OBJECTIVE IMAGE AND VIDEO QUALITY ASSESSMENT WITH APPLICATIONS

by

QIANG LI

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2009

For Ge Song, without whom nothing would be much worth doing.

ACKNOWLEDGEMENTS

First, I thank my supervising professor Dr. Zhou Wang. This dissertation could not have been written without Dr. Zhou Wang who constantly motivates and encourages me and gives me invaluable advice throughout my doctoral studies.

Second, I wish to thank Dr. Qilian Liang, Dr.Michael T. Manry, Dr. K. R. Rao and Dr Saibun Tjuatja for their interest in my research and for taking time to serve in my dissertation committee.

Finally, I would like to express my deep gratitude to my family. I extend my heartfelt gratitude to my wife who is always there to support me. To Mom, Dad, I am extremely grateful for their encouragement and patience. I also thank my friends, especially Philip for his generous help.

Apr,27,2009

ABSTRACT

OBJECTIVE IMAGE AND VIDEO QUALITY ASSESSMENT WITH APPLICATIONS

Qiang Li, Ph.D.

The University of Texas at Arlington, 2009

Supervising Professor: Zhou Wang

Objective image and video quality assessment(IQA/VQA) aims to automatically measure the quality degradation perceived by the human eyes. It is of fundamental importance to address a wide variety of problems in image and video processing. Based on the availability of the information about the reference image, IQA/VQA models can be classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) IQA/VQA methods. This dissertation focuses on FRIQA/VQA, RRIQA/VQA, as well as their applications in perceptual image coding and video interpolation.

First, we propose novel metrics for FRIQA/VQA based on structural similarity (SSIM) and the information theoretical weighting. The spatial information weights for image and the spatial-temporal information weights for video are computed respectively in an information theoretical framework. For FRIQA, the spatial information weight is computed as the mutual information using natural scene statistics (NSS) models. For FRVQA, we incorporate the prior and likelihood models of human visual speed perception to compute the spatial-temporal information weight as a sum of

information content and *perceptual uncertainty*. Moreover, our metrics employ the perceptual weights for multiscale SSIM based on subjective tests.

Second, we propose general-purpose RRIQA algorithms which estimate perceptual image quality degradations with partial information about the "perfect-quality" reference image. Considering the dependence in the natural images, joint statistical model is applied to RRIQA, which can handle more general distortions than marginal statistics. A novel RRIQA method is proposed using the statistics of the perceptually and statistically motivated image representation. By using a Gaussian scale mixture statistical model of image wavelet coefficients, we compute a divisive normalization transformation (DNT) for images and evaluate the quality of a distorted image by comparing a set of reduced-reference statistical features extracted from DNT-domain representations of the reference and distorted images, respectively. This leads to a generic or general-purpose RRIQA method, in which no assumption is made about the types of distortions occurring in the image being evaluated. To address the problem of RRVQA, a novel statistical prior to measure the motion regularity of the natural image sequences is adopted. We investigate the temporal variations of local phase structures in the complex wavelet transform domain. It is observed that natural image sequences exhibit strong prior of temporal motion smoothness, by which local phases of wavelet coefficients can be well predicted from their temporal neighbors. We study how such a statistical regularity is interfered with "unnatural" image distortions and demonstrate the potentials of using temporal motion smoothness measures for RRVQA.

Third, we apply our IQA/VQA methods for perceptual image coding and video interpolation. Typically, perceptual image coding algorithms impose perceptual modelling in a preprocessing stage. A perceptual normalization model is often used to transform the original image signal into a perceptually uniform space, in which all the

transform coefficients have equal perceptual importance. Standard coding schemes are then applied uniformly to all coefficients. Here we use a different approach, in which we iteratively reallocate the available bits over the image space based on a *maximum of minimal structural similarity criterion*. We demonstrate the proposed method by incorporating it with the bitplane coding scheme in the set partitioning in hierarchical trees algorithm. Finally, we propose a video frame interpolation method by using the prior knowledge about temporal motion smoothness measured in the complex wavelet domain. This allows us to avoid the time-consuming motion estimation process, and thus largely reduces the computational complexity of video interpolation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

DCT: Discrete Cosine Transform

DCQ: Dynamic Contrast-based Quantization

DMOS: Difference Mean Opinion Score

DNT: Divisive Normalization Transform

FF: Fast Fading

FRIQA/VQA: Full-Reference Image Quality Assessment/Video Quality Assessment

GGD: Generalized Gaussian Distribution

GOP: Group Of Pictures

GSM: Gaussian Scale Mixture

HVS: Human Vision System

IVC: Image Video Communication

JND: Just Noticeable Differences

JPG/JPEG: Joint Photographic Experts Group

JP2: JPEG 2000

KLD: KullbackCLeibler Divergence

LIVE: Laboratory for Image and Video Engineering

LPCC: Linear Pearson Correlation Coefficient

MAE: Mean Absolute Error

MOS: Mean Opinion Score

MSE: Mean Squared Error

MSSIM: Mean Structural SIMilarity

NRIQA/VQA: No-Reference Image Quality Assessment/Video Quality Assessment

NSS: Natural Scene Statistics

PSNR: Peak Signal-to-Noise Ratio

RMSE: Root Mean-squared Error

RRIQA/VQA: Reduced-Reference Image Quality Assessment/Video Quality Assessment

SROCC: Spearman Rank Order Correlation Coefficient

UIQI: Universal Image Quality Index

VCL: Video Communication Laboratory

VIF: Visual Information Fidelity

VQEG: Video Quality Experts Group

CHAPTER 1

INTRODUCTION

Today, digital image and video become indispensable means to represent and communicate information in daily life. But distortions and artifacts are inevitably introduced during image acquisition, compression, transmission, processing and reproduction. In order to maintain and improve the quality, it is important to design the image and video quality assessment (IQA/VQA) metrics that can automatically predict the perceptual quality degradation as the human vision system (HVS) perceives [9]. Generally speaking, IQA/VQA metrics can be applied to: 1. evaluate the performance of different image and video processing systems and algorithms. For example, the rate-distortion curve have been used for years in the video compression community; 2. optimize and design the image and video processing systems and algorithms, such as perceptual image coding methods [10] to improve the visual quality.

Based on the availability of the information from reference image, IQA/VQA methods can be classified into full-reference, reduced-reference and no-reference approaches.

1.1 Full-Reference Image and Video Quality Assessment (FRIQA/VQA)

FRIQA/VQA is to predict the quality of the distorted image or video with the full access to the reference image that is assumed to be of perfect quality. Essentially, it measures "image fidelity". Most quality assessment methods in the literature [9, 10, 11] belong to this category.

1.1.1 Why and Why Not Mean Squared Error

The mean squared error (MSE) and the related measurement, the peak signal-to-noise ratio (PSNR), are the most extensively and intensively used FRIQA/VQA metrics. They are defined as:

$$MSE(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2, \tag{1.1}$$

$$PSNR(\mathbf{x}, \mathbf{y}) = 10 \log_{10} \frac{L^2}{MSE} \tag{1.2}$$

where $\mathbf{x}$ is the distorted image and $\mathbf{y}$ is the original image; N is the number of pixels; L is the maximum possible pixel value of the image.

The advantages of MSE/PSNR include: (1) the computational complexity is low; (2) the nice mathematical convexity guarantees a closed-form optimization solution. However, they are also the most criticized metrics, because MSE/PSNR are poorly correlated with human perception especially across a broad range of distortion types [12, 9]. Fig. 1.1 clearly shows the failure of MSE to predict the different distortions in agreement with human quality judgements. In [13], the reasons why MSE/PSNR fails to be a good quality prediction metric are thoroughly discussed, one of which is that MSE/PSNR ignores the correlation between the distortions and the original image resulting in the disagreement with human judgement.

Here we study the MSE/PSNR from a different point of view by addressing the image quality assessment problem in a Bayesian framework. Bayesian approaches have achieved great success in a wide range of image and video processing problems, for example, image denoising[14], motion estimation[15] and content-based image retrieval [16].

Given the distorted image $\mathbf{x}$, reference image $\mathbf{y}$ and the prior probability distribution of the reference image as $p(\mathbf{y})$ and that $p(\mathbf{x}|\mathbf{y})$ is the likelihood, the Bayesian posterior is

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})} \tag{1.3}$$

Eq.(1.3) is used as an indication of the quality of $\mathbf{x}$, which means that given $\mathbf{x}$, the more probable $\mathbf{y}$ is, the better quality $\mathbf{x}$ has. Take natural logarithm of Eq.(1.3):

$$\ln p(\mathbf{y}|\mathbf{x}) = \ln \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})} = \ln p(\mathbf{x}|\mathbf{y}) + \ln p(\mathbf{y}) - \ln p(\mathbf{x}) \tag{1.4}$$

In order to demonstrate the relationship between $\ln p(\mathbf{y}|\mathbf{x})$ and MSE/PSNR, some assumptions are made.

First, there is no prior knowledge, or $p(\mathbf{x})$ and $p(\mathbf{y})$ are constants.

$$\ln p(\mathbf{y}|\mathbf{x}) = \ln p(\mathbf{x}|\mathbf{y}) + C \tag{1.5}$$

where $C = \ln p(\mathbf{y}) - \ln p(\mathbf{x})$

Second, the distortion between $\mathbf{x}$ and $\mathbf{y}$ is modelled as additive white noise:

$$\mathbf{x} = \mathbf{y} + \mathbf{w} \tag{1.6}$$

where $\mathbf{w}$ is Gaussian distributed and independent of $\mathbf{y}$ with probability distribution function as p_w.

Thus,

$$p(\mathbf{x}|\mathbf{y}) = p_w(\mathbf{x} - \mathbf{y}) = \frac{1}{\sqrt{(2\pi)^N |C_w|}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{y})^T C_w^{-1} (\mathbf{x}-\mathbf{y})} \tag{1.7}$$

where N is the length of $\mathbf{y}$ and C_w is the covariance matrix of $\mathbf{w}$.

Third, assume $C_w = {\sigma_w}^2 \mathbf{I}$ where σ_w is constant and $\mathbf{I}$ is the identity matrix. (C_w actually serves as a local weighting map. An intensive study of local weighting strategies is reported in [6]). Then,

Table 1.1. Performance of IQA metrics[4]: ROCC

	JP2K#1	JP2K#2	JPEG#1	JPEG#2	WN	GBlur	FF	All data
PSNR	0.9263	0.8549	0.8779	0.7708	0.9854	0.7823	0.8907	0.8755
JND	0.9646	0.9608	0.9599	0.9150	0.9487	0.9389	0.9045	0.9291
DCTune	0.8335	0.7209	0.8702	0.8200	0.9324	0.6721	0.7675	0.8032
PQS	0.9372	0.9147	0.9387	0.8987	0.9535	0.9291	0.9388	0.9304
NQM	0.9465	0.9393	0.9360	0.8988	0.9854	0.8467	0.8171	0.9049
Fuzzy S7	0.9316	0.9000	0.9077	0.8012	0.9199	0.6056	0.9074	0.8291
BSDM (S4)	0.9130	0.9378	0.9128	0.9231	0.9327	0.9600	0.9372	0.9271
SSIM(MS)	0.9645	0.9648	0.9702	0.9454	0.9805	0.9519	0.9395	0.9527
IFC	0.9386	0.9534	0.9107	0.9005	0.9625	0.9637	0.9556	0.9459
VIF	0.9721	0.9719	0.9699	0.9439	0.9828	0.9706	0.9649	0.9584

$$\begin{aligned} \ln p(\mathbf{y}|\mathbf{x}) &= \ln \frac{1}{\sqrt{(2\pi)^N {\sigma_w}^2}} - \frac{1}{2{\sigma_w}^2}(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + C \\ &= A(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + B \end{aligned} \tag{1.8}$$

where $A = -\frac{1}{2{\sigma_w}^2}$ and $B = \ln \frac{1}{\sqrt{(2\pi)^N {\sigma_w}^2}} + C$.

Eq.(1.8) reveals that $\ln p(\mathbf{y}|\mathbf{x})$ is linearly correlated with the sum of squared error, $(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})$, which means the Bayesian posterior is directly related to the MSE/PSNR. In Table 1.1 based on LIVE database [17], it is demonstrated that for the distortion as additive white noise, the PSNR correlated with the Bayesian posterior performs statistically equally well compared with the other state-of-the-art image quality assessment methods. However, if the distortion is not simply the additive white noise or is related to the original signal, the posterior would not be correlated with the PSNR which performs poorly for other distortions also shown in Table 1.1. In order to compute the correct posterior to indicate the quality, the likelihood and the prior models must be utilized.

1.1.2 IQA/VQA based on Error Visibility Models

Since the ultimate observer of the image and video is human, the most straightforward way to design IQA/VQA metrics is to simulate the human visual system (HVS). By modelling the early visual pathway of the human eye-brain system such as the retina, lateral geniculate nucleus and Area V1 of the cortex, a lot of image and video quality assessment metrics [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]are proposed; refer to [32, 33] for detailed surveys.

A general framework of IQA/VQA based on error visibility models of the HVS is drawn in Fig 1.2. Firstly, a preprocessing stage is applied to the reference and distorted images/videos including color space transformation and calibration. Secondly, the reference and the distorted images/videos are decomposed into multiple spatial frequency channels. For video, a temporal filtering process is applied to model the sustained and transient mechanisms[34] of the HVS. Spatial and temporal contrast sensitivity functions, luminance masking, contrast masking features and motion error masking are then adopted to create a threshold map of the human visual error visibility for each channel, for example, the just noticeable differences (JND) threshold map[22]. Finally, the differences between the reference and the distorted images in all the channels are normalized by these threshold maps and pooled to generate an objective quality map or a single score.

Although the IQA/VQA metrics based on the error visibility models have been used in the past few decades, there are still many unanswered questions and problematic assumptions [9, 35]. Particularly for VQA, the phase 1 test of Video Quality Experts Group (VQEG)[36] reported that most popular video quality metrics based on HVS models performs statistically equivalent to the PSNR. The fundamental drawback of the approach based on error visibility models is that the simplified computational models of the neurons of the early HVS can not fully predict the behavior

Figure 1.1. "Lena" with different kinds of distortions by courtesy of [1]: (a) original image MSE=0,MSSIM=1; (b) Impulsive Salt-Pepper Noise MSE=225,MSSIM= 0.7227; (c) Additive Gaussian Noise MSE=225,MSSIM= 0.4508; (d) Multiplicative Speckle Noise MSE=225,MSSIM= 0.5009; (e) Mean Shift MSE=225,MSSIM=0.9890; (f) Contrast Stretching MSE=225,MSSIM=0.9494; (g) Blurring MSE=225,MSSIM=0.6880; (h) JPEG Compression MSE=215,MSSIM= 0.6709.

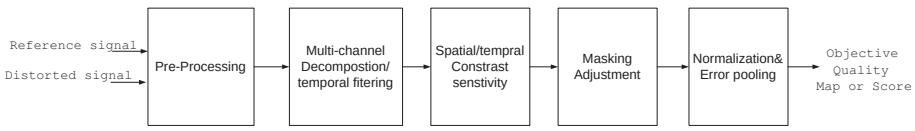

Figure 1.2. Framework of IQA/VQA based on error visibility models.

of the complicated human visual perception. For example, it is well known that the higher level of visual area MT plays an important role in motion perception. However, MT neuron models have rarely been incorporated into VQA algorithms. Our recent study of RRVQA is based on the statistical models [37] of the MT neurons for motion perception.

1.1.3 IQA/VQA based on Structural Similarity

Different from the traditional "bottom-up" methods based on human error visibility models, the Structural SIMilarity (SSIM) index is designed based on a new "top-down" philosophy. Under the assumption that the human visual system is highly adapted to extract structural information from the viewing field, the SSIM index measures the structural information as a good approximation to perceived image distortion. The SSIM index in spatial pixel domain contains three components: luminance similarity l, contrast similarity c and structure similarity s, which are defined as below:

$$\begin{aligned}
SSIM(\mathbf{x},\mathbf{y})) &= [l(\mathbf{x},\mathbf{y})]^{\alpha}\cdot[c(\mathbf{x},\mathbf{y})]^{\beta}\cdot[s(\mathbf{x},\mathbf{y})]^{\gamma} \\
l(\mathbf{x},\mathbf{y}) &= \frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1} \\
c(\mathbf{x},\mathbf{y}) &= \frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2} \\
s(\mathbf{x},\mathbf{y}) &= \frac{2\sigma_{xy}+C_3}{\sigma_x\sigma_y^2+C_3} \\
\mu_x &= \frac{1}{N}\sum_{i=1}^{N}x_i \\
\sigma_x &= [\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)^2]^{(1/2)}
\end{aligned}$$

where $\mathbf{x},\mathbf{y}$ are patches of the reference and the distorted image, α,β and γ are the weighting constants, C_1,C_2 and C_3 are the constants of small value to stabilize SSIM,

μ is the mean and σ is the standard variance of the image patch. In practical usage, with setting $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$, SSIM can be simplified as

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{1.9}$$

Fig. 1.3 demonstrates the framework of the SSIM. By implementing with a local

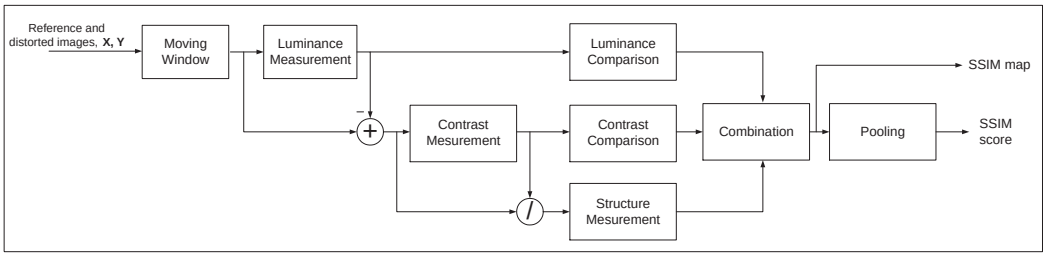

Figure 1.3. Framework of SSIM.

moving window, a spatial SSIM map between the reference and distorted images is calculated. A final score is computed by either simple averaging or more advanced pooling strategies [6]. Fig. 1.1 displays the mean SSIM scores for different distorted images, which show significant agreement with the human perception of the quality compared with the MSE. SSIM achieves substantial success compared with the MSE/PSNR in terms of the correlation with the HVS. Its low computational complexity compared with other IQA metrics makes it more appropriate for practical applications. The latest version of video codec software based on H264 standard, x.264 (www.videolan.orgdevelopersx264.html)) reports mean SSIM as a measurement of quality.

Since SSIM is efficient to compute and achieves great success, there has been a lot of research effort to extend the idea. The complex-wavelet SSIM index [38] is to tackle the problem that the spatial SSIM index is sensitive to the small translations or rotations. In order to take viewing condition into account, the weights for multi-scale

SSIM are proposed based on subjective tests [5]. In [39], a perceptually weighted complex wavelet SSIM implementation is proposed based on the HVS models, where the weights for each wavelet subband are computed according to contrast frequency sensitivity functions. SSIM has been applied to VQA by incorporating luminance masking and motion error masking[40]. A recent proposed statistical model of motion perception[37] is employed to compute information theoretic weights applied to SSIM map for VQA [41, 42]. Also, the design of SSIM motivates some latest VQA methods [43].

1.1.4 IQA/VQA based on Information Theory

Another "top-down" approach of IQA/VQA method is based on information theory and modelling the HVS as a communication channel.

Some basic knowledge of the information theory [44] is briefly reviewed. Here we use natural logarithm which can be extended to the logarithm of any base.

1. The self-information or surprisal as a measure of the information content:

$$S(x) = -\ln p(x) \tag{1.10}$$

where $p(x)$ is the probability of x.

2. The entropy, H, of a discrete random variable X is a measure of the amount of uncertainty associated with the value of X.

$$H(X) = -\sum_{x} p(x) \ln p(x) \tag{1.11}$$

3. Mutual information measures the amount of information that can be obtained about one random variable by observing another.

$$I(X;Y) = \sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \tag{1.12}$$

4. The Kullback-Leibler divergence (or information divergence, information gain, or relative entropy) is a way of comparing two distributions: a "true" probability distribution $p(X)$, and an arbitrary probability distribution $q(X)$.

$$D_{KL}(p(X)||q(X)) = \sum_x p(x) \ln \frac{p(x)}{q(x)} \tag{1.13}$$

The visual information fidelity (VIF) [7] views FR IQA as an information fidelity problem under the hypothesis that visual quality is related to the amount of information that the HVS can extract from an image. Fig.1.4 shows the framework of the VIF. The VIF [11] is also applied to FRVQA.

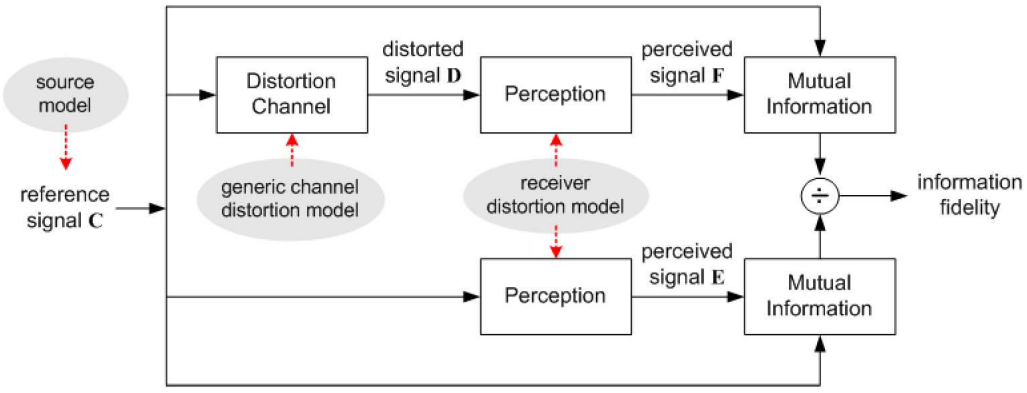

Figure 1.4. Framework of VIF.

Information theory has been applied to design pooling strategies for FRIQA/VQA. The basic assumption is that the regions with more information content are more likely attract visual attention; thus, they should be given more computational effort. Information content weights [6] are computed based on a simple Gaussian source model. Using the statistical models of human motion perception [37], an information pooling strategy considering both the information content and perceptual uncertainty are proposed for FRVQA in [41, 42].

1.2 Reduced-reference Image and Video Quality Assessment (RRIQA/VQA)

Reduced-reference (RR) IQA/VQA methods predict the quality degradation of an image with only partial information about the reference image, in the form of a set of RR features [9]. RRIQA methods provide a practically useful and convenient tool for real-time visual information communication and networking systems, where they can be used to track image quality degradations and control the streaming resources on the fly. The frame work of the RR quality assessment system is shown in Fig. 1.5.

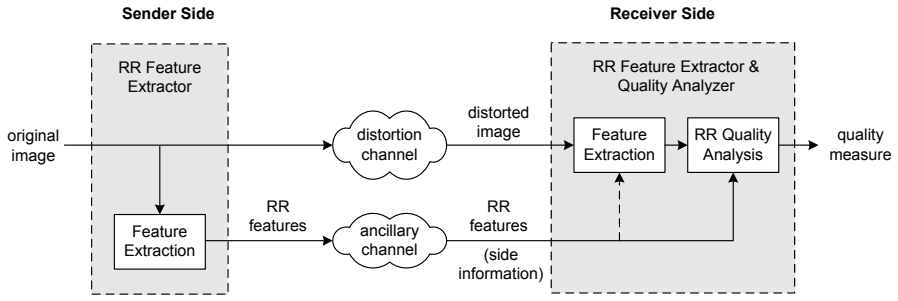

Figure 1.5. RR quality assessment system.

The major challenge in the design of RRIQA/VQA algorithms is to find appropriate RR features which are desirable to

1. provide an efficient summary of the reference image;
2. be sensitive to a variety of image distortions;
3. be relevant to the visual perception of image quality.

Another important aspect that has to be kept in mind in the selection of RR features is to maintain a good balance between the data rate of RR features and the accuracy of image quality prediction. With a high data rate, one can include a large amount of information about the reference image, leading to a more accurate estimation of image quality degradations, but it also becomes a heavy burden to transmit the RR

features to the receiver. On the other hand, a lower data rate makes it easier to transmit the RR information, but more difficult for accurate quality estimation. In practical implementation and deployment, the maximal allowed RR data rate is often given and must be observed. Overall, the merits of an RRIQA/VQA system should not be gauged only by the quality prediction accuracy, but by a tradeoff between the accuracy and the RR data rate.

Three different but related types of approaches have been employed in existing RRIQA/VQA methods [45, 46, 47, 48, 49, 50, 51]. The first type of approaches are based on *modeling image distortions* and are mostly developed for specific application environments [45, 46, 47, 48, 49]. For example, when the distortion type is known to be standard image/video compression, a set of typical distortion artifacts such as blurring, blocking and ringing may be identified, and image features may be defined that are particularly useful to quantify these artifacts [46, 47]. For another example, in [45, 48], a set of spatial and temporal features have been found to be effective in measuring the distortions occurring in standard video compression and communication environment. The second type of approaches are based on *modeling the human visual system* [50], where perceptual features motivated from computational models of low level vision were extracted to provide a reduced description of the image. One advantage of these approaches is that the perceptual features being employed are not directly related to any specific distortion system. As a result, RRIQA methods built upon them could potentially be extended for general purpose. They may also be trained on different types of distortions and produce a variety of distortion-specific RRIQA algorithms under the same general framework. However, no study has been reported so far that applies these methods to the images with generic distortions except for JPEG and JPEG2000 compression [50]. The third type of approaches are based on *modeling natural image statistics* [51]. The basic assumption behind

these approaches is that most real-world image distortions disturb image statistics and make the distorted image "unnatural". The unnaturalness measured based on models of natural image statistics can then be used to quantify image quality degradation. In [51], a generalized Gaussian density function [52] as in Eq. (1.14) is used to model the marginal statistics of the linear coefficients in wavelet subbands.

$$p_m(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x|/\alpha)^\beta} \tag{1.14}$$

where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$(for $a > 0$) is the Gamma function, and β and α are called the scale and power factors, respectively. Fig. 1.6 shows that the GGD model fits the marginal distribution of wavelet coefficients of the natural scene image very well. The parameters of the fitting model are employed as RR features. This general-purpose RRIQA approach has achieved somewhat surprising success, as it does not require any training, and has a rather low RR data rate, but still supplies reasonable performance when tested with a wide range of image distortion types [51].

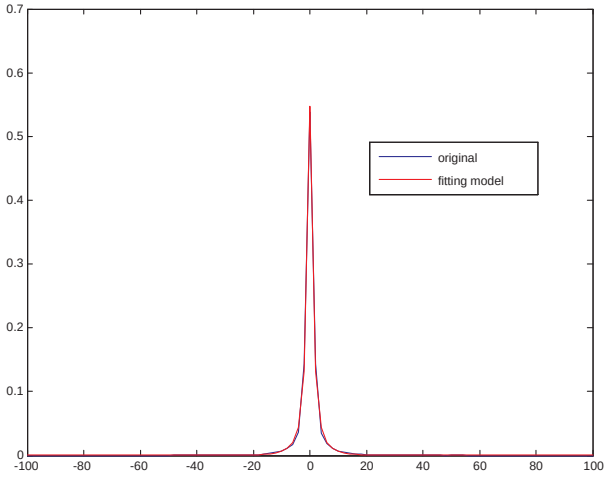

Figure 1.6. Marginal distribution of wavelet coefficients and fitting GGD model: blue line is the marginal distribution of one suband of natural image Fig.1.1 (a) in steerable wavelet domain, red line is the fitting model, x-axis is the range of wavelet coefficients.

Based on a more comprehensive and successful model of the natural scene statistics (NSS), Gaussian scale mixture model [14, 53], divisive normalization transform is proposed in order to further reduce the dependence in the wavelet domain. The statistics of the coefficients in the divisive normalized transform domain is employed for RRIQA[54]. The success of application of the NSS to RRIQA motivates new RRVQA methods based on statistics of natural video sequences. A novel prior of natural video sequences, namely *temporal motion smoothness*, is presented and modelled to capture the temporal regularity in the natural video, which is adopted for the general-purpose RRVQA [55].

1.3 No-Reference Image and Video Quality Assessment (NRIQA/VQA)

NRIQA/VQA methods evaluate the quality without the requiring knowledge about the reference image. Most NRIQA/VQA methods in the literature are dependent on the knowledge of specific distortions. For instance, the blur and blocking artifacts of block-based compression standards including JPEG and MPEG are perceptually annoying to human eyes, which inspire NRIQA/VQA methods to measure those artifacts to predict the quality degradation in [56, 57, 58, 59]. In [60], a specific RRIQA metric was proposed for JPEG2000 compressed images. More types of distortions are accounted for including white noise, impulse noise, blurness, blockness and ringing in [61]. But, there is no computational model to combine those measurements for general-purpose NR quality assessment.

It is interesting to note that the general-purpose NR quality assessment is not difficult for human eyes. Usually, one can tell not only the levels of distortions but also the types of distortions in the distorted image without the original image. One possible explanation is that the prior knowledge about the natural images are acquired and stored when the HVS adapts to the surrounding natural environment through

the evolution process. Once given unnatural or distorted images, one will perceive the disturbance based on the prior knowledge of the tiny cluster of the natural images. Therefore, the Bayesian framework introduced in Section 1.1.1 for quality assessment problem may lead to new general-purpose NRIQA methods by studying the prior and likelihood functions of the natural signals.

1.4 Validation of Perceptual Quality Assessment Methods

The performance of the objective quality assessment methods are evaluated by the following steps:

(1) carrying out subjective tests to acquire subjective scores including mean opinion score (MOS) or difference mean opinion Score ($DMOS$). The subjective testing procedures for visual quality assessment have been suggested in [62] including double stimulus continuous quality scale method, double stimulus impairment scale method and single stimulus continuous quality evaluation. The MOS is the arithmetic mean value of all the subjective scores. The DMOS is the quality score difference between the reference and distorted image, which means that the quality decreases with the increasing vales of DMOS.

(2) applying regression functions to the objective scores to compute the predicted subjective score. Three widely used regression functions are listed in the following equations. Eq.(1.15) and Eq.(1.16) are proposed by the VQEG [63]. Eq.(2.7) is used in [64] to evaluate different state-of-the-art FRIQA algorithms.

$$s_p = ao^3 + bo^2 + co + d \tag{1.15}$$

$$s_p = \frac{b1}{1 + e^{-b2*(o-b3)}} \tag{1.16}$$

$$s_p = \beta_1 logistic(\beta_2, (o - \beta_3)) + \beta_4 o + \beta_5, \tag{1.17}$$

$$logistic(\tau, o) = \frac{1}{2} - \frac{1}{1 + \exp(\tau o)}$$

where o is the objective score from the quality assessment methods; s_p is the predicted subjective score; the parameters are estimated by maximizing the correlation between S_p and subjective scores such as MOS or $DMOS$.

(3) computing evaluation metrics between the predicted subjective scores and the subjective scores. Four metrics [36, 64] are generally accepted to evaluate the performance of the objective quality assessment methods:

1) Linear Pearson correlation coefficient (LPCC)

$$R = \frac{\sum_i (s_i - \hat{s}) * (s_{pi} - \hat{s_p})}{\sqrt{\sum_i (s_i - \hat{s})^2 * \sum_i (s_{pi} - \hat{s_p})^2}} \tag{1.18}$$

where s_i denotes the subjective scores, MOS or $DMOS$ and s_{pi} denotes the predicted subjective score based on the above regression functions.

2) Spearman rank order correlation coefficient (SROCC)

$$r = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{1.19}$$

where d_i is the difference between the i-th image/video's ranks in subjective and objective evaluations and N denotes the total number of the image/video.

3) Mean absolute error (MAE)

$$MAE = \frac{\sum |s_i - s_{pi}|}{N} \tag{1.20}$$

4) Root mean-squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{N-d}\sum(s_i - s_{pi})} \tag{1.21}$$

where d is the degrees of freedom of regression function.

Among the above 4 metrics, LPCC is to evaluate the *prediction accuracy*; SROCC is an evaluation of *prediction monotonicity*. In particular, SROCC is the only one robust to the non-linear regression function, i.e. adjusting the parameters of the function will not affect the SROCC value. There is more detail of the metrics in [36]. The better objective quality assessment metric should have higher SROCC and LPCC while lower MAE and RMSE.

We have attempted to collect all the publicly-accessible image and video databases with subjective opinion scores and obtained 4 IQA databases and 1 VQA databases.

Among the 4 IQA databases, the UT-Austin LIVE database[17] is the most comprehensive one that includes 779 distorted images generated from 29 natural images. The distortions are JPEG2000, JPEG compression, white noise, Gaussian blur and Fast Fading Rayleigh Channel. The IVC database [65] contains 10 original images and 235 distorted images generated from 4 different processing including JPEG, JPEG2000,LAR coding and Blurring. The Toyama database [66] includes 182 JPEG and JPEG2000 compressed images. The Cornell-VCL A57 database [67] has 60 distorted images with 6 types of distortions.

