SSIM-Inspired Quality Assessment, Compression, and Processing for Visual Communications

by

Abdul Rehman

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Objective Image and Video Quality Assessment (I/VQA) measures predict image/video quality as perceived by human beings - the ultimate consumers of visual data. Existing research in the area is mainly limited to benchmarking and monitoring of visual data. The use of I/VQA measures in the design and optimization of image/video processing algorithms and systems is more desirable, challenging and fruitful but has not been well explored. Among the recently proposed objective I/VQA approaches, the structural similarity (SSIM) index and its variants have emerged as promising measures that show superior performance as compared to the widely used mean squared error (MSE) and are computationally simple compared with other state-of-the-art perceptual quality measures. In addition, SSIM has a number of desirable mathematical properties for optimization tasks. The goal of this research is to break the tradition of using MSE as the optimization criterion for image and video processing algorithms. We tackle several important problems in visual communication applications by exploiting SSIM-inspired design and optimization to achieve significantly better performance.

Firstly, the original SSIM is a Full-Reference IQA (FR-IQA) measure that requires access to the original reference image, making it impractical in many visual communication applications. We propose a general purpose Reduced-Reference IQA (RR-IQA) method that can estimate SSIM with high accuracy with the help of a small number of RR features extracted from the original image. Furthermore, we introduce and demonstrate the novel idea of partially repairing an image using RR features. Secondly, image processing algorithms such as image de-noising and image super-resolution are required at various stages of visual communication systems, starting from image acquisition to image display at the receiver. We incorporate SSIM into the framework of sparse signal representation and non-local means methods and demonstrate improved performance in image de-noising and super-resolution. Thirdly, we incorporate SSIM into the framework of perceptual video compression. We propose an SSIM-based rate-distortion optimization scheme and an SSIM-inspired divisive optimization method that transforms the DCT domain frame residuals to a perceptually uniform space. Both approaches demonstrate the potential to largely improve the rate-distortion performance of state-of-the-art video codecs. Finally, in

real-world visual communications, it is a common experience that end-users receive video with significantly time-varying quality due to the variations in video content/complexity, codec configuration, and network conditions. How human visual quality of experience (QoE) changes with such time-varying video quality is not yet well-understood. We propose a quality adaptation model that is asymmetrically tuned to increasing and decreasing quality. The model improves upon the direct SSIM approach in predicting subjective perceptual experience of time-varying video quality.

Acknowledgments

First and foremost, I thank Allah Almighty for giving me grace and privilege to pursue my dream and His blessing in giving me life, health and intelligence. It is through His mercy that I can fathom the true meaning of life and the purpose to do research and explore the world, not merely for myself, but rather to contemplate and share the beauty of His creation and, hopefully, to serve humanity.

In particular, I wish to express my sincere appreciation and gratitude to my supervisor, Dr. Zhou Wang. He is undoubtedly a superb advisor and I am very fortunate to have learned under his guidance. He gave me the freedom to explore on my own, and at the same time the direction to a new road when I faced a deadlock. Dr. Wang has always been patient, quick to compliment and slow to criticize. His observations and comments helped me establish the overall direction of my research and to move forward with my studies in depth. It has been my pleasure and a tremendous opportunity to work with Dr. Wang and a talented team of researchers.

Thank you to the members of my committee, Dr. Oleg Michailovich, Dr. En-Hui Yang, Dr. David Clausi, and Dr. Z. Jane Wang. I offer my sincere gratitude for your help, cooperation, and reviewing this thesis.

The University of Waterloo offers a rich and productive environment to explore new ideas. I am grateful to have had the chance to study amidst a greatly supportive community and be surrounded by wonderful colleagues. I would also like to thank my friends, Adeel Akhtar and Shafiq-ur-Rahman for their support throughout the course of my studies.

My family has been an integral part of my academic life. Despite being many thousands of miles away, their constant courage and support has carried me through these many years as a student. A very special word of gratitude goes to my sisters, Asma and Fatima, and my brother, Abdullah. Most of all, my heartfelt appreciation goes to my mother and father, whose patience and understanding gave me the strength to persevere. I will forever be indebted to my mother for her constant love, encouragement and prayers. You have made me the man I am today and I hope that this work makes you proud.

I lovingly dedicate this thesis to my precious wife, Aysha, whose help and encouragement saw me through this degree. Through her love, patience and unwavering belief in

me, I have been able to complete this long dissertation journey. There are no words that can express my gratitude and appreciation for all you have done and been for me. Thank you and I love you with all my heart.

Dedicated to the best Mom in the whole wide world!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In recent years, images and videos have become integral parts of our lives. The current applications range from casual documentation of events and visual communication, to the more serious surveillance and medical fields. This expansion has led to an ever-increasing demand for accurate and visually pleasing visual content. Over the past years, we observed an exponential increase in the demand for image and video services. Every minute, over 48 hours of video content is being uploaded to YouTube and over two million video clips are being downloaded [195]. These constitute only about 40% of the exponentially growing Internet video streaming data, among which a significant portion is accessed via mobile devices. Video traffic has emerged as the dominant traffic in today's Internet and it is predicted to increase much faster than other applications in the years to come. Particularly, streaming traffic (which consists of Live and video-on-demand streaming, but excludes the downloads of video content like P2P) counts for the biggest share in the whole video traffic (predicted to count for more than 90% Internet traffic by 2016 according to source from Cisco [34]). By 2017, mobile video will represent 66% of all mobile data traffic [35]. Figure 1.1 shows the trend of mobile video data growth over wireless networks. Since mobile video content has much higher bit rates than other mobile content types, mobile video

will generate much of the mobile traffic growth through 2017. Mobile video will grow at a Compound Annual Growth Rate (CAGR) of 75% between 2012 and 2017, the highest growth rate of any mobile application category that we forecast. Of the 11.2 Exabytes per month crossing the mobile network by 2017, 7.4 Exabytes will consist of video [35]. Recent advances in video capturing and display technologies will increase the presence of high resolution and quality contents in digital video coding applications. The storage space and bandwidth capacity involved in visual content production, storage, and delivery will be stressed to fulfill the new resolution and quality requirements. The following are among the main challenges the technology is increasingly encountering:

- The networks in service are not designed to accommodate the current traffic trends. In practice, the multimedia content delivered over the networks suffers from various kinds of distortions on the way to its destination. It is important for service providers to be able to identify and quantify the quality degradations in order to maintain the required Quality of Service (QoS). This situation gives rise to the desire for accurate and efficient perceptual image and video quality assessment algorithms that can estimate the subjective quality of the visual content at the receiver side under various kinds of distortions;

- The volume of digital video data is notoriously huge. Transmission of raw video data over communication channels of limited bandwidth is implausible. Video encoders are primarily characterized in terms of the throughput of the channel and perceived distortion of the reconstructed video. The main task of a video encoder is to convey the sequence of images with minimal possible perceived distortion within available bit rate. Distortion model used by a video encoder should ideally be in perfect coherence with the actual "receiver" of video content, the Human Visual System (HVS). The current video compression techniques do not use the distortion models which correlate well with subjective scores and as a result optimize for the wrong quality measure. The compression performance of video encoders can be improved significantly by using a Video Quality Assessment (VQA) method that can help the encoder to squeeze video data to just the "information" relevant to the HVS;

- Images and videos captured by modern cameras are invariably corrupted by noise.

With increasing pixel resolution of image and video capturing devices, but more or less the same aperture size, noise suppression has become more relevant. This creates the need for better image restoration algorithms that can recover an image which is *perceptually* as close as possible to the distortion-free image.

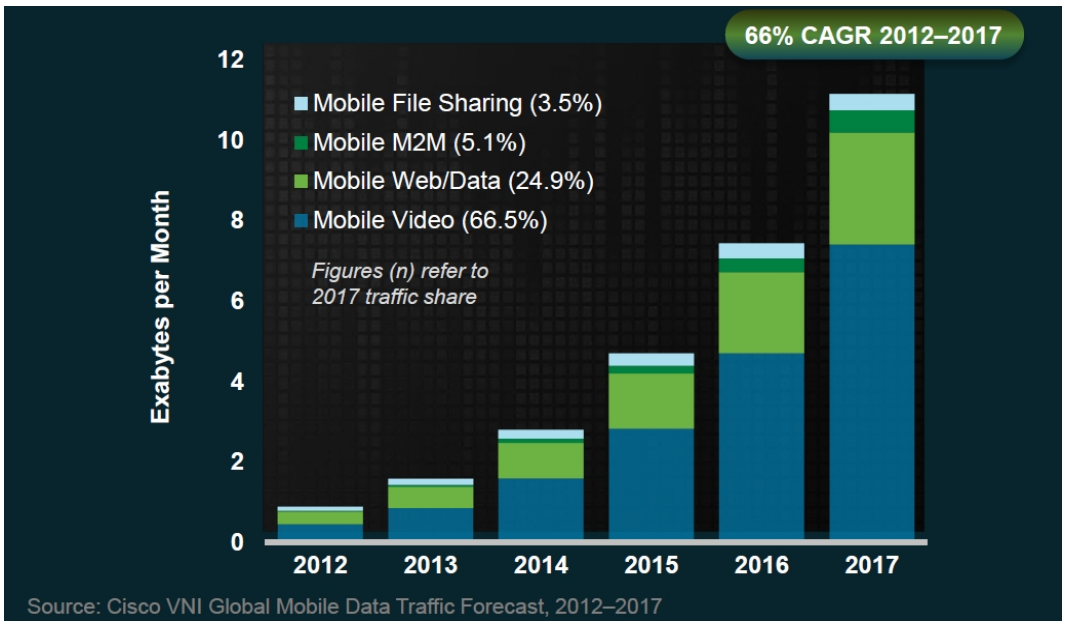


Figure 1.1: Global mobile video data traffic forecast (2012 - 2017)

Central to the image and video processing algorithms, designed to deal with challenges mentioned above, is a constrained optimization problem. The solution to this problem aims to generate an output that is as close as possible to the optimal, by maximizing its similarity with the desired signal in the presence of constraint(s). Depending on the type of application, the constraint term is defined based on factors such as available resources, prior knowledge about the underlying unknown signal, among others. Central to such an optimization problem is the way the similarity is measured, because an image can only be as good as it is optimized for. Since the ultimate receivers of images are human eyes, the correct optimization goal of image processing algorithms should be perceptual quality.

In most of the image and video processing algorithms, mean squared error (MSE) has been the preferred choice as the optimization criterion due to its ease of use and

popularity, irrespective of the nature of signals involved in a problem. Algorithms are developed and optimized to generate the output image that has minimal MSE with respect to the target image. MSE has been known as a poor indicator of perceived image and video quality, however it is widely employed, mostly because of its simplicity and sound mathematical properties for optimization purposes. The structural similarity (SSIM) index [165], a recently proposed computationally simple image similarity measure, has shown superior performance over MSE in predicting perceptual quality and has a number of desirable mathematical properties for optimization tasks. The SSIM index has a great potential as an optimization criterion for image and video processing applications.

1.2 Objectives

The main goal of this thesis is to demonstrate SSIM's potential as a perceptual quality measure for optimization of primary image and video processing algorithms in visual communications. We provide SSIM-inspired novel methods for quality assessment, restoration, and compression of visual data and demonstrate that output of the proposed methods is visually more pleasing than that of the corresponding state-of-the-art MSE-optimal methods.

1.3 Contributions

The main contributions of this thesis are summarized as follows:

1. a general-purpose RR-IQA measure and an image repairing algorithm using reduced-reference features;
2. SSIM-inspired sparsity based image restoration and non-local means image de-noising;
3. SSIM-Quantization model for rate-distortion optimization and SSIM-inspired adaptive quantization method for video compression;

4. an asymmetric adaptation objective model to quantify the perceptual experience of time-varying video quality.

1.4 Thesis Outline

The objective of this work is to break the trend of using MSE as the optimization criterion for image and video processing algorithms and demonstrate that significantly better performance can be achieved when SSIM is employed for the design and optimization of image/video processing algorithms. This work targets several main optimization problems in visual communications that are solved conventionally using MSE as the distortion measure. SSIM-inspired novel solutions and algorithms are devised to solve the optimization problems.

