UNIVERSITY OF CALIFORNIA SAN DIEGO

ON THE RATE & DISTORTION: CONFORMITY WITH THE STATISTICS OF NATURAL IMAGES AND VISUAL PERCEPTION IN HUMANS

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Koohyar Minoo

Committee in charge:

Professor Truong Nguyen, Chair
Professor Gert Cauwenberghs
Professor Pamela Cosman
Professor David J. Kreigman
Professor Bhaskar Rao

2008

The Dissertation of Koohyar Minoo is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2008

DEDICATION

To my loving parents.

To my beloved wife and daughter.

To my gracious teachers and mentors.

And to all the wonderful people who have touched my life with grace.

EPIGRAPH

When, into the mirror of the cup, the reflection of Thy face fell,

From the laughter of wine, into the crude desire of the cup, the Aref fell.

With that splendor that in the mirror, the beauty of Thy face made,

All this picture into the mirror of fancy fell.

All this reflection of wine and varied picture that have appeared

Is a splendor of the face of the Said that, into cup fell.

Hafez Shirazi,

Original Translation by Henry Wilberforce Clarke (1840-1905).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AWGN:	Additive White Gaussian Noise
BRE:	Basic Rate Entity
CABAC:	Context Adaptive Binary Arithmetic Coding
CBO:	Current Basic-block Only
CBP:	Coded Block Pattern
CPB:	Current and Pas Basic-blocks
CSF:	Contrast Sensitivity Function
DCT:	Discrete Cosine Transform
DMOS:	Differential Mean Opinion Score
EOB:	End Of Block
EOP:	Expected Optimal Parameter
FCD:	Feature (or Channel) Decomposition
FFRCE:	Fast Fading Rayleigh Channel Error
FFT:	Fast Fourier Transform
FRIQ:	Full Reference Image Quality
GBL:	Gaussian Blur
GOP:	Group Of Pictures
HVS:	Human Visual System
IIJND:	Independent Just Noticeable Distortion
IIND:	Intensity Independent Noticeable Distortion
IIUD:	Intensity Independent Un-noticeable Distortion
JND:	Just Noticeable Differences (Distortion)
LEV:	Local Error Visibility

LGN:	Lateral Geniculate Nucleus
LoG:	Laplacian of Gaussian
LSS:	Local Structural Similarity
LTS:	Local Texture Spread
MAE:	Mean Absolute Error
MAV:	Mean Absolute Value
MIIND:	Minimum Intensity Independent Noticeable Distortion
MIIUD:	Maximum Intensity Independent Un-noticeable Distortion
MLPE:	Maximum Likelihood Parameter Estimation
MNZAV:	Mean of Non-Zero Absolute Values
MSE:	Mean Squared Error
NRIQ:	No Reference Image Quality
pdf:	Probability Density Function
PPIQ:	Probabilistic Perceptual Image Quality
PPIQ:	Probabilistic Perceptual Image Quality
PRIQ:	Partial Reference Image Quality
PSF:	Point Spread Function
R-D:	Rate-Distortion
RF:	Receptive Field
RMSE:	Root Mean Squared Error
ROI:	Region of Interest
SATD:	Sum of Absolute Transform domain Differences
SAV:	Sum of Absolute Values
SLB:	Shannon Lower Bound

SSIM:	Structural SIMilarities
TMF:	Texture Masking Factor
TPSA:	Texture Perception Support Area
UQEDZ:	Uniform Quantizer with Extended Dead Zone
VLC:	Variable Length Coding

ACKNOWLEDGEMENTS

The humble scientific achievements in this presentation had not been possible without contributions from a very diverse group of people, for most of whom, the technical language of this writing is alien. Pondering on whom I should thank and who needs to be acknowledged in the prelude of this dissertation, led me to realize another beauty within this body of work. A reality which is much more splendid than the scientific glamour one can find in this work. When I start delving into the history and tracing back the events which took me to this point, I saw many beautiful souls who unselfishly came to assist me when I needed guidance and help.

From the streets of Tehran to the campus of Stanford, from the cubical offices in the San Francisco Bay area to the offices of high-tech companies in the beautiful British Columbia, Canada and from living a corporate life to living a student life at UCSD, my life has been touched by so many people from different countries, different colors, different religions and different backgrounds. When I look at the pages of my dissertation, I find a great lesson: When humans reach out to understand each other, no hurdle can stop them from achieving the fulfillment of the individuals' desires and realization of the utopian society.

I can not pay the due acknowledgment and tribute to many great teachers and mentors with whom I have had the privilege of apprenticeship and without whom, I would have not been where I am today. Words can not express my gratitude for what the have done for me and for many generations who benefited from their wisdom and knowledge. Although it is not possible to name them all here, I can not leave this section

without naming some of the great mentors to whom I feel indebted forever. Mr. Hadi H. Saeedi and the recently deceaseds Akbar Radi and Ebrahim Nouri from my high school years, Amir M. Pezeshk, my undergrad thesis advisor, Cyrus Hazari and Greg Wallace, my professional mentors are the names I have to acknowledge here.

I would like to specially thank my Ph.D. advisor, Prof. Truong Nguyen, whose great personality was a perfect complement to his vast technical knowledge and wisdom. The rare combination of being patient with the students and at the same time extremely motivational for those who are struggling to find their path to an authentic research topic, make him the best advisor I could have ever wished for. Prof. Nguyen's style of scientific advising and academic management made my studies at UCSD, one of the best and most memorable times I have ever had in my life. I am indebted to him for giving me the opportunity to achieve my long awaited dreams.

I would also like to thank the members of my dissertation committee, Prof. Gert Cauwenberghs, Prof. Pamela Cosman, Prof. David J. Kreigman and Prof. Bhaskar Rao, who kindly accepted to furnish their time and invaluable comments to improve the quality of this work.

I would like to thank all my lab-mates over the past 5 years at the Video Processing Laboratory at UCSD, from whom I learnt a great deal. I would like to especially thank Ryan Prendergast, Shay Har-Noy and Nickolaus Mueller for reading my papers and providing their comments for which the results can be seen in this writing.

Here I would like to acknowledge Motorola Inc. for their generous financial support through the Motorola Partnership in Research Grant, during my Ph.D. studies at UCSD. I would like to especially thank Dr. Ajay Luthra and Dr. David Baylon who bent

over backward to make this fellowship grant getting approved every year for the past 4 years, despite all the bureaucratic odds. I am also grateful to David Baylon who has been reading many of my papers and providing his valuable comments to enhance the technical and presentational aspects of those papers.

No one had a greater impact on my life than my parents. They not only set most of the values and morals that I dearly cherish in my life, but also gave me the courage to believe in myself when I have to face new challenges in life. They have thought me to be happy about things which I have achieved and be happily content about what is in my destiny.

Also for the past ten years, the joy of living with my wonderful wife, Shadi Sagheb, has brightened every aspect of my life. The work in front of you is a witness to many sacrifices she made to help me reach my goals. When we were blessed by our daughter, Salma, in November of 2004, my wife worked extra hard to manage the family affairs, while she was completing her graduate studies at UCSD from Sept-2006 to June-2008. Shadi's support goes beyond taking extra responsibility at home while I was studying. In fact she was instrumental in conducting many experiments we performed to verify the soundness of many of the theoretical works in this dissertation. Some of Shadi's contributions are reflected in the paper we jointly submitted on the PPIQ framework for image quality assessment.

The love and support I have received everyday from my greater family, including main and foremost, my parents: Javad Minoo and Mahdokht Ehsan, my wife and her family, my daughter, my brother: Dr. Parham Minoo and his family, and finally my

sister: Dr. Niusha Minoo and her family, have been empowering me to stand many challenges to earn this academic degree.

Before concluding this section I would like to acknowledge the following copyright matters:

Chapter 3, in part, contains segments from the following submitted papers:

- Minoo, K.; Nguyen, T.Q., “Optimal Mode selection via Maximum Likelihood rate estimation: Application IN fast mode-selection within H.264,” Signals, Systems and Computers, the Forty-second Asilomar Conference on, ACSSC 2008. Accepted for publication, 2008.

- Minoo, K.; Truong Nguyen, "Maximum Likelihood Rate Estimation: With Applications in Image and Video Compression," Data Compression Conference, 2008. DCC 2008, vol., no., pp.535-535, 25-27 March 2008.

Chapter 2 and Chapter 3, in parts, are currently being prepared for submission as the following paper:

- Minoo, K.; Truong Nguyen, "Maximum Likelihood Entropy Rate Estimation and Its Application in Image & Video Compression," Circuits and Systems for Video Technology, IEEE Transactions on, Sept 2008.

Chapter 5, in part, has been submitted for the following publication:

- Minoo, K.; Nguyen T., “A Perceptual Image Quality Metric Based on Local Texture Spread,” Image Processing, IEEE Transactions on, submitted for publication, 2008.

Chapter 6, in part, has been submitted for the following publication:

- Minoo, K.; Sagheb, S.; Nguyen T., “PPIQ: A Probabilistic Perceptual Image Quality Metric”, Selected Topics in Signal Processing, IEEE Journal of, Visual Media Quality Assessment April 2009. Submitted for publication, 2008.

I would like to thank Alan Bovik and his team for sharing their subjective test results with us and the larger video processing community. Without their contribution, it would have been a very long and costly process to verify the experimental results and certainly to publish this work. I would like to especially thank Hamid Sheikh for providing us with the access to LIVE image quality database [SWCBOL].

At the end I need to acknowledge that this work was supported, in part, by the “Motorola Partnership in Research Grant” and by the matching fund from the “UC Discovery Grant”.

VITA

1990 Bachelor of Science, Sharif University of Technology, Tehran, Iran.

1994 Master of Science, Stanford University , Stanford, CA, USA.

2008 Doctor of Philosophy, University of California, San Diego, CA, U.S.A

PUBLICATIONS

Journals:

Minoo, K.; Sagheb, S.; Nguyen T., "PPIQ: A Probabilistic Perceptual Image Quality Metric", Selected Topics in Signal Processing, IEEE Journal of, Visual Media Quality Assessment April 2009. Submitted for publication, 2008.

Minoo, K.; Nguyen T., "A Perceptual Image Quality Metric Based on Local Texture Spread," Image Processing, IEEE Transactions on, submitted for publication, 2008.

Minoo, K.; Truong Nguyen, "Reciprocal Subpixel Motion Estimation: Video Coding With Limited Hardware Resources," Circuits and Systems for Video Technology, IEEE Transactions on , vol.17, no.6, pp.707-718, June 2007

Conferences:

Minoo, K.; Nguyen, T.Q., "Optimal Mode selection via Maximum Likelihood rate estimation: Application IN fast mode-selection within H.264," Signals, Systems and Computers, the Forty-second Asilomar Conference on, ACSSC 2008. Accepted for publication, 2008.

Minoo, K.; Nguyen T., "A perceptual metric for blind measurement of blocking artifacts with applications in transform-block-based image and video coding," Image Processing, 2008 IEEE International Conference on. Accepted for publication, 2008.

Minoo, K.; Truong Nguyen, "Maximum Likelihood Rate Estimation: With Applications in Image and Video Compression," Data Compression Conference, 2008. DCC 2008 , vol., no., pp.535-535, 25-27 March 2008.

Minoo, K.; Nguyen, T.Q., "Rate Estimation, Using Forward Adaptive Quantization: H.264 Fast Intra Mode Selection at High Data Rates," Signals, Systems and Computers, 2007. ACSSC 2007. Conference Record of the Forty-First Asilomar Conference on , vol., no., pp.235-238, 4-7 Nov. 2007.

Koohyar Minoo; Nguyen, T.Q., "Reverse, Sub-Pixel Block Matching: Applications within H.264 and Analysis of Limitations," Image Processing, 2006 IEEE International Conference on , vol., no., pp.3161-3164, 8-11 Oct. 2006.

Minoo, K.; Nguyen, T.Q., "Perceptual Video Coding with H.264," Signals, Systems and Computers, 2005. Conference Record of the Thirty-Ninth Asilomar Conference on , vol., no., pp. 741-745, October 28 - November 1, 2005.

ABSTRACT OF THE DISSERTATION

ON THE RATE & DISTORTION:
CONFORMITY WITH THE STATISTICS OF NATURAL IMAGES
AND VISUAL PERCEPTION IN HUMANS

by

Koohyar Minoo

Doctor of Philosophy in Electrical Engineering
(Signal and Image Processing)

University of California, San Diego, 2008

Professor Truong Nguyen, Chair

In this dissertation the subjects of entropy coding and quality assessment in the context of natural image processing and compression have been revisited. Both subjects are amongst the most fundamental concepts which have been extensively studied under the theories of source coding and signal processing. In this dissertation, it will be demonstrated how conformity to the statistical properties of natural image data, makes it possible to estimate the entropy rate of such data with high accuracy and very low

complexity. A maximum likelihood parameter estimation framework is proposed which not only is enabling the design of a fast and efficient entropy rate estimator, but also unifies the legacy rate estimation methods, namely the heuristic low-data-rate methods and the analytical high-data-rate methods.

The concept of entropy rate crosses the concept of image quality measure, or distortion metric (fidelity criterion), most often under the subject of lossy source coding to measure the optimality of a compression scheme. However the distortion metrics are amongst the most basic concepts for evaluation of other image processing algorithms, beyond the image compression. Underlined by numerous publications, the need for a perceptual quality metric that reflects the perception of humans on the subject of visual quality is unanimously agreed upon. The endeavor to find a suitable image quality metric has resulted in the introduction of many image quality assessment methods.

The contribution of this work on the subject of image quality is a modest step forward in unifying many of the legacy methods under a "probabilistic perceptual image quality" framework. It will be shown that different methods such as contrast sensitivity, channel decomposition and structural similarity methods are different realizations of the proposed framework. This framework not only unifies the legacy methods, but also provides means for comparing different legacy methods. Furthermore, the proposed framework creates opportunities to enhance most of the legacy perceptual image quality measures. Finally the probabilistic nature of image quality in the proposed method lends itself to extending the quality metric beyond image quality assessment with full-reference image. It also covers the quality assessment when there is no access to the reference image.

Chapter 1
Introduction

The subjects of entropy coding and quality assessment are amongst the most fundamental concepts in the filed of image and video processing. For legacy reasons these two concepts have been treated in image processing applications very undesirably. Conventional entropy coding methods often use the same generic tools that have been studied and developed under the broad class of source coders. The non-stationary nature of image data requires the generic entropy coders to be either inefficient in the case of non-adaptive entropy coding schemes, or computationally complex in the case of adaptive arithmetic entropy coders in order to achieve desired coding efficiency. In this dissertation it will be demonstrated how the conformity of image data with certain classes of distributions provides great opportunities for fast entropy rate estimation and efficient entropy coding.

The other legacy concept which is used in many image processing applications is the concept of Euclidian distance for measuring the quality of images. The legacy applications in generic signal processing and the facilitating nature of Mean Squared Error (MSE) for derivation of the theorems in lossy source coding have been the main reasons for the prevalence of MSE distortion in most image processing applications.

A relevant distortion metric (or a fidelity criterion) which measures the distance between a test image and a reference image should be representative of the application which uses the metric. In the case of quality evaluation by a human observer, the MSE has been proved to be undesirable [Giro93].

The contributions of this dissertation are as follows:

1. Introduction of a fast and robust entropy rate estimation method, based on a parametric distribution of image data for natural scenes and the maximum likelihood parameter estimation technique.
2. Introduction of the local texture activity for measuring the masking effect of image content.
3. Introduction of the Probabilistic Perceptual Image Quality framework, which unifies the conventional image quality metrics and provides a simple way to define accurate quality metric for different applications.

1.1 Entropy Rate

One of the most fundamental techniques, used in all modern image and video coding schemes, is entropy coding. Entropy coding exploits the statistical redundancies in the transmitted data to reduce the number of required bits for transmission of that data in a lossless manner [CoTh91]. Huffman coding, Elias coding, Golomb coding and arithmetic coding are examples of entropy coding schemes. Since the efficiency of compression schemes will enhance with the efficiency of entropy coders, designing high efficiency entropy coders, with reasonable complexity (e.g., to achieve real-time performance within the realm of conventional computing power) is of great interest.

In Chapter 2 we review the principles of entropy coding for visual data. In Chapter 3 it will be shown that an analytical approach to rate estimation by means of maximizing the likelihood of observed data in a block of image data, provides a rich theoretical framework. In this framework some of the legacy ad hoc and heuristic

approaches to rate estimation at low data rates can be evaluated. This framework also affirms the theoretical results which had been previously proved to be true based on the high data rate assumption.

1.2 Distortion (Image Quality)

Most of the image and video processing applications rely on MSE to measure the distortion. Although MSE has proven to be a suitable measure for the development of many theoretical results in the area of image and video processing, it falls short of representing how a human observer perceives the distortion in an image or a sequence of pictures. In cases where the final judgment on efficiency of a coding scheme comes from a human observer, it becomes very important to find an objective perceptual distortion measure which closely predicts how, on the average, observers grade the quality of a distorted image.

Chapter 4 reviews the concepts and methods which have been previously studied for the purpose of image quality assessment. In Chapter 5, the Local Texture Spread (LTS) measure is introduced as an alternative for measuring the texture activity in different areas of an image. It is argued why the LTS is a better choice for representing the Human Visual System (HVS) compared to alternative block based methods in terms of representing the HVS and the computational complexity. Experimental results confirm that the LTS is a suitable measure to predict the Just Noticeable Distortion (JND) in the presence of image texture. The results of subjective tests are presented to prove the validity of a normalized MSE as a perceptual metric in supra-threshold regimes, when the normalization is done according to the LTS measure. To that end a parametric image

quality model is introduced and optimized based on the empirical studies at JND regime to find the MSE normalization factor at supra-threshold regimes according to the LTS measure.

Chapter 5 further shows that the proposed distortion metric not only achieves higher accuracy to predict the subjective test results, but also requires lower computational complexity compared to most of the conventional perceptual image quality metrics.

In Chapter 6 a probabilistic framework is introduced for measuring image quality that unifies many of the previously suggested distortion metrics. This Probabilistic Perceptual Image Quality (PPIQ) framework is developed based on the known principles of how visual "features" are formed and perceived in human visual system. Based on this framework, a generic image quality metric is introduced which corresponds to a simple contrast feature (Laplacian of Gaussian). It will be shown that this simple contrast feature within the PPIQ framework performs better than most quality metrics with comparable computational complexity, such as structural similarity index [WBSS04].

Before concluding this chapter the respected readers are reminded that a list of all acronyms, used in this dissertation, can be found at the end of this document, before the list of published references.

Chapter 2
A Review of Concepts and Methods for Entropy Coding of Visual data

One of the most fundamental techniques, used in all modern image and video coding schemes, is entropy coding. Entropy coding exploits statistical redundancies in the transmitted data to reduce the number of required bits for transmission of that data in a lossless manner [Gray90]. Huffman coding, Elias coding, Golomb coding and arithmetic coding are examples of entropy coding schemes, which are used in conventional image and video compression algorithms [TaMa02], [WOQZ02].

In image and video compression, the subject of entropy rate is of interest from two perspectives. First, the design of fast and efficient entropy coding increases the efficiency and performance of the overall compression scheme. Secondly, in many of compression tasks such as selection of quantization parameter (e.g. for rate control) and selection of spatiotemporal prediction choice, the knowledge about rate is required along with the distortion for making optimal decisions. In this chapter we first review entropy coding from the image coding perspective and then elaborate on techniques for estimating the entropy rate for the purpose of making R-D optimal compression choices, without the need to perform the actual coding.

2.1 Entropy Coding Techniques in Image and Video Compression

Modern image and video coding schemes employ entropy coders to compress the quantized coefficients of (usually) transformed residual data. In this setup, Uniform Quantization (UQ) has been shown to be optimal for high data rates. This means that

entropy constrained quantization requires uniform quantization [GeGr91]. Optimal entropy coding, ultimately requires the knowledge about the distribution of the data which will be entropy coded. For this purpose we first review the type of information that would be entropy coded in conventional image and video compression applications, in 2.1.1. Observing that entropy coding (on the encoder side) and the entropy decoding (on the decoder side) need to have the same knowledge about the probability distributions of the transmitted data, poses a technical problem. In 2.1.2, a number of techniques are discussed which can be used to extract the probability distributions both at the encoder and the decoder.

2.1.1 Coded Data in Image Compression

In video and image coding schemes, entropy coding relies on the knowledge of the distribution of the quantized data. In most image and video coding standards, the residuals (after some temporal or spatial prediction of the image content) go through a transform, e.g. Discrete Cosine Transform (DCT) or Wavelet, to form the residual coefficient (coefficients). The coefficients then will be quantized to form some indices which identify the quantization bin. The study of statistical properties of these indices has led to the adoption of the following entropy coded data partitioning (grouping) in most conventional image and video entropy coding techniques [Wall92], [H_264_]. The coded indices are typically divided into three groups.

1. Significance map: This data group represents the position of data which are not quantized to zero.

2. Non-zero values: This data group represents the absolute values of non-zero quantized data.
3. Sign value: The third group contains the information on the signs of quantized data (only non-zero).

The statistical behavior of the residual coefficients (for images) before quantization makes it more efficient to entropy code these three groups of data independently in the following ways:

1. Coding of Significance map: The positions of non-significant values are highly correlated. Therefore higher order entropy coding which considers the joint probability of neighboring quantized coefficients can extensively enhance the coding efficiency. Separation of the significance map from the value of coded data allows a simple binary entropy coder to create the probability models which requires a relatively small amount of data for updating the probabilistic model parameters.
2. Coding of non-zero values: Since one of the goals of transform coding is to make the coded data (coefficients in transform domain) uncorrelated, it is reasonable to perform first order entropy coding on this type of data.
3. Coding of Sign bits for non-zero coefficients: As will be shown in chapter 3.1, the sign value data group has entropy of 1 bit per sample due to the symmetric nature of the residual coefficients and symmetry of the conventional quantization schemes around zero. So in practice we don't need to keep track of the probabilities of the sign values for entropy coding. Here, the readers need to be reminded that despite the fact that

most entropy coding schemes in current image and video coding schemes follow the argument we put forth, it has been shown in [DeHe03] that performing higher order entropy coding on the joint probability of sign bits can result in marginal efficiency of the entropy coding (at least for Wavelet compression schemes).

2.1.2 Probability of Coded Data

According to the Shannon theory of communication, the amount of information (e.g., measured in bits) to asymptotically represent a symbol is given by the expected value of the logarithm (e.g. in base two) of the probability of that symbol (or the frequency by which the symbol appears in a transmitted message). To decode the entropy coded data the receiver needs to know the correspondence between the symbols and the codewords. In general this can be done in one of the following three ways.

1. Static approach: In this approach the encoder and the decoder are provided with the same dictionary at the beginning of the transmission.
2. Adaptive approach with no side information: In this approach both the encoder and the decoder, synchronously, build the same dictionary based on the statistics of the data which already received by the decoder.
3. Adaptive approach with side information: This approach benefits from the advantages of the adaptive approach, but in occasions when the transmission cost of extra information is justified, the codebook (or the parameters of the probability model) will get updated at the encoder and sent to the decoder as side information.

In the context of video coding, the first two approaches have been used by different coding standards. For example JPEG and MPEG-2 standards use static look up tables for entropy coding (also called Variable Length Coding or: VLC) of the (run, length) pairs for the quantized DCT coefficients of the residual data after prediction (for MPEG-2 it is mainly temporal prediction with the exception of spatial prediction of the DC coefficients). More recently the new video coding standards have added an alternative entropy coding option which adaptively builds context-based probability models as new data is being received and decoded at the decoder (approach number two above). For example the standard H.264/AVC uses a Context Adaptive Binary Arithmetic Coders (CABAC) scheme to achieve higher coding efficiency of 5% to 15% (in terms of rate reduction) at a price of 150% more computational complexity [MaSW03].

2.2 Entropy Rate Estimation

Rate-distortion optimized coding decisions are amongst the most time consuming tasks of image and video encoding. They require an actual encoding and decoding for every possible coding option. This includes the entropy coding of the residual signal for all possible coding scenarios. Finding the exact rate is especially very time consuming in the state of the art video coding schemes, because modern video encoders employ very sophisticated entropy coders to achieve high coding efficiency for the lossless compression of quantized residual data.