The VQEG phase 1 test database [68] is the only publicly-accessible and the most extensively used database with $DMOS$ for video quality assessment. The video database includes 20 SDTV reference video sequences and 320 decoded video sequences.

1.5 Applications of Objective Quality Assessment Method

The objective quality assessment methods are desirable for many practical applications. They can be used to benchmark signal processing systems and algorithms and to monitor the quality of service (QoS) in real time. More importantly, they can be used as criterion to optimize signal processing algorithms and systems. The MSE is extensively employed as the optimization criterion because of its mathematical convenience. But the MSE is also problematic to indicate the true fidelity or quality, especially for image and video signals. So in order to maximize the perceptual quality, we should use perceptually more meaningful quality assessment metrics as the optimization criterion in the design of image and video processing algorithms and systems.

Perceptual image coding [10] is to improve the visual quality within the limitation of the bit rate. Most methods in the literature [10] use the following approach: First, some low level HVS features including contrast sensitivity and error masking are adopted in the preprocessing stage. Second, a perceptual normalization model is used to transform the original image signal into a perceptually uniform space, in which all the transform coefficients have equal perceptual importance. Finally, standard coding schemes are applied uniformly to all coefficients in order to achieve a uniform distortion map. If we model perceptual image coding as an optimization process to maximize the visual quality, the above methods based on HVS lack an explicit criterion to guide the optimization. We propose a different approach, in which we iteratively reallocate the available bits over the image space based on a *maximum of minimal structural similarity criterion*[69]. The proposed method is demonstrated by incorporating it with the bitplane coding scheme in the set partitioning in hierarchical trees algorithm.

The prior of *temporal motion smoothness* in natural image sequences has been successfully applied to RRVQA[55]. We further employed it as an optimization criterion to restore the distorted video sequence with frame dropping. Under the criterion to satisfy the regularity of *temporal motion smoothness*, the temporal interpolation is implemented in log complex wavelet domain. This allows us to avoid the time-consuming motion estimation process, and thus largely reduces the computational complexity of video interpolation.

CHAPTER 2

INFORMATION THEORETIC WEIGHTING FOR FRIQA/VQA

2.1 Pooling strategy of FRIQA/VQA

FR IQA/VQA methods generally contain two steps as shown in Fig. 2.1.

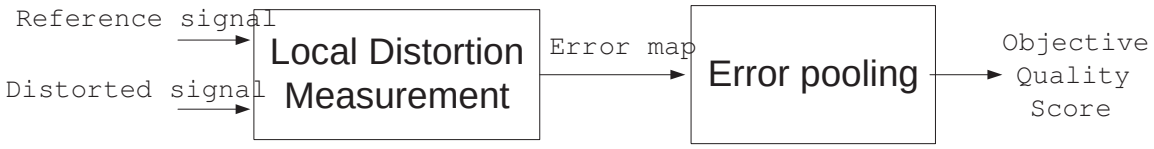

Figure 2.1. General framework of FR IQA/VQA.

First, a local distortion measurement map is computed based on the reference and distorted signals. For example, the simple spatial absolute difference map, the Sarnoff's JNDmap[22], the SSIM[70] index map and the VIF map in wavelet domain [7].

It is believed that the information represented in the primary visual cortex is integrated in the subsequent brain areas. But, unfortunately, little is know about the nature of the actual integration taking place in the brain. However, in order to simulate this process, a pooling step is applied to the local distortion measurement map by giving different weights for the regions in the map so as to compute a scalar objective quality score. Although a lot of research effort has been put into investigating the perceptual error map, much less has been done for studying the pooling strategy.

The most widely-used pooling strategy is in Minkowski form:

$$O = \frac{1}{N}(\sum_{i}^{N} {e_i}^p) \tag{2.1}$$

where $p \in [1, \infty]$ and e_i is the coefficient at location of i in the error map, N is the number of the coefficients. If $p = 2$ and the error map is the spatial absolute difference map, then O is the MSE which can be monotonically mapped to the widely used peak signal-to-noise ratio (PSNR). The Minkowski pooling method for IQA is intensively studied in [6].

An advanced pooling approach is proposed in [6] based on information theory. It is hypothesized that the human visual perception is to efficiently extract information from the natural scene. The regions with more information content are more likely to attract visual attention, thus should be given more computational effort. In [6], by modelling the image intensity as a local Gaussian source and the HVS as an additive Gaussian channel with constant noise, the information content weight is computed as

$$w(x, y) = \log_2(1 + \frac{\sigma_x}{C})(1 + \frac{\sigma_y}{C}) \tag{2.2}$$

where x, y are patches in the reference and the distorted images; C is a constant to indicate the strength of noise.

Although the pooling strategy based on information content in [6] shows great improvement, it is under a problematic assumption that the natural image is Gaussian distributed. It is well known that the statistics of natural scene images is not Gaussian [71, 52, 53], for instance the marginal statistics of natural images has a higher peak and longer tail when compared with Gaussian. We give an example in Fig. 2.2 which draws the marginal statistics in spatial domain and the Gaussian distribution with the same mean and variance. The obvious departure between these two distributions demonstrates the non-Gaussian property of the natural image.

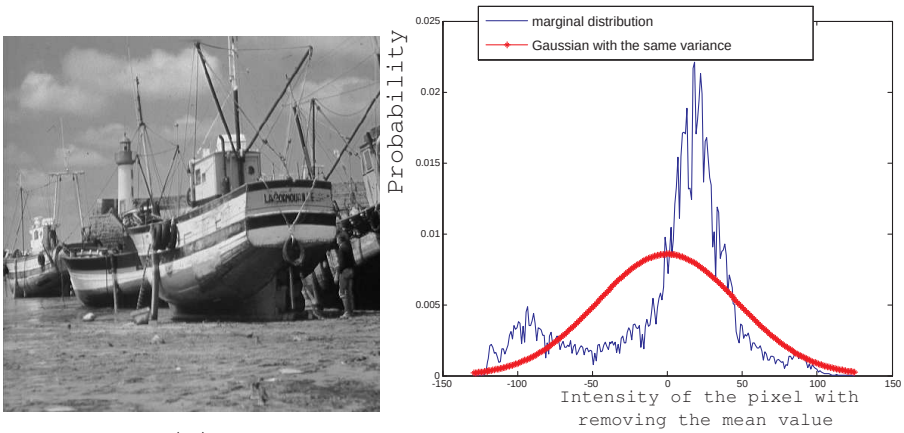

Figure 2.2. Comparison of marginal statistics from an example image with the Gaussian distribution: (a) image "Boats", (b) The blue line is the marginal distribution of intensity of the pixel in (a) with removing the mean value; The red line is the Gaussain distribution with the same mean and variance .

In the following sections, we respectively present our information theoretic weighting methods for FRIQA and FRVQA based on statistical models. For FRIQA, a more successful and comprehensive statistical model for natural image, Gaussian Scale Mixture [53] is adopted to compute the information content weights. For FRVQA, the spatial-temporal information theoretic weights are calculated as the sum of information content and perceptual uncertainty based on the prior and likelihood models of the motion perception [37].

2.2 Information Theoretic Weighting For FRIQA

2.2.1 Gaussian Scale Mixtures

A natural image is decomposed into a set of subbands of different scales and orientations, loosely speaking in wavelet domain, for example, using steerable pyramid transform [72]. X_i^j is a vector of coefficients of the coefficients in the subband at scale

i and orientation j, which can be modelled as the Gaussian scale mixtures (GSM) model[14, 53].

$$X_i^j = \sqrt{S_i^j} U \tag{2.3}$$

where U is normally distributed and S_i^j is an independent random variable.

The GSM model can successfully capture both the marginal and the joint statistical behavior of natural images in wavelet domain. The former has non-Gaussian characteristics of a high peak and long tail and the latter shows strong dependence among the subbands, which is shown in Fig. 2.3[14]

2.2.2 Information Theoretical Weighting Based on GSM model

In [7], the distorted image is modelled as the one with some process of the reference image:

$$Y_i^j = g_i^j X_i^j + V_i^j \tag{2.4}$$

where X_i^j is a vector of coefficients of the reference image at scale i and orientation j; Y_i^j is a vector of coefficients of the distorted image; g_i^j and V_i^j can be regarded as contrast and luminance distortions.

HVS is assumed as an additive Gaussian communication channel with constant noise. If the X_i^j and Y_i^j are transmitted through the channel, the received or perceived signals will be

$$\begin{aligned} E_i^j &= X_i^j + N_1 \\ F_i^j &= Y_i^j + N_2 = Y_i^j = g_i^j X_i^j + V_i^j + N_2 \end{aligned} \tag{2.5}$$

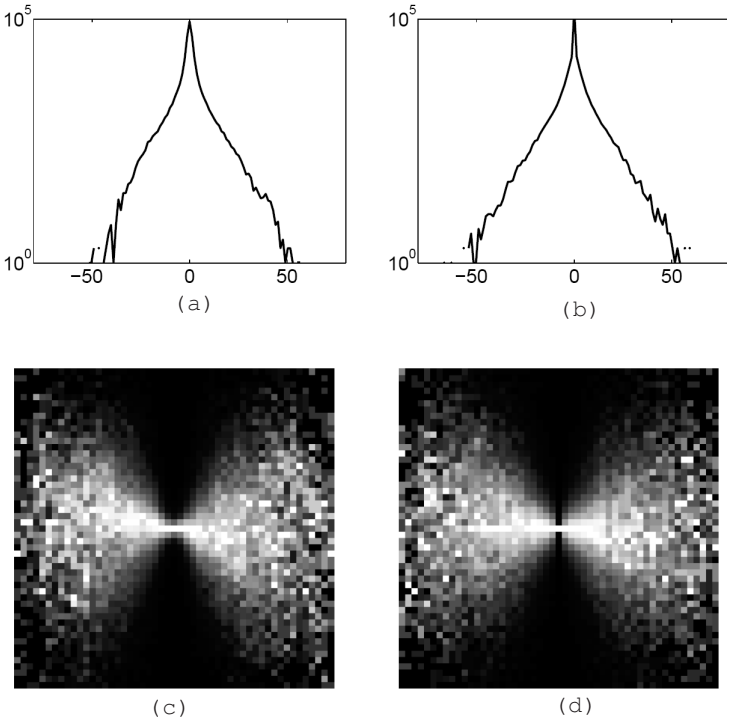

Figure 2.3. Comparison of coefficient statistics from an example image subband (a vertical subband of Fig.2.2 (a), left panels) with those arising from simulation of a local GSM model (right panels). Model parameters (covariance matrix and the multiplier prior density) are estimated by maximizing the likelihood of the observed set of wavelet coefficients. (a,b) Log marginal histograms. (c,d) Conditional histograms of two spatially adjacent coefficients. Brightness corresponds to probability, except that each column has been independently rescaled to fill the range of display intensities.

The mutual information between E_i^j and X_i^j,and that between F_i^j and X_i^j are computed as

$$\begin{aligned} I(E_i^j; X_i^j | S_i^j) &= \frac{1}{2}\sum_{k=1}^{M} \log_2(1 + \frac{s_i^j \lambda_k}{\sigma_N^2}) \\ I(F_i^j; X_i^j | S_i^j) &= \frac{1}{2}\sum_{k=1}^{M} \log_2(1 + \frac{{g_i^j}^2 s_i^j \lambda_k}{\sigma_{V_i^j}^2 + \sigma_N^2}) \end{aligned} \tag{2.6}$$

where λ_k is the eigenvalue of C_U and $\sigma_{N_1} = \sigma_{N_2} = \sigma_n$.

Since E_i^j and F_i^j are correlated because of X_i^j, the mutual information between E_i^j and F_i^j is

$$I(E_i^j; F_i^j|S_i^j) = \frac{1}{2}\log_2\left(\frac{|s_i^j C_U + \sigma_n^2 I||{g_i^j}^2 s_i^j C_U + \sigma^2_{v_i^j} I + \sigma_n^2 I|}{|s_i^j C_U(\sigma^2_{v_i^j} + \sigma_n^2) + {g_i^j}^2 s_i^j C_U \sigma_n^2 + \sigma_n^2(\sigma_n^2 I + \sigma^2_{v_i^j} I)|}\right) \tag{2.7}$$
$$= \frac{1}{2}\sum_{k=1}^{M}\log_2\left(\frac{(s_i\lambda_k + \sigma_n^2)({g_i^j}^2 s_i^j \lambda_k + \sigma^2_{v_i^j} + \sigma_n^2)}{(s_i^j\lambda_k(\sigma^2_{v_i^j} + \sigma_n^2) + {g_i^j}^2 s_i^j \lambda_k \sigma_n^2 + \sigma_n^2(\sigma_n^2 + \sigma^2_{v_i^j}))}\right)$$

The detailed derivation of Eq.(2.7) is given here.

First, we review the mutual information between two Gaussian. If X, Y are Gaussian distributed, then the mutual information of X, Y is

$$I = \frac{1}{2}\log_2(\frac{|C_X||C_Y|}{|C|})$$

where C_X and C_Y are covariance matrices of X and Y. C is defined as $C = \begin{bmatrix} C_X & C_{XY} \\ C_{YX} & C_Y \end{bmatrix}$

Second, given the general case

$$Y_1 = a_1 X + N_1 \tag{2.8}$$
$$Y_2 = a_2 X + N_2$$

where N_1, N_2, X are independent Gaussian with zero mean and a_1, a_2 are known scalars.

If X, N_1 and N_2 are independent and identically-distributed, then Y_1, Y_2 are Gaussian distributed.

$$C_{Y_1} = a_1^2 C_X + \sigma_{n1}^2 I \tag{2.9}$$

$$C_{Y_2} = a_2^2 C_X + \sigma_{n2}^2 I$$

$$C_{Y_1 Y_2} = E(Y_1 Y_2^T) = a_1 a_2 C_X$$

$$C = \begin{bmatrix} a_1^2 C_X + \sigma_{n1}^2 I & a_1 a_2 C_X \\ a_1 a_2 C_X & a_2^2 C_X + \sigma_{n2}^2 I \end{bmatrix}$$

$$|C| = |a_1^2 C_X \sigma_{n2}^2 + a_2^2 C_X \sigma_{n1}^2 + \sigma_{n2}^2 \sigma_{n1}^2 I| \tag{2.10}$$

The mutual information between Y_1 and Y_2 are:

$$\begin{aligned} I(Y_1; Y_2) = \frac{1}{2}\log_2(\frac{|C_{Y_1}||C_{Y_2}|}{|C|}) &= \frac{1}{2}\log_2(\frac{|a_1^2 C_X + \sigma_{n1}^2 I||a_2^2 C_X + \sigma_{n2}^2 I|}{|C|}) \\ &= \frac{1}{2}\log_2(\frac{|a_1^2 C_X + \sigma_{n1}^2 I||a_2^2 C_X + \sigma_{n2}^2 I|}{|a_1^2 C_X \sigma_{n2}^2 + a_2^2 C_X \sigma_{n1}^2 + \sigma_{n2}^2 \sigma_{n1}^2 I|}) \end{aligned} \tag{2.11}$$

Third, given

$$\begin{aligned} E_i^j &= X_i^j + N_1 = \sqrt{S_i^j} U + N_1 \\ F_i^j &= Y_i^j + N_2 = Y_i^j = g_i^j X_i^j + V_i^j + N_2 = g_i^j \sqrt{S_i^j} U + V_i^j + N_2 \end{aligned} \tag{2.12}$$

and according to Eq.(2.11), the mutual information between E_i^j and F_i^j, $I(E_i^j; F_i^j | S_i^j)$ is computed as Eq. (2.7).

Fig. 2.4 shows the relationship among $I(E_i^j; X_i^j | S_i^j)$, $I(F_i^j; X_i^j | S_i^j)$ and $I(E_i^j; F_i^j | S_i^j)$.

Consequently, the overall information extracted from the reference image in the subband is:

$$W(X_i^j, Y_i^j) = I(\vec{E}_i^j; \vec{X}_i^j | S_i^j) + I(\vec{F}_i^j; \vec{X}_i^j | S_i^j) - I(\vec{E}_i^j; \vec{F}_i^j | S_i^j) \tag{2.13}$$

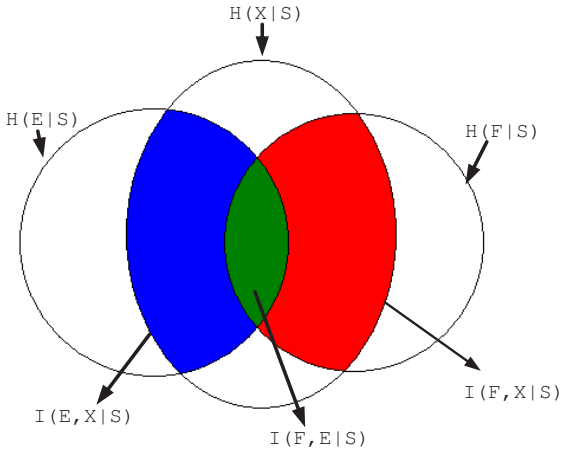

Figure 2.4. The relationship among $I(E_i^j; X_i^j | S_i^j)$, $I(F_i^j; X_i^j | S_i^j)$ and $I(E_i^j; F_i^j | S_i^j)$.

The total information at scale i is defined as the summation of $W(X_i^j, Y_i^j)$ along all the orientation.

$$IW(X_i, Y_i) = \sum_{j \in orientations} W(X_i^j, Y_i^j) \tag{2.14}$$

Fig.2.5 (b) gives an example of information theoretic weighting map at the finest scale computed by Eq. (2.14), which shows the larger weights in the areas such as the eyes and the mouth and smaller weights in the smooth regions, for example the background.

2.2.3 Perceptual Multi-scale Weights

It is observed that the frequency sensitivity of the HVS varies with viewing conditions including the resolution and the view distances. There is an interesting example to demonstrate this characteristic of the HVS. An artificial hybrid image Fig. 2.6, generated by Dr. Aude Oliva [2] from MIT, combines the detailed features such as wrinkles and the mustache of Einstein with the rough outline of Monroe. When viewing the picture in a near distance, we see Einstein. While at a far distance (e.g. more than 1 meter), we find a smiling Monroe.

(a) (b)

Figure 2.5. Information theoretical weighting map. (a) original image, (b) Information theoretical weighting map at scale 0. The brighter regions indicate larger weights.

Therefore, it is reasonable to assume the image quality of the human perception is also related to the view conditions. In order to find the right scale (view distance), Wang et al. [5] proposed the multiscale SSIM index as

$$SSIM_m(X,Y) = [l_M(X,Y)]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(X,Y)]^{\beta_j} [s_j(X,Y)]^{\gamma_j} \tag{2.15}$$

where M is the highest scale and α, β and γ are weighting parameters gauged by psychovisual experiments.

2.2.4 Information Weighted Multi-Scale SSIM Index

Incorporating the information theoretic weighting and the perceptual scale weighting, we define a novel IQA metric, information weighted multi-scale SSIM (IW_MultiSSIM) index, as

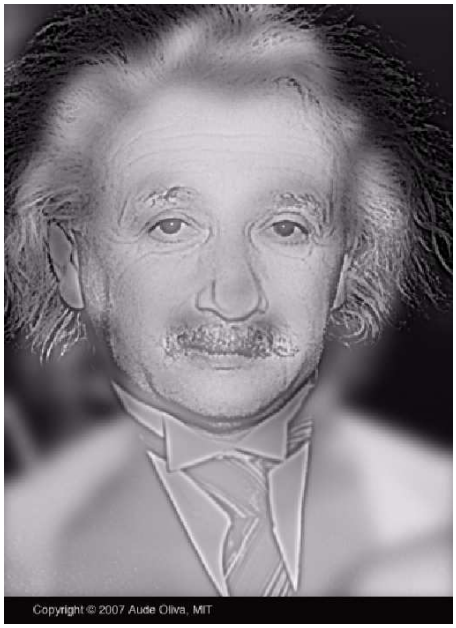

Figure 2.6. Hybrid Einstein-Monroe image from [2].

$$IW\_MultiSSIM(X,Y) = [E(l(X_M, Y_M))]^{SW_M} \prod_{i=1}^{M} \left[\frac{\sum IW(X_i, Y_i) c(X_i, Y_i) s(X_i, Y_i)}{\sum IW(X_i, Y_i)} \right]^{SW_i} \quad (2.16)$$

where $l(X_M, Y_M)$ is the luminance similarity map between X and Y at scale M as

$$l(X_M, Y_M) = \frac{2\mu_{X_M}\mu_{Y_M} + C_1}{{\mu_{X_M}}^2 + {\mu_{Y_M}}^2 + C_1} \quad (2.17)$$

$c(X_i, Y_i)$ is the contrast similarity map between X and Y at scale i as

$$c(X_i, Y_i) = \frac{2\sigma_{X_i}\sigma_{Y_i} + C_2}{{\sigma_{X_i}}^2 + {\sigma_{Y_i}}^2 + C_2} \quad (2.18)$$

$s(X_i, Y_i)$ is the structural similarity map between X and Y at scale i as

$$s(X_i, Y_i) = \frac{\sigma_{X_i Y_i} + C_3}{\sigma_{X_i}\sigma_{Y_i} + C_3} \quad (2.19)$$

$IW(X_i, Y_i)$ is the information theoretic weight defined in Eq.(2.14).

The perceptual weights for five scale are estimated from subjective test in [5], which are shown in Table 2.1.

Table 2.1. Multi-scale weights for SSIM [5]

weight	scale 1	scale 2	scale 3	scale 4	scale 5
SW	0.0448	0.2856	0.3001	0.2363	0.1333

2.2.5 Test

In order to validate the IW_MultiSSIM index, the tests are extensively carried out across four publicly-accessible image quality assessment databases including UT-Austin LIVE database[17], the Cornell-VCL A57 database[67], the IVC database[65] and the Toyama database[66]. The result of IW_MultiSSIM index is compared with other state-of-the-art perceptual IQA metrics.

The results of LPCC and SROCC are shown in Table.2.2 and Table.2.3 for UT-Austin LIVE database.

The results of the Cornell-VCL A57 database are in Table 2.4.

Table 2.2. LPCC after nonlinear regression based on LIVE database. SSIM(MS): multi-scale SSIM[5], WSSIM: SSIM with local information contend based on Gaussian model[6]; VIF: visual information fidelity[7], VSNR: visual SNR[8](*for JPEG and JP2K, the results are the combination of 1 and 2)

	JP2K-1	JP2K-2	JPEG-1	JPEG-2	WN	GBlur	FF	All data
PSNR	0.9332	0.8740	0.8856	0.9167	0.9859	0.784	0.8895	0.8709
SSIM(MS)	0.9702	0.9711	0.9699	0.9879	0.9737	0.9487	0.9304	0.9393
SSIM	0.9648	0.9684	0.9667	0.9813	0.9852	0.9454	0.9540	0.9436
WSSIM[6]	0.9714	0.9738	0.9645	0.9835	0.9869	0.9733	0.9675	0.9426
VIF	0.9791	0.9787	0.9714	0.9885	0.9877	0.9762	0.9704	0.9533
VSNR	0.957*		0.923*		0.978	0.934	0.902	0.889
IW_MultiSSIM	0.9738	0.9774	0.9707	0.9855	0.9883	0.9699	0.9503	0.9565

Table 2.3. SROCC based on LIVE database. SSIM(MS): multi-scale SSIM[5],WSSIM: SSIM with local information contend based on Gaussian model[6]; VIF: visual information fidelity[7], VSNR: visual SNR[8](*for JPEG and JP2K, the results are the combination of 1 and 2)

	JP2K-1	JP2K-2	JPEG-1	JPEG-2	WN	GBlur	FF	All data
PSNR	0.9263	0.8549	0.8779	0.7708	0.9854	0.7823	0.8907	0.8755
SSIM(MS)	0.9645	0.9648	0.9702	0.9454	0.9805	0.9519	0.9395	0.9527
SSIM	0.9545	0.9636	0.9598	0.9028	0.9737	0.9497	0.9546	0.9475
WSSIM[6]	0.9612	0.9743	0.9591	0.9401	0.9776	0.9716	0.9659	0.9494
VIF	0.9721	0.9719	0.9699	0.9439	0.9828	0.9706	0.9649	0.9584
VSNR	0.946*		0.908*		0.979	0.941	0.906	0.889
IW_MultiSSIM	0.9647	0.9745	0.9654	0.9442	0.9832	0.9696	0.9514	0.9627

Table 2.4. Results of Cornell database

	LPCC	MAE	RMS	SROCC
PSNR	0.6347	0.1607	0.1899	0.6205
SSIM (downsampled by 2)	0.7994	0.1209	0.1476	0.8061
SSIM (Multiscale)	0.8802	0.0945	0.1166	0.8587
WSSIM[6]	0.8886	0.0908	0.1127	0.8671
VSNR	0.9146	0.0809	0.0994	0.9360
VIF	0.6139	0.1421	0.1940	0.6225
IW_MultiSSIM	0.9303	0.0770	0.0901	0.9007

The results of the IVC database are in Table 2.5 and the scatter plots are displayed in Fig.2.7

For Toyama database, the scatter plots of different IQA methods are displayed in Fig.2.8. The results are shown in Table. 2.6.

From the results based on all the four databases, we can see the IW_MultiSSIM index always performs statistically equivalent to the best methods. More importantly, the consistent improvement of our method compared with SSIM[70], multi-scale SSIM[5] and weighted SSIM [6] demonstrates the validity of our information theoretical weighting strategy.

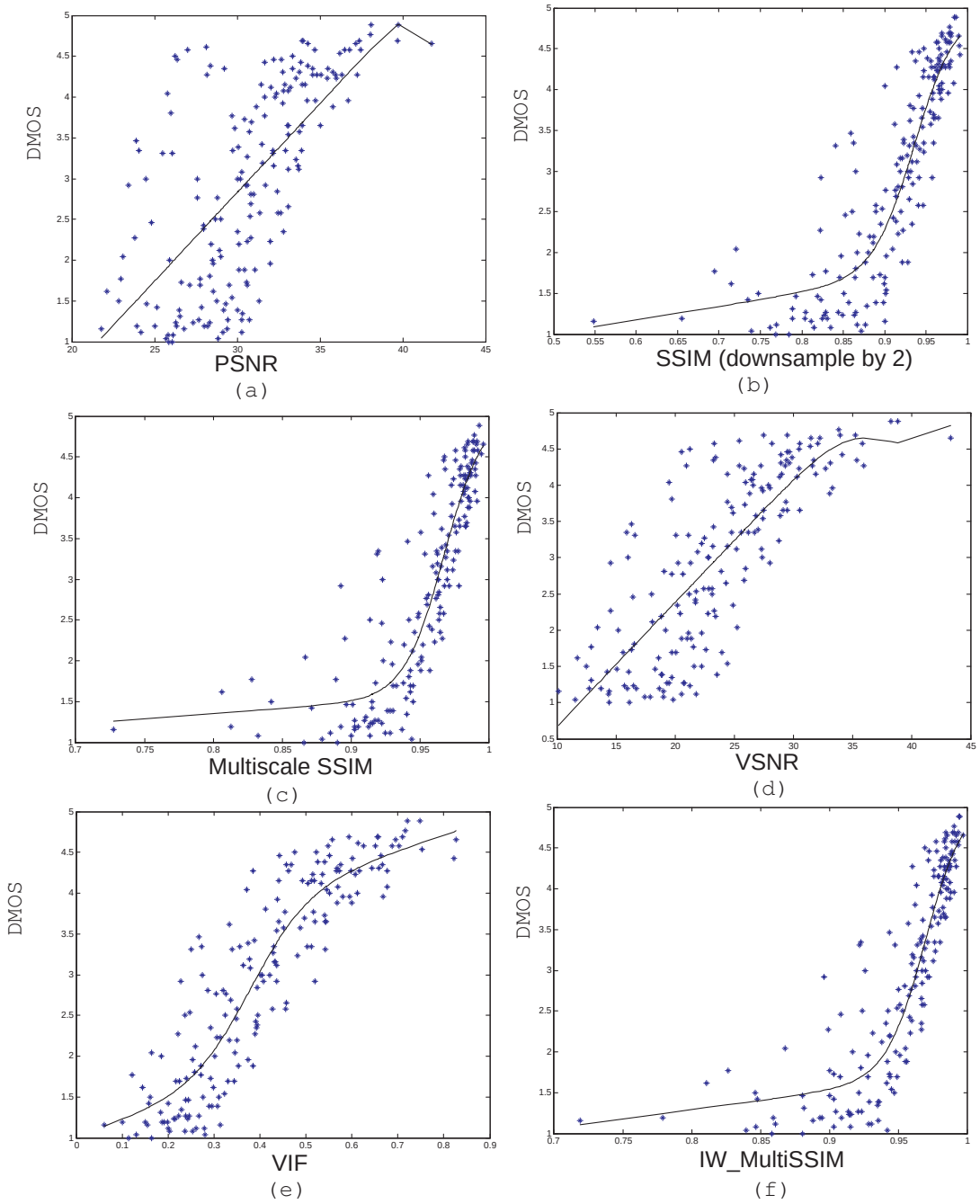

Figure 2.7. IVC database: scatter plot between objective scores and subjective scores without fitting. (a) DMOS against PSNR, (b) DMOS against SSIM(downsample by 2), (c) DMOS against Multiscale SSIM, (d) DMOS against VSNR, (e) DMOS against VIF, (f) DMOS against IW_MultiSSIM .

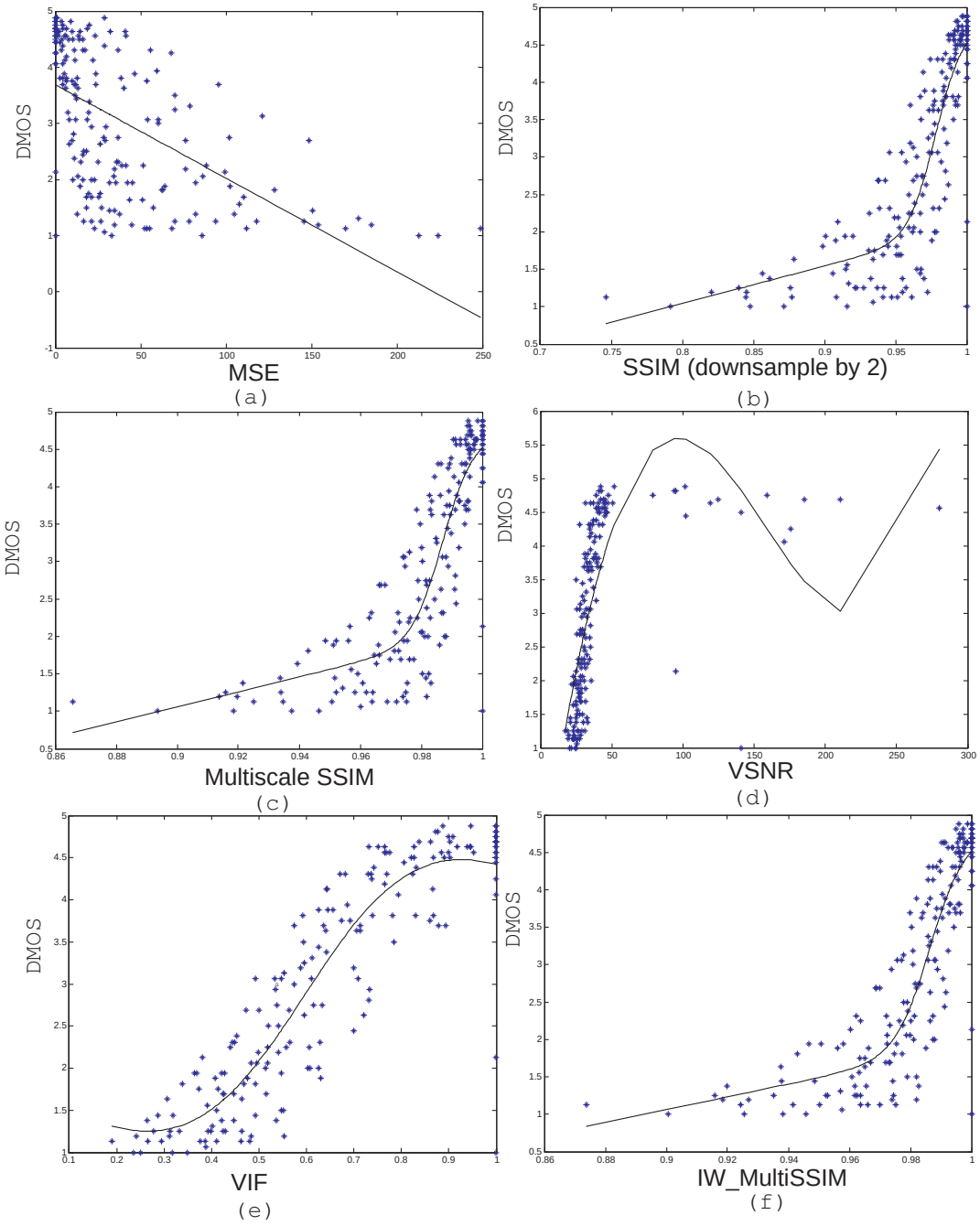

Figure 2.8. Toyama database: scatter plot between objective scores and subjective scores without fitting. (a) DMOS against PSNR, (b) DMOS against SSIM(downsample by 2), (c) DMOS against Multiscale SSIM, (d) DMOS against VSNR, (e) DMOS against VIF, (f) DMOS against IW_MultiSSIM .

