as a vector random variable. This ‘blocking’ of coefficients results in an upper bound on the IFC:

$$\begin{aligned} I(C^N; D^N | s^N) &\leq \sum_{j=1}^{N/M} I(\vec{C}_j; \vec{D}_j | s^N) \\ &\leq \sum_{j=1}^{N/M} I(\vec{C}_j; \vec{D}_j | s_j) \end{aligned}$$

where $\vec{C}_j = (C_{j,i}, i = 1 \dots M)$ is a vector of M wavelet coefficients that form the j -th neighborhood. All such vectors, associated with non-overlapping neighborhoods, are assumed to be uncorrelated with each other. I then model the wavelet coefficient neighborhood as a vector GSM. Thus, the vector RF $\mathcal{C} = \{\vec{C}_i : i \in I'\}$ on a lattice I' is a product of a *scalar* RF \mathcal{S} and a zero-mean Gaussian *vector* RF $\mathcal{U} = \{\vec{U}_i : i \in I'\}$ of covariance $\mathbf{C}_{\vec{U}}$. The noise \mathcal{V} is also a zero-mean vector Gaussian RF of the same dimensionality as \mathcal{C} , and has covariance $\mathbf{C}_{\vec{V}}$. If we assume that \vec{U}_i is independent of \vec{U}_j , $\forall i \neq j$, it is quite easy to show (by using differential entropy for multivariate Gaussian vectors) that:

$$I(C^N; D^N | s^N) \leq \sum_{j=1}^{N/M} I(\vec{C}_j; \vec{D}_j | s_j) \quad (4.15)$$

$$= \frac{1}{2} \sum_{j=1}^{N/M} \log_2 \left(\frac{|g_j^2 s_j^2 \mathbf{C}_{\vec{U}} + \mathbf{C}_{\vec{V}}|}{|\mathbf{C}_{\vec{V}}|} \right) \quad (4.16)$$

where the differential entropy of a continuous vector random vector \vec{X} distributed as a multivariate Gaussian $\mathcal{N}(\vec{\mu}, \Sigma)$, $h(\vec{X}) = 1/2 \log_2 (2\pi e)^d |\Sigma|$ where $|\cdot|$ denotes the determinant and d is the dimension of \vec{X} [15]. Recalling that $\mathbf{C}_{\vec{U}}$ is symmetric and can be factorized as $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ with orthonormal \mathbf{Q} (that is $\mathbf{Q}\mathbf{Q}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, and $|\mathbf{Q}| = |\mathbf{Q}^T| = 1$) and eigenvalues λ_k [105], and that for a distortion model where $\mathbf{C}_{\vec{V}} = \sigma_V^2 \mathbf{I}$, the IFC simplifies as follows³:

$$I(C^N; D^N | s^N) \leq \sum_{j=1}^{N/M} I(\vec{C}_j; \vec{D}_j | s_j) \quad (4.17)$$

$$= \frac{1}{2} \sum_{j=1}^{N/M} \log_2 \left(\frac{|g_j^2 s_j^2 \mathbf{C}_{\vec{U}} + \sigma_V^2 \mathbf{I}|}{|\sigma_V^2 \mathbf{I}|} \right) \quad (4.18)$$

$$= \frac{1}{2} \sum_{j=1}^{N/M} \log_2 \left(\frac{|g_j^2 s_j^2 \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T + \sigma_V^2 \mathbf{I}|}{\sigma_V^{2M}} \right) \quad (4.19)$$

$$= \frac{1}{2} \sum_{j=1}^{N/M} \log_2 \left(\frac{|\mathbf{Q}| |g_j^2 s_j^2 \mathbf{\Lambda} + \sigma_V^2 \mathbf{I}| |\mathbf{Q}^T|}{\sigma_V^{2M}} \right) \quad (4.20)$$

$$= \frac{1}{2} \sum_{j=1}^{N/M} \log_2 \left(\frac{|g_j^2 s_j^2 \mathbf{\Lambda} + \sigma_V^2 \mathbf{I}|}{\sigma_V^{2M}} \right) \quad (4.21)$$

$$= \frac{1}{2} \sum_{j=1}^{N/M} \sum_{k=1}^M \log_2 \left(1 + \frac{g_j^2 s_j^2 \lambda_k}{\sigma_V^2} \right) \quad (4.22)$$

where the numerator term inside the logarithm of (4.21) is the determinant of a diagonal matrix and hence equals the product of the diagonal terms. The bound in (4.17) shrinks as M increases. In my simulations I use vectors from 3×3 spatial neighborhoods and achieve good performance. Equation (4.22)

³Utilizing the structure of $\mathbf{C}_{\vec{U}}$ and $\mathbf{C}_{\vec{V}}$ helps in faster implementations through matrix factorizations.

is the form that is used for implementation.

For the vector GSM model, the maximum-likelihood estimate of s_j^2 can be found as follows [106, 122]:

$$\hat{s}_j^2 = \frac{\vec{C}_j^T \mathbf{C}_{\mathbf{u}}^{-1} \vec{C}_j}{M} \quad (4.23)$$

where M is the dimensionality of \vec{C}_j . Estimation of the covariance matrix $\mathbf{C}_{\vec{U}}$ is also straightforward from the reference image wavelet coefficients [106]:

$$\hat{\mathbf{C}}_{\vec{U}} = \frac{M}{N} \sum_{j=1}^{N/M} \vec{C}_j \vec{C}_j^T \quad (4.24)$$

In (4.23) and (4.24), $E[S_i^2]$ is assumed to be unity without loss of generality [106, 122].

4.3.3.2 Downsampling

The second approximation to obtain uncorrelated GSM is to use a subset of the coefficients obtained by *downsampling* \mathcal{C} . Downsampling reduces the correlation between coefficients. I assume that the downsampled subband is approximately uncorrelated, and then use (4.11) for scalar GSM on the downsampled subband. The underlying assumption in the downsampling approach is that the quality prediction from the downsampled subbands should be approximately the same as the prediction from the complete subband. This downsampling approach has an additional advantage that it makes it possible to substantially reduce the complexity of computing the wavelet decomposition, since only a fraction of the subband coefficients need to be computed. In

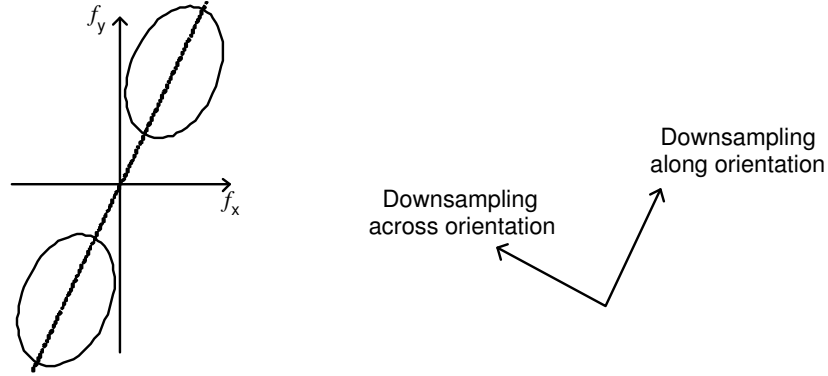


Figure 4.2: Downsampling a subband for reducing correlation between coefficients. The orientation selective filters in the steerable pyramid decomposition select image frequencies along the principal orientation (left). Downsampling along and across the principal orientation is a good way of reducing correlations in the filtered signal (right). In my simulations, I downsample by 3 along orientation and by 5 across orientation.

my simulations I discovered that the wavelet decomposition is the most computationally expensive step. Significant speedups are possible with the typical downsampling factors of twelve or fifteen in my simulations. Since the filters in the analysis filter-bank are orientation selective, a subband is downsampled along and across the principal orientations of the respective filters. This is shown in Figure 4.2. In my simulations, the downsampling was done using nearest-neighbor interpolation.

Further specifics of the estimation methods used in my testing are given in Section 4.5.

4.4 Subjective QA Study for Validation

In order to calibrate and test the algorithm, an extensive subjective QA study was conducted. In these experiments, a number of human subjects were asked to assign each image with a score indicating their assessment of the quality of that image, defined as the extent to which the artifacts were visible and annoying. The details of the study are given in Appendix A. A total of 982 images, out of which 203 were the reference images, were evaluated by human subjects, and the raw scores were processed to give Mean Opinion scores (MOS) and Difference Mean Opinion Scores (DMOS) for all distorted image.

In order to compare one quality assessment algorithm against another the subjective database on which they are evaluated needs to be the same. Most researchers report the performance of their methods on their own small-scale studies that sometimes do not provide a diverse enough set of distortion types, image content or quality range. It therefore becomes difficult to compare methods across databases. One of my goals was to provide a larger data sets for QA researchers. It is my hope that the freely available database would be used by other researchers to compare and contrast quality assessment methods [93].

4.5 Results

In this section I present results on validation of the IFC on the database presented in Section 4.4, and comparisons with other quality assessment algo-

rithms. Specifically, I will compare the performance of the proposed algorithm against PSNR, SSIM [131], and the well known Sarnoff model (Sarnoff JND-Metrix 8.0 [84]). I present results for different versions of the IFC: scalar GSM, scalar GSM with downsampling by three along the principal orientation and five across, vector GSM, and vector GSM using the horizontal and vertical orientations only, vector GSM using horizontal and vertical orientations and only one eigenvalue in the summation of (4.22). Table 4.1 and 4.2 summarize the results for the quality assessment methods for MOS and DMOS respectively. Although the tables lead to identical conclusions regarding the performance of different methods, some readers may prefer comparisons with MOS instead of DMOS.

4.5.1 Simulation Details

Some additional simulation details are as follows. Although full color images were distorted in the subjective evaluation, the QA algorithms (except Sarnoff's) operated upon the luminance component only. For the scalar GSM with no downsampling, a 5×5 moving window was used for local variance estimation (\hat{s}_i^2), and 16×16 non-overlapping blocks were used for estimating parameters g_l and $\sigma_{V,l}^2$. The blocking was done in order for the stationarity assumptions on the distortion model to approximately hold. For the scalar GSM with downsampling, all parameters were estimated on the downsampled signals. A 3×3 window was used for variance estimation, while 8×8 blocks were used for the distortion model estimation. For vector GSM, vectors were

constructed from non-overlapping 3×3 neighborhoods, and the distortion model was estimated with 18×18 non-overlapping blocks. In all versions of the IFC, only the subbands at the finest level were used in the summation of (4.12). Since the sizes of the images in the database were different, the IFC was normalized by the number of pixels in each image. MSSIM (Mean SSIM) was calculated on the luminance component after decimating (filtering and downsampling) it by a factor of 4 [131].

4.5.2 Calibration of the Objective Score

It is generally acceptable for a QA method to stably predict subjective quality within a non-linear mapping, since the mapping can be compensated for easily. Moreover, since the mapping is likely to depend upon the subjective validation/application scope and methodology, it is best to leave it to the final application, and not to make it part of the QA algorithm. Thus in both the VQEG Phase-I and Phase-II testing and validation, a non-linear mapping between the objective and the subjective scores was allowed, and all the performance validation metrics were computed *after* compensating for it [120, 121]. This is true for the results in tables 4.1 and 4.2, where a five-parameter non-linearity (a logistic function with additive linear term) is used for all methods except IFC, for which I used the mapping on the logarithm of IFC. The fitting of the logistic curve to some of the methods tested is shown in Figure 4.3, while the quality predictions after compensating for the mapping are shown in Figure 4.4. The mapping function used is given in (4.25), while

Validation against MOS					
Model	CC	MAE	RMS	OR	SROCC
PSNR	0.828	7.244	8.999	0.217	0.821
Sarnoff	0.899	5.305	7.007	0.086	0.899
MSSIM	0.912	4.977	6.592	0.076	0.910
IFC (no ds)	0.907	5.129	6.753	0.098	0.903
IFC (ds 3/5)	0.909	5.030	6.667	0.096	0.904
IFC (vec)	0.913	4.955	6.527	0.091	0.911
IFC (h/v, vec)	0.915	4.897	6.467	0.089	0.913
IFC (h/v, 1 ev)	0.928	4.538	5.980	0.058	0.925

Table 4.1: Validation scores for different quality assessment methods. The methods tested were PSNR, Sarnoff JND-Metrix 8.0 [84], MSSIM [131], IFC for scalar GSM without downsampling, IFC for scalar GSM with downsampling by 3 along orientation and 5 across, IFC for vector GSM, and IFC for vector GSM using horizontal and vertical orientations only, IFC for vector GSM and horizontal/vertical orientations with only the smallest eigenvalue in (4.22). The methods were tested against MOS from the subjective study after a non-linear mapping. The validation criteria are: correlation coefficient (CC), mean absolute error (MAE), root mean squared error (RMS), outlier ratio (OR) and spearman rank-order correlation coefficient (SROCC).

the fitting was done using MATLAB’s *fminsearch*.

$$\text{Quality}(x) = \beta_1 \text{logistic}(\beta_2, (x - \beta_3)) + \beta_4 x + \beta_5 \quad (4.25)$$

$$\text{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1 + \exp(\tau x)} \quad (4.26)$$

4.5.3 Discussion

Tables 4.1 and 4.2 show that the IFC, even in its simplest form, is competitive with all state-of-the-art FR QA methods presented in this chapter. The comparative results between MSSIM and Sarnoff’s JND-Metrix are

Validation against DMOS					
Model	CC	MAE	RMS	OR	SROCC
PSNR	0.826	7.272	9.087	0.114	0.820
Sarnoff	0.901	5.252	6.992	0.046	0.902
MSSIM	0.912	4.979	6.616	0.035	0.910
IFC (no ds)	0.911	5.078	6.652	0.041	0.908
IFC (ds 3/5)	0.913	5.009	6.587	0.041	0.909
IFC (vec)	0.917	4.919	6.437	0.039	0.915
IFC (h/v, vec)	0.919	4.855	6.366	0.032	0.918
IFC (h/v, 1 ev)	0.929	4.523	5.941	0.059	0.928

Table 4.2: Validation scores for different quality assessment methods. The methods tested were PSNR, Sarnoff JND-Metrix 8.0 [84], MSSIM [131], IFC for scalar GSM without downsampling, IFC for scalar GSM with downsampling by 3 along orientation and 5 across, IFC for vector GSM, and IFC for vector GSM using horizontal and vertical orientations only, IFC for vector GSM and horizontal/vertical orientations with only the smallest eigenvalue in (4.22). The methods were tested against DMOS from the subjective study after a non-linear mapping. The validation criteria are: correlation coefficient (CC), mean absolute error (MAE), root mean squared error (RMS), outlier ratio (OR) and spearman rank-order correlation coefficient (SROCC).

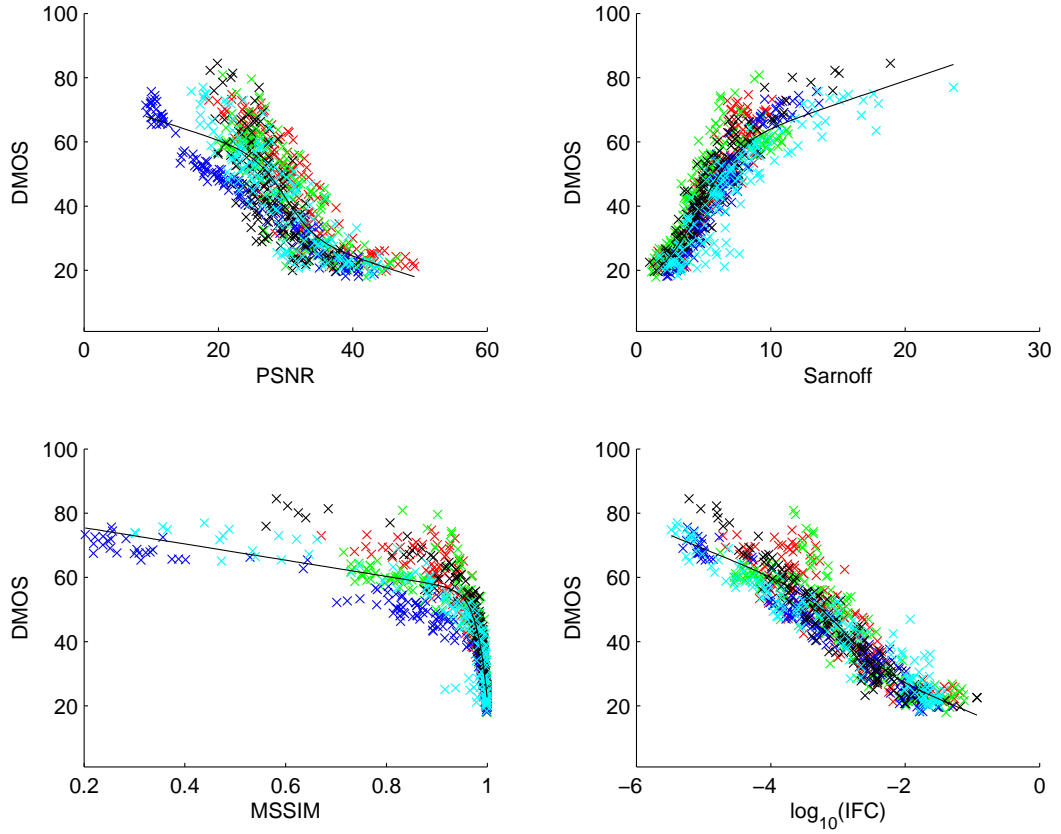


Figure 4.3: Scatter plots for the four objective quality criteria: PSNR, Sarnoff's JND-Metrix, MSSIM, and $\log(\text{IFC})$ for vector GSM using horizontal/vertical orientations. The IFC shown here uses only the horizontal and vertical subbands at the finest scale, and only the smallest eigenvalue in (4.22). The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).

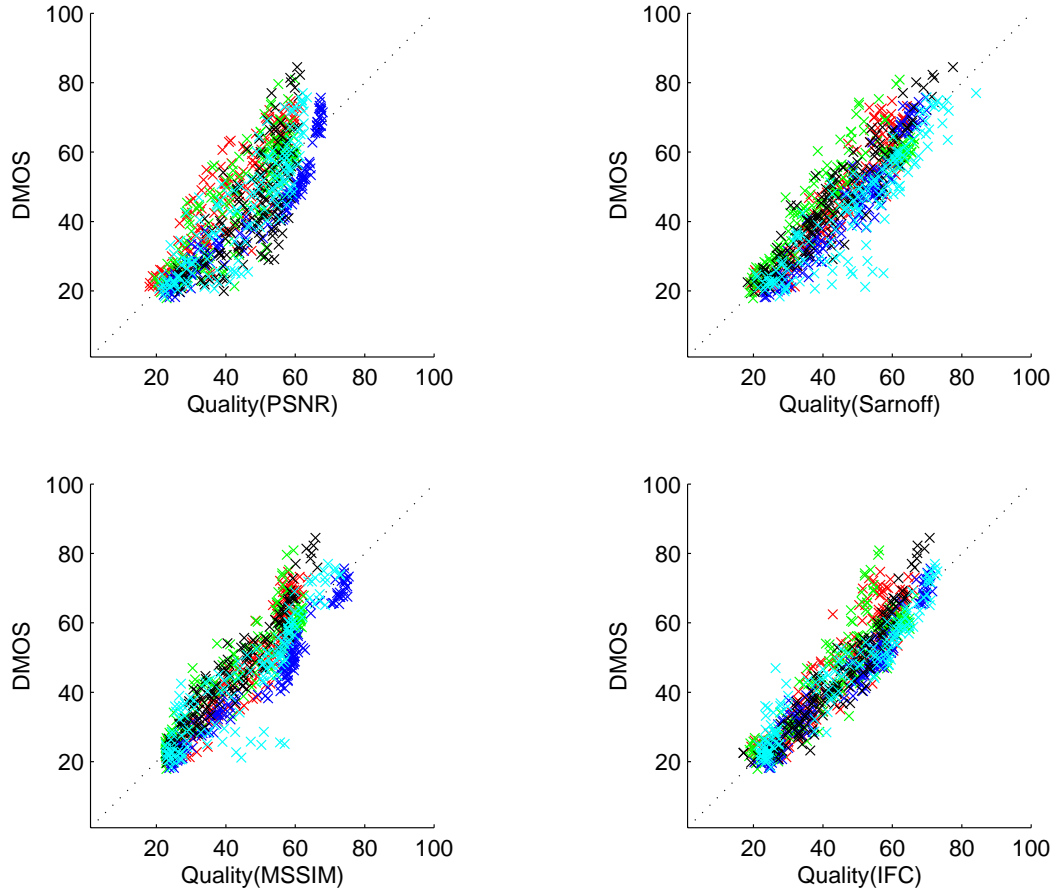


Figure 4.4: Scatter plots for the quality predictions by the four methods after compensating for quality calibration: PSNR, Sarnoff's JND-Metrix, MSSIM, and IFC for vector GSM using horizontal/vertical orientations. The IFC shown here uses only the horizontal and vertical subbands at the finest scale, and only the smallest eigenvalue in (4.22). The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).

qualitatively similar to those reported in [131], only that both of these methods perform poorer in the presence of a wider range of distortion processes than reported in [131]. However, MSSIM still outperforms JND-Metrix by a sizeable margin using any of the validation criteria in Tables 4.1 and 4.2.

The IFC also performs demonstrably better than Sarnoff’s JND-Metrix under all of the alternative implementations of IFC. The vector-GSM form of IFC outperforms even MSSIM. Note that the downsampling approximation performs better than scalar IFC without downsampling, even though the downsampled version operates on signals that are fifteen times smaller, and hence it is a computationally more feasible alternative to other IFC implementations at a reasonably good performance. Also note that IFC as well as MSSIM use only the luminance components of the images to make quality predictions, whereas the JND-Metrix uses all color information. Extending the IFC to incorporate color could further improve performance.

An interesting observation is that when only the smaller eigenvalues are used in the summation of (4.22), the performance increases dramatically. The last rows in Table 4.1 and 4.2, and Figures 4.3 and 4.4 show results when only the smallest eigenvalue is used in the summation in (4.22). The performance is relatively unaffected up to an inclusion of six smallest eigenvalues (out of a total of nine). One hypothesis that could explain this observation is that a systematic measurement noise could be present in the IFC whose strength depends upon the strength of the signal used in the computation of IFC. Thus, ignoring components with high signal strength (corresponding to summing over

low eigenvalues only in (4.22)) could lower the noise in IFC if the relationship between the measurement noise variance and the signal variance is super-linear. Thus, an increase in signal strength would cause a *decrease* in the signal-to-noise ratio in such a case.

Another interesting observation is that when only the horizontal and vertical subbands are used in the computation of the IFC in (4.12) for the vector GSM IFC, the performance increases remarkably⁴. I first thought that this was due to the presence of JPEG distorted images in the database since the blocking artifact is represented more in the horizontal and vertical subbands than at other orientations. However, I discovered that the performance increase was consistent for *all* distortion types present in the database, and most notably for the JPEG2000 distortion. Also I do not get this increase in performance when I sum over other subbands; the performance in fact worsens. Table 4.3 gives the performance change of the IFC for horizontal and vertical subbands and the corresponding performance change when orientations of ± 60 degrees were summed in (4.12). I feel that this performance increase is due to the importance that the HVS gives to horizontal and vertical edge information in images in comparison with other orientations. I will discuss similarities between the IFC and the HVS later in Section 4.6.