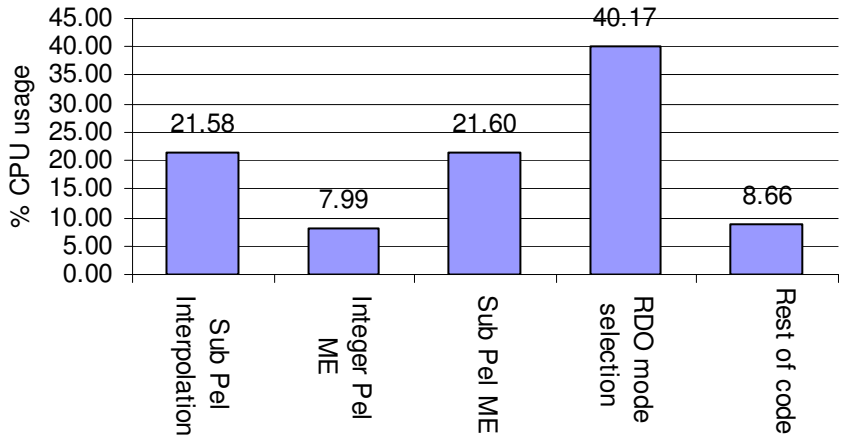


Figure 2.1 Percent of CPU usage by different coding modules.

The use of a robust rate estimation method (instead of the actual rate calculation) for rate-distortion optimized video coding, has been previously considered and many proposals based on this concept have been published [KiKA05], [KaAM05], [HMit02]. In these methods the unconstrained Lagrangian cost function is minimized based on the estimated rate and the estimated distortion. There have been other methods of rate estimation based on assuming different distributions for the coded signal, which are as effective as the accuracy of assumed model. These estimation methods work fine as long as the distribution model is valid, but when the assumed model fails to capture the statistical characteristics of the data, the coding efficiency degrades significantly.

2.2.1 Rate Estimation at High Data Rates

Most rate estimation schemes work on the premise of operation in high data rate regimes. In the context of lossy source coding the “high data rate” condition implies constant probability distribution over each quantization bin. It has been shown [FaMo84] that for many classes of memoryless distributions, uniform quantization is optimal for bit

rates as low as 0.5 bits per sample (bps). The fact that in transform coding of image and video data, the distribution of coded coefficients has a sharp peak around zero, combined by the extended quantization bin around zero (dead zone), violates the assumption of high data rates in typical operation rates (only applicable for very small quantization step sizes, which are not used in many practical cases). Also many high-data-rate approaches to rate estimation that rely on a certain probability density function (pdf) for coded data [KaAM05], compensate for the non-stationary nature of the image and video data, by employing a parametric model that should be updated after coding of each block of data. This adds to the complexity of the rate-estimator.

2.2.2 Rate Estimation at Low Data Rates

On the other hand there are some heuristic methods for rate estimation at very low data rates [HeKM01], [HeMi02]. These methods usually rely on a number of parameters which are obtained empirically by means of statistical regression from a training set, which offers no guarantee that the model parameters hold true for all contents. The other problem with algorithms such as [HeKM01] is that they only perform well under low and mid-low data rate regimes. The low data rates rate estimation methods are driven by experimental observations and only predict well when the number of data for which the rate is estimated is large (e.g. the overall rate for the entire image).

The high data rate and low data rate methods for rate estimation each have their own problems. For example in the case of high data-rate-based approaches, these methods have to update the model parameters after encoding each block of data. In video coding, most often, different types of predictive schemes introduce residual data with

remarkably different distributions. This requires the rate estimation techniques to maintain different model parameters for different image coding types such as I, P and B pictures, which adds to the complexity of model updating.

In Chapter 3, we discuss how the rate for a group (block) of data can be estimated, based on the maximum likelihood parameter estimation, for a well accepted Laplacian distribution [SmRo96]. Furthermore, it will be shown that the proposed rate estimation method in Chapter 3 is capable of accurately estimating the rate over the entire range of operational rates from very low data rates to very high.

This chapter, in part, contains segments from the following paper which is being prepared for submission:

- Minoo, K.; Truong Nguyen, "Optimal Entropy Coding via Parameter Estimation And Its Application in Image & Video Compression," Circuits and Systems for Video Technology, IEEE Transactions on, Sept 2008.

Chapter 3
Entropy Rate Estimation via Maximum Likelihood Parameter Estimation

In this chapter we consider the statistical properties of the entropy coded data, (e.g. the quantized coefficients of transformed, residual image data after some temporal or spatial prediction) to estimate the rate produced by an efficient adaptive entropy coder (e.g. CABAC). The statistical model, which is defined in a parametric manner, lends itself to defining the rate estimation as a Maximum Likelihood Parameter Estimation.

3.1 Statistics of coded data for Natural Images

In scientific literature one can find many candidates for the classes of probability distribution for the residual data in predictive video coding schemes, especially for residual coefficients of DCT. Cauchy [KaAM05], Laplacian [SmRo96], DC coefficients Gaussian + AC coefficients Laplacian [ReGi83], generalized Gaussian [Mull93], Gaussian mixture models [EGCD94], etc. In [LaGo00], Lam and Goodman showed that an infinite Gaussian mixture would result in a Laplacian distribution. In this work we assume a Laplacian distribution for representing the coded data before quantization. This choice is made based on the previous studies [SmRo96], [LaGo00] and the fact that the specific choice of Laplacian makes it possible to solve related mathematical equations and find the exact results (look at 3.2 and 3.3 for details), as opposed to methods such as [KaAM05] for which the closed form can not be found and simplification should be made which then questions any possible advantage of using a specific distribution in the first place with the promise of achieving marginal improvements.

The Laplacian distribution assumption assigns the probability distribution $f(x)$ to the transform coefficient (with value x) of the residual data as follows:

$$f(x)=\frac{\lambda}{2}\cdot e^{-\lambda|x|} \tag{3.1}$$

In contemporary video coding schemes the transformed coefficients are quantized, typically with an un-bounded Uniform Quantizer with Extended Dead Zone (UQEDZ) to generate the coded data (c) based on the following rule:

$$c=\begin{cases}0 & when\ |x|\le(\Delta+\delta)\\ j & when\ \ |j\cdot\Delta+\delta|<|x|\le|(j+1)\cdot\Delta+\delta|\end{cases} \tag{3.2}$$

The combination of (3.1) and (3.2) results in the probability mass function of the coded data c as follows:

$$P(c=i)=\begin{cases}1-e^{-\lambda\cdot(\Delta+\delta)} & j=0\\ \frac{1}{2}\cdot\left(1-e^{-\lambda\cdot\Delta}\right)\cdot e^{-\lambda\cdot(|j|\cdot\Delta+\delta)} & j\neq 0\end{cases} \tag{3.3}$$

To simplify the notation we use the following substitutions:

$$\begin{aligned}\alpha&=1-e^{-\lambda\cdot(\Delta+\delta)}\\ \beta&=e^{\lambda\cdot(\Delta)}\end{aligned} \tag{3.4}$$

This results in the following probability mass function for the quantized coefficient with Laplacian distribution.

$$P(c=j)=\begin{cases}\alpha & j=0\\ \frac{1}{2}\cdot(1-\alpha)\cdot(\beta-1)\cdot\beta^{-j} & j>0\\ \frac{1}{2}\cdot(1-\alpha)\cdot(\beta-1)\cdot\beta^{j} & j<0\end{cases} \tag{3.5}$$

3.2 Estimation of the Entropy Rate and Maximum Likelihood Parameter Estimation

In this chapter, we are interested in assessing the entropy rate for transmission of a block of data $x_1, x_2, ...x_N$. If we assume the coded data have a distribution of $P(x_1, x_2, ...x_N; \Theta)$, with known parameter Θ, then the asymptotically optimal number of bits for representing the block of data is given by Shannon entropy rate as:

$$R = \frac{1}{N} \cdot H(x_1, x_2, \cdots x_N) = \frac{-1}{N} \cdot \log_2 \left(P(x_1, x_2, \cdots x_N; \Theta) \right) \quad (3.6)$$

The goal of entropy coding is to reduce the number of bits given by (3.6). If the coded data and the family of distribution for that data are given, the only parameter we can manipulate to influence the rate is the parameter of the probability model, Θ. Mathematically this can be written as:

$$\Theta^* = \underset{all \ \Theta}{\arg\min} \ \{ R = \frac{-1}{N} \log_2 \left(P(x_1, x_2, \cdots x_N; \Theta) \right) \}$$

or (3.7)

$$\Theta^* = \Theta^*_{ML} = \underset{all \ \Theta}{\arg\max} \ \{ \log_2 \left(P(x_1, x_2, \cdots x_N \mid \Theta) \right) \}$$

Note that the minimization of the entropy rate resulted in maximizing the log likelihood of the coded data in Equation (3.7). This suggests that minimum entropy can be achieved by estimating the parameter Θ which maximizes the (log) likelihood. Once the optimal Θ is known, the entropy can be done on the encoder side.

3.3 First Order Entropy Rate Estimation for Transmission of Quantized Image Data

A rate estimation based on the first order entropy of the rate in (3.7) is given by:

$$\Theta^* = \underset{all\ \Theta}{\arg\min} \{R = \frac{1}{N}\sum_{i=1}^{N} r_i = \frac{-1}{N}\cdot\sum_{i=1}^{N}\log_2\left(P(x_i;\Theta)\right)\} \tag{3.8}$$

Based on the probability model in (3.5) we can re-write (3.8) as follows:

$$\left(\alpha^*,\beta^*\right) = \underset{all(\alpha,\beta)}{\arg\min} \{R = \frac{1}{N}\cdot\sum_{i=1}^{N} r_i = \frac{-1}{N}\cdot\sum_{i=1}^{N}\log_2\left(P(c = x_i;\alpha,\beta)\right)\} \tag{3.9}$$

Before expanding (3.9) we introduce a few notational conventions.

1. $f(j)$: The number of times that symbols are observed with an absolute value of j in the block which is coded $f(j) = \sum_{i=1}^{N} I\left(|x_i| = j\right)$. Note that $\sum_{j=0}^{\infty} f(j) = N$.

2. ρ: The number of zero symbols in the block of coded data divided by the number of symbols in the block (i.e. N). By definition $\rho = f(0)/N$ therefore we refer to this value as the percentage of zero data. Note that $\frac{1}{N}\cdot\sum_{j=1}^{\infty} f(j) = 1-\rho$

3. γ: The Sum of Absolute Values (SAV) of the coded data for the entire block divided by the number of symbols or Mean Absolute Value (MAV) (i.e. $\gamma = \frac{\sum_{j=1}^{\infty} j\cdot f(j)}{N}$).

With the above notation one can rewrite (3.9) as follows:

$$R = \frac{-1}{N}\cdot\sum_{i=1}^{N}\log_2\left(P(c = j;\alpha,\beta)\right) = \frac{-1}{N}\cdot\left(\left(f(0)\cdot\log_2(\alpha)\right) + \left(f(j)\cdot\left(\log_2\left(\frac{1}{2}\cdot(1-\alpha)\cdot(\beta-1)\cdot\beta^{-j}\right)\right)\right)\right) \tag{3.10}$$

Substituting the probability model from (3.5) in the above and using the fact that the probability is symmetric around zero, yields:

$$R=-\left(\left(\frac{f(0)}{N}\cdot\log_2(\alpha)\right)+\left(\sum_{j=1}^{\infty}\frac{f(j)}{N}\cdot\left(-1+\log_2(1-\alpha)+\log_2(\beta-1)+\log_2\left(\beta^{-j}\right)\right)\right)\right) \qquad (3.11)$$

Further simplification and replacing terms with the notation introduced above yields:

$$R=-\left((\rho\cdot\log_2(\alpha))+((1-\rho)\cdot(-1+\log_2(1-\alpha)+\log_2(\beta-1))-\gamma\cdot\log_2(\beta)\right) \qquad (3.12)$$

To find the optimal parameters (α^*,β^*) in (3.9) which minimize the rate, we take derivatives of (3.12) with respect to α and β and set them to zero to find the optimal parameter values as follows:

$$\begin{aligned}\frac{\partial(R)}{\partial\alpha}&=-\log_2(e)\cdot\left(\frac{\rho}{\alpha}-\frac{(1-\rho)}{(1-\alpha)}\right)=0\Rightarrow\alpha^*=\rho\\ \frac{\partial(R)}{\partial\beta}&=-\log_2(e)\cdot\left(\frac{(1-\rho)}{(\beta-1)}-\frac{\gamma}{\beta}\right)=0\Rightarrow\beta^*=\frac{\gamma}{\gamma-(1-\rho)}\end{aligned} \qquad (3.13)$$

Substituting these optimal values in (3.12) gives the following estimation of the rate based on the first order entropy of the coded symbols.

$$R=-\left((\rho\cdot\log_2(\rho))+\left((1-\rho)\cdot\left(-1+\log_2(1-\rho)-\log_2\left(\frac{\gamma}{1-\rho}-1\right)\right)+\gamma\cdot\log_2\left(1-\frac{1-\rho}{\gamma}\right)\right)\right) \qquad (3.14)$$

3.4 Higher Order Entropy Rate Estimation for Transmission of Quantized Image Data

One can use (3.3) to derive the probability mass function for the three groups of coded image data (consult 2.1.1) as follows.

1. Significance map information: $P_{si\_ma}(\cdot)$

$$P_{si\_ma}(0) = P(c = 0) = \alpha$$
$$P_{si\_ma}(1) = P(c \neq 0) = (1-\alpha) \tag{3.15}$$

2. Absolute value of non-zero coefficients: $P_{nz\_ab}(\cdot)$

$$P_{nz\_ab}(j) = P(|c| = j \mid c \neq 0) = \frac{P(|c| = j)}{P(c \neq 0)} = (\beta - 1)\cdot \beta^{-j} \qquad \forall\ j > 0 \tag{3.16}$$

3. Distribution of sign values: $P_{nz\_si}(\cdot)$

$$P_{nz\_si}(1) = P(c > j \mid c \neq 0) = \frac{1}{2}$$
$$P_{nz\_si}(0) = P(c < j \mid c \neq 0) = \frac{1}{2} \tag{3.17}$$

One can calculate the rate, associated with each of the three categories of data, based on the probability mass function given in (3.15) to (3.17), as follows:

$$\begin{aligned} R_{si\_ma} &= \frac{-1}{N}\left(f(0)\cdot \log_2(\alpha) + \left((N - f(0))\cdot \log_2(1-\alpha)\right)\right) \\ &= -\left(\rho \cdot \log_2(\alpha) + \left((1-\rho)\cdot \log_2(1-\alpha)\right)\right) \end{aligned} \tag{3.18}$$

$$\begin{aligned} R_{nz\_ab} &= \left(\frac{-1}{N}\sum_{j=1}^{\infty} f(j)\cdot\left(\log_2\left(P_{nz\_ab}(j)\right)\right)\right) = -\sum_{j=1}^{\infty}\left(\frac{f(j)}{N}\cdot\left(\log_2\left((\beta-1)\cdot\beta^{-j}\right)\right)\right) \\ &= -\left(\left((1-\rho)\cdot\log_2(\beta-1)\right) - \left(\gamma\cdot\log_2(\beta)\right)\right) \end{aligned} \tag{3.19}$$

$$R_{nz\_si} = \frac{1}{N}\sum_{j=1}^{\infty} f(j) = 1 - \rho \tag{3.20}$$

A comparison between (3.12) and (3.18) through (3.20) reveals that as expected we have the following equality: $R = R_{si\_ma} + R_{nz\_ab} + R_{nz\_si}$. Because of this equality the optimal parameters (α^*, β^*) in (3.18) through (3.20) should be the same as we derived in (3.13). Consequently the following expressions are the estimated rate for transmission of different groups of data within a block of size N.

$$R_{si\_ma} = -\left(\rho\cdot\log_2(\rho) + \left((1-\rho)\cdot\log_2(1-\rho)\right)\right) \tag{3.21}$$

$$R_{nz\_ab} = (1-\rho)\cdot\left(\log_2(\bar{\gamma}-1)-\left(\bar{\gamma}\cdot\log_2\left(1-\frac{1}{\bar{\gamma}}\right)\right)\right) \quad (3.22)$$

$$R_{nz\_si} = \frac{1}{N}\sum_{j=1}^{\infty} f(j) = 1-\rho \quad (3.23)$$

Note that in (3.22) we have introduced a new notation $\bar{\gamma}$ which is the SAV divided by the number of non-zero symbols or the Mean of Non-Zero Absolute Values (MNZAV) (i.e. $\bar{\gamma} = \frac{\gamma}{(1-\rho)}$). The observation that overall rate optimization yields the same results as optimizing individual rate components, above, suggests that there is no need to divide the coded data into the three groups of: significance map, non-zero absolute values and non-zero sign values. However as we discussed in 2.1.1, the decomposition of the entropy rates into the aforementioned three groups will allow us to take advantage of the statistical properties of the coded data for natural images and videos. As we will see in 3.5.1 the first order entropy would be a very close estimate of the rate for non-zero absolute values, when compared to the more sophisticated entropy coding schemes such as CABAC in the H264/AVC video coding standard. Also because of the symmetry in probability distribution of the coded data around zero and the un-correlated nature of sign bits, one bit per non-zero data is required (i.e. no entropy coding is needed). This only leaves the coding of the significance map to require higher order statistics. In practice (e.g., in CABAC) this has been achieved by keeping different probability contexts based on the value of the significance map in the neighboring locations.

Another key advantage of separation of rate terms as in (3.18) through (3.20), is to show that the Maximum Likelihood Parameter Estimation (MLPE) for β only

depends on, and at the same time, minimizes the rate for coding the absolute values of non-zero symbols R_{nz_ab}. In the same manner the parameter α is only influenced by and affects the rate for transmission of the significance map R_{si_ma}.

3.5 Rate Estimation Applications: Mode Selection within H.264

In this section we apply the result of the proposed rate estimation based on MLPE to perform rate-distortion optimized mode selection. For this purpose we outline two different cases. In the first case the mode selection is done based on an estimated rate for each macroblock from the symbol values in that macroblock. The second approach performs rate estimation for smaller blocks (within a macroblock) based on the data in the block which is being coded and a number of its surrounding blocks.

3.5.1 Macroblock level mode selection

To observe the efficiency of the proposed rate estimation algorithm we use the results in (3.14) to estimate the rate within the H.264/AVC reference encoder to choose the optimal coding mode (prediction mode) at macroblock level. First, we investigated how well the proposed scheme in (3.14) can predict the actual rate for coding of the texture data. For this purpose, we perform DCT followed by UQEDZ on the residual for every macroblock and calculate the values ρ and $\bar{\gamma}$. Note that since we need to perform the DCT and quantization we may as well calculate the exact distortion in the DCT (due to the unitary property of the DCT).

3.5.1.1 Study of empirical data

Comparing empirical rate (using CABAC in H.264) with that estimated by (3.21) through (3.23) for each of three types of data for any given macroblock, we made the following observations.

1. Inaccuracy of first order entropy for rate estimation of significance map:
2. Accuracy of first order entropy for rate estimation of non-zero values (absolute and sign values):

The above observations motivated us to further break down the rate required for transmission of the significance map into two parts:

1. Rate for transmission of the zero positions in the significance map: $R_{si\_ma\_z} = -\rho \cdot \log_2(\rho)$.
2. Rate for transmission of the non-zero positions in the significance map: $R_{si\_ma\_nz} = -(1-\rho) \cdot \log_2(1-\rho)$.

The experimental results show that $R_{si\_ma\_nz}$ is a close estimate of the required rate for transmission of the significance map. The reason for this observation is that video coding algorithms do not use first order entropy to compress the zero codes (significance map). It has been noted that the position of zeros in a transform block are highly correlated. For example in zig-zag scan order, if a coefficient is zero then there is higher chance that the rest of the coefficients thereafter would be zero too. For this reason, in block based transform coding schemes such as MPEG-2 or H.264/AVC (in VLC), the zeros are coded by entropy coding the number of consecutive zeros (in some scan order). Furthermore, every non-zero quantized DCT symbol includes a flag bit to indicate the End Of Block (EOB) if the rest of the symbols (in a given scan order) are zero. Another

efficient method of conveying the zeros in conventional block based compression schemes is to use coded block patterns to indicate if a block is completely zero. Because of all the aforementioned strategies for transmission of zero locations in the significance map, we drop the $R_{si\_ma\_z}$ term from the total estimated rate for the calculation of the cost function and only use the rate for sending non-zero bits of significance map, as follows:

$$R \approx R_{nz} = R_{nz\_ab} + R_{nz\_si} + R_{si\_ma\_nz}$$
$$R \approx (1-\rho)\cdot\left[1-\log_2(1-\rho)+\log_2(\bar{\gamma}-1)-\left(\bar{\gamma}\cdot\log_2\left(1-\frac{1}{\bar{\gamma}}\right)\right)\right] \quad (3.24)$$

As mentioned, to calculate the cost function for mode selection at the macroblock level, the distortion would be measured in the DCT domain which is very close to the exact distortion (i.e. not estimated). One should note that the distortion might be slightly different in the pixel domain due to truncation of pixel values. Also for minimization of the Lagrangian R-D cost function we use the same Lagrangian multiplier that the rate-distortion optimization with the exact (actual) rate uses. After calculating the Lagrangian cost function based on the estimated rate and distortion, we select the mode that minimizes this estimated cost function. Figure 3.1 shows the result for high data rate (Qp=25) and low data (Qp=37) rate regimes.

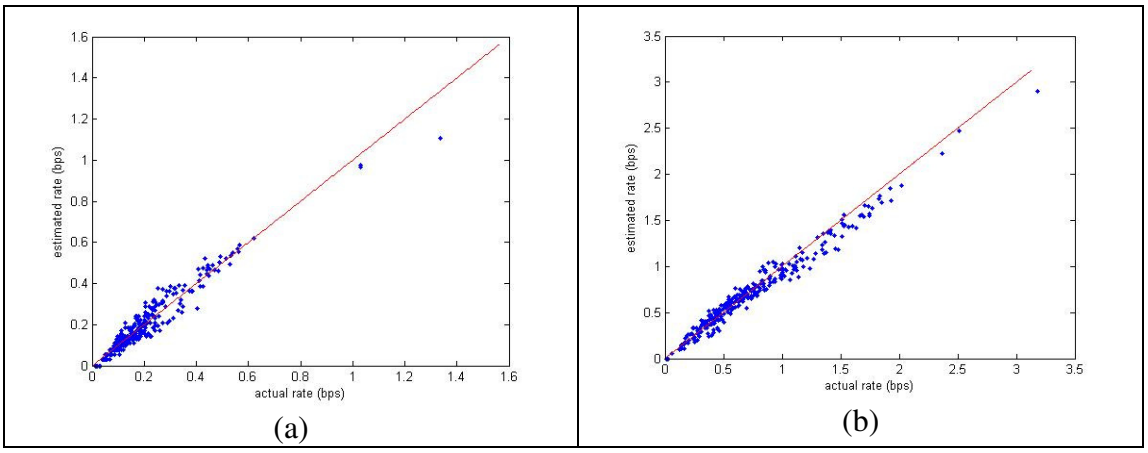


Figure 3.1 The estimated rate vs. the actual rate for foreman sequence(CIF size). (a) is the low data rate (Qp=37) and (b) is the high data rates (Qp=25).

To prove that (3.24) is a reasonable estimation of the rate, a test was conducted, using version 13.0 of the JM reference H.264/AVC coded [JM_REF] to compress video sequences. In this experiment the actual rate and the estimated rate, using (3.24), were logged when different video sequences were coded with different quantization parameters for different picture types (I, P and B). The comparison between the actual rate and the estimated rate is performed by a scatter plot of actual vs. estimated rate as depicted in Figure 3.1. As can be seen in Figure 3.1 the R_{nz} provides a very close estimate of the actual rate over the whole spectrum of operational rates for different video content and different coding strategies.