Table 2.5. Results of IVC database

	LPCC	MAE	RMS	SROCC
PSNR	0.6718	0.7192	0.9025	0.6947
SSIM (downsampled by 2)	0.9004	0.3978	0.5301	0.8977
SSIM (Multiscale)	0.9082	0.3837	0.5100	0.8992
WSSIM[6]	0.8936	0.4094	0.5468	0.8902
VSNR	0.7959	0.5778	0.7376	0.8112
VIF	0.9026	0.4098	0.5245	0.9004
IW_MultiSSIM	0.9150	0.3736	0.4914	0.9075

Table 2.6. Results of Toyama database

	LPCC	MAE	RMS	SROCC
MSE	0.5748	0.9408	1.0801	0.6773
SSIM (downsampled by 2)	0.8780	0.4495	0.6319	0.8640
SSIM (Multiscale)	0.8746	0.4656	0.6399	0.8645
WSSIM[6]	0.7213	0.6777	0.9143	0.6874
VSNR	0.7536	0.7159	0.8677	0.8563
VIF	0.9030	0.4148	0.5670	0.8868
IW_MultiSSIM	0.8852	0.4390	0.6141	0.8697

2.3 Information Theoretical Weighting Based on a Statistical Model of Human Visual Speed Perception for FRVQA

2.3.1 Previous application of motion in Video Quality Assessment

The capability of representing motion is probably the most critical characteristic that distinguishes a natural video sequence from a stack of independent still image frames. If we believe that the central goal of vision is to extract useful information from the visual scene, then the perception of motion information would play an important role in the perception of natural video. Since the main purpose of objective video quality assessment (VQA) is to predict human behavior in the evaluation of video quality, it would be essential for a successful VQA system to effectively take into account motion information.

Nevertheless, in the literature of VQA, motion information has typically been employed *indirectly*. The most frequently used method is temporal filtering [32, 73], where linear filters or filter banks are applied along the temporal direction (or along the spatial and the temporal directions simultaneously), and the filtered signals are normalized to reflect the effect of the temporal contrast sensitivity function [74] (the variation of human visual sensitivity as a function of temporal frequency). Advanced models may also include the temporal masking effects (the reduction of visibility of one image component due to the existence of its neighboring components) [73] or statistics of the temporal filter coefficients [31]. Since motion in the visual scene may cause variations in signal intensity along the temporal direction, temporal filtering can, to some extent, capture motion. However, representing motion using temporal filtering responses is indirect, inaccurate, and in some sense problematic. First, motion may not be the sole reason for temporal signal intensity variations. The change of lighting conditions is an obvious counterexample. Therefore, the temporal filter coefficients are indeed a mixture effect of motion together with many other reasons. Second, the speed of motion cannot be directly related to the strength of temporal filter responses. For example, two objects with the same speed of motion but different texture and contrast would result in different speeds of temporal intensity variation, and thus different temporal filter responses. Third, many visual experiments that measure temporal visual sensitivities were done with flickering patterns [32], which do not reflect any physical motion of the objects. Moreover, since the motion and speed information is not represented explicitly, a lot of knowledge about motion perception cannot be directly used within such a temporal filtering framework.

Only a relatively small number of existing VQA algorithms detect motion explicitly and use motion information directly. Wang *et al.* proposed a heuristic weighting model [40], which was combined with the structural similarity (SSIM)[70] based

quality assessment method to take into account the fact that the accuracy of visual perception is significantly reduced when the speed of motion is extremely large. A set of heuristic fuzzy rules was proposed by Lu *et al.* [30] that use both absolute and relative motion information to account for visual attention and motion suppression. It was shown that these rules are effective in improving VQA performance of the standard mean squared error (MSE)/peak signal-to-noise ratio (PSNR) measures as well as the SSIM [70] approach. In two recent papers by Seshadrinathan and Bovik, local motion information obtained from optical flow computation is employed to adaptively guide the orientation of a set of three-dimensional Gabor filters [43, 75]. The adopted Gabor filter responses are then incorporated into the SSIM [70, 38] and the visual information fidelity (VIF) [7] measures for the purpose of VQA.

2.3.2 Perceptual Motion Information

Our approach is largely inspired by the recent psychophysical study by Stocker and Simoncelli on human visual speed perception [3]. Based on a Bayesian optimal observer hypothesis, Stocker and Simoncelli [3] measured the prior and the likelihood probability distributions of speed perception simultaneously from a set of carefully designed psychovisual experiments. These measured probability distributions are consistent across human subjects and can be modelled using simple parametric functions. These results are substantially different from previous statistical models of visual speed perception [15, 76, 77], where the prior distributions are often assumed rather than measured. Our information theoretical approach has greatly benefited from these results, because the statistical models derived from them provide the essential ingredients in the computation of the perceptual motion information including the information content and the perceptual uncertainty. Our method is based on the following assumptions and observations.

First, we believe that the human visual system (HVS) is an efficient encoder or information extractor (subject to certain physical constraints such as power consumption), as widely hypothesized in computational vision science [78, 7]. To achieve such efficiency, it is natural to assume that the areas in the visual scene that contain more information should be more likely to attract visual attention and fixations [79, 80]. Such *information content* can be quantified using statistical information theory, provided that a statistical model about the information source is available. In fact, information content-based method has already shown to be useful in still image quality assessment (IQA) [6].

Second, as in a number of previous papers [31, 7, 6], we model visual perception as an information communication process, where the information source (the video signal) passes through an error-prone communication channel (the HVS). The key difference from the previous IQA/VQA models is that the noise level in the communication channel is not fixed here. This is motivated by the empirical observation that the HVS does not perceive all the information content with the same degree of certainty. For example, when the background motion in a video sequence is very large (or the head/camera motion is very large), the HVS cannot identify the objects presented in the video with the same accuracy as in static background images, i.e., the video signal is perceived with higher uncertainty. Again, such *perceptual uncertainty* can be quantified based on information theory, by relating the stochastic channel distortion model with the speed of motion. In particular, the psychophysical study by Stocker and Simoncelli [3] suggests that the internal noise of human visual speed perception increases with the true stimulus speed and decreases with the stimulus contrast.

In the following section, we elaborate how to compute locally (in both space and time) the information content and the perceptual uncertainty based on the physical

motion information estimated from the video sequence. Perceptual motion information is defined as a spatio-temporal function of the information content and the perceptual uncertainty, which is incorporated as weighting factors into any local VQA algorithm that produces a quality/distortion map over space and time. Fig.2.9 shows the framework to extract the perceptual motion information.

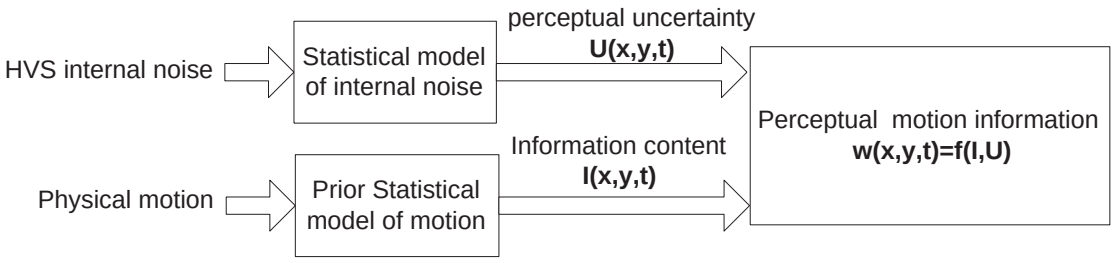

Figure 2.9. Perceptual motion information.

2.3.3 Method

The motion information in a video sequence can be represented as a three-dimensional field of motion vectors, where each spatial location (x, y) and time instance t is associated with a motion vector $\vec{v}(x, y, t) = [v_x(x, y, t)\ v_y(x, y, t)]^T$. For notational convenience, we often drop the space and time indices and write a motion vector as $\vec{v}$. For a given video sequence, we consider three types of motion fields – absolute motion, background motion, and relative motion. An illustration is given in Fig. 2.10, where the absolute motion $\vec{v}_a$ is estimated as the absolute pixel movement at each spatial location between two adjacent video frames. By contrast, the background motion $\vec{v}_g$ is global, which is often caused by the movement of the image

acquisition system. We also define a relative motion $\vec{v}_r$ at each spatial location as the vector difference between the absolute and the global motion, i.e.,

$$\vec{v}_r = \vec{v}_a - \vec{v}_g \,. \tag{2.20}$$

The speed of motion can be computed as the length of the motion vector, which, for convenience, we denote as $v = \|\vec{v}\|_2$, L_2-norm of motion vector. Thus, v_g, v_a and v_r represent the speed of the background motion, the absolute motion, and the relative motion, respectively.

A recent approach in understanding human visual speed perception is to use a Bayesian optimal observer model, in which the visual system judges the speed of motion by "optimally" combining some prior knowledge of the visual world together with the current noisy measurements [15, 76, 3]. It has been shown that this approach can successfully explain a number of psychovisual phenomena where the visual system tends to give biased judgements on the speed of retinal motion [15, 76, 3]. Fig. 2.11 describes this approach in an information communication framework, where the stimulus speed information v passes through a noisy front-HVS channel. This results in the internal noisy measurement m, which is associated with a statistical noise model, or a likelihood function. The visual system gives an estimate of the stimulus speed $\hat{v}$ not only from m, but also based on some prior information about the probability distribution of the stimulus speed. It is assumed that the prior distribution has been established beforehand in the brain by sufficient statistics about the natural visual environment. Fig. 2.11 also shows the prior distribution and the noise likelihood function measured by Stocker and Simoncelli [3]. Here, we will describe how these measurements help us develop models to compute both the information content and the perceptual uncertainty of speed perception and how to combine them for VQA.

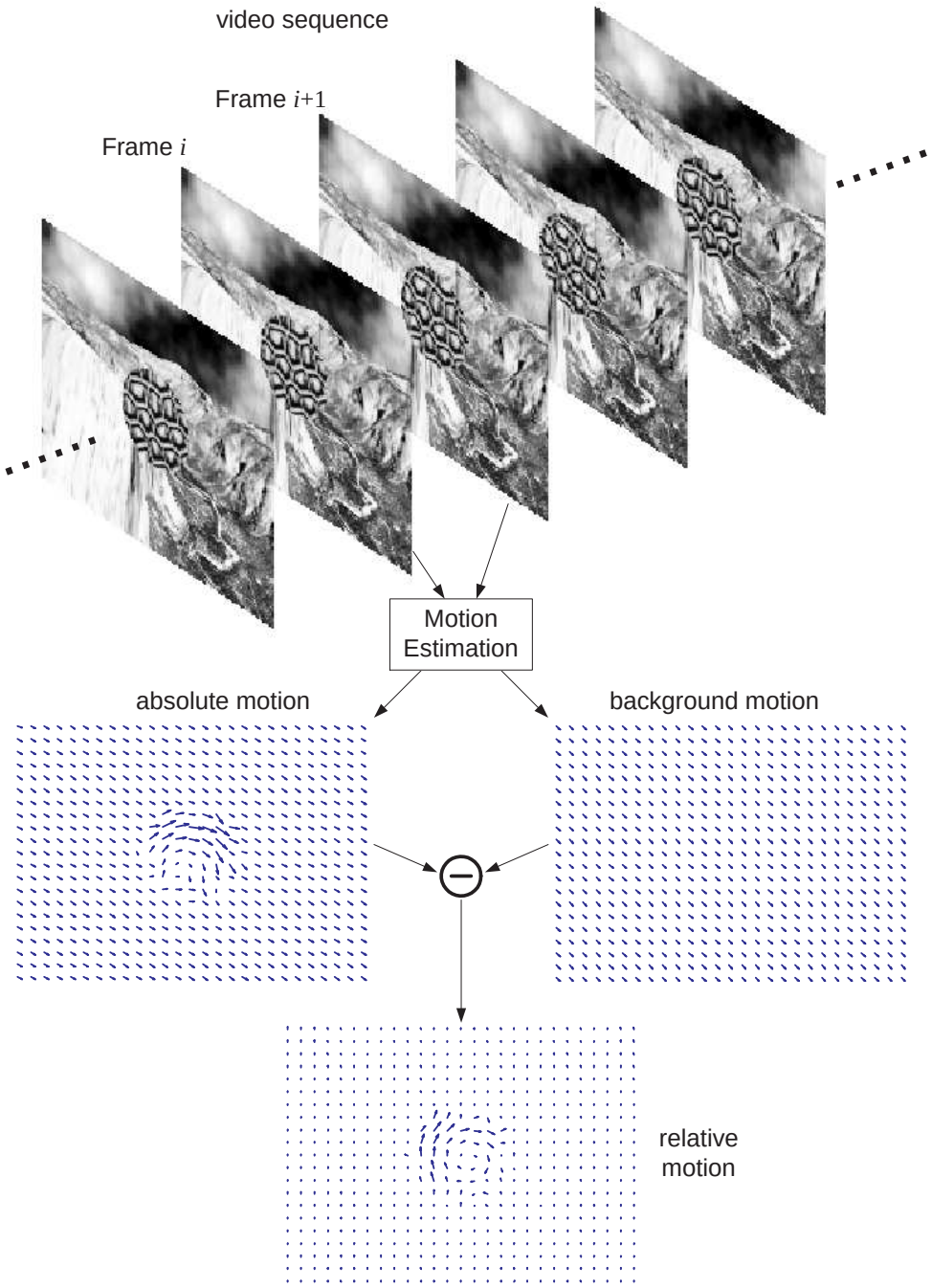

Figure 2.10. Illustration of absolute motion, background motion and relative motion estimated from two consecutive frames of a video sequence.

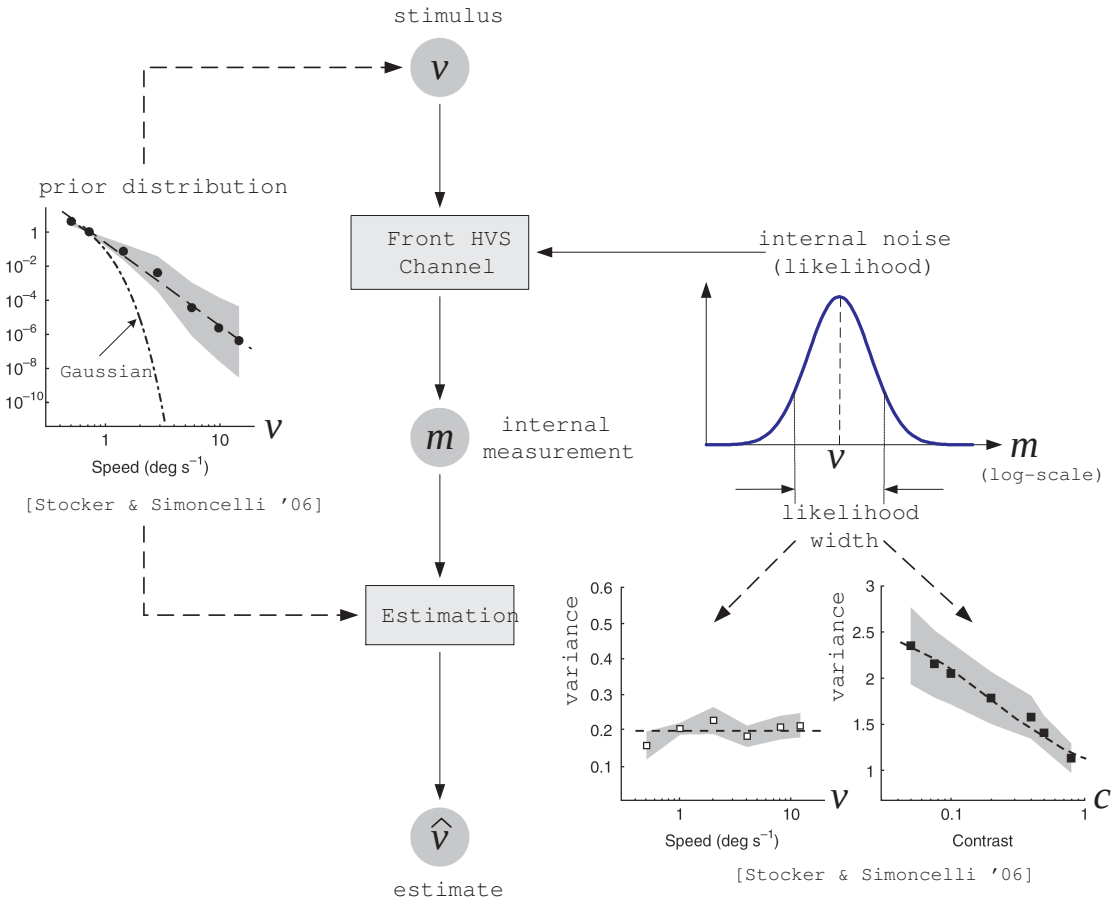

Figure 2.11. Bayesian visual speed perception in an information communication framework. v: stimulus speed; m: noisy measurement; $\hat{v}$: estimated speed; c: stimulus contrast. Adapted from [Stocker & Simoncelli '06] [3].

2.3.3.1 Information Content

It is believed that object motion is associated with visual attention and can be used for predicting visual fixations. This is intuitively sensible because statistically, most of the objects in the visual world are static (or close to static) relative to the background. As a result, an object with significant motion with respect to the background would be a strong *surprisal* to the visual system. If the HVS is an efficient information extractor, as discussed in section 2.3.2, then it should pay more attention

to such a surprising event. This intuitive idea may be converted into a quantitative measure of motion information content (or how surprising the event is), provided that the prior probability distribution about the speed of motion is known. Early work on Bayesian speed perception has assumed Gaussian distribution for the speed prior [15], but the recent result by Stocker and Simoncelli [3] suggests that the distribution has a much longer tail than Gaussian, as shown in Fig. 2.11. Indeed, it can be well fitted with a straight line in the log-log domain (see Fig. 2.11). This leads us to assume a power-law function for the prior distribution of relative motion:

$$p(v_r) = \frac{\tau}{v_r^{\alpha}}, \tag{2.21}$$

where τ and α are two positive constants. Since the power-law function does not sum to a finite number, this is not a strictly valid probability density function and can only be used when v_r is away from 0. For any observed motion v_r, we can then estimate the information content associated with it by computing its self-information or surprisal as

$$I = -\ln p(v_r) = \alpha \ln v_r + \beta, \tag{2.22}$$

where $\beta = -\ln \tau$ is a constant. Eq. (2.22) suggests that the motion information content increases with the speed of relative motion, which is consistent with our intuition discussed earlier.

2.3.3.2 Perception Uncertainty

If we model visual perception as an information communication process, then the amount of information that can be received (perceived) at the receiver end will largely depend on the noise in the distortion channel (the HVS). In other words, the internal noise in the HVS, or the likelihood function of the noisy measurement, determines the perceptual uncertainty. It was found that for a given stimulus speed,

a log-normal distribution can provide a good description of the likelihood function [3]:

$$p(m|v_s) = \frac{1}{\sqrt{2\pi}\sigma m} \exp\left[\frac{-(\ln m - \ln v_s)^2}{2\sigma^2}\right], \tag{2.23}$$

where v_s and m are the speeds of the true stimulus motion and the measurement, respectively. Furthermore, the experimental results by Stocker and Simoncelli [3] suggest that in the logarithmic speed domain, the width parameter σ in the log-normal distribution is roughly constant for any stimulus speed v_s and inversely dependent on the stimulus contrast c, as illustrated in Fig. 2.11. Note that the width here is represented in the log-domain, and thus it indeed scales linearly with v_s in the linear speed domain. Mathematically, we model it as

$$\sigma = \frac{\lambda}{c^\gamma}, \tag{2.24}$$

where λ and γ are both positive constants.

For a given video sequence, we assume that the underlying stimulus speed v_s is the speed of the background motion v_g. A natural way to quantify the level of the internal noise, or the perceptual uncertainty, is the entropy of the likelihood function, which can be computed as

$$\begin{aligned} U &= -\int_{-\infty}^{\infty} p(m|v_g)[\ln p(m|v_g)]dm \\ &= \frac{1}{2} + \frac{1}{2}\ln(2\pi\sigma^2) + \ln v_g \\ &= \ln v_g - \gamma \ln c + \delta, \end{aligned} \tag{2.25}$$

where $\delta = \frac{1}{2} + \frac{1}{2}\ln(2\pi) + \ln\lambda$ is a constant. Again, this perceptual uncertainty measurement is consistent with our intuition. On the one hand, it increases with the background motion of the video frame, suggesting that when the background motion is very large, the HVS cannot extract the structural information about the objects presented in the video with the same accuracy as in static images. On the other

hand, it decreases with the stimulus contrast, implying that higher contrast objects are perceived with lower uncertainty.

2.3.3.3 VQA Based on Perceptual Motion Information

We compute the motion information content and the perceptual uncertainty at every spatial location and time instance (x, y, t) in the video sequence. Based on the efficient coding hypothesis about the HVS, the importance of a visual event should increase with the information content, and decrease with the perceptual uncertainty. Therefore, perceptual motion information is defined as the following spatiotemporal importance weight function at every (x, y, t)

$$w = I - U = (\alpha \ln v_r + \beta) - (\ln v_g - \gamma \ln c + \delta)\,. \tag{2.26}$$

The calculation of the information content, the perceptual uncertainty and the importance weighting function is demonstrated in Fig. 2.12, where two consecutive video frames are extracted from the "Mobile Calendar" sequence, and the motion field as well as the maps for I, U and w are computed. It is observed from the video sequence that the toy train moves from the right to the left with respect to the moving background. Note that although the absolute motion of the train is almost static, their relative motion is significant. Thus, based on our model, the region associated with the train is given larger weights relative to the background.

The importance weight function alone cannot serve as a VQA algorithm. However, it can be incorporated into a local image quality/distortion measure as a weighting function. The local image quality/distortion measure must provide a 3D quality/distortion map of the video sequence being evaluated. Let $q(x, y, t)$ be the qual-

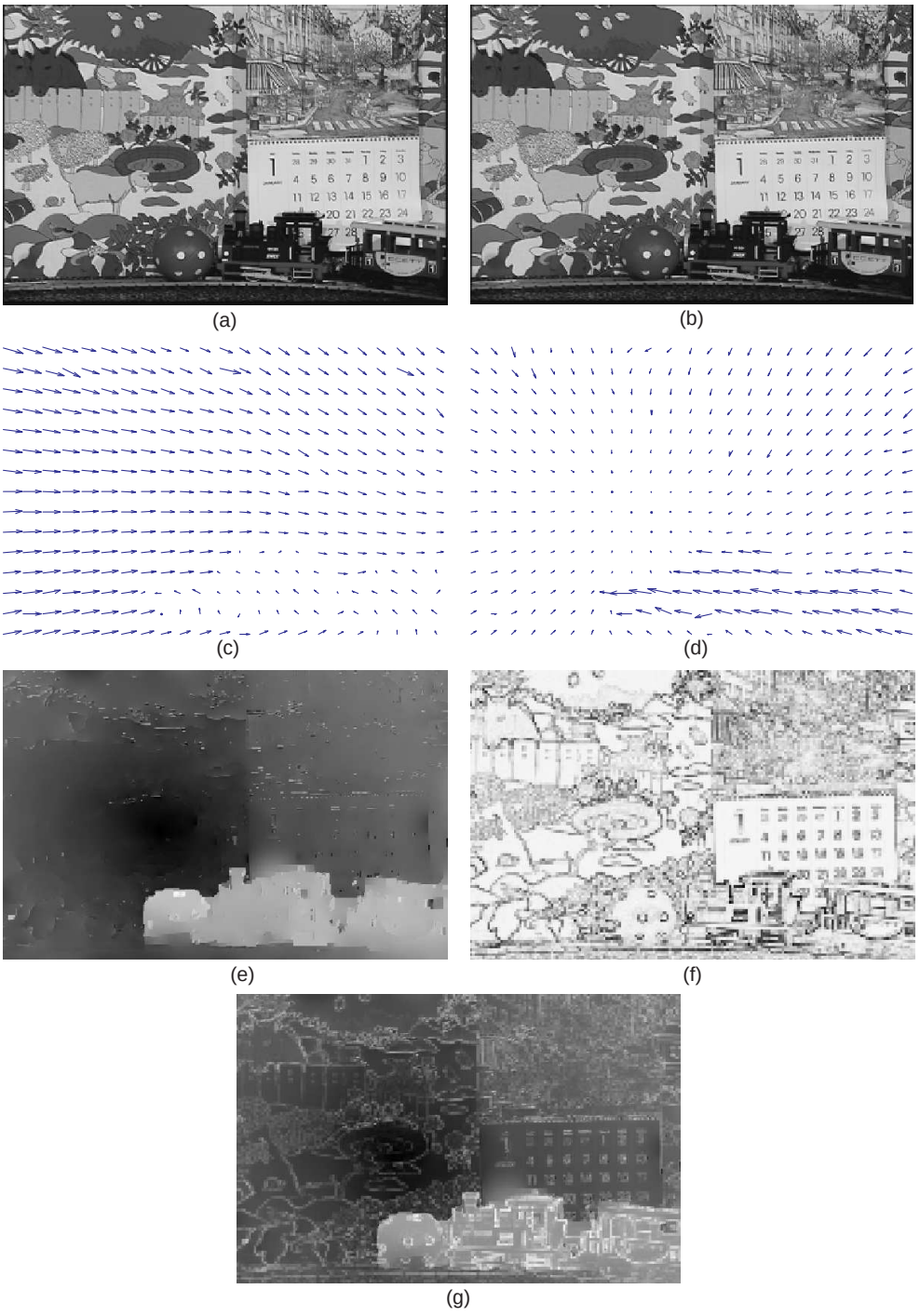

Figure 2.12. (a),(b) Two consecutive frames extracted from the "Mobile Calendar" sequence; (c) Estimated absolute motion field; (d) Estimated relative motion field; (e) Estimated local information content map; (f) Estimated local perceptual uncertainty map; (g) Estimated local weighting factor map.

ity/distortion map given by the local quality/distortion metric, the final VQA score is computed as

$$Q = \frac{\sum_t \sum_x \sum_y w(x,y,t)\, q(x,y,t)}{\sum_t \sum_x \sum_y w(x,y,t)}\,. \tag{2.27}$$

where x, y are spatial coordinates and t is the temporary time index.

2.3.4 Implementation

To build a real VQA system based on the proposed approach, several implementation issues need to be resolved. First, we need to estimate the motion vector field. Rather than using block matching-based motion estimation as in previous work [40], here we choose to use a optical flow method for motion estimation, which avoids the computationally intensive block search procedures and provides smoother motion vector field. In particular, we compute the absolute motion field using Black and Anandan's multi-layer optical flow estimation algorithm [81] with a five-level pyramid decomposition. The background motion is obtained by a maximum likelihood estimation to identify the peak of the histogram of motion vectors on the 2D grid [82]. The relative motion vector $\vec{v}_r$ is then computed using Eq. (2.20).

Second, the local contrast needs to be computed at each spatial location and time instance. Although contrast is an extensively used term throughout the filed of visual psychophysics and physiology, mathematical definition of local contrast for complex natural images is a nontrivial issue [83]. Here we compute the local contrast as the ratio between the local standard deviation normalized by the local mean, i.e., for a given local image patch p, we define

$$c' = \frac{\sigma_p}{\mu_p + \mu_0} \tag{2.28}$$

where σ_p and μ_p are the standard deviation and the mean computed within the local patch, respectively, and μ_0 is a small constant to avoid instability near 0. In addition,

as in previous models [84, 25], to take into account the contrast response saturation effect at small and large contrast values, we pass the contrast computation through a pointwise nonlinear function given by

$$c = 1 - e^{-(c'/\theta)^{\rho}}, \tag{2.29}$$

where ρ and θ are two constants that control the slope and the position of the function, respectively.

The third practical issue in the implementation of the algorithm is that the background motion v_g, the relative motion v_r, and the local contrast c may be close to zero. This could result in unstable evaluation of the weight function. To avoid this, and to take into account the Weber-Fechner law, we take a similar approach as in the Stocker and Simoncelli paper [3]. That is, instead of computing $\ln v_r$, $\ln v_b$, and $\ln c$, we replace them with $\ln(1+v_r/v_0)$, $\ln(1+v_b/v_0)$, and $\ln(1+c/c_0)$, respectively, where v_0 and c_0 are both small positive constants. Furthermore, to avoid the situation that the the weight might go negative, we threshold it at 0. Therefore, the final importance weight function we are computing is given by

$$w = \max\left\{0, \left[\alpha \ln\left(1 + \frac{v_r}{v_0}\right) + \beta\right] - \left[\ln\left(1 + \frac{v_g}{v_0}\right) - \gamma \ln\left(1 + \frac{c}{c_0}\right) + \delta\right]\right\}. \tag{2.30}$$

Since the motion vectors are in the unit of pixels/frame, the parameter v_0 also needs to be in the same unit. In our implementation, we assume a 32 pixels/degree of viewing distance, and as in Stocker and Simoncelli paper[3], we fix $v_0 = 0.3$ degree/sec. We can then convert v_0 based on the frame rate of the video sequence. For example, if the frame rate is 30 frames/sec, then $v_0 = 0.3 \times 32/30 = 0.32$ pixle/frame. If the frame rate is 25 frames/sec, then $v_0 = 0.3 \times 32/25 = 0.384$ pixle/frame. The other parameters are hand-picked and we find that the following parameters give reasonable results and use them in all the experiments reported later in this paper: $\alpha = 0.2$, $\beta = 0.09$, $\gamma = 2.5$, $\delta = 2.25$, $\mu_0 = 6$, $\theta = 0.05$, $\rho = 2$, and $c_0 = 0.07$.

2.3.5 Test

To validate the proposed model with real VQA algorithms, we incorporate the proposed weighting method with two types of image distortion/quality maps. The first is the squared error map defined as

$$q(x, y, t) = |I_r(x, y, t) - I_d(x, y, t)|^2 , \tag{2.31}$$

where $I_r(x, y, t)$ and $I_d(x, y, t)$ are the pixel intensity values at spatial location (x, y) and time t in the original video sequence (as a perfect-quality reference) and the distorted video sequence (quality to be evaluated), respectively. The PSNR/MSE is calculated from error map $q(x, y, t)$. With the proposed weighting approach based on perceptual motion information being taken into account, a weighted MSE measure can be computed using Eq. (2.27). This can then be further converted to a weighted PSNR.

The second type of image quality map is created using the SSIM. Again, the standard SSIM measure is a simple average of the SSIM map over all space and time, and a weighted SSIM measure can be computed by incorporating SSIM into Eq. (2.27).

The proposed method is tested using the VQEG Phase I database [36]. We use the Spearman rank order correlation coefficient between the subjective and objective scores to evaluate the performance of the VQA algorithms. Table 2.7 shows the SROCC test results of three datasets − the 50Hz dataset, the 60Hz dataset, and all data combined. The results suggest that the proposed weighting method is quite effective. It gives clear and consistent improvement to all test datasets with two completely different types of image distortion/quality maps. Similar results are also obtained with all the other VQEG test metrics [36]. Figs. 4.1.3 (a), (b), (c), (d) show the scatter plots of the subjective/objective comparisons on all VQEG test video

Table 2.7. SROCC results of VQA algorithms. PSNR(w [6]): with the spatial information content weighting as in [6]. PSNR(w): PSNR with proposed weighting; SSIM(w [6]): SSIM with the spatial information content weighting as in [6]. SSIM(w): SSIM with proposed weighting.

Dataset	MSE/PSNR	PSNR(w [6])	PSNR(w)	SSIM	SSIM(w[6])	SSIM(w)
50Hz	0.8152	0.8211	0.8278	0.8301	0.8544	0.8948
60Hz	0.7112	0.7120	0.7303	0.7680	0.7692	0.7985
All	0.7818	0.7887	0.8048	0.8127	0.8287	0.8621

sequences for PSNR, PSNR with proposed weighting, SSIM, and SSIM with proposed weighting, respectively. These scatter plots confirm the SROCC results shown in Table 2.7. It can be clearly seen that after applying the proposed weighting method, the clusters of sample points (each associated with a video sequence) become much tighter, which implies better consistency between subjective and objective quality evaluations.