I would like to point out the most salient feature of the IFC: it does not require any parameters from the HVS or viewing configuration, training data

⁴It does so for other IFC forms but I will not report those results here since they are mirrored by the ones presented.

RMS in prediction against DMOS			
Distortion	All orientations	Hor./Vert.	± 60 deg.
JPEG2000	6.899	6.017	7.565
JPEG	6.542	6.237	6.927
White Noise	3.589	3.444	3.698
Gauss. Blur	4.166	3.873	4.521
Fast-fading	4.448	4.416	4.661

Table 4.3: Validation scores for the vector GSM IFC using all orientations versus using: only the horizontal and vertical orientations, and the subbands oriented at ± 60 deg. Only the smallest eigenvalue has been used in (4.22) for generating this table.

or stabilizing constants. In contrast, the JND-Metrix requires a number of parameters for calibration such as viewing distance, display resolution, screen phosphor type, ambient lighting conditions etc. [84], and even SSIM requires two hand-optimized stabilizing constants. Despite being parameterless, the IFC outperforms both of these methods. It is reasonable to say that the performance of the IFC could improve further if these parameters, which are known to affect perceptual quality to some degree, were incorporated into the IFC as well.

4.6 Similarities with HVS Error-Sensitivity Based QA Methods

I will now compare and contrast IFC with HVS error sensitivity measures. Figure 4.5 shows an HVS error sensitivity based quality measurement system that computes the error signal between the processed reference and

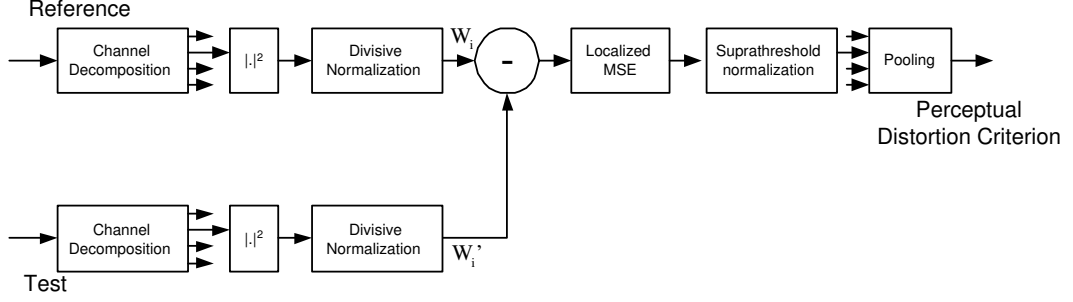


Figure 4.5: An HVS error-sensitivity based quality measurement system. I show that this HVS model is the dual of the IFC under the assumptions regarding source and distortion models.

test signals, and then processes the error signal before computing the final perceptual distortion measure. A number of key similarities with most HVS QA methods are immediately evident. These include a scale-space-orientation channel decomposition, response exponent, masking effect modeling, localized error pooling, suprathreshold effect modeling and a final pooling into a quality score.

In Appendix B show the following relationship between the scalar version of IFC in (4.11) and the HVS model of Figure 4.5 for one subband:

$$I(C^N; D^N | s^N) \approx \alpha \sum_{i=1}^N \log_2(\text{MSE}(W_i, W'_i | s_i)) + \beta \quad (4.27)$$

where W_i and W'_i are as shown in Figure 4.5. The MSE computation in Figure 4.5 and (4.27) is a *localized* error strength measure. The logarithm term can be considered to be modeling of the suprathreshold effect. Suprathreshold effect is the name given to the fact that the same amount of distortion becomes perceptually less significant as the overall distortion level increases. Thus a

change in MSE of, say, 1.0 to 2.0 would be more annoying than the same change from 10.0 to 11.0. Researchers have previously modeled suprathreshold effects using visual impairment scales that map error strength measures through concave non-linearities, qualitatively similar to the logarithm mapping, so that they emphasize the error at higher quality [140]. Also, the pooling in (4.27) can be seen to be Minkowski pooling with exponent 1.0. Hence with the stated components, the IFC can be considered to be a particular HVS error-sensitivity based quality assessment algorithm, the Perceptual Distortion Criterion (PDC), within multiplicative and additive constants that could be absorbed into the calibration curve:

$$\text{PDC} = \sum_{k \in \text{subbands}} \sum_{i=1}^{N_k} \log_2(\text{MSE}(W_{k,i}, W'_{k,i} | s_{k,i})) \quad (4.28)$$

$$\text{IFC} \approx \alpha(\text{PDC}) + N_{\text{sub}}\beta \quad (4.29)$$

where k denotes the index of the k -th subband, and N_{sub} is the number of subbands used in the computation.

One could make the following observations regarding the perceptual distortion criterion (PDC) of (4.28), which is the HVS dual of the IFC (using the scalar GSM model), in comparison with other HVS error sensitivity based FR QA methods:

- Some components of HVS are not modeled in Figure 4.5 and (4.29), such as the optical point spread function and the contrast sensitivity function.
- The exponent of the non-linearity is 2. Different researchers choose different values (2 or 3) for the response exponent [144].

- The masking effect is modeled differently from some HVS based methods. While the divisive normalization mechanism for masking effect modeling has been employed by some QA methods [32, 70, 113], most methods divisively normalize the *error* signal with visibility thresholds that are dependent on the neighborhood signal strength.
- Minkowski error pooling occurs in two stages: first a localized pooling in the computation of the localized MSE (with exponent 2) and then a global pooling with an exponent of unity after the suprathreshold modeling. Thus the perceptual error calculation is different from most methods, in that it happens in two stages with suprathreshold effects in between.
- In (4.28), the non-linearity that maps the MSE to a suprathreshold-MSE is a logarithmic non-linearity and it maps the MSE to a suprathreshold distortion that is later pooled into a quality score. Watson et al. have used threshold power functions to map objective distortion into *subjective* JND by use of two-alternative forced choice experiments [140]. However, their method applies the suprathreshold non-linearity *after* pooling, as if the suprathreshold effect only comes into play at the global quality judgement level. The formulation in (4.28) suggests that the suprathreshold modeling should come *before* a global pooling stage but after localized pooling, and that it affects judgement of quality at a *local* level.
- One significant difference is that the IFC using the scalar GSM model, or

the perceptual distortion criterion of (4.28), which are dual of each other, is notably inferior to the vector GSM based IFC. I believe that this is primarily due to the underlying assumption about the uncorrelated nature of the wavelet coefficients being inaccurate. This dependence of perceptual quality on the correlation among coefficients is hard to investigate or model using the HVS error-sensitivity paradigm, but the task is greatly simplified by approaching the same problem with NSS modeling. Thus I feel that HVS based QA methods need some modifications to incorporate the fact that natural scenes are correlated within subbands, and that this inter-coefficient correlation affects human perception of quality⁵.

- Another significant difference between IFC/PDC, and other HVS based methods is distinct modeling of signal attenuation. Other HVS based methods ignore signal gains and attenuations, constraining $g = 1$, whereas a generalized attenuation value g in IFC/PDC allows for signal attenuations to be handled differently from the additive noise components.
- One could conjecture that the conditioning on \mathcal{S} in the IFC is paralleled in the HVS by the computation of the local variance and divisive normalization. Note that the high degree of self-correlation present in \mathcal{S} enables its adequate estimation from \mathcal{C} by local variance estimation.

⁵Equation (4.22) suggests that the same noise variance would cause a greater loss of information fidelity if the wavelet coefficients of the reference image were correlated than if they were uncorrelated.

Since this divisive normalization occurs quite early in the HVS model⁶ and since the visual signal is passed to the rest of the HVS after it has been *conditioned* by divisive normalization by the estimated s_i^2 , one could hypothesize that the rest of the HVS analyzes the visual signal *conditioned on the prior knowledge of \mathcal{S}* , just as the IFC analyzes the mutual information between the test and the reference conditioned on the prior knowledge of \mathcal{S} .

- One question that should arise when one compares the IFC against the HVS error model is regarding HVS model parameters. Specifically, one should notice that while functionally the IFC captures HVS sensitivities, it does so without using actual HVS model parameters. I believe that some of the HVS model parameters were either incorporated into the calibration curve, or they did not affect performance significantly enough under the testing and validation experiments reported in this chapter. Parameters such as the characteristics of the display devices or viewing configuration information could easily be understood to have approximately similar affect on all images for all subjects since the experimental conditions were approximately the same. Other parameters and model components, such as the optical point spread function or the contrast sensitivity function, which depend on viewing configuration parameters as well, are perhaps less significant for the scope and range of

⁶Divisive normalization has been discovered to be operational in the HVS [102, 103].

quality of my validation experiments. It is also reasonable to say that incorporating these parameters could further enhance the performance of IFC. In the next chapter, I present an extension of IFC using a unified model that consists of source, distortion and HVS models, and this extension indeed gives improvements in the performance of the algorithm.

- I would like to emphasize at this point that although the IFC can be shown to be similar to an HVS error sensitivity based distortion measure, it has *not* been derived using any HVS knowledge, and its derivation is completely independent. The similarities exist due to the similarities between NSS and HVS models. The difference is subtle, but profound!
- The relationship (4.29) between the IFC of (4.11) and (4.12) and the HVS based perceptual distortion criterion of (4.28) is based on assumptions regarding the source and the distortion models. It may not hold for other sources or distortions and the IFC and the perceptual distortion criterion of (4.28) may be quantifying different aspects of distortion and fidelity.

In this section I have demonstrated the dual nature of the visual quality assessment problem. It could be approached from an HVS based error-sensitivity paradigm as well as from an NSS based information fidelity paradigm. Even in its rudimentary form, the IFC outperforms the current state of the art QA methods under the testing conditions. And despite all the similarities that the IFC has with HVS based methods, the IFC is parameterless: it does

not need training data, psychophysical experiment data, viewing configuration measurements or stabilizing constants.

4.7 Conclusions

In this chapter I presented an information fidelity criterion for image quality assessment using natural scene statistics. I showed that using signal source and distortion models, one could quantify the Shannon information between the reference and the test images, and that this quantification, the information fidelity criterion, quantifies perceptual quality. The IFC was demonstrated to be better than a state-of-the-art HVS based method, the Sarnoff's JND-Metrix, as well as a state-of-the-art structural fidelity criterion, the structural similarity (SSIM) index in my testing. I showed that despite its competitive performance, the IFC is parameterless. I also showed that the IFC, under certain conditions, is quantitatively similar to an HVS error-sensitivity based QA method, and I compared and contrasted the two approaches and hypothesized directions in which HVS based methods could be refined and improved.

In the next chapter, I take the notion of information fidelity one step further and propose a novel visual information fidelity measure. This new measure not only performs better than the IFC, but it also has interesting properties that, to the best of my knowledge, are not present in any other QA method.

Chapter 5

Image Information and Visual Quality

Scattered all over the city are
worn out and disfigured dreams
of which the city dwellers are oblivious!
I roam around the city day and night to collect them
Heat them in the furnace of my heart
so that the old rust on them comes off
Their limbs come out nice and clean
Their lips, cheeks and heads start shimmering
Like the desires of freshly dressed bridegrooms.
So that once again these dreams may find a direction!

"Dreams for sale, dreams ..."

As the morning dawns I go calling out in the streets....

"Are these dreams real or fake?"

They check them out as if there isn't anyone more adept at judging them!

A dream maker I'm not either, just a face-lifter....

But yes dreams are the source of my livelihood!

Evening settles in and I call out again....

"Free everyone, free, these dreams of gold..."

Hearing "free", people get even more frightened
and slip away lip-tightened...

"Well he says they are 'free'

"Could it be a sham?

"Some hidden deception?

"They may break on reaching home, or just melt away, these dreams?

"Disappear with a pop somehow, or cast upon us some spell, these
dreams?

"No sir, what use could they be?

"Dreams of this hawker?

"Junk dreams of this blind hawker!!"

Night sets in

carrying heaps of dreams over my head, I reach home disappointed

Mumble all night again "Take these dreams...

"and take from me their price as well

"Take these dreams, dreams... my dreams ...

"dreams my dreams dr..ee..eaams ...

"their pr...iii...ceee as welllll....."

— "*The blind hawker*" by Noon Meem Rashid

5.1 Introduction

In the previous chapter, I presented an information fidelity criterion for image quality assessment. In this FR QA method, the quality assessment problem is treated as an information fidelity problem in which a source sends data to a receiver through a channel that limits the amount of information that could flow through it. The source is a stochastic natural image source whose output is the reference image, the channel is the image distortion operator, and the receiver is a human observer. The output of the channel is the test image, and mutual information between the input and the output of the channel quantifies how much information about the reference image is present in the test image. In contrast to the HVS error-sensitivity and the structural approaches, the statistical approach, used in an information-theoretic setting, yielded an FR QA method that did not rely on any HVS or viewing geometry parameters, nor any constants requiring optimization.

In this chapter, I further explore the connections between image information and visual quality¹. Specifically, I will model the reference image as being the output of a stochastic natural source that passes through the HVS channel and is processed later by the brain. Thus I consider the HVS to be a channel as well that imposes limits on how much information could pass through it to the brain (receiver). I quantify the information content of

¹Copyright 2004 IEEE. Some of the material in this chapter has been reproduced, with permission, from: H. R. Sheikh and A. C. Bovik, “Image Information and Visual Quality,” IEEE Trans. Image Processing, submitted December 2003.

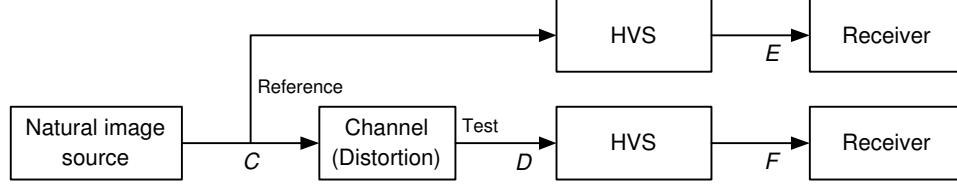


Figure 5.1: Mutual information between \mathcal{C} and \mathcal{E} quantifies the information that the brain could ideally extract from the reference image, whereas the mutual information between \mathcal{C} and \mathcal{F} quantifies the corresponding information that could be extracted from the test image.

the reference image as being the mutual information between the input and output of the HVS channel. This is the information that the brain could ideally extract from the output of the HVS. I then quantify the same measure in the presence of an image distortion channel that distorts the output of the natural source before it passes through the HVS channel, thereby measuring the information that the brain could ideally extract from the test image. This is shown pictorially in Figure 5.1. I then combine the two information measures to form a visual information fidelity measure that relates visual quality to *relative* image information [89].

Section 5.2 presents the development of the image information measure and the proposed visual information fidelity criterion. Implementation and subjective validation details are provided in Sections 5.3 and 5.4, while the results are discussed in Section 5.5. I conclude the chapter in Section 5.6.

5.2 Visual Information Fidelity for Image Quality Assessment

Natural images of perfect quality can be modeled as the output of a stochastic source. In the absence of any distortions, this signal passes through the HVS channel of a human observer before entering the brain, which extracts cognitive information from it. For distorted images, I assume that the reference signal has passed through another ‘distortion channel’ before entering the HVS. This is shown pictorially in Figure 5.1. The visual information fidelity criterion that I propose in this chapter is derived from a quantification of two mutual information quantities: the mutual information between the input and the output of the HVS channel when no distortion channel is present (I call this the *reference image information*) and the mutual information between the input of the distortion channel and the output of the HVS channel for the test image. I discuss the components of the proposed method in this section.

5.2.1 The Source Model

As in the case of IFC, (Section 4.2), the NSS model that I use is the GSM model in the wavelet domain. In this chapter I will only consider the vector GSM model. It is convenient to deal with one subband of the wavelet decomposition at this point and later generalize this for multiple subbands. A GSM is a random field (RF) that can be expressed as a product of two independent RFs [125]. That is, a GSM $\mathcal{C} = \{\vec{C}_i : i \in I\}$, where I denotes

the set of spatial indices for the RF, can be expressed as:

$$\mathcal{C} = \mathcal{S} \cdot \mathcal{U} = \{S_i \cdot \vec{U}_i : i \in \mathcal{I}\} \quad (5.1)$$

where $\mathcal{S} = \{S_i : i \in \mathcal{I}\}$ is an RF of positive scalars and $\mathcal{U} = \{\vec{U}_i : i \in \mathcal{I}\}$ is a Gaussian vector RF with mean zero and covariance \mathbf{C}_U . \vec{C}_i and \vec{U}_i are M dimensional vectors, and I assume that for the RF \mathcal{U} , \vec{U}_i is independent of \vec{U}_j , $\forall i \neq j$. Each subband of a scale-space-orientation wavelet decomposition (such as the steerable pyramid [98]) of an image is modeled as a GSM. I partition the subband coefficients into non-overlapping blocks of M coefficients each, and model block i as the vector \vec{C}_i .

One could easily make the following observations regarding the above model: \mathcal{C} is normally distributed given \mathcal{S} (with mean zero and covariance $\mathcal{S}_i^2 \mathbf{C}_U$), C_i are independent of $S_j \forall j \neq i$ given S_i , and given \mathcal{S} , \vec{C}_i are conditionally independent of \vec{C}_j , $\forall i \neq j$ [125].

5.2.2 The Distortion Model

As in the case of the IFC, the distortion model that I use in this chapter is also the signal attenuation and additive noise model in the wavelet domain:

$$\mathcal{D} = \mathcal{G}\mathcal{C} + \mathcal{V} = \{g_i \vec{C}_i + \vec{V}_i : i \in \mathcal{I}\} \quad (5.2)$$

where \mathcal{C} denotes the RF from a subband in the reference signal, $\mathcal{D} = \{\vec{D}_i : i \in \mathcal{I}\}$ denotes the RF from the corresponding subband from the test (distorted) signal, $\mathcal{G} = \{g_i : i \in \mathcal{I}\}$ is a deterministic scalar gain field, and $\mathcal{V} = \{\vec{V}_i :$

$i \in \mathcal{I}$ is a stationary additive zero-mean Gaussian noise RF with covariance $\mathbf{C}_V = \sigma_v^2 \mathbf{I}$. The RF \mathcal{V} is white, and is independent of \mathcal{S} and \mathcal{U} . I constrain the field \mathcal{G} to be slowly-varying.

5.2.3 The Human Visual System Model

The HVS model that I use is also described in the wavelet domain. Since HVS models are the dual of NSS models [102], many aspects of the HVS are already modeled in the NSS description, such as a scale-space-orientation subband decomposition, response exponent, and masking effect modeling [90]. The components missing include, among others, the optical point spread function (PSF), luminance masking, the contrast sensitivity function (CSF) and internal neural noise sources. I found from experiments that just modeling the distortions in the HVS channel as additive noise gives marked improvement in performance in terms of the ability of the overall algorithm to predict visual quality.

The model for the HVS channel noise that I use is a stationary, zero mean, additive white Gaussian noise model in the wavelet domain. It is widely known that the neurons of the primary visual cortex are orientation and frequency selective, much like the subbands in a wavelet decomposition [126]. Thus I model the HVS noise in the wavelet domain as a stationary RF $\mathcal{N} = \{\vec{N}_i : i \in \mathcal{I}\}$, where \vec{N}_i are zero-mean uncorrelated multivariate

Gaussian with the same dimensionality as \vec{C}_i :

$$\mathcal{E} = \mathcal{C} + \mathcal{N} \text{ (reference image)} \quad (5.3)$$

$$\mathcal{F} = \mathcal{D} + \mathcal{N} \text{ (test image)} \quad (5.4)$$

where \mathcal{E} and \mathcal{F} denote the visual signal at the output of the HVS model from the reference and the test images in one subband respectively, from which the brain extracts cognitive information (5.1). The RF \mathcal{N} is assumed to be independent of \mathcal{U} , \mathcal{S} , and \mathcal{V} . I model the covariance of the additive noise as:

$$\mathbf{C}_N = \sigma_n^2 \mathbf{I} \quad (5.5)$$

where σ_n^2 is a model parameter (variance of the HVS noise).

5.2.4 The Visual Information Fidelity Criterion

With the source, distortion, and HVS models as described above, the visual information fidelity criterion that I propose can be derived. Let $\vec{C}^N = (\vec{C}_1, \vec{C}_2, \dots, \vec{C}_N)$ denote N elements from \mathcal{C} . Let S^N , \vec{D}^N , \vec{E}^N and \vec{F}^N be correspondingly defined. In this section I will assume that the model parameters \mathcal{G} , σ_v^2 and σ_n^2 are known. As in the case of IFC, I will analyze the conditional mutual information between \mathcal{C} and \mathcal{E} (or \mathcal{F}) given \mathcal{S} .

For the reference image, I can analyze $I(\vec{C}^N; \vec{E}^N | s^N)$, where s^N denotes a *realization* of S^N . With the stated assumptions on \mathcal{C} and the distortion

model (5.2), I get:

$$I(\vec{C}^N; \vec{E}^N | s^N) = \sum_{j=1}^N \sum_{i=1}^N I(\vec{C}_i; \vec{E}_j | \vec{C}^{i-1}, \vec{E}^{j-1}, s^N) \quad (5.6)$$

$$= \sum_{i=1}^N I(\vec{C}_i; \vec{E}_i | s_i) \quad (5.7)$$

$$= \sum_{i=1}^N (h(\vec{C}_i + \vec{N}_i | s_i) - h(\vec{N}_i | s_i)) \quad (5.8)$$

$$= \frac{1}{2} \sum_{i=1}^N \log_2 \left(\frac{|s_i^2 \mathbf{C}_U + \sigma_n^2 \mathbf{I}|}{|\sigma_n^2 \mathbf{I}|} \right) \quad (5.9)$$

where we get (5.6) from chain rule [15], and (5.7) from the conditional independence of \mathcal{C} and \mathcal{N} given \mathcal{S} , and $|\cdot|$ denotes the determinant. Similarly one could show that for the test image

$$\begin{aligned} I(\vec{C}^N; \vec{F}^N | s^N) &= \sum_{i=1}^N (h(g_i \vec{C}_i + \vec{V}_i + \vec{N}_i | s_i) - h(\vec{V}_i + \vec{N}_i | s_i)) \end{aligned} \quad (5.10)$$

$$= \frac{1}{2} \sum_{i=1}^N \log_2 \left(\frac{|g_i^2 s_i^2 \mathbf{C}_U + (\sigma_v^2 + \sigma_n^2) \mathbf{I}|}{|(\sigma_v^2 + \sigma_n^2) \mathbf{I}|} \right) \quad (5.11)$$

Since \mathbf{C}_U is symmetric, it can be factored as $\mathbf{C}_U = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T$, where \mathbf{Q} is an orthonormal matrix, and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues λ_k . As in Section 4.3.3.1, one can use this matrix factorization to show:

$$I(\vec{C}^N; \vec{E}^N | s^N) = \frac{1}{2} \sum_{i=1}^N \sum_{k=1}^M \log_2 \left(1 + \frac{s_i^2 \lambda_k}{\sigma_n^2} \right) \quad (5.12)$$

$$I(\vec{C}^N; \vec{F}^N | s^N) = \frac{1}{2} \sum_{i=1}^N \sum_{k=1}^M \log_2 \left(1 + \frac{g_i^2 s_i^2 \lambda_k}{\sigma_v^2 + \sigma_n^2} \right) \quad (5.13)$$

$I(\vec{C}^N; \vec{E}^N | s^N)$ and $I(\vec{C}^N; \vec{F}^N | s^N)$ represent the information that could ideally be extracted by the brain from a particular subband in the reference and the test images respectively. I call $I(\vec{C}^N; \vec{E}^N | s^N)$ the *reference image information*. Intuitively, visual quality should relate to the amount of image information that the brain could extract from the test image *relative* to the amount of information that the brain could extract from the reference image. For example, if the information that could be extracted from the test image is 2.0 bits per pixel, and if the information that could be extracted from the corresponding reference image is 2.1 bits per pixel, then most of the information content of the reference image has been communicated by the test image. By contrast, if the corresponding reference image information were, say, 5.0 bits per pixel, then 3.0 bits of information have been lost to the distortion channel, and the visual quality of the test image should be inferior.