3.5.1.2 Experimental results for compression efficiency of ML rate estimation within the H.264/AVC

To use the proposed rate estimation for optimal mode selection, the distortion was calculated in the DCT domain to obtain the "almost" exact distortion. (Note that distortion might be slightly different in the pixel domain due to truncation of pixel values.). For the aforementioned reason, the $R_{si\_ma\_z}$ term was dropped from the total

estimated rate, for calculation of the cost function. Also for calculation of the R-D cost function, the Lagrangian multiplier is the same as the one used for brute force rate-distortion optimization.

To evaluate the performance of the proposed rate estimation methods, the JM13.0 version of the reference H.264/AVC encoder [JM_REF] was modified to accommodate the adaptation of the new rate estimation methods. As the compression of symbols takes place in the DCT domain, we define the concept of basic-block to be an image block that goes through the DCT transform. In [JM_REF] there are two options for transform size, 4x4 and 8x8. Throughout this chapter and the next chapter we only considered the smaller block size for two reasons. 1- This block size is more prevalent in contemporary coding applications. 2- The smaller size blocks are more difficult to use for estimation as they provide relatively smaller statistical sample size for optimization of parameters in (3.13).

To perform the rate estimation, for the proposed method in this section and the ones in 3.5.2 and 3.6, the parameter estimation in (3.13) was done separately for five different data contexts where each data context is associated with a data type and a basic-block size as follows:

1. Intra_16x16_DC (sixteen samples per basic-block).
2. Intra_16x16_AC (fifteen samples per basic-block).
3. Luminance_4x4 (sixteen samples per basic-block).
4. Chroma_AC (fifteen samples per basic block).
5. Chroma_DC (variable samples per basic-block depending on the chroma format).

After calculating the Lagrangian cost function based on the estimated rate and distortion, we select the mode that minimizes this estimated cost function. Using the JM13.0 version of the reference H.264 encoder, on a PC platform, we can compare the coding speed and coding efficiency of the proposed algorithm versus that of the conventional method, using the actual rate for mode selection and the approach which uses the fast but not so efficient cost function of Sum of Absolute Transform domain Differences (SATD).

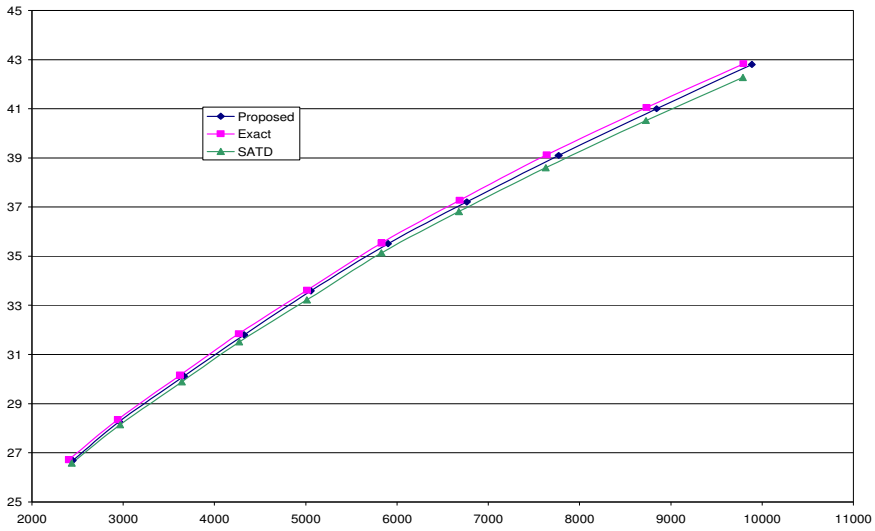


Figure 3.2 The PSNR comparison for mobile sequence at CIF resolution between the proposed method and the SATD method and the brute force rate-distortion calculations.

Figure 3.2 shows the PSNR curve for the proposed rate estimation method in this section. As can be seen, the proposed method performs better than the SATD with almost the same complexity. Also in terms of coding efficiency it performs very close to R-D optimized macroblock mode selection where the conventional rate-distortion mode selection method within the H.264/AVC JM 13.0 reference is used to perform the actual entropy coding to find the exact rate for minimization of the Lagrangian cost function for

each choice of mode. It is important to note that for almost the same coding efficiency, the proposed rate estimation method significantly reduces the required computational complexity when compared with the conventional rate-distortion mode selection within the H.264/AVC JM 13.0 reference codec (almost 10 time faster).

3.5.2 Block level mode selection

Within the H.264/AVC standard a macroblock consists of 256 pixels. This number of symbols provides a large enough sample size to use for parameter estimation in (3.13).

In the previous section, we introduced a rate estimation algorithm via MLPE which provides good results for mode selection when the size of observed data is at least one macroblock. However, it was noted that applying the same method for mode selection of 4x4 or 8x8 blocks (i.e. basic-block as explained in 3.5.1.2) produces sub-optimal results. In this section, the proposed algorithm in 3.4 is extended to estimate the rate for an arbitrary basic-block size. One of the benefits of this rate estimation is its adaptive nature which matches the non-stationary nature of image and video data at a cost of slightly higher complexity.

3.5.2.1 Optimal parameters for probability distribution model

The new approach applies the MLPE to the observed data from the current basic-block and the surrounding N basic-blocks to find the optimal parameters of α and β in (3.13). To that end we consider and compare three methods for rate optimal parameter estimation. In the following, ρ_i is the ratio of zero symbols and n_i is the total number of symbols, in the ith basic-block before the current basic-block (in this notation the 0th

basic-block is the current basic-block for which the rate is being estimated). Also ρ_t is the total ratio of zero coefficients for all past N basic-blocks and the current basic-block. γ_i is the MAV of the i th basic-block before the current basic-block and $\bar{\gamma}_t$ is the MNZAV for all the N basic-blocks.

1. Current Basic-block Only (CBO): This method uses only the data in the coded basic-block to estimate the optimal parameters ($\alpha^*_{CBO}, \beta^*_{CBO}$) of the probability distribution via MLPE for which the results are given by (3.25) based on (3.13).

$$\alpha^*_{CBO} = \rho_0$$
$$\beta^*_{CBO} = \frac{\bar{\gamma}_0}{\bar{\gamma}_0 - (1-\rho_0)} \tag{3.25}$$

2. Current and Past Basic-blocks (CPB): This method uses the observations made for the current basic-block and the N previously coded basic-blocks (total of $N+1$ basic-blocks in coding order) to estimate the optimal parameters ($\alpha^*_{CPB}, \beta^*_{CPB}$) of the probability distribution via MLPE, as suggested in (3.13).

$$\alpha^*_{CPB} = \frac{1}{\sum_{i=0}^{N} n_i} \cdot \sum_{i=0}^{N} n_i \cdot \rho_i = \rho_t$$
$$\beta^*_{CPB} = \frac{\sum_{i=0}^{N} (n_i \cdot \gamma_i)}{\sum_{i=0}^{N} (n_i \cdot \gamma_i) - \left(\sum_{i=0}^{N} n_i - \sum_{i=0}^{N} (n_i \cdot \rho_i) \right)} = \frac{\bar{\gamma}_t}{\bar{\gamma}_t - (1-\rho_t)} \quad \text{and} \quad \bar{\gamma}_t = \frac{\sum_{i=0}^{N} (n_i \cdot \gamma_i)}{\sum_{i=0}^{N} n_i} \tag{3.26}$$

3. Expected Optimal Parameters (EOP): In this approach the optimal parameters ($\alpha^*_{EOP}, \beta^*_{EOP}$) of the probability model are the averages of corresponding

parameters (via MLPE for each individual basic-block) amongst $N+1$ basic-blocks including the current basic-block and the past N basic-blocks. These optimal values can be derived by the following equations.

$$\alpha^*_{EOP} = \frac{1}{N+1} \cdot \sum_{i=0}^{N} \alpha^*_{CBO}$$
$$\beta^*_{EOP} = \frac{1}{N+1} \cdot \sum_{i=0}^{N} \beta^*_{CBO} \quad (3.27)$$

3.5.2.2 Experimental results

The following describes the selection of data which is used to find the optimal parameters via the three estimation methods, discussed above, for different data contexts (as explained in 3.5.1.2).

1. CBO: The quantized coefficients, from each context of the basic-block (for which the rate is being estimated) is used to predict the optimal parameters for rate estimation of the same context of the same basic-block via (3.25).
2. CPB: The quantized coefficients for each context from N preceding basic-blocks (in coding order) and the current basic-block (for which the rate estimation is performed) are used to estimate the optimal parameters for rate estimation via (3.26). We used N =3 to generate the reported results in this section.
3. EOP: The optimal CBO parameters for each context of the basic-block for which the rate is being estimated and the N previous basic-blocks (in coding order) are averaged to estimate the optimal parameters as suggested in (3.27).

As in the case of CPB, we chose N =3 for experimental results, reported in this section.

Table 3.1 shows the relative coding efficiency and the relative computational complexity of the three proposed MLPE rate estimation methods for optimal mode selection. The relative coding efficiency is measured based on the increased coding bit-rate if the mode selection would have performed with the actual CABAC rate. To find the bit rate increase, the recommendation in [Bjorn01] was followed. Also the relative computational complexity is measured by calculating the percentage of CPU cycles saved by performing mode selection task, using the estimated rate, if the reference value would be the CPU cycles required to perform the optimal mode selection using the actual rate calculated by CABAC.

Table 3.1 Experimental Results: Coding efficiency of different rate estimation methods compared with exact rate calculations.

Video Sequence	Change Percentile	CBO	CPB	EOP	SATD
Mobile 352x240	Rate increase %	1.3	0.8	0.7	3.54
	Speed increase %	94	93	94	96
Mobile 720x1280	Rate increase %	1.8	0.9	0.9	4.11
	Speed increase %	93	92	93	95
Cycling 1280x720	Rate increase %	5.3	2.0	1.8	7.42
	Speed increase %	93	92	92	95
Walking Couple 1920x1080	Rate increase %	6.1	2.3	2.1	8.23
	Speed increase %	93	92	92	94

To put the test results in perspective, Table 3.1 also provides information on relative coding efficiency and relative computational complexity for the case that the optimal mode selection is done by minimizing the SATD. This method treats the sum of absolute values of Hadammard coefficients of the residual block data as the cost function. This method is part of the reference implementation of the H.264/AVC codec to provide a low complexity approach to mode selection (and also motion estimation).

As can be seen, the CBO is out performed by CPB and EOP. While CPB and EOP performance are very comparable, one may choose EOP due to simpler parameter estimation scheme. Also Table I shows that all three rate estimation methods are comparable to SATD in terms of computational complexity, while they significantly out perform the SATD in terms of coding efficiency.

3.6 Rate Estimation Applications: Rate Control within H.264

In 3.5.1, we observed that the proposed rate estimation method based on MLPE can accurately predict the entropy rate for coding the data within a macroblock or even a block of 4x4. This observation is encouraging to apply the MLPE rate estimation method for rate control. The purpose of rate control is to assign bit-budget to entities that constitute a coded video or image bitstream. Examples of such entities are: Group Of Pictures (GOPs), frames, slices and macroblocks (in the case of block based compression schemes) or sub-bands (in the case of sub-band and wavelet compression schemes). Based on the assigned bit-budget, a quantization parameter would be decided for compression of the Basic Rate Entities (BREs). A BRE is the smallest coding entity for which a unique quantization parameter can be assigned (BRE is a.k.a. “basic-units” in the H.264/AVC reference codec). The goal of rate control is to assign the quantization parameters to each BRE. Note that in the H.264/AVC a BRE can be as small as a macroblock.

3.6.1 A Bit-Budget Allocation Algorithm based on MLPE

To find the quantization step size (quantization parameter) for each BRE the following steps should be taken:

1. Estimation of Laplacian distribution parameter: Within each BRE one can calculate the pdf parameter λ of the assumed Laplacian distribution. For this purpose the image should be quantized with an arbitrary but small step size Δ_0. Assuming that the MAV and the percentage of zero symbols for the BRE of interest are γ_0 and ρ_0, respectively, one can use (3.4) along with (3.13) to find λ as follows:

$$\lambda = \frac{1}{\Delta_0} \ln\left(\frac{\gamma_0}{\gamma_0 - (1-\rho_0)} \right) \tag{3.28}$$

2. Calculation of ρ at a given quantization step size: There are two options to calculate the ρ value. The data driven approach finds the value of ρ by building a histogram of symbols in step 1 (when the compression is performed with quantization step size of Δ_0) and finding what percentage of symbols is smaller than $(\Delta+\delta)$. Alternatively one can calculate ρ from (3.4) which yields:

$$\rho = 1 - e^{-\lambda(\Delta+\delta)} \tag{3.29}$$

Note that for the first approach, to have sufficient resolution one needs to choose Δ_0 to be very small compared to the range of possible Δs. Also note that for the second approach λ is calculated in step 1 and $(\Delta+\delta)$ is given by the quantization parameter.

3. Calculation of γ at arbitrary quantization step size: By replacing the β from (3.4) in (3.13), one can find the γ at quantization step size Δ, as follows:

$$\gamma = \frac{(1-\rho)}{(1-e^{-\lambda\cdot\Delta})} \Rightarrow \bar{\gamma} = \frac{1}{(1-e^{-\lambda\cdot\Delta})} \tag{3.30}$$

Note that distribution parameter λ is calculated in step 1, and the quantization step size Δ is given by the quantization parameter, and percentage of zero indices ρ is known by step 2.

4. Estimated rate for a given step size: Replacing the ρ and $\bar{\gamma}$ values (found in steps 2 and 3) in (3.21) through (3.23) yields the rate for different components of coded data as in (3.31) through (3.33). The rate would be given by (3.24).

$$R_{si\_ma\_z} = -(1-\rho)\cdot\log_2(1-\rho) \tag{3.31}$$

$$R_{nz\_ab} = -(1-\rho)\cdot\left(\log_2\left(e^{\lambda\cdot\Delta}-1\right) - \frac{\lambda\cdot\Delta\cdot\log_2(e)}{1-e^{-\lambda\cdot\Delta}}\right) \tag{3.32}$$

$$R_{nz\_si} = 1-\rho \tag{3.33}$$

5. Finding the quantization step size which matches a given rate: In order to find a suitable quantization size, one can try different step size and repeat steps 1 to 4 above, till a quantization parameter is found that matches the total estimated rate given by (3.24).

3.6.2 Comparison with ρ-domain Rate Estimation technique

As mentioned in 2.2.2 the ρ-domain rate estimation [HeKM01] is one of the most accepted rate-estimation methods for rate control. This method provides reasonable accuracy in estimation of rate at low data rates. Here we review how the method works and then compare this method against the one which was proposed in 3.6.1 based on the MLPE.

1. Rate indicator for non-zero coefficients, Q_{nz}: It has been observed that the average number of bits for binary representation of the non-zero coefficients (Q_{nz}) has a linear relationship with the percentage of non-zero coefficients $(1-\rho)$ or $Q_{nz} = \theta \cdot (1-\rho)$. Note that θ can be assumed constant for low data rates for a given image.
2. Rate indicator for zero coefficients Q_z: This indicator has a polynomial relationship with θ. Furthermore, coefficients of the polynomial depend on the percentage of zeros by the following equation: $Q_z = \sum_{n=0}^{3} \alpha_n(\rho) \cdot \theta^n$. $\alpha_n(\rho)$s are given based on statistical regression for certain values of ρ.
3. Finally the rate can be estimated by a linear combination of Q_{nz} and Q_z as follows: $\tilde{R} = \beta_z(\rho) \cdot Q_z + \beta_{nz}(\rho) \cdot Q_{nz} + \beta_0(\rho)$. The linear model parameters $(\beta_z(\rho), \beta_{nz}(\rho), \beta_0(\rho))$ are chosen based on statistical regression for certain values of ρ.

A heuristic justification for derivation of the above steps can be found in [HMit02]. Here we compare the ρ-domain method to the proposed method by making the following observations:

1. Range of operation: Notice that 0.89 to 1 is the only range of ρ that the parameter values ($\alpha_n(\rho)$ and $(\beta_z(\rho), \beta_{nz}(\rho), \beta_0(\rho))$) are given in [HeKM01]. Also the mathematical justification for the linear model in [HMit02] is conditioned on the value of ρ being very close to 1. In contrast the proposed

method in 3.6.1 works for the entire range of rates. Note should be taken that although equation (13) in [HMit02] suggests that the linear relation is a good estimate for all values of ρ (i.e. all the rates) the flaw in proving equation (13) dismisses this claim. Further explanation on this faulty proof is given in bullet 5 below.

2. Order of Entropy: The derivation in [HMit02] relies on the first order entropy of all coefficients. As discussed in 3.5 most video compression schemes compress the location of zeros of the transmitted symbols with higher order entropies. The proposed method considers this fact in 3.5. The method in [HeKM01] attempts to compensate this shortcoming by a heuristic mapping of θ to Q_z through a third order polynomial regressor.
3. Lack of generality: The parameters for linear estimation of the rate $(\beta_z(\rho), \beta_{nz}(\rho), \beta_0(\rho))$ and the parameters ($\alpha_n(\rho)$s) for regression of Q_z are derived by statistical regression on a given set of test images, for a specific compression scheme (H.263). To use the same method with another compression scheme, one would need to compute all these values by learning those parameters for every encoder, which is not a trivial task.
4. Probability model: Although both methods in [HeKM01] and in 3.6.1 assume a Laplacian distribution of non-quantized residual coefficients (symbols), the MLPE method discussed in 3.3 and 3.4 can be generalized to any distribution, including the generalized Gaussian and Mixture of distribution models (Look

at 3.1). On the other hand the adoption of any other probabilistic model (other than the Laplacian case) in the heuristic method in [HeKM01] is ambiguous.

5. Mathematical proof: The heuristic proof given in [HMit02] to derive the linear relationship between rate and the value of ρ based on equation (7) in the same paper is faulty. The flaw begins in Section IV part A of [HMit02] where the Shannon Lower Bound (SLB) R-D function is used for a non MSE distortion. There are two problems with this proof. First, the distortion used for proof in equations (8) through (13) is based on Mean Absolute Error (MAE). However the R-D optimal coding in video coders is done according to the MSE not MAE. Secondly the SLB is not proven to be converging to the entropy rate of a UQEDZ scheme. In fact it has been shown [GeGr91] that for the MSE distortion metric and for a Gaussian source, the SLB converges to the R-D which can be achieved by a uniform quantizer only at high data rates. In contrast to [HMit02] the proposed method in this work (3.3 and 3.4) is based on a solid mathematical foundation provided by MLPE.

In conclusion the proposed rate estimation method in 3.6.1 is a robust rate estimator across the whole rate-spectrum of rates from very low to very high data rates. Also the proposed method is founded on a mathematical ground based on the MLPE which does not depend on some statistical regression on a limited set of data for a given compression scheme. Again it needs to be emphasized that for every set of compression scheme and video data, which results in quantization of symbols with Laplacian distribution the proposed method will work well.

3.7 MLPE Rate Estimation Beyond Laplacian Distribution

Using a more general probability distribution function such as a generalized Gaussian density function or a Gaussian mixture model can improve the accuracy of the rate estimation. Considering that the Gaussian mixture model results in the Laplacian distribution of the coded data [LaGo00], it would be interesting to observe if employing one of the other candidate distributions for image and video data (look at 3.1) would help the accuracy of rate estimation, discussed in 3.6.1.

This chapter, in part, contains segments from the following submitted papers:

- Minoo, K.; Nguyen, T.Q., “Optimal Mode selection via Maximum Likelihood rate estimation: Application IN fast mode-selection within H.264,” Signals, Systems and Computers, the Forty-second Asilomar Conference on, ACSSC 2008. Accepted for publication, 2008.

- Minoo, K.; Truong Nguyen, "Maximum Likelihood Rate Estimation: With Applications in Image and Video Compression," Data Compression Conference, 2008. DCC 2008 , vol., no., pp.535-535, 25-27 March 2008.

Chapter 4
A Review of Concepts and Methods for Visual Quality Assessment

During the past three decades, many perceptual models for objective assessment of visual quality have been proposed. Most of these models are either too complex to be implemented in real-time applications, or they are too application-specific (e.g., as in medical imaging) to be of any practical use in general cases. Most practical perceptual models for image and video processing (coding) applications, in recent literature, are based on the concept of JND. The principal idea for introduction of the JND concept comes from the fact that the HVS can tolerate a certain level of change in an image (distortion) before the change becomes observable by average human observer. The amount of this change is the JND and it depends on the spatial masking characteristics of the HVS.

In this chapter, we first review the results of psychophysical experiments which quantify the HVS sensitivity at JND level to simple patterns, in the absence of any image texture. Next we review the studies which relate the effect of image texture on the HVS sensitivity. We conclude this chapter by reviewing different classes of image quality methods and metrics.

4.1 HVS Sensitivity

Many psychophysical experiments have been conducted to better understand the sensitivity of the HVS to changes in different attributes of a visual experience. Most of these psychophysical experiments focus on a particular characteristic of the HVS which

influences the perceived notion of distortion. In this section we review some of the relevant characteristics of the HVS which influence the perceived distortion.

4.1.1 Luminance Sensitivity

In 1834, Weber discovered that the perceived change in the weight, lifted by a person, is inversely proportional to the weight of the object before the change. Later on, Weber's observation was formalized by Fechner and proved to be correct for other physiological sensory stimuli. In the case of HVS, the Weber-Fechner law states that the just noticeable changes in the intensity of an area, is proportional to the average intensity of that area. If ΔI is the statistical minimum perceived change (or JND) in the luminance and I_o is the average luminance, then the Weber law can be formulated as $\Delta I / I_o = \alpha$ where α is the Weber constant which is reported to be around 0.02 for intensities in photopic range. Note that the HVS sensitivity is the inverse of ΔI, and as such, the Weber-Fechner law captures the masking effect of the intensity, which assigns a higher JND to an area with higher average intensity value.

The Weber-Fechner law implies that human's neural system perceives a sensory stimulus based on a power-law. This was proved to be the case by published studies of Stanly S. Stevens [Stev57], who popularized the Stevens' power-law based on psychophysical experiments over a wide range of stimuli types. The general form is expressed as $P(I) \propto I^a$ where $P(I)$ is the perceived intensity, I is the intensity and a is a constant which, for experiments in photopic ranges, is assumed to be around 0.5.

Luminance masking is one the most effective types of masking. In fact, it has been suggested that other types of perceptual distortion masking such as contrast or

texture masking need to include the effect of luminance masking in order to find a better sensitivity measure to distortion [KaKi95]. It is important to note that for calculating the intensity masking from image pixel values, one has to take into consideration the gamma effect of the display devices and the gamma correction which is introduced by recording devices. As noted in [TaMa02], because of a fortunate coincidence, the gamma correction and luminance perception work inversely to cancel each other. This allows the gamma corrected pixel values to better represent perceived changes in luminance by the HVS.

4.1.2 Contrast Sensitivity

A large body of published work in the area of perceptual image and video processing are results of psychophysical experiments that measure the average human sensitivity to spatial intensity contrast superimposed on a uniform background by a sine wave (usually of vertical orientation) at different angular frequencies[*]. The experiments of Robson and Kelly [Robs66], [Kell77] showed that the JND varies based on the spatial angular frequency of the superimposed stimulus. Based on these observations, a spatial Contrast Sensitivity Function (CSF) can be defined which estimates the minimum contrast that is observable by average humans, (can be thought of as JND) for a given angular frequency. The CSF has received extensive attention for perceptual quality assessment of natural images based on the concept of normalized error in the wavelet domain [WYSV97] or the DCT domain [AhPe92], or the Fast Fourier Transform (FFT) domain [MENS06].

In all the aforementioned transform domains, each error coefficient would be normalized based on its position which signifies the spatial frequency and orientation of

[*] Angular frequency is expressed as the number of oscillations per unit angle and measured by cycles per degree (cpd).

that error. A norm-p of the normalized error would pool the distortion across different coefficients. As explained in [HoKa02] this norm-p summation can be justified by the probability summation rule [RoGr81], [Wats79].