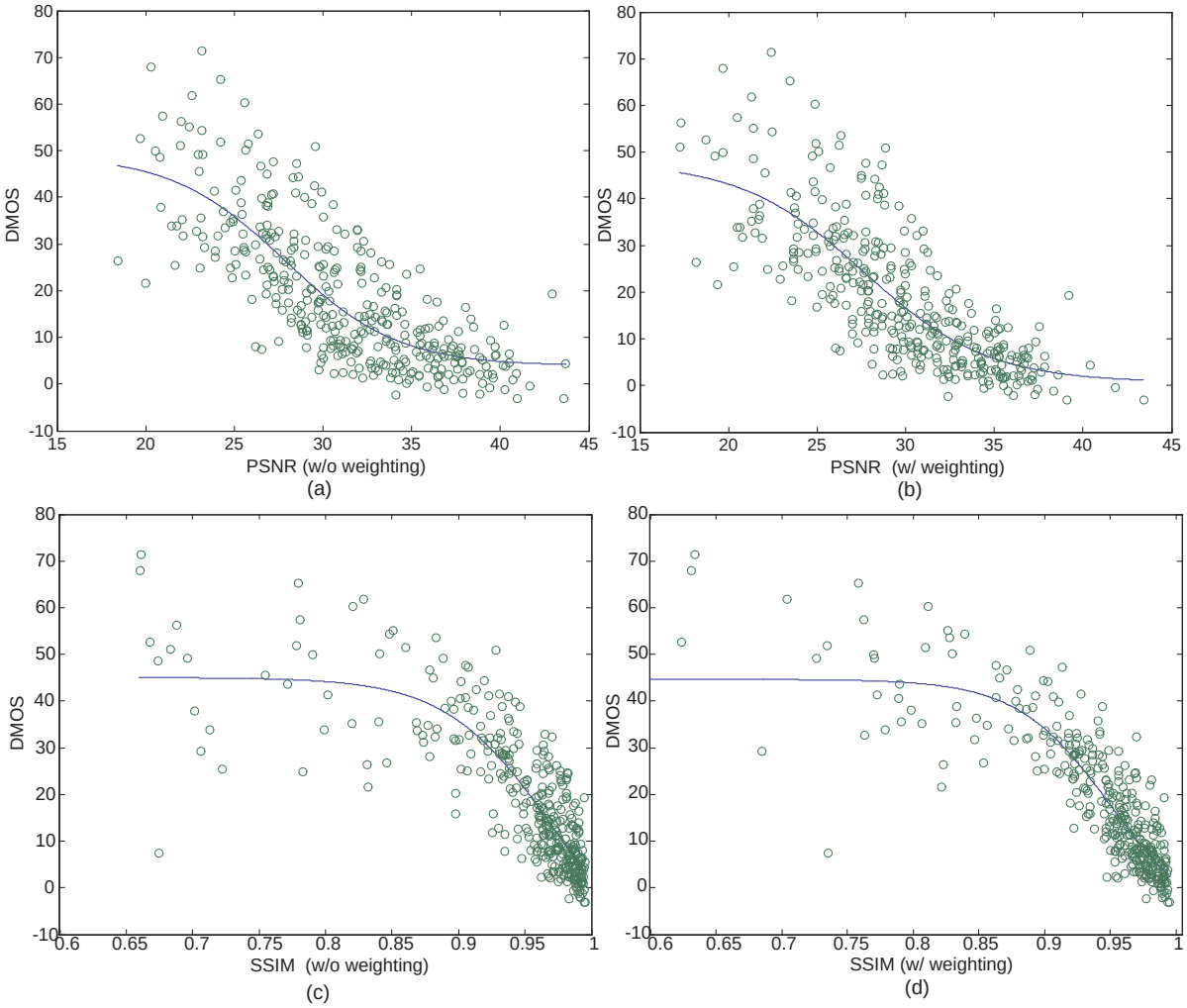

Figure 2.13. Scatter plots of subjective/objective scores on VQEG Phase I test database (all video sequences included). The vertical and horizontal axes represent the subjective and the objective scores, respectively. Each sample point represents one test video sequence. (a) PSNR; (b) PSNR with proposed weighting method; (c) SSIM; (D) SSIM with proposed weighting method. All SSIM values were raised to the 8th power for better visualization.

CHAPTER 3

RRIQA/VQA BASED ON NATURAL SCENE STATISTICS (NSS)

A general RRIQA method based on NSS introduced in [51] achieved notable success. However, our further investigation has revealed some important limitations.

First, although the method performed quite well when tested with individual distortion types (e.g., JPEG[85] or JPEG 2000 [86] compression, blurring, or noise contamination), its performance degrades significantly when images with different types of distortions are tested together.

Second, it uses a rather weak model of natural image statistics, as only marginal distributions of wavelet coefficients are considered. It has been widely noticed that there exist strong dependencies between neighboring wavelet coefficients, which has been completely ignored by this method. In order to show the limitation of the RRIQA method based on only marginal statistics, a counter example is given in Fig. 3.1: 1) The original image (a) is decomposed into any wavelet domain with perfect reconstruction [1] of 4 scales as (c) using the Haar wavelet transform; 2) We randomly permute the coefficients in the H4, V3 and D2 subbands as shown in the wavelet domain,(c) to generate distorted coefficients; 3) Distorted image (b) is reconstructed using this set of distorted wavelet coefficients; 4) (b) is then decomposed into the Haar wavelet domain as (d). Because of the perfect reconstruction, (d) has the same coefficients as the distorted wavelet coefficients in step 2). So by this means, distorted image Fig. 3.1(b) shares the same marginal statistics as the original image.

[1] Wavelet with perfect reconstruction is used to generate more prominent counterexample although similar effect is also observed with wavelet with nearly perfect reconstruction [15]

Fig. 3.2(a-c) shows the negligible difference between the marginal distributions of the subbands of H4,V3,and D2 in the original and the distorted images. Therefore, the RRIQA method based on marginal statistics can not predict this type of distortion, which does not satisfy the requirement that RR features are sensitive to general image distortions.

Third, it also uses a rather weak model for perceptual image representation, as wavelet decomposition is linear and cannot reflect the nonlinear mechanisms used by the biological visual systems.

In the following sections, we propose two general RRIQA methods based NSS. One is using the joint statistics of the natural images by considering the dependencies in the image, the other is employing a statistically and perceptually motivated image representation.

3.1 RRIQA Based on Joint Statistics of Natural Image

The joint statistics of natural images contains more information and displays the dependencies between the subbands, which makes it a more precise model to describe the natural images. For example, given the distorted image in Fig.3.1 (b), it can be observed that the joint statistics is dramatically corrupted as shown in Fig. 3.2 (d)-(i). The KLD between the joint distributions in Fig 3.2(d) and 3.2(g) is 0.1235 that is more significant than the KLD between marginal distributions. This means that the joint statistics can capture the distortion in Fig.3.1(b), where the marginal statistics fails. Hence, we propose a RRIQA method based on the joint statistics of the natural image.

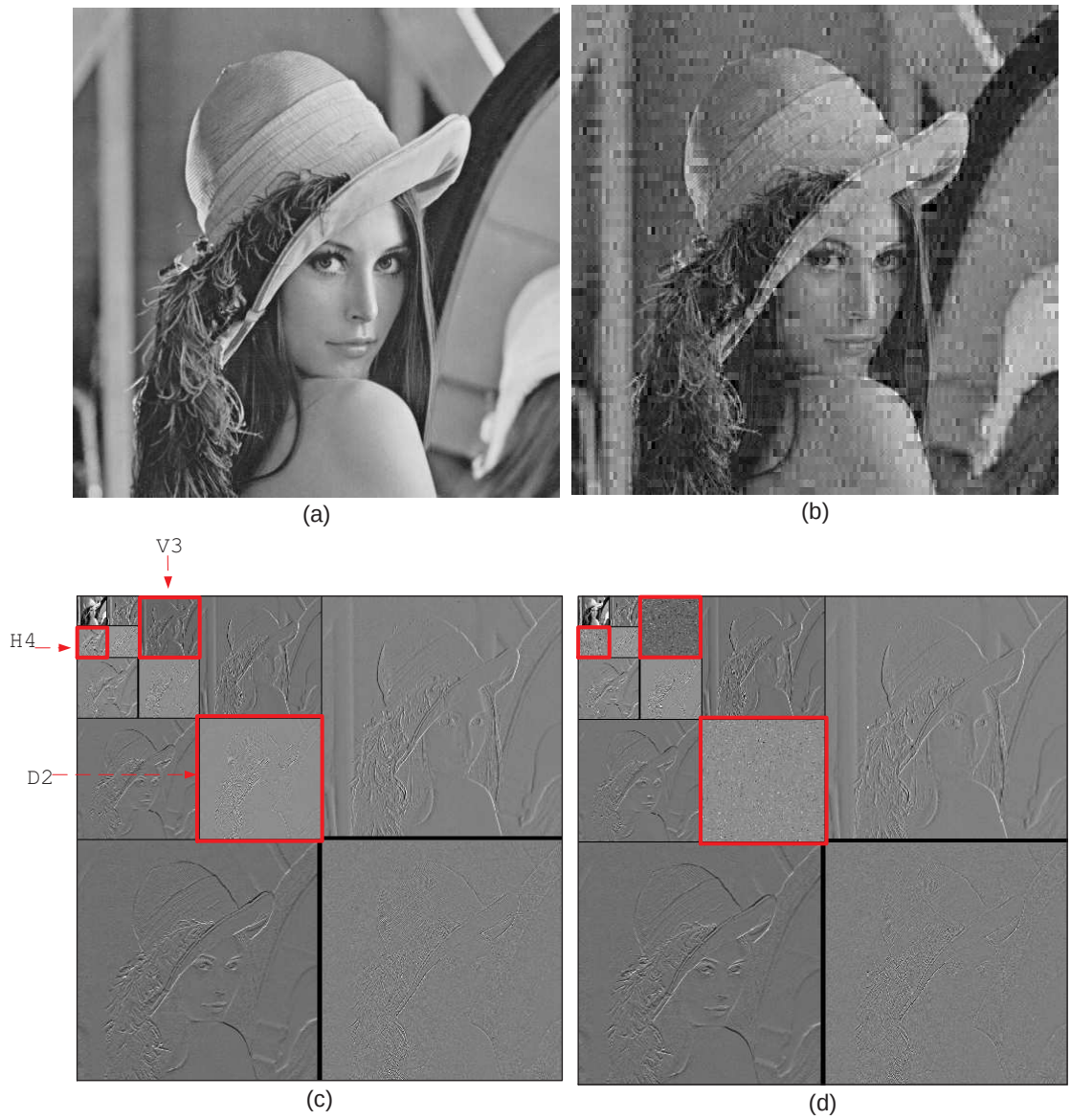

Figure 3.1. A counterexample for marginal distribution to measure distortion (a) Original image; (b) Distorted image as counterexample; (c) Haar wavelet transform of (a); (d) Haar wavelet transform of (b).

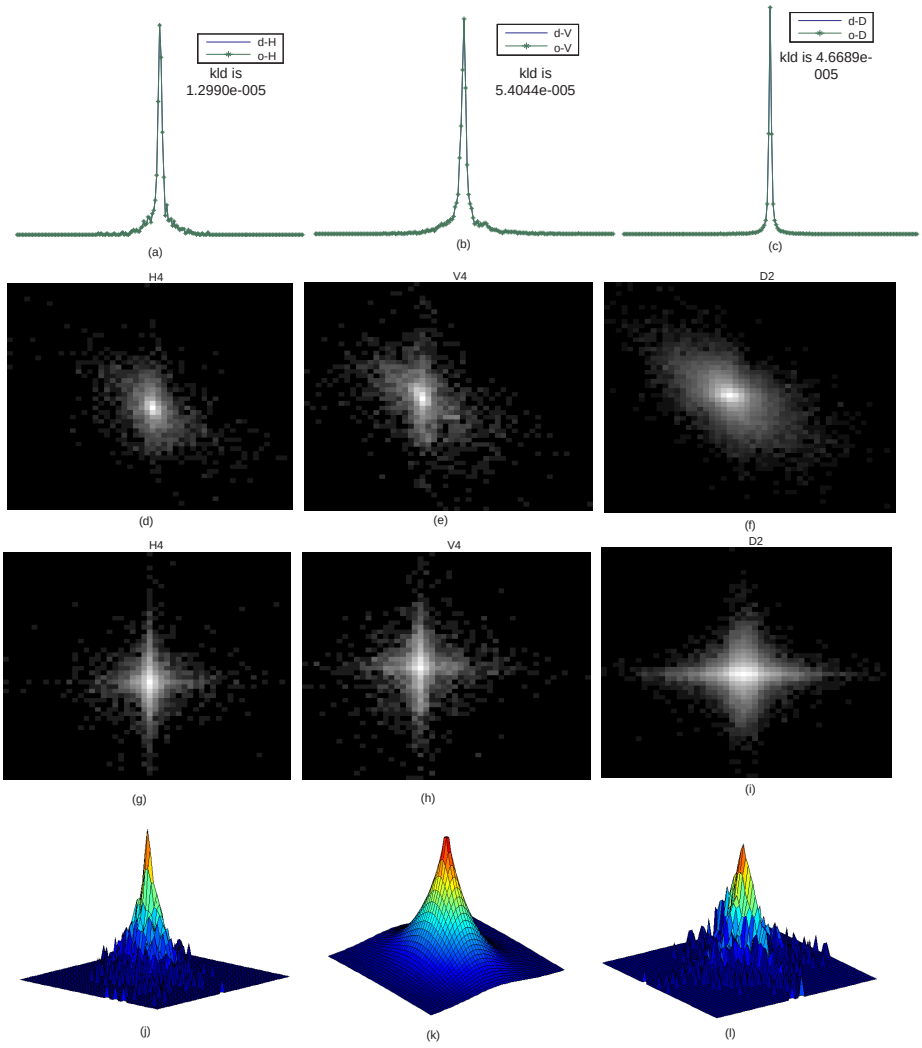

Figure 3.2. (a-c) marginal distributions between Fig3.1(c) and Fig3.1(d) in H4,V3 and D2 subbands. KLDs are 1.2990e-005, 5.4044e-005 and 4.6689e-005. (d) joint distribution of Fig3.1(c) between H4-H3 . (e) joint distribution of Fig3.1(c) between V4-V3. (f) joint distribution of Fig3.1(c) between D2-D1.(g) joint distribution of Fig3.1(d) between H4-H3 . (h) joint distribution of Fig3.1(d) between V4-V3. (i) joint distribution of Fig3.1(d) between D2-D1. (j) surf plot of joint distribution of Fig3.1(c) between H4-H3.(k) fitting model of (j). (l) surf plot of joint distribution of Fig3.1(d) between H4-H3.

3.1.1 Joint Statistical Models and Image Distortion Measurement

First, we need an explicit statistical model to fit the joint statistics of the natural image. In [52], a 2D joint statistical model between the child and parent subbands is presented as a form of "generalized Laplacian" distribution.

$$p_m(x_c, x_p) = \frac{1}{Z} e^{-\alpha r^{\beta}} \tag{3.1}$$

where x_p, x_c are wavelet coefficients in the parent and the child subbands. $r = \sqrt{{x_c}^2 + {x_p}^2}$ and α, β are the constant parameters and Z is a normalization factor. It is found that this model fits fairly well to the image for $\alpha = 1$ and $\beta = 0.5$. Given p_m, the KLD between p_m and the original joint statistics, p is

$$kld(p_m||p) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_m(x_c, x_p) \ln \frac{p_m(x_c, x_p)}{p(x_c, x_p)} dx_c dx_p \tag{3.2}$$

Thus, the KullbackCLeibler divergence[44] (KLD) between the original distribution, p, and distorted distribution, q is approximated as

$$kld(p||q) \approx \hat{kld}(p||q) = kld(p_m||q) - kld(p_m||p) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_m(x_c, x_p) \ln \frac{p_m(x_c, x_p)}{q(x_c, x_p)} dx_c dx_p \tag{3.3}$$

Given the distorted image, $kld(p_m||q)$ can be easily computed. Finally, the overall objective score to predict the distortions is defined as

$$D = \sum_{N} \hat{kld}(p||q) \tag{3.4}$$

where N is the number of joint distributions.

3.1.2 Implementation and Test

We decompose the original image into Haar wavelet domain with 4 scales. Six joint statistical distributions between the parent subbands including H4, V4, D4,

Table 3.1. RR features

	α	β	w_c	w_p	$kld(p_m\|\|p)$
H4-H3	1.0081	0.5411	1.9631	1.3401	7.0298e-008
V4-V3	1.0408	0.44732	2.6133	1.9459	8.1815e-009
D4-D3	1.06	0.54963	1.255	0.91941	1.2113e-008
H2-H1	1.0239	0.50361	1.6267	1.5232	2.8705e-007
V2-V1	1.0606	0.4341	2.4105	2.2525	6.4667e-010
D2-D1	1.0534	0.57656	0.80596	0.96338	3.0457e-008

H2, V2 and D2 and the child subbands including H3, V3, D3, H1, V1 and D1 are computed. In this way, all the subbands are taken into account. In order to estimate the parameters, α, β, a nonlinear optimization method is used to minimize the KLD between the original and the fitting distributions. For each joint statistics, we extract the RR features including α, β of the fitting model, the width of bin for parent and child subbands, w_p, w_c. The total number of RR features is 24. Table 3.1 shows the parameters of the model for Fig3.1 (a). The very small $kld(p_m||p)$ indicates that the fitting model can precisely describe the original joint distribution.

The test is carried out on the data set at [1] of different distortions with the same MSE but quite different perceptual quality including the distorted image of Fig. 3.1(b). Fig 3.3 shows that our method is highly consistent with human perception for image quality assessment and has comparable performance with the Universal Image Quality Index(UIQI) [87].

3.2 RRIQA based on the Statistics of Divisive Normalization Transform

We propose a new RRIQA method that is inspired by the recent success of the divisive normalization transform (DNT) as a perceptually and statistically motivated image representation [53, 88]. In computational vision science, it has long been hypothesized that the purpose of early visual sensory processing is to increase

the statistical independence between neuronal responses [89, 90]. However, linear decompositions, such as Fourier- and wavelet-types of transformations, only reduce the first-order correlation, but cannot reduce the higher order statistical dependencies [91]. In the literature of neural physiology, it has been shown that a local gain-control divisive normalization model is powerful in accounting for the neuronal responses in biological visual systems [92, 93]. This nonlinear gain-control mechanism is built upon linear transform models, where each neuronal response (or linear transform coefficient) is normalized (divided) by the energy of a cluster of neighboring neuronal responses (neighboring coefficients). This process has been shown to significantly reduce the statistical dependencies of the original linear representation [91] and produce approximately Gaussian marginal distributions [94]. Similar models have also been employed in real world image processing applications, including image compression [95] and image enhancement [88]. The strong perceptual and statistical relevance of divisive normalization representation (as compared to linear decompositions) motivated us to switch from the linear wavelet transform domain (as in [51]) to DNT domain in the design of our RRIQA method.

3.2.1 Computation of Divisive Normalization Transformation

A divisive normalization transform (DNT) is built upon a linear image decomposition, followed by a divisive normalization stage. The linear transformations may be discrete cosine transform (DCT) (as in [95]) or wavelet-type of transforms (as in [53, 91, 88]). Here, we assume a wavelet image decomposition, which provides a convenient framework for localized representation of images simultaneously in space, frequency (scale) and orientation. Let y represent a wavelet coefficient, then a normalized coefficient is computed as $\tilde{y} = y/p$, where p is a positive divisive normalization

factor that is calculated as the energy of a cluster of coefficients that are close to the coefficient y in space, scale and orientation.

Several different approaches have been used to compute the normalization factor p [53, 91, 95, 88]. Most of them use a weighted sum of the squared neighboring coefficients plus a positive constant [91, 95, 88]. This involves several parameters (the weights and the constant) that are sometimes difficult to determine. They may be hand-picked (as in [95]) or chosen to maximize the independence of the normalized response to an ensemble of natural images [91]. In [88], a global model of Markov random field over the wavelet coefficients is assumed and the parameters were derived by learning the model parameters using natural images. A more convenient approach is to derive the factor p through a local statistical image model. In particular, the Gaussian scale mixtures (GSM) model has found to be very useful in this context [53]. A length-N random vector Y is a GSM if it can be expressed as the product of two independent components: $Y \doteq zU$, where $\doteq$ denotes equality in probability distribution, U is a zero-mean Gaussian random vector with covariance C_U, and z is a scalar random variable called a mixing multiplier. In other words, the GSM model expresses the density of a random vector as a mixture of Gaussians with the same covariance structure (C_U) but scaled differently (by z). Suppose that the mixing density is $p_z(z)$, then the probability density of Y can be written as

$$p_Y(Y) = \int \frac{1}{[2\pi]^{\frac{N}{2}} |z^2 C_U|^{1/2}} \exp\left(-\frac{Y^T C_U^{-1} Y}{2z^2}\right) p_z(z) dz \,. \tag{3.5}$$

This GSM model has shown to be very useful to account for both the marginal and joint statistics of the wavelet coefficients of natural images [53], where the vector Y is formed by clustering a set of neighboring wavelet coefficients within a subband, or across neighboring subbands in scale and orientation. The GSM model has also

found successful applications such as image desnoing [14], image restoration [96] and image quality assessment [7].

The general form of the GSM model allows for the mixing multiplier z to be a continuous random variable at each location of the wavelet subbands. To simplify the model, we assume that z only takes a fixed value at each location (but varies over space and subbands). The benefit of this simplification is that when z is fixed, Y is simply a zero-mean Gaussian vector with covariance $z^2 C_U$. As a result, it becomes natural to define the normalization factor p in the DNT representation as an estimate of the multiplier z from the neighboring coefficient vector Y. The coefficient cluster Y moves step by step as a sliding window across a wavelet subband, resulting in a spatially varying normalization factor p. In our implementation, the normalization factor computed at each step is only applied to the center coefficient y_c of the vector Y, and the normalized new coefficient becomes $\tilde{y}_c = y_c/\hat{z}$, where $\hat{z}$ is the estimate of z. A convenient method to obtain $\hat{z}$ is by a maximum likelihood estimation [53] given by

$$\begin{aligned} \hat{z} &= \arg\max_z \{\ln p(Y|z)\} \\ &= \arg\min_z \{N \ln z + Y^T C_U^{-1} Y / 2z^2\} \\ &= \sqrt{Y^T C_U^{-1} Y / N}\,, \end{aligned} \tag{3.6}$$

where the covariance matrix $C_U = E[UU^T]$ is estimated from the entire wavelet subband before the estimation of local z, and N is the length of vector Y, or the number of neighboring wavelet coefficients.

3.2.2 Image Statistics in Divisive Normalization Transform Domain

As will be shown in the next section, our RRIQA approach is essentially based on the statistics of the transform coefficients in DNT domain and how they vary

with image distortions. Before the development of the specific RRIQA algorithm, it is useful to observe variations of image statistics before and after the DNT is applied. In Fig. 3.4, we compare the marginal distributions of an original wavelet subband computed from a steerable pyramid decomposition [72] (Fig. 3.4(a)) and the same subband after DNT (Fig. 3.4(b)). In Fig. 3.4(c), the original wavelet coefficient histogram is compared with a Gaussian shape that has the same standard deviation. The significant gap between the two curves indicates that the original wavelet coefficients are highly non-Gaussian. It has been shown that such histograms can be well-fitted with a generalized Gaussian density (GGD) function given by [97]

$$p_{GGD}(x) = \frac{\lambda}{2\mu\Gamma(1/\lambda)} e^{-(|x|/\mu)^{\lambda}}, \tag{3.7}$$

where $\Gamma(a) = \int_0^{\infty} t^{a-1}e^{-t}dt$ (for $a > 0$) is the Gamma function, and λ and μ are called the scale and power factors, respectively. The Gaussian density is a special case of GGD when λ is fixed to be 2. However, for the histograms of the wavelet coefficients of natural images, the best fitting value of λ typically lies between 0.5 and 1.0 [98]. By contrast, the histogram of the coefficients after DNT can be well-fitted with a Gaussian, as demonstrated in Fig. 3.4(d). Similar behavior is observed for other natural images. To provide a quantitative measure, we compute the KLD [44] between the histogram and the best-fitting Gaussian curve before and after DNT for a set of natural images. The results are shown in Table 3.2, where we can see that Gaussian fit is consistently better in DNT domain for all test images.

Figure 3.5 and Figure 3.6 demonstrate the impact of DNT on the joint statistics of wavelet coefficients. In Figs. 3.5 and Figs. 3.6, we show the conditional histograms of the coefficients extracted from two neighboring subbands (a parent band and a child band) in the original wavelet decomposition and in the DNT representation, respectively. It can be observed that in the conditional histogram ($histo(C|P)$ in

Table 3.2. KLD between the marginal distributions of wavelet/DNT coefficients and Gaussian fit

Image	Wavelet domain	DNT domain
Lena	0.4143	0.0009
Barbara	0.4301	0.0125
Barco	0.5226	0.0058
Boat	0.3848	0.0098
House	0.4084	0.0106
Peppers	0.4722	0.0082
Fingerprint	0.0123	0.0029
Flintstones	0.2436	0.0034

Fig. 3.5), the variance of a child coefficient (vertical axis) is highly dependent on the magnitude of its parent coefficient (horizontal axis). Such strong second-order variance dependency is confirmed by the significant difference between the widths of two cross-sections of the conditional histogram. By contrast, in the DNT representation, the histogram of the child coefficients makes little difference when conditioned on the magnitudes of the parent coefficients, as can be seen in Fig. 3.6. This demonstration clearly shows that the DNT representations can significantly reduce the second-order dependencies between the transform coefficients.

3.2.3 Perceptual Relevance of Divisive Normalization Representation

The DNT image representation is not only an effective way to reduce the statistical redundancies between wavelet coefficients, it is also highly relevant to biological vision. First, based on the widely accepted hypothesis that the early visual sensory processing is optimized to increase the statistical independence between neuronal responses (subject to certain physical limitations such as power consumption) through the evolution and development processes, the modeling of the biological visual system and the modeling of natural scene statistics are dual problems [89, 91, 90]. Second,

in the context of neural physiology, it has been found that divisive normalization provides an effective model to account for many recorded data of cell responses in the visual cortex [92, 93]. It is also a useful framework in explaining the adaptations of neural responses with respect to the variations of the visual environment [99]. Third, in psychophysical vision, it has been shown that the divisive normalization procedure can well explain the visual masking effect [100, 101], where the visibility of an image component (e.g., a wavelet coefficient) is reduced in the presence of large neighboring components (e.g., the wavelet coefficients close in space, scale and orientation). Furthermore, the perceptual relevance of DNT image representation has also been demonstrated by testing its resilience to noise contamination as well as its effectiveness in image compression and image contrast enhancement [88].

3.2.4 DNT-Domain Statistics of Distorted Images

The strong perceptual and statistical relevance of DNT image representation provides good justifications for the use of DNT for RRIQA. In addition to that, we must also show that the statistics of DNT coefficients are sensitive to various image distortions. To study this, we apply DNT to a set of images with different types of distortions and observe how these distortions alter the statistics of the coefficients in DNT domain. This is demonstrated in Fig. 3.7, where the histogram of the DNT coefficients of a wavelet subband can be well-fitted with a Gaussian model (Fig. 3.7(a)). However, when we draw the same Gaussian model together with the histogram of the DNT coefficients computed from Gaussian noise contaminated image (Fig. 3.7(b)), Gaussian blurred image (Fig. 3.7(c)), or JPEG compressed image (Fig. 3.7(d)), significant changes are observed. It is also interesting to see that the way the distribution changes varies with the distortion type. For example, Gaussian noise contamination increases the width of the histogram, but maintains the

shape of Gaussian. By contrast, Gaussian blur reduces the width of the histogram and creates a much peakier distribution than Gaussian. These observations are important because our RRIQA algorithm is based on quantifying the variations of DNT-domain image statistics as a measure of image quality degradation.

3.2.5 Reduced-Reference Image Quality Assessment Algorithm

We propose an RRIQA algorithm by working with the marginal distributions of DNT coefficients. Although this algorithm still works with marginal distributions only (no explicit joint statistical model is employed, as in [51]), it does take into account the dependencies between the original neighboring wavelet coefficients because of the involvement of the divisive normalization process. We consider this as a major advantage of the proposed approach (as compared to [51]) in capturing the joint statistics of wavelet coefficients while maintaining the simplicity of the algorithm. Moreover, the algorithm has a low data rate, as only a small set of RR features are extracted from the reference image and are employed in quality evaluation of the distorted image.

A convenient approach to measure the variations of the marginal probability distributions of the DNT coefficients between the original and distorted images (as being observed in Fig. 3.7) is to compute the KLD between them:

$$d(p\|q) = \int p(x) \ln \frac{p(x)}{q(x)} dx \,, \tag{3.8}$$

where $p(x)$ and $q(x)$ are the probability density functions of the DNT coefficients in the same subband of the original and distorted images, respectively. To accomplish this, the DNT coefficient histograms of both the reference and distorted images must be available. The latter can be easily computed from the distorted image, which is always available. The difficulty is in obtaining the DNT coefficient histogram of the

original image. Using all the histogram bins as RR features would result in either a heavy RR data rate (when the bin size is fine) or a poor approximation accuracy (when the bin size is coarse). To overcome this problem, we make use of the important property that the probability density function $p(x)$ of the original DNT coefficients can be well approximated with a zero-mean Gaussian model (as has been observed in Figs. 3.4(d) and 3.7(a)):

$$p_m(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{3.9}$$

This model provides a very efficient means to summarize the DNT coefficient histogram of the original image, such that only one parameter σ is needed to describe it (as opposed to all the histogram bins). Furthermore, to account for the variations between the model and the true distribution, we compute the KLD between $p_m(x)$ and $p(x)$ as

$$d(p_m||p) = \int p_m(x) \ln \frac{p_m(x)}{p(x)} dx \tag{3.10}$$

and use it as an additional RR feature. This is computed for each subband independently, resulting in 2 parameters (σ and $d(p_m||p)$) for each subband.

In order to evaluate the quality of a distorted image, we estimate the KLD between the probability density function $q(x)$ of the DNT coefficients computed from the distorted image and the model $p_m(x)$ estimated from the original image:

$$d(p_m||q) = \int p_m(x) \ln \frac{p_m(x)}{q(x)} dx\,. \tag{3.11}$$

Combining this with the available RR feature $d(p_m||p)$, we obtain an estimate of the KLD between $p(x)$ and $q(x)$:

$$\hat{d}(p||q) = d(p_m||q) - d(p_m||p)\,. \tag{3.12}$$

It can be easily shown that

$$\hat{d}(p\|q) = \int p_m(x) \ln \frac{p_(x)}{q(x)} dx\,. \tag{3.13}$$

The estimation error can then be calculated as

$$d(p\|q) - \hat{d}(p\|q) = \int [\, p(x) - p_m(x) \,] \ln \frac{p_(x)}{q(x)} dx \,. \tag{3.14}$$

This error is small when $p_m(x)$ and $p(x)$ are close, which is true for typical natural images. With the additional cost of adding one more RR parameter $d(p_m||p)$, Eq. (3.13) not only delivers a more accurate estimate of $d(p\|q)$ than Eq. (3.36), but also provides a useful feature that when there is no distortion between the original and distorted images (which implies that $p(x) = q(x)$ for all x), then both the targeted distortion measure $d(p\|q)$ and estimated distortion measure $\hat{d}(p\|q)$ are exactly zero.

In addition to $\hat{d}(p\|q)$, we also found the following measures useful in improving the accuracy of image quality evaluation:

$$d_\sigma = |\sigma_o - \sigma_d| \,, \tag{3.15}$$

$$d_k = |k_o - k_d| \,, \tag{3.16}$$

$$d_s = |s_o - s_d| \,, \tag{3.17}$$

where σ_o, k_o, s_o, and σ_d, k_d, s_d are the standard deviation, the kurtosis (the fourth-order central moment divided by the fourth power of the standard deviation and then minus 3) and the skewness (the third-order central moment divided by the third power of the standard deviation) of the DNT coefficients computed from the original and distorted images, respectively. These measures provide further information about the shape changes of the probability density functions. In particular, two images with the same KLD with respect to the original image may have different types of distortions, and visual quality assessment varies across distortion types. Adding these features not only provides new means to quantify the amount of distortions, but also supplies new information that helps the algorithm differentiate distortion types. We have also carried out experiments to compare our IQA algorithm with and without these

features, and we found that adding these features lead to significant improvement in terms of the performance of image quality prediction. Since σ_d, k_d, s_d can be computed from the available distorted image and σ is already acquired when fitting the Gaussian model of Eq. (3.9), only two new RR features, k_o and s_o, are added. Indeed, both of them are close to zero because the probability distribution of DNT coefficients of the original image is approximately Gaussian, which has zero skewness and kurtosis.