I discovered that a simple *ratio* of the two information measures relates very well with visual quality. It is easy to motivate this choice of relationship between image information and visual quality. When a human observer sees a distorted image, he has an idea of the amount information that he expects to receive in the image (modeled through the known \mathcal{S} field), and it is natural to expect the proportion of the expected information actually received from the distorted image to relate well with visual quality. This motivation is also strengthened by empirical evidence in terms of the excellent performance of the quality assessment algorithm.

Also I have only dealt with one subband so far. One could easily in-

corporate multiple subbands by assuming that each subband is completely independent of others in terms of the RFs as well as the distortion model parameters. Thus, the VIF that I propose in this dissertation is given by:

$$\text{VIF} = \frac{\sum_{j \in \text{subbands}} I(\vec{C}^{N,j}; \vec{F}^{N,j} | s^{N,j})}{\sum_{j \in \text{subbands}} I(\vec{C}^{N,j}; \vec{E}^{N,j} | s^{N,j})} \quad (5.14)$$

where we sum over the subbands of interest, and $\vec{C}^{N,j}$ represent N elements of the RF \mathcal{C}_j that describes the coefficients from subband j , and so on.

The VIF given in (5.14) is computed for a collection of $N \times M$ wavelet coefficients that could either represent an entire subband of an image, or a spatially localized region of subband coefficients. In the former case, the VIF is one number that quantifies the information fidelity for the entire image, whereas in the latter case, a sliding-window approach could be used to compute a *quality map* that could visually illustrate how the visual quality of the test image varies over space. Note that in contrast with most previous quality assessment methodologies, the VIF is *not* a Minkowski norm of the quality map. Thus, say, a mean of the VIF quality map is not the correct measure of image quality.

The VIF has a number of interesting features. Firstly, note that VIF is bounded below by zero (such as when $I(\vec{C}^N; \vec{F}^N | s^N) = 0$ and $I(\vec{C}^N; \vec{E}^N | s^N) \neq 0$), which indicates that all information about the reference image has been lost in the distortion channel. Secondly, in case the image is not distorted at all, and VIF is calculated between the reference image and its copy, VIF

is *exactly* unity². This is because $g_i = 1 \ \forall i$, and $\sigma_v^2 = 0$, and therefore $I(\vec{C}^N; \vec{F}^N | s^N) = I(\vec{C}^N; \vec{E}^N | s^N)$. Thus for all practical distortion types, VIF would lie in the interval $[0, 1]$. Thirdly, and this is where I feel that VIF has a distinction over traditional quality assessment methods, a linear contrast enhancement of the reference image that does not add noise to it will result in a VIF value *larger* than unity, thereby signifying that the enhanced image has a *superior* visual quality than the reference image! It is common observation that contrast enhancement of images increases their perceptual quality unless quantization, clipping, or display non-linearities add additional distortion. Theoretically, contrast enhancement results in a higher signal-to-noise ratio at the output of the HVS neurons, thereby allowing the brain to have a greater ability to discriminate objects present in the visual signal. The VIF is able to capture this improvement in visual quality. To the best of my knowledge, no other quality assessment algorithm has the ability to predict if the visual image quality has been enhanced by a contrast enhancement operation. I envision extending the notion of quantifying improvement in visual quality of

²The keen reader would have noticed that a constant mean-shift will also result in VIF being exactly unity signifying zero loss of visual quality. However, perceptual quality does vary with mean-shifts. One reason that this loss of quality is not captured by VIF measure is that I have not included any modeling of display non-linearities and luminance adaptations in the HVS model. It is well known that the response of many display devices to increasing pixel values is sub-linear. Hence a positive mean shift would *reduce* the contrast of the image, and this reduction of contrast would then be captured by VIF. Similarly, an increase in the luminance (average value) without increase in the contrast would cause the luminance adaptation mechanisms in the HVS to reduce the perceived contrast. However, for practical distortion types, I discovered that one could use a simpler HVS model that does not include display device modeling or luminance adaptation and still get good performance. Note also that the VIF has been derived in the wavelet domain, and the kernels are zero-mean. This means that the version of the VIF presented here would ignore mean shifts completely.

images by image enhancement operations using a similar information-theoretic paradigm.

It is interesting to see a few test cases that illustrate these properties of VIF visually. The implementation details of VIF are given in the next section; here I only wish to illustrate the above discussion pictorially. Figure 5.2 shows a reference image that has been distorted with three different types of distortion, all of which have been adjusted to have about the same mean squared error with the reference image. The distortion types illustrated are contrast stretch, Gaussian blur and JPEG compression. In comparison with the reference image, the contrast enhanced image has a better visual quality despite the fact that the ‘distortion’ (in terms of a perceivable difference with the reference image) is clearly visible. A VIF value larger than unity captures the improvement in visual quality. In contrast, both the blurred image and the JPEG compressed image have clearly visible distortions and poorer visual quality, which is captured by a low VIF measure for both. Notice that VIF has an interesting information-theoretic interpretation: VIF indicates the relative image information that is present in the distorted image. Thus, for example in Figure 5.2, VIF indicates only 7% of the reference image information is present in the blurred image.

Figure 5.3 illustrates the behavior of VIF with spatial quality maps. Figure 5.3(a) shows a reference image and Figure 5.3(b) the corresponding JPEG2000 compressed image. Note that the distortions are clearly visible. Figure 5.3(c) shows the reference image information map in the same loca-

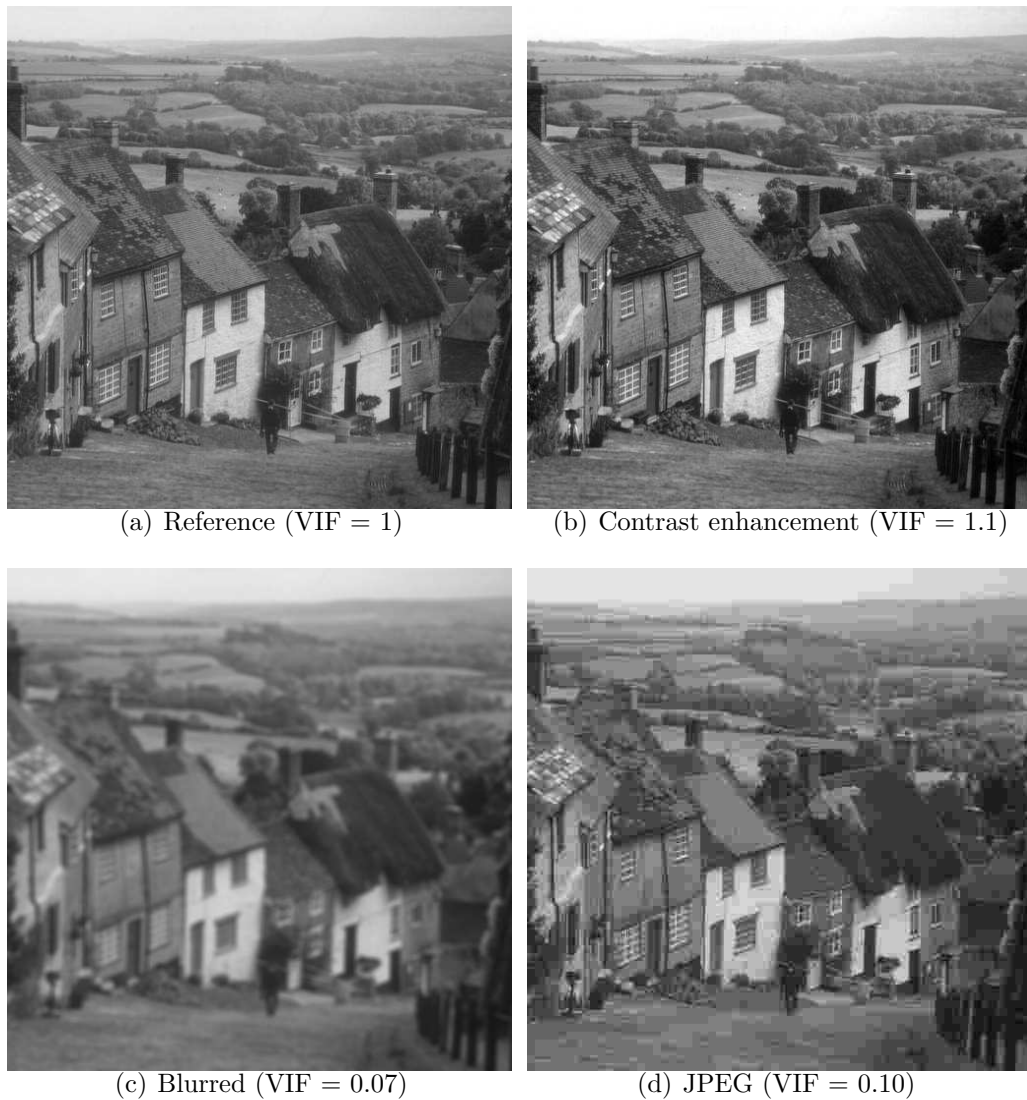


Figure 5.2: The VIF can capture the quality improvement from linear contrast enhancement. A VIF value greater than unity indicates this improvement, while a VIF value less than unity signifies a loss of visual quality. The MSE between the reference image and the test images is approximately the same in this figure.

tion. The information map shows the spread of structural information in the reference image. In flat image regions, the information content of the image is low, whereas in textured regions and regions containing strong edges, the image information is high. The quality map in Figure 5.3(d) shows the proportion of the image information that has been lost to JPEG2000 compression.

5.3 Implementation Issues

In order to implement VIF criterion in (5.14) a number of assumptions are needed about the source, distortion, and HVS models. I outline them in this section.

5.3.1 Assumptions About the Source Model

Ergodicity assumptions about the RF's is the same as in the previous chapter in Section 4.3.

The source model parameters that need to be estimated from the data consist of the field \mathcal{S} . For the vector GSM model, the maximum-likelihood estimate of s_i^2 can be found as follows [106]:

$$\hat{s}_i^2 = \frac{\vec{C}_i^T \mathbf{C}_U^{-1} \vec{C}_i}{M} \quad (5.15)$$

Estimation of the covariance matrix \mathbf{C}_U is also straightforward from the reference image wavelet coefficients [106]:

$$\hat{\mathbf{C}}_U = \frac{1}{N} \sum_{i=1}^N \vec{C}_i \vec{C}_i^T \quad (5.16)$$

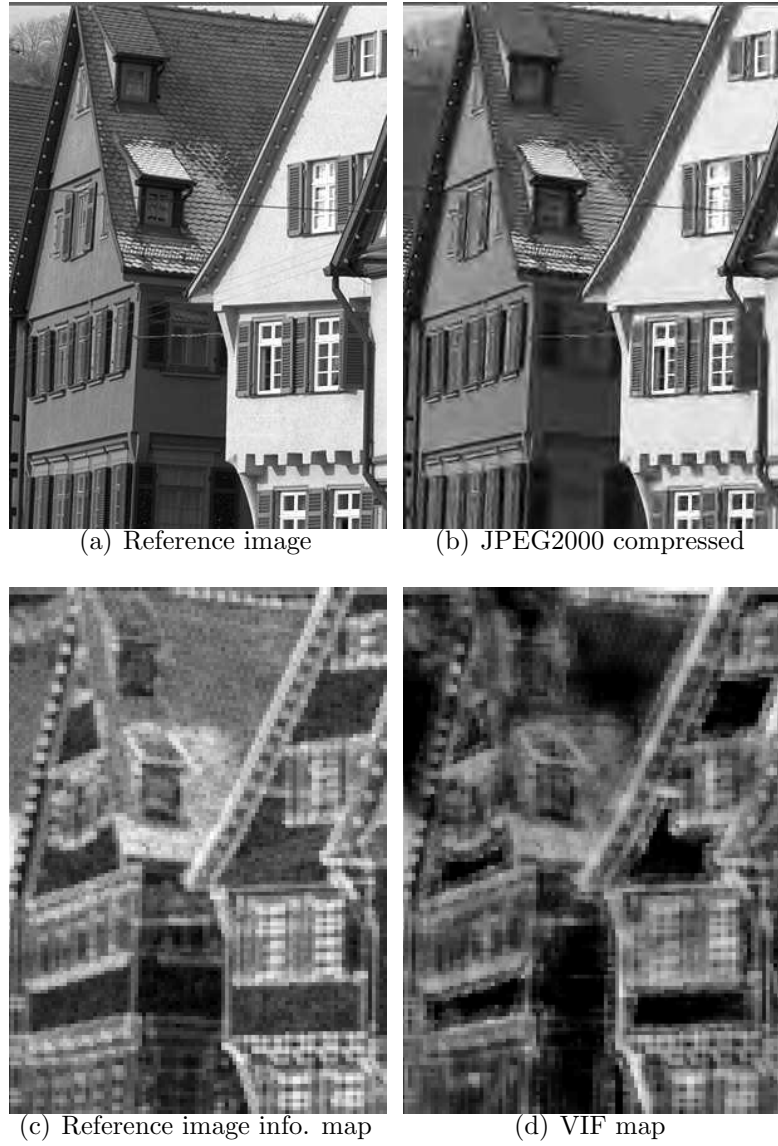


Figure 5.3: Spatial maps showing how VIF captures spatial information loss.

In (5.15) and (5.16), $E[S_i^2]$ is assumed to be unity without loss of generality [106].

5.3.2 Assumptions About the Distortion Model

In order for the assumptions on the distortion operator to hold, I estimate the parameters of the distortion channel *locally*. Hence I will use a $B \times B$ window centered³ at coefficient i to estimate g_i and σ_v^2 at i . The value of the field \mathcal{G} over the block centered at coefficient i , which I denote as g_i , and the variance of the RF \mathcal{V} , which I denote as $\sigma_{v,i}^2$, are fairly easy to estimate (by linear regression) since both the input (the reference signal) as well as the output (the test signal) of the system (5.2) are available:

$$\hat{g}_i = \widehat{\text{Cov}}(C, D) \widehat{\text{Cov}}(C, C)^{-1} \quad (5.17)$$

$$\hat{\sigma}_{v,i}^2 = \widehat{\text{Cov}}(D, D) - \hat{g}_i \widehat{\text{Cov}}(C, D) \quad (5.18)$$

where the covariances are approximated by sample estimates using sample points from the corresponding blocks centered at coefficient i in the reference and the test signals.

5.3.3 Assumptions About the HVS Model

The only parameter that I need to use in the HVS model is σ_n^2 . I chose to hand-optimize the value of the parameter σ_n^2 in my simulations by running

³A computationally simpler approach is to partition the subbands into non-overlapping blocks and assume that the field \mathcal{G} is constant over such blocks, as are the noise statistics σ_v^2 . This approach was taken for the IFC in Chapter 4, and causes a slight degradation in performance.

the algorithm over a range of values and observing its performance. I noted that while the performance is affected by the choice of σ_n^2 , the algorithm's overall performance continues to be highly competitive with other methods for a wide range of values. I will discuss the dependence of the performance of VIF later in Section 5.5.3.3.

Further specifics of the estimation methods used in my testing are given in Section 5.5.

5.4 Subjective QA Study for Validation

In order to calibrate and test the algorithm, an extensive subjective quality assessment study was conducted. In these experiments, a number of human subjects were asked to assign each image with a score indicating their assessment of the quality of that image, defined as the extent to which the artifacts were visible and annoying. The details of the study are given in Appendix A. In this study, a total of 982 images, out of which 203 were the reference images, were evaluated by human subjects, and the raw scores for each subject processed to give Mean Opinion scores (MOS) and a Difference Mean Opinion Score (DMOS) for each distorted image.

5.5 Results

In this section I present results on validation of VIF on the database presented in Section 5.4, and present comparisons with other quality assessment

algorithms. Specifically, I compare the performance of VIF against PSNR, SSIM [131], and the well known Sarnoff model (Sarnoff JND-Metrix 8.0 [84]). I present results for two versions of VIF: VIF using the finest resolution at all orientations, and using the horizontal and vertical orientations only. Table 5.1 summarizes the results for the quality assessment methods, which are discussed in Section 5.5.3.

5.5.1 Simulation Details

Some additional simulation details are as follows. Although full color images were distorted in the subjective evaluation, the QA algorithms (except Sarnoff's) operated upon the luminance component only. GSM vectors were constructed from non-overlapping 3×3 neighborhoods, and the distortion model was estimated with an 18×18 sliding window. Only the subbands at the finest level were used in the summation of (5.14). MSSIM (Mean SSIM) was calculated on the luminance component after decimating (filtering and downsampling) it by a factor of 4 (see [131]).

5.5.2 Calibration of the Objective Score

A calibration procedure similar to the one described in Section 4.5.2 was used, except that instead of *fminsearch*, I used *fminunc* for numerical curve-fitting.

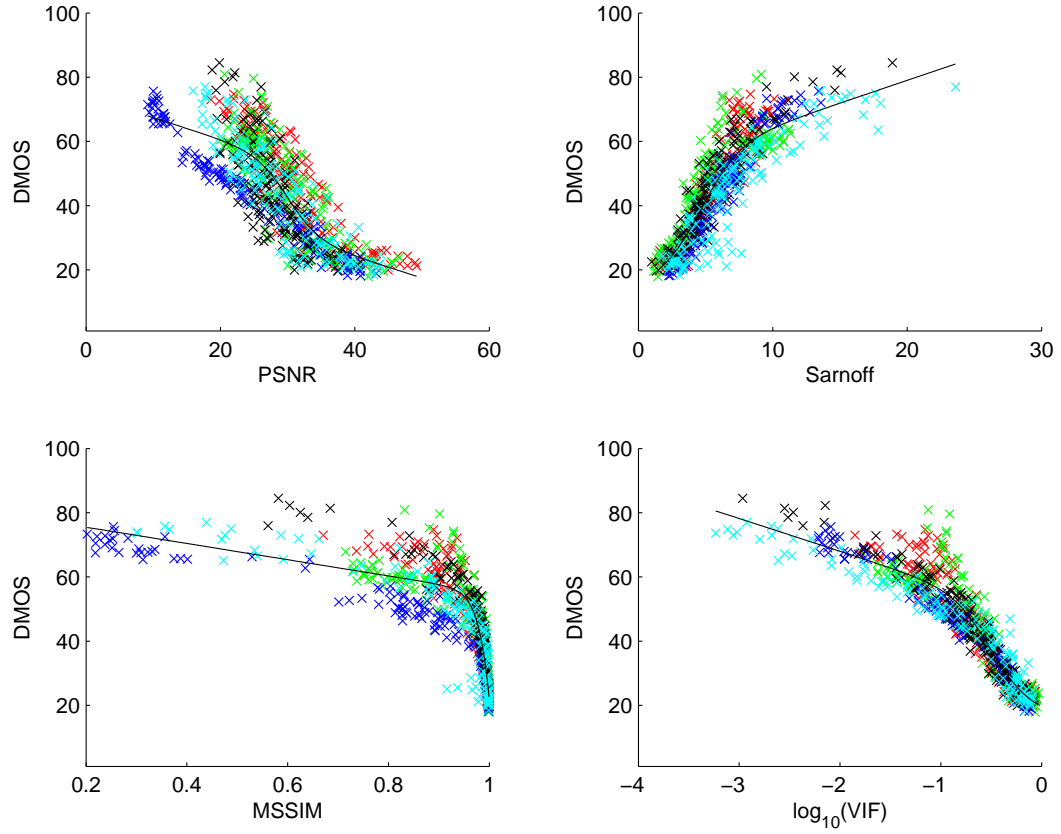


Figure 5.4: Scatter plots for the four objective quality criteria: PSNR, Sarnoff's JND-Metrix, MSSIM, and $\log(VIF)$ for VIF using horizontal/vertical orientations. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).

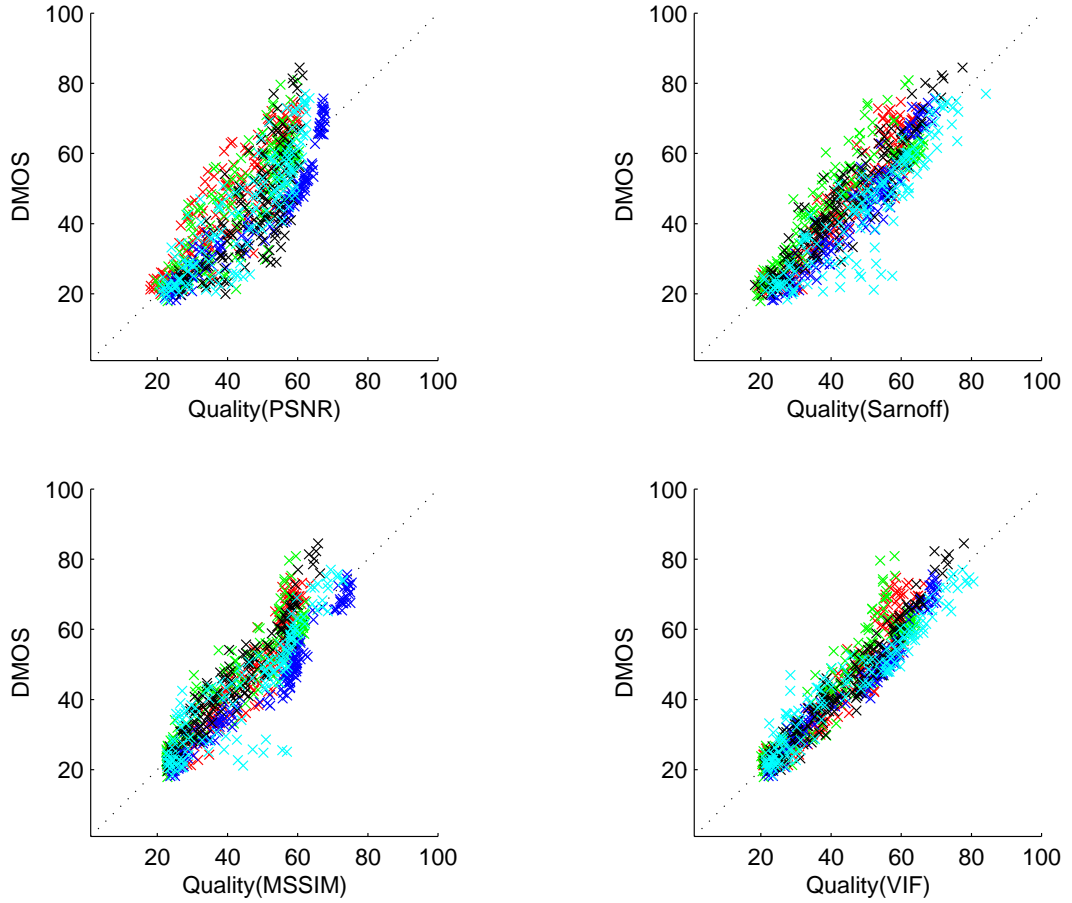


Figure 5.5: Scatter plots for the quality predictions by the four methods after compensating for quality calibration: PSNR, Sarnoff's JND-Metrix, MSSIM, and VIF using horizontal/vertical orientations. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan).