The layout of this thesis is organized as follows. Chapter 2 discusses the related work on the topics addressed in the thesis. We briefly describe characteristics of natural images in the beginning of the chapter in order to provide an introduction to the properties of visual data. We highlight the importance of perceptual image and video quality assessment in the context of optimal image and video processing and compare the SSIM index to the MSE as a model of image perception through examples and psycho-physical experiments. An overview of previous work done on visual quality assessment, restoration, and compression is presented at the end of the chapter.

SSIM is a Full-Reference IQA (FR-IQA) scheme that requires full availability of the reference image in order to estimate the quality of the distorted image. This makes it impractical in visual communication applications, where we have no access to the reference image at the receiver side. Reduced-reference image quality assessment (RR-IQA) provides a practical solution for automatic image quality evaluations in various applications where only partial information about the original reference image is accessible. We propose a general purpose RR-IQA method in Chapter 3 that can estimate the SSIM index with high accuracy. We introduce the novel idea of partially repairing an image using RR features and use de-blurring as an example to demonstrate its application.

Image processing algorithms such as image de-noising and image super-resolution are

required at various stages of visual communication starting from image acquisition to image display at the receiver. In Chapter 4, we incorporate SSIM into the framework of sparse signal representation and approximation. Specifically, the proposed optimization problem solves for coefficients with minimum $\mathcal{L}_0$ norm and maximum SSIM index value. Furthermore, a gradient descent algorithm is developed to achieve SSIM-optimal compromise in combining the input and sparse dictionary reconstructed images. We demonstrate the performance of the proposed method by using image de-noising and super-resolution methods as examples.

Chapter 5 presents a novel SSIM-based non-local means image de-noising algorithm. We incorporate SSIM into the framework of non-local means (NLM) image de-noising. Specifically, a de-noised image patch is obtained by weighted averaging of neighboring patches, where the similarity between patches as well as the weights assigned to the patches are determined based on an estimation of SSIM.

Video compression is absolutely necessary for visual communication. We propose a rate-distortion optimization (RDO) scheme based on the SSIM index in Chapter 6 . At the frame level, an adaptive Lagrange multiplier selection method is proposed based on a novel reduced-reference statistical SSIM estimation algorithm and a rate model[1] that combines the side information with the entropy of the transformed residuals. At the macroblock level, the Lagrange multiplier is further adjusted based on an information theoretical approach[1] that takes into account both the motion information content and perceptual uncertainty of visual speed perception.

In Chapter 7, we propose a perceptual video coding framework based on the divisive normalization scheme, which was found to be an effective approach to model the perceptual sensitivity of biological vision, but has not been fully exploited in the context of video coding. At the macroblock (MB) level, we derive the normalization factors based on the SSIM index as an attempt to transform the DCT domain frame residuals to a perceptually uniform space. We further develop an MB level perceptual mode selection scheme[1] and a frame level global quantization matrix optimization method.

We study the perceptual experience of time-varying video quality in Chapter 8. In real-

[1]proposed in collaboration with S. Wang, a visiting Ph.D. student from Peking University, Beijing.

world visual communications, it is a common experience that end-users receive video with significantly time-varying quality due to the variations in video content/complexity, codec configuration, and network conditions. The way by which the human visual quality of experience (QoE) changes with such time-varying video quality is not yet well-understood. To investigate this issue, we conduct subjective experiments designed to examine the quality predictability between individual video segments of relatively constant quality and combined videos consisting of multiple segments that have significantly different quality. We propose a quality adaptation model that is asymmetrically tuned to increasing and decreasing quality.

Finally, Chapter 9 concludes the thesis and discusses different avenues for future research.

The research performed in this thesis lead to the development of state-of-the-art image and video processing algorithms. The significant improvement in the performance of the image and video processing algorithms due to the use of SSIM provides strong evidence for convincing researchers to replace MSE with SSIM in image and video processing applications.

Chapter 2

Background

This chapter starts with a brief discussion about the characteristics of visual data followed by an introduction of perceptual image and video processing. The SSIM index will then be presented and compared to the MSE as a model of image perception through examples and psycho-physical experiments. This chapter also performs a brief overview of previous work done on visual quality assessment, restoration, and compression. The review is by no means comprehensive and only summarizes relevant literature, while leading the interested reader to more comprehensive reviews.

2.1 Characteristics of Natural Images

Natural or typical images and videos (a stack of images) refer to the visual data obtained from a camera - these include pictures of physical scenes, man-made objects and natural environments. The "amount" of incoming photons entering a camera, through an open aperture, is recorded on an array of charge-coupled device (CCD) receptors. The analog values measured in the form of difference of voltage are converted to digital form using an analog-to-digital converter. The digital data is then transformed into an array of pixels (picture elements).