At each subband, we define the overall image distortion measure as a linear combination of $\hat{d}(p||q)$, d_σ, d_k and d_s in the logarithmic domain:

$$D_{band} = \alpha \ln(\hat{d}(p||q)) + \beta \ln(d_\sigma) + \gamma \ln d_k + \delta \ln d_s = \ln\left((\hat{d}(p||q))^\alpha (d_\sigma)^\beta (d_k)^\gamma (d_s)^\delta\right), \tag{3.18}$$

where α, β, γ and δ are weighting parameters. Finally, the overall distortion of the distorted image is computed as the sum of the distortion measures of all subbands:

$$D = \sum_{all\ subbands} D_{band}\,. \tag{3.19}$$

3.2.5.1 Implementation Issues

To compute the DNT representation of an image, we first decompose the image using a steerable pyramid [72] with three scales and four orientations, as shown in Fig. 3.8. For each center coefficient y_c at each subband, we define a DNT neighboring vector Y that contains 13 coefficients, including 9 from the same subband (including the center coefficient itself), 1 from the parent band, and 3 from the same spatial location in the other orientation bands at the same scale. An illustration is given in Fig. 3.8. These coefficients are selected from the direct neighbors of the center coefficient because the magnitudes of clusters of wavelet coefficients tend to scale together [90] and thus are more likely to share the same scale factor z in the GSM model

described earlier. Increasing the size of the neighborhood will increase the computational complexity of DNT calculation (specifically, the estimation of $\hat{z}$ in Eq. (3.6)), but will not add extra RR features (because it only affects the DNT computation and all other processes after DNT remain unaltered). In our experiments, we did not observe significant variations of the overall performance of the algorithm under slight changes of the neighborhood, but more careful study on this issue remains for future work. After the DNT computation, four RR features are extracted from each subband of the original image, including σ, $d(p_m||p)$, k_o and s_o. This results in a total of 48 scalar RR features for each original image.

The evaluation of the KLD between probability density functions needs to be done numerically using histograms. For example, for Eq. (3.10), we compute:

$$d(p_m\|p) = \sum_{i=1}^{L} P_m(i) \ln \frac{P_m(i)}{P(i)}\,, \tag{3.20}$$

where $P(i)$ and $P_m(i)$ are the normalized heights of the i-th histogram bins, and L is the number of bins in the histograms.

One problem with the subband quality measure of Eq. (3.18) is that when $\hat{d}(p\|q)$, d_σ, d_k or d_s is close to zero, the measure becomes unstable. In our implementation, to avoid such instability, we compute

$$D_{band} = \ln\left(1 + \frac{(\hat{d}(p||q))^\alpha (d_\sigma)^\beta (d_k)^\gamma (d_s)^\delta}{D_0}\right)\,, \tag{3.21}$$

where D_0 is a positive constant. Another useful property of this formulation is that the resulting distortion measure is always non-negative, and is zero when the original and distorted images are exactly the same.

Before applying the proposed algorithm for image quality assessment, five parameters, α, β, γ, δ and D_0, need to be learned from the data. It is important to cross-validate these parameters with different selections of the training and testing

data. Details will be given in the next section. For a given set of training images and the associated subjective scores, we use the Matlab nonlinear optimization routine *fminsearch* in the optimization toolbox to find the optimal parameters.

3.2.6 Validation

To validate the proposed RRIQA algorithm, two publicly-accessible subject-rated image databases are used, which are the LIVE database [17] developed at Laboratory for Image and Video Engineering at The University of Texas at Austin and the Cornell-VCL A57 database [65] developed at the Visual Communications Laboratory at Cornell University. The LIVE database contains seven datasets of 982 subject-rated images created from 29 original images with five types of distortions at different distortion levels. The distortion types include 1) JP2: JPEG2000 compression (2 sets); 2) JPG: JPEG compression (2 sets); 3) Noise: white noise contamination; 4) Blur: Gaussian blur; and 5) FF: fast fading channel distortion of JPEG2000 compressed bitstream. The subjective test was carried out with each of the seven data sets individually. A cross-comparison set that mixes images from all distortion types is then used to help align the subject scores across different data sets. The subjective scores of all images are then adjusted according to this alignment. The alignment process is rather crude. However, the aligned subjective scores (all data) are still very useful references, which are particularly important for testing general-purpose IQA algorithms, for which cross-distortion comparisons are highly desirable. In the Cornell-VCL database, there are totally 60 distorted images generated from 3 original images. Six different types of distortions are included, which are 1) FLT: quantization of the LH subbands of a 5-level DWT of the image using the 9/7 filters, where the bands were quantized via uniform scalar quantization with step sizes chosen such that the RMS contrast of the distortions was equal; 2) NOZ: additive Gaussian

white noise; 3) JPG: baseline JPEG compression; 4) JP2: JPEG2000 compression using the 9/7 wavelet [86] and no visual frequency weighting; 5) DCQ: JPEG2000 compression using the 9/7 wavelet [86] with the dynamic contrast-based quantization algorithm, which applies greater quantization to the fine spatial scales relative to the coarse scales in an attempt to preserve global precedence; and 6) BLT: blurring by using a Gaussian filter.

Three criteria are used to evaluate how well the objective scores predict the subjective scores: 1) LPCC; 2) SROCC; 3) Outlier ratio is used to evaluate prediction consistency, which is defined as the percentage of predictions outside the range of ± 2 standard deviations between subjective scores. These criteria were used in the previous tests conducted by the VQEG [102]. Since we do not have access to the raw subjective scores of the Cornell-VCL database, the standard deviations between subjective scores for each test image cannot be computed. Therefore, only LPCC and SROCC comparisons are included for the Cornell-VCL database.

Our validation work has two major purposes. The first is to verify that using DNT image representation is beneficiary for the improvement of IQA algorithms. The second is to compare the performance of the proposed method with existing IQA algorithms.

To show the impact of DNT representation, we compare the performance of the proposed RRIQA algorithm implemented in the wavelet domain (linear steerable pyramid decomposition) and in the DNT domain (linear steerable pyramid decomposition, followed by the nonlinear DNT process). Specifically, GGD is used to model the marginal distribution of wavelet coefficients and Gaussian density is employed to model that of DNT coefficients. All other aspects of the algorithm, including the standard deviation, skewness and kurtosis features, the KLD measure, the subband and overall quality measurement approach, and the training data and process, are exactly

Table 3.3. Wavelet and DNT domain comparison of the proposed methods using the LIVE database

LIVE data set		JP2(1)	JP2(2)	JPG(1)	JPG(1)	Noise	Blur	FF	All data
		LPCC (prediction accuracy)							
Proposed	wavelet + GGD	0.9115	0.9422	0.8501	0.9354	0.9401	0.8773	0.9243	0.8930
Proposed	DNT + Gaussian	0.9485	0.9655	0.8203	0.9579	0.9654	0.9562	0.9464	0.9173
		SROCC (prediction monotonicity)							
Proposed	wavelet + GGD	0.9081	0.9239	0.8389	0.8734	0.9316	0.8608	0.9237	0.9093
Proposed	DNT + Gaussian	0.9478	0.9610	0.8143	0.8937	0.9559	0.9584	0.9443	0.9287
		Outlier Ratio (prediction consistency)							
Proposed	wavelet + GGD	0.0230	0.0122	0.0805	0.1250	0.0345	0.0483	0.0345	0.1853
Proposed	DNT + Gaussian	0.0115	0.0122	0.1149	0.0341	0.0000	0.0000	0.0207	0.1069

Table 3.4. Wavelet and DNT domain comparison of the proposed methods using the Cornell-VCL database

Cornell-VCL data set		FLT	JPG	JPG2	DCQ	BLR	NOZ	All data
		LPCC (prediction accuracy)						
Proposed	wavelet + GGD	0.4592	0.8303	0.7802	0.8808	0.9270	0.7748	0.5125
Proposed	DNT + Gaussian	0.7630	0.9108	0.8185	0.9095	0.9340	0.9900	0.6635
		SROCC (prediction monotonicity)						
Proposed	wavelet + GGD	0.4167	0.7833	0.8333	0.8833	0.7500	0.7333	0.5134
Proposed	DNT + Gaussian	0.5000	0.7667	0.8000	0.6667	0.8000	0.9833	0.7018

the same. The test results on the LIVE database and the Cornell-VCL database are shown in Tables 3.3 and 3.4, respectively, where the training data are the full LIVE database and the full Cornell-VCL database, respectively. It can be concluded from these tables that the overall performance is clearly improved from wavelet-domain to DNT-domain implementations.

The performance comparison with other IQA algorithms is shown in Tables 3.5 and 3.6. To the best of our knowledge, the only other RRIQA algorithm that has a comparable low RR data rate and is designed for general-purpose is the method proposed in [51]. In addition to this method, we have also included peak signal-to-noise-ratio (PSNR), which is still the most widely used full-reference IQA measure. Although such comparison is highly unfair to the proposed method and the method in [51] (PSNR requires full access to the original image, as opposed to the 48 scalar

features in the proposed method), it provides a useful indication of the relative performance of the proposed algorithm. For any IQA algorithm that involves a training process of the parameters, it is important to verify that the model is not overtrained. In other words, the performance of the algorithm should not change dramatically with different training data set. Therefore, in both Tables 3.5 and 3.6, we have included two versions of the proposed DNT-domain algorithm, where the only difference between them is that their model parameters (α, β, γ, δ and D_0) are trained with the LIVE database or the Cornell-VCL database (using all images in both cases). Such a cross-validation process is useful to test the robustness of the model. Not surprisingly, the test results are better when the parameters are trained with the same database than the results obtained by cross-training the parameters (Note that some image distortion types included in one database may not be included in the other). However, in both cases and for both databases, the proposed algorithm performs better than the method in [51]. In particular, it can be seen from both Table 3.5 and Table 3.6 that for the all-data cases, where all the images with different distortion types are mixed together, the method in [51] does not perform well, and the improvement of the proposed method is quite significant. Indeed, its LPCC and SROCC values (for all-data cases) are comparable or even higher than the full-reference PSNR measure.

3.3 RRVQA based on Statistics of Natural Image Sequences: Temporal Motion Smoothness

3.3.1 Statistics of Natural Image Sequences

While great effort has been made to study the statistical regularities of static natural images [78], much less has been done for natural image sequences. One approach is to compute the autocorrelation function of the image sequence along both spatial and temporal directions. Assuming spatial and temporal stationarities, such

Table 3.5. Performance comparison of IQA algorithms using the LIVE database

LIVE data set	JP2(1)	JP2(2)	JPG(1)	JPG(1)	Noise	Blur	FF	All data
	LPCC (prediction accuracy)							
PSNR	0.9337	0.8948	0.9015	0.9136	0.9866	0.7742	0.8811	0.8709
Wang *et al.* [51]	0.9353	0.9490	0.8452	0.9695	0.8889	0.8872	0.9175	0.8226
Proposed (training: Cornell-VCL)	0.9115	0.9422	0.8501	0.9354	0.9401	0.8773	0.9243	0.8930
Proposed (training: LIVE)	0.9485	0.9655	0.8203	0.9579	0.9654	0.9562	0.9464	0.9173
	SROCC (prediction monotonicity)							
PSNR	0.9231	0.8816	0.8907	0.8077	0.9855	0.7729	0.8785	0.8755
Wang *et al.* [51]	0.9298	0.9470	0.8332	0.8908	0.8639	0.9145	0.9162	0.8437
Proposed (training: Cornell-VCL)	0.9081	0.9239	0.8389	0.8734	0.9316	0.8608	0.9237	0.9093
Proposed (training: LIVE)	0.9478	0.9610	0.8143	0.8937	0.9559	0.9584	0.9443	0.9287
	Outlier Ratio (prediction consistency)							
PSNR	0.0805	0.0976	0.092	0.1818	0.0000	0.2069	0.1517	0.2373
Wang *et al.* [51]	0.0690	0.0366	0.1839	0.0341	0.1793	0.1172	0.0621	0.2311
Proposed (training: Cornell-VCL)	0.0230	0.0122	0.0805	0.1250	0.0345	0.0483	0.0345	0.1853
Proposed (training: LIVE)	0.0115	0.0122	0.1149	0.0341	0.0000	0.0000	0.0207	0.1069

Table 3.6. Performance comparison of IQA algorithms using the Cornell-VCL database

Cornell-VCL data set	FLT	JPG	JPG2	DCQ	BLR	NOZ	All data
	LPCC (prediction accuracy)						
PSNR	0.9100	0.7008	0.7957	0.5637	0.5904	0.9340	0.6347
Wang *et al.* [51]	0.4939	0.8575	0.7880	0.9357	0.7687	0.6252	0.3166
Proposed (training: LIVE)	0.4864	0.9183	0.8813	0.8847	0.8602	0.9489	0.5385
Proposed (training: Cornell-VCL)	0.7630	0.9108	0.8185	0.9095	0.9340	0.9900	0.6635
	SROCC (prediction monotonicity)						
PSNR	0.9000	0.6333	0.8000	0.5000	0.4667	0.9500	0.6205
Wang *et al.* [51]	0.1000	0.7667	0.5333	0.8000	0.6667	0.7333	0.2948
Proposed (training: LIVE)	0.1500	0.7833	0.8667	0.7333	0.7500	0.9500	0.5110
Proposed (training: Cornell-VCL)	0.5000	0.7667	0.8000	0.6667	0.8000	0.9833	0.7018

an autocorrelation function can be studied more conveniently in the Fourier transform domain as a spatiotemporal power spectrum [103]. It has been found that the spatiotemporal power spectrum of natural image sequences demonstrate interdependence between spatial and temporal frequencies, and the interdependence may be accounted for by assuming a static power spectrum and a rotationally invariant distribution of velocities [103]. Independent component analysis has also been applied to local 3-D blocks extracted from natural image sequences [104]. It was shown that the components obtained by optimizing independence are filters localized in space and time, spatially oriented, and directionally selective. Similar shapes of linear components

were also obtained by optimizing sparseness via a matching pursuit algorithm [105]. Other prior models about natural image sequences have also been assumed, though not directly measured. For example, in the literature of optical flow estimation, it is often assumed that image motion or optical flow is spatially smooth [106]. As a result, the motion or optical flow vectors measured locally should vary smoothly across space. Explicit prior models in favor of lower speed of motion has also been assumed [107, 103, 76] and applied to Bayesian optical flow estimation [107]. In a recent study [37], the shape of the "biological" speed prior was inferred directly from psychophysical speed perception experiment under the existence of noise. The inferred prior verifies the strong preference of slower motion and shows significantly heavier tails than a Gaussian.

Besides the preference for lower-speed and spatially-smooth motion, here we are interested in another type of statistical regularity of natural image sequences − the smoothness of motion along temporal direction. Figure 3.9 gives an illustration, where (a) and (b) are motion vector fields estimated from three consecutive frames of the "Susie" sequence. It can be observed that the motion vectors are slowly-varying not only over space, but also over time, which is confirmed by the difference motion vector fields shown in (c). The histograms of the vertical and horizontal components of (c) are plotted in (d) and (e), respectively, where the high peaks at 0 indicate the statistical preference of temporal motion smoothness.

Figure 3.9 suggests a direct method to capture temporal motion smoothness, i.e., estimating the motion vector fields of consecutive video frames and then measuring the variations of the motion vectors along temporal direction. However, motion estimation is a computationally expensive task, which often involves a complicated search procedure (e.g., in block-matching motion estimation algorithms [108]) or re-

quires solving adaptive equations at each spatial location (e.g., in optical flow-based motion estimation methods [106]).

We propose to investigate temporal motion smoothness in the complex wavelet transform domain, where the magnitudes of complex wavelet coefficients exhibit translation invariance properties [109], and the relative phase patterns between the coefficients have found to be the most informative in describing local image structures [110, 38]. In previous work, global (Fourier) and local (wavelet) phases have been found to carry important information about image structures [111, 112, 113, 110]. The local phase structure of static natural images demonstrates clear statistical regularities and has intriguing perceptual implications [110]. In the computer vision literature, local phase has been used in a number of applications such as estimation of image disparity [114] and motion [115, 116], description of image textures [117], and recognition of persons using iris patterns [118]. However, the behaviors of local phase variations over time, whether such behaviors can be used to characterize "natural" image sequences, and how "unnatural" image distortions interfere with such behaviors have not been deeply investigated.

3.3.2 Temporal Motion Smoothness by Local Phase Correlations

Let $f(x)$ be a given real static signal, where x is the index of spatial position. When $f(x)$ represents an image, x is a 2-D vector. For simplicity, in the derivations below, we assume x to be one dimensional. However, the results can be easily generalized to two and higher dimensions. A time varying image sequence can be created from the static image $f(x)$ with rigid motion and constant variations of average intensity:

$$h(x,t) = f(x + u(t)) + b(t)\,. \tag{3.22}$$

Here $u(t)$ indicates how the image positions move spatially as a function of time. $b(t)$ is real and accounts for the time-varying background luminance changes. This formulation can be viewed as a generalization of the brightness constancy assumption [119, 106] (in which $b(t) \equiv 0$), but the inclusion of the luminance change improves flexibility and stability of the representation. For example, when the lighting condition of a fixed scene changes over time, the brightness constancy assumption would not hold, but the situation would be better described with this formulation.

Now consider a family of symmetric complex wavelets whose "mother wavelets" can be written as a modulation of a low-pass filter $w(x) = g(x)\, e^{j\omega_c x}$, where ω_c is the center frequency of the modulated band-pass filter, and $g(x)$ is a slowly varying and symmetric function. The family of wavelets are dilated/contracted and translated versions of the mother wavelet:

$$w_{s,p}(x) = \frac{1}{\sqrt{s}}\, w\left(\frac{x-p}{s}\right) = \frac{1}{\sqrt{s}}\, g\left(\frac{x-p}{s}\right)\, e^{j\omega_c (x-p)/s}\,, \tag{3.23}$$

where $s \in R^+$ is the scale factor, and $p \in R$ is the translation factor. Considering the fact that $g(-x) = g(x)$, and using the convolution theorem and the scaling and modulation properties of the Fourier transform, we can compute the complex wavelet transform of a given signal $f(x)$ as

$$\begin{aligned} F(s,p) &= \int_{-\infty}^{\infty} f(x)\, w^*_{s,p}(x)\, dx = \left[f(x) \star \frac{1}{\sqrt{s}}\, g\left(\frac{x}{s}\right)\, e^{j\omega_c x/s} \right]_{x=p} \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \sqrt{s}\, G(s\,\omega - \omega_c)\, e^{j\omega p}\, d\omega\,, \end{aligned} \tag{3.24}$$

where $F(\omega)$ and $G(\omega)$ are the Fourier transforms of $f(x)$ and $g(x)$, respectively. Applying such a complex wavelet transform to both sides of Eq. (3.22) at a given time instance t, we have

$$\begin{aligned} H(s,p,t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\sqrt{s}\,G(s\,\omega-\omega_c)\,e^{j\omega(p+u(t))}\,d\omega \\ &= \frac{e^{j(\omega_c/s)u(t)}}{2\pi}\int_{-\infty}^{\infty} F(\omega)\sqrt{s}\,G(s\omega-\omega_c)e^{j\omega p}e^{j(\omega-\omega_c/s)u(t)}d\omega \\ &\approx F(s,p)\,e^{j(\omega_c/s)u(t)}\,. \end{aligned} \tag{3.25}$$

Here $b(t)$ is eliminated because of the bandpass nature of the wavelet filters. The approximation is valid when the envelope window $g(t)$ is slowly varying and the motion $u(t)$ is small. In the extreme case, the approximation becomes exact when $g(x) \equiv 1$, i.e., $G(\omega) = \delta(\omega)$, or when there is no motion, i.e., $u(t) = 0$. A more convenient way to understand Eq. (3.25) is to take a logarithm on both sides, which gives

$$\ln H(s,p,t) \approx \ln F(s,p) + j(\omega_c/s)u(t)\,. \tag{3.26}$$

Note that the first term of the right-hand-side does not change over time. The key property of Eq. (3.26) is that at a given scale s and a given spatial position p, the imaginary part of the logarithm of the complex wavelet coefficient changes linearly with $u(t)$. In other words, the local phase structures over time can be fully characterized by the movement function $u(t)$. Taylor series expansion of $u(t)$ at a specific time instance t_0 yields

$$u(t) = u(t_0) + u'(t_0)(t-t_0) + \frac{u''(t_0)}{2}(t-t_0)^2 + \cdots + \frac{u^{(n)}(t_0)}{n!}(t-t_0)^n + \cdots\,. \tag{3.27}$$

where $u'(t_0)$ is the first order derivative, $u''(t_0)$ is the second order derivative and $u^{(n)}(t_0)$ is the nth order derivative of $u(t)$ at t_0.

We call $u(t)$ N-th order smooth if its $(N+1)$-th and higher order derivatives with respect to t are all zeros. For instance, zero-order smooth motion implies no

motion [$u(t)$ is a constant over time], first-order smooth motion corresponds to constant speed [$u'(t)$ is a constant], and second-order smooth motion leads to constant acceleration [$u''(t)$ is a constant], and so on. Notice that here the definition of motion smoothness is different from the notion of motion smoothness typically used in optical flow estimation [106], where motion smoothness refers to the slow variations of motion vectors over space. We believe that *temporal motion smoothness* is a better term to describe the concept we are discussing here.

In order to relate temporal motion smoothness with the time-varying complex wavelet transform relationship of Eq. (3.26), we must examine the complex wavelet coefficients at multiple time instances. A convenient choice is to start from a time instance t_0 and sample the sequence at consecutive time steps $t_0 + n\Delta t$ for $n = 0, 1, ..., N$ (Δt is the time interval). The N-th order derivatives of $u(t)$ at t_0 can be approximated by the following N-th order differentiator:

$$u^{(N)}(t_0) = \frac{1}{(\Delta t)^N} \sum_{n=0}^{N} (-1)^{n+N} \binom{N}{n} u(t_0 + n\Delta t)\,. \tag{3.28}$$

where $\binom{N}{n}$ denotes the number of n-combinations (each of size n) from a set with N elements (size N). Now we define the N-th order *temporal correlation function* as follows:

$$L_N(s,p) = \sum_{n=0}^{N} (-1)^{n+N} \binom{N}{n} \ln H(s,p,t_0 + n\Delta t)\,. \tag{3.29}$$

By Eq. (3.26), we have

$$\begin{aligned} L_N(s,p) &\approx \sum_{n=0}^{N} (-1)^{n+N} \binom{N}{n} \left[\ln F(s,p) + j(\omega_c/s) u(t_0 + n\Delta t)\right] \\ &= (-1)^N \ln F(s,p) \left[\sum_{n=0}^{N} (-1)^n \binom{N}{n}\right] + j\frac{w_c}{s} \left[\sum_{n=0}^{N} (-1)^{n+N} \binom{N}{n} u(t_0 + n\Delta t)\right] \\ &= j\frac{w_c (\Delta t)^N}{s} u^{(N)}(t_0)\,, \end{aligned} \tag{3.30}$$

where we have used Eq. (3.28) and the fact that $\sum_{n=0}^{N}(-1)^n\binom{N}{n} = 0$. Now suppose that the motion is (N-1)-th order smooth, then $u^{(N)}(t_0) = 0$, and therefore

$$L_N(s,p) \approx 0\,. \tag{3.31}$$

It needs to be kept in mind that this approximation is achieved based on the ideal formulation of Eq. (3.22) and the ideal assumption of (N-1)-th order temporal motion smoothness. Real natural image sequences are expected to deviate from these assumptions. However, by looking at the statistics of $L_N(s,p)$ (especially its imaginary part, which is a measure of temporal local phase correlation), one may be able to quantify such deviation and use it as an indication of the strength of temporal motion smoothness.

In addition, we define the following temporal weighted averaging function in the log-complex wavelet domain:

$$M_N(s,p) = \sum_{n=0}^{N}\binom{N}{n}\ln H(s,p,t_0+n\Delta t)\,. \tag{3.32}$$

We find it also helpful in characterizing the statistical properties of natural image sequences and will demonstrate its usefulness in the next section.

3.3.3 Image Sequence Statistics

For a given image sequence, we decompose each frame using the complex version [117] of the steerable pyramid [72], a multi-scale wavelet decomposition whose basis functions are spatially localized, oriented, and roughly one octave in bandwidth. Specifically, a 3-scale 2-orientation pyramid is computed, resulting in six oriented subbands, a highpass residual band, and a lowpass residual band. By aligning the oriented subbands at the same orientation and scale but across different frames, we obtain a discrete (in both space and time) version the function $H(s,p,t)$ for a partic-

ular scale and orientation. We then compute $L_N(s,p)$ and $M_N(s,p)$ for $N = 1,2,3,4$ for all the coefficients within the subband.

To study temporal motion smoothness, we first examine the marginal distribution of the imaginary part of the temporal correlation coefficient $imag\{L_N(s,p)\}$. The histograms of $imag\{L_N(s,p)\}$ for $N = 1,2,3,4$ of the "Susie" sequence are shown in Fig. 3.10. It can be observed that all the histograms peak near zero, and the peaks move toward zero with the increasing order of the temporal correlation function. Although Fig. 3.10 only shows the statistical results from a single image sequence, similar results were obtained for most of the other sequences we tested[2]. This demonstrates strong prior of temporal motion smoothness of natural image sequences. Another important observation is that the histograms are quite peaky, much more than the von Mises distribution widely used in describing statistics of circular data [120]. We empirically found that a four-parameter function that can almost always well describe the data is given by

$$p_m(\theta) = \frac{1}{Z}\left\{\exp\left[-\left(\frac{|\sin[(\theta-\theta_0)/2]|}{\alpha}\right)^{\beta}\right] + C\right\} \tag{3.33}$$

where θ is the phase variable, Z is a normalization constant, and the four parameters θ_0, α, β and C control the center position, width, peakedness and the baseline of the function, respectively. We numerically fit the histograms with the model by minimizing the Kullback-Leibler distance [44] (KLD) between the observed and the model distributions. Some fitting results are demonstrated in Figure 3.10. We have used this fitting model for reduced-reference image quality assessment, which will be detailed in Section 3.3.5.

[2]Exceptions were observed for the image frames across scene changes and for the image frames with very large motion (where the distances of moving objects between frames are beyond the coverage of the wavelet filter envelops).

We have also studied the relationship between temporal motion smoothness and the strength of the underlying local signal. In particular, we generate the conditional histogram of the imaginary part of $L_N(s,p)$ versus the real part of $M_N(s,p)$, which provides a useful measure of local signal strength. The result is demonstrated in Figure 3.11(b), where each column in the 2-D histogram is normalized to one. Again, the histogram shows strong temporal motion smoothness, and such a statistical regularity becomes stronger with the increase of local signal strength. This is not surprising because small magnitude coefficients typically come from the smooth background regions in an image and are easily disturbed by background noise.

3.3.4 Interference with "Unnatural" Distortions

The merit of natural image prior models should be evaluated by their capabilities of distinguishing natural and unnatural images. Here we simulate a set of "unnatural" image distortions that often occur in real-world applications and examine how these distortions interfere with the temporal motion smoothness prior.

The distortions being tested are divided into two categories. The first category of distortions do not change individual pixel values but directly disturb temporal motion smoothness by shifting the positions of pixels. Specifically, we investigated the effects of line jittering, frame jittering and frame dropping distortions, each of which is associated with certain real-world scenario. In particular, line jittering occurs when two fields of interlaced video signals are not synchronized, frame jittering is often caused by irregular camera movement such as hand shaking, and frame dropping usually happens when the bandwidth of a real-time communication channel drops and some video frames have to be discarded to reduce the bit rate of the video signal being transmitted. To simulate line jittering, we shift each line in a video frame horizontally by a random amount uniformly distributed between a range of $[-S, S]$,

where S defines the level of jittering distortion. Figure 3.12 shows the results of line jittering. Comparing the marginal and conditional histograms (Fig. 3.12(a) and (b)) with those in Fig. 3.11, we observe that the distributions of temporal phase correlation coefficients become almost flat, which implies that the prior structure of temporal motion smoothness shown in Fig. 3.11 is severely disturbed. Frame jittering is simulated in a similar way, only that the entire frame (rather than each line in the frame) is shifted together. Again, the statistical regularity of temporal motion smoothness has been destroyed, as demonstrated in Fig. 3.13. To simulate frame dropping, we discard N out of every $N + 1$ frames and use N to define the level of frame dropping. The dropped frames will then be filled by repeating their previous frames. Figure 3.14 shows the effect of frame dropping. It can be seen that the sharpness of the marginal and conditional histograms is significantly reduced and the centers of the peaks in the distributions are shifted away from 0, demonstrating a clear disruption of temporal motion smoothness.

The second category of distortions directly alters the values of individual image pixels. In particular, we studied the effects of additive white Gaussian noise contamination and Gaussian blur distortion. Although they do not directly change the motion information contained in the video, they reduce the sharpness of local image structures, and thus affect the local phase correlations across frames. In Fig. 3.15, white Gaussian noise is added to each frame of the video sequence, where the noise level is defined as the standard deviation of the Gaussian distribution. In Figure 3.16, each video frames is blurred spatially by convolving with a linear filter of Gaussian shape, where the standard deviation of the Gaussian filter defines the blur level. It can be observed that in both cases, the strong prior of temporal motion smoothness is significantly reduced.

3.3.5 Application to Reduced-Reference Video Quality Assessment

From the study in previous sections, we observe that temporal motion smoothness is a common feature of natural image sequences but disrupted by various types of "unnatural" image distortions. One direct application of such a feature is to use it for reduced-reference video quality assessment (RRVQA), which aims to estimate video quality degradations with only partial information about the "perfect-quality" reference video (This is different from full-reference video quality measures such as peak signal-to-noise ratio and the structural similarity index [70] that require full access to the original video). The idea is to use temporal motion smoothness measures extracted from the reference video signal as the RR features and then quantify video quality degradations based on the variations of these RR features in the distorted video signal.

For a given image sequence, we first divide it into groups of pictures (GOPs), each containing 3 consecutive frames. For each GOP, we apply a complex steerable pyramid decomposition to all 3 frames and compute the second order temporal correlation function $L_2(s, p)$ for each oriented subband. The observations of the marginal histograms shown in Figs. 3.12 to 3.16 suggest that the variations in the marginal distributions of $imag\{L_2(s, p)\}$ between the original and distorted image sequences can be used as a measure of image distortions. A convenient way to quantify such variations is to compute the KLD [44] between them:

$$d(p\|q) = \int p(\theta) \ln \frac{p(\theta)}{q(\theta)} d\theta \,, \tag{3.34}$$

where $p(\theta)$ and $q(\theta)$ are the probability density functions of $imag\{L_2(s, p)\}$ of the original and distorted signals, respectively. To accomplish this, the histograms of both the original and distorted signals must be available. The latter can be easily computed from the distorted signal, which is always available. The difficulty is in

obtaining the histogram of the reference signal. Using all the histogram bins as RR features would result in either a heavy RR data rate (when the bin size is fine) or a poor approximation accuracy (when the bin size is coarse). To overcome this problem, we make use of the fitting model of Eq. (3.33), such that only four parameter (θ_0, α, β and C) are needed to describe it (as opposed to all the histogram bins). Furthermore, to account for the variations between the model and the true distribution, we compute the KLD between $p_m(\theta)$ and $p(\theta)$ as

$$d(p_m \| p) = \int p_m(\theta) \ln \frac{p_m(\theta)}{p(\theta)} d\theta \tag{3.35}$$

In summary, a total of 5 RR features (4 features to describe $p_m(\theta)$ together with $d(p_m \| p)$) are extracted from each subband of the original signal.

To evaluate the quality of the distorted image sequence, we first estimate the KLD between the probability density function $q(\theta)$ of the $imag\{L_2(s,p)\}$ coefficients computed from the distorted signal and the model $p_m(\theta)$ estimated from the original signal:

$$d(p_m || q) = \int p_m(\theta) \ln \frac{p_m(\theta)}{q(\theta)} d\theta \,. \tag{3.36}$$

Combining this with the available RR feature $d(p_m||p)$, we obtain an estimate of the KLD between $p(\theta)$ and $q(\theta)$:

$$\hat{d}(p||q) = d(p_m||q) - d(p_m||p) = \int p_m(\theta) \ln \frac{p(\theta)}{q(\theta)} d\theta \,. \tag{3.37}$$

With the additional cost of adding one more RR parameter $d(p_m||p)$, Eq. (3.37) not only delivers a more accurate estimate of $d(p\|q)$ than Eq. (3.36), but also provides a useful feature that when there is no distortion between the original and distorted signals (which implies that $p(\theta) = q(\theta)$ for all θ), both the targeted distortion measure

$d(p\|q)$ and estimated distortion measure $\hat{d}(p\|q)$ are exactly zero. Finally, the overall quality degradation of the distorted image sequence is computed as

$$D = \frac{1}{K} \sum_{GOPs} \sum_{subbands} \hat{d}(p\|q)\,, \tag{3.38}$$

where K is the number of GOPs in the image sequence.