4.2 Texture Masking and Divisive Gain Control

In [Wats93], Watson acknowledged that the CSF based on the contrast sensitivity for individual DCT coefficients on a smooth background does not fully exploit the HVS properties for efficient perceptual video coding. In practical coding scenarios the perceived distortion is influenced by the background texture for the same spatial location. To include the effect of texture masking on the overall JND for each coefficient, Watson considered the masking effect, only from the collocated coefficient in the same block. This formulation implies that the only texture masking factor is the collocated coefficient in the DCT domain. Obviously this assumption ignores the effect of overall background texture which is influenced by other coefficients in the same block and some of the neighboring blocks.

Divisive gain control (or divisive normalization) [Wats97] extends the influence of image texture on the perceived changes (error/distortion) in an image. In this model a set of linear filters (e.g. Gabor filters) are applied to an image to measure the strength of different features (represented by different Gabor filters). Finally the divisive gain control reduces the error in feature strength based on the presence of features in surrounding locations (pooled by a linear kernel) via a power law.

Divisive gain control can be interpreted as a generic model for considering the effect of texture masking. In this generic model the changes (in the pixel domain or the

feature domain) are normalized based on the presence of other features or pixel activities. In this section we consider some of the existing models for texture masking.

4.2.1 Divisive Gain Control for Block Transform Features

Malo et al. in [MENS06] proposed divisive normalization for efficient image coding. In their model, the features are the coefficients of a block-based transform such as Fourier transform. The block transform coefficients are then normalized based on the weighted sum of all the coefficients within the same transform block. The weights are assigned to fit the result of psychophysical experiments where the mutual masking effect between two coefficients is measured.

One point of concern in the proposed divisive gain control of [MENS06] is that the attenuation term is calculated based on a weighted sum of all coefficients, where the weighting factor is derived from experiments only involving two coefficients. It would be most likely the case that the interaction of three or more coefficients introduces a different weighting for summation of the effect of other coefficients in the perception of any given coefficient.

There is yet another concern with block based transform distortion measures such as [Wats93] and [MENS06]. The concern originates from the fact that in these measures, each coefficient-error (e.g., in Fourier or DCT domain) gets normalized by a function of the coefficients, solely within the same transform block. The spatial-frequency uncertainty principle causes a large transform-block size to have less accurate local information in the pixel domain but more accurate transform domain information and vice versa. As a result, the selection of the right size block for transformation is very

crucial. A given transform-block size is considered too large if different parts of that block, in the pixel domain, show different perceptual properties (e.g. intensity or texture masking properties). This fact results in a condition where a change in the value of a given coefficient would be perceived differently based on the perceptual properties of different parts of the image within the same block. A given transform-block size is considered too small, for a given viewing condition, if the perceptual properties for each block (such as intensity and texture masking) are also influenced by the neighboring blocks. In such cases the divisive gain control based on the value of the coefficients within the same block is not descriptive of the texture masking effect.

4.2.2 Divisive Gain Control Based on Block Texture Activity

An alternative approach to include the effect of image content (texture) on the perceived distortion is to use the concept of error (distortion) normalization based on a generic texture activity measure. The dominant perceptual coding methods based on texture activity masking often normalize the MSE distortion by a function of texture activity to represent the distortion in a perceptual sense [RaHe01], [Lubi97], [ZhLX03]. In this context one needs to define a texture measure and a function that relates the defined texture measure to a normalization factor. Here the goal is to find a normalized MSE metric that resembles the perceived distortion by the HVS.

There are many proposed candidates for the texture measure. In this work we use the most prevalent definition of texture measure which is based on the standard deviation of pixel values in an image block. To normalize the MSE based on this texture measure, threshold functions have been widely used in practice. These functions use the texture

measure and a set of threshold values to classify a block of the image into one of several classes and assign a fixed distortion-normalization factor to the corresponding block based on that class [ZhLX03], [CLCR93], [TaPN96].

4.3 Review of Image Quality Assessment

There are many published works on the subject of image quality assessment. Here a brief review of the methods and the measures which dominate the field of image quality assessment is presented.

4.3.1 Image Quality Methods

In terms of how the quality of a test image is assessed, one can distinguish the following three classes:

1. Full Reference Image Quality (FRIQ): In this category, the quality assessment is done with complete information about a reference frame. Most often the methods in this category measure a "distance" between the test image and the reference image. The methods in this category are often used for comparison of different image processing or image coding techniques.
2. No Reference Image Quality (NRIQ): In this category, a single test image is evaluated for quality assessment. Most often the methods of this category measure the "deviation" of test image from an expected set of statistics. The methods in this category are often used to perform image post processing enhancement and restoration. In image and video communication this class of measures can be used to efficiently request re-transmission of the corrupted or missing data, if it causes significant degradation of quality.

3. Partial Reference Image Quality (PRIQ): In this category a single test image is evaluated based on partial information of the reference image. The partial information could be a sub-sampled representation of the reference image in some transform domain, or it can be some statistical information on the "features" (as will be explained in 4.3.2) of the reference image.

Although these three methods traditionally have been treated separately (by utilizing different metrics, as will be explained in 4.3.2), in Chapter 6 a framework is introduced by which these three methods can be viewed as different instances of a more general framework which we call: Probabilistic Perceptual Image Quality (PPIQ) framework.

4.3.2 Image Quality Metrics

Image quality metrics can be discussed based on the "basic quality" which would be measured and the approach by which these "basic qualities" collectively influence the overall quality of the image (or video).

4.3.2.1 Basic Quality Measure

The basic quality which is measured in different quality assessment methods can be categorized into one of the following three classes:

1. Local Error Visibility (LEV): The fundamental characteristic which is measured in this class of image quality metrics is the difference (error) between the color component values of a test image and those of a reference image. Alternatively, more advanced methods measure the difference between

a filtered version of the test and reference images. Examples of those filters are referred to as “contrast sensitivity filters” in [WaAh05].

2. Feature (or Channel) Decomposition (FCD): In this category the images are evaluated in a feature space. The features (or equivalently, channels) can be derived by decomposing the image through a transform, where the coefficients of the transform each constitute a feature. The measure can evaluate the difference (error) between corresponding features of test and reference images (in FRIQ methods). The measure can also quantify the deviation of the statistics for a given feature of a test image from an expected statistics (e.g., in NRIQ and PRIQ). Note that depending on the nature of features, the image representation in the feature space can have greater (over complete), fewer (under complete) or equal (complete) number of features compared to the number of pixels in the image.
3. Local Structural Similarity (LSS): The basic characteristic which is measured in this category is the deviation of local area of a test image from the corresponding area of a reference image. In contrast to LEV, the measure of deviation in LSS is not based on the difference between the two images, rather the cross-correlation between them. As one can see, the LSS class is similar to the LEV class in the sense that each local area in the test and the reference image is convolved with a filter. The filter in LEV case is a CSF and in LSS case is a matched filter, based on the corresponding area in the reference image. Then the error between the convolved local signals of the test image

and reference image is subtracted to find the deviation of the test image from the reference image.

4.3.2.2 Measure Pooling:

In order to assign a single value to the quality of an image, it is required to influence the effect of each quality measure on the overall quality of the image. According to 4.3.2.1, there are two fundamental spaces for which we have to pool (combine) the effect of individual basic measures.

1. Spatial pooling: LEV methods require that the individual local errors (basic measures) be pooled to form a single value that covers the image quality across the entire image.
2. Feature (channel) pooling: FCD methods require that the individual feature quality measures be pooled to form a single value that represents the overall quality of the image for all the features.

Note that since most perceptually effective FCD methods use localized transforms (such as wavelet and block-based DCT) to define the features, in practice error pooling for FCD methods should take place across both spatial and feature spaces. In 6.2 when the PPIQ framework is introduced, we address perceptual pooling strategies in more detail.

Chapter 5
A Perceptual Distortion Metric Based on the Local Texture Spread

The problem with the method of texture masking as explained in [MiNg05] has to do with the fact that the HVS is capable of tuning to localized features within a block if the angular size of the block is large. For example a block containing checkerboard pattern and another block with a simple vertical contrast of half black and half white pixels will be classified to the same texture activity class while these two blocks have different texture masking properties.

This observation in [MiNg05] led us to suggest a new texture masking measure based on the texture spread within the block of interest (i.e., a macroblock in [MiNg05]). The main logic behind this proposal was as follows: A block is classified as Texture (with higher resiliency towards distortion), if the average texture value for smaller constituting sub-blocks within that block is high and most sub-blocks have almost the same texture (high) values (i.e. the variance of texture values is low). The blocks with a mixture of high and low texture sub-blocks (i.e. the variance of texture values is high) contain detail information where distortion is more obvious compared to the previous case. Finally if an area is smooth, it means that the average of texture values from all sub-blocks is very small.

The idea of fixed sized, block classification (e.g., macroblock classification) introduces several problems. First, the idea of assigning all regions within a block the same perceptual masking effect is not quite realistic. A macroblock might be too large or

too small for perceptual consideration as previously explained in 4.2.1. Second, in general for classification algorithms, different threshold values would be needed for different image sizes, or for the same image sizes being observed from different distances, resulting in different angular resolutions. The need to find different empirical threshold values for different viewing conditions becomes especially cumbersome at high angular resolutions (e.g. high resolution images). For this type of viewing condition, all blocks appear to be smooth (low texture activity). This fact results in texture values that are very small for all sub-blocks within any macroblock. The small texture values in turn, yield threshold values, for distinguishing between different classes, to be very close to each other. This fact makes the classification very sensitive to possible model errors for a given texture measure.

5.1 Local Texture Spread as a Measure of Texture Masking

The justification provided in [MiNg05] to utilize the average and the variance amongst sub-blocks in a larger block for calculating the amount of masking, combined with legitimate concerns about fixed block-based perceptual classification methods (elaborated in the previous section), have led us to propose a better method for measuring and applying the distortion masking property for different regions of an image. The new proposed method defines a measure, called Local Texture Spread (LTS), and a function (called mapping function) to predict the HVS's sensitivity to distortion based on the LTS for different regions of the image. Subsequently the predicted distortion-sensitivity will be used to define a perceptual distortion metric.

5.1.1 JND According to the LTS

In essence, the LTS is the ratio of the average textures to the standard deviation of those textures amongst the sub-block (basic blocks) within a larger area (texture perceptual support area). To calculate the LTS, we start with the introduction of a perceptual basic block. A perceptual basic block or simply a “basic block” is the smallest rectangular area which shows similar perceptual texture masking property. Obviously the size of the basic block (in terms of number of pixels) is chosen based on the given angular resolution, which is calculated from the viewing distance and pixel pitch. For example in the subjective test environment, (specifics are given in 5.1.4 and 5.1.5), the optimal size of basic blocks was found to be 4x4 pixels. As a result, if for a given viewing condition the angular resolution doubles, compared to the test setting, then the size of basic blocks needs to be doubled too (basic blocks would then have a size of 8x8). Calculation of the LTS starts with partitioning a picture into basic blocks. One way to perform this partitioning is to start at the top left corner of the image and span the whole image in raster scan order from left to right and top to bottom. The image might be extended, symmetrically, to provide an integer number of basic blocks per row and column, as shown in Figure 5.1.

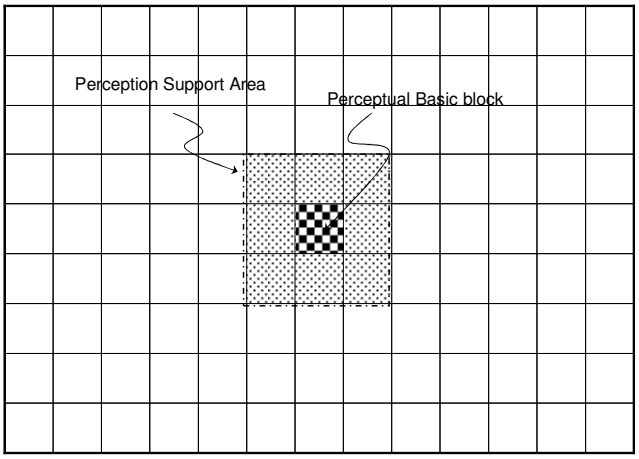


Figure 5.1 Perceptual basic block (checker board), texture perception range (dotted blocks).

Once the partitioning of the image is done, each basic block will be assigned a LTS value, based on the average of texture values and the variance of texture values amongst the neighboring basic blocks of the basic block of interest. We call these neighboring areas the Texture Perception Support Area (TPSA). It is intuitive to assume the size of the TPSA is fixed in terms of the number of basic blocks they contain, for different viewing distances and display resolutions. Note that the size of the TPSA, in terms of the number of pixels, actually changes as a result of changes in the size of the basic block for different viewing conditions.

The average and the standard deviation of the texture values for basic blocks, within the TPSA of the k-th basic block, is denoted by $\mu_{txt}(k)$ and $\sigma_{txt}(k)$ respectively.

Denoting $\sigma_p(k) = \left(\frac{1}{N_{bb}} \cdot \sum_{i \in bb(k)} \left(x_i^{\,2}\right) - \left(\frac{1}{N_{bb}} \cdot \sum_{i \in bb(k)} x_i \right)^2 \right)^{1/2}$ as the standard deviation of the

pixel values inside the k-th basic block ($bb(k)$) when the number of all pixels in each basic block is N_{bb}, we have:

$$\mu_{txt}(k) = \frac{1}{N_{TPSA}} \sum_{j \in I_{TPSA}(k)} \sigma_p(j)$$
$$\sigma_{txt}(k) = \left(\frac{1}{N_{TPSA}} \sum_{j \in I_{TPSA}(k)} (\sigma_p(j) - \mu_{txt}(k))^2 \right)^{1/2} \quad (5.1)$$

where N_{TPSA} is the number of basic blocks in a TPSA and $I_{TPSA}(k)$ is the set of indices for all the basic blocks which belong to the TPSA of the k-th basic block. Note that this set includes "k ".

We observe that the k-th basic block is located in a highly textured area (tolerating more distortion), if almost all basic blocks in the corresponding TPSA have large texture (σ_p) values. This situation translates to observing a large $\mu_{txt}(i)$ value and a relatively small $\sigma_{txt}(i)$ value. Similarly a basic block would belong to a detail area (with medium tolerance for distortion), when there are some scattered edges and lines on an otherwise smooth area (e.g. when that area contains alphabetical or numerical characters). Therefore a basic block belongs to a detail area if some of the basic blocks in the corresponding TPSA have high texture (σ_p) values (because of lines or edges) and the other basic blocks in that TPSA are smooth (having low σ_p values). This arrangement of basic blocks suggests a μ_{txt} which is smaller and a σ_{txt} which is larger than the corresponding values for a texture basic block. Finally for smooth areas which are the most sensitive to distortion, all basic blocks are smooth (having low σ_p s) which results in a μ_{txt} that is much lower than the value for texture and detail areas.

Based on the preceding argument, it is proposed to define the LTS measure (noted by ψ) as the indicator of sensitivity towards the distortion as follows:

$$\psi(k) = \frac{\mu_{txt}(k)}{\sigma_{txt}(k) + c} \tag{5.2}$$

where $\psi(k)$ is the LTS for the k-th basic block and c is a small constant to regulate the behavior of the $\psi(k)$ when $\sigma_{txt}(k)$ is very small (i.e. for smooth areas). This constant term also prevents a possible division by zero. Next, we describe the context in which the LTS measure is used for normalization of the MSE distortion to perceptually assess the quality of a reconstructed image.

5.1.2 Relevance of the LTS Measure in Image Quality Assessment

As explained previously, the LTS measure seems to be a good indicator for JND based on texture masking. On the other hand many distortion models have suggested the use of the JND [HoKa02], [Wats93], [ZhLX03] for normalization of MSE (or other norms of the error) to define an objective distortion metric. At first we establish the relevance of the LTS to the perceptual distortion metric by estimating the JND from the LTS measure which is later going to be used to derive the perceptual weighting factor, $\xi(i, j)$ in (5.3).

$$d_{ROI}^{p} = \frac{1}{N_{ROI}} \sum_{(i,j) \in S_{ROI}} \frac{\left| x_{tst}(i, j) - x_{ref}(i, j) \right|^2}{\xi(i, j)} \tag{5.3}$$

In (5.3), S_{ROI} is the set of pixel coordinates in a given Region Of Interest (ROI), N_{ROI} is the number of pixels and d_{ROI}^{p} is the overall perceptual distortion in that ROI.

$x_{ref}(i,j)$ and $x_{tst}(i,j)$ are pixel values at location (i,j) in the reference image and the test image respectively.

5.1.3 JND According to the LTS Measure and Weber Law

To find the JND in the pixel domain, we need to consider two basic types of masking by the HVS. The first type is texture masking which, in this work, is conjectured to be a function of the LTS measure and the second type of masking is luminance masking. By adopting a similar approach as in [Wats93], the combined effect of these two types of masking can be computed by the following equation:

$$\xi(i,j) = T_{JND}(k) = T_{TMF}(k) \cdot \left(\frac{I(k)}{I_o}\right)^{a_T} \quad (i,j) \in k \tag{5.4}$$

In the above expression, $\xi(i,j)$ is the normalization factor which will be used to perceptually adjust the squared-error at pixel position (i,j). The term $I(k)$ is the average intensity over the TPSA of the k-th basic block which includes the pixel position (i,j). $T_{JND}(k)$ denotes the minimum average distortion, over the k-th basic block (measured by the MSE), which would become noticeable by average human observer. Note that based on this definition (and throughout the rest of this chapter) we use the term "JND" to refer to $T_{JND}(k)$. In (5.4), $T_{TMF}(k)$ is the JND if the average intensity is I_o. The parameter a_T captures the non-linear nature of the luminance masking which includes Steven's power law and the gamma effect of the display device (as explained in 4.1.1).

The concept of the JND at a reference intensity allows us to treat $T_{TMF}(k)$ as the Texture Masking Factor (TMF). Therefore equation (5.4) separates the effects of texture

masking and luminance masking. The independence of texture masking and luminance masking in the formulation of the JND as suggested by equation (5.4) enables the estimation of the function $S(.)$ which maps the LTS value, $\psi(k)$, to the TMF value, $T_{TMF}(k)$, as follows:

$$T_{TMF}(k) = S(\psi(k)) \tag{5.5}$$

Next, the methodology which was followed to find the function $S(.)$ is explained.

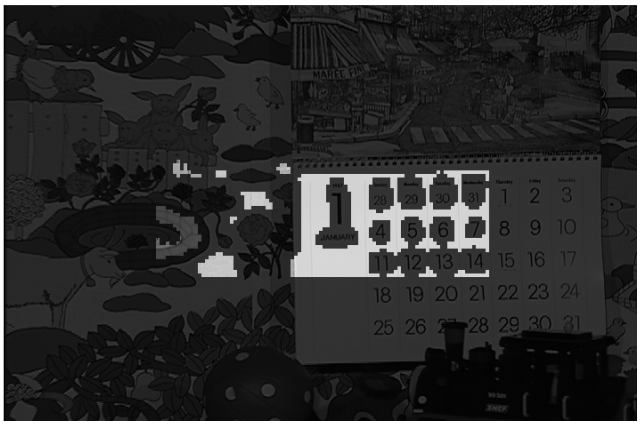
5.1.4 Methodology to Find TMF as a Function of LTS

For the purpose of experimental studies, the range of possible LTS values was divided into N_{LTS} bins (one can think of this as a non-uniform quantization of the LTS value). For any given original image, N_{LTS} test-masks were generated, where each mask corresponds to one of the LTS bin-indices and is referenced by the corresponding bin-index. Note that these test masks will be used to find the JND based on the LTS value. The procedure to generate the test-masks follows these steps:

1. Each original image is first partitioned into basic blocks (as mentioned in 5.1.1) and based on (5.2), a real value (the LTS measure) is assigned to each basic block.
2. To generate the m-th mask, initially, a 2D mask is created as an empty mask (with the same size as the original picture) by assigning zero to all the pixels in that mask.
3. A value of one is assigned to all the pixels that belong to a basic block with a LTS value which would be quantized to the m-th bin.

4. If the number of pixels with value one (active pixels) is less than 5% of the total pixels for a test-mask, then that mask would be set to null by changing all the pixel values to zero.
5. If the number of active pixels in a test-mask is more than 5% of the total pixels, the number of active pixels is reduced by admitting only the active pixels from the basic blocks with smaller distance from the center of the image, until the total number of admitted active pixels is 5% of the total number of pixels. (Remaining pixels would be reset to zero).

Enforcing steps 4 and 5 above guarantees that the number of pixels in all active (non-null) masks, would be the same. The significance of this requirement is explained later when the actual test procedure is discussed. Furthermore, since natural images tend to have a larger number of basic blocks with lower values of the LTS (i.e. larger smooth areas), a non-uniform quantization of the LTS values was employed. Selection of the bin-boundaries is done with two goals. First, to make sure that each bin contains at least 5% of all pixels (to have a good size statistics for all bins, especially at larger LTS values), and second: the quantization is not very coarse (especially for smaller LTS values) so that we get a meaningful representation of LTS value by each bin. In the experiments ten bins were used with the following set of quantization boundaries: {0, 0.2, 0.5, 0.9, 1.4, 1.9, 2.4, 3.0, 4.0, 5.5, ∞}. Figure 5.2 is an example of two test-masks for frame #45 of the mobile and calendar sequences.



(a)

(b)

Figure 5.2 Test-masks for frame # 45 of Mobile&Calendar 720x480 for LTS ranges of (a) [0,0.2] and b) [1.4, 1.9].

In order to create test-images for the experiments, first, several distorted versions of each original image (at different distortion levels) were generated. To create these distorted images, an original image was reconstructed by employing different quantization step sizes in a block-based DCT coding scheme (JPEG). Then one test-image was created for each pair of one distorted image and one valid (non-null) test-mask (both corresponding to the same original image). Each test-image would be identical to the original image except for the active pixel locations of the test-mask, where the image

is identical to the distorted image. In other words, each test-image exhibits a certain level of distortion, only at pixel locations with certain (quantized) LTS value.

To find the TMF, a number of comparative tests were performed, where in each test, one test-image was presented to a test subject (a human observer) for twenty seconds and the test-subject was asked to identify the most obviously distorted area of the test-image (by drawing a polygon using a computer mouse). For every comparative test, the test-subject could compare the test-image vs. the original image by toggling between the two images. The test-image and the original image would appear on the same location on the screen and they would be separated by a monotone gray image which appears on the same location for 0.5 second between each toggle. This strategy to separate the original and the test-image was employed to make sure that the temporal changes between the two images would not hint to the observer to locate the distorted areas. Furthermore, for every test, the first image on the screen was randomly chosen between the original image and the test-image so the experiment would be a blind test. The outcome of each comparative test is an intensity-independent distortion value: $d^{ii} = \frac{1}{N_{ROI}} \sum_{k \in S_{ROI}} \left(\frac{I_o}{I(k)}\right)^{a_T} \cdot \left(d^{mse}(k)\right)$.

The above equation is derived based on (5.3) and (5.4) where $d^{mse}(k)$ is the MSE for the k-th basic block. S_{ROI} is a set of indices for the basic blocks in a given ROI and N_{ROI} is the number of basic blocks in the same region. Moreover, the ROI depends on the test-subject answer in the following way:

1. If the test-subject correctly identifies a distorted area, the ROI consists of all distorted basic blocks within the selected polygon. In this case we denote d^{ii} as the Intensity Independent Noticeable Distortion (IIND).
2. If the test-subject can not find any difference between the original image and the test-image, then the ROI consists of all the distorted basic blocks in the test-image. In this case we denote d^{ii} as the Intensity Independent Un-noticeable Distortion (IIUD).

A test-session consists of a set of comparative tests to find the TMF based on the LTS measure, for one original image and performed on one test-subject. In a test-session, the computer program adaptively arranges a series of comparative tests to narrow the gap between the Maximum IIUD (MIIUD) and the Minimum IIND (MIIND), observed for all test-images which are generated from the same mask and for the given original image. Therefore at the end of a test-session we have a pair of (MIIUD, MIIND) for each test-mask (which corresponds to a quantized LTS value). The Just Observable Intensity Independent Just Noticeable Distortion (IIJND) for each quantized LTS measure is the minimum of the pair: (MIIND and (MIIND + MIIUD)/2) for the corresponding mask.