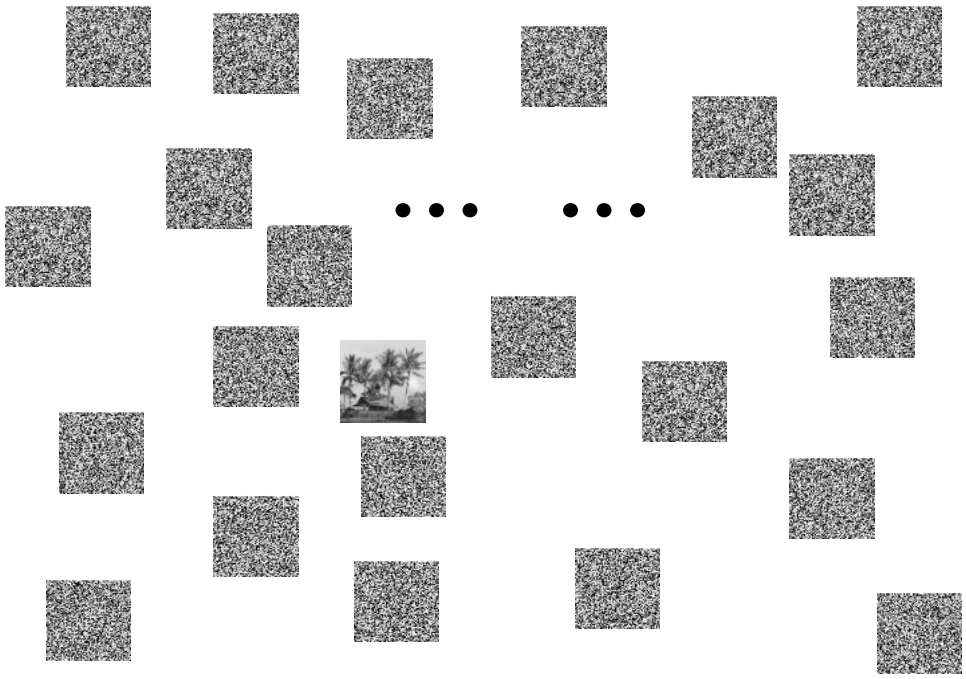


Figure 2.1: Depiction of the image space

The knowledge of the *nature* of visual data, obtained as a result of the process explained above, is very crucial for efficient processing, transmission, and storage of images and videos. An important question that arises here is: how to fully describe the image statistics? That is, given any combination of pixel values (supposing an already sampled image), can we find the probability that this image could be taken by a digital camera at any time in history? There are approximately 10^{1000} possible 65×65 gray-scale images. This gives us a good idea about how big is the space of all the images. Typical (natural) images occupy an extremely tiny (and unknown-shape) space in the space of all images as demonstrated by Figure 2.1. As an attempt to understand the statistical properties of natural images, let us first take a look at the marginal distribution of two images from the scene as shown in Figure 2.2. We can observe that the marginal distribution of the two images vary significantly even though the images belong to the same scene. Therefore, we

can conclude that marginal statistics of natural images may not be useful in defining the space of all natural images. Next, we look at the joint statistics of closely located pixels in the image shown in 2.3(a). Specifically, we draw a scatter plot between the pixel intensity values which are located at a horizontal distance of one, two, and four pixels. Majority of the points in the scatter plots lie along the diagonal line which is an evidence of strong correlation between intensity values of neighboring pixels. The spread of points increases as the distance between pixels increases, depicting the trend of decrease in correlation with increase in pixel distance. We can conclude from Figure 2.3 that pixels in an image are not independent of one another and there exists some sort of structure since there is a strong correlation between neighboring pixels.

The current state of research in the area of natural image statistics has not yet been able to find a good natural image model. The main observations include:

- second-order pixel correlations [50];
- importance of phases [16];
- optimal approximation by piecewise smooth functions [96];
- heavy-tail non-Gaussian marginals in wavelet domain [32, 37];
- near elliptical shape of joint densities in wavelet domain [131];
- decay of dependency in the wavelet domain [82].

2.2 Perceptual Image and Video Quality Assessment

The Human Visual System (HVS) is optimized for processing the spatial information in natural visual images [111]. We have learned in the previous section that visual content exhibits certain properties which are specific to natural images and videos. This knowledge can applied to design better image and video quality assessment methods that

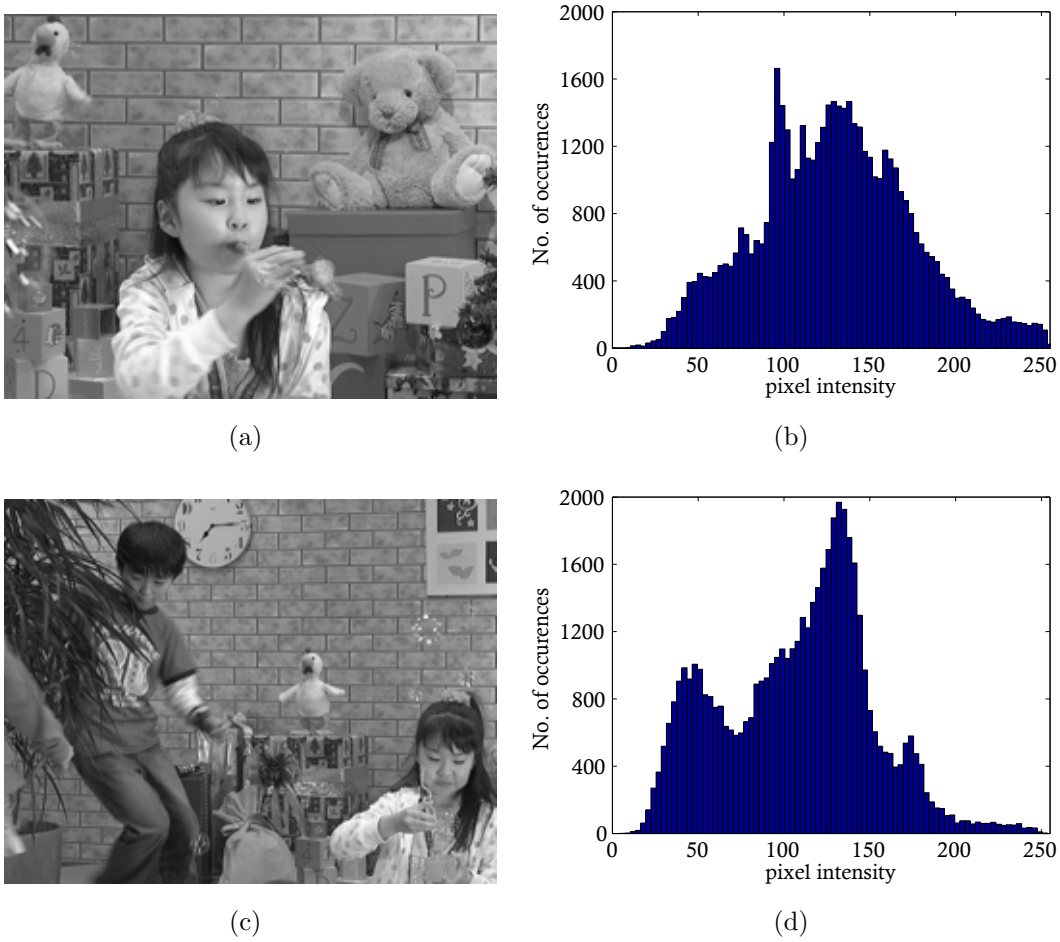


Figure 2.2: Marginal statistics of pixel intensity

correlate well with the HVS. In the classical approach for image and video quality assessment, the researchers use a bottom-up approach that builds the computational system of the HVS in order to reach a realistic model of image quality perception [39, 80]. A top-down philosophy, towards image and video quality assessment, makes hypotheses about the overall functionality of the HVS. The main purpose of such an approach is to use a simpler solution by treating HVS as a black-box and concentrating only on its input-output relationship [127, 165].