We test the proposed algorithm using five types of distortions, including line jittering, frame jittering, frame dropping, additive white Gausssian noise contamination and Gaussian blur, as described in Section 3.3.4. The results for the "Susie" image sequence of the five distortion types are shown in Figs. 3.12 to 3.16 (c), respectively, where the horizontal axes indicate the distortion levels and the vertical axes show the distortion measure computed using Eq. (3.38). It can be observed that the same objective distortion measure D is consistently increasing with the strength of each individual type of distortion. Similar results were obtained for other image sequences we tested. This demonstrates the potential of the proposed method for general-purpose RRVQA, which is different from most VQA approaches in the literature where ad-hoc features tuned to specific distortion types (such as blocking [57] and ringing [121] artifacts) are often used, and thus limit their application scope. Another interesting observation is regarding the frame jittering and frame dropping distortions. Notice that with these two types of distortions, the quality of each individual frame remains high quality, and thus frame-by-frame quality assessment approaches would give high quality scores to the image sequences undergoing these distortions, but the proposed method can capture them quite effectively without any specific change of the algorithm.

Figure 3.3. Results of RR IQA based joint distribution. (a) Original image: D=4.0874e-007, UIQI=1, MSE=0 (b) Fig.3.1 b: D=0.3715, UIQI=0.460, MSE=580 (c) Multiplicative Speckle Noise:D=0.4555 UIQI=0.4408, MSE=225 (d) Mean shift: D= 0.0468, UIQI=0.9894,MSE=225 (e) Contrast stretching: D=0.1039, UIQI=0.9372, MSE=225 (f) blurring: D=1.8112, UIQI=0.3261,MSE=225 (g) JPEG Compression: D=4.4830, UIQI=0.2876, MSE=215 (h) Additive Gaussian noise: D=0.3286, UIQI=0.3891,MSE=225.

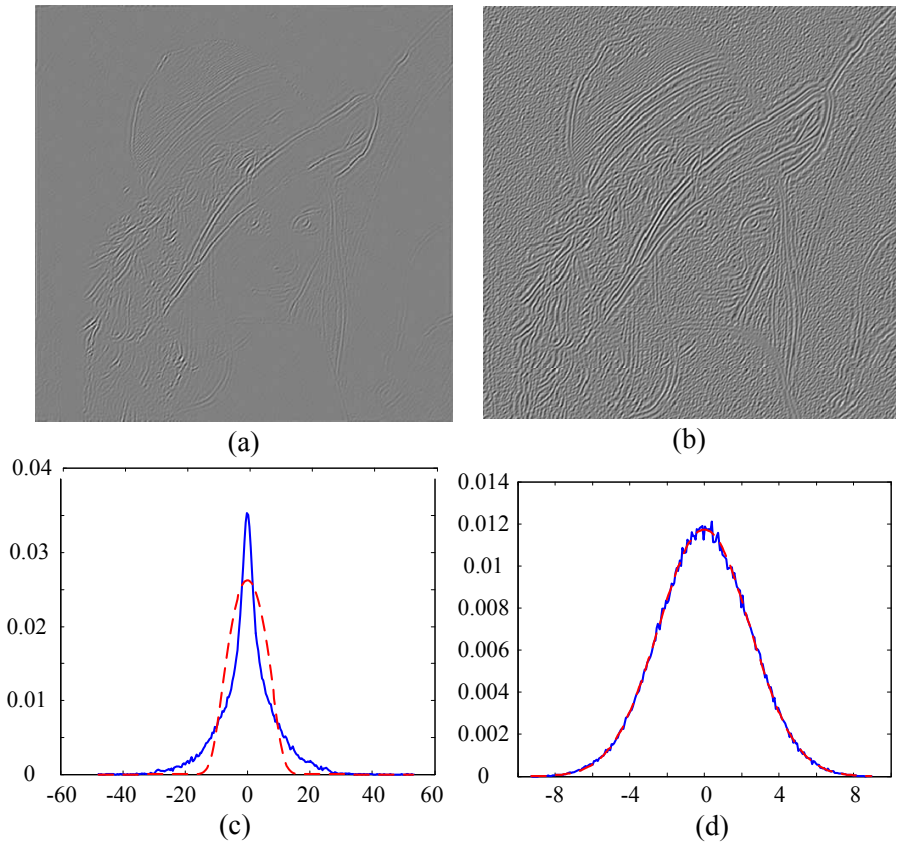

Figure 3.4. (a) original wavelet coefficients; (b) DNT coefficients; (c) histogram of original coefficients (solid curve) and a Gaussian curve with the same standard deviation (dashed curve); (d) histogram of DNT coefficients (solid) fitted with a Gaussian model (dashed).

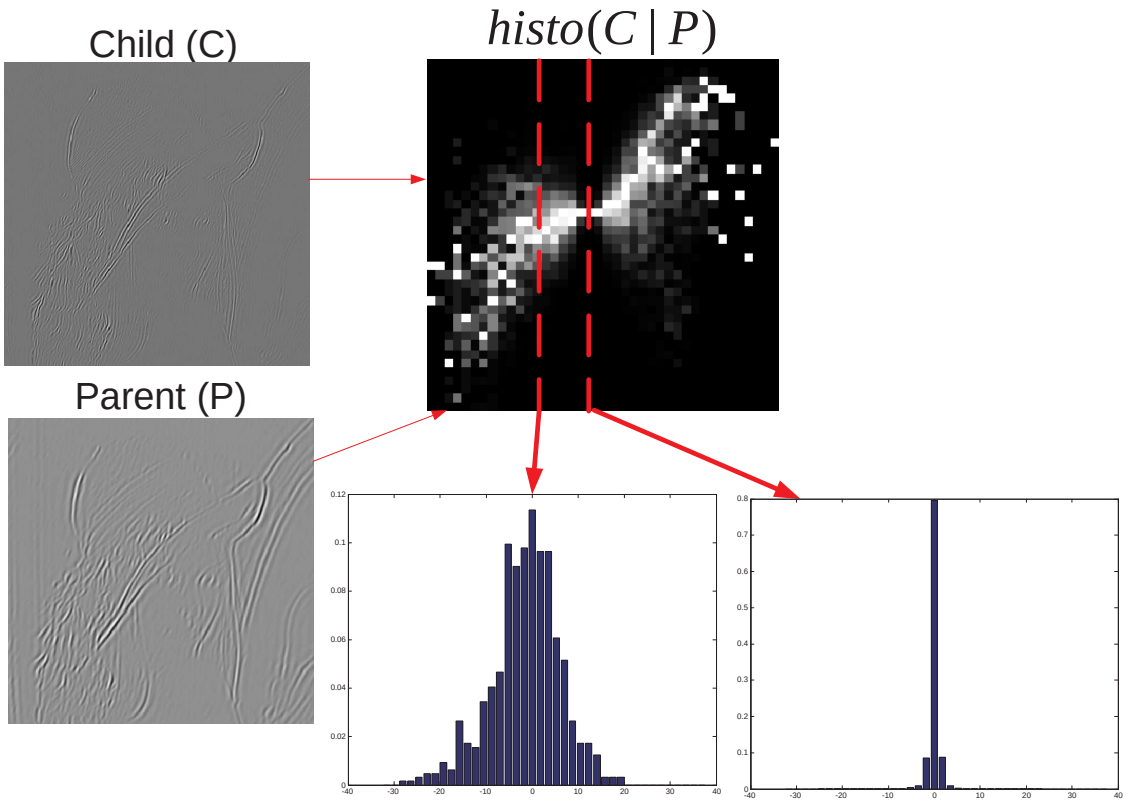

Figure 3.5. Conditional histograms between a parent and a child coefficients extracted from the original wavelet representation.

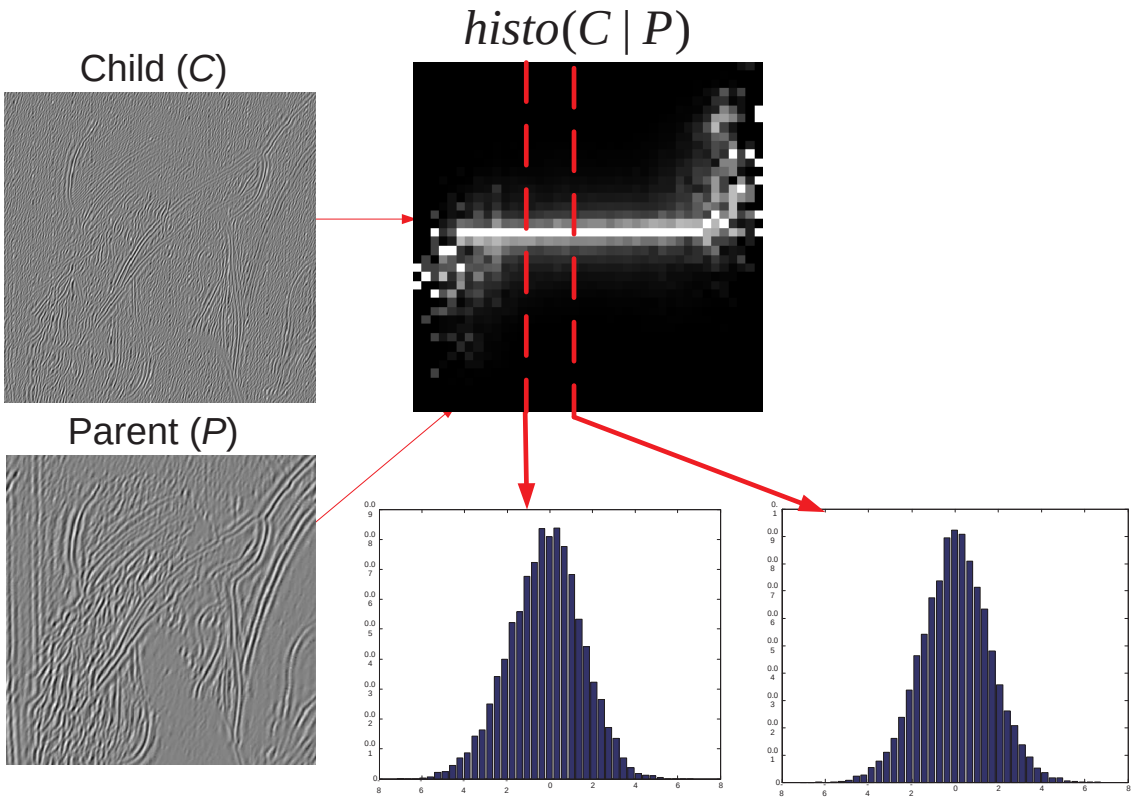

Figure 3.6. Conditional histograms between a parent and a child coefficients extracted from the corresponding DNT representation (b).

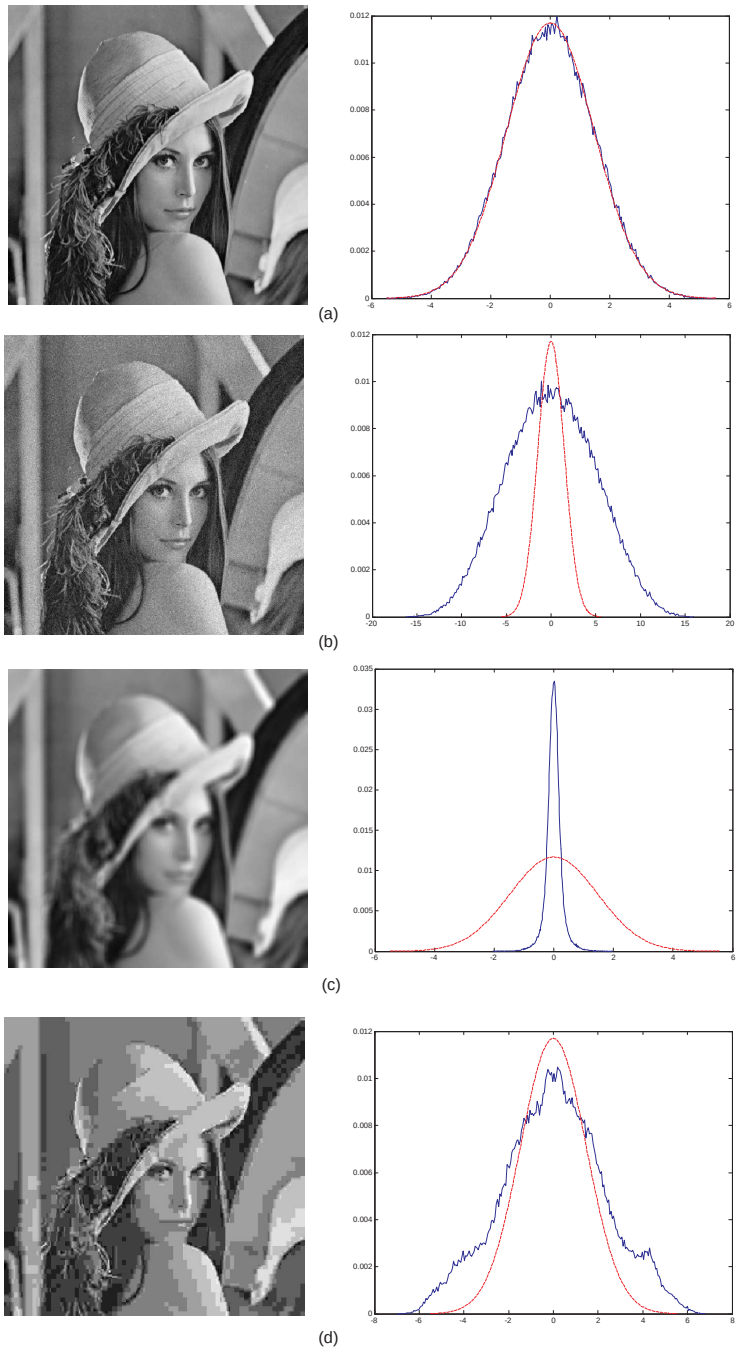

Figure 3.7. Histograms of DNT coefficients in a wavelet subband under different types of image distortions. (a) original "Lena" image; (b) Gaussian noise contaminated image; (c) Gaussain blurred image; (d) JPEG compressed image. Solid curves: histograms of DNT coefficients. Dashed curves: the Gaussian model fitted to the histogram of DNT coefficients in the original image. Significant departures from the Gaussian model is observed in the distorted images (b), (c) and (d).

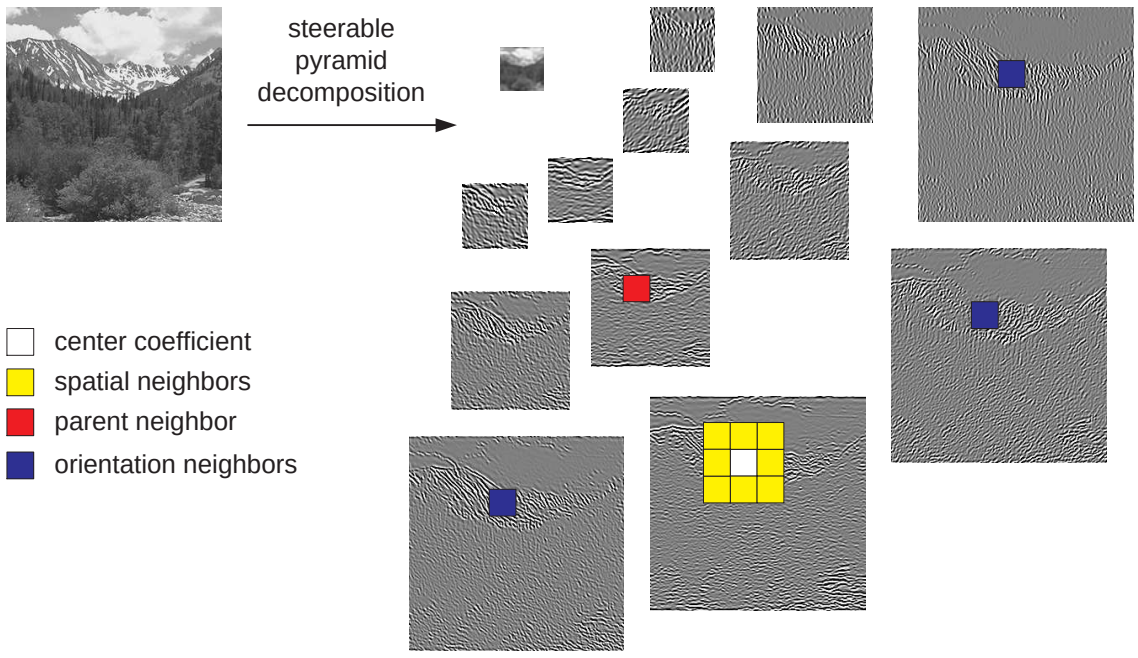

Figure 3.8. Illustration of steerable pyramid decomposition and the selection of DNT neighbors. The neighboring coefficients include the 3×3 spatial neighbors within the same subband, one parent neighboring coefficient and three orientation neighboring coefficients.

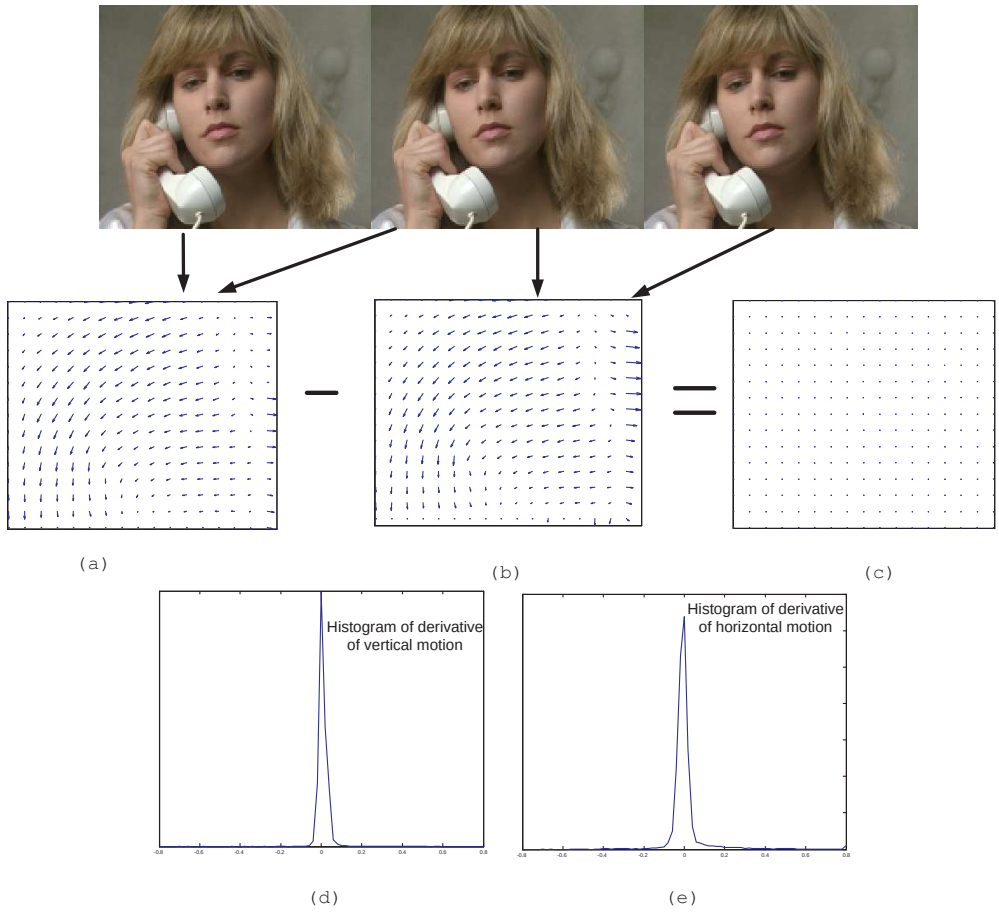

Figure 3.9. Illustration of motion smoothness of natural image sequences. The motion vector fields estimated for consecutive video frames are slowly varying over both space and time .

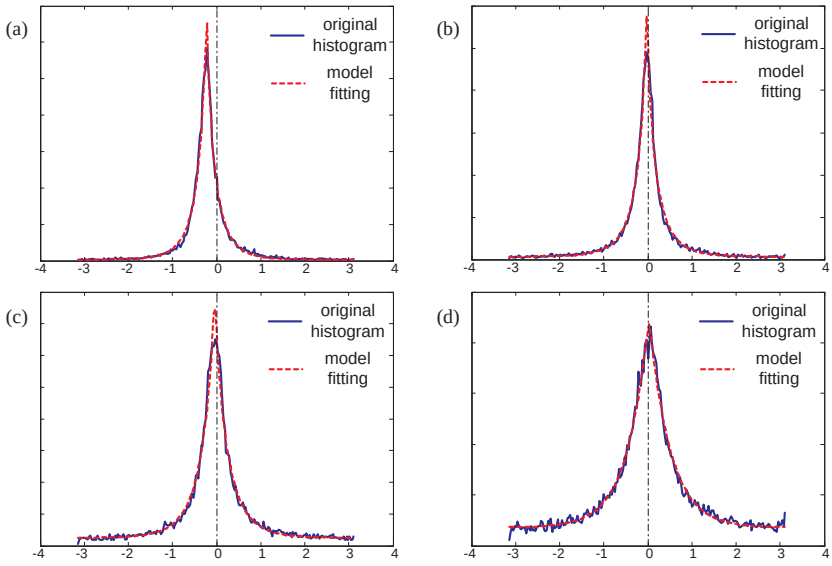

Figure 3.10. Marginal statistics of the imaginary parts of the first-order (a), second-order (b), third-order (c), and fourth-order (d) temporal correlation functions $L_N(s,p)$. The image sequence demonstrates strong temporal motion smoothness.

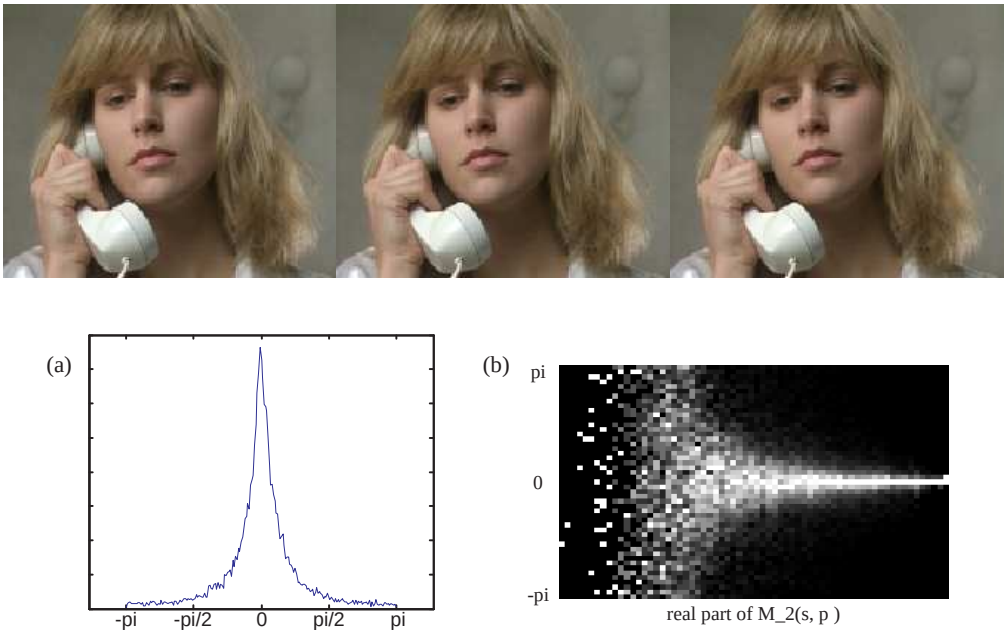

Figure 3.11. Three consecutive frames of the image sequence "Susie" and statistics of the second-order temporal correlation function $L_2(s,p)$. (a) Marginal histogram of the imaginary part; (b) Histogram of the imaginary part of $L_2(s,p)$ conditioned on the real part of $M_2(s,p)$.

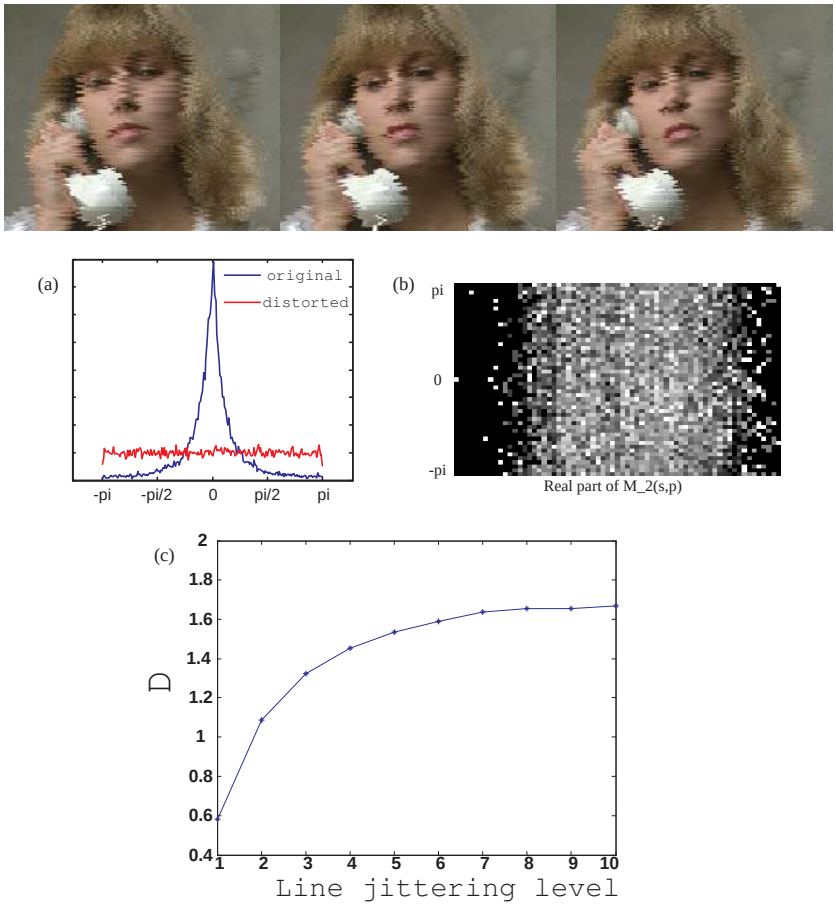

Figure 3.12. Three consecutive frames of the image sequence "Susie" distorted with line jittering and statistics of the second-order temporal correlation function $L_2(s,p)$. (a) Marginal histogram of the imaginary part; (b) Histogram of the imaginary part of $L_2(s,p)$ conditioned on the real part of $M_2(s,p)$; (c) Objective RRVQA score D as a function of line jittering level.

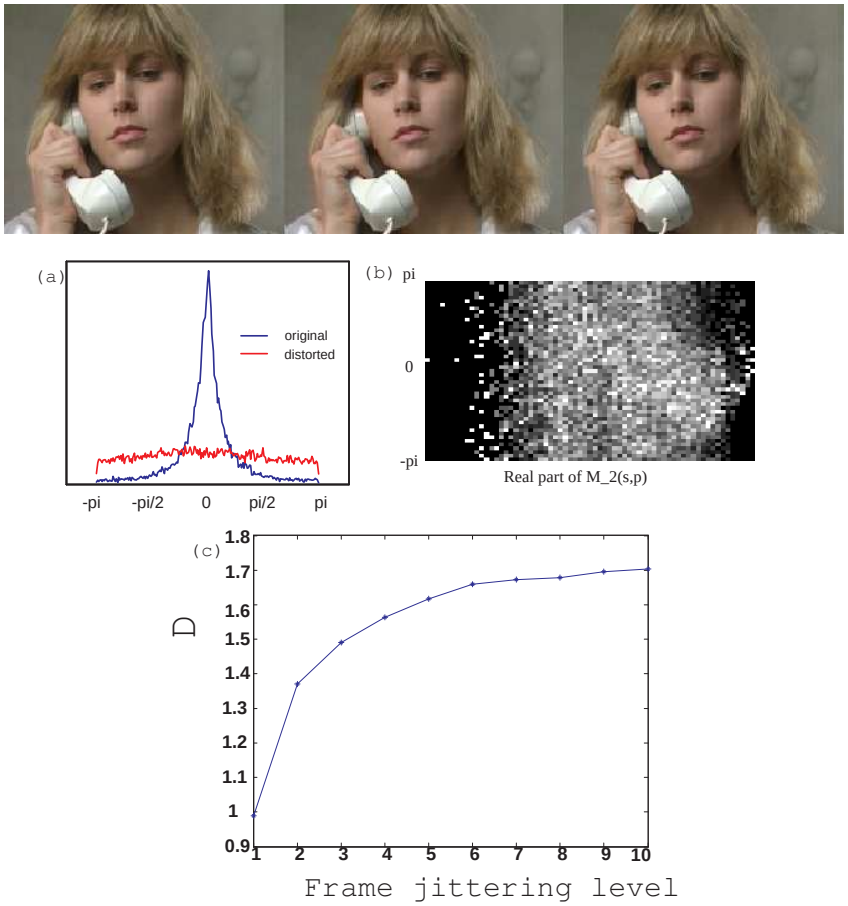

Figure 3.13. Three consecutive frames of the image sequence "Susie" distorted with frame jittering and statistics of the second-order temporal correlation function $L_2(s,p)$. (a) Marginal histogram of the imaginary part; (b) Histogram of the imaginary part of $L_2(s,p)$ conditioned on the real part of $M_2(s,p)$; (c) Objective RRVQA score D as a function of frame jittering level.

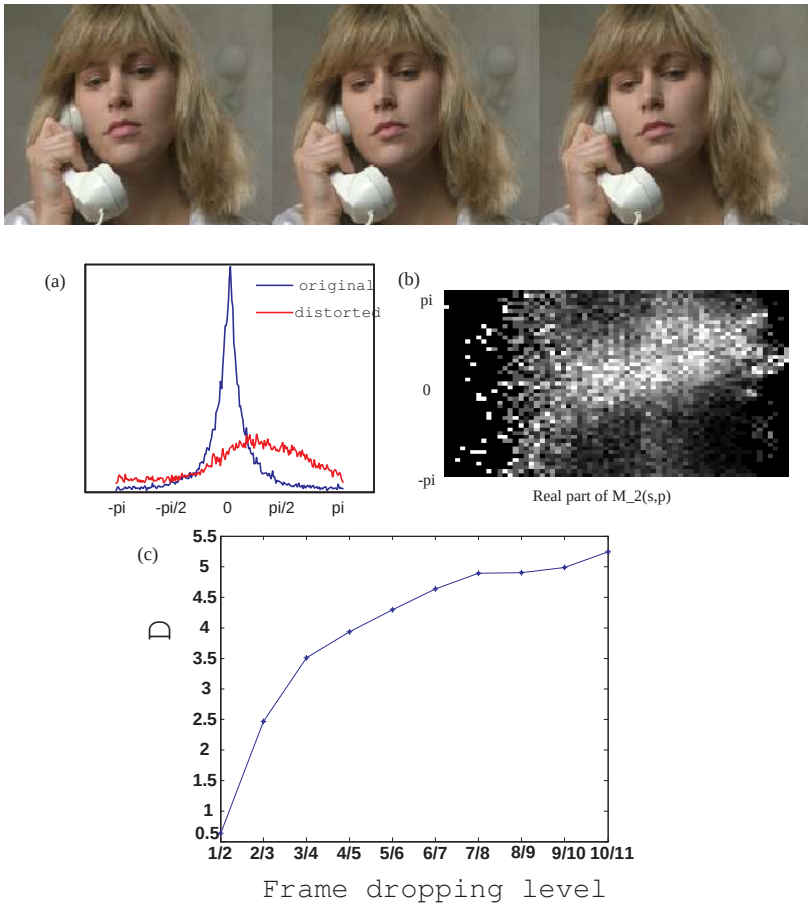

Figure 3.14. Three consecutive frames of the image sequence "Susie" with frame dropping distortion and statistics of the second-order temporal correlation function $L_2(s,p)$. (a) Marginal histogram of the imaginary part; (b) Histogram of the imaginary part of $L_2(s,p)$ conditioned on the real part of $M_2(s,p)$; (c) Objective RRVQA score D as a function of frame dropping level.