Validation against DMOS					
Model	CC	MAE	RMS	OR	SROCC
PSNR	0.826	7.272	9.087	0.114	0.820
Sarnoff	0.901	5.252	6.992	0.046	0.902
MSSIM	0.912	4.979	6.616	0.035	0.910
VIF	0.949	3.878	5.083	0.013	0.949
VIF (hv)	0.950	3.820	5.025	0.013	0.950

Table 5.1: Validation scores for different quality assessment methods. The methods tested were PSNR, Sarnoff JND-Metrix 8.0 [84], MSSIM [131], VIF, and VIF using horizontal and vertical orientations only. The methods were tested against DMOS from the subjective study after a non-linear mapping. The validation criteria are: correlation coefficient (CC), mean absolute error (MAE), root mean squared error (RMS), outlier ratio (OR) and spearman rank-order correlation coefficient (SROCC).

5.5.3 Discussion

5.5.3.1 Overall performance

Table 5.1 shows that VIF is competitive with all state-of-the-art FR QA methods presented in this chapter and outperforms them in my simulations by a sizeable margin. Also note that VIF and MSSIM use only the luminance components of the images to make quality predictions, whereas the JND-Metrix uses all color information. Extending VIF to incorporate color could further improve performance.

As noted in Chapter 4, the performance of VIF improves slightly when only the horizontal and vertical orientations are used in the summation in Chapter (5.14), although the improvement is less marked than in Chapter 4. Nevertheless, the reduced computational complexity makes this a much more

RMSE performance on specific distortions.				
Distortion	PSNR	Sarnoff	MSSIM	VIF
JPEG2000	7.187	5.028	4.693	4.873
JPEG	8.173	5.451	5.511	5.366
White noise	2.583	3.967	2.702	2.338
Gaussian blur	9.774	5.104	5.178	3.446
FF	7.517	6.713	6.990	4.000

Table 5.2: RMSE performance of the QA methods on individual distortion types.

attractive implementation option.

5.5.3.2 Cross-distortion performance

It is interesting to study the performance of VIF on specific distortion types. Many image QA methods perform well on single distortion types, but their limitations show up on a broader validation study involving different distortion types. Nevertheless, it is sometimes interesting from an application perspective to restrict the quality measures to a single distortion type. Table 5.2 shows the performance of VIF and other measures on each of the five distortion types. Note that while the JND-Metrix, MSSIM and VIF perform quite well on individual distortion types, their performance worsens in cross-distortion validation, with VIF's worsening the least. Note that VIF performs better than (or at par with) JND-Metrix and MSSIM in cross-distortion validation (Table 5.1) as well as for individual distortion types (Table 5.2).

Figure 5.6 shows graphically why is it important for a QA measure to perform well across distortions. Figure 5.6 shows the DMOS calibration

curves for each of the five distortion types present in the database ⁴. Ideally for a QA method, these curves should lie on top of each other. If this were the case, then the QA measure could stably predict quality across distortion types. For the PSNR scale for example, we see that good quality images (where DMOS is around 20), have PSNR values that lie in the approximate interval from 40 to 50 dBs, which is roughly 25% of the entire range of values that the PSNR takes. In contrast, we see that for good to medium quality images (DMOS values between 20 and 40), VIF curves are very close to each other, signifying that the mapping of VIF to visual quality is more stable, and has a smaller dependence on the underlying distortion type. Note that the distortion types present in the database are quite diverse, including linear blur, blocking, white noise as well as blurring/ringing from JPEG2000 compression, and transmission error in JPEG2000 bit stream.

At poorer quality ranges, the calibration curves for all four methods diverge, as shown in Figure 5.6 (one could note by visual inspection that the curves for VIF diverge far less than those for PSNR). One reason for this could be the lack of proper judgement scales in human observers for bad quality images, or psychometric scale warping effects at the lower end of quality.

⁴The non-linearity used for MSSIM is different from the one used in Figure 5.4 and Table 5.2 for illustrative purposes.

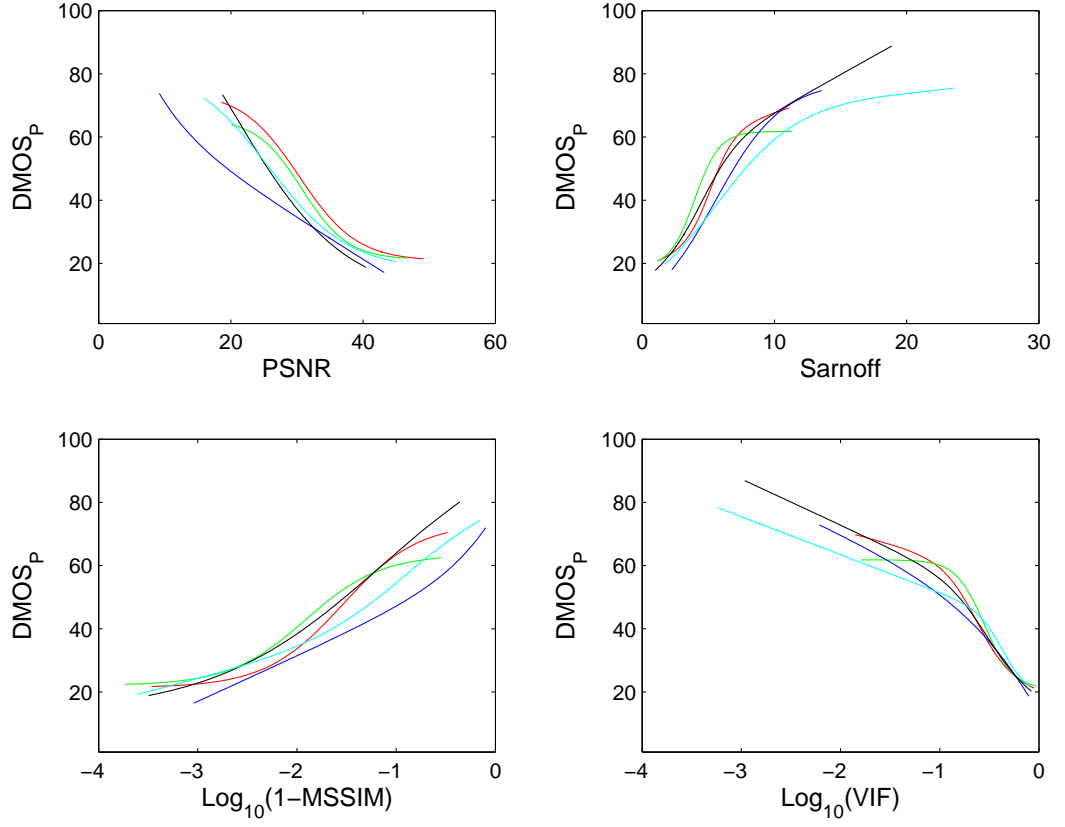


Figure 5.6: Calibration curves for the four quality assessment methods for individual distortion types. The distortion types are: JPEG2000 (red), JPEG (green), white noise in RGB space (blue), Gaussian blur (black), and transmission errors in JPEG2000 stream over fast-fading Rayleigh channel (cyan). Note that VIF can be stably calibrated for predicting quality for a wider range of distortion types. The mapping used for MSSIM in this figure is $\text{log}_{10}(1 - \text{MSSIM})$ for illustrative purposes.

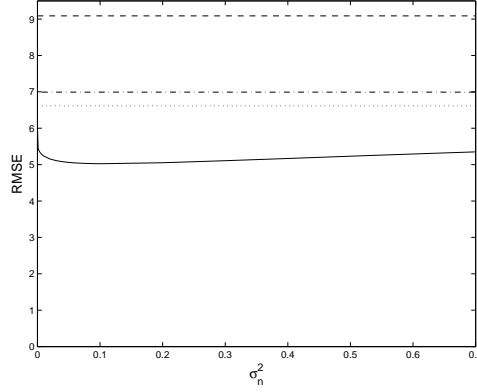


Figure 5.7: Dependence of VIF performance on the σ_n^2 parameter. Note that VIF performs better than other methods against which it is compared in this chapter for all range of values of σ_n^2 shown this figure: VIF (solid), PSNR (dashed), Sarnoff JND-Metrix 8.0 (dash-dot), and MSSIM (dotted) .

5.5.3.3 Dependence on the HVS parameter

It was mentioned in Section 5.3 that the value of HVS noise variance, σ_n^2 was hand-optimized. It is instructional to study the dependence of the performance of VIF on σ_n^2 . Ideally, σ_n^2 should depend on the dynamic range of the input, and a multiplicative constant should instead be tuned, as was done in [131], but here I only wish to show that the performance of VIF is relatively robust to small changes in the value of the parameter σ_n^2 . Figure 5.7 shows how the RMSE varies with σ_n^2 . It can be seen that VIF performs better than all the methods compared against in this chapter for the entire range of values of σ_n^2 shown in Figure 5.7 (see Table 5.1), with an approximate minimum occurring at 0.10.

5.5.3.4 Computational Complexity

The VIF has one disadvantage when compared against PSNR or MSSIM: it has a higher computational complexity. Most of this complexity comes from computing the wavelet decomposition, and the parameters of the distortion model. In Chapter 4, one version of the fidelity criterion using downsampling was presented, which has the potential to substantially reduce the computational complexity of the algorithm. Also, many estimation methods presented in the chapter could be simplified greatly at the cost of slight reduction in performance. Nevertheless, even without these improvements, the VIF using the horizontal and vertical subbands with MATLAB implementation takes about 16 seconds to run for a 512×768 image on a Pentium III, 1 GHz Machine.

5.6 Conclusions

In this chapter I presented a novel notion of image information and demonstrated its relationship with image quality. The VIF, which was derived from a statistical model for natural scenes, a model for image distortions, and a human visual system model in an information-theoretic setting, outperformed traditional image QA methods in our simulations by a sizeable margin. The VIF was demonstrated to be better than a state-of-the-art HVS based method, the Sarnoff's JND-Metrix, as well as a state-of-the-art structural fidelity criterion, the structural similarity (SSIM) index, in my testing. I demonstrated that VIF performs well in single-distortion as well as in cross-distortion scenarios.

The IFC presented in Chapter 4, and the VIF presented in this chapter are both full-reference quality assessment methods that are based on natural scene statistics. In the next chapter, I show how NSS models could be used for NR QA of images using *a priori* information about the distortion process as well.

Chapter 6

No-Reference Quality Assessment Using Natural Scene Statistics: JPEG2000

There is no evil in the world

It's just a joy-ride of continuity, that's what I am saying

I am not evil, I am not the world

I am not a joy-ride of continuity

What would I know what's with evil or what's with the world

And I would go on to say that whatever stays alone is destined to perish!

Evil, goodness, world, continuity,

all of these come from the House of Immortality,

and I don't have anything to do with any such place!

I am one, and I am alone, I am a stranger

This city, this wilderness, these flowing roads and rivers

This horizon

A tall building suddenly coming into vision

These decaying cemeteries, and in them

like ever-coming death itself, undertakers

These innocent laughing children

Run over by a car this dying blind traveler
The winds, the greenery, and clouds in the skies floating from here to there
What's all this?
This is what the world is!
It is a joy-ride of continuity!
The city, the wilderness, the roads and rivers,
the skies, the buildings, undertakers, travelers,
winds, greenery, and clouds in the skies floating from here to there,
all of these, each one of them,
comes from MY creed
I AM the world
My existence fuels an endless joy-ride of continuity!
But there is no evil in me
How should I say that?
That in me both mortality and eternity have come together.

— *"Uniqueness" by Meeraji*

6.1 Introduction

The standard approach for quality assessment over the years has been that of image fidelity measurement using the full-reference quality assessment paradigm. But human beings do not need to have access to the reference to make judgements about quality. Given that human beings can do quality

assessment so easily without the reference image, can we design computer algorithms to do the same? This is the problem of blind or No-Reference (NR) quality assessment, and it has recently received a great deal of attention, since the reference signal may not be available for many applications, or may be too expensive to provide. But given the limited success that FR quality assessment has achieved, it should come as little surprise that NR quality assessment is a very hard problem indeed, and is far from being a mature research area.

The problem of NR quality assessment may seem hopelessly difficult at first. How can a computer be expected to assess the quality of an image or video and to ascertain the degradation in its information content without understanding it first? There is solace, however, in the fact that natural scenes belong to a small set in the space of all possible signals, and many researchers have developed statistical models to describe natural scenes, and that most distortions that are prevalent in image/video processing systems are not natural in terms of such statistics. Thus, one could consider the following alternative philosophy for NR QA: *all images are perfect, regardless of content, unless distorted during acquisition, processing or reproduction*. This philosophy assigns equal quality to all natural visual stimuli that human beings could possibly encounter, and the task of NR QA is reduced to blindly measuring the distortion (using signal or distortion models) that has possibly been introduced during the stages of acquisition, processing or reproduction, and then calibrating this measurement against human judgements of quality.

In this chapter¹, I propose to use natural scene statistics models for assessing the quality of images and videos blindly. Using the above philosophy, I will demonstrate the use of an NSS model for blindly measuring the quality of images compressed by JPEG2000 [91, 92, 94].

6.2 The JPEG2000 Image Compression Standard

Among the important compression algorithms that one encounters daily is the (lossy) JPEG image compression standard [69]. The JPEG compression algorithm is based on the block-based Discrete Cosine Transform (DCT). Over the years researchers have found that greater compression could be achieved at the same visual quality if the DCT is replaced by the Discrete Wavelet Transform (DWT). The research activity in wavelet based image compression techniques has resulted in the JPEG2000 compression standard [111], standardized and approved only recently.

While there are many interesting features of the JPEG2000 image compression algorithm, we are only interested in the distortion process that occurs during compression, which is the chief source of artifacts in most cases. Although other sources may also be present, such as distortions resulting from bit errors during transmission, I ignore them in this work. In brief, JPEG2000 compression, operating in the baseline lossy mode, computes the DWT using

¹Copyright 2004 IEEE. Some of the material in this chapter has been reproduced, with permission, from: H. R. Sheikh, A. C. Bovik and L. K. Cormack, “No-Reference Quality Assessment Using Natural Scene Statistics: JPEG2000,” IEEE Trans. Image Processing, revised December 2003.



Figure 6.1: Uncompressed image.

the biorthogonal 9/7 wavelet [12, 111]. The DWT coefficients are quantized using a scalar quantizer, with possibly different step sizes for each subband. This quantization causes many small DWT coefficients to become zero. The result is that the reconstruction from the quantized DWT coefficients contains blurring (since small high-frequency DWT coefficients, which are responsible for image details and sharpness, may have become zero) and ringing artifact (due to zeroing of coefficients and mapping of actual DWT coefficients to quantized values). Figure 6.4 shows the reconstruction from the quantized DWT coefficients.

The key difference between the distortion introduced by JPEG2000 and

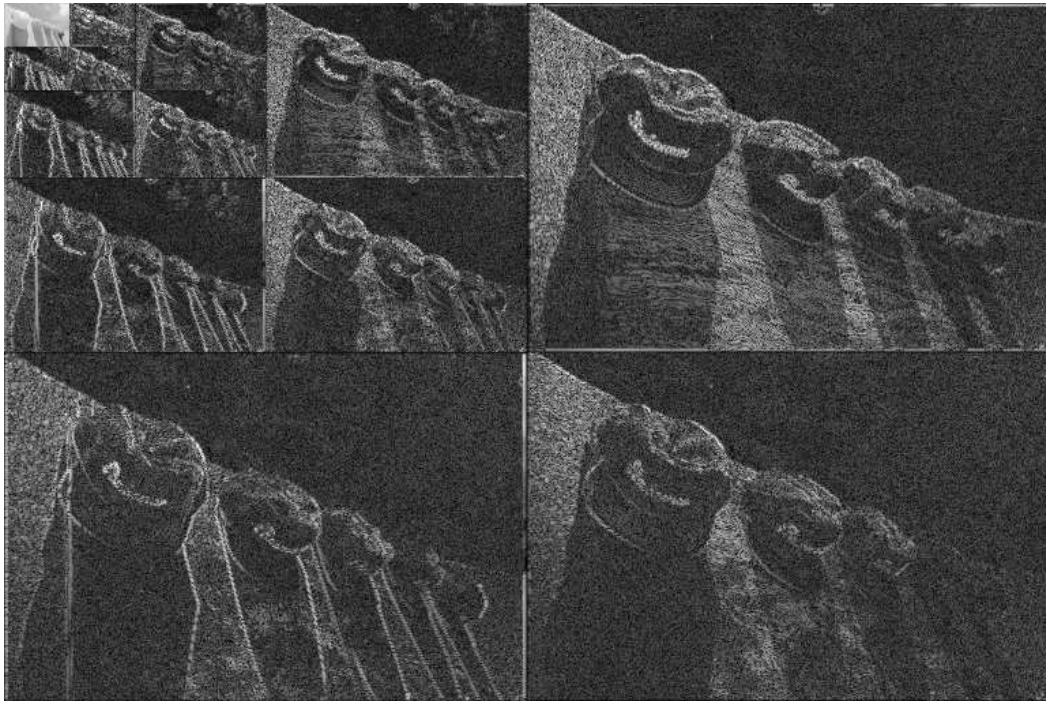


Figure 6.2: Four-level DWT coefficients (magnitude) of the uncompressed image using the 9/7 biorthogonal Wavelet. The subband coefficients are contrast stretched and non-linearly enhanced for better display.

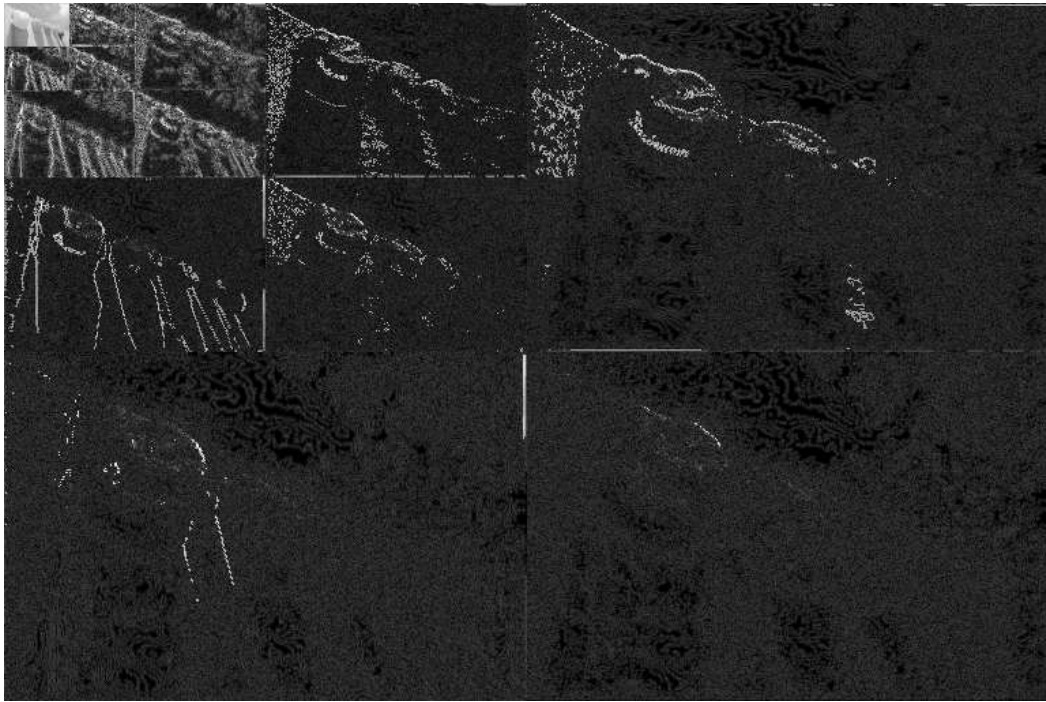


Figure 6.3: DWT coefficients (magnitude) of the image after compression by JPEG2000. The subband coefficients are contrast stretched and non-linearly enhanced for better display.



Figure 6.4: Image compressed by JPEG2000

that by JPEG is that in the case of JPEG2000, the distortion does not show regularity, such as periodic occurrences or directional preference (blocking occurs as horizontal/vertical edges every eight pixels). Rather, the distortion is image dependent, occurring mostly in highly textured areas or around strong edges, with the structure of the distortion being locally dependent on the image structure and the compression ratio. This lack of regularity in the distortion complicates the task of quantifying it without reference. However, if one were to use natural scene statistics models, especially those in the wavelet domain, one could identify and quantify the distortion introduced by the quantization of the wavelet coefficients.

6.3 Blind Measurement of Image Naturalness

6.3.1 Statistical Model for Natural Images in the Wavelet Domain

One particularly useful model for natural scene statistics has been presented in [8, 96]. It captures the statistics of wavelet coefficients of natural images in a given subband and their correlations with other wavelet coefficients across scales and orientations. I noted that this model is suitable for measuring the effect of quantization of wavelet coefficients of natural images, since quantization pushes wavelet coefficients at finer scales towards zero. This results in a greater probability of zero coefficients in any subband than expected for natural images.

The statistical model proposed in [8, 96] models the wavelet coefficient's magnitude, C , conditioned on the magnitude of the linear prediction

of the coefficient, P , and is given in (6.1) where M and N are assumed to be independent zero mean random variables:

$$C = MP + N \quad (6.1)$$

$$P = \sum_{i=1}^n l_i C_i \quad (6.2)$$

where the coefficients C_i come from an n coefficient neighborhood of C in space, scale, and orientation, and l_i are linear prediction coefficients [8].