To summarize, the outcome of a test-session is a set of (LTS, IIJND) pairs, for a given original image and a given test-subject. It is important to note that in the conducted experiments JND is measured as the smallest visible distortion measured by MSE. This approach is not the conventional approach in psychophysical experiments to measure contrast sensitivity, where JND is the minimum visible contrast value.

By averaging IIJNDs (for each quantized LTS) across all test-subjects and all test images we can find the function $S(.)$ by interpolating between the quantized LTS values

we found empirically. This function would map an LTS value to the intensity-independent JND which would be equivalent to the TMF by the definition in (5.5). Next, the experimental set up to empirically find the function $S(.)$ will be discussed.

5.1.5 Experimental Results

A total of 12 test-subjects participated in the conducted experiments from which only three had prior knowledge of image coding. Each test subject participated in two or three test-sessions to evaluate the function $S(.)$ for different original images. There were four original images, two at 720x480 (frame #45 and frame #145 of mobile & calendar sequences) and two images at 512x512 (Lena and Boat). All test-sessions were conducted on two monitors with pixel pitch of 0.294mm, with a preferred viewing distance of 50cm~60cm.

In the conducted experiments it was assumed that $a_T = 0.65$ based on the results in [AhPe92] and the fact that the monitors' gamma correction factors were set to 1.0. In the experiments a basic block size of 4x4 and a TPSA of size 3x3 basic blocks were used.

Following the explained procedure (in 5.1.4) to find $S(.)$, the TMF vs. LTS curves for each of the four original images were collected. The plot in Figure 5.3, shows the $S(.)$ functions, which were empirically found based on the results from all four of the original images and 12 test-subjects.

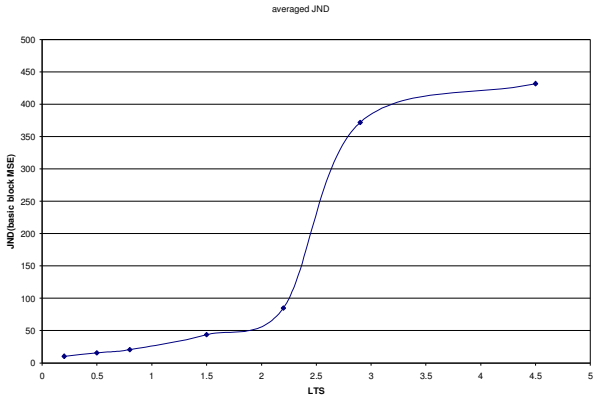


Figure 5.3 Average JND as a function of Local Texture Spread, averaged over all 12 test subjects and 4 images.

An interesting observation is that the sigmoid shape in Figure 5.3 supports the idea of threshold-classification, where each basic block can be classified to one of the two categories based on the LTS value. In this setup one class has higher resilience towards distortion (i.e. higher TMF) and the other one has lower tolerance for hiding distortion (i.e. lower TMF). The advantage of using a smooth continuous function as opposed to a hard threshold is to take advantage of a more effective masking property of the perceptual basic blocks with mid-LTS values (i.e. in the transition region).

5.2 Supra Threshold Distortion Metric According to the LTS

So far we have explored the relevancy of the LTS as a suitable measure to predict the JND based on texture masking. A naive conclusion, as suggested in 5.1.2, would be to use the function $S(.)$ as derived in the previous section to relate the LTS to the JND for the purpose of normalization of the pixel errors between two images. However this approach begs the question that if the normalization of pixel errors based on JND is a valid option in Supra-Threshold regimes, where the perceptual metric should predict the image quality when the distortion is quite visible. The second concern comes from the

fact that in supra-threshold regimes the texture characteristics of the reference image and the test image can be quite different. This raises the question of how we should define the LTS measure. Should it be calculated based on the reference image, test image or a combination of both? In this section these two questions are answered and by a more comprehensive subjective test results the validity of those answers is verified.

5.2.1 Non-linear Mapping from the LTS to Error Normalization Factor

The first concern is addressed by exploring the possibility of normalizing the pixel errors by a parameterized non-linear function, instead of the JND-mapped values in Figure 5.3. In this work it is proposed that in supra-threshold regimes the non-linear mapping function keeps the same generic shape of the $S(.)$ (which was derived empirically for the JND case in the previous section). However the function's parameters would be learned to relate LTS to a normalization factor $\xi(i, j)$ that results in a reasonable estimation of subjective test results in supra-threshold regimes, when (5.3) is used as the distortion metric.

In order to learn the parameters for the new non-linear function, the LIVE database (release 2.0) [SWCBOL] was used. This database includes the subjective test results for a set of 982 images, distorted with various sources of distortion and different distortion strengths (the details of the subjective image quality experiments can be found in [ShBo06]).

There are several classes of functions which can resemble the sigmoid shape curve in Figure 5.3. The main features of that curve are:

1. A low saturation level.

2. A high saturation level.
3. A point at which the transition from low saturation to high saturation reaches the midway between the two saturation levels.
4. A slope which indicates how fast the transition from low to high saturation levels takes place.

Many functions were considered, including a shifted and scaled logistic function for the mapping function. Based on the experimental results it was decided to use the following function in (5.6). Note that (5.6) is obtained by scaling and shifting the visibility probability model which has been used in many perceptual distortion studies, involving the probability summation concept [HoKa02], [RoGr97].

$$T_{TMF}(k) = a + (b-a)\cdot\left(1-\exp\left(-\left[\frac{|\psi(k)|}{\alpha}\right]^{\beta}\right)\right) \tag{5.6}$$

In (5.6), parameter a controls the level of low saturation. Parameter b controls the level of high saturation. Parameter α controls where the transition takes place and finally parameter β controls how steep the transition is.

Once we have these four parameters in (5.6) and parameter c in (5.2), we can find $T_{TMF}(k)$ in (5.5). Using (5.4) and a given a_T, we can find the normalization factor $\xi(i,j)$. The perceptual distortion would be immediately available using (5.3).

5.2.2 Distortion Metric's Symmetry and Selection of LTS from Reference and Test Image

To answer the second question asked at the beginning of 5.2, we start by discussing an example. Imagine a case where an area of the reference image is highly

textured (high LTS), but due to distortions such as quantization (with dead zone) or other types of smoothing process, the same area in the test image has very little texture activity (low LTS). The perceived distortion via error normalization according to the reference image is low and the same perceptual distortion according to the test image is higher. It should be also pointed out that error normalization based only on the reference image (or any non-symmetric function of reference and test image) makes the distortion metric non-symmetric. This means that at supra-threshold regimes the perceptual distortion (distance) between two images depends on which one is the reference frame and which is the test frame.

Since a metric, by definition, has to be symmetric (not dependent on the order of the images being compared), it is more desirable to define error weighting based on a symmetric function of the reference and the test images. Based on this argument three approaches were studied to define the overall normalization factor MSE between the two images, as follows:

1. Error Summation: This approach works based on the conjecture that overall perceptual distortion is the average perceptual distortion according to each of the reference and test images as follows: $d^p = \frac{d_{ref}^p + d_{tst}^p}{2} = \frac{(e(i,j))^2}{2 \cdot \xi_{ref}(i,j)} + \frac{(e(i,j))^2}{2 \cdot \xi_{tst}(i,j)} = \frac{(e(i,j))^2}{\xi(i,j)}$ where $e(i,j)$ is the error between the two pixels in the test and reference images at position (i,j) and $\xi_{ref}(i,j)$ and $\xi_{ref}(i,j)$ are the MSE normalization factors according to the

reference image and the test image respectively. This approach suggests the following formulation: $\xi(i,j) = \frac{2 \cdot \xi_{ref}(i,j) \cdot \xi_{tst}(i,j)}{\xi_{ref}(i,j) + \xi_{tst}(i,j)}$.

2. Max LTS: This approach normalizes the MSE by the larger texture masking normalization factor between the corresponding basic blocks in the two images. This approach can be justified by conjecturing that when two images are compared (e.g., in a double stimuli experiment), the human's perception of distortion is influenced by how much that distortion can be perceived in the image with more masking. For example if the distortion makes the image smoother (less texture activity) as in the case of low pass filtering of an image then one has to use the reference image's texture properties for the purpose of finding the MSE normalization factor. This approach suggests the following: $\xi(i,j) = \max\left(\xi_{ref}(i,j), \xi_{tst}(i,j)\right)$.

3. Min LTS: This approach normalizes the MSE by the smaller texture masking normalization factor between the corresponding basic blocks in the two images. This approach can be justified by conjecturing that when two images are compared (e.g., in a double stimuli experiment), the human's perception of distortion is influenced by how much that distortion can be perceived in the image with less masking. For example if the distortion makes the image noisier (more texture activity) as in case of high pass filtering of an image (or just adding white Gaussian noise) then one has to use the distorted image's texture properties for the purpose of finding the MSE normalization factor. This approach suggests the following: $\xi(i,j) = \min\left(\xi_{ref}(i,j), \xi_{tst}(i,j)\right)$.

In 5.3, the subjective test results were used to verify that Max LTS approach is the most viable choice to assess the distance (distortion) between two images based on the MSE normalization metric suggested in (5.3).

5.3 Empirical Study of LTS to Assess Image Quality at Supra Threshold Distortion Levels

In this section we use the images along with the subjective test results in the LIVE database [SWCBOL] to optimize the model parameters in (5.9) and to find the optimal approach to combine different texture masking from the reference and the test images. Furthermore we use the same database to evaluate the accuracy of the proposed image quality metric against three other choices in prediction of the subjective test results. The LIVE database contains subjective test results measured in Differential Mean Opinion Score (DMOS) for five categories of distorted images along with the corresponding test and reference images. The distortion categories are JPEG compression distortion, JPEG2000 compression distortion, Additive White Gaussian Noise (AWGN) distortion, Gaussian BLur (GBL) distortion and finally JPEG2000 image distortion due to Fast Fading Rayleigh Channel Error (FFRCE) and certain error recovery scheme. The details of the subjective test experiments can be found in [ShBo06]. In the rest of this section the experimental analysis for six distinct distortion groups is discussed. The first five distortion groups represent the data analysis according to each of the five individual categories of distortion in the LIVE database, while the sixth distortion group represents the results based on the data analysis for all the images as a General distortion group.

5.3.1 Optimal Model Parameters via Non-linear Regression

To find the optimal set of parameters for the proposed distortion model we need to associate a cost to a given set of parameters. To that end, for a given set of parameters in (5.6), the distortion is calculated for each test image in the same distortion group and according to (5.3), (5.4) and (5.5). The cost associated with this set of parameters is calculated based on the Root Mean Squared Error (RMSE) to regress the subjective test results from the perceptual distortion calculated for all images in a distortion group (for the given set of parameters). The optimal parameters are those which yield the lowest RMSE. For regression, a non-linear regressor is used as follows:

$$y = a_0 + \frac{a_1}{\exp(a_2 + a_3 \cdot x)} \tag{5.7}$$

Since the regression function in (5.7) automatically takes care of the gain factor in (5.6), we can reduce the number of independent model parameters by one and re-write (5.4) to derive the normalization factor in the supra-threshold regimes as follows:

$$\xi(k) = \left(\frac{I(k)}{I_o}\right)^{a_T} \cdot \left(1 + \gamma \cdot \left(1 - \exp\left(-\left[\frac{|\psi(k)|}{\alpha}\right]^{\beta}\right)\right)\right) \quad \gamma \geq 0 \tag{5.8}$$

The use of data in the LIVE database for optimization of the five parameters (a_T, γ, α, β, c), resulted in the observation that the parameter a_T constantly converges to zero. This observation can be justified by one of the following two arguments. The first argument is to assume the ambient lighting was the dominant factor in each test and the local intensity from the image was not playing a significant part in masking the changes. The second argument has to do with the fact that the formulation of intensity masking as we used in (5.8) has come from the psychophysical experiments at JND regime. One can argue that this will not be as valid in the supra-threshold regimes or when the background

is not a simple monotone gray light and light adaptation time is longer than test time. Given this observation we simplify the equation in (5.8) as follows in the experimental studies.

$$\xi(k)=\left(1+\gamma\cdot\left(1-\exp\left(-\left[\frac{|\psi(k)|}{\alpha}\right]^{\beta}\right)\right)\right)\quad \gamma\geq 0 \tag{5.9}$$

5.3.2 Combining the LTS-based Normalization Factors from the Reference and the Test Image

In 5.2.2 three different approaches were proposed to combine different texture masking normalization according to the reference image and the test image. By optimizing the distortion model parameters (γ, α, β, c) according to each of the three suggested approaches and comparing the RMSE for perceptual distortion metrics, we compare the suggested approaches. Table 5.1 shows the result of this comparison. The result suggests that although MAX LTS is comparable with the other two approaches for image quality assessment in the case of JPEG 2000, JPEG, Gaussian Blur and the error concealment in FFRCE, it significantly outperforms the distortion assessment for white noise and the general distortion case. Consequently, MAX LTS was chosen to perform other studies as follows.

Table 5.1 RMSE for DMOS Regression according to normalization factors combination for the two images

Distortion Category	ERROR SUMMATION	MIN LTS	MAX LTS
JPEG 2000	4.27	4.40	4.46
JPEG	5.08	5.09	5.17
White Noise	3.81	4.49	2.46
Gaussian Blur	4.58	5.03	4.35
JPEG 2000 with FFRCE	5.52	5.67	5.11
General Distortion	7.40	7.55	5.47

5.3.3 Comparison with Other Image Quality Metrics

We compare the proposed image quality metric against two other choices. The first choice is the MSE. The second choice is the SSIM index [WBSS04] using the default parameters in the Matlab implementation of the code [SSIMMC]. To assign the model parameters for the proposed distortion metric, the parameter c in (5.2) was fixed at 20. The optimization was then performed to find the optimal parameters for the mapping function in (5.9). A distinct set of optimal parameters was found for each distortion category. The optimal mapping functions are depicted in Figure 5.4 for each distortion category, including the general distortion case.

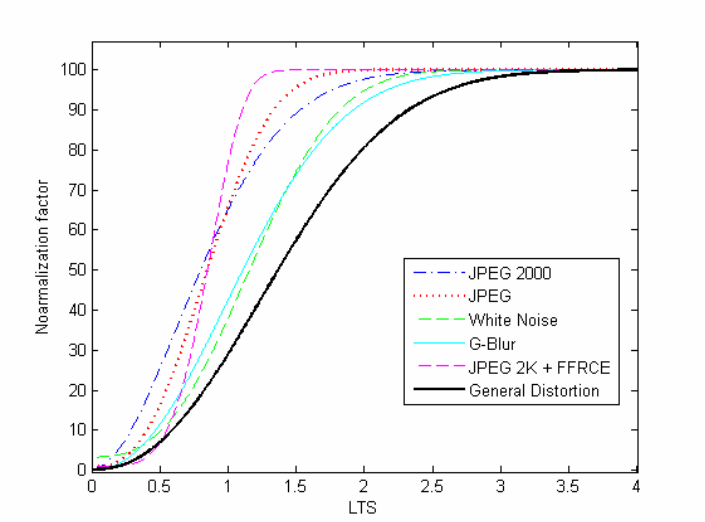


Figure 5.4 Optimal LTS mapping function for different types of distortion.

The results of the comparison between the LTS measure and other distortion metrics are given in Table 5.2. The numbers in Table 5.2 indicate the RMSE of the non-linear regression function in (5.7) for predicting the subjective test result, (in DMOS) from each of the examined image quality metrics, explained above. As observed, the LTS

measure significantly improves the accuracy for predicting the subjective test results in every distortion category, available in LIVE database. To better demonstrate the advantage of the proposed distortion metric, the scatter plot of DMOS vs. each of the rival quality metrics is shown in Figure 5.5. This comparison reveals that SSIM index is doing relatively well to cluster the scatter data for each category, when it is compared with the MSE metric. On the other hand one can observe that SSIM is not doing so well for ranking the image quality between two different categories of distortions. This is most obvious for the case of white noise, when the same SSIM index needs to be mapped to different DMOS depending on the type of distortion. The preceding argument is the reason why SSIM index shows a large RMSE in the General distortion category (close to that of MSE) in Table 5.2. For example a distorted image with white noise (wn\image91.bmp in the LIVE database) has an SSIM index of 0.22 with a DMOS of 50.2 while another image distorted with JPEG compression, (jpeg\imag218.bmp) has an SSIM index of 0.62 and a DMOS value of 60.4, i.e. while SSIM suggests a lower quality for the white noise corrupted image, the subjective test result shows the contrary.

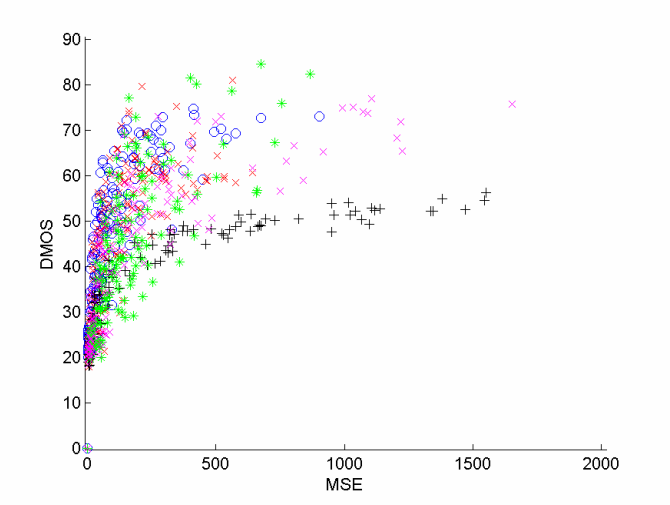
Table 5.2 RMSE for DMOS Regression based on different objective metrics

Distortion Category	MSE	SSIM	LTS
JPEG 2000	7.32	5.79	4.46
JPEG	8.16	6.12	5.23
White Noise	5.01	3.80	2.46
Gaussian Blur	9.93	7.96	4.35
JPEG 2000 with FFRCE	7.82	5.66	5.14
General Distortion	9.22	8.20	5.48

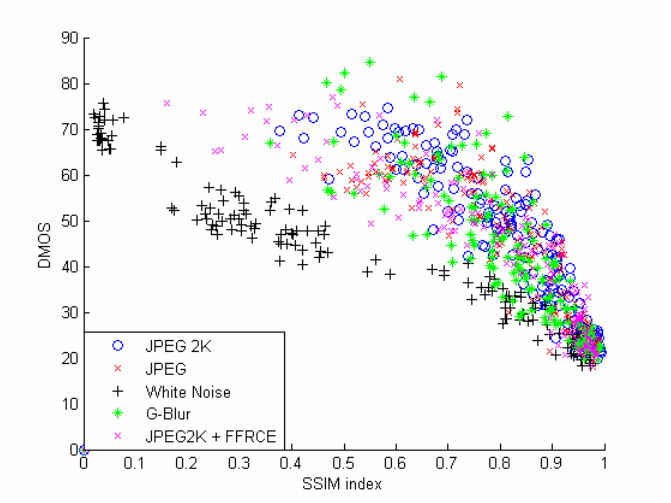
In contrast to SSIM, the LTS measure not only provides a better clustering of subjective data within each distortion category, but it also provides good clustering of all the data across different distortion categories. This is why the RMSE performs so well in the General distortion category. The scatter plot in Figure 5.5 (c) is derived by the following parameter models.

$a_T = 0,\ \gamma = 1000,\ \alpha = 1.9,\ \beta = 2.2,\ c = 20$

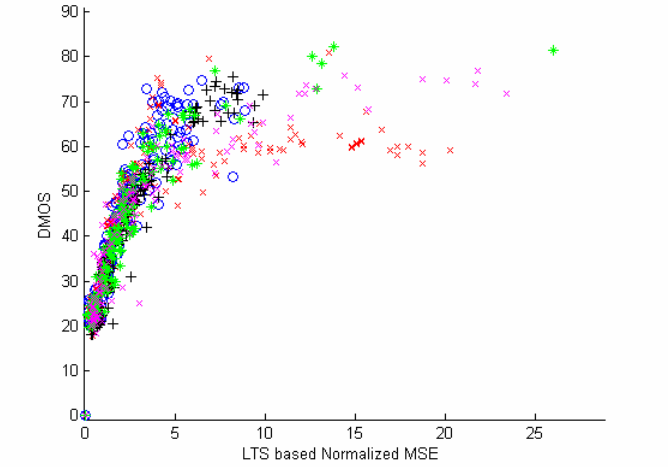
Note that these are the same parameters for reporting the performance of the proposed metric for the General distortion case. Remember that we only optimized the mapping function based on three parameters of γ, α and β, for all the data in the database (i.e., for all types of distortion).



(a)



(b)



(c)

Figure 5.5 Scatter plots of DMOS vs. different quality metrics for different distortion types. (a) MSE. (b) SSIM. (c) Proposed distortion metric. Note: The legend for all figures is given inside image (b).

5.4 Further Considerations for Distortion Metric based on the LTS

5.4.1 Investigation on more Complex LTS to Normalization Factor Mapping Functions

In (5.6) a generic mapping function was proposed that relates the LTS measure to the MSE normalization factor, based on the shape of the JND as a function of the LTS measure which we empirically found in 5.1.5 (in Figure 5.3). This generalization on the

texture masking properties of the HVS from the JND regime to the supra-threshold regime legitimately raises the question on the validity of the proposed mapping function in (5.6). This question becomes more concerning when we observe that prior work based on block texture masking such as [ZhLX03] and [TOVE98] amongst others, used several classes for the purpose of MSE normalization. As explained at the end of 5.1.5 the general shape of Figure 5.3 at JND regime, resembles a binary classification scheme and consequently the proposed mapping function based on the optimal parameters depicted in Figure 5.4, suggests a simple normalization of the MSE, based on a constant which can be determined according to a binary classification based on the LTS measure.

To make sure that the simple mapping function in (5.9) is suitable to define the normalization factor based on the LTS measure, a more complex mapping function was considered which provided the possibility of creating more than two distinct levels. The examined function is given in (5.10) which is a summation of two simple "S" shape functions as in (5.9). Note that in (5.10) we have twice the number of parameters to optimize compared to (5.9). If a multi-category classification (normalization) function is to provide any advantage, then we expect learning of the model parameters in (5.10), using the subjective test result for the LIVE database, will provide an overall function with three such flat levels.

$$\xi(k)=\left(1+\gamma_1\cdot\left(1-\exp\left(-\left[\frac{|\psi(k)|}{\alpha_1}\right]^{\beta_1}\right)\right)+\gamma_2\cdot\left(1-\exp\left(-\left[\frac{|\psi(k)|}{\alpha_2}\right]^{\beta_2}\right)\right)\right) \tag{5.10}$$

The two important parameters in the conducted experiments were α_1 and α_2 which indicate where the transition from one level to the next level takes place. The optimization for different distortion classes in the LIVE database revealed that depending

on the distortion class and the initial parameter set, one of the two outcomes can be expected. In one case both α_1 and α_2 converge to the same value which was found under simple mapping function in (5.9). In this case the overall shape of the function in (5.10) is the same as function (5.9), however it is more convenient to work with (5.9) due to smaller number of loose parameters. In the other case it was observed that α_1 and α_2 converge to distinct values, however the optimization process moves one of them to a value very high (e.g. 50~100) where the maximum of the LTS measure ($\psi(k)$) can not reach (the typical value is well below 7). Given this result we conclude that the function in (5.9) provides a reasonable mapping from the LTS measure to the normalization factor ξ. We conjecture that a continuous transition from high sensitivity to low sensitivity is a better choice compared to defining three regions of high, low and medium sensitivity. The justification is that if this conjecture were not right, the optimization process should have resulted in distinct α_1 and α_2 in the range of valid LTS values while the β_1 and β_2 parameters should converge to a large number (fast transition).