The use of image and video quality assessment methods in the design and optimiza-

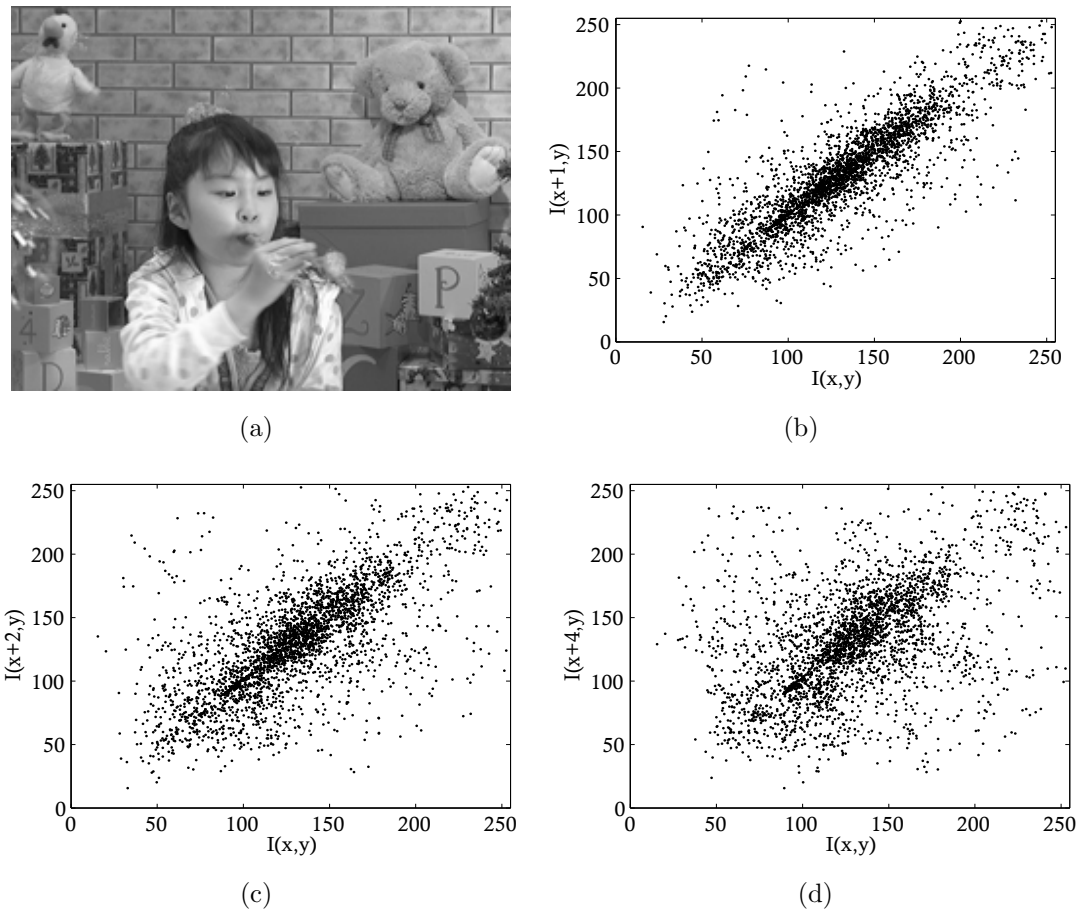


Figure 2.3: Second order statistics of pixel intensity

tion of image/video processing algorithms and systems is more desirable, challenging, and fruitful as compared to benchmarking and monitoring of visual data. Existing image and video quality assessment methods do not directly qualify for this job. To be deemed fit for optimization of image and video processing algorithms, there are four desirable properties in an image and video quality assessment method:

1. high correlation with subjective scores;

2. low computational complexity so that the algorithm is practically usable;

3. accurate local quality prediction that can help determine varying local quality level based on content;

4. good mathematical properties that can help in solving optimization problems i.e. a valid distance metric that satisfies convexity, differentiability, symmetry, etc.

Most of the existing image and video quality assessment methods lack at least one of the above mentioned characteristics. For example, Video Quality Metric (VQM) [104] is good at predicting perceived video quality but is computationally very complex, does not provide local quality map, and does not satisfy the desirable mathematical properties. According to the best of our knowledge, there are only two image and video quality assessment methods that satisfy all four requirements: MSE/PSNR (Peak Signal-to-Noise Ratio) and SSIM. The remainder of this section introduces MSE and SSIM and also provides their comparison based on the four points mentioned above.

2.2.1 Mean Squared Error

The goal of an image/video fidelity measure is to provide a quantitative comparison between two images/videos, where one of the image/video is considered pristine or treated as a reference. The most widely method to measure image/video fidelity is PSNR, a monotonic function of MSE.

The MSE between two images $\mathbf{y}$ and $\hat{\mathbf{y}}$ is

$$\mathrm{MSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{L_1 L_2} \sum_{i_1=1}^{L_1} \sum_{i_2=1}^{L_2} \left(\mathbf{y}(i_1, i_2) - \hat{\mathbf{y}}(i_1, i_2)\right)^2, \tag{2.1}$$

where L_1 and L_2 are the length and the width of the images respectively. As MSE computation is based on the error signal, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, between the reference image, $\mathbf{y}$, and its distorted version, $\hat{\mathbf{y}}$, therefore it can regarded as a measure of image quality. In image and video processing literature, MSE is often converted to PSNR using the expression:

$$\mathrm{PSNR}(\mathbf{y}, \hat{\mathbf{y}}) = 10 \log_{10} \frac{R^2}{\mathrm{MSE}(\mathbf{y}, \hat{\mathbf{y}})}, \tag{2.2}$$

(a) MSE = 0, SSIM = 1 (b) MSE = 200, SSIM = 0.99 (c) MSE = 200, SSIM = 0.99

(d) MSE = 200, SSIM =0.68 (e) MSE = 198, SSIM = 0.65 (f) MSE = 200, SSIM = 0.72

Figure 2.4: Comparison between distorted images with the same MSE. (a) Original image; (b) Global brightness shift; (c) Global contrast stretch; (d) Gaussian noise; (e) Gaussian blur; (f) JPEG compression.

where R is the dynamic range of image pixel intensities e.g. for an 8-bit/pixel gray-scale image, $R = 2^8 - 1 = 255$. The only advantage of PSNR over MSE, as a perceptual quality measure, is its capability to handle images with different dynamic ranges.