In [8, 96], the authors use an empirical distribution for M and assume N to be Gaussian of unknown variance. The linear prediction, P , comes from a set of neighboring coefficients of C at the same scale and orientation, different orientations at the same scale, and coefficients at the parent scales. Zero-tree based wavelet image coding algorithms [82, 88, 111] also try to capture the non-linear relationship of a coefficient with its parent scales, and hence are conceptually similar to (6.1).

Figure 6.5 shows the joint histograms of $(\log_2(P), \log_2(C))$ of an image at different scales and orientations. The strong non-linear dependence between C and P is clearly visible on the logarithmic axes. As can be seen from the figure, the model can describe the statistics across subbands and orientations. The authors of [8, 96] report (and I observed) that the model is stable across different images as well.

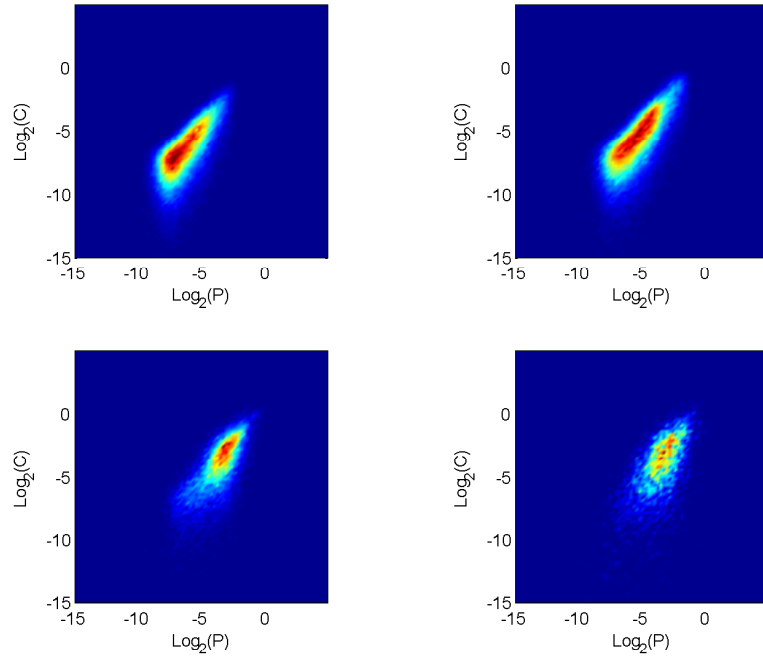


Figure 6.5: Joint histograms of $(\log_2 P, \log_2 C)$ for an uncompressed natural image at different scales and orientations of its wavelet decomposition. Top left: Diagonal subband at the finest scale. Top right: Horizontal subband at the finest scale. Bottom left: Vertical subband at the second-finest scale. Bottom right: Diagonal subband at the third-finest scale.

6.3.2 Compressed Natural Images

The model in (6.1) (Figure 6.5) is not very useful for modeling compressed images however, since quantization of the coefficients significantly affects the distribution. Figure 6.6 shows the joint histograms of a subband from the uncompressed and compressed versions of an image at different bit rates. The effects of the quantization process are obvious, in that quantization pushes the coefficients towards zero, and disturbs the dependencies between C and P ².

I propose to use a simplified two-state model of natural scenes in the wavelet domain. These two states correspond to a coefficient or its predictor being significant or insignificant. The joint two-state model is motivated by the fact that the quantization process in JPEG2000, which occurs in all subbands, results in more of P and C values being insignificant than expected for natural images. Hence, a good indicator for unnaturalness and the perceptual effects of quantization is the proportion of significant P and C . Thus, if the proportion of significant P and C is low in an image, it could be a result of quantization in the wavelet domain.

The details of the model are as follows: Two image-dependent thresholds, one for P and the other for C , are selected for each subband for bina-

²The histograms for compressed images shown in Figure 6.6 have open bins to take care of $\log_2(0)$. Since I computed the DWT on reconstructed JPEG2000 images, the bulk of the insignificant coefficients are not exactly zero due to computational residues in the calculations of DWT/inverse-DWT and color transformations. However, the algorithm will not be affected by these small residues, as will become apparent shortly.

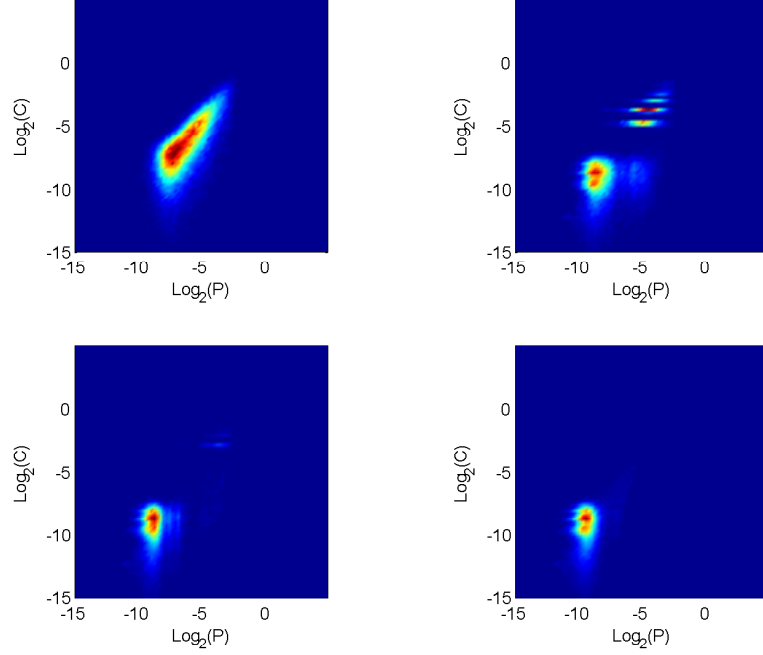


Figure 6.6: Joint histograms of $(\log_2 P, \log_2 C)$ for one subband of an image when it is compressed at different bit rates using the JPEG2000 compression algorithm. Top left: No compression. Top right: 2.44 bits/pixel. Bottom left: 0.78 bits/pixel. Bottom right: 0.19 bits/pixel.

rization. The details of threshold computation will be given in Section 6.3.4. A coefficient (or its predictor) is considered to be significant if it is above the threshold³. Consequently, we obtain a set of four empirical probabilities, $p_{ii}, p_{is}, p_{si}, p_{ss}$, corresponding to the probabilities that the predictor/coefficient

³The binarized model may remind some readers of two-state hidden Markov tree models for the wavelet coefficients of images, in which a coefficient is associated with the state of a two-state Markov process describing whether the coefficient is insignificant or significant [77]. However, the proposed model is conceptually and computationally simpler, since it does not require complicated parameter estimation associated with hidden Markov tree models, but still provides good performance.

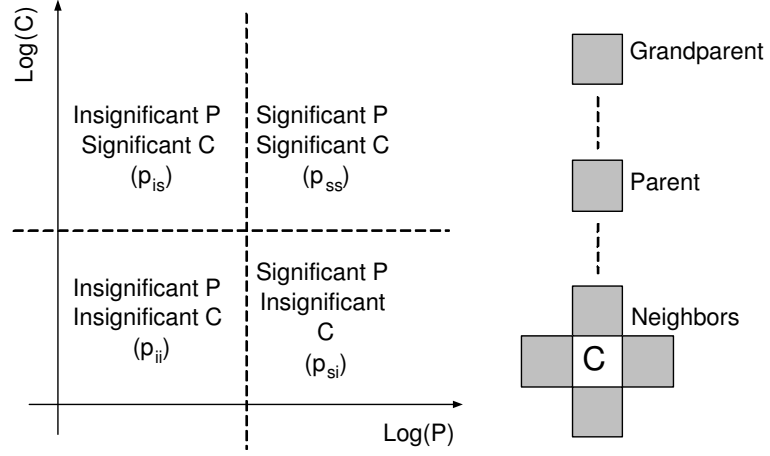


Figure 6.7: Partition of the (P, C) space into quadrants. Also, the set of coefficients from which P is calculated in my simulations.

pair lies in one of the four quadrants, as depicted in Figure 6.7. Obviously the sum of all these probabilities for a subband is unity. The set of neighbors from which P is calculated in my simulations is also shown in Figure 6.7.

6.3.3 Features for Blind Quality Assessment

I observed from experiments that the subband probabilities p_{ss} give the best indication of the loss of visual quality, in terms of minimizing the quality prediction error. In my simulations, I compute the p_{ss} feature from six subbands: horizontal, vertical and diagonal orientations at the second-finest resolution; horizontal, vertical and diagonal orientations at the finest resolution. Since the subband statistics are different for different scales and orientations, a non-linear combination of these features is required. This is because the p_{ss} feature from a finer subband decreases faster with increas-

ing compression ratio than the p_{ss} feature from a coarser subband. In order to compensate for these differences, I propose to nonlinearly transform each subband feature independently, before combining them linearly. The nonlinear transformations is designed to improve correspondence among the features coming from different subbands before a simple average rule is applied.

I train the transformations using training data by fitting image quality to the p_{ss} features from individual subbands as follows:

$$q_i = K_i \left(1 - \exp \left(-\frac{(p_{ss,i} - u_i)}{T_i} \right) \right) \quad (6.3)$$

where q_i is the transformed feature (predicted image quality) for the i -th subband, $p_{ss,i}$ is the p_{ss} probability for the i -th subband, and K_i, T_i and u_i are curve fitting parameters for the i -th subband that are learned from the training data. Since each subband feature is mapped by a nonlinearity for that subband to subjective quality, the transformed features from all subbands are approximately aligned with subjective quality, and hence with each other as well.

A weighted average of the transformed feature is used for quality prediction. Due to the similarity in the statistics of horizontal and vertical subbands at a particular scale, I constrain the weights to be the same for these orientations at a given scale. Thus, the six-dimensional subband quality vector, $q = \{q_i | i \in 1 \dots 6\}$, is modified into a four-dimensional vector q' by averaging the quality predictions from horizontal and vertical subbands at a given scale,

and the final quality prediction is taken to be a weighted average of q' :

$$\begin{aligned} \begin{bmatrix} q'_1 \\ q'_2 \\ q'_3 \\ q'_4 \end{bmatrix} &= \begin{bmatrix} (q_1 + q_2)/2 \\ q_3 \\ (q_4 + q_5)/2 \\ q_6 \end{bmatrix} \\ Q &= q'^T w \end{aligned} \tag{6.4}$$

where the weights w are learned by minimizing quality prediction error over the training set.

6.3.4 Image-dependent Threshold Calculations

One important prerequisite for a feature for quality assessment is that it should adequately reflect changes in visual quality due to compression while being relatively robust to changes in the image content. However, the feature proposed in Section 6.3.2 varies not only with the amount of quantization, but also with the variation in image content. For example, if the thresholds were held constant, then a low p_{ss} feature could either signify a heavily quantized image, or a relatively smooth image (say the sky and the clouds) that was lightly quantized. In order to make the QA features more robust to changes in the image content, I propose to adapt the thresholds with image content, so that they are lower for smooth images, and higher for highly textured images. Once again I will use NSS models in conjunction with distortion modeling to achieve this goal.

An interesting observation is that when the means of (\log_2 of) subband coefficient amplitudes are plotted against an *enumeration* of the subbands,

the plot is approximately linear. This is shown in Figure 6.8(b) for a number of uncompressed natural images. The graphs have approximately the same slope, while the intercept on the y -axis varies from image to image. This approximately linear fall-off is expected for natural images since it is well known that natural image amplitude spectra fall off approximately as $\frac{1}{f}$, which is a straight line on log-log axes. Another aspect of NSS is that horizontal and vertical subbands have approximately the same energy, whereas the diagonal subbands have lower energy at the same scale. The interesting observation is that a diagonal subband sits between the horizontal/vertical subband at its scale and the one finer to it.

Quantized images are not natural however, and hence the corresponding plots for them will not have approximately linear fall-off (Figure 6.8(c), solid lines). However, we know that the quantization process in wavelet based compression systems (such as the JPEG2000) is designed in such a way that the subband means for coarser subbands are less affected by it, whereas the means for finer subbands are affected more. Hence, from the coarser subbands, one could predict the line that describes the energy fall-off for the image by estimating its intercept (assuming that all natural images have the same slope). This line yields the estimated means for the finer subbands in the unquantized image from which the compressed image whose quality is being evaluated was derived. This is shown in Figure 6.8(c) as well, where the means of (\log_2 of) subband coefficients are plotted for an image compressed at different bit rates (as well as the uncompressed image). Notice that predicted subband means

(shown by dotted lines) are quite close to the actual means of the uncompressed image (top solid line).

I use the above observation to calculate the image dependent thresholds as follows:

$$\text{Threshold} = \text{Estimated Subband Mean} + \text{Offset} \quad (6.5)$$

The slope of the line can be learned from uncompressed natural images in the training set, while the offsets (one for P and one for C for each subband) can be learned by a minimization process that attempts to minimize the error in the quality predictions over the training set (using MATLAB's command *fminsearch*). In this way, the proposed algorithm utilizes NSS models in concert with modeling the salient features of the distortion process to make the QA feature more robust to changes in the image content.

6.3.5 Simplified Marginal Model

Since the computation of coefficient predictions P is expensive, I also consider the marginal distribution of the binarized wavelet coefficients C , as opposed to the joint distribution of binarized P and C . Figure 6.9 shows the histogram of the logarithm (to the base 2) of the horizontal wavelet coefficient magnitude for one image, and the histogram for the same compressed image at the same scale and orientation. Quantization shifts the histogram of $\log_2(C)$ towards lower values. I divide the histogram into two regions: insignificant coefficients and significant coefficients. Again, the probability of a coefficient being in one of the regions at a certain scale and orientation is a good feature

to represent the effect of quantization. I map the probabilities by (6.3) and then take a weighted average as in (6.4).

6.4 Results

In this section I will present the performance of the methods presented in this chapter.

6.4.1 Subjective QA Study for Training and Testing

In order to calibrate, train, and test the NR QA algorithm, an extensive subjective QA study was conducted. Details of the study are provided in Appendix A. In these experiments, a number of human subjects were asked to assign each image with a score indicating their assessment of the quality of that image, defined as the extent to which the artifacts were visible. A total of 198 JPEG2000 images were ranked in the study (including the reference images).

6.4.2 Simulation Details

For training and testing, the database was divided into two parts. The training database consisted of fifteen randomly selected images (from the total 29) and all of their distorted versions. The testing database consisted of the other fourteen images and their distorted versions. This way there was no overlap between the training and the testing databases. The algorithm was run several times, each time with a different (and random) subset of the original

29 images for training and testing, with 15 and 14 images (and their distorted versions) in the training and testing sets respectively.

The algorithm was run on the luminance component of the images only, which was normalized to have a Root-Mean-Squared (RMS) value of 1.0 per pixel. The biorthogonal 9/7 wavelet with four levels of decomposition is used for the transform. The slope of the line for estimating the subband coefficient means in (6.5) is learned from the uncompressed images in the training set. The weights w in (6.4) are learned using non-negatively constrained least-squares fit over the training data (MATLAB command *lsqnonneg*). The minimization over the threshold offsets in (6.5), as well as for the fitting parameters in (6.3) is done by unconstrained non-linear minimization (MATLAB command *fminsearch*).

6.4.3 Quality Calibration

The quality calibration used for the NR simulations is the same as in Section 4.5.2.

6.4.4 Quality Prediction Results

Figure 6.10(a) shows the predictions of the algorithm on the testing data for one of the runs against the MOS for the joint binary simplification of the model. Figure 6.10(b) shows the normalized histogram of the RMSE (which I use as a measure of the performance of the NR metric) between the quality prediction and the MOS, for a number of runs of the algorithm.

The mean RMSE is 8.05, with a standard deviation of 0.66. Figure 6.11(a) shows the histogram of the linear correlation coefficient for the joint model. The average linear correlation coefficient between the prediction and the MOS for all the runs is 0.92 with a standard deviation of 0.013. Figure 6.10(c) shows the RMSE histogram for the marginal binary simplification of the model. The mean RMSE is 8.54 and the standard deviation is 0.83. Figure 6.11(b) shows the histogram of the linear correlation coefficient for the joint model. The average linear correlation coefficient is 0.91 with a standard deviation of 0.018. Figure 6.12 shows how the RMSE varies with the threshold offset for one simulation for the binarized marginal model.

6.4.5 Discussion

It is apparent from the above figures that the proposed NR algorithm is able to make predictions of the quality of images compressed with JPEG2000 that are consistent with human evaluations. The average error in quality assignment for a human subject is 7.04, while for the algorithm it is 8.05. *It is therefore performing close to the limit imposed on useful prediction by the variability between human subjects*, that is the variability between the proposed algorithm and MOS is comparable to the variability between a human and the MOS on average. The average gap between the quality prediction ability of an average human and the algorithm is only about 1.0 on a scale of 1 - 100. Another interesting figure is the standard deviation of the RMSE of 0.66 on a scale of 1 - 100, which indicates that the algorithm's performance is

stable to changes in the training database. As a comparison, we can compare the performance of the algorithm against a recent no-reference perceptual blur metric presented in [58]. The blur metric reports an average correlation coefficient of 0.86 on the same database ⁴. This shows that the blur caused by the JPEG2000 is of a much different nature than that caused by linear kernels, and is much more difficult to quantify using blur detection algorithm. Quantization in the wavelet domain, as in the JPEG2000, not only causes blur, but also introduces ringing distortion, which is in fact a high-frequency distortion close to strong edges, that would fool blur detection algorithms.

It is also interesting to analyze the performance of the algorithm for different images in order to expose any content-dependence. Figure 6.13(a) shows the prediction-MOS graph for one image for which the algorithm consistently performed well over the entire range of quality. Figure 6.13(b) shows the case where the algorithm consistently over-predicts the quality and Figure 6.14 shows the case where the algorithm consistently under-predicts the quality. I believe that the spatial *spread* of image details and texture over the image affects the performance of the algorithm and makes it content dependent.

6.4.6 Implementation Complexity

The computational complexity of the proposed NR algorithm is small, especially the simplification based on the wavelet coefficients marginals. The

⁴The authors did not report partitioning the database into non-overlapping sets for training and testing. They randomly chose training and testing sets without specifying if the source images for the two sets were disjoint or not.

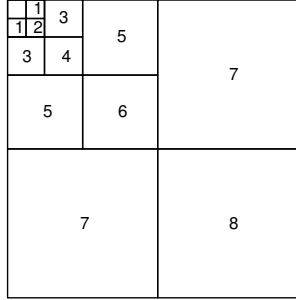
proposed algorithm takes about 20 seconds per image for the joint model and about 5 seconds per image for the marginal model with un-optimized MATLAB implementation running on a Pentium-III, 1 GHz machine. If the DWT coefficients are already available, say directly from the JPEG2000 stream, the marginal model computation reduces essentially to counting the number of significant coefficients, which is computationally very efficient.

6.5 Conclusions

In this chapter, I proposed to use Natural Scene Statistics for measuring the quality of images and videos without the reference image. I claimed that since natural images and videos come from a tiny subset of the set of all possible signals, and since the distortion processes disturb the statistics that are expected of them, a deviation of a signal from the expected natural statistics can be quantified and related to its visual quality. I presented an implementation of this philosophy for images distorted by JPEG2000 (or any other wavelet based) compression. JPEG2000 compression disturbs the non-linear dependencies that are present in natural images. Adapting a non-linear statistical model for natural images by incorporating quantization distortion modeling, I presented an algorithm for quantifying the departure of compressed images from expected natural behavior, and calibrated this quantification against human judgements of quality. I demonstrated the metric on a data set of 198 images, calibrated and trained the algorithm on data from human subjects, and demonstrated the stability and accuracy of the algorithm to changes in the

training/testing content. The algorithm performs close to the limit imposed by variability between human subjects on the prediction accuracy of an NR algorithm. Specifically, I achieved an average error of 8.05 (on a scale of 1-100) from the MOS, whereas subjective human opinion is expected be deviant by 7.04 from the MOS on average.

I feel that NSS modeling should be an important component of image and video processing algorithms that operate on natural signals. Efforts need to be made into designing NR QA algorithms that can predict quality for a broader class of distortion types, such as noise, wireless channel errors, watermarking etc., by posing the problem in an estimation-theoretic framework based on statistical models for natural scene.



(a) Subband labels n .

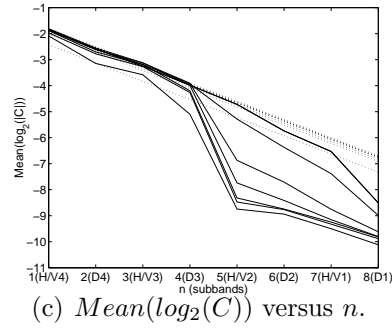
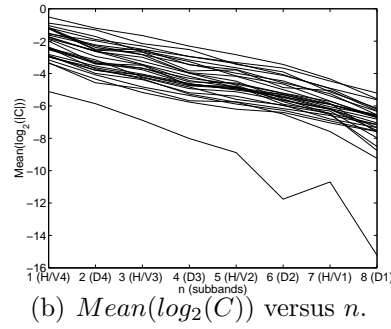


Figure 6.8: $Mean(\log_2(C))$ versus subband enumeration index. The means of horizontal and vertical subbands at a given scale are averaged. (a) The subband enumeration n used in (b) & (c). (b) Uncompressed natural images. $Mean(\log_2(C))$ falls off approximately linearly with n . (c) $Mean(\log_2(C))$ for an image at different bit rates (solid) and the corresponding linear fits (dotted). The fits are computed from $n = 1 \dots 4$ only. Note that the estimated fits are quite close to the uncompressed image represented by the top most solid line. These linear fits are used for computing the image-dependent thresholds for the corresponding images.

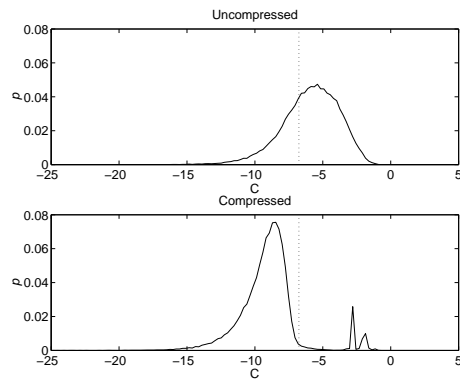


Figure 6.9: Histograms of the logarithm (to the base 2) of the horizontal subband the finest scale for one image, before and after compression (at 0.75 bits per pixel). The dotted line denotes the threshold at -6.76.