5.4.2 Computational Complexity Considerations

The proposed metric has very little complexity compared to the block transformed methods and pre-filtering methods for linear or non-linear feature extraction methods [PMAC99]. The complexity of the proposed metric is also less than those of other methods which require some pre or post filtering, such as contrast sensitivity filtering in [WaAh05] and apparatus filtering used in SSIM. To better enumerate the computational advantage of the LTS method let's assume that the image size is M by N pixels, and for the proposed LTS method each basic block has the size m by n pixels and the perceptual

support area covers p by q basic blocks. The computational operations have a complexity of $O(M \cdot N)$ to calculate σ_p and a complexity of $O\left(\frac{p \cdot q}{m \cdot n} \cdot M \cdot N\right)$ to calculate μ_{txt} and σ_{txt} for all basic blocks based on (5.1). Moreover, to calculate ψ for all basic blocks in the image we require $O\left(\frac{M \cdot N}{m \cdot n}\right)$ operations based on (5.2). The complexity of mapping from ψ to ξ (e.g. through look up tables) is also $O\left(\frac{M \cdot N}{m \cdot n}\right)$. Note that in typical viewing conditions where $p \cdot q < m \cdot n$ (in fact as the angular resolution increases the m and n values get larger while the p and q remain constant) the complexity order of the proposed approach is $O(M \cdot N)$.

The complexity for a block transform method with the same image sizes and non-overlapping transform blocks of size $n \times n$, for fast transforms algorithms such as FFT is $O(M \cdot N \cdot \log_2(n))$ for performing the transformation. Also if the transform coefficient size remains the same as the pixel size (which is typically the case) and the number of interacting coefficients on each coefficient is r then the number of operations for normalization is $O(M \cdot N \cdot r)$, ignoring the error pooling, the overall complexity of such a block transform method is $O(M \cdot N \cdot \max(\log_2(n), r))$ (note that as opposed to the proposed method the complexity increases as the angular resolution increases as result of bigger n even if r remains constant). Finally the complexity of perceptual distortion methods which require 2-D filtering is $O(M \cdot N \cdot m.n)$, if the filter has a 2D span of m by n. Again note that as the filter dimension gets larger due to higher angular resolution the complexity would increase.

In essence the most time consuming part of the proposed metric is finding the variance of pixel values for each basic block, which has a complexity comparable to finding the MSE for residual error.

5.5 Deficiencies and Possible Improvements

The distortion metric proposed in this chapter has two deficiencies which can be easily fixed as explained below.

1. Average intensity shift: If the average intensity in basic blocks of the reference image is shifted by a constant value to form the test image, the weighted MSE predicts a perceptual distortion which is proportional to the amount of average intensity change. However it is a known fact that this treatment of the distortion does not represent the HVS, as the visual system is not as sensitive to a shift in image luminance. A low complexity approach to solve this problem is to weight the "variance" of the error in each basic block as opposed to the weighting the average squared error (or MSE).
2. Influence of the Point Spread Function (PSF): The visual apparatus in humans shows a PSF impulse response. This fact causes the contrasts at very high angular resolution to blend (the same principle is used in display devices for half-toning or to represent different colors with only three color components). This behavior is not considered in the distortion model we introduced so far. For example, assume a case where the image of interest is a grating pattern with a constant spatial pitch in terms of pixels.

As the distance between the eye and the display increases, the proposed model increases the size of the basic block, however the texture activity (as the variance of the pixels within the basic block) will not change. Consequently the proposed texture masking factor will not change as the angular resolution changes. In this case the proposed model fails to consider the fact that at some distance the grating pattern turns into a smooth surface. To fix this problem one can pre-filter the reference and the test image prior to the calculation of the LTS and the MSE, using an appropriate PSF (low-pass filter) based on the angular resolution. This smoothing is the function of the aperture module in many psychophysical studies such as [WaAh05].

One solution that can remedy both of the problems mentioned above is to pre-filter both the reference and the test image, using a band-pass filter. This possibility was studied for a Laplacian of Gaussian (LoG) filter with parameter σ, which controls the peak gain frequency of the band-pass filter.

$$LoG(x,y)=\frac{1}{\pi\cdot\sigma^4}\left(\frac{x^2+y^2}{2\sigma^2}-1\right)\exp\left[-\left(\frac{x^2+y^2}{2\sigma^2}\right)\right] \qquad (5.11)$$

In the simulations, the reference and the test images in the LIVE database were pre-filtered prior to calculating the LTS and the MSE. Using the same methodology, it was attempted to optimize the combination of the proposed model parameters (γ, α, β, c) and σ. It was observed that adding the parameter σ did not significantly improve the results reported in Table 5.2. This observation was attributed to the data which was used for the optimization purposes. In fact a closer look at the type of distortions used for

subjective tests in release 2.0 of the LIVE database indicates that all of the distortion types add errors with zero mean. Therefore there is no need for a filter to correct the first effect expressed above. Moreover, the need for modeling the PSF property of the HVS in the conducted experiments is limited as all the subjective tests were conducted for the same optimal angular resolution. Future work on a more robust quality metric need to use a more comprehensive database with subjective test results for images observed at different angular resolutions (different pixel pitch or different distances) and distortion types which offer non-zero mean errors in arbitrary-size image blocks. Furthermore it is anticipated that replacing the LoG filter with a band pass filter which is more representative of the HVS can be beneficial. Examples of these filters have been studied as contrast sensitivity filters, in [WaAh05]. We believe research on addition of an optimal band-pass filter to the proposed distortion model is complementary to what is proposed in this chapter.

This chapter, in part has been submitted for the following publication:

- Minoo, K.; Nguyen T., “A Perceptual Image Quality Metric Based on Local Texture Spread,” Image Processing, IEEE Transactions on, submitted for publication, 2008.

Chapter 6
PPIQ: A Probabilistic Perceptual Image Quality Framework

In this chapter we introduce a distortion metric which is based on the probability of detecting discrepancies between two images in displaying the same visual feature at a given spatial coordinates. As an example a visual feature can exist in one picture to reflect an impression from the natural world (therefore the lack of such a feature would be considered a distortion). Alternatively a feature can exist in one picture due to a modification process (e.g. quantization), changing the natural scene's content (i.e., the detection of this feature is considered a distortion). In 6.1, we formalize a generic probabilistic metric model for measuring the distance between two corresponding regions of interest in two images in terms of the probability of finding a discrepancy between the two images for showing the same feature. This generic model can adopt an arbitrary complex feature detection model from linear + non-linear transform models for detection of spatial contrasts with different shapes, orientations and spatial frequencies [MENS06], [KaKi95], [WaAh05], [PMAC99] to more complex models such as the hierarchical temporal memory model for detection of sophisticated features such as body, face, etc. [GeHa05].

In 6.2, we discuss error pooling. Specifically we consider two types of error pooling: 1) error pooling across different features and 2) error pooling across the spatial domain. We will see that the proposed probabilistic distortion metric model and spatial error pooling, together, facilitate the inclusion of foveated distortion [KoGe96], [SaPB01]. In the proposed probabilistic model, the foveated distortion concept translates

to a spatial pooling strategy that gives higher probability to errors (detection of feature discrepancies) in the area of an image with higher possibility of becoming the center of attention.

Obviously the overall "goodness" of the proposed distortion metric, to predict the subjective test result, would depend on the underlying feature-detection model. To demonstrate the efficiency of the proposed model in practice, in 6.3, a Probabilistic Perceptual Image Quality (PPIQ) metric is proposed which is based on detection of non-directional contrasts. The proposed feature model is derived based on the properties of physiological neural structures in the early vision of the HVS and the Receptive Fields (RFs) at higher levels of the visual neural pathway. The proposed measure inherently factors in viewing conditions such as the angular resolution of the image and the well established psychophysical properties of the HVS such as contrast sensitivity and texture masking (as elaborated in Chapter 4) through the concept of receptive field. These properties include intensity masking, contrast sensitivity and texture masking.

The distortion models have been introduced in a parameterized manner so one can assess the optimal parameters for different applications. In 6.4, the subjective test results for different types of distortions will be used to optimize the model's parameters such that the proposed distortion metric best matches the results from the subjective experiments. We will discuss how the optimal parameter values relate to the type of distortion. Also it will be shown how the optimal parameters depend on the viewing conditions and how they should change if the viewing condition changes (without the need to find a different set of optimal parameters for every new viewing condition).

6.1 The Probabilistic Metric Model

Many features and properties of the HVS for perceptual evaluation of an image can be accurately explained by the properties of receptive fields. It has been shown that through the concept of receptive field one can explain how the HVS system can detect different visual features from simple intensity contrasts to more complex shapes to very complex tasks of detection of complex objects such as human faces. In this section we introduce a framework based on the theory of receptive field which enables us to define a distortion (or similarity) metric for images, relative to a reference image. First, we do a high level review of the theory of receptive field for detection of visual features and then derive a probabilistic feature oriented similarity or dissimilarity (distortion) measure between the corresponding locations of two images.

6.1.1 A Review of Receptive Field Model

In general, a receptive field is a neural connection configuration, where a number of neurons send electrical impulses to the same node (ganglion cell). The ganglion cell performs a signed-weighted summation on the input signals and produces an output signal for the next ganglion cell in a hierarchical fashion. This description of receptive field ignores possible feedback from a higher level cell to a lower level one, but for the purpose of our discussion it suffices. Figure 6.1 follows what Hubel showed in [HUBE63], where it is assumed that every visual feature in an image corresponds to a specific receptive field in the visual pathway. Typically the receptive fields closer to the sensory cells are typically simpler (retinal and Lateral Geniculate Nucleus (LGN)

receptive fields) performing simple contrast detection, and higher level receptive fields use these simple contrast basis functions to respond to more complex visual stimuli.

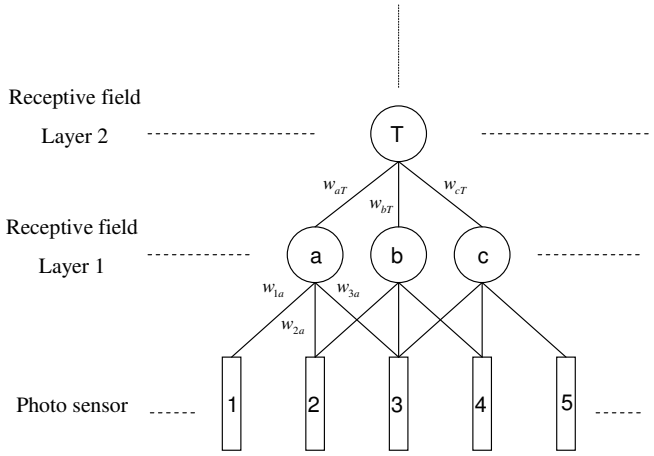


Figure 6.1 A hierarchical receptive field structure for feature detection.

For example a retinal receptive field at the fovea is usually connected only to a few cone sensors. The location map of these cones on the fovea forms a round or elongated circle which can be modeled by a region of in-phase gains (all positive or all negative), in the center, surrounded by another region of out-phase gains (opposing sign, compared to the central region). Figure 6.2 shows that at the resting condition, when there is constant light over the whole RF, the ganglion cell fires pulses at the resting-rate. When the light shone on the center increases, the ganglion cells' pulse rate changes depending on the type of RF. The "on-center" RF type fires electric pulses much faster compared to the resting-rate only when the center area is excited. The "off-center" RFs fire slower pulses compared to the resting-rate when the center is lit. When the light shone on the surrounding area the situation would be the exact opposite, i.e., slower

pulses for on-center ganglion cells and faster pulses for off-center ganglion cells [HUBEOL].

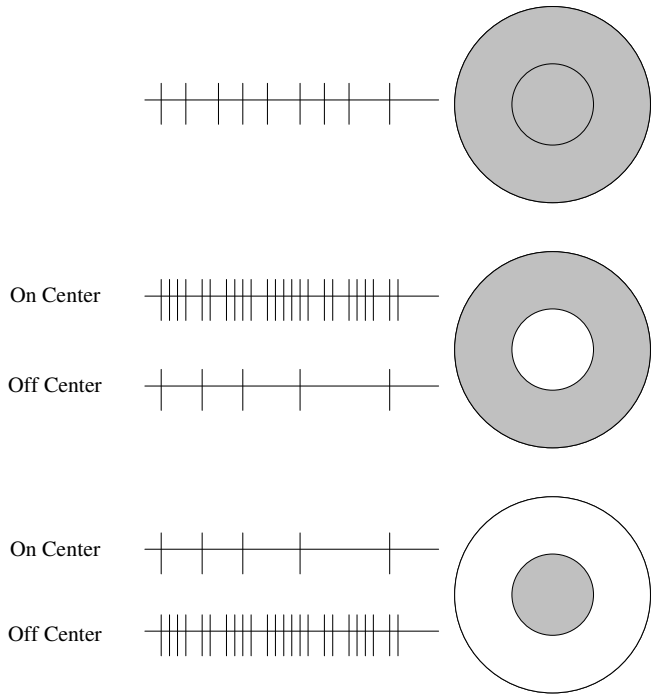


Figure 6.2 Receptive Field's impulse activity responding to a stimulus. Top is impulse rate at resting. Middle when the center is excited. Bottom is when the surround is excited.

Our reference receptive model consists of a hierarchy of nodes, e.g. a feed forward or recurring spiking neural network. Each node performs a linear operation on the input nodes (filtering) and then a non-linear operator (decision function) sets the output (pulse rate activity) of that node. Each node at different levels represents a certain visual feature. The shapes of these features in lower visual levels are usually simpler (simple contrasts) and as we get to the higher levels of the visual pathway the feature shape becomes more complex (Figure 6.1).

6.1.2 Probability of Feature Detection

Our simple receptive field based model for feature detection consists of a feature extraction transform $T_f(a)$ (not necessarily a linear transform) and a nonlinear feature detection function. $T_f(a)$ maps the intensities $i(\cdot)$ from the image to a feature domain value $r(f,a)$, which reflects the impulse rate (neural activity) at the receptive field which corresponds to a desired feature (left side of Figure 6.3). $r(f,a)$ can vary around the resting pulse rate from zero to a maximum number (look at the rightmost block in Figure 6.3).

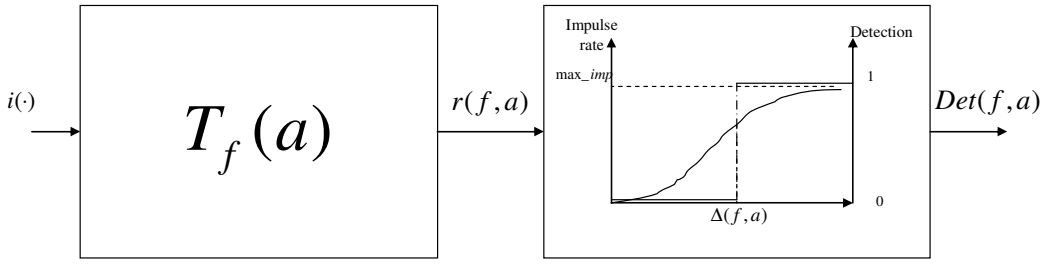


Figure 6.3 Feature detection block diagram. The left block represents the feature extraction transform and the right block represents the receptive field neural impulse activity and the corresponding feature detection decision.

In the proposed feature detection model, when a specific receptive field is assessed for the presence of a feature, a hard-limit threshold at $\Delta(f,a)$ is used to set the detection decision to either 0 (not detected) or 1 (detected). We represent the detection of the feature f, at location a in an image by $Det(f,a)$. To formalize the notation we assume that for a person with a threshold decision of $\Delta(f,a)$ the detection function can be defined as:

$$Det(f,a)=\begin{cases}1 & |r(f,a)|>\Delta(f,a)\\ 0 & |r(f,a)|\le\Delta(f,a)\end{cases} \qquad (6.1)$$

In (6.1), $r(f,a)$ is the activity of the receptive field (corresponding to the desired feature f at location a). Realizing that the value of $\Delta(f,a)$ is different from person to person, a probabilistic measure is chosen to predict the statistical nature for the outcome of many subjective tests. We assign a cumulative probability to the detection threshold value ($\Delta(f,a)$) amongst a large number of test subjects as $P_{\Delta(f,a)}(\Delta) = Prob(\Delta(f,a) \le \Delta)$. The feature detection in a probabilistic manner can be interpreted as follows:

$$Det(f,a) = \begin{cases} 1 & with\ prob = \ P_{\Delta(f,a)}(|r(f,a)|) \\ 0 & with\ prob = \ 1 - P_{\Delta(f,a)}(|r(f,a)|) \end{cases} \tag{6.2}$$

In the proposed generic probabilistic model we only assume that $P_{\Delta(f,a)}(\Delta)$ is a cumulative probability function, (non-decreasing function which goes from zero to one). In 6.3, we will consider specific functions to represent $P_{\Delta(f,a)}(\Delta)$ and discuss how the chosen function reflects known psychophysical properties of the HVS such as luminance masking and texture masking.

6.1.3 Probability of Detecting Feature Discrepancies Between Two Images

In the model, a feature detection discrepancy between two images (usually a reference image and a test image) happens when the same feature, f, at the same location a is detectable in one picture and not in the other. We assign a probability to this error-detection in terms of the percentage of people who notice a feature discrepancy between the two images. It is easy to derive this probability from (6.2) as follows:

$$\begin{aligned} P_{dis}(ref,tst,f,a) = & \max\left(P_{\Delta_{ref}(f,a)}(|r_{ref}(f,a)|), P_{\Delta_{tst}(f,a)}(|r_{tst}(f,a)|)\right) \\ & - \min\left(P_{\Delta_{ref}(f,a)}(|r_{ref}(f,a)|), P_{\Delta_{tst}(f,a)}(|r_{tst}(f,a)|)\right) \end{aligned} \tag{6.3}$$

In (6.3), $P_{dis}(ref, tst, f, a)$ is the probability of detecting a dissimilarity between the reference and the test images at location a for the feature f. $r_{ref}(f, a)$ and $r_{tst}(f, a)$ are the feature extraction transform responses in the reference and the test image, respectively. Also $P_{\Delta_{ref}(f,a)}(\cdot)$ and $P_{\Delta_{ref}(f,a)}(\cdot)$ are the corresponding cumulative probabilities for detection threshold at the two images. Note that based on the same concept we can define $P_{sim}(ref, tst, f, a)$ as the probability of finding similarity between the two images for the same feature f at the same location a.

$$P_{sim}(ref, tst, f, a) = 1 - P_{dis}(ref, tst, f, a) \tag{6.4}$$

Note that since $P_{\Delta(f,a)}(\cdot)$ is a cumulative distribution function, $P_{dis}(ref, tst, f, a)$ has all the appealing properties of being a metric to show the distance in terms of detecting features at the same location in two images :

Non-negativity:

$$P_{dis}(ref, tst, f, a) \geq 0$$

Identity of indiscernibles:

$$P_{dis}(ref, tst, f, a) = 0 \quad iff \quad P_{\Delta_{ref}(f,a)}(|r_{ref}(f, a)|) = P_{\Delta_{tst}(f,a)}(|r_{tst}(f, a)|)$$

which implies the feature is detectable with the same probability in both images.

Symmetry:

$$P_{dis}(ref, tst, f, a) = P_{dis}(tst, ref, f, a)$$

Triangle inequality:

$$P_{dis}(x, y, f, a) \leq P_{dis}(x, z, f, a) + P_{dis}(y, z, f, a).$$

The proof is straightforward considering the properties of $\min(\cdot)$ and $\max(\cdot)$ functions in (6.3).

6.2 Error Pooling

So far we have derived a basic distortion metric which is the probability of observing an error in detection of a given feature at a given location between two images by equation (6.3). However before we can use this result to assign a distortion (or similarity) metric to an image we need to address error pooling across feature space and spatial space as discussed in 4.3.2.2. The concept of feature detection error pooling has been extensively studied in linear transform spaces (e.g. FFT, DCT and Wavelet). Most of these studies deal with very specific features which are typically spatial contrasts, produced by different basis functions of a transform space. Karam in [LiKW03] and [HoKa02] proposed a distortion metric based on the linear feature gain and foveal probability summation ([RoGr81], [Wats79]) for Wavelet and DCT coefficients, respectively. A more general treatment of contrast based quality metric can be found in [PMAC99] where Pons considered the case of error pooling in a general feature transform space. Most feature (error) pooling schemes such as [LiKW03], [HoKa02], [PMAC99], [WaSo97], use a norm-p to assign an overall "size" value to a vector of features across feature space and foveal-spatial space ([PMAC99] uses a weighted norm-β). The fundamental principle for this pooling scheme comes from the concept of probability summation in foveal feature detection which reasonably states that the probability of detecting a feature (error) is equal to detecting at least one feature (error) [RoGr81]. Although this concept matches well the results of psychophysical studies

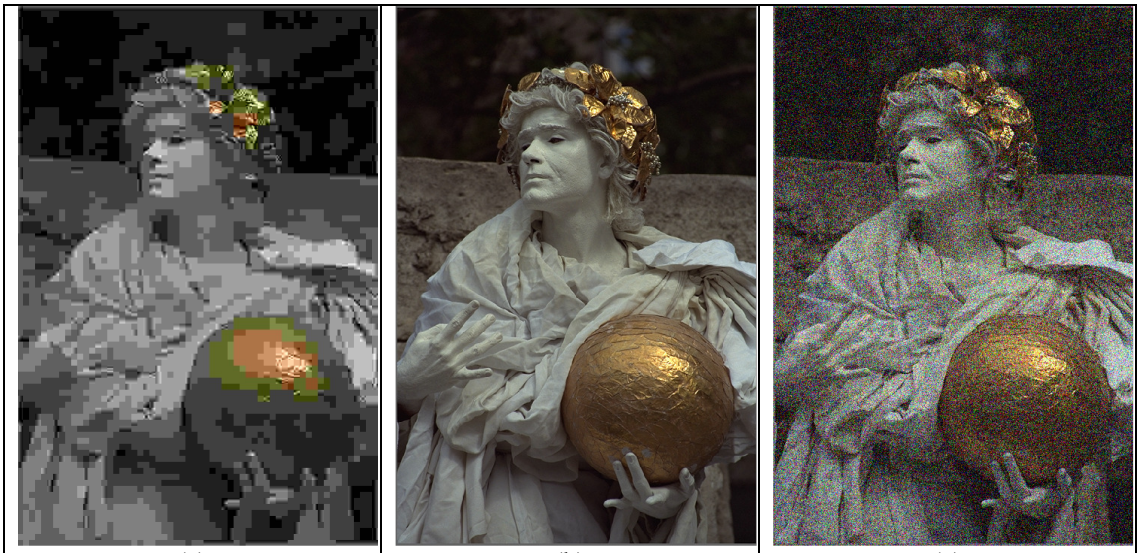
which are concerned with the concept of "Just Noticeable (feature) Detection", they fall short of representing the image quality at supra-threshold cases [RaHe01], where the feature or distortion is beyond the "just-visibility" point and one has to predict how people rate the quality of such an image with many visible features (errors). In the rest of this section the proposed strategy for pooling across feature and spatial spaces is presented.

6.2.1 Pooling Across Feature Space

We argue that a generic pooling across all features is not representative of how people evaluate image quality. In fact the subjective criteria for judging image quality depends on the application in which the image will be used (e.g. medical diagnosis quality vs. print quality), people's expectations (e.g. based on their expertise). This suggests that a good error pooling strategy should depend on the application. We further justify this argument in the following two aspects:

1. Overall perception of distortion: When human subjects are asked to describe the distortion in an image, they would describe the image quality in terms of appearance of certain classes of distortions (features). This is why people complain about blocking (edge) artifact, ringing artifact, salt & pepper noise, among other attribution, when they are asked to describe the quality of a distorted image. Therefore, a natural way to evaluate an image is by quantifying the image distortion based on the severity of individual feature-specific distortions.

2. Mapping from feature detection to quality distortion: In order to measure the quality of an image, the detection of a distorted feature needs to be mapped to a distortion quality (or quantity). As mentioned in the beginning of this section, this task is not straightforward. In fact this is the similar argument that Wang [WBSS04] put forth to reason why PSNR is not descriptive of the distortion and offered Structural SIMilarities (SSIM) index to predict the subjective quality of an image. However, as it can be seen in Figure 6.4, the two images from the LIVE database [SWCBOL], with different distortion type or error feature (added white noise vs. DCT based artifacts) have almost the same DMOS, but significantly different SSIM or PSNR metrics (in fact the white noise distortion has a better subjective quality).