MSE has been ubiquitously used in the literature as a signal fidelity measure and its use as an image/video quality assessment metric has become a convention. The poor performance of MSE as an image/video quality assessment method is ignored on the expense of its attractive features such as simplicity, low computational cost, and memorylessness [166]. MSE serves very well in solving design an optimization problems for the following reasons: it is a valid distance metric in $\mathcal{R}^N$; it preserves energy after any orthogonal (or unitary) linear transformation (Parseval's theorem); it is convex, and differentiable; it often provides closed form or iterative numerical solutions to optimization problems; it is additive

for independent sources of distortions. In spite of having sound mathematical properties, MSE should not be used unquestioned as a perceptual quality measure in image and video processing applications. Figure 2.4 provides an illustrative example and rationale for not trusting MSE's judgment of perceptual quality. The reference image is shown in Figure 2.4(a). The rest of images are created from the reference image by introducing same level of various distortions in terms of MSE. We can readily observe that the perceptual quality of the distorted images differs significantly, although MSE wrongly predicts a similar quality. According to the MSE, the image in Figure 2.4(e) has the best quality among the five distorted images. However, according to the HVS, the images in Figures 2.4(b) and 2.4(c) have the least perceptual distortion.

The MSE does not account for a number of important psychological and physiological features of the HVS [163]. The reason behind failure of MSE in providing accurate perceived quality prediction lies in various questionable assumptions made when used as an image/video quality measure: 1) the spatial relationship between pixels is irrelevant as far the perceptual quality is concerned, therefore, the distortion in each pixel can be calculated individually; 2) the error signal, $\mathbf{e}$, introduces the same level of distortion when introduced in any reference image; 3) the perceptual quality evaluation is insensitive to the sign of $\mathbf{e}$; 4) all the pixels of an image are equally important for perceptual image quality. All of these assumptions have been proven wrong [166].

2.2.2 Structural Similarity

SSIM is based on measuring the similarities of luminance, contrast and structure between local image patches $\mathbf{x}$ and $\mathbf{y}$ extracted from a reference and a distorted image:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{2.3}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{2.4}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{2.5}$$

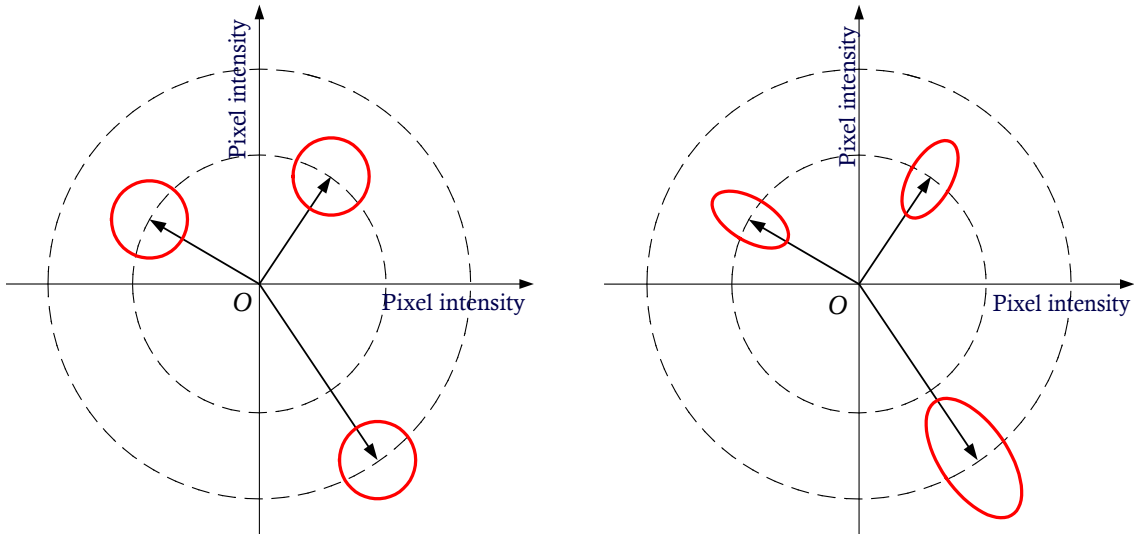


Figure 2.5: Comparison of level sets; (a) MSE measure; (b) SSIM measure.

where μ, σ and σ_{xy} represent the mean, standard derivation and covariance of the image patches, respectively, and C_1, C_2 and C_3 are positive constants used to avoid instability when the denominators are close to zero. Subsequently, the local SSIM index is defined as the product of the three components, which gives

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \big[l(\mathbf{x},\mathbf{y})\big]^{\alpha}\big[c(\mathbf{x},\mathbf{y})\big]^{\beta}\big[s(\mathbf{x},\mathbf{y})\big]^{\gamma} \tag{2.6}$$

The SSIM index is usually simplified by taking $\alpha = \beta = 1$, and $C_3 = C_2/2$. Equation 2.6 then reduces to

$$\begin{aligned} \mathrm{SSIM}(\mathbf{x},\mathbf{y}) &= \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right)\left(\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right), &(2.7)\\ &= \mathrm{S}_1(\mathbf{x},\mathbf{y})\mathrm{S}_2(\mathbf{x},\mathbf{y}). &(2.8)\end{aligned}$$

The SSIM index of the whole image is obtained by averaging (or weighted averaging) the local SSIM indices obtained using a sliding window that runs across the image.

Figure 2.5 gives a graphical explanation in the vector space of image components, which can be pixels, wavelet coefficients, or features extracted from the reference image.