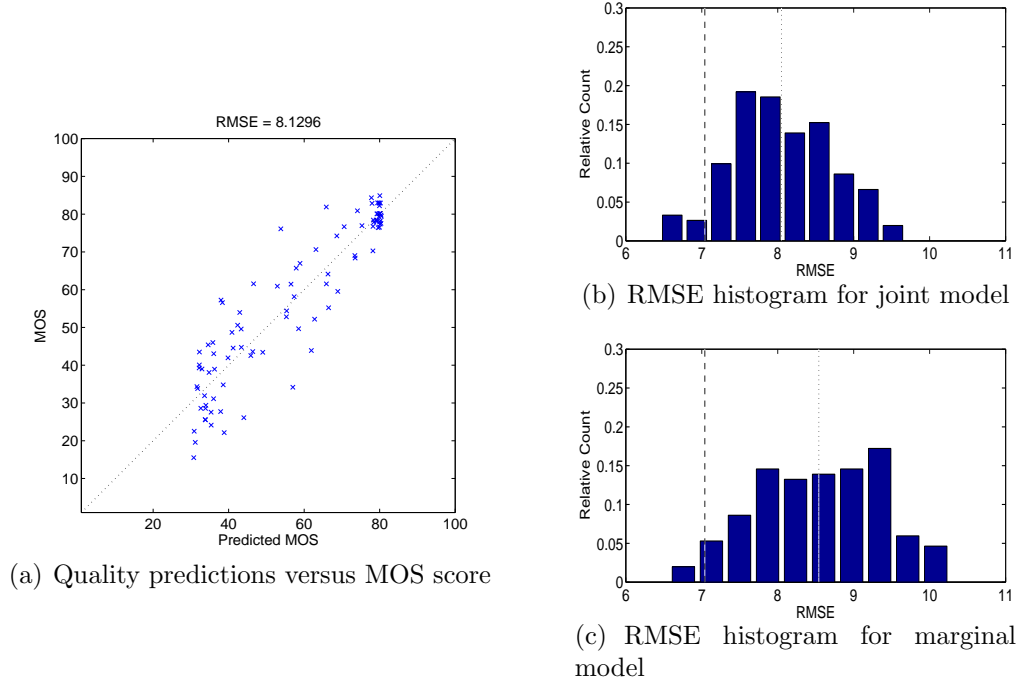


Figure 6.10: Results: (a) Quality predictions versus MOS for one run of the algorithm. Normalized histograms of the RMSE for several runs of the algorithm using the joint statistics of P and C (b) and using the marginal statistics of C only (c). The dotted line in the histograms show the mean value, while the dashed line shows the standard deviation of human scores.

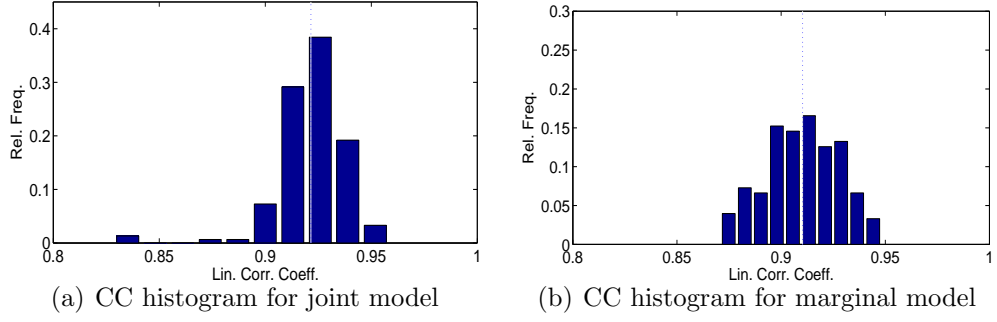


Figure 6.11: Results: Normalized histograms of the linear correlation coefficient for several runs of the algorithm using the joint statistics of P and C (b) and using the marginal statistics of C only (c). The dotted lines in the histograms show the mean value.

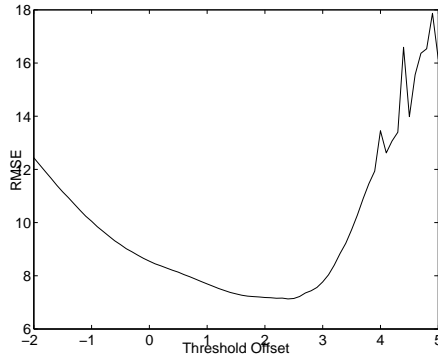
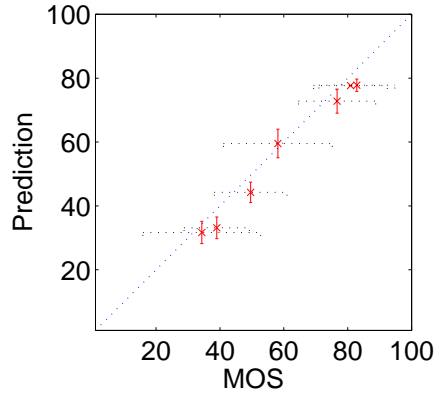
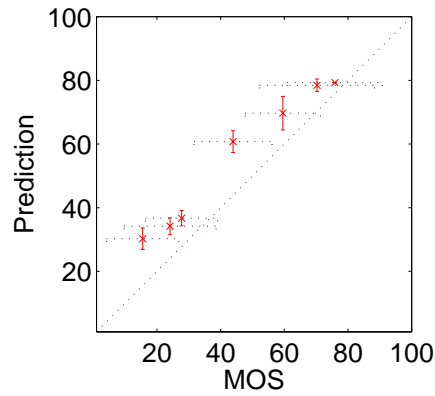


Figure 6.12: Variation of the RMSE with threshold offset in (6.5) for one simulation involving the entire database.



(a)



(b)

Figure 6.13: Content-dependence of the performance of the proposed NR QA algorithm over the quality range: (a) Algorithm performs consistently well over the range of quality. (b) The algorithm consistently over-predicts the quality. The MOS-Prediction point ‘ \times ’ and the ± 2 standard deviation intervals are shown for each figure. The variance in the MOS is due to inter-human variability and the variation in the prediction is due to changes in the training set.

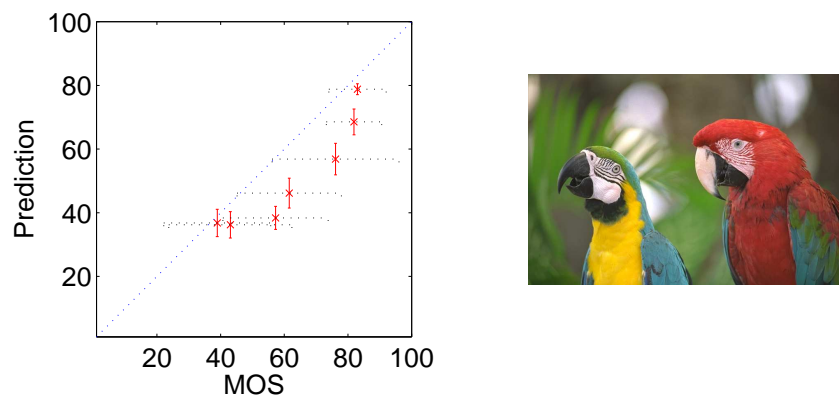


Figure 6.14: Content-dependence of the performance of the proposed NR QA algorithm over the quality range: The algorithm consistently under-predicts the quality. The MOS-Prediction point ‘ \times ’ and the ± 2 standard deviation intervals are shown for each figure. The variance in the MOS is due to inter-human variability and the variation in the prediction is due to changes in the training set.

Chapter 7

Conclusions and Future Work

So you think I am
Senseless with drunken arrogance
Consumed by a burning vengeance
Or drenched in destructive intolerance!
Listen warriors, soldiers!
Betray not the meekness of my desires
For if I stumble to the mountain foot, I fear
that I would eat you all alive!
For you are all syrupy men of self-righteousness!
You would laugh at my helplessness, laugh on
I would pray: O Lord of light and color and music
have mercy on their souls!
(Lord, Lord of the new light, the new song, the new scent!
Lord, Lord of the winds, Lord of the waters, Lord of the unsaid words!
Lord of the pen, Lord of the flute!
Lord of the new miracle of life!)

—from “Men of self-righteousness” by Noon Meem Rashid

7.1 Dissertation Summary

In this dissertation, I presented image quality assessment algorithms, full-reference and no-reference, that were based on statistical models for natural images. Images of the three dimensional visual environment constitute a very tiny subset in the set of all possible images. I claimed that the distortions present in real-world image processing systems make these images unnatural, and that the departure from expected natural behavior could be quantified for prediction of visual quality.

In this direction, I proposed two FR and one NR method for image quality assessment. The FR methods were based on a novel information theoretic framework based on natural scene statistics, which posed the QA problem as an information-fidelity problem rather than a signal fidelity problem. This novel approach yielded FR methods that outperformed state-of-the-art methods in my simulations.

I also proposed using NSS models for NR quality assessment. In this direction, I presented an algorithm for NR QA of JPEG2000 compressed images that was based on adapting a wavelet-domain NSS model using characteristics of the distortion process. This joint modeling of source and distortion yielded a QA algorithm whose predictions of image quality were in close agreements with human judgements.

For the purpose of training and validating the QA methods, I conducted an extensive subjective quality assessment study that resulted in a large set of

images that were ranked by a number of human observers. The database was diverse in terms of the image content, distortion type as well as the distortion range. It proved to be of fundamental importance in my study of the QA problem.

7.2 Contributions

The specific contributions of this dissertation are as follows:

1. For full-reference quality assessment, I proposed a novel information-theoretic framework based on natural scene statistics, which approached the QA problem as an information fidelity problem. In this framework, a source of natural images communicates to a receiver (cognitive processes in the brain) through a channel that imposes limitations on the amount of information that could flow through it. I claimed that visual quality should relate with a measure of success of this communication. Using natural scene statistics models and a channel model that I proposed in this dissertation, I showed that the mutual information, which quantifies the amount of information about the source that is received by the receiver, is a good measure of image information fidelity that relates with visual quality.
2. I proposed an image degradation model that consisted of describing image distortions locally as combinations of blur and additive noise. This model, being in the wavelet domain, is especially suited for use in concert

with popular wavelet based NSS models for analytical purposes.

3. I proposed the use of conditional mutual information given the knowledge of the ‘hidden variance field’. I hypothesized that such a conditioning has two advantages: it tunes the source model to suit a particular reference image in question, and also allows analytical tractability.
4. I derived closed-form conditional mutual information expressions for the chosen source and distortion models.
5. The IFC that I proposed in Chapter 4 quantified the amount of information shared between a reference and a test image. I demonstrated that IFC is competitive with state-of-the-art quality assessment algorithms, and outperformed them in my simulations. Furthermore, the proposed IFC is parameterless like MSE, being independent of training data, data from psychometric experiments, viewing configuration information, or stabilizing constants.
6. I showed that IFC is a dual of a QA method based on the human visual system. This duality of two QA methods derived from entirely different perspectives is certainly reassuring. Based on IFC and this duality, I hypothesized directions in which HVS based methods could be improved.
7. I extended the notion of quantifying information fidelity by proposing to view the HVS as a channel as well that places limits on the amount of information that could flow through it to the rest of the brain. This

allows one to define the notion of *image information*, which quantifies the statistical information content of the reference image. I thus combined source, distortion, *and* HVS models in one framework that related visual quality to the *relative* image information present in the test image.

8. The VIF presented in Chapter 5 utilized two aspects of image information: the information fidelity between the test and the reference image, and the information content of the reference image itself. The VIF, being a ratio of these two information quantities, had the desirable property of lying in the interval $[0, 1]$ for distorted images, and being larger than unity for contrast enhanced images.
9. For the first time in the field of image quality assessment, to my knowledge, a QA method (the VIF) was able to quantify improvements in visual quality over the reference image.
10. I showed via simulations that VIF outperformed state-of-the-art QA methods by a sizeable margin.
11. I proposed an NR quality assessment method for measuring the quality of images distorted by JPEG2000 compression. The wavelet quantization in JPEG2000 is especially hard to quantify without a reference since it causes two perceptual distortion artifacts: blurring and ringing. By using joint modeling of natural images and characteristics of the distortion process, I designed a feature that quantified the degree of unnaturalness

in an image caused by wavelet quantization, and used it to predict visual quality.

12. In order to train and test the algorithms presented in this dissertation, I conducted extensive experiments to collect subjective quality scores for a diverse set of images distorted using five distortion types. To my knowledge, this study is one of the most extensive studies conducted in the QA literature, containing about 22,000 subjective quality evaluations by human subjects. This database proved to be of fundamental importance to my exploration of the quality assessment problem.

7.3 Future Work

7.3.1 Full-Reference Image Quality Assessment

There are a number of directions in which my dissertation research on full reference image quality assessment could be extended:

- The distortion model that I used for the development of IFC and VIF needs to be examined more closely. In particular, I feel that the relationships between the model that I chose, ‘natural distortions’, that is image distortions that are present in nature, and natural scene statistics need to be explored further.
- The HVS model presented in Chapter 5 for the development of image information measures and the VIF needs to be more sophisticated. In particular, I feel that the optics of the eye and the contrast sensitivity

function (CSF) could be explicitly modeled in addition to the noise model already used in Chapter 5. Furthermore, a more sophisticated noise model should also be explored. It is well known that the noise in the neurons depends upon the signal as well [30]. This signal dependence could be further explored for QA purposes.

- A display device model could also be incorporated into the development of IFC and VIF.
- In the development of both IFC and VIF, inter-subband correlations were ignored, and different subbands of an image were treated as being independent. It is common knowledge that subbands are not independent, and this dependence needs to be incorporated into IFC and VIF.
- The estimation method used for estimating the distortion parameters \mathcal{G} and σ_v^2 , as well as the hidden field \mathcal{S} , was the maximum-likelihood method. Researchers have worked on MAP estimates for estimating \mathcal{S} using different prior distributions for \mathcal{S} [123, 125]. Similar estimates could also be explored for distortion model parameters, and incorporated into IFC and VIF.
- A more sophisticated approach for handling correlations between the wavelet coefficients within a subband could be explored. The ‘blocking’ approach provides great improvements over those versions of IFC in which the correlations were ignored. This leads one to wonder if further

improvements are possible by reducing the correlations between subband coefficients even more.

- The IFC and VIF could be optimized for faster implementation and reduced complexity by exploring different transforms, such as the DCT, or separable wavelet decompositions.
- Recently, complex wavelet analysis for natural scene models has gained popularity [42, 86, 87]. These models introduce another statistic, local phase, that has recently been shown to play important role in modeling natural image characteristics [135], as well as texture synthesis [73]. Such models could also be explored for quality assessment.
- Color is an important component of visual stimulus that was unfortunately ignored in my treatment for the sake of simplicity. The QA methods presented in this dissertation should certainly be extended using color NSS and HVS models.

7.3.2 No-Reference Image Quality Assessment

No-reference quality assessment methods are most desirable for quality monitoring applications in multimedia systems. But unfortunately, the generic NR QA problem is largely unsolved. People have resorted to restricting the scope of these methods to, say, the blocking artifact, or, as in my dissertation, to blurring and ringing caused by wavelet domain quantization.

The reason NR problem is so hard is that distortions occurring in image

and video processing systems are manmade, and it is quite likely that some hitherto unknown distortion would appear due to the processing systems of the future. This generality in distortion type may not bother a human observer however. It is easy to imagine a thought experiment in which an image distorted by some unknown distortion is shown to a human observer. It would be a trivial task for the human observer to say if the quality of the image is good or bad. This motivates one to consider the possibility of generic NR QA algorithms.

A set of complementary image features that could construct a multidimensional space in which natural images reside in one subspace and unnatural images in another should be explored. Such an approach of heuristically constructing spaces for image representation has been used effectively in [73] for texture analysis and synthesis, and I feel that the generic paradigm could be used for NR QA methods as well. Once such a space is constructed, one could then quantify the distance between a test image and the subspace of natural images. The underlying assumption here is that a set of image features exists that would allow the construction of such a space. However, finding this set of features and the subspace of natural images in the space spanned by these features would basically be a trial-and-error exercise based on existing distortion types, and still no guarantees could be given regarding the generic applicability of QA methods based on this approach. Hopefully, such an undertaking would give further insights into solving the generic NR problem.

One thing is however certain in my mind: natural scene models would

be at the heart of any generic NR QA method.

7.3.2.1 Future Improvements for NR QA of JPEG2000

A number of ideas could be used for improving the performance of the NR QA algorithm for JPEG2000 compressed images presented in Chapter 6 to further narrow the gap between the human and machine assessment. To this end, I propose investigating a number of extensions:

1. Although NSS models are global, images seldom have stationary characteristics. The algorithm is, as yet, blind to the spread of spatial detail in images, or to the presence of different textures within an image, which may have different statistics from one another. Incorporation of spatial adaptation into the algorithm by constructing ‘quality maps’, and investigations into optimal pooling methods could improve algorithm performance.
2. The proposed simplification of the NSS model reduces continuous variables to binary ones. More sophisticated models for characterizing natural images and their compressed versions could result in more stable features and improved algorithm performance.

7.3.3 Reduced-Reference Image Quality Assessment

A compromise between the infeasibility of FR methods and the unavailability of NR methods is the so called reduced-reference (RR) approach to the QA problem. In this approach, *some* information extracted from the

reference image is made available to the QA algorithm. The whole idea is that the amount of information should be much smaller than the reference signal for the sake of feasibility. Research into RR methods requires identification of features to be extracted from the reference image as well as the actual quality assessment algorithms. Limited amount of work has been done in RR QA, with basically two kinds of approaches. In the first approach, information is hidden in the reference image before it is processed by the distorting system. The output, the test image, is used to recover this information. The RR method consists of comparing the extracted information against the originally inserted information, which is made available through the RR channel [24, 107]. In the second approach, image features are extracted from the reference image and transmitted to the RR algorithm running at the output of an image distorting system [145, 148], and compared for QA.

Once again, RR methods proposed so far do not make explicit use of NSS models, and the use of these models for RR QA needs to be explored.

7.3.4 Video Quality Assessment

One of the aims of research into image quality assessment is to discover feasible methods for video quality assessment. Due to the tremendous resource demands that realtime video processing systems have, research into efficient systems is in dire need of accurate, yet feasible, video quality assessment algorithms. Many researchers have worked in the past on the problem of FR QA for video using HVS models [109, 110, 115, 116, 139, 143, 147, 155] as

well as using signal fidelity or feature based methods [152]. A survey of video distortions can be found in [134, 156], while a review of HVS modeling for video quality metrics is presented in [146]. A structural approach for video QA has also been attempted [132], while a hybrid HVS-structural measure has been presented in [154]. Some RR and NR methods for video for the blocking artifact have also been explored [29, 44, 107, 145, 148, 150]. I feel that the information-theoretic framework that I presented in this dissertation is ripe for extensions to video QA. This would require incorporation of the temporal dimension into the analysis, for the source model as well as for the distortion and HVS models. Researchers have worked on discovering the temporal characteristics of natural ‘videos’ [19, 100], which could prove helpful.

7.3.5 Image Information Metrics

An interesting byproduct of my dissertation research is the development of a statistical measure for image information. This measure represents the amount of information that could ideally be extracted from an image by the brain. This metric is a combination of NSS and HVS models in an information theoretic settings, and is an intriguing entity in itself. Combined with other features of the HVS, such as foveation (nonuniform sampling of the visual stimulus in the retina), color, motion, and stereo NSS and HVS models, this framework for quantifying statistical image information might have a much wider impact beyond the quality assessment problem. Perhaps it could help us explore attraction of human gaze by image features, or help us in

understanding how humans solve the stereo correspondence problem.

Appendices

Appendix A

Subjective Quality Assessment Study

In this appendix I will describe the details of the psychometric experiments used for training and testing the algorithms proposed in this dissertation.

A.1 Scope of the Subjective QA Study

The design of a subjective study needs to accurately reflect the intended scope of application of the quality assessment algorithms that would be trained and tested on it. For example, if a QA method is required to operate only on JPEG compressed images of human faces in the FR framework, then the subjective study should only involve images of human faces that have been compressed using JPEG. Training and testing such an algorithm using generic images distorted by blurring artifacts would obviously be useless. Unfortunately, this dependence on the final application limits the scope of subjective studies. Subjective QA studies are cumbersome and resource-consuming, and a compromise between the applicability of its results and the cost of the study is unavoidable.

In the ideal situation, a generic QA algorithm is sought that can predict

image quality with high accuracy for all types of images and image distortions. For the FR case, such a generic method is highly desirable since a tremendously large amount information (the reference image) is made available to the QA method at a substantial resource cost. Thus, the subjective study designed for testing and validating a generic FR metric should be as diverse as possible, both in terms of the image content as well as the distortion types. For NR methods, it is generally acceptable to limit the scope of the algorithm to a specific distortion type, primarily because the NR problem is largely an unsolved problem. The corresponding experiments would therefore involve only the target distortion type at the target distortion range.

The scope of the QA algorithms that I had envisioned for my dissertation was multifarious. I was aiming for a study that could simultaneously be useful for the FR methods proposed in this dissertation, as well as NR methods. Moreover, I sought a database with generic photographic images depicting various scenes containing different types of content, and which were distorted using a wide variety of distortion types. If the testing were however conducted in several sessions, each with only one distortion type, then such a database could simultaneously prove useful for generic FR as well as NR algorithms, and my testing methodology reflects this desire.

A.2 Subjective Quality Assessment Methodologies

A number of alternative methodologies are available for doing subjective quality assessment studies, each having its own domain and advantages

[14, 57, 117, 118, 140]. A set of standardized methodologies are also available for subjective quality assessment of television pictures [34, 35].

There are three main types of experimental methodologies that could be used for subjective quality assessment studies: double-stimulus methods, single-stimulus methods, and forced-choice methods. Double-stimulus methods are specifically suited for testing and validating FR QA methods. In this method of testing, the test and the corresponding reference image or video are shown one after the other¹, and the human assessor assigns a quality score to both signals. This way the *relative* quality of the test image or video can be ascertained, and any systematic sources of error that could result from, say, the variation in image content, are minimized. Double-stimulus methods have the advantage of being more robust to contextual effects, and the results are considered to be more stable and reproducible. However, double-stimulus methods assign each distorted video with a *quality difference* score, and as such these methods are not suited for scenarios where a quality score is to be assigned to an image or video in the absence of the reference. Single-stimulus methods, in which only the test image or video is evaluated for quality by the human assessor, are more suitable for such tasks.

The third subjective test methodology is a two-alternative forced-choice (2AFC) setup between two images or videos that could be perceptually different from each other. The goal is to make a judgement whether they are

¹The order of the presentation of the reference and the test signals is randomized.

different or not. The two signals are shown one after the other in a random order, and the subject chooses one of the two alternative choices to rank them, such as better/worse, good/bad, undistorted/distorted etc. Thus a quality evaluation is not made, only a quality comparison, and the test and the reference images cannot be assigned a quality score. Such methods are known to be most stable to contextual effects, and are the most reliable among the three types of testing. However, the 2AFC methods are very cumbersome, and require many trials, and its hard to convert these results into a quality scale.

An interesting stimulus synthesis methodology for comparison of two image quality assessment metrics against one another has recently been proposed, but is limited in application to differentiable QA methods [136], and in my view, local minima/maxima and limitations of finite precision during iterative minimization and maximization in very high dimensional spaces may render this method inconclusive.