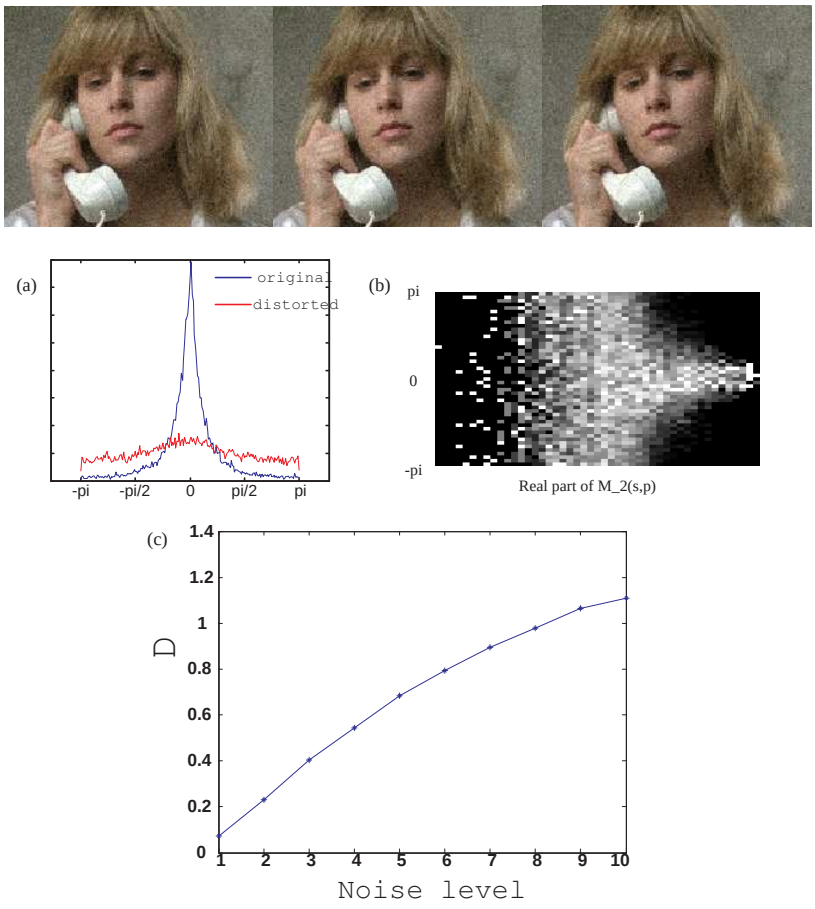

Figure 3.15. Three consecutive frames of the image sequence "Susie" contaminated with different levels of white Gaussian noise and statistics of the second-order temporal correlation function $L_2(s,p)$. (a) Marginal histogram of the imaginary part; (b) Histogram of the imaginary part of $L_2(s,p)$ conditioned on the real part of $M_2(s,p)$; (c) Objective RRVQA score D as a function of noise level.

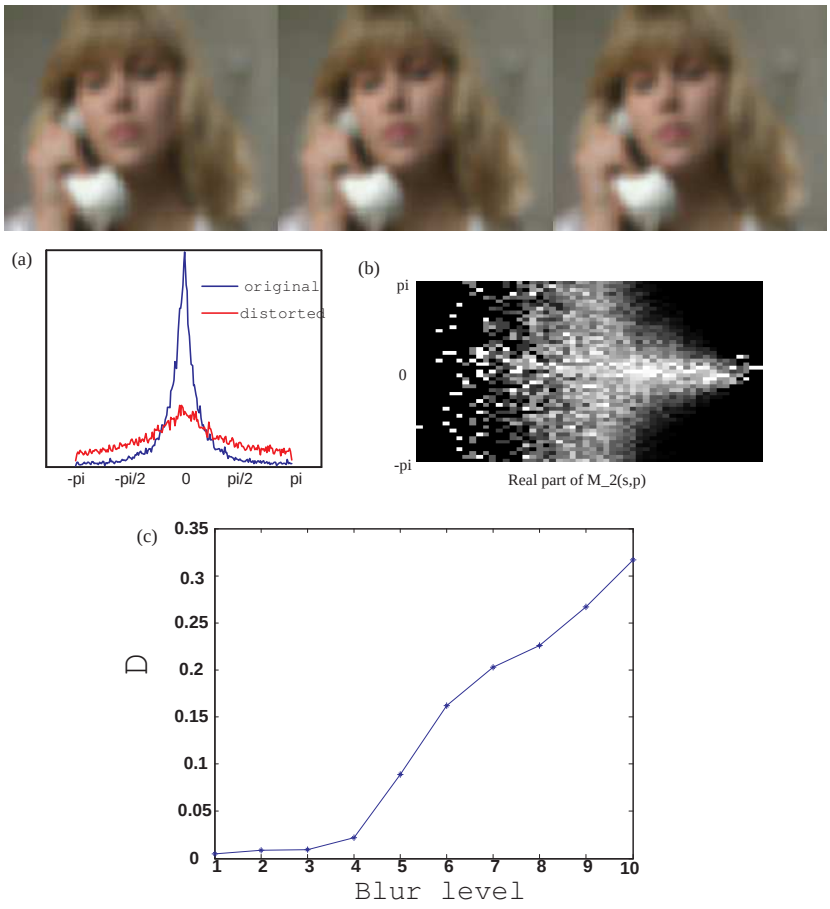

Figure 3.16. Three consecutive frames of the image sequence "Susie" distorted with different levels of Gaussian blur and statistics of the second-order temporal correlation function $L_2(s,p)$. (a) Marginal histogram of the imaginary part; (b) Histogram of the imaginary part of $L_2(s,p)$ conditioned on the real part of $M_2(s,p)$; (c) Objective RRVQA score D as a function of blur level.

CHAPTER 4

APPLICATIONS OF THE OBJECTIVE QUALITY ASSESSMENT METHOD

4.1 Perceptual Image Coding Based on a Maximum of Minimal Structural Similarity Criterion

4.1.1 Perceptual Image Coding

Image coding algorithms have been traditionally optimized to achieve the minimal mean squared error (MSE) under the constraint of a limited bit budget. However, MSE has been widely criticized for poorly correlating with visual perception of image quality [9]. An example is shown in Fig. 4.1, where a JPEG compressed image is evaluated locally to create the the absolute error map and the structural similarity (SSIM) index [70] map. Both maps use brighter pixels to indicate better quality, but they give substantially different evaluations. Careful inspection of the distorted image together with the quality maps concludes that absolute error (which is the basis for all Minkowski error metrics, including MSE) is not a good indicator of local image quality when compared with the SSIM index (e.g., the SSIM map clearly points out the annoying blocking artifacts in the sky).

The poor performance of MSE motivated researchers to incorporate perceptual models in image coding [122]. Most perceptual coding methods first decompose the image signal using a linear transform (e.g., a DCT or a wavelet transform) and then normalize (rescale) each transform coefficient with a *perceptual weight* before a *uniform* quantization and entropy coding scheme is applied. These weights may be determined by a number of psychophysical features of the human visual system (HVS), typically including the contrast sensitivity function and the contrast masking

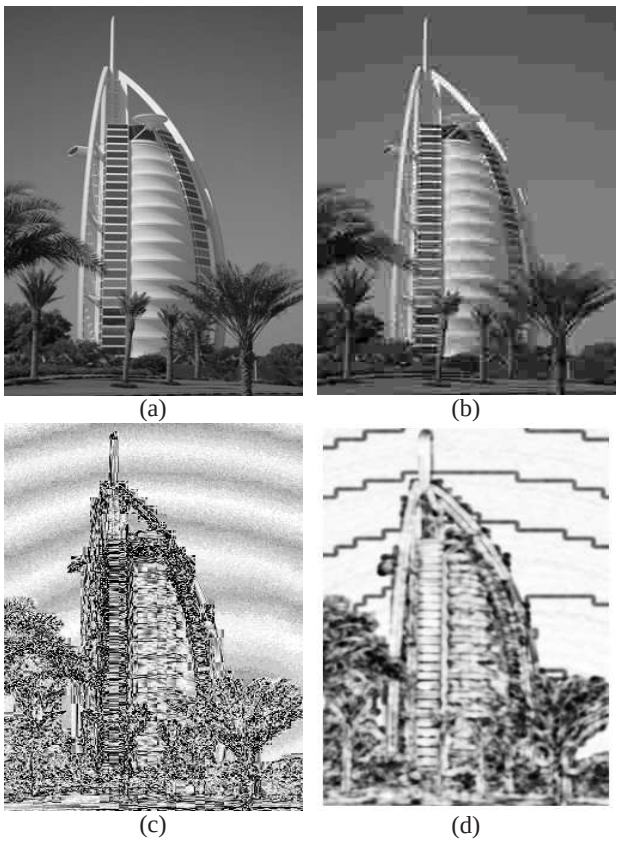

Figure 4.1. (a) Original image; (b) distorted image (by JPEG compression); (c) absolute error map − brighter indicates better quality (smaller absolute difference); (d) SSIM index map − brighter indicates better quality (larger SSIM value).

effect [122, 9]. An equivalent method is to design a *nonuniform quantization* scheme, where the quantization steps of the transform coefficients are proportional to their perceptual weights. This general design principle has been employed in many existing algorithms (e.g., [123, 122, 124, 125]), with variations in the linear transforms being used and the way the perceptual weights are computed. It has also been used in the design of the JPEG quantization table [85] and the visual optimization tools in JPEG2000 [86, 126]. This approach is appealing because it completely separates perceptual modeling from the subsequent processes, making it convenient to implement. Nevertheless, its accuracy is questionable. For example, when the masking effect is

considered, the perceptual weight of a given coefficient is computed from its neighboring coefficients. However, after the subsequent nonlinear quantization process, the coefficient and all of its neighboring coefficients have been changed. As a result, the computed masking effect and the corresponding perceptual weight would not be accurate anymore.

4.1.2 Method of Image Coding Based on a Maximum of Minimal Structural Similarity Criterion

We propose a different approach for perceptual image coding. First, we use the SSIM index map as a local perceptual quality indicator, which, to the best of our knowledge, has not been used directly for image coding before. Second, we do not attempt to impose perceptual modeling using one single normalization process. Instead, we encode the image iteratively. Within each iteration, we operate on the bit allocation scheme that redistributes the available bits over the image space according to the SSIM quality map obtained from the last iteration. Third, our scheme aims to improve the worst case scenario, such that the quality at the lowest quality region in the image is enhanced. In other words, we use a *maximum of minimal structural similarity criterion* as our optimization goal. This is justified based on the observation that human visual attention is often attracted to the image regions with extremely annoying artifacts (very low quality) that could dominate the quality evaluation of the entire image.

The central idea of our method is to iteratively redistribute the available bits based on local image quality measures. We find that an easy way to implement the idea is to incorporate it into an embedded bitplane coding algorithm. Embedded bitplane coding [127, 128, 86] has received wide acceptance in the past fifteen years. It encodes images into continuously scalable bit streams that can be truncated at

arbitrary places to create multiple versions of decoded images with variable bit rate and quality. Moreover, the encoded information bits are naturally organized according to their importance. Figure 4.2(a) provides a simple illustration of a regular bitplane coding scheme. The image components (typically wavelet coefficients) are binary represented and aligned to bitplanes, and the bitplanes are scanned and coded from the most significant bitplane (MSB) to the least significant bitplane. The scanning and coding process may stop at any place when a target bit budget is reached. This is equivalent to setting all the remaining bits to zero.

The bitplane coding scheme is flexible in the sense that the importance of image components (coefficients) can be easily adjusted. There are two ways to accomplish this. The first approach emphasizes the important coefficients by shifting them up in the bitplane representation (or equivalently, shifting the unimportant coefficients down). Examples include the MaxShift [86] and the BbBShift [129] methods. The second approach, which we use in this paper, is the bitplane-trimming method demonstrated in Fig. 4.2(b), where the coefficients are trimmed from the bottom such that the bits below certain level are all set to zero. The trimming level is variable based on the importance of the coefficient. This is equivalent to quantizing the coefficient at that level. A regular bitplane coding scheme is applied to the trimmed coefficients, leading to a variable bit allocation over the image space. One advantage of this method is that the decoder does not need to reconstruct the trimming function, and thus no overhead is needed to encode the information about the trimming function.

Let $\mathbf{x}$ and $\mathbf{y}$ be the original and the decoded images respectively. Let $\mathbf{tr}$ denotes the trimming function, i.e., it is a function of the coefficient index that defines the trimming level of all coefficients. Let $\mathbf{TR}$ be the set of all possible trimming functions. Let $\mathbf{C}$ represent the entire image encoding and decoding operator that takes a given

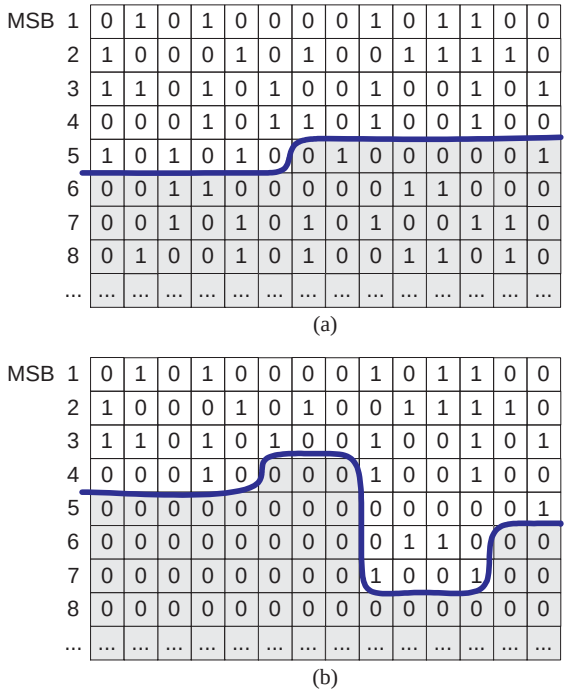

Figure 4.2. (a) Regular bitplane coding. Bitplanes are scanned and coded until a target bitrate is reached. The result is equivalent to setting all bits in the gray region to zero; (b) Bitplane-trimming based coding. The gray region is set to all zero before regular bitplane coding.

original image $\mathbf{x}$, a given bit rate R, and a given trimming function $\mathbf{tr}$ as the input, and creates a decoded image $\mathbf{y}$ as the output:

$$\mathbf{y} = \mathbf{C}(\mathbf{x}, R, \mathbf{tr})\,. \tag{4.1}$$

Let $S_{\mathbf{x},\mathbf{y}}$ denote the operator that computes the SSIM index map between $\mathbf{x}$ and $\mathbf{y}$, and thus $S_{\mathbf{x},\mathbf{y}}(n)$ represents the SSIM index value at spatial location n. Under the maximum of minimal structural similarity criterion, our task is to find the best trimming function that maximizes the minimal value in the SSIM index map $S_{\mathbf{x},\mathbf{y}}$. Mathematically, this can be expressed as

$$\mathbf{tr}_{opt} = \underset{\mathbf{tr}\in\mathbf{TR}}{\operatorname{argmax}}\{\min_{n}[S_{\mathbf{x},\mathbf{C}(\mathbf{x},R,\mathbf{tr})}(n)]\}\,. \tag{4.2}$$

Since the minimal SSIM value in an image has a highly nonlinear relationship with respect to the trimming function $\mathbf{tr}$, it is difficult to find the optimal solution $\mathbf{tr}_{opt}$ analytically. Here we propose an iterative approach given as follows:

1. Initiate the iteration number $i = 0$. For the given target bit rate R, create a constant initial trimming function $\mathbf{t}_0$, i.e., the trimming level (bitplane) is uniform for all coefficients. The trimming level should be high enough such that the bit rate needed to encode all bits above the level is lower than R.
2. Encode and decode the image to create $\mathbf{y}_i = \mathbf{C}(\mathbf{x}, R, \mathbf{tr}_i)$.
3. Compute the SSIM map $S_{\mathbf{x},\mathbf{y}_i}$ between the original and the decoded image.
4. Find the minimal value and location in $S_{\mathbf{x},\mathbf{y}_i}$. If the minimal SSIM (Min-SSIM) value does not change for several iterations, stop the iteration and report $\mathbf{tr}_i$ as the optimized trimming function. Otherwise, update $\mathbf{tr}_i$ by adding one more bits for all the coefficients around the spatial location corresponding to the Min-SSIM value (In wavelet domain, these include a set of neighboring coefficients in all subbands). Let $i = i + 1$ and go to Step 2.

Our algorithm converges only when the bits introduced by $\mathbf{tr}_i$ can not be encoded by C. Figure 4.3 gives a demonstration about how the minimal SSIM value is updated over iterations and how the iterative algorithm converges.

4.1.3 Test

We test the proposed approach using 8bits/pixel gray scale images. The set partitioning in hierarchical trees (SPIHT) [128] algorithm is used as the basic bitplane encoding and decoding operator $\mathbf{C}$. The images are coded to a range of bit rates, from 0.2 to 0.9 bit/pixel using both SPIHT and the proposed method. The Min-SSIM results for the "Lighthouse" image are shown in Fig. 4.4. It can be seen that the proposed method achieves significantly higher Min-SSIM values than SPIHT over a

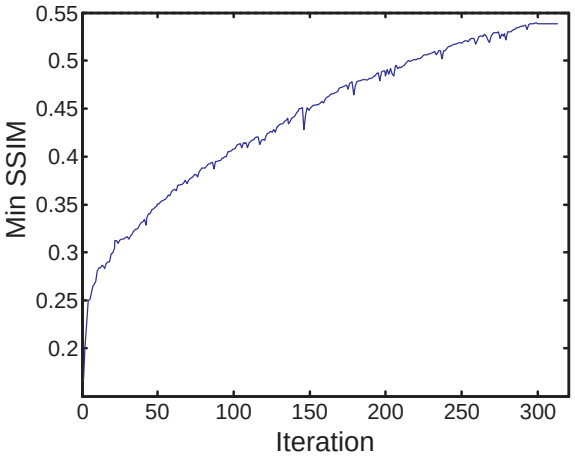

Figure 4.3. Min-SSIM as a function of iteration for the "Lighthouse" image coded at 0.5bit/pixel.

wide range of bit rates. Similar results are also obtained for the other images being tested.

In Fig. 4.5, we compare the coding results of the "Lighthouse" image provided by SPIHT and the proposed algorithms at 0.2bit/pixel, respectively. The SSIM maps indicate that the quality of the image coded by the proposed method is more uniformly distributed over the image space than that of the SPIHT coded image. Since the proposed method mainly focuses on the worst case scenario, the regions with the worst quality in the SPIHT coded image obtain the most improvement. For better visualization, we also enlarged two regions in the images. It can be observed that more detailed structures are exhibited in the image coded by the proposed method.

4.2 Temporal Interpolation based on Temporal Motion Smoothness

4.2.1 Introduction

Temporal frame interpolation has emerged as a potential solution for two problems associated with limitation of the video communication channel and unreliable

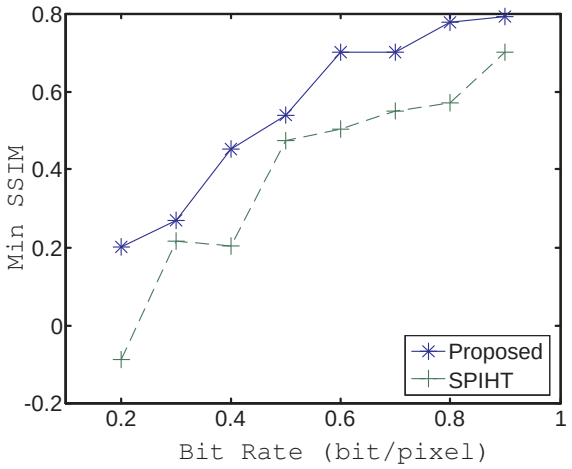

Figure 4.4. Min-SSIM comparison of SPIHT and the proposed method for "Lighthouse" image coded at different bit rates.

packet-based network architectures: First, in order to reduce the bit rate, the video is temporally subsampled by skipping some frames at the transmitter side, which unfortunately can result in significantly reduced temporal resolution, thus possibly causing several possible undesirable artifacts, such as jittering and flickering, during playback. In order to improve the temporal resolution, temporal interpolation is used at receiver side. Second, a missing packet in the video stream results in losing the visual information at macro block or frame level. Temporal interpolation can, to some extend, restore the distorted video by using the redundant temporal information.

In the literature, most of the temporal interpolation methods are based on motion information [130, 131, 132, 133]. The common assumption of these approaches is that the motion among consecutive frames is linearly continuous. Thus, the skipping frames are reconstructed by the motion compensation based on the linear-interpolated motion vectors according to the estimated motion vectors from the available frames. The disadvantage of these methods is that the motion estimation is of high computational complexity, which motivates researchers[130, 131, 132, 133] to employ different

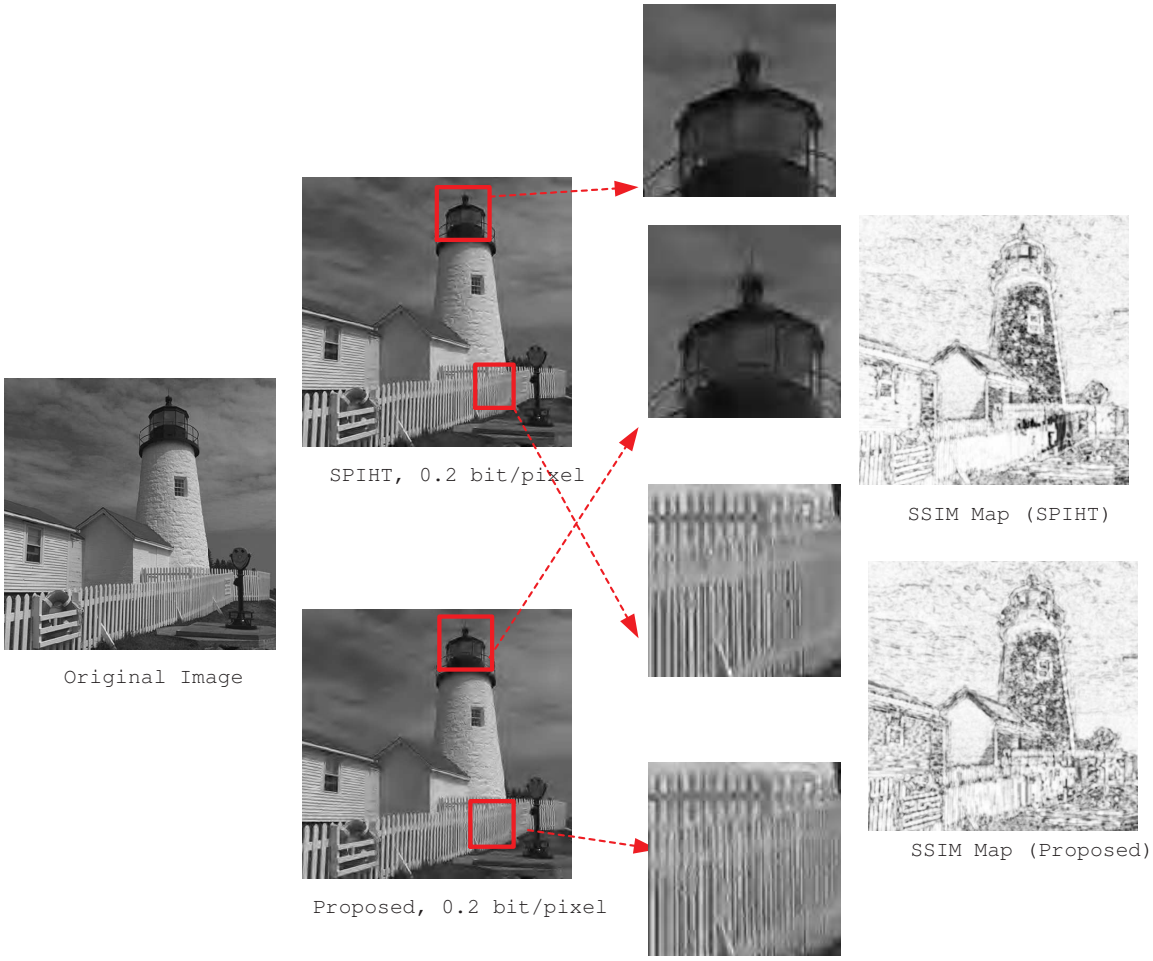

Figure 4.5. Comparison of coding results by SPIHT and the proposed algorithms at 0.2bit/pixel.

motion estimation methods. The basic assumption of continuous motion shared by these methods is reasonable. The prior of temporal motion smoothness is statistically measured and applied to RRVQA in [55]. It is worthwhile to note that temporal motion smoothness [55] is presented by the temporal phase correlation in complex wavelet domain without estimating the explicit motion vector. In this section, we further adopt the prior as an optimization criterion to restore the distorted video sequence with frame dropping. Under the criterion to satisfy the regularity of *temporal*

motion smoothness, the temporal interpolation is implemented in complex wavelet domain. This allows us to avoid the time-consuming motion estimation process, and thus largely reduces the computational complexity of temporal interpolation.

4.2.2 Method of Video Interpolation based on Temporal Motion Smoothness

In Chapter 3, we observe and measure a prior of natural video sequences, *temporal motion smoothness*, which can be presented by the temporal phase correlation function, $L_N(s,p)$. If the image sequences strictly satisfy $N-1$-th order temporal motion smooth, then

$$L_N(s,p) \approx 0 \tag{4.3}$$

Figure 4.6 (b) draws the marginal histogram of the imaginary part of the $L_2(s,p)$ from a natural video, where the high peak around zero indicates the strong statistical preference of the *temporal motion smoothness*. Meanwhile, Figure 4.6 (d) shows how the frame dropping distortion disturbs the prior of temporal motion smoothness.

Inspired by the strong prior of *temporal motion smoothness*, we propose a method to interpolate the dropped frames with criterion to satisfy the prior of *temporal motion smoothness*. The interpolation is carried out in complex wavelet domain by regularizing the temporal phase correlation function, $L_N(s,p)$ so as to avoid the motion estimation.

The diagram of our algorithms for temporal interpolation is shown in Fig 4.7. The distorted video with frame dropping is divided into group of pictures (GOP), for example here a GOP consists of 3 frames. The three continuous frames in the k-th GOP are noted as $h(3(k-1)+1)$, $h(3(k-1)+2)$ and $h(3(k-1)+3)$. For notional convenience, we use h_1 to denote $h(3(k-1)+1)$, h_2 to denote $h(3(k-1)+2)$, and h_3 to denote $h(3(k-1)+3)$.

In order to do temporal interpolation, h_1 and h_3 are transformed to complex wavelet domain as $H_1(s,p)$ and $H_3((s,p))$ at scale s and position p. We assume that the original video sequences satisfies the first order motion smoothness, which means the second order temporal correlation function approximates 0.

$$L_2(s,p) = \ln H_1(s,p) - 2\ln H_2(s,p) + \ln H_3(s,p) \approx 0 \tag{4.4}$$

To satisfy the temporal motion smoothness, $H_2(s,p)$ is interpolated as the linear interpolation in log domain:

$$H_2(s,p) = \exp \frac{\ln H_1(s,p) + \ln H_3(s,p)}{2} \tag{4.5}$$

Eq. (4.5) is inversely transformed to the spatial domain. In additional to this, we also compute the direct temporal interpolation in the spatial domain. Temporal motion smoothness is employed as an objective criterion select one from the two interpolated frames. The criterion is based on the histograms of the imagery part of $L_2(s,p)$ that are computed respectively using these two frames. The one with higher peak around 0 is selected as the interpolated frame.

4.2.3 Test

We artificially generate the distorted video sequences by dropping N frames from $N+1$ continuous frames. For example with $N = 1$,given the original image sequences as: f1 f2 f3 f4 f5 f6..., the distorted image sequences will be f1 f1 f3 f3 f5 f5.... So the dropping rate is 50%. The distorted video sequence is based on "Suzie". Other videos are also tested and have similar results. Our temporal interpolation algorithm is applied to the distorted video. However, it is a difficult task to evaluate the performance of such algorithms. Our subjective evaluation indicates that the

Table 4.1. Performance comparison of our temporal interpolation method

	Distorted	Spatial interpolation	our method
MSE	81.1117	34.7823	33.7877
PSNR	38.5854	40.4240	40.4870
MSSIM	0.8914	0.9279	0.9294

quality of the interpolated video is improved since the motion is temporally smoother and the perceptual annoying jerkiness is reduced. Besides, we propose some objective quantitative measurements based on the original video to test the efficiency of our method.

First, since our optimal criterion is to enforce the temporal motion smoothness which can be clearly indicated by the statistics of temporal phase correlation, we use Kullback-Leibler Distance (KLD) between the statistics of the phase correlation function in interpolated video and that of the original video as an evaluation. Fig. 4.6 (e),(f) show the interpolated frame and the statistics of temporal phase correlation, from which we can see that the KLD is greatly reduced compared with the distorted video, Fig. 4.6 (c). This means that the prior of temporal motion smoothness is strengthened.

Second, we compute the MSE and mean SSIM (MSSIM)index [70] based on the original video. The results of our method are compared with those of the interpolated frame using direct spatial interpolation as $(h_1 + h_3)/2$. Table 4.1 shows the MSE and MSSIM computed from all the interpolated frame. Fig. 4.8 displays the MSE and MSSIM for each interpolated frame. The consistent improvement is observed. Especially Fig. 4.8(b,d) shows that for those frame with larger motion, the improvement is more significant.

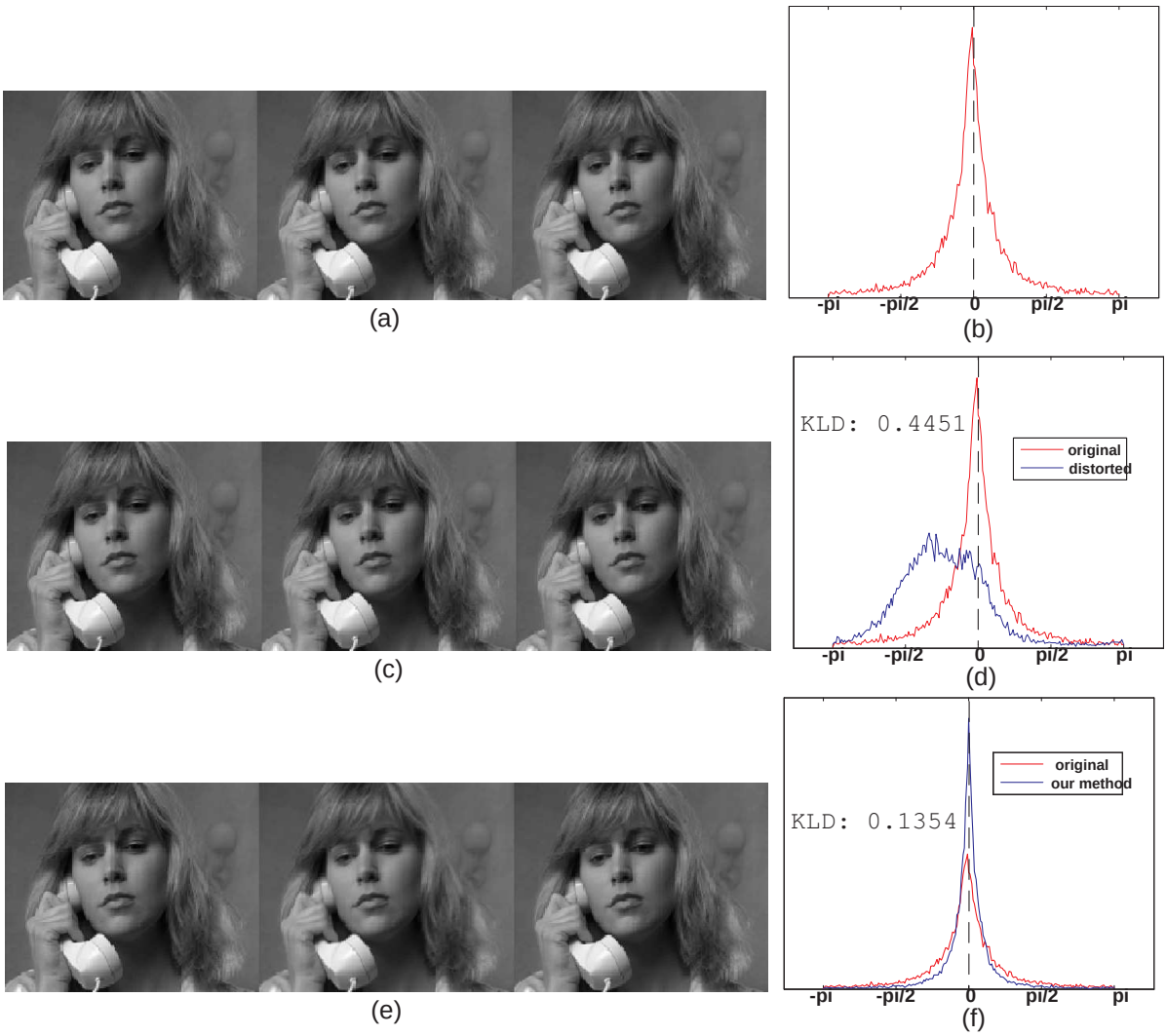

Figure 4.6. The marginal statistics of the imaginary part of $L_2(s,p)$. (a), original video sequences; (b), the marginal histogram of the imaginary part of $L_2(s,p)$ of (a); (c), video sequences with frame dropping; (d), the marginal histogram of the imaginary part of $L_2(s,p)$ of (c); (e), interpolated video sequences using our method; (f), the marginal histogram of the imaginary part of $L_2(s,p)$ of (e).