For the purpose of illustration, two-dimensional diagrams are shown here. However, the actual dimensions may be equal to the number of pixels or features being compared. The three vectors represent three reference images and the contours around them represent the images with the same distortion level using (a) MSE and (b) SSIM as the distortion/quality measures, respectively. The critical difference is in the shapes of the contours. Unlike MSE (where all three contours have the same size and shape), SSIM is adaptive according to the reference image. In particular, if the "direction" of distortion is consistent with the underlying reference (aligned with the direction of the reference vector), the distortion is non-structural and is much less objectionable than structural distortions (the distortions perpendicular to the reference vector direction). The formulation of SSIM in (2.6) provides a flexible framework for adjusting the relative importance between structural (the last term) and non-structural (first two terms) distortions. As explained in [165], the luminance term of the SSIM index is related to Weber's Law. According to the this law, the perception of the change of a stimulus is proportional to the intensity of the stimulus. Weber's Law not only applies to the luminance but also to the image contrast i.e., the ratio of contrasts is constant for a constant SSIM index value. Figure 2.4 provides an example of SSIM's capability to differentiate between structural and non-structural distortion. Global brightness shift (Figure 2.4(b)) and global contrast shift (Figure 2.4(b)) introduce non-structural distortion in the reference image (Figure 2.4(a)) and as a result are penalized less by SSIM index as compared to the structural distortions namely Gaussian noise (Figure 2.4(d)), Gaussian blur (Figure 2.4(e)), and JPEG compression (Figure 2.4(f)).

Equation (2.6) does not take into account the viewing distance of the observer. Therefore, the performance of the SSIM index depends on the scale it is applied to. A multi-scale approach that incorporates SSIM index at various scales, Multi-Scale Structural Similarity (MS-SSIM), has been proposed in [177]. The relative importance/weight of each scale was decided based on psychovisual experiments. Interestingly, the weights determined based on the experiments were found to be consistent with the philosophy of contrast sensitivity function [177]. In the general form, the MS-SSIM can be written as

$$\mathrm{MS-SSIM}(\mathbf{x},\mathbf{y}) = \prod_{r=1}^{R} \left[l(\mathbf{x}_r,\mathbf{y}_r)\right]^{\alpha_r} \left[c(\mathbf{x}_r,\mathbf{y}_r)\right]^{\beta_r} \left[s(\mathbf{x}_r,\mathbf{y}_r)\right]^{\gamma_r}, \tag{2.9}$$

where $\mathbf{x}_r$ and $\mathbf{y}_r$ are the image $\mathbf{x}$ and $\mathbf{y}$, respectively, at resolution r.

Initially, a simple average over the local SSIM scores was adapted as the pooling strategy [165]. Information content based weighting can yield more accurate quality prediction as compared to minkowski, local quality/distortion-based, saliency-based, and object-based pooling. Information content Weighted SSIM (IW-SSIM) has been shown to outperform the basic spatial domain SSIM index [168].

The major drawback of the spatial domain SSIM index is its high-sensitivity to translation, scaling, and rotation of images, which are also non-structural distortions. The CW-SSIM measure was proposed in [121, 175], which was built upon local phase measurements in complex wavelet transform domain. The underlying assumptions behind CW-SSIM are that local phase pattern contains more structural information than local magnitude, and non-structural image distortions such as small translations lead to consistent phase shift within a group of neighboring wavelet coefficients. Therefore, CW-SSIM is designed to separate phase from magnitude distortion measurement and impose more penalty to inconsistent phase distortions.

Consider a mother wavelet $w(u) = g(u)e^{jw_c u}$, where w_c is the center frequency of the modulated band-pass filter and $g(u)$ is a slowly varying symmetric function. The family of wavelets are dilated and translated versions of $w(u)$ given by

$$w_{s,p}(u) = \frac{1}{\sqrt{s}} w\left(\frac{u-p}{s}\right) = \frac{1}{\sqrt{s}} g\left(\frac{u-p}{s}\right) e^{jw_c(u-p)/s} \tag{2.10}$$

where scale factor $s \in \mathcal{R}^+$ and translation factor $p \in \mathcal{R}$. The continuous wavelet transform of a real signal $x(u)$ is

$$X(s,p) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(w)\sqrt{s}G(sw - w_c)e^{jwp}dw \tag{2.11}$$

where $X(w)$ and $G(w)$ are the Fourier transforms of $x(u)$ and $g(u)$, respectively. The discrete wavelet coefficients are sampled versions of the continuous wavelet transform. Please note that this is a specific way of defining wavelets that best suits the target application. Interested reader should refer to [89] for a comprehensive description of wavelet transforms.

Given two sets of complex wavelet coefficients $\mathbf{c}_x = \{c_{x,i} | i = 1, ..., M\}$ and $\mathbf{c}_y = \{c_{y,i} | i = 1, ..., M\}$ extracted at the same spatial location in the same wavelet subbands of the two

images being compared, the local CW-SSIM index is defined as

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2|\sum_{i=1}^{M} c_{x,i} c_{y,i}^*| + K}{\sum_{i=1}^{M} |c_{x,i}|^2 + \sum_{i=1}^{M} |c_{y,i}|^2 + K}. \tag{2.12}$$

where c^* denotes the complex conjugate of c, and K is a small positive stabilizing constant. The value of the index ranges from 0 to 1, where 1 implies no structural distortion (but could still have a small spatial shift). The global CW-SSIM index $\tilde{S}(I_x, I_y)$ between two images I_x and I_y is calculated as the average of local CW-SSIM values computed with a sliding window running across the whole wavelet subband and then averaged over all subbands. It was demonstrated that CW-SSIM is simultaneously insensitive to luminance change, contrast change, and small geometric distortions such as translation, scaling and rotation [121, 175]. This makes CW-SSIM a preferred choice for image classification tasks because it is versatile and largely reduces the burden of preprocessing steps such as contrast and mean adjustment, pixel shifting, deskewing, zooming and scaling.

The SSIM index and its extensions have found a wide variety of applications, ranging from image/video coding, i.e., H.264 video coding standard implementation [67], image classification [53, 113], restoration and fusion [103], to watermarking, de-noising and biometrics (See [166] for a list of references). In most existing works, however, SSIM has been used for quality evaluation and algorithm comparison purposes only. SSIM possesses a number of desirable mathematical properties, making it easier to employ in optimization tasks than other state-of-the-art perceptual IQA measures [11]. However, much less work has been done on using SSIM as an optimization criterion in the design and optimization of image processing algorithms and systems [26, 84, 98, 114, 116, 119, 157, 169, 187].