The details of the experimental procedure that I chose for my research is presented in Section A.4.

A.3 Designing the Image Database

A.3.1 Choice of Input Images

Since I sought to design QA methods that had generic applicability, I chose to use a set of images that reflected adequate diversity in image content. Twenty-nine high resolution and high quality (as per my personal judgement) color images were collected from the Internet and photographic

CD-ROMs. These images included pictures of faces, people, animals, close-up shots, wide-angle shots, nature scenes, man-made objects, images with distinct foreground/background configurations, and images without any specific object of interest. Some of the images had high activity, while some were mostly smooth. These images were resized to a reasonable size for display on the screen resolution of 1024×768 that I had chosen for the experiments. Most images were 768×512 pixels in size.

A.3.2 Distortion Types

I chose to distort the images using five different image distortion types that could occur in real-world applications. The distortion types were: JPEG compression, JPEG2000 compression, additive white Gaussian noise in the RGB components, Gaussian blur in the RGB components, and bit errors in the JPEG2000 bitstream when it is transmitted over a simulated wireless channel (more specifically, a fast-fading Rayleigh channel). The distortion parameters varied for generating the database are listed in Table A.1. These distortions reflect a broad range of image impairments, from smoothing, to structured distortion, image-dependent distortions, and random noise. The level of distortion was varied to generate images at a broad range of quality, from imperceptible levels to high levels of impairment that would significantly hamper cognitive understanding of image content. Moreover, I ensured that each image was distorted by each distortion type over the entire range of quality. I also tried to keep the distribution of scores over the quality range as uniform as

Distortion	Parameter
JPEG2000	Bit rate
JPEG	Bit rate
White Noise	Noise variance
Gaussian Blur	Gaussian kernel variance
Fast-fading wireless	Receiver SNR

Table A.1: Distortion types and parameter used to control the degree of distortion. A JPEG2000 bitstream at a rate of 2.5 bits per pixel was transmitted over a simulated channel for the wireless channel distortion.

possible. In all, a total of 982 images, out of which 779 were distorted images and 203 were reference images, were evaluated in the study.

A.4 Details of the Experiments

A.4.1 Test Methodology

The experimental setup that I used was derived from standardized recommendations according to my specific needs and resources. In particular, I chose to use a single stimulus methodology in which the reference images were also evaluated for quality by the human assessors in the same experimental session as the test images. Since reference images were also evaluated by each subject in each session, a quality difference score can be derived for all distorted images. This approach is a midway approach between single stimulus methods and double stimulus methods, and allows me to use the database for training and validation of no-reference methods, as well as for full-reference methods.

A.4.2 Equipment and Display Configuration

The experiments were conducted using identical Microsoft Windows workstations. A web-based interface showing the image to be ranked and a Java scale-and-slider applet for assigning a quality score was used to mimic real-world quality judgement conditions as closely as possible. The workstations were placed in an office environment with normal indoor illumination levels. The display monitors were all 21-inch CRT monitor displaying at a resolution of 1024×768 pixels. Although the monitors were not calibrated, they were all approximately the same age, and set to the same display settings. Subjects were requested to view the monitors from an approximate viewing distance of 2-2.5 screen heights. The software slider included features that would not allow users to skip an image without making a judgement, and would also reset the slider to zero when a new image was displayed. Figure A.1 shows the user interface of the quality assessment software.

The experiments were conducted in seven sessions: two sessions for JPEG2000, two for JPEG, and one each for white noise, gaussian blur, and channel errors. Each session included the full set of reference images randomly placed among the distorted images. The number of images in each session are shown in Table A.2.

A.4.3 Human Subjects, Training, and Testing

The bulk of the subjects taking part in the study were recruited for course credit from the Digital Image and Video Processing (undergraduate)



Figure A.1: Interface for the subjective study.

Session	Number of images	Number of Subjects
JPEG2000 #1	116	29
JPEG2000 #2	111	25
JPEG #1	116	20
JPEG #2	117	20
White Noise	174	23
Gaussian Blur	174	24
Fast-fading wireless	174	20
Total	982	22.8 (average)

Table A.2: Subjective evaluation sessions: number of images in each session and the number of subjects participating in each session. The reference images were included in each session.

and the Digital Signal Processing (graduate) classes at the University of Texas at Austin. The subject pool consisted of mostly male students inexperienced with image quality assessment and image impairments. Due to resource constraints, the subjects were not screened for color blindness or vision problems, and their verbal expression of the soundness of their (corrected) vision was considered sufficient. I decided to compensate for this shortcoming by increasing the number of subjects ranking each image to about 25, whereas experts consider 15 subjects to be the minimal requirement [35]. Some expert viewers (members of LIVE) were also requested to participate in the study.

Each subject was individually briefed about the goal of the experiment, and given a demonstration of the experimental procedure. A short training showing the approximate range of quality of the images in each session was also presented to each subject. Generally, each subject participated in one session only. Subjects were shown images in a random order; the randomization was different for each subject. Subjects usually took about twenty minutes per session, including the time used for briefing and the training. This is well within the 30 minute limit considered appropriate by QA experts [35].

The subjects reported their judgments of quality by dragging a slider on a quality scale. The quality scale was unmarked numerically. It was divided into five equal portions, which were labeled with adjectives: “Bad”, “Poor”, “Fair”, “Good”, and “Excellent”. The position of the slider after the subject ranked the image was converted into a quality score by linearly mapping the entire scale to the interval $[1, 100]$, rounding to the nearest integer. In this

way, the raw quality scores consisted of integers in the range 1 – 100.

A.5 Processing of Raw Scores

The processing of the raw scores was adapted to the type of quality assessment algorithms that were to be trained and validated. The processing of raw scores was done to compute the mean opinion score (MOS) and the difference mean opinion score (DMOS) for each image. MOS was used for training and validation of NR methods, while DMOS was used for training and testing FR methods.

A.5.1 Outlier Detection and Subject Rejection

A simple outlier detection and subject rejection algorithm was chosen. A raw score for an image was considered to be an outlier if it was outside an interval of width Δ standard deviations about the mean score (computed from the raw scores from all subjects) for that image, and all quality evaluations of a subject were rejected if more than R of his evaluations were outliers. This outlier rejection algorithm was run twice. A minimization algorithm was run for each distortion type that varied Δ and R to minimize the average width of the 95% confidence interval for each distortion type. A total of three subjects from all sessions were rejected.

A.5.2 MOS Scores

For the calculation of MOS scores, the raw scores (after outlier removal and subject rejection) r_{ij} for the i -th subject and j -th image were converted into Z-scores [117]:

$$z_{ij} = \frac{r_{ij} - \bar{r}_i}{\sigma_i} \quad (\text{A.1})$$

where \bar{r}_i is the mean of the raw score over all images ranked by the subject i , and σ_i is the standard deviation. The Z-scores were then re-scaled to 1 – 100 range before being averaged across subjects to give the MOS scores.

A.5.3 DMOS Scores

The procedure of calculating the DMOS score was similar to that for the MOS scores, except that the raw scores were first converted to difference quality scores between the distorted images and their corresponding reference images:

$$d_{ij} = r_{iref(j)} - r_{ij} \quad (\text{A.2})$$

where $r_{iref(j)}$ denotes the raw quality score assigned by the i -th subject to the reference image corresponding to the j -th distorted image. Outliers were then removed (a total of four subjects were rejected in all sessions), and Z-scores were calculated using d_{ij} with (A.1), linearly re-scaled to 1 – 100 range and then averaged across subjects to yield the DMOS scores.

A.6 Results

In this section, I will discuss some characteristics of the study and the quality scores.

A.6.1 Sample Images

Figure A.2 show some sample images from the twenty-nine reference images used to derive the database.

A.6.2 Confidence Interval and RMS Error

The confidence interval is an indication of the confidence in the estimate of the MOS or DMOS. For a normally distributed process (with 25 sample points), the 95% confidence interval about the mean μ , $[\mu - \delta, \mu + \delta]$ is [63]:

$$\delta \approx \frac{2.1\sigma}{\sqrt{N}} \quad (\text{A.3})$$

where σ is the standard deviation of the samples, and N is the number of samples used in the calculation of the mean. The average width of the confidence interval (average δ for all images) is shown in Table A.3 for MOS and DMOS.

The RMS error between the MOS and human subjects is an indication of the spread of the human quality score of an image about its MOS value. It indicates the degree of disagreement between an ideal quality predictor that exactly predicts the MOS/DMOS (also known as the *null* predictor) and individual human assignments. Table A.3 gives the RMSE between the null model (MOS and DMOS) and human subjects.



Figure A.2: Sample pictures from the set used to derive the database.

δ (MOS)	2.53
δ (DMOS)	2.74
RMSE (MOS)	5.49
RMSE (DMOS)	5.92

Table A.3: Average size of the 95% confidence interval about the mean μ , $[\mu - \delta, \mu + \delta]$ and RMSE for MOS and DMOS on a 1-100 quality scale.

A.6.3 Dependence of Quality on Distortion Parameters

It is interesting to note the relationship between image quality and the distortion parameters. Figures A.3 through A.5 graphically show these relations.

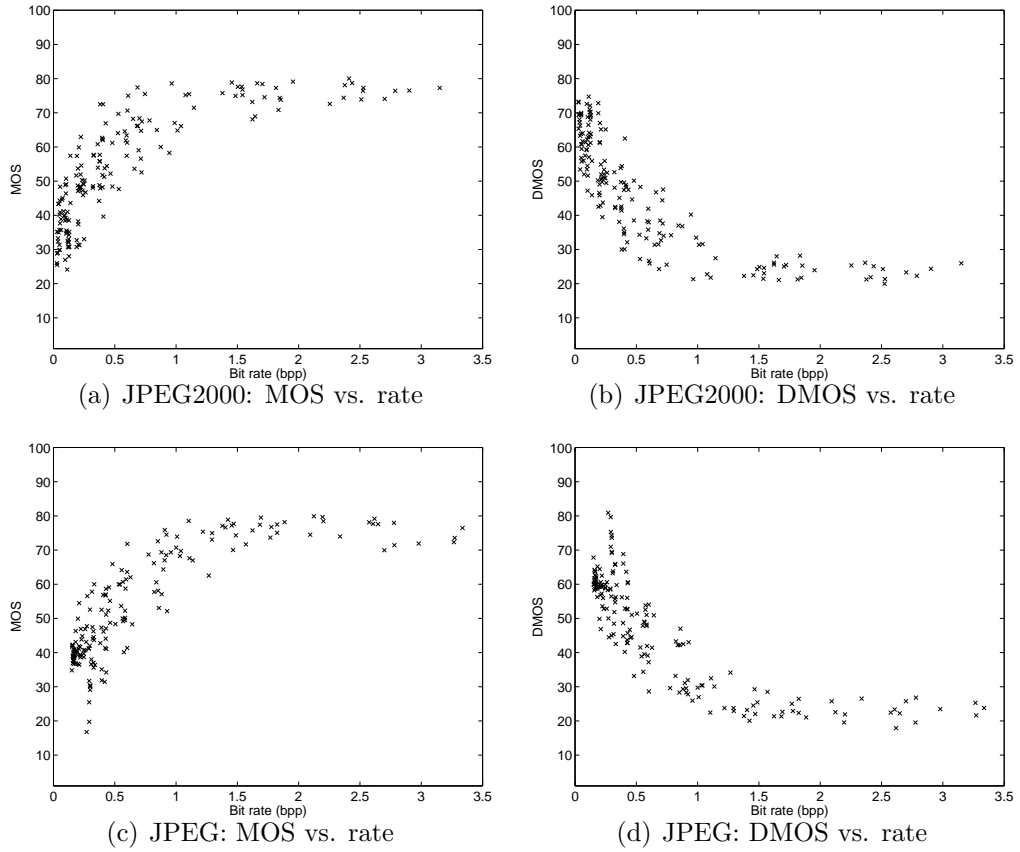


Figure A.3: Dependence of quality on distortion parameters for different distortion types: JPEG2000 and JPEG.

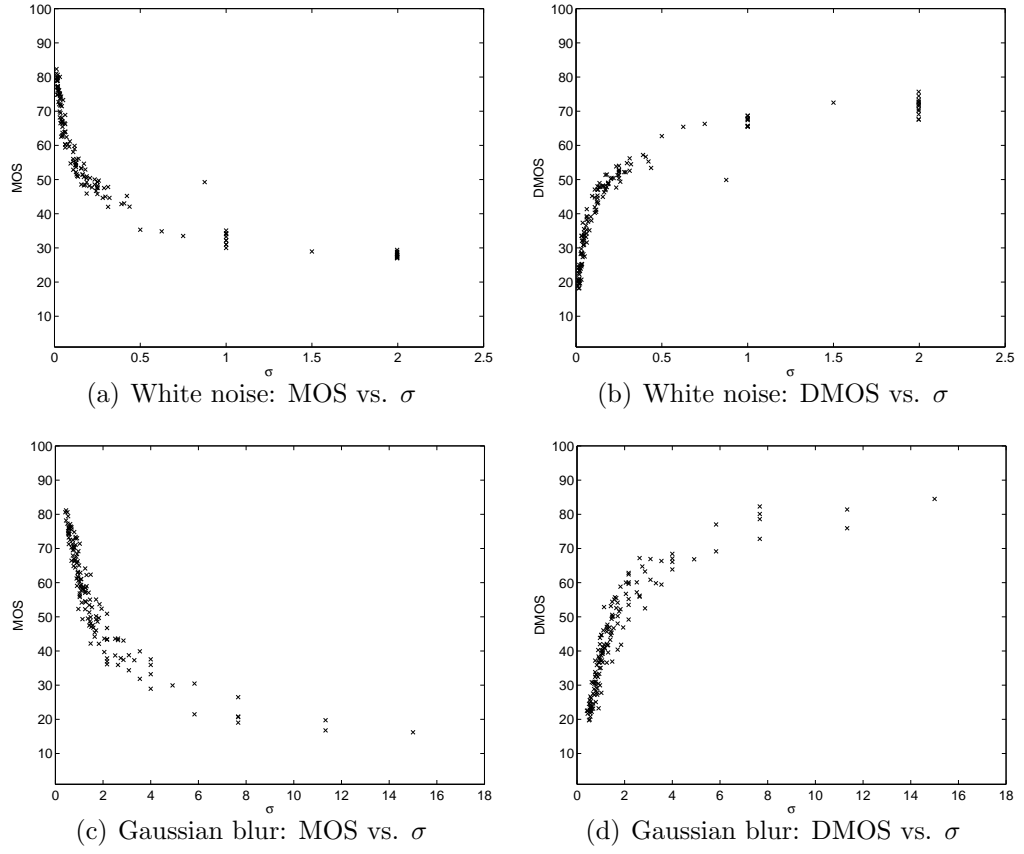


Figure A.4: Dependence of quality on distortion parameters for different distortion types: White Noise and Gaussian Blur.

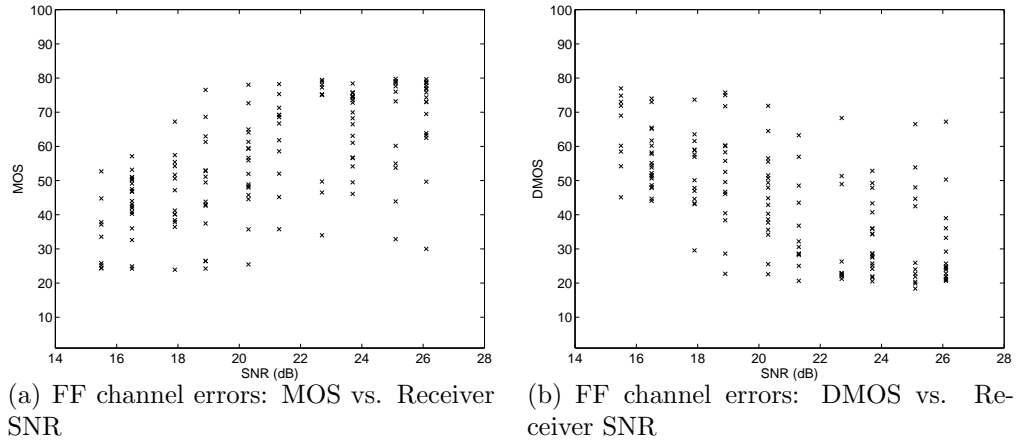


Figure A.5: Dependence of quality on distortion parameters: bit errors in the JPEG2000 bit stream in a fast-fading Rayleigh channel.

Appendix B

Similarities between IFC and HVS based FR QA Methods

In this appendix I will quantify the similarities between the scalar GSM version of the IFC of (4.11) and the HVS based QA assessment method shown in Figure 4.5. The model in Figure 4.5 is based on calculating MSE in the perceptual space and then processing it further to yield the final perceptual distortion measure. Here I will only deal with coefficients in one subband and with the scalar GSM model.

I start by giving the formulation for the divisive normalization stage, which divides the input by its localized average. Considering the input to the squaring block, this turns out to be normalization by the estimated local variance of the input of the squaring block:

$$W_i = C_i^2 \left(\frac{1}{K} \sum_{j \in \mathcal{N}(i)} C_j^2 \right)^{-1} \approx \frac{C_i^2}{s_i^2} = U_i^2 \quad (\text{B.1})$$

$$W'_i = D_i^2 \left(\frac{1}{K} \sum_{j \in \mathcal{N}(i)} D_j^2 \right)^{-1} \approx \frac{D_i^2}{g_i^2 s_i^2 + \sigma_V^2} \quad (\text{B.2})$$

Here I have assumed that $s_j \approx s_i$ for $j \in \mathcal{N}(i)$, that is, the variance is approximately constant over the K pixels neighborhood of i , which I denote

by $\mathcal{N}(i)$. Also note that the term inside the parentheses in an estimate of the conditional local variance of C (or D) at i given $S_i = s_i$, which could be approximated by the actual value when considered in concert with the expectation operator below; I make this approximation here in order to avoid clutter. I have also assumed, without loss of generality, that $\mathbb{E}[U_i^2] = \sigma_U^2 = 1$, since any non-unity variance of \mathcal{U} could be absorbed into \mathcal{S} . The MSE between W_i and W'_i given $S_i = s_i$ could now be analyzed:

$$\text{MSE}(W_i, W'_i | s_i) = \mathbb{E}[(W'_i - W_i)^2 | s_i] \quad (\text{B.3})$$

$$\approx \mathbb{E} \left[\left(\frac{D_i^2}{g_i^2 s_i^2 + \sigma_V^2} - U_i^2 \right)^2 | s_i \right] \quad (\text{B.4})$$

$$= \mathbb{E} \left[\frac{(V_i^2 + 2g_i C_i V_i - \sigma_V^2 U_i^2)^2}{(g_i^2 s_i^2 + \sigma_V^2)^2} | s_i \right] \quad (\text{B.5})$$

where I have used $D_i = g_i C_i + V_i$ and that given $S_i = s_i$, $C_i = s_i U_i$. Expanding the above expression and taking expectation, and using independence between \mathcal{U} and \mathcal{V} , the fact that \mathcal{C} , \mathcal{U} , and \mathcal{V} are all zero-mean, and the fact that for zero-mean Gaussian variables $\mathbb{E}[X^4] = 3\sigma^4$, where σ^2 is the variance of X , we get:

$$\text{MSE}(W_i, W'_i | s_i) \approx \frac{4\sigma_V^2}{g_i^2 s_i^2 + \sigma_V^2} \quad (\text{B.6})$$

The goal of this derivation is to compare the information fidelity crite-

tion of (4.11) and HVS based MSE criterion:

$$I(C^N; D^N | s^N) = \frac{1}{2} \sum_{i=1}^N \log_2 \left(1 + \frac{g_i^2 s_i^2}{\sigma_V^2} \right) \quad (\text{B.7})$$

$$= -\frac{1}{2} \sum_{i=1}^N \log_2 \left(\frac{\sigma_V^2}{g_i^2 s_i^2 + \sigma_V^2} \right) \quad (\text{B.8})$$

$$\approx -\frac{1}{2} \sum_{i=1}^N (\log_2(\text{MSE}(W_i, W'_i | s_i)) - \log_2 4) \quad (\text{B.9})$$

Hence we have an approximate relation between the information fidelity criterion and the HVS based MSE:

$$I(C^N; D^N | s^N) \approx \alpha \sum_{i=1}^N \log_2(\text{MSE}(W_i, W'_i | s_i)) + \beta \quad (\text{B.10})$$

where α and β are constants.

Bibliography

- [1] *Special Issue on the H.264/AVC Video Coding Standard*, volume 13. IEEE Trans. Circuits and Systems for Video Tech., July 2003.
- [2] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102, 1974.
- [3] Ismail Avcibas, Bülent Sankur, and Khalid Sayood. Statistical evaluation of image quality measures. *Journal of Electronic Imaging*, 11(2):206–23, April 2002.
- [4] Peter G. J. Barten. Evaluation of subjective image quality with the square-root integral method. *Journal of Optical Society of America*, 7(10):2024–2031, October 1990.
- [5] A. C. Bovik and S. Liu. DCT-domain blind measurement of blocking artifacts in DCT-coded images. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, volume 3, pages 1725 –1728, May 2001.
- [6] Alan C. Bovik, Marianna Clark, and Wilson S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(1):55–73, January 1990.

- [7] A. P. Bradley. A wavelet visible difference predictor. *IEEE Trans. Image Processing*, 5(8):717–730, May 1999.
- [8] R. W. Buccigrossi and E. P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. Image Processing*, 8(12):1688–1701, December 1999.
- [9] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Trans. Communications*, 31:532–540, April 1983.
- [10] H. Choi and R. G. Baraniuk. Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Trans. Image Processing*, 10(9):1309–1321, September 2001.
- [11] C. H. Chou and Y. C. Li. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Trans. Circuits and Systems for Video Tech.*, 5(6):467–476, December 1995.
- [12] A. Cohen, I. Daubechies, and J. C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Commun. Pure Appl. Math.*, 45:485–560, 1992.
- [13] L. K. Cormack. Computational models of early human vision. In A. Bovik, editor, *Handbook of Image and Video Processing*. Academic Press, May 2000.

- [14] P. Corriveau, Christina Gojmerac, Bronwen Hughes, and Lew Stelmach. All subjective scales are not created equal: The effects of context on different scales. *Signal Processing*, 77:1–9, 1999.
- [15] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.
- [16] S. Daly. The visible difference predictor: An algorithm for the assessment of image fidelity. In *Proc. SPIE*, volume 1616, pages 2–15, 1992.
- [17] S. Daly. The visible difference predictor: An algorithm for the assessment of image fidelity. In Andrew B. Watson, editor, *Digital images and human vision*, pages 179–206. The MIT Press, Cambridge, Massachusetts, 1993.
- [18] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik. Image quality assessment based on a degradation model. *IEEE Trans. Image Processing*, 4(4):636–650, April 2000.
- [19] Dawei W. Dong and Joseph J. Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6(3):345–358, April 1995.
- [20] D. L. Donoho and A. G. Felisia. Can recent innovations in harmonic analysis ‘explain’ key findings in natural image statistics. *Vision Research*, 12(3):371–393, 2001.