(a) (b) (c)

Figure 6.4 Images from LIVE database [SWCB02]. (a) JPEG compressed image218.bmp with PSNR =31dB, SSIM =0.62 and DMOS=60.4. (b) the original image statue.bmp. (c) white noise distorted image91.bmp with PSNR =28.2dB and SSIM =0.22 and DMOS=50.2.

The above justifications are encouraging to evaluate an image based only on one feature and assign the quality metric for observable, feature-specific, discrepancies between each image and the corresponding reference image. Based on this argument the proposed generic distortion measure consists of a vector of distortion metrics which shows the probability of detecting discrepancies for each feature in a set of features which are relevant to a specific application.

Here it should be also acknowledged that if a single quantity should be reported as the "global image quality metric", for an application, there should be a combining (pooling) and comparison scheme, tailored for the specific application which is going to use this "global image quality metric" (e.g., based on the type of distortion, etc.). In fact in 6.3 we discuss a global distortion metric based on different feature detection parameters and regression function parameters based on different type of distortions to predict DMOS which can be considered a global (subjective) metric for image quality assessment.

6.2.2 Spatial Error Pooling

In general there are two approaches to spatial pooling in conventional distortion models. The first approach as in [LiKW03] and [HoKa02], uses a probability summation concept (norm-p Minkowski distance) in foveal regions around each pixel and then either uses the maximum of all these foveal error values or uses the average of the error values to define the distortion metric for the image. The second approach such as the one used in [WBSS04] uses the mean of feature (similarity or distortion) values across the whole image. In the previous discussion (at the beginning of this section) we argued that the

foveal error pooling methods are a reasonable choice when we are concerned with the "just-visible" error (feature). Intuitively, in supra-threshold when we have many visible error features in the foveal region, a better choice to grade the image quality would be to find the average of severity of (probability of detecting) the error features.

The verbal justification of the proposed spatial pooling strategy can be formalized by defining a quality metric that depends on the average response time to find the first distortion. A shorter average response time indicates a more distorted image. Note that so far what we found was a probability assigned to detection of the feature of interest at a given location in the image. We now consider the act of eye fixation on a given spot in the image. This act can also be described as a random event with a given probability. In fact many studies have been conducted to find the probability that a given area of the image becomes the center of attention [KoGe96], [SaPB01]. Of course this requires knowledge about the image content and several factors which can draw viewers' attention. Examples of these factors include proximity to the center of the image, detection of skin color, recognition of a human face or existence of motion in a given region of an image amongst other things. With less knowledge about these elements, the distribution becomes flatter (i.e. all locations are equally likely). In general we assign a probability to location a to become the center of attention as $P_a(a)$. Note that the act of finding the first discrepancy between the two images, involves a sequence one random fixations, where each fixation on location a in one image has the probability of $P_a(a)$. In here we assume that the second fixation on the same spot in the other image happens with probability 1 (as the test subjects tries to find the same feature in the same location). As a result we can assume that the probability of finding a feature discrepancy at location a in

one fixation has a probability of: $P_{dis}(ref, tst, f, a) \cdot P_a(a)$. The probability of finding a distortion in one fixation $P_{dis\_f}(ref, tst, f)$ becomes:

$$P_{dis\_f}(ref, tst, f) = \sum_{a \in image} P_{dis}(ref, tst, f, a) \cdot P_a(a) \tag{6.5}$$

Now if we assume that people make several fixations, each taking a time t_f (in seconds), then we can represent the distortion by the expectation of the time it takes to find the first distortion. In order to perform this task, note that the probability of observing the first distortion after n fixations is:

$$P(t = n \cdot t_f) = (1 - P_{dis\_f}(ref, tst, f))^{n-1} \cdot P_{dis\_f}(ref, tst, f) \tag{6.6}$$

Note that in (6.6) it is assumed that each location can be revisited with the same probability in each fixation. The expected time to find the first distortion then becomes:

$$t_{f\_ed} = \left[\sum_{1}^{\infty} n \cdot t_f P(t = n \cdot t_f) \right] = \frac{t_f}{P_{dis\_f}(ref, tst, f)} \tag{6.7}$$

Assuming that fixation time is constant, the distortion for the whole image with respect to feature f ($D(ref, tst, f)$) which is inversely proportional to the time that it takes to find the first error can be represented by:

$$D(ref, tst, f) = P_{dis\_f}(ref, tst, f) \tag{6.8}$$

Note that the distortion which is equal to $P_{ed\_f}(ref, tst, f)$ in (6.5) is a weighted mean of the proposed error detection metric for individual pixels locations in (6.3). As such it affirms the verbal argument above that in supra-threshold scenarios the expected error detection probability is a good choice for spatial error pooling (instead of finding the maximum probability of error or finding the probability of at least one error in the foveal area).

6.3 A Generic PPIQ Metric

The proposed model so far does not assume any particular realization of the feature detection probability function $P_{\Delta(f,a)}(\Delta)$, nor does it assume any feature extraction transform $T_f(a)$ (receptive field shape). In the first part of this section we consider a probability model that captures the properties of receptive fields in the Lateral Geniculate Nucleus (LGN). In the second part, the proposed general receptive model is specialized to define a distortion (or a similarity) metric for a generic distortion (feature). The generic feature is an omni-directional contrast feature. The result of the first two parts provides us with a parameterized distortion metric for the suggested detection probability function and the generic feature we consider in 6.3.2.

6.3.1 Detection Probability Function Model

So far we have only assumed that $P_{\Delta(f,a)}(\Delta)$ has a general S shape (as a cumulative density function) which resembles a Sigmoid or Logistic function (Figure 6.3). A good choice of feature detection probability function should include parameters which represent important features from the psychophysical and physiological point of view. To that end, we note that one of the important attributes (parameters) of the $P_{\Delta(f,a)}(\Delta)$ (from a statistical point of view) would be the normal-threshold (Δ_α) where $P_{\Delta(f,a)}(\Delta_\alpha) = \alpha/100$. This parameter represents the value of receptive field activity (at feature detection node) where $\alpha\%$ of subjects from the whole subject population would detect the feature. Another important attribute (parameter) of $P_{\Delta(f,a)}(\Delta)$ is how fast this

function saturates around the Δ_α (this parameter is also referred to as the slope of the $P_{\Delta(f,a)}(\Delta)$). A conventional practice is to express $P_{\Delta(f,a)}(\Delta) = P_{\Delta(f,a)}(|r(f,a)|)$ as follows:

$$P_{\Delta(f,a)}(|r(f,a)|) = 1 - \exp\left(-\left[\frac{|r(f,a)|}{\Delta_{norm}(f,a)} \right]^{\beta(f,a)} \right) \qquad (6.9)$$

Note that $\Delta_{norm}(f,a)$ is the receptive field activity ($|r(f,a)|$) at which 63% of people can detect the existence of feature f at location a. $\beta(f,a)$ decides the slope factor, i.e. how fast or slowly the probability of feature detection saturates. In general both $\Delta_{norm}(f,a)$ and $\beta(f,a)$ are functions of the desired feature f and the image content, especially around the location a (as the notation of these values implies). In practice to be able to parameterize the detection probability function, $P_{\Delta(f,a)}(\Delta)$, we choose $\beta(f,a)$ to be a constant for each feature.

To choose $\Delta_{norm}(f,a)$, we note that the detection of a feature in the presence of other features (those created by distortion or the ones belong to the original image), in the same spatial vicinity is highly dependent on the strength and type of the distorted feature, itself, and the strength and type of other existing features. While some features in an image boost the perception of each other (excitatory effect), most features tend to attenuate the presence of each other (inhibitory effect, similar to texture masking) [Ring04]. The divisive gain control strategy has been employed in many studies [WaSo97], [PMAC99], [MENS06] to factor in this excitatory and inhibitory aspect of the HVS in detection of features. The divisive gain control in equation (6.10) performs a non-linear operation on the activity of each feature (neural cell) to include the excitatory and

inhibitory interactions from other features (cells) in the neighborhood of the feature (cell) of interest.

$$r_{gc}(f,a)=\frac{|r(f,a)|^{p}}{b(f,a)+|H_{f,a}(r(.,.))|^{q}} \tag{6.10}$$

In (6.10), $r_{gc}(f,a)$ is the gain controlled response of the receptive field and $H_{f,a}(.)$ is a function (not necessarily linear) to capture interaction of all features (neural activities) in the spatial vicinity of point a on the receptive field that extracts feature f. Also $b(f,a)$ indicates the linear gain before the inhibitory effects become noticeable. Based on (6.10), the decision on (probability of) detection of a feature can be derived by:

$$P_{\Delta(f,a)}(r(f,a))=1-\exp\left(-\left[\frac{|r(f,a)|^{p}}{b(f,a)+|H_{f,a}(r(.,.))|^{q}}\right]^{\beta}\right) \tag{6.11}$$

Equation (6.11) suggests that $\Delta_{norm}(f,a)=\sqrt[p]{b(f,a)+|H_{f,a}(r(.,.))|^{q}}$. This formulation of $\Delta_{norm}(f,a)$ allows us to easily capture the masking effects of the image content in detection of any feature, namely the texture masking [ZhLX03] and the Local Texture Spread, discussed in Chapter 5 and the luminance masking [Stev57], [LaRP97], [FPSG96], [Ward94] effects. Keeping $H_{f,a}(.)$ in (6.11) provides the proposed model with more flexibility, however it makes optimization of the distortion model parameters (as explained in 6.3 and 6.4) more difficult. Consequently, the function $H_{f,a}(.)$ is dropped from (6.11) to reach (6.12).

$$P_{\Delta(f,a)}(r(f,a))=1-\exp\left(-\left[\frac{|r(f,a)|^{p}}{b(f,a)}\right]^{\beta}\right) \tag{6.12}$$

With this simplified distortion model with only linear gain control, two concerning issues for image quality evaluation should be addressed. First we have to discuss how the proposed model in (6.12) captures the texture masking effect (surround inhibition) and secondly how the intensity masking effect (Weber law) should be included in the model.

6.3.1.1 Texture masking

A partial inclusion of texture masking on the detection of a feature discrepancy between two images is to some extent given in the proposed model. Observe that the proposed distortion metric for an individual feature at a location in (6.3) subtracts the probabilities of feature detection in each of the original and test frames. On the other hand if the original image results in a feature response $r_{ref}(f,a)$ and the test image results in a feature response $r_{tst}(f,a)=r_{ref}(f,a)+\delta_r(f,a)$, because of the shape of (6.12) which saturates around $b(f,a)$, the probability of observing a discrepancy not only depends on the increased activity feature response $\delta_r(f,a)$, but also depends on the $r_{ref}(f,a)$. For example when $r_{ref}(f,a)$ is well above the $b(f,a)$ (i.e. visible feature with probability 1) then $\delta_r(f,a)$ (the effect of error on feature extraction transform response) should be in the order $b(f,a)-r_{ref}(f,a)$ so that it can bring the probability of detection below one. Note that this is similar to the error weighting scheme proposed in [Wats93] which takes into account the value of a DCT coefficient in order to weight the error for that coefficient. Of course the above argument only justifies the effect of a feature at a given point on the error detection for the same feature at the same location. However as we see in 6.4 (when we learn the model's parameters from a set of images with

corresponding subjective test results), the optimal parameters suggest a larger spatial support size for the feature extraction transform in the proposed model when compared to the expected receptive field size in the fovea. The reason for this larger feature transform (receptive field) support size is to allow the surrounding textures saturate the receptive field response, which as explained in previous paragraph translates to a masking effect on detection of the feature of interest (also look at [Ring04]).

6.3.1.2 Luminance masking

Following the logic in [Wats93], the dependency on the average intensity in the proposed model is formulated as a power function with parameter η so the effect of display gamma factor can be conveniently included in the calculations.

$$b(f,a) = b(f,I_o)\cdot\left(\frac{I(a)}{I_o}\right)^{\alpha} \tag{6.13}$$

In (6.13), $b(f,I_o)$ is the linear gain for feature f at a known average intensity I_o and $I(a)$ is the average intensity at location a.

Replacing (6.13) in (6.12) and replacing $b(f,I_o)/I_o^{\alpha}$ with $b(f)$ results in the final detection probability function.

$$P_{\Delta(f,a)}(r(f,a)) = 1-\exp\left(-\left[\frac{|r(f,a)|}{b(f)I^{\alpha}(a)}\right]^{\beta}\right) \tag{6.14}$$

6.3.2 Omni-Directional Contrast Feature

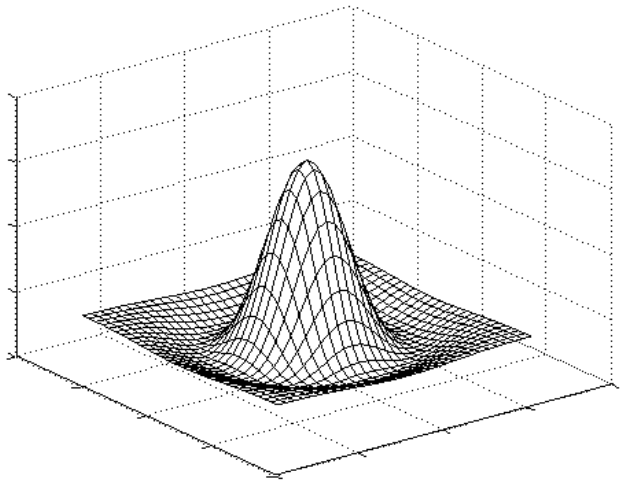
As noted in 6.1, the receptive fields in the higher hierarchical levels (which respond to more complex features) get their inputs from the lower level receptive fields. The retinal receptive fields lie at the bottom of this hierarchy, right after the sensory cells

(first layer in Figure 6.1). These receptive fields are not only responsible for detection of simple contrast features, but they also provide the input for detection of more complex features in the higher layers in the visual pathway. Next, a parameterized receptive field model is introduced. This model can be utilized in the proposed distortion metric.

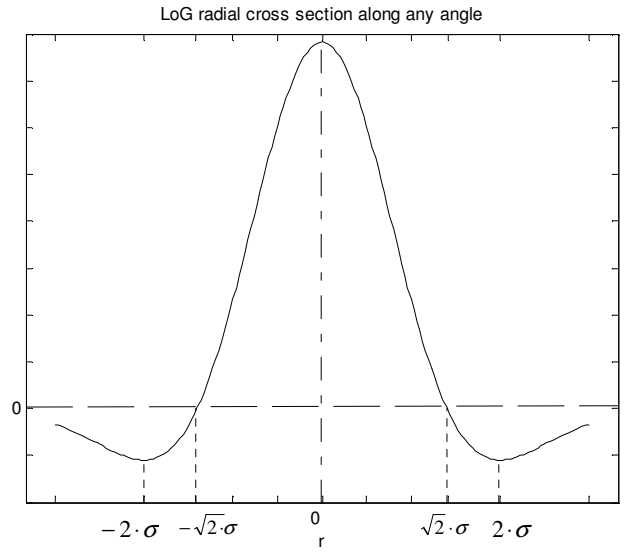
The proposed model here, offers the simplest type of retinal receptive field, which is an omnidirectional contrast detector. This type of receptive field can be observed more frequently in the foveal receptive fields which are responsible for highest acuity vision (Photopic vision). In the proposed model a symmetric Laplacian of Gaussian (LoG) filter is used to represent the feature extraction transform part of the receptive field. The parameterized model uses equation (6.15) to generate the 2-D FIR filter for feature extraction.

$$T_{oc}(x,y)=\frac{1}{\pi\cdot\sigma^4}\left(\frac{x^2+y^2}{2\sigma^2}-1\right)\exp\left[-\left(\frac{x^2+y^2}{2\sigma^2}\right)\right] \tag{6.15}$$

The only parameter associated with this filter is σ which defines the band-pass characteristics of the receptive filter in the spatial frequency domain. Figure 6.5 shows a realization of such a filter for $\sigma=2$. It should be noted that σ also defines the spatial span of the receptive field on the display, for example in terms of number of pixels. Since this number should reflect the actual size of the corresponding receptive fields on the retina, one can use the value of optimal σ for a given viewing distance and pixel pitch and calculate the optimal σ for a different viewing distance or pixel pitch. The adaptability to viewing conditions is one of the advantages of the proposed distortion metric which is not provided directly by other distortion metrics such as SSIM or PSNR.



(a)



(b)

Figure 6.5 (a) Impulse response of a Laplacian of Gaussian filter (inversed). (b) a cross section of LoG impulse response along any angle which crosses the origin. The filter response resembles the shape of the simple receptive field right after retina.

Next, we need to optimize the parameters for the proposed generic distortion metric which is proposed for an omni-directional contrast feature. We will see how the experimental results affirm findings of psychophysical experiments. However, before proceeding, it should be emphasized that the proposed generic distortion model

framework can accommodate more distortion metrics by choosing different feature extraction transforms and cumulative probability functions. For example one can define a distortion metric by employing an edge extraction (detection) filter [MaHi80], [Cann86], [Lind98] or a feature extraction filter by using 7/9 wavelet basis for detection of ringing artifacts in JPEG2000 compressed images [WYSV97]. Note that Gabor filters can also be used for feature extraction purposes in the proposed model (they have been a valid filter option for contrast sensitivity studies [WaAh05]).

6.4 Empirical Study of Generic PPIQ Metric

The proposed generic metric model has been parameterized to obtain the optimal metric model based on the subjective test result. To that end the subjective test results from the LIVE database version 2.0 [SWCBOL] were used to optimize the proposed distortion metric models. This database includes subjective test results for five categories of distorted images along with the test and the reference images. The distortion categories include JPEG compression distortion, JPEG 2000 compression distortion, AWGN distortion, GBL distortion and finally JPEG 2000 image distortion due to FFRCE model. The details of the subjective test experiments can be found in [ShBo06]. In the followings the results are presented and the significance of these results is discussed. For parameter optimization, a third order polynomial regression method is used to find the RMSE performance of the model for each set of parameters [Roha00].

6.4.1 Optimal Model Parameters

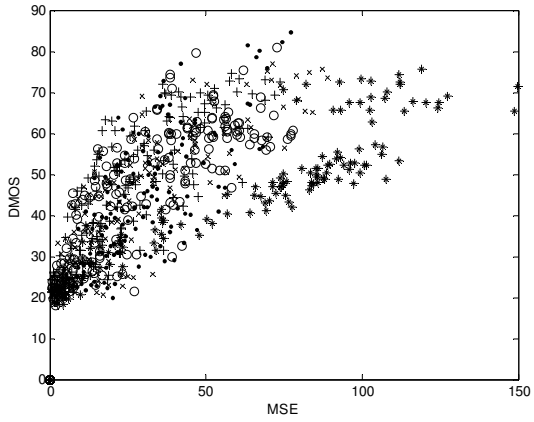
The proposed distortion model has four parameters. One of the parameters defines the feature extraction filter (σ) while the other three define the detection probability

function (b, β and α) as explained in the previous section. First a set of PPIQ objective distortion metric parameters was found, which minimizes the RMSE for DMOS estimation for all the images in the database (the values of these parameters are given in the last row of Table 6.2). Using the same parameters for the PPIQ and the default parameters for the SSIM metrics we compare the performance of three quality metrics, the PPIQ and the SSIM and the MSE in dB ($10*\log_{10}(MSE)$). The results for this comparison are shown in Table 6.1 where the numbers in the first five rows are the RMSE for the optimal regresser to fit the value of each metric to the DMOS observation for each distortion category. Note that although both PPIQ and SSIM metrics use the same set of model parameters for all different distortion categories, the regresser for each distortion category is different and it is optimized to map the quality metric value to the DMOS, only for the distortion of interest. For this reason the sixth row was added where the overall performance of each metric is evaluated for the best regresser which fits the metric value for all distortion categories.

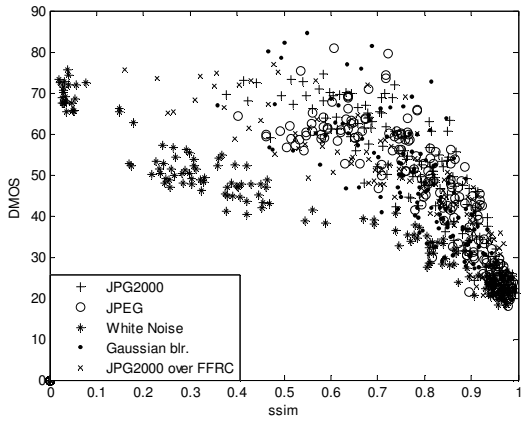
Table 6.1 RMSE for DMOS Regression based on different objective metrics

Distortion Category	# OF IMAGES	PPIQ	SSIM	10*log(MSE)
JPEG 2000	227	4.76	5.79	8.10
JPEG	233	5.94	6.14	8.59
White Noise	174	4.59	3.02	5.87
Gaussian Blur	174	4.30	7.96	10.40
JPEG 2000 with FFRCE	174	4.90	5.63	8.31
All categories	982	6.55	7.97	9.79

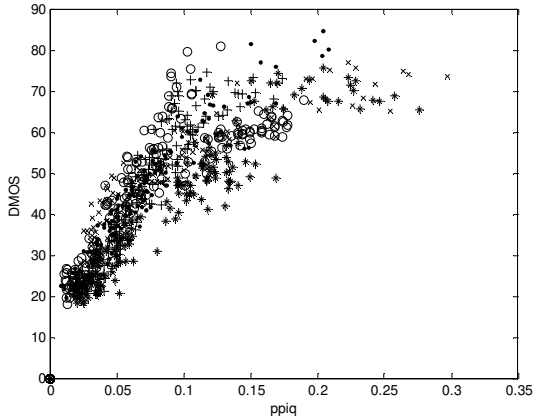
Figure 6.6 depicts DMOS vs. different metrics for different distortion types. Visual inspection of Figure 6.6 suggests that MSE and SSIM to a large extent and PPIQ to a much lesser extent require different regressers for the white noise distortion (note the distinct clustering of the white noise from the rest of distortions in PPIQ and MSE cases). This observation undermines the effort to have a global metric which can be applied to any image regardless of its distortion type. In fact the distorted images in Figure 6.4 exemplify why different distortion type can not be compared only based on the value of the quality metric (e.g., SSIM or PSNR in that example). Note that this behavior has also been observed with other quality metrics such as the Sarnoff quality metric, reported in [ShBo06].



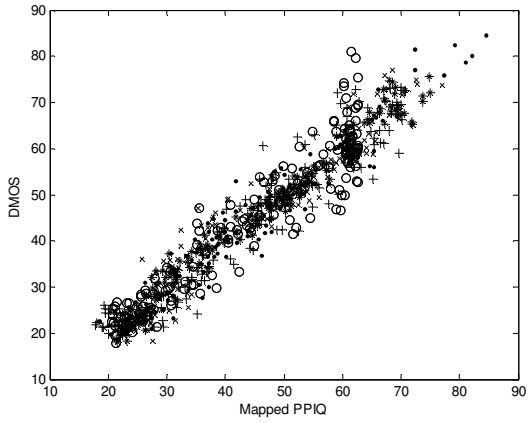
(a)



(b)



(c)



(d)

Figure 6.6 DMOS vs. different quality metrics for different distortion types. (a) MSE in dB. (b) SSIM. (c) PPIQ. (d) PPIQ mapped to DMOS using optimal parameters for individual distortion types. Note that the legends for distortion type are given in picture (b).

To solve this problem and to use the distortion metric for comparing the quality of different images, either one should only use the metric within a family of relevant distortions or use different regressers based on the type of distortion to map the quality metric to a subjective value such as DMOS. However if we use different regressers for each distortion category, we may as well use different metric model parameters for different distortion categories. Table 6.2 shows these optimal distortion parameters and their corresponding RMSE for PPIQ metric. As can be seen, the optimal parameters are different for each type of distortion which signifies the fact that the HVS would use different basic contrast features (the proposed metric feature here) to evaluate the image quality. This affirms the discussion in 6.2.1 on the subject of feature pooling. Figure 6.6 (d), depicts the DMOS vs. a predicted DMOS (non-linear regression from PPIQ to DMOS) using a different set of receptive field parameters (as in Table 6.2) and non-linear regressers for each class of distortion. In the following we observe that an advantage of the PPIQ quality metric model is that one can directly relate each optimal parameter to the specific distortion class properties.