2.2.3 Comparison between MSE and SSIM

One the main differences between SSIM and MSE is the divisive normalization [10, 158]. This normalization is conceptually consistent with the light adaptation (also called luminance masking) and contrast masking effect of HVS and has been recognized as an efficient perceptual and statistical non-linear image representation model [81, 154]. Moreover, it provides a useful framework that accounts for the masking effect in the HVS, which refers to

the reduction of the visibility of an image component in the presence of large neighboring components [51, 179]. Divisive normalization is also powerful in modeling the neuronal responses in the visual cortex [59, 135], and has been successfully applied in image quality assessment [76, 115], image coding [90], video compression [118, 158, 160], and image de-noising [109].

Correlation with subjective scores

The ultimate goal of an IQA algorithm is to predict subjective quality scores of images. Therefore, an important comparison between MSE and SSIM is based on how well they can predict subjective scores. For this purpose, we use six publicly available databases and four evaluation metrics to compare the performance of MSE and SSIM. The evaluation metrics are Pearson Linear Correlation Coefficient (PLCC), Root Mean Squared Error (RMSE), Spearman's rank correlation coefficient (SRCC), and Kendall's rank correlation coefficient (KRCC). The detail of these metrics can be found in Chapter 3. The performance of improved versions of SSIM, MS-SSIM and IW-SSIM, is also provided in Table 2.1. We can observe from the results that SSIM and its variants perform significantly better than PSNR, simply a remapping of MSE, in predicting subjective scores of all the databases. As a result, we can conclude that SSIM is a better perceptual quality measure as compared to MSE.

Local quality prediction

Optimization of image and video processing applications require accurate local perceptual quality prediction for the following reasons:

- Image distortions may not be uniform across the whole image or uniform image distortions may introduce space-variant degradation;
- Statistical features for a typical image are significantly non-stationary across space;
- A human observer can only perceive a small high-resolution region in an image at one time due to the nonuniform retinal sampling feature of HVS.

Table 2.1: Performance comparison between PSNR and SSIM using publicly available databases

	LIVE Database [130]				Cornell A57 Database [23]			
IQA	PLCC	MAE	SRCC	KRCC	PLCC	MAE	SRCC	KRCC
PSNR	0.8723	10.51	0.8756	0.6865	0.6347	0.1607	0.6189	0.4309
SSIM [165]	0.9449	6.933	0.9479	0.7963	0.8017	0.1209	0.8066	0.6058
MS-SSIM [177]	0.9489	6.698	0.9513	0.8044	0.8603	0.1007	0.8414	0.6478
IW-SSIM [168]	0.9522	6.470	0.9567	0.8175	0.9034	0.0892	0.8709	0.6842
	IVC Database [74]				Toyama-MICT Database [60]			
IQA	PLCC	MAE	SRCC	KRCC	PLCC	MAE	SRCC	KRCC
PSNR	0.6719	0.7191	0.6884	0.5218	0.6329	0.7817	0.6132	0.4443
SSIM [165]	0.9119	0.3777	0.9018	0.7223	0.8887	0.4386	0.8794	0.6939
MS-SSIM [177]	0.9108	0.3813	0.8980	0.7203	0.8927	0.4328	0.8874	0.7029
IW-SSIM [168]	0.9231	0.3694	0.9125	0.7339	0.9248	0.3677	0.9202	0.7537
	TID 2008 Database [106]				CSIQ Database [72]			
IQA	PLCC	MAE	SRCC	KRCC	PLCC	MAE	SRCC	KRCC
PSNR	0.5223	0.8683	0.5531	0.4027	0.7512	0.1366	0.8058	0.6084
SSIM [165]	0.7732	0.6546	0.7749	0.5768	0.8612	0.0992	0.8756	0.6907
MS-SSIM [177]	0.8451	0.5578	0.8542	0.6568	0.8991	0.0870	0.9133	0.7393
IW-SSIM [168]	0.8579	0.5276	0.8559	0.6636	0.9144	0.0801	0.9213	0.7529

Figure 2.7 compares absolute error (the bases for $\mathcal{L}_p$, MSE, PSNR, etc.) and SSIM maps in terms of their accuracy in predicting local quality of an image. The reference image, shown in Figure 2.4(a), is degraded with different types of distortion. The distorted images are shown in the left column of Figure 2.7. The absolute error map is adjusted so that brighter indicates better predicted quality.

Global brightness shift introduces a non-structural distortion in the reference image and affects the quality minutely as shown by the SSIM map in Figure 2.7(c). However, the absolute error map in Figure 2.7(b) predicts severe uniform local distortion which contradicts the perceived local quality of the distorted image.

In Figure 2.7(d), the distortion is not obvious as global contrast shift is a non-structural distortion. Therefore, the SSIM map predicts good quality throughout the image. On the other hand, according to the absolute error map quality prediction, the bright areas in the image are severely distorted which does not conform with the perceived quality of the distorted image.

In Figure 2.7(g), introducing noise severely degrades the quality of low variance regions such as the face and the box, which is accurately predicted by the SSIM map. However, the absolute error map is completely independent of the underlying image structures.

In Figure 2.7(j), details of the face are preserved relatively better compared to the wall/bricks. This is clearly indicated by the SSIM index map, but again, not well predicted by the absolute error map.

Maximum Differentiation

Figure 2.4 provides a convincing example and rationale for not trusting MSE's judgment of perceptual quality. According to MSE, the quality of all the distorted images is similar, but visually they do not have the same perceptual quality. For a fair comparison, we should devise another example with all the distorted images with the same SSIM index value. The methodology for comparing computational models of perceptual quantities through Maximum differentiation (MAD) competition was proposed in [176]. We use this methodology to create an example by keeping the quality level of the input image fixed according to one of the IQA methods (MSE/SSIM) and synthesizing output images with maximum/minimum quality according to the other IQA method, and vice versa. The perceptual quality assessment method whose maximum/minimum quality pairs are easier for subjects to discriminate is the better method. The reference and input images used in the experiment are shown in Figures 2.8(a) and 2.8(b), respectively. Images 2.8(c) and 2.8(d) have the maximum and minimum quality, respectively, according to MSE but the same SSIM index value as the input image. Images 2.8(e) and 2.8(f) have the maximum and minimum quality, respectively, in the SSIM sense