- [21] M. P. Eckert and A. P. Bradley. Perceptual quality metrics applied to still image compression. *Signal Processing*, 70(3):177–200, November 1998.
- [22] P. Corriveau *et al.* Video quality experts group: Current results and future directions. *Proc. SPIE Visual Comm. and Image Processing*, 4067, June 2000.
- [23] A. M. Eskicioglu and P. S. Fisher. Image quality measures and their performance. *IEEE Trans. Communications*, 43(12):2959–2965, December 1995.
- [24] M. C. Q. Farias, S. K. Mitra, M. Carli, and A. Neri. A comparison between an objective quality measure and the mean annoyance values of watermarked videos. In *Proc. IEEE Int. Conf. Image Proc.*, September 2002.
- [25] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of Optical Society of America*, 4(12):2379–2394, 1987.
- [26] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13:891–906, 1991.
- [27] D. R. Fuhrmann, J. A. Baro, and J. R. Cox Jr. Experimental evaluation

- of psychophysical distortion metrics for JPEG-encoded images. *Journal of Electronic Imaging*, 4:397–406, October 1995.
- [28] J. A. Garcia, J. Fdez-Valdivia, R. Rodriguez-Sanchez, and X. R. Fdez-Vidal. Performance of the Kullback-Leibler information gain for predicting image fidelity. In *Proc. IEEE International Conference on Pattern Recognition*, volume 3, pages 843–848, 2002.
 - [29] P. Gastaldo, S. Rovetta, and R. Zunino. Objective assessment of MPEG-video quality: a neural-network approach. In *Proc. IJCNN*, volume 2, pages 1432–1437, 2001.
 - [30] W. S. Geisler and D. G. Albrecht. Visual cortex neurons in monkeys and cats: Detection, discrimination, and identification. *Visual Neuroscience*, 14:897–919, 1997.
 - [31] Peter J. B. Hancock, Roland J. Baddeley, and Leslie S. Smith. The principal components of natural images. *Network: Computation in Neural Systems*, 3:61–70, 1992.
 - [32] David J. Heeger and Patrick C. Teo. A model of perceptual image fidelity. In *Proc. IEEE Int. Conf. Image Proc.*, pages 343–345, 1995.
 - [33] Jinggang Huang and David Mumford. Statistics of natural images and models. In *Proc. IEEE CVPR*, pages 1541–1547, June 1999.
 - [34] ITU-R Rec. BT. 1082-1. *Studies toward the unification of picture assessment methodology*.

- [35] ITU-R Rec. BT. 500-11. *Methodology for the Subjective Assessment of the Quality for Television Pictures*.
- [36] T. J. W. M. Jannsen and F. J. J. Blommaert. Predicting the usefulness and naturalness of color reproductions. *Journal of Imaging Science and Technology*, 44(2):93–104, 2000.
- [37] Thomas Kailath, Ali H. Sayed, and Babak Hassibi. *Linear Estimation*. Prentice Hall, 2000.
- [38] S. A. Karunasekera and N. G. Kingsbury. A distortion measure for image artifacts based on human visual sensitivity. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, volume 5, pages 117–120, 1994.
- [39] S. A. Karunasekera and N. G. Kingsbury. A distortion measure for blocking artifacts in images based on human visual sensitivity. *IEEE Trans. Image Processing*, 4(6):713–724, June 1995.
- [40] Vishwakumara Kayargadde and Jean-Bernard Martens. Perceptual characterization of images degraded by blur and noise: experiments. *Journal of Optical Society of America*, 13(6):1166–1177, 1996.
- [41] Vishwakumara Kayargadde and Jean-Bernard Martens. Perceptual characterization of images degraded by blur and noise: model. *Journal of Optical Society of America*, 13(6):1178–1188, 1996.

- [42] Nick Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Applied and Computational Harmonic Analysis*, 10: 234–253, 2001.
- [43] T. D. Kite, N. Damera-Venkata, B. L. Evans, and A. C. Bovik. A high quality, fast inverse halftoning algorithm for error diffused halftones. In *Proc. IEEE Int. Conf. Image Proc.*, volume 2, pages 59–63, October 1998.
- [44] M. Knee. A robust, efficient and accurate signal-ended picture quality measure for MPEG-2. available at <http://www-ext.crc.ca/vqeg/frames.html>, 2001.
- [45] Y. K. Lai and C.-C. J. Kuo. A Haar wavelet approach to compressed image quality measurement. *Journal of Visual Communication and Image Representation*, 11:17–40, March 2000.
- [46] Edmund Y. Lam and Joseph W. Goodman. A mathematical analysis of the DCT coefficient distributions for images. *IEEE Trans. Image Processing*, 9(10):1661–66, October 2000.
- [47] Ann B. Lee, David Mumford, and Jinggang Huang. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision*, 41(1/2):35–59, 2001.
- [48] S. Lee, M. S. Pattichis, and A. C. Bovik. Foveated video quality assessment. *IEEE Trans. Multimedia*, 4(1):129–132, March 2002.

- [49] B. Li, G. W. Meyer, and R. V. Klassen. A comparison of two image quality models. In *Proc. SPIE*, volume 3299, pages 98–109, 1998.
- [50] Xin Li. Blind image quality assessment. In *Proc. IEEE Int. Conf. Image Proc.*, Rochester, September 2002.
- [51] J. Liu and P. Moulin. Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients. *IEEE Trans. Image Processing*, 10(11):1647–1658, November 2001.
- [52] Vhi-Min Liu, Jine-Yi Lin, and Kuo-Guan Chung-Neng Wang. Objective image quality measure for block-based DCT coding. *IEEE Trans. Consumer Electronics*, 43(3):511–516, June 1997.
- [53] J. Lubin. The use of psychophysical data and models in the analysis of display system performance. In Andrew B. Watson, editor, *Digital images and human vision*, pages 163–178. The MIT Press, Cambridge, Massachusetts, 1993.
- [54] J. Lubin. A visual discrimination mode for image system design and evaluation. In E. Peli, editor, *Visual Models for Target Detection and Recognition*, pages 207–220. World Scientific Publishers, Singapore, 1995.
- [55] J. Malo, A. M. Pons, and J. M. Artigas. Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain. *Image and Vision Computing*, 15:535–548, 1997.

- [56] J. L. Mannos and D. J. Sakrison. The effects of a visual fidelity criterion on the encoding of images. *IEEE Trans. Information Theory*, 4:525–536, 1974.
- [57] J.-B. Martens and L. Meesters. Image dissimilarity. *Signal Processing*, 70(3):155–176, November 1998.
- [58] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. Perceptual blur and ringing metrics: Application to JPEG2000. *Signal Processing: Image Communication*, 19(2):163–172, February 2004.
- [59] A. Mayache, T. Eude, and H. Cherifi. A comparison of image quality models and metrics based on human visual sensitivity. In *Proc. IEEE Int. Conf. Image Proc.*, pages 409–413, 1998.
- [60] Lydia Meesters and Jean-Bernard Martens. A single-ended blockiness measure for JPEG-coded images. *Signal Processing*, 82:369–387, 2002.
- [61] M. Kivanç Mihçak, Igor Kozintsev, Kannan Ramachandran, and Pierre Moulin. Low-complexity image denoising based on statistical modeling of wavelet coefficients. *IEEE Signal Processing Letters*, 6(12):300–303, December 1999.
- [62] M. Miyahara, K. Kotani, and V. R. Algazi. Objective Picture Quality Scale (PQS) for image coding. *IEEE Trans. Communications*, 46(9):1215–1225, September 1998.

- [63] Douglas C. Montgomery and George C. Runger. *Applied Statistics and Probability for Engineers*. Wiley-Interscience, 1999.
- [64] S. H. Oguz, Y. H. Hu, and T. Q. Nguyen. Image coding ringing artifact reduction using morphological post-filtering. In *1998 IEEE Second Workshop on Multimedia Signal Processing*, pages 628–633, 1998.
- [65] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7:333–339, 1996.
- [66] W. Osberger, N. Bergmann, and A. Maeder. An automatic image quality assessment technique incorporating high level perceptual factors. In *Proc. IEEE Int. Conf. Image Proc.*, pages 414–418, 1998.
- [67] T. N. Pappas and R. J. Safranek. Perceptual criteria for image quality evaluation. In A. Bovik, editor, *Handbook of Image & Video Proc.* Academic Press, 2000.
- [68] E. Peli. Contrast in complex images. *Journal of Optical Society of America*, 7(10):2032–2040, October 1990.
- [69] W. B. Pennebaker and J. L. Mitchell. *JPEG: Still Image Data Compression Standard*. Kluwer Academic Publishers, 1992.
- [70] A. M. Pons, J. Malo, J. M. Artigas, and P. Capilla. Image quality metric based on multidimensional contrast perception models. *Displays*, 20:93–110, 1999.

- [71] J. Portilla, M. Wainwright V. Strela, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans. Image Processing*, 12(11):1338–1351, November 2003.
- [72] Javier Portilla and Eero P. Simoncelli. Image denoising via adjustment of wavelet coefficient magnitude correlation. In *Proc. IEEE Int. Conf. Image Proc.*, September 2000.
- [73] Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–71, 2000.
- [74] Javier Portilla, Vasily Strela, Martin J. Wainwright, and Eero P. Simoncelli. Adaptive weiner denoising using a gaussian scale mixture model in the wavelet domain. In *Proc. IEEE Int. Conf. Image Proc.*, October 2001.
- [75] U. Rajashekar, L. K. Cormack, A. C. Bovik, and W. S. Geisler. Image properties that draw fixation [abstract]. *Journal of Vision*, 2(7):730a, 2002. <http://journalofvision.org/2/7/730/>, DOI 10.1167/2.7.730.
- [76] Pamela Reinagel and Anthony M. Zador. Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, 10:1–10, 1999.
- [77] J. K. Romberg, H. Choi, and R. Baraniuk. Bayesian tree-structured image modeling using wavelet-domain hidden markov models. *IEEE Trans. Image Processing*, 10(7):1056–1068, July 2001.

- [78] Daniel L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, November 1994.
- [79] Daniel L. Ruderman. The origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, December 1997.
- [80] R. J. Safranek. A comparison of the coding efficiency of perceptual models. In *Proc. SPIE*, volume 2411, pages 83–91, 1995.
- [81] R. J. Safranek and J. D. Johnston. A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 1945–1948, May 1989.
- [82] A. Said and W. A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. Circuits and Systems for Video Tech.*, 6(3):243–250, June 1996.
- [83] Sarnoff Corporation. Sarnoff JND Vision Model Algorithm Description and Testing. Download available: <http://www-ext.crc.ca/vqeg/downloads.html>, 1997.
- [84] Sarnoff Corporation. JND-Metrix Technology. Evaluation Version available: http://www.sarnoff.com/products_services/video_vision/jndmetrix/downloads.asp, 2003.

- [85] Heiko Schwarz and Thomas Wiegand. The emerging JVT/H.26L video coding standard. In *Proc. of IBC*, Amsterdam, Neatherlands, September 2002.
- [86] Ivan W. Selesnick. Hilbert transform pairs of wavelet bases. *IEEE Signal Processing Letters*, 8(6):170–173, 2001.
- [87] Ivan W. Selesnick. The design of approximate hilbert transform pairs of wavelet bases. *IEEE Trans. Signal Processing*, 50(5):1144–1152, 2002.
- [88] J. M. Shapiro. Embedded image coding using zerotrees of wavelets coefficients. *IEEE Trans. Signal Processing*, 41:3445–3462, December 1993.
- [89] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Trans. Image Processing*, December 2003. Submitted.
- [90] H. R. Sheikh, A. C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Processing*, March 2004. Accepted.
- [91] H. R. Sheikh, L. Cormack, and A. C. Bovik. Blind quality assessment for JPEG2000 compressed images using natural scene statistics. In *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, November 2003.
- [92] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. Blind quality assessment for JPEG2000 compressed images. In *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, November 2002.

- [93] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. LIVE image quality assessment database. Available at <http://live.ece.utexas.edu/research/quality>, 2003.
- [94] Hamid R. Sheikh, Alan C. Bovik, and Lawrence Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Trans. Image Processing*, December 2003. Revised.
- [95] D. A. Silverstein and J. E. Farrel. The relationship between image fidelity and image quality. In *Proc. IEEE Int. Conf. Image Proc.*, volume 1, pages 881–884, 1996.
- [96] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, November 1997.
- [97] E. P. Simoncelli. Modeling the joint statistics of images in the wavelet domain. In *Proc. SPIE*, volume 3813, pages 188–195, July 1999.
- [98] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. IEEE Int. Conf. Image Proc.*, pages 444–447, October 1995.
- [99] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multi-scale transforms. *IEEE Trans. Information Theory*, 38: 587–607, 1992.

- [100] Eero P. Simoncelli. Local analysis of visual motion. In Leo M. Chalupa and John S. Werner, editors, *The Visual Neurosciences*. MIT Press, 2003.
- [101] Eero P. Simoncelli. Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13, April 2003.
- [102] Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–216, May 2001.
- [103] Eero P. Simoncelli and Ordelia Schwartz. Modeling surround suppression in V1 neurons with a statistically-derived normalization model. In *Proc. NIPS-98, Advances in Neural Information Processing Systems*, pages 153–159, December 1998.
- [104] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18:17–33, 2003.
- [105] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley Cambridge Pr, 1993.
- [106] Vasily Strela, Javier Portilla, and Eero Simoncelli. Image denoising using a local Gaussian Scale Mixture model in the wavelet domain. *Proc. SPIE*, 4119:363–371, 2000.
- [107] O. Sugimoto, R. Kawada, M. Wada, and S. Matsumoto. Objective measurement scheme for perceived picture quality degradation caused by

- MPEG encoding without any reference pictures. *Proc. SPIE*, 4310:932–939, 2001.
- [108] K. T. Tan and M. Ghanbari. Frequency domain measurement of blockiness in MPEG-2 coded video. In *Proc. IEEE Int. Conf. Image Proc.*, volume 3, pages 977–980, September 2000.
 - [109] K. T. Tan and M. Ghanbari. A multi-metric objective picture-quality measurement model for MPEG video. *IEEE Trans. Circuits and Systems for Video Tech.*, 10(7):1208–1213, October 2000.
 - [110] K. T. Tan, M. Ghanbari, and D. E. Pearson. An objective measurement tool for MPEG video quality. *Signal Processing*, 70(3):279–294, November 1998.
 - [111] D. S. Taubman and M. W. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Kluwer Academic Publishers, 2001.
 - [112] Patrick C. Teo and David J. Heeger. Perceptual image distortion. In *Proc. IEEE Int. Conf. Image Proc.*, pages 982–986, 1994.
 - [113] Patrick C. Teo and David J. Heeger. Perceptual image distortion. In *Proc. SPIE*, volume 2179, pages 127–141, 1994.
 - [114] M. G. A. Thomson. Beats, kurtosis and visual coding. *Network: Computation in Neural Systems*, 12:271–287, 2001.

- [115] C. J. van den Branden Lambrecht, D. M. Costantini, G. L. Sicuranza, and M. Kunt. Quality assessment of motion rendition in video coding. *IEEE Trans. Circuits and Systems for Video Tech.*, 9(5):766–782, August 1999.
- [116] Christian J. van den Branden Lambrecht. A working spatio-temporal model of the human visual system for image restoration and quality assessment applications. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 2291–2294, 1996.
- [117] A. M. van Dijk, J. B. Martens, and A. B. Watson. Quality assessment of coded images using numerical category scaling. *Proc. SPIE*, 2451: 90–101, March 1995.
- [118] Antoon M. van Dijk and Jean-Bernard Martens. Subjective quality assessment of compressed images. *Signal Processing*, 58:235–252, 1997.
- [119] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. In *Proceedings of the Royal Society of London, Series B*, pages 265:359–366, 1998.
- [120] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment. <http://www.vqeg.org/>, March 2000.

- [121] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II. *ftp://ftp.its.bldrdoc.gov/dist/ituvidq/frtv2_final_report/VQEGII_Final_Report.pdf*, August 2003.
- [122] Martin J. Wainwright and Eero P. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. *Advances in Neural Information Processing Systems*, 12:855–861, 2000.
- [123] Martin J. Wainwright, Eero P. Simoncelli, and Alan S. Wilsky. Random cascades of gaussian scale mixtures and their use in analyzing and modeling natural images. In *Proceedings of the 45th Annual Meeting of the SPIE*, August 2000.
- [124] Martin J. Wainwright, Eero P. Simoncelli, and Alan S. Wilsky. Random cascades of gaussian scale mixtures and their use in modeling natural images with applications to denoising. In *Proc. IEEE Int. Conf. Image Proc.*, September 2000.
- [125] Martin J. Wainwright, Eero P. Simoncelli, and Alan S. Wilsky. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied and Computational Harmonic Analysis*, 11:89–123, 2001.
- [126] B. A. Wandell. *Foundations of Vision*. Sinauer Associates, Inc., 1995.

- [127] Z. Wang and A. C. Bovik. Embedded foveation image coding. *IEEE Trans. Image Processing*, 10(10):1397–1410, October 2001.
- [128] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, March 2002.
- [129] Z. Wang, A. C. Bovik, and B. L. Evans. Blind measurement of blocking artifacts in images. In *Proc. IEEE Int. Conf. Image Proc.*, volume 3, pages 981–984, September 2000.
- [130] Z. Wang, A. C. Bovik, and L. Lu. Why is image quality assessment so difficult? In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Orlando, May 2002.
- [131] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error measurement to structural similarity. *IEEE Trans. Image Processing*, January 2004. To appear.
- [132] Z. Wang, L. Lu, and Alan C. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, 19(2), February 2004.
- [133] Z. Wang, H. R. Sheikh, and A. C. Bovik. No-reference perceptual quality assessment of JPEG compressed images. In *Proc. IEEE Int. Conf. Image Proc.*, Rochester, September 2002.

- [134] Z. Wang, H. R. Sheikh, and A. C. Bovik. Objective video quality assessment. In B. Furht and O. Marques, editors, *The Handbook of Video Databases: Design and Applications*. CRC Press, 2003.
- [135] Zhou Wang and Eero Simoncelli. Local phase coherence and the perception of blur. In *Adv. Neural Information Processing Systems*. 2004. Oral presentation in NIPS 2003.
- [136] Zhou Wang and Eero P. Simoncelli. Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics. In *Proc. SPIE*, volume 5292, 2004.
- [137] A. B. Watson. DCT quantization matrices visually optimized for individual images. In *Proc. SPIE*, volume 1913, pages 202–216, 1993.
- [138] A. B. Watson. DCTune: A technique for visual optimization of dct quantization matrices for individual images. In *Society for Information Display Digest of Technical Papers*, volume XXIV, pages 946–949, 1993.
- [139] A. B. Watson, J. Hu, and J. F. III. McGowan. DVQ: A digital video quality metric based on human vision. *Journal of Electronic Imaging*, 10(1):20–29, 2001.
- [140] A. B. Watson and L. Kreslake. Measurement of visual impairment scales for digital video. In *Human Vision, Visual Processing, and Digital Display*, *Proc. SPIE*, volume 4299, 2001.

- [141] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor. Visibility of wavelet quantization noise. *IEEE Trans. Image Processing*, 6(8):1164–1175, August 1997.
- [142] Andrew B. Watson. The cortex transform: rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing*, 39:311–327, 1987.
- [143] Andrew B. Watson and Jesus Malo. Video quality measures based on the standard spatial observer. In *Proc. IEEE Int. Conf. Image Proc.*, 2002.
- [144] Andrew B. Watson and J. A. Solomon. Model of visual contrast gain control and pattern masking. *Journal of Optical Society of America*, 14(9):2379–2391, 1997.
- [145] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf. An objective video quality assessment system based on human perception. *Proc. SPIE*, 1913:15–26, 1993.
- [146] S. Winkler. Issues in vision modeling for perceptual video quality assessment. *Signal Processing*, 78:231–252, 1999.
- [147] S. Winkler. A perceptual distortion metric for digital color video. *Proc. SPIE*, 3644:175–184, 1999.

- [148] S. Wolf and M. H. Pinson. Spatio-temporal distortion metrics for in-service quality monitoring of any digital video system. *Proc. SPIE*, 3845: 266–277, 1999.
- [149] J. W. Woods and S. D. O’Neil. Subband coding of images. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 34(5):1278–1288, October 1986.
- [150] H. R. Wu and M. Yuen. A generalized block-edge impairment metric for video coding. *IEEE Signal Processing Letters*, 4(11):317–320, November 1997.
- [151] W. Xu and G. Hauske. Picture quality evaluation based on error segmentation. *Proc. SPIE*, 2308:1454–1465, 1994.
- [152] T. Yamashita, M. Kameda, and M. Miyahara. An objective picture quality scale for video images (PQS_{video}) - definition of distortion factors. *Proc. SPIE*, 4067:801–809, 2000.
- [153] Seungjoon Yang, Yu-Hen Hu, Truong Q. Nguyen, and Damon L. Tull. Maximum-likelihood parameter estimation for image ringing-artifact removal. *IEEE Trans. Circuits and Systems for Video Tech.*, 11(8):963–973, August 2001.
- [154] Susu Yao, Weisi Lin, Zhongkang Lu, EePing Ong, and Minoru Etoh. Objective quality assessment for compressed video. In *Proc. IEEE Int. Sym. Circuits and Systems*, 2003.

- [155] Z. Yu, H. R. Wu, S. Winkler, and T. Chen. Vision-Model-Based impairment metric to evaluate blocking artifact in digital video. *Proceedings of the IEEE*, 90(1):154–169, January 2002.
- [156] M. Yuen and H. R. Wu. A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal Processing*, 70(3):247–278, November 1998.

Vita

Hamid Rahim Sheikh was born in Lahore, Pakistan, in December of 1974. He completed his Bachelor's degree in Electrical Engineering from the University of Engineering and Technology, Lahore, Pakistan, in 1998 and worked for a few months for And-Or Logic, Islamabad, Pakistan, before joining The University of Texas at Austin in Fall 1999. He joined the Laboratory for Image and Video Engineering (LIVE) in January 2000 as a graduate research assistant. Hamid completed his MS in May 2001 and started his PhD in Fall 2001. His MS work was on real-time implementation of foveated video coding. He interned at Texas Instruments, Dallas, Texas, in summer 2001, working with their Imaging Business Unit on video codecs for TI's Digital Still Camera (DSC) line of processors. His research interests include full-reference and no-reference quality assessment, application of natural scene statistics models and human visual system models for solving image and video processing problems, and image and video codecs and their embedded implementation.

He enjoys translating Urdu poetry into English, some of which could be found somewhere on the web.

Permanent address: 151 Ataturk Block
New Garden Town
Lahore 54600
Pakistan.
Search the web for “Hamid Rahim Sheikh” for a
quicker contact.

This dissertation was typeset with L^AT_EX[†] by the author.

[†]L^AT_EX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth’s T_EX Program.