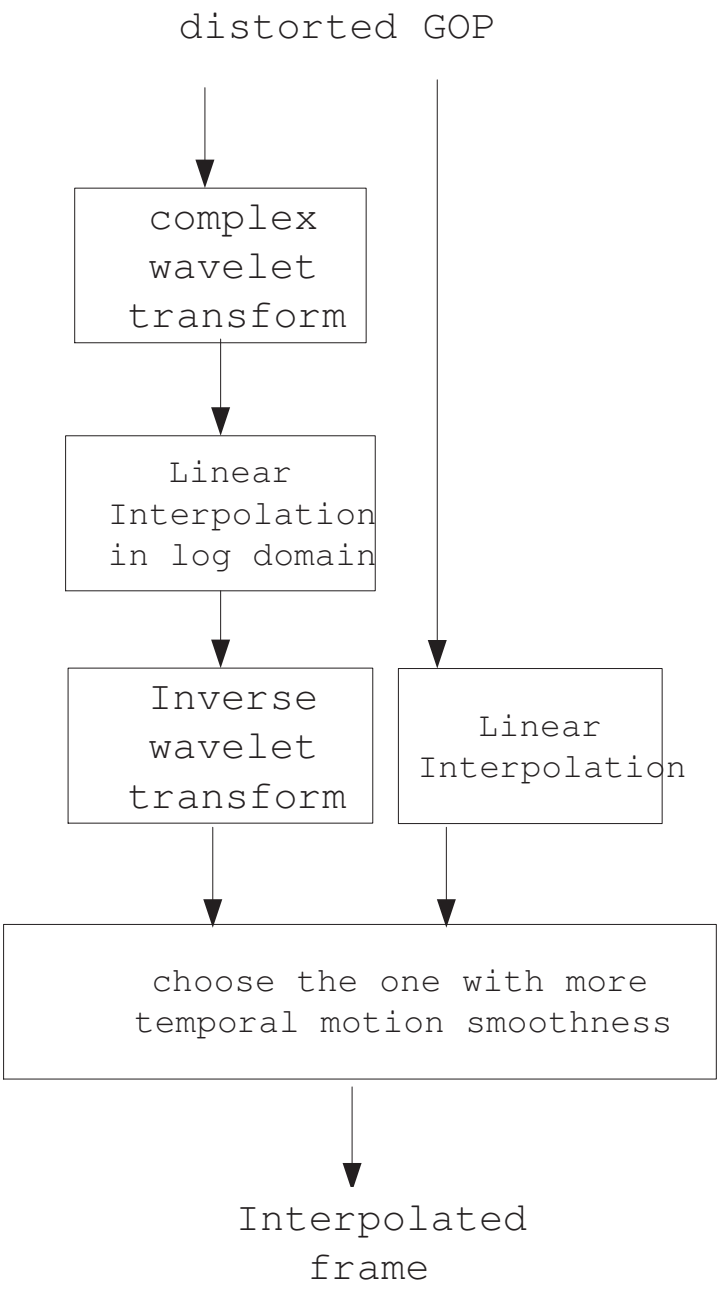

Figure 4.7. The diagram of our interpolation method .

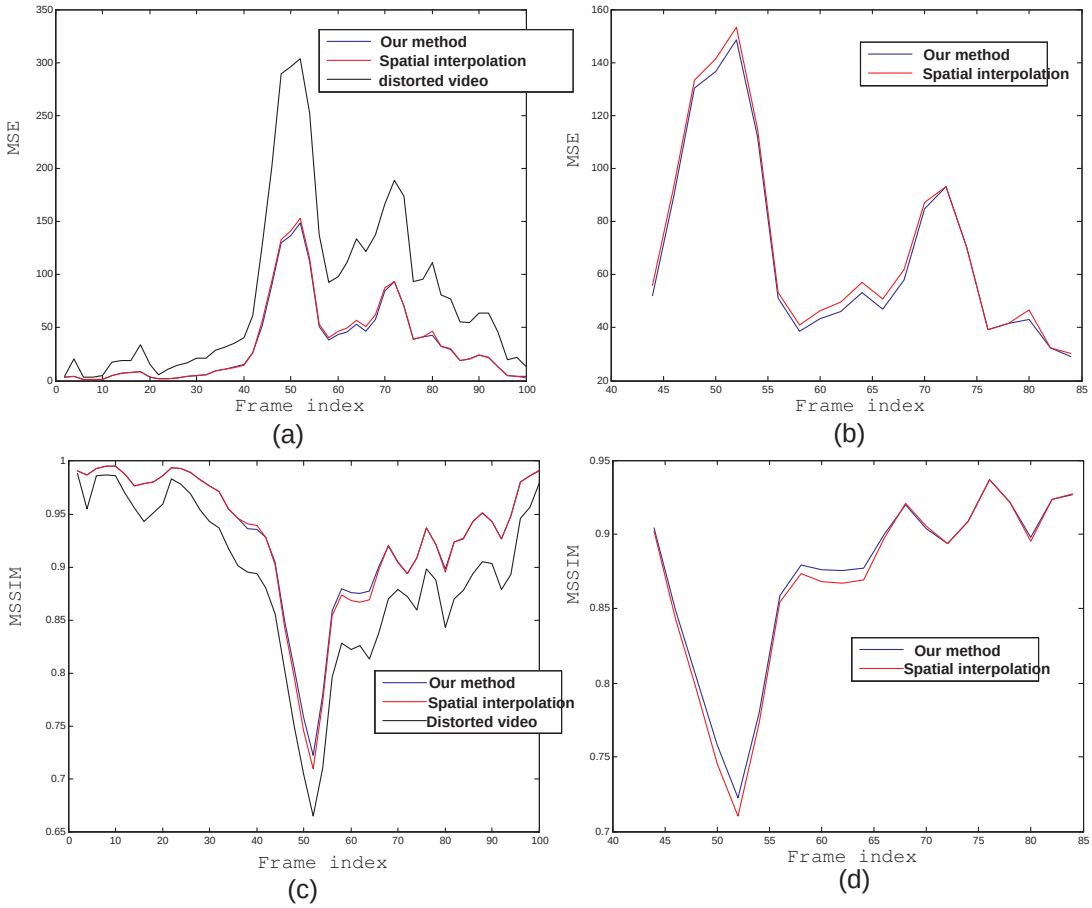

Figure 4.8. Comparison MSE and MSSIM of interpolated frames. (a) MSE for every frame, (b) MSE for frames with large motion, (c) MSSIM for every frame, (d) MSSIM for frames with large motion.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In this dissertation, we propose novel methods for perceptual IQA/VQA and apply the objective metrics to optimize various algorithms including image coding and video interpolation. This chapter draws some conclusions by summarizing the contributions of our work and discusses the directions for future work.

5.1 Conclusions

We propose information theoretic weighting methods for FRIQA and FRVQA respectively by modelling the overall HVS as an information communication channel. In FRIQA, different from other approach using simple Gaussian model [6], the mutual information based on the reference image is measured using an advanced NSS model, GSM [14] at different scales. Under the assumption that the region with more information content will require more computational effort, the mutual information is employed as the weighting coefficients incorporating with the multi-scale spatial SSIM map. Extensive test shows the validity of this information theoretic weighting strategy. We propose a new method to incorporate motion information in an information theoretic framework in FRVQA. Our tests with the VQEG Phase I dataset show that the information theoretic weighting function computed based on our model is effective and consistent in improving VQA algorithms. A distinctive feature of our approach, as compared to the heuristic methods proposed in [40, 30], is that the use of motion information is well justified from the recent findings in psychophysical stud-

ies of human motion perception [3] and is formulated using an information theoretic framework.

The models of the NSS are employed to design new RRIQA/VQA methods in this dissertation. The joint statistics of wavelet coefficients captures the statistical dependence in natural images and accounts for more general distortions than the marginal distribution [51]. We design a RRIQA method to extract the RR features based on the joint statistics. We propose an RRIQA algorithm using statistical features extracted from a divisive normalization-based image representation. The simultaneous perceptual and statistical relevance of this new representation leads to improved performance for image quality assessment. A novel statistical model of natural image sequences, temporal motion smoothness is investigated. This statistical regularity can be represented by the distribution of the temporal phase correlation. We observe that the strong prior of temporal motion smoothness is spoiled by typical unnatural image distortions. With an empirical probability model to describe the prior of temporal motion smoothness, we show the potential to design a RRVQA algorithm based on this prior. To sum up, the advantages of our proposed RRIQA/VQA methods include: 1) the data rates of RR features extracted from the NSS models are low; 2) they are general-purpose models without any assumption about the image distortion types.

It is worthwhile to point out that both our FR and RR quality assessment methods extensively use NSS. The modelling of HVS and the modelling of NSS are believed to be dual problems [9] because the HVS is hypothesized to highly adapted to the statistics of the natural surroundings during the long process of evolution. Thus, studying the NSS provides an indirect but powerful way to investigate the HVS. Since the ultimate goal of perceptual quality assessment is to stimulate the performance of the HVS, it is no wonder the NSS plays an important role to design effective

algorithms, which is verified by the results in this dissertation. Appropriate use of NSS models may also help to reduce the complexity of the image quality assessment algorithms. For example, some RR features of low bit rate in our RRIQA/VQA methods are directly extracted from the parameters of the statistical models of NSS.

The objective quality assessment metrics are applied as optimization criterions to perceptual image coding and video interpolation. We propose a novel perceptual coding method that incorporates a maximum of minimal SSIM criterion into bit-plane coding through an iterative optimization process. The test results show that the proposed method significantly improves the worst case scenario (worst quality region in the image) and the coded image appears to have more uniform quality over the image space. Although the proposed scheme is currently implemented with the SPIHT algorithm only, the general design principle may be generalized for other bit-plane coding schemes.We propose a video interpolation algorithm based on a strong prior, temporal motion smoothness, in the natural image sequence. Different from other video interpolation approaches to explicitly estimate the motion vectors, our interpolation method is carried out using the phase information in complex wavelet domain since the prior of temporal motion smoothness is clearly represented by the distribution of temporal phase correlation functions. This significantly reduces the computational complexity by avoiding the time-consuming motion estimation process. Besides the subjective evaluation, we objectively measure the temporal phase correlation function, MSE and SSIM to validate our method and observe consistent improvement.

5.2 Future work

The information theoretic weighting strategy for FRIQA proposed in Chapter 2 may be improved/extended in many ways. First, although the Gaussian scale

mixture model achieves certain amount of success to describe the natural image, it is essentially a limited local description that may not precisely capture natural image structures. More advanced models, for example, a global, fields of Gaussian scale mixtures [134] model, could be adopted. Second, the extracted information from neighboring subbands at the same scale is implicitly supposed to be equally important. But the HVS is generally believed to have different sensitivity for different spatial frequency. The CSF may be considered to design weights for these subbands. Third, the mutual information of each subband is extracted from the Gaussian channel which is assumed to be independent. This assumption may not be accurate since there is significant dependency among the subbands [91]. It needs further research effort to precisely measure the mutation information by accounting for the dependencies or redundances. Finally, the information theoretic weighting is incorporated with the spatial SSIM map. It can also be applied to other metrics including CW-SSIM[38] and VIF[7]. We propose an algorithm of information theoretical weighting based on a statistical model of human visual speed perception for FRVQA. Future work may include: First, there might be better ways to combine the information content and the perceptual uncertainty measures. Second, the computation of local image contrast and the estimation of motion vectors may be improved. For example, we frequently observe instabilities in the current optical flow-based motion estimation algorithm, especially in the video frames with large background motion. This implies that more robust motion estimation method is needed in the existence of noise and large motion. Third, the sophistication and the high-level nature of the proposed model make its parameters difficult to calibrate. More careful psychovisual studies are still needed. Fourth, the weighting function computed based on our model is effective and consistent in improving the performance of VQA algorithms in all the tests we have done so far (with MSE/PSNR and SSIM). Other VQA algorithms may

also be included to further validate the model. Finally, the general idea of quality map weighting does not constrain itself to be used for full-reference VQA only, as being tested in Chapter 2 (Note that both the MSE/PSNR and the SSIM calculations require access to the original video sequence as a reference). If a no-reference method is available that can provide us with a quality map without using any reference video, the same weighting approach is also applicable. Such no-reference or blind VQA systems are highly desirable in the real world and are yet to be developed in the future.

In Chapter 3, we introduce several RRIQA/VQA algorithms which share a similar framework because all of them are based on the statistical models of the NSS. In order to improve/extend these methods, several further questions may be asked. First, while the statistical features used in the proposed algorithm seem to be perceptually relevant and useful, is there any better means to combine them into a single scalar quality measure of the distorted image? Second, can the statistical models being used sufficiently represent the nature of images? One of our RRIQA methods is based on the DNT which can effectively reduce the variance dependency. But there are many other types of dependencies between neighboring wavelet coefficients that are still missing, for instance, local phase coherence[110] . Is there any efficient way to incorporate these dependencies as well? Third, using the proposed RRIQA/VQA measure, together with the statistical properties (RR features) about the perfect-quality original image, can we design image quality enhancement method that can correct or improve the quality of the distorted image being evaluated? Finally, since the proposed RRIQA/VQA method is relevant to the quantification of the naturalness of images and does not use any knowledge about image distortion types, would it be possible to further develop it into a general-purpose NRIQA/VQA method?

The perceptual image coding method in Chapter 4 uses an iterative approach to allocate the bits under the criterion to maximize the minimal SSIM index. This increases the complexity at the encoder. It is desirable to estimate a nearly-optimal trimming function based on the bit rate and image content before initial iteration. The proposed video interpolation method is developed based on the second order temporal phase correlation. It is observed that the higher order temporal phase correlation function shows stronger temporal motion smoothness. Therefore, higher order temporal phase correlation functions may be used to impose the prior of temporal motion smoothness.

The general-purpose NRIQA/VQA methods aim to predict the quality based on only distorted signals without any information from original signals and the types of distortions. The problem is extremely difficult but also very important in real world, for example, in visual communications where the channel is open to any distortions and the quality of the received signal is required to be monitored. Although RRIQA/VQA can be a compromising solution, the RR features need to be transmitted so that they are also subjective to distortions. We believe the modelling of NSS has the potential to be used in the design of the general-purpose NRIQA/VQA algorithms. This may lead us to a better understanding of the biological visual system and the statistics of the natural images. In Chapter 1, we proposed a Bayesian framework for IQA. If we can model all related aspects of the HVS, including its built-in knowledge about the visual environment by the likelihood and prior functions , then the posterior based on likelihood and prior models could be used to design computationally tractable algorithms of general-purpose NRIQA/VQA.

REFERENCES

[1] Z. Wang, "Demo images and free software for 'a universal image quality index'," *http://anchovy.ece.utexas.edu/~zwang/research/quality_index/demo.html*, 2001.

[2] A. Oliva, 2007, http://cvcl.mit.edu/hybrid_gallery/monroe_einstein.html.

[3] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature Neuroscience*, vol. 9, pp. 578–585, Mar. 2006.

[4] K. Seshadrinathan and A. C. Bovik, "New vistas in image and video quality assessment," *Proc. SPIE Conf. on Human Vision and Electronic Imaging XII*, vol. 6492, pp. 1–13, Jan. 2007.

[5] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, vol. 2, Nov 2003, pp. 1398– 1402.

[6] Z. Wang and X. L. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. IEEE Int. Conf. Image Proc.*, Oct. 2006, pp. 2945 – 2948.

[7] H. R. Sheikh and A. C. Bovik, "Imageinformation and visual quality," *IEEE Trans. Image Processing*, vol. 15, pp. 430–444, Feb. 2006.

[8] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise-ratio for natural images," *IEEE Trans. Image Processing*, vol. 16, no. 9, pp. 2284–2298, Sept. 2007.

[9] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment.* New York: Morgan and Claypool Publishers, 2006.

[10] H.R. Wu and K.R. Rao, ed., *Digital Video Image Quality and Perceptual Coding.* The Taylor and Francis, 2006.

[11] E. A. Bovik, *Handbook of Image and Video Processing, Second Edition.* Elsevier Academic Press, 2002.

[12] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*, A. B. Watson, Ed. the MIT press, 1993, pp. 207–220.

[13] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? - a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[14] J. Portilla, V. S. M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. Signal Processing*, vol. 12, pp. 1331–1338, Nov. 2003.

[15] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, "Probability distributions of optical flow," in *IEEE Inter. Conf. Computer Vision and Pattern Recognition*, June 3-6 1991, pp. 310–315.

[16] N. Vasconcelos and A. Lippman, "A unifying view of image similarity," in *Proc. IEEE Int. Conf. Pattern Recognition*, vol. 1, Sept. 2000, pp. 38–41.

[17] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, "Subjective quality assessment live database," 2001, http://live.ece.utexas.edu/research/Quality/.

[18] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Digital images and human vision*, A. B. Watson, Ed. Cambridge, Massachusetts: The MIT Press, 1993, pp. 179–206.

[19] J. Lubin, "A visual discrimination mode for image system design and evaluation," in *Visual Models for Target Detection and Recognition*, E. Peli, Ed. Singapore: World Scientific Publishers, 1995, pp. 207–220.

[20] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," in *Society for Information Display Digest of Technical Papers*, vol. XXIV, 1993, pp. 946–949.

[21] A. Pons, J. Malo, J. M. Artigas, and P. Capilla, "Image quality metric based on multidimensional contrast perception models," *Displays*, vol. 20, pp. 93–110, 1999.

[22] JNDmetrix Technology at Sarnoff Corporation. http://www.sarnoff.com/.

[23] C. H. Chou and Y. C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 5, no. 6, pp. 467–476, Dec. 1995.

[24] S. A. Karunasekera and N. G. Kingsbury, "A distortion measure for image artifacts based on human visual sensitivity," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 5, pp. 117–120, Apr. 1994.

[25] D. J. Heeger and P. C. Teo, "A model of perceptual image fidelity," *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, pp. 343–345, Oct. 1995.

[26] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. SPIE*, vol. 2668, San Jose, LA, 1996, pp. 450–461.

[27] A. B. Watson, "Toward a perceptual video quality metric," in *Proc. SPIE Conf. on Human Vision and Electronic Imaging III*, vol. 3299, Jan. 1998, pp. 139–147.

[28] S. Winkler, "A perceptual distortion metric for digital color video," *Proc. SPIE*, vol. 3644, pp. 175–184, 1999.

[29] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual qualtiy measure using a spatio-temporal model of the human visual system," *Proc. SPIE*, vol. 2668, pp. 450–461, 1996.

[30] Z. K. Lu, W. Lin, X. K. Yang, E. P. Ong, and S. S. Yao, "Modeling visual attention's modulatory afterreffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Processing*, vol. 14, pp. 1928–1942, Nov. 2005.

[31] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessement," *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan 2005.

[32] T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Proc Second Edition.*, A. Bovik, Ed. Academic Press, 2005.

[33] S. Winkler, *Digital Video Quality - Vision Models and Metrics.* The John Wiley and Sons Press, 2005.

[34] G. E. Legge, "Sustained and transient mechannisms in human vision: Temporal and spatial properties," *Vision Research*, vol. 18, no. 1, pp. 69–81, 1978.

[35] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 4, May 2002, pp. 3313–3316.

[36] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Apr. 2000, http://www.vqeg.org/.

[37] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature Neuroscience*, vol. 9, no. 4, pp. 578–585, April 2006.

[38] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, Mar. 2005, pp. 573– 576.

[39] A. C. Brooks, X. Zhao, and T. N. Pappas, "Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions," *IEEE Trans. Image Processing*, vol. 17, pp. 1261–1273, Aug. 2008.

[40] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *IEEE Trans. Signal Processing*, vol. 19, pp. 121–132, Feb. 2004.

[41] Q. Li and Z. Wang, "Video quality assessment by incorporating a motion perception model," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, Sept. 2007, pp. 173–176.

[42] Z.Wang and Q.Li, "Video quality assessment using a statistcial model of human visual speed perception," *J. Opt. Soc. Am. A*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[43] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, Apr. 2007, pp. 869–872.

[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley-Interscience, 1991.

[45] S. Wolf and M. H. Pinson, "Spatio-temporal distortion metrics for in-service quality monitoring of any digital video system," *Proc. SPIE*, vol. 3845, pp. 266–277, Sept. 1999.

[46] I. P. Gunawan and M. Ghanbari, "Reduced reference picture quality estimation by using local harmonic amplitude information," in *Proc. London Communicationa Symposium*, Sept. 2003, pp. 137–140.

[47] T. M. Kusuma and H.-J. Zepernick, "A reduced-reference perceptual quality metric for in-service image quality assessment," in *Joint First Workshop on Mobile Future and Symposium on Trends in Communications*, Oct. 2003, pp. 71–74.

[48] S. Wolf and M. Pinson, "Low bandwidth reduced reference video quality monitoring system," in *Int. Workshop Video Proc. and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, Jan. 2005.

[49] P. Le Callet, C. Viard-Gaudin, and D. Barba, "Continuous quality assessment of MPEG2 video with reduced reference," in *Int. Workshop Video Proc. and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, Jan. 2005.

[50] M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 3, Sept. 2003, pp. 185–188.

[51] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Processing*, vol. 15, no. 6, pp. 1680–1689, 2006.

[52] J. Huang and D. Mumford, "Statistics of natural images and models," in *Proc. IEEE Int. Conf Computer Vision and Pattern Recognition*, vol. 1, June 1999, p. 547.

[53] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Adv. Neural Information Processing Systems*, vol. 12. Cambridge, MA: MIT Press, May 2001, pp. 855–861.

[54] Q. Li and Z. Wang, "General-purpose reduced-reference image quality assessment based on perceptually and statistically motivated image representation," in *Proc. IEEE Int. Conf. Image Proc.*, Oct. 2008, pp. 1192 – 1195.

[55] B. Hiremath, Q. Li, and Z. Wang, "Quality-aware video," in *Proc. IEEE Int. Conf. Image Proc.*, Oct. 2007, pp. 469–472.

[56] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 1, Sept. 2002, pp. 477–480.

[57] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 3, Sept. 2000, pp. 981–984.

[58] K. T. Tan, M. Ghanbari, and D. E. Pearson, "An objective measurement tool for MPEG video quality," *Signal Processing*, vol. 70, no. 3, pp. 279–294, Nov. 1998.

[59] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters*, vol. 4, pp. 317–320, Nov. 1997.

[60] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Blind quality assessment for JPEG2000 compressed images," in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, vol. 2, Nov. 2002, pp. 1735–1739.

[61] X. Li, "Blind image quality assessment," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 1, Sept. 2002, pp. 449–452.

[62] "Itu-r recommendation bt.500-10: Methodology for the subjective assessment of the quality of television pictures," Mar. 2002.

[63] Video Quality Experts Group: http://www.vqeg.org/.

[64] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Processing*, vol. 15, pp. 3440–3451, Nov. 2006.

[65] P. Le Callet and F. Autrusseau, "Subjective quality assessment irccyn/ivc database," 2005, http://www.irccyn.ec-nantes.fr/ivcdb/.

[66] "Subjective quality assessment toyama database," 2005, http://www.irccyn.ec-nantes.fr/ touranch/ToyamaDatabase.rar.

[67] "Subjective quality assessment database of the cornell visual communications lab," http://foulard.ece.cornell.edu/.

[68] VQEG, "Subjective video quality assessment database," ftp://ftp.crc.ca/crc/vqeg.

[69] Z. Wang, Q. Li, and X. Shang, "Perceptual image coding based on a maximum of minimal structural similarity criterion," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, Oct. 2007, pp. 121–124.

[70] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity,," *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[71] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–1216, May 2001.

[72] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Information Theory*, vol. 38, pp. 587–607, Mar. 1992.

[73] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*, B. Furht and O. Marques, Eds. CRC Press, 2003.

[74] B. A. Wandell, *Foundations of Vision.* Sinauer Associates, Inc., 1995.

[75] K. Seshadrinathan and A. C. Bovik, "An information theoretic video quality metric based on motion models," in *Third Inter. Workshop on Video Proc. and Quality Metrics for Consumer Electronics*, Jan 2007.

[76] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, "Motion illusions as optimal percepts," *Nature Neuroscience*, vol. 5, pp. 589–604, May 2002.

[77] F. Hürlimann, D. C. Kiper, and M. Carandini, "Testing the bayesian model of perceived speed," *Vision Research*, vol. 42, pp. 2253–2257, Sept. 2002.

[78] E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–1216, May 2001.

[79] R. Raj, W. S. Geisler, R. A. Frazor, and A. C. Bovik, "Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy," *Vision Research*, vol. 22, pp. 2039–2049, Oct. 2005.

[80] J. Najemnik and W. S. Geisler, "Optimal eye movement strategies in visual search," *Nature*, vol. 434, pp. 387–391, Mar. 2005.

[81] M. J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, vol. 63, pp. 75–104, 1996.

[82] T.Vlachos, "Simple method for estimation of global motion paramenters using spares translational miton vector fields," *Electronics letters*, vol. 34, pp. 90–91, Jan. 1998.

[83] E. Peli, "Contrast in complex images," *Journal of Optical Society of America*, vol. 7, no. 10, pp. 2032–2040, Oct. 1990.

[84] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. SPIE*, vol. 2179, Nov. 1994, pp. 127–141.

[85] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard*. Kluwer Academic Publishers, 1992.

[86] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Kluwer Academic Publishers, 2001.

[87] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[88] S. Lyu and E. P. Simoncelli, "Statistically and perceptually motivated nonlinear image representation," *Proc. SPIE Conf. on Human Vision and Electronic Imaging XII,*, Jan. 2007.

[89] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, W. A. Rosenblith, Ed. MIT Press, 1961, pp. 217–234.

[90] E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–1216, May 2001.

[91] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature: Neuroscience*, vol. 4, no. 8, pp. 819–825, Aug. 2001.

[92] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Visual Neural Science*, vol. 9, pp. 181–198, Aug. 1992.

[93] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vision Research*, vol. 38, no. 5, pp. 743–761, Mar. 1998.

[94] D. L. Ruderman, "The statistics of natural images, network: Computation in neural systems," *Network: Computation in Neural Systems*, vol. 5, pp. 517–548, Nov. 1996.

[95] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli, "Non-linear image representation for efficient perceptual coding," *IEEE Trans. Image Processing*, vol. 15, no. 1, pp. 68–80, Jan. 2006.

[96] J. Portilla and E. P. Simoncelli, "Image restoration using Gaussian scale mixtures in the wavelet domain," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, Barcelona, Spain, Sept. 2003, pp. 965–968.

[97] S. G. Mallat, "Multifrequency channel decomposition of images and wavelet models," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 2091–2110, Dec. 1989.

[98] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc 3rd IEEE Int'l Conf on Image Proc*, vol. I. Lausanne: IEEE Sig Proc Society, September 16-19 1996, pp. 379–382.

[99] M. J. Wainwright, "Visual adaptation as optimal information transmission," *Vision Research*, vol. 39, pp. 3960–3974, 1999.

[100] J. Foley, "Human luminance pattern mechanisms: Masking experiments reuqire a new model," *Journal of Optical Society of America*, vol. 11, no. 6, pp. 1710–1719, 1994.

[101] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *Journal of Optical Society of America*, vol. 14, no. 9, pp. 2379–2391, 1997.

[102] P. Corriveau, *et al.*, "Video quality experts group: Current results and future directions," *Proc. SPIE Visual Comm. and Image Processing*, vol. 4067, June 2000.

[103] D. W. Dong and J. J. Atick, "Statistics of natural time-varying images," *Network: Computation in Neural Systems*, vol. 6, pp. 345–358, 1995.

[104] J. H. van Hateren and D. L. Ruderman, "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex," *Proc Royal Society: Biological Sciences*, vol. 265, pp. 2315–2320, Dec. 1998.

[105] B. A. Olshausen, "Learning sparse, overcomplete representations of time-varying natural images," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 1, Sept. 2003, pp. 41–44.

[106] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Computing Surveys*, vol. 27, no. 3, pp. 433–467, Sept. 1995.

[107] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, "Probability distributions of optical flow," in *Proc Conf on Computer Vision and Pattern Recognition*. Maui, Hawaii: IEEE Computer Society, June 3-6 1991, pp. 310–315.

[108] F. Dufaux and F. Moscheni, "Motion estimation techniques for digital TV: a review and a newcontribution," *Proceedings of the IEEE*, vol. 83, no. 6, pp. 858–876, June 1995.

[109] I. Selesnick, R. Baraniuk, and N. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, Nov. 2005.

[110] Z. Wang and E. P. Simoncelli, "Local phase coherence and the perception of blur," in *Adv. Neural Information Processing Systems (NIPS03)*, vol. 16. Cambridge, MA: MIT Press, May 2004.

[111] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. of the IEEE*, vol. 69, pp. 529–541, May 1981.

[112] M. C. Morrone and D. C. Burr, "Feature detection in human vision: A phase-dependent energy model," *Proc. R. Soc. Lond. Biological Sciences*, vol. 235, pp. 221–245, Dec. 1988.

[113] P. Kovesi, "Phase congruency: A low-level image invariant," *Psych. Research*, vol. 64, pp. 136–148, 2000.

[114] D. J. Fleet, "Phase-based disparity measurement," *CVGIP: Image Understanding*, vol. 53, no. 2, pp. 198–210, Mar. 1991.

[115] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *Int'l J Computer Vision*, vol. 5, no. 1, pp. 77–104, Aug. 1990.

[116] J. F. A. Magarey and N. G. Kingsbury, "Motion estimation using a complex-valued wavelet transform," *IEEE Trans. Signal Proc.*, vol. 46, no. 4, pp. 1069–1084, Apr. 1998.

[117] J. Portilla and E. P. Simoncelli, "Texture modeling and synthesis using joint statistics of complex wavelet coefficients," *IEEE Workshop on Statistical and Computational Theories of Vision*, vol. 22, June 1999.

[118] J. Daugman, "Statistical richness of visual phase information: update on recognizing persons by iris patterns," *Int'l J Computer Vision*, vol. 45, no. 1, pp. 25–38, Oct. 2001.

[119] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[120] N. I. Fisher, *Statistical analysis of circular data*. New York: Cambridge University Press, 2000.

[121] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163–172, Feb. 2004.

[122] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Proc.*, A. Bovik, Ed. Academic Press, 2005.

[123] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, May 1989, pp. 1945–1948.

[124] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.

[125] D. M. Chandler and S. S. Hemami, "Additivity models for suprathreshold distortion in quantized wavelet-coded images," in *Proc. SPIE Conf. on Human Vision and Electronic Imaging VII*, vol. 4662, Jan. 2002, pp. 742–753.

[126] W. Zeng, S. Daly, and S. Lei, "Visual optimization tools in JPEG 2000," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, Oct. 2000, pp. 37–40.

[127] J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[128] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 6, no. 3, pp. 243–250, June 1996.

[129] Z. Wang and A. C. Bovik, "Bitplane-by-bitplane shift (BbBShift) - A suggestion for JPEG 2000 region of interest coding," *IEEE Signal Processing Letters*, vol. 9, no. 5, pp. 160–162, May 2002.

[130] A. M. Tourapis, H. Y. Cheong, M. L. Liou, , and O. C. Au, "Temporal interpolation of video sequences using zonal based algorithms," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 3, Oct. 2001, pp. 895–898.

[131] G. Schuster and A. Katsaggelos, "An optimal quadtree-based motion estimation and motion-compensated interpolation scheme for video compression," *IEEE Trans. Image Processing*, vol. 7, no. 11, pp. 1505–1523, Nov. 1998.

[132] M. Chahine and J. Konrad, "Motion-compensated interpolation using trajectories with acceleration," *Electronic Imaging Science and Technology, Digital Video Compression: Algorithms and Technologies*, vol. 2419, pp. 124–131, Feb. 1995.

[133] C. Cafforio, F. Rocca, and S. Tubaro, "Motion compensated image interpolation," *IEEE Transactions on Communications*, vol. 38, no. 2, pp. 215–222, Feb. 1990.

[134] S. Lyu and E. P. Simoncelli, "Statistical modeling of images with fields of gaussian scale mixtures," in *Adv. Neural Information Processing Systems (NIPS06)*, vol. 19. Cambridge, MA: MIT Press, May 2007.

BIOGRAPHICAL STATEMENT

Qiang Li was born in HanDan, China, in 1978. He received his B.S. and M.S. degrees from Beijing institute of technology in 2000 and 2003 respectively. Now he is a Ph.D student in university of Texas at Arlington. His current research interest is in the area of perceptual image and video quality assessment with applications. He interned in Qualcomm in spring, 2008. He won the "IBM Student Paper Award" at International Conference of Image Processing, 2008.