Table 6.2 Optimal Distortion Metric parameters for different distortion classes

Distortion Category	b	σ	β	RMSE
JPEG 2000	8.6	3.2	0.6	4.58
JPEG	2.4	4.6	0.8	5.13
White Noise	3.0	2.8	0.8	2.50
Gaussian Blur	1.7	2.1	0.4	3.78
JPEG 2000 with FFRCE	1.8	1.3	0.5	4.36
All categories	11.00	1.66	0.4	6.55

6.4.2 Discussion on Optimal Model Parameters

One possible explanation for the observation of optimal metric model parameters considers the fact that a given set of parameters is best suited for representation of a specific feature which is caused by the specific distortion type. For example the optimal value of σ for the JPEG category is expected to create a LoG filter which is most sensitive for detection of intensity changes for an 8x8 block. According to Figure 6.5 (b) the expected value of σ would be given by the diagonal size of the block ($\sqrt{2}\times 8$) divided by $2\times\sqrt{2}$, which results in $\sigma = 4$. Note that this is almost the same optimal value in Table 6.2 which was found by minimizing the RMSE for data regression using the subjective test results. With some comments on the optimal parameters of the proposed metric model, we conclude this section.

First of all, the optimal exponent parameter of the probability function (β in Table 6.1) is different from what one would find in other literature which uses the probability summation for error pooling. Here the optimal value is around 0.6 while in other literature a value of 4.0 [WaSo97] or 2.2 [WaAh05] has been suggested. The reason for this difference is that we do not follow the same error pooling policy which has been only proved to be optimal for sub-threshold feature detection applications.

Secondly the optimization results show that the proposed metric is independent of the average intensity (an area of two times the receptive field feature size was used to calculate the average intensity). This seems to be a conflicting result from luminance masking which is a certain fact from the psychophysical studies. Here we conclude that

the combined effects of ambient light and the supra-threshold distortion regimes were causing the operation intensity to be same for all images, regardless of the locality of each point.

The final note is on the size of the receptive field which is a function of σ. The receptive size (which is roughly $6\times\sigma$) is around 15 pixels for detection of White noise distortion. Given the normal viewing distance and the pixel pitch we arrive at a rough estimation of the average receptive field size of 0.5^o (angular degree). This value seems to be higher than what has been reported in physiological studies. However the larger receptive field can be justified in two respects. First, as explained in 6.3.1, it allows for inclusion of texture masking in the proposed model [Fiel87]. Secondly by noting that white noise is visible through receptive fields of much smaller size, the larger optimal receptive field in Table 6.2 indicates that the perception of distortion would be related to features (structures) of larger sizes (about 0.5^o in terms of angular size). In other words, one can speculate that receptive fields of smaller size at the fovea join together to form a larger receptive field at higher levels of the visual pathway that influences the perception of distortion by HVS.

6.5 PPIQ Beyond Full-Frame Image Quality Metric

The PPIQ provides a unifying model that relates the findings of psychophysical experiments with the techniques used for image quality assessment at supra threshold. The distortion metric based on this model has proved to perform very well under different distortion conditions. The main aspects of the PPIQ metric model can be summarized as follows:

1. Perceptual representation: The experimental results show that the proposed quality metric is a relatively good representative of the HVS impression of the image quality (compared to the alternative perceptual metrics).
2. Low computational complexity: Compared to the alternatives such as SSIM, the proposed generic metric requires the same, if not less, computational processing power. Note that it only takes two 2-D convolutions and two point exponential functions on each image to calculate the metric value.
3. Perceptually meaningful and ease of adaptation to different viewing conditions: The model parameters relate to the physiological or statistical properties of the HVS. This not only facilitates the adaptation of the model's parameters when the viewing condition changes, but also provides an insight on how the HVS system might operate, at least for evaluation of image quality.
4. Flexibility of the metric model: The flexibility of the proposed model allows the inclusion of more sophisticated distortion models for specific applications such as medical diagnosis or machine vision applications. In these applications the distortion metric would be the probability of misclassifying a feature (e.g. an object).

6.5.1 PPIQ and blind Image Quality assessment

PPIQ possesses unique properties to support blind image quality assessment. One research direction, in the field of image processing, would include a treatment of probabilistic distortion metric for blind distortion metric. This treatment of image quality

assessment provides a metric which uses the same measure to assess image quality in both cases where there is a full-reference image and when there is no-reference image (blind). Note that the conventional full-reference quality metrics use a distance metric, while conventional blind quality metrics gauge the deviation of the test image in statistics of a given feature, compared to that of uncorrupted natural images for example in [GaCr07], [XinL02], [ShBC03], [ShBo02] and [PeMa05].

The advantageous characteristic of the PPIQ metric for quality assessment of a test image, with or without the full knowledge about the reference frame, roots in the fact that the PPIQ metric offers a probabilistic notion of distortion. Note that in the absence of a reference image the distortion depends on the natural statistics of image features (e.g., edges, etc.) and the statistical properties of the source of distortion (e.g. white-noise, quantization noise, etc.). These statistical priors make it possible to assign a probability to the existence of a feature in the absent, original image. It needs to be emphasized that the knowledge about these priors can be inferred from the statistics of natural images and also the properties of distortion as explained in [MNIC08] where a blind metric is defined for blocking artifact in DCT coded images.

This chapter, in part, has been submitted for the following publication:

- Minoo, K.; Sagheb, S.; Nguyen T., “PPIQ: A Probabilistic Perceptual Image Quality Metric”, Selected Topics in Signal Processing, IEEE Journal of, Visual Media Quality Assessment April 2009. Submitted for publication, 2008.

Chapter 7
Conclusion and Future Research Directions

7.1 Summary of Research

In this work two concepts of rate and distortion were revisited within the context of image processing and compression. The following provides a brief summary of what is presented in this dissertation.

7.1.1 On Entropy Rate

By taking advantage of a parametric model for representing the probability distribution of visual data, the problem of rate estimation was posed as a parameter estimation problem. It was shown that in the context of rate estimation, the MLPE can be used for robust rate estimation which works for all data rate conditions (from very low to very high data rates).

It was shown that modeling the data distribution as Laplacian results in a very simple, closed form expression for rate estimation of the non-zero symbols after uniform quantization (with possible extended dead zone).

To overcome the issues related to the non-stationary nature of image data and also lack of insufficient samples for parameter estimation, different approaches for estimation of the probabilistic model's parameters were proposed and examined. Experimental results show the robustness of the propose rate estimation for different image contents at different data sizes (as small as 4x4 pixels).

7.1.2 On Distortion

In this work, two approaches to image quality assessment were presented. First by introducing the concept of a local-texture-spread measure, a framework for measuring the amount of perceptual distortion was developed. Subjective tests proved that the LTS measure is capable of predicting the perceptual distortion not only at just-noticeable-distortion, but also at supra-threshold distortion conditions. The benefit of a perceptual distortion metric, based on the LTS measure, is in the simplicity of utilizing this distortion metric within the context of rate-distortion optimal image compression.

Next PPIQ was introduced which is a probabilistic perceptual framework for modeling the perception of distortion in an image according to the human visual system. It was shown that this framework not only unifies many of the legacy distortion methods, but also provides means of comparing those methods, and even suggests simple steps to enhance them. To experimentally show the power of PPIQ, a generic distortion metric based on a simple contrast sensitivity function was proposed within the PPIQ framework. It was shown that this simple model outperforms the contending distortion metrics.

We discussed how the probabilistic nature of PPIQ framework can be used to introduce perceptual distortion metrics for “full-reference” and “blind” image quality applications.

7.2 Future Research Direction

Conducting research for completion of this dissertation in the past year has left me with a list of issues I would have liked to address and include in the present writing on the following subjects.

7.2.1 Entropy Rate Estimation

It is interesting to perform a comparative study between different probability distribution models, such as Gaussian, Mixture of Gaussian, Cauchy, etc. in the context of MLPE. Note that this study is by nature different from those which try to predict the closest form of distribution to an observed sample set from natural images. In this suggested context, it is more important to find a distribution which results in a smaller estimated entropy rate, when the MLPE technique, as explained in this dissertation, is used on a large database of different image contents.

7.2.2 Full Reference Distortion Metrics

The simple contrast feature, used for the generic distortion metric which was introduced in Chapter 6, can certainly be improved with one of the more complex contrast sensitivity filters [WaAh05]. Note should be taken that the higher number of parameters in those complex CSFs require a larger number of training set, beyond the sample provided by the second version of LIVE image quality database [SWCBOL].

Beyond the contrast sensitivity models which effectively represent the low-level perceptions at retinal and LGN levels, a future research may include features which are more representative of higher perceptions at the cortex level. It is desirable to use a feature model that captures the hierarchical and temporal notion of perception in the cortex [GeHa05].

7.2.3 Blind Distortion Metrics

As demonstrated in [MNIC08], PPIQ provides a framework suitable for blind assessment of image quality for particular compression artifacts introduced by a specific

compression scheme. Future research into generic blind distortions, which does not depend on the information about the statistical properties of the distortion, is an interesting area of research.

Certainly a key element for realization of a generic blind quality metric is to find a feature or set of features under which the natural images in that feature space have consistent and distinguishing statistics.

7.2.4 Distortion Metrics Suitable for R-D optimization

Image compression is one of the main applications which benefits from a perceptual distortion metric. In this class of applications it is important to be able to find a closed form that relates the rate (usually an entropy rate) to the distortion (perceptual distortion). The probabilistic nature of the distortion metrics, defined within the PPIQ framework, is facilitating to relate the rate and distortion through an intermediate variable such as quantization step size.

It would be interesting to find a closed form that relates the quantization step size for a specific compression scheme, to the expected probability of distortion detection. The entropy rate can also be related to the quantization step size, using the MLPE rate estimation method, discussed in Chapter 3.

7.2.5 Distortion Metrics for Video

Defining a perceptual distortion metric that reflects the temporal aspects of visual perception can be done instantly within the PPIQ framework. The spatiotemporal characteristics of receptive fields can be exploited to define a set of spatiotemporal features, based on which the probability of finding discrepancies can be calculated. An

interesting research subject in this context would be concerned with finding suitable features that represent the spatial and temporal characteristics of the HVS.

Bibliography

[AhPe92] Ahumada, A. J., Jr and Peterson, H. A., ”Luminance-model-based DCT quantization for color image compression," Human Vision, Visual Processing, and Digital Display III, 1992.

[Bjon01] Bjontegaard, G., “Calculation of average PSNR differences between RD-curves,” Tech. Report. ITU-T SG16 Doc. VCEG-M33, March 2001.

[Cann86] Canny, J., “A computational approach to edge detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 8, no. 6, pp. 679-698, November 1986.

[CLCR93] Chun, K.W.; Lim, K.W.; Cho H.D.; Ra, J.B., "An adaptive perceptual quantization algorithm for video coding,"IEEE Transactions on Consumer Electronics, vol. 39, pp. 555-558, 1993.

[CoTh91] Cover, T. M.; Thomas, J.A., “Elements of Information Theory,” Wiley-Interscience, August 1991.

[DeHe03] Deever, A.T.; Hemami, S.S., "Efficient sign coding and estimation of zero-quantized coefficients in embedded wavelet image codecs," Image Processing, IEEE Transactions on , vol.12, no.4, pp. 420-430, April 2003

[EGCD94] Eude, T.; Grisel, R.; Cherifi, H.; Debrie, R., "On the distribution of the DCT coefficients," Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on , vol.v, no., pp.V/365-V/368 vol.5, 19-22 Apr 1994.

[FaMo84] Farvardin, N., Modestino, J.; "Optimum quantizer performance for a class of non-Gaussian memoryless sources," *Information Theory, IEEE Transactions on* , vol.30, no.3, pp. 485-497, May 1984.

[Fiel87] Field, DJ., “Relations between the statistics of natural images and the response properties of cortical cells,” Journal of Optical Society of America. A 4(12): pp. 2379–94, 1987.

[FPSG96] Ferwerda, J. A.; Pattaniak, S. N.; Shirley, P.; Greenberg, D. P.; “A Model of Visual Adaptation for Realistic Image Synthesis”, Computer Graphics, pp. 249-258, 1996.

[GaCr07] Gabarda, S.; Cristóbal, G., “Blind image quality assessment through anisotropy,” *Journal of the Optical Society of America. A, Optics, image science, and vision,* 24(12), 2007.

[GeGr91] Gersho, A., Gray, R. M.; "Quantization and entropy coding", in “*Vector Quantization and Signal Compression”,* Kluwer Academic Publishers, Norwell, MA: 1991, pp. 295-302.

[GeHa05] George, D.; Hawkins, J., "A hierarchical Bayesian model of invariant pattern recognition in the visual cortex," *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on* , vol.3, no., pp. 1812-1817 vol. 3, 31 July-4 Aug. 2005

[Giro93] Girod, B., “What's wrong with mean-squared error?,” In: A.B. Watson, Editor, Digital Images and Human Vision, MIT Press (1993).

[Gray90] Gray R.M.; “Source Coding Theory,” Kluwer Academic Press, Boton, 1990.

[H_264_] Wiegand, T.; Sullivan, G.J.; Bjontegaard, G.; Luthra, A., "Overview of the H.264/AVC video coding standard," *Circuits and Systems for Video Technology, IEEE Transactions on* , vol.13, no.7, pp.560-576, July 2003.

[HeKM01] He, Z.; Kim, Y. K.; Mitra, S.K., "Low-delay rate control for DCT video coding via ρ-domain source modeling," Circuits and Systems for Video Technology, IEEE Transactions on , vol.11, no.8, pp.928-940, Aug 2001.

[HeMi02] He, Z.; Mitra, S.K., "A unified rate-distortion analysis framework for transform coding: a summary," *Circuits and Systems Magazine, IEEE* , vol.2, no.3, pp. 46-49, Third Quarter 2002.

[HMit02] He, Z.; Mitra, S.K., "A linear source model and a unified rate control algorithm for DCT video coding," *Circuits and Systems for Video Technology, IEEE Transactions on* , vol.12, no.11, pp. 970-982, Nov 2002.

[HoKa02] Hontsch, I.; Karam, L. J., "Adaptive image coding with perceptual distortion control," Image Processing, IEEE Transactions on, vol. 11, pp. 213-222, 2002.

[HUBE63] HUBEL, DH., "The visual cortex of the brain," *Scientific American*, 209, pp. 54-62, 1963.

[HUBEOL] Hubel, D.H., "Eye, Brain, and Vision", [On line book]. Available: http://hubel.med.harvard.edu/index.html

[JM_REF] Reference Model JM13.0, JVT. Available: http://iphome.hhi.de/suehring/tml/

[KaAM05] Kamaci, N.; Altunbasak, Y.; Mersereau, R.M., "Frame bit allocation for the H.264/AVC video coder via Cauchy-density-based rate and distortion models," Circuits and Systems for Video Technology, IEEE Transactions on , vol.15, no.8, pp. 994-1006, Aug. 2005.

[KaKi95] Karunasekera, S.A.; Kingsbury, N.G., "A distortion measure for blocking artifacts in images based on human visual sensitivity," Image Processing, IEEE Transactions on , vol.4, no.6, pp.713-724, Jun 1995.

[Kell77] KELLY, DH. (1977). Visual Contrast Sensitivity. Journal of modern optics, 24(2), 107-129.

[KiKA05] Kim, H., Kamaci, N., Altunbasak, Y.; "Low-complexity rate-distortion optimal macroblock mode selection and motion estimation for MPEG-like video coders," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 15, pp. 823-834, 2005.

[KoGe96] Kortum, P. T.; Geisler, W. S., "Implementation of a foveated image-coding system for bandwidth reduction of video images," SPIE Proceedings: Human Vision and Electronic Imaging, vol. 2657, pp. 350-360, 1996.

[LaGo00] Lam, E.Y.; Goodman, J.W., "A mathematical analysis of the DCT coefficient distributions for images," Image Processing, IEEE Transactions on , vol.9, no.10, pp.1661-1666, Oct 2000.

[LaRP97] Larson, G. W.; Rushmeier, H.; Piatko, C., "A Visibility Matching Tone Reproduction Operator for High Dynamic Range Scenes," Technical Report LBNL-39882, Lawrence Berkeley National Laboratory, March 1997.

[LiKW03] Liu, Z.; Karam, L.J.; Watson, A.B., "JPEG2000 encoding with perceptual distortion control," *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on* , vol.1, no., pp. I-637-40 vol.1, 14-17 Sept. 2003.

[Lind98] Lindeberg, T., "Edge detection and ridge detection with automatic scale selection," International Journal of Computer Vision, 30, 2, pp 117--154, 1998.

[Lubi97] Lubin, J., "A human vision system model for objective picture quality measurements," Broadcasting Convention, 1997. International , vol., no., pp.498-503, 12-16 Sep 1997

[MaHi80] Marr, D.; Hildreth, E., "Theory of Edge Detection," Proceedings of the Royal Society of London, vol. 207, pp. 187, 1980.

[MaSW03] Marpe, D., Schwarz H., Wiegand, T.; "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard", Circuits and Systems for Video Technology, IEEE Transactions on, vol. 13, pp. 620-636, 2003.

[MENS06] Malo, J.; Epifanio, I.; Navarro, R.; Simoncelli, E.P., "Nonlinear image representation for efficient perceptual coding," Image Processing, IEEE Transactions on , vol.15, no.1, pp. 68-80, Jan. 2006.

[MiNg05] Minoo, K.; Nguyen, T.Q., "Perceptual video coding with H.264," Asilomar Conference in 2005, pp. 741-745.

[Mitr71] Mitra, S. K., "On the probability distribution of the sum of uniformly distributed random variables," SIAM J. Appl. Math., 20, (2), 195—198, 1971.

[MNIC08] Minoo, K.; Nguyen T., "A percetpbual metric for blind measurement of blocking artifacts with applications in transform-block-based image and video coding," *Image Processing, 2008 IEEE International Conference on.* Accepted for publication. Available: http://videoprocessing.ucsd.edu/personal/kminoo/jstsp_08/icip_08.pdf

[Mull93] Muller, F., "Distribution shape of two-dimensional DCT coefficients of natural images," Electronics Letters , vol.29, no.22, pp.1935-1936, 28 Oct. 1993.

[PeMa05] Perra, C.; Massidda, F., "Image blockiness evaluation based on Sobel operator," Image Processing, 2005. ICIP 2005. IEEE International Conference on, vol. 1, pp. I-389-92, 2005.

[PMAC99] Pons, A. M.; Malo, J.; Artigas, J. M.; Capilla, P., "Image quality metric based on multidimensional contrast perception models," Displays Vol 20, Issue 2, , pp. 93-110, 25 August 1999.

[RaHe01] Ramos, M. G.; Hemami,S S., "Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis." Journal of the Optical Society of America. A, Optics, Image Science, and Vision, vol. 18, pp. 2385, 2001.

[ReGi83] Reininger, R.; Gibson, J., "Distributions of the Two-Dimensional DCT Coefficients for Images," Communications, IEEE Transactions on [legacy, pre - 1988] , vol.31, no.6, pp. 835-839, Jun 1983.

[Ring04] Ringach, D. L.; "Mapping receptive fields in primary visual cortex," J Physiol (Lond), vol. 558, no. 3, pp. 717-728, August 2004.

[Robs66] Robson, J.G.: Spatial and temporal contrast-sensitivity functions of the visual system. J. Opt. Soc. Am 56, 1141–1142 (1966).

[RoGr81] Robson, J.G.; Graham, N., "Probability summation and regional variation in contrast sensitivity across the visual field," Vision Research Volume 21, Issue 3, Pages 409-418, 1981.

[Roha00] Rohaly, Ann M.; et al, "Video Quality Experts Group: current results and future directions," *Proc. SPIE* 4067, 742 , DOI:10.1117/12.386632, 2000.

[SaPB01] Sanghoon, L.; Pattichis, M.S.; Bovik, A.C., "Foveated video compression with optimal rate control" IEEE Trans. on Image Processing, July 2001 Page(s):977 - 992.

[ShBC03] Sheikh, H.R.; Bovik, A.C.; Cormack, L., "Blind quality assessment of JPEG2000 compressed images using natural scene statistics," Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on, vol. 2, pp. 1403-1407, 9-12 Nov. 2003.

[ShBo02] Shizhong. L.; Bovik, A.C., “Efficient DCT-domain blind measurement and reduction of blocking artifacts,” Circuits and Systems for Video Technology, IEEE Transactions on, vol.12, no.12, pp. 1139-1149, Dec 2002.

[ShBo06] Sheikh, H.R.; Bovik, A.C., "Image information and visual quality," Image Processing, IEEE Transactions on , vol.15, no.2, pp. 430-444, Feb. 2006.

[SmRo96] Smoot, S. R.; Rowe, L. A., “Study of DCT coefficient distributions,” in Proceedings of the SPIE Symposium on Electronic Imaging, vol 2657, San Jose, CA, January 1996.

[SSIMMC] Available at: http://www.ece.uwaterloo.ca/~z70wang/research/ssim/ssim_index.m

[Stev57] Stevens, S. S., “On the psychophysical law” Psychological Review, 64, pp. 153-181, 1957.

[SWCBOL] Sheikh, H.R.; Wang, Z.; Cormack, L.; Bovik, A.C., “LIVE Image Quality Assessment Database Release 2", Available: http://live.ece.utexas.edu/research/quality.

[TaMa02] Taubman, D., Marcellin, M.W., JPEG2000: Image Compression Fundamentals, Standards and Practice. Kluwer, Boston, 2002.

[TaPN96] Tan, S.H.; Pang, K. K.; Ngan, K. N., "Classified perceptual coding with adaptive quantization," IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, pp. 375-388, 1996.

[ToVe98] Tong, H.H.Y.; Venetsanopoulos, A.N., "A perceptual model for JPEG applications based on block classification, texture masking, and luminance masking," Image Processing, 1998. 1998 International Conference on , vol., no., pp.428-432 vol.3, 4-7 Oct 1998.

[WaAh05] Watson, A. B.; Ahumada, A. J. Jr., “A standard model for foveal detection of spatial contrast,” Journal of Vision, 5(9):6, pp.717-740, 2005.

[Wall92] Wallace, G.K., "The JPEG still picture compression standard," *Consumer Electronics, IEEE Transactions on* , vol.38, no.1, pp.xviii-xxxiv, Feb 1992.

[Ward94] Ward, G., “A Contrast-Based Scale factor for Luminance Display”, Graphics Gems IV, Ed. by P. S. Heckbert, pp. 415-421, 1994.

[WaSo97] Watson, A. B.; Solomon, J. A.; “A model of visual contrast gain control and pattern masking,” J. Opt. Soc. Amer., vol. 14, pp. 2397–2391, 1997.

[Wats79] Watson, A. B., “Probability summation over time,” Vision Research, vol. 19, pp. 515–522, 1979.

[Wats93] Watson, A.B., "DCT quantization matrices visually optimized for individual images," in Proc. SPIE Vol. 1913, p. 202-216, Human Vision, Visual Processing, and Digital Display IV, Jan P. Allebach; Bernice E. Rogowitz; Eds. 1993, pp. 202-216.

[Wats97] Watson, A.B., “Model of visual contrast gain control and pattern masking.” Journal of the Optical Society of America. B, Optical physics, 1997, Vol.14, issue 9, 2379-2391.

[WBSS04] Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P., “Image quality assessment: from error visibility to structural similarity,” Image Processing, IEEE Transactions on , vol.13, no.4, pp. 600-612, April 2004.

[WOQZ02] Wang, Y., Ostermann, J.; Zhang, Y.; “Video Processing and Communications,” Prentice Hall, New Jersey 2002

[WYSV97] Watson, A. B.; Yang, G. Y.; Solomon, J. A.; Villasenor, J., “Visibility of wavelet quantization noise,” IEEE Trans. On Image Processing, pp. 1164-1175, 1997.

[XinL02] Xin Li, “Blind image quality assessment,” *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, pp. I-449-I-452, 2002.

[ZhLX03] Zhang, X.H.; Lin, W.S.; Xue, P., "A new DCT-based just-noticeable distortion estimator for images," Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on , vol.1, no., pp. 287-291 Vol.1, 15-18 Dec. 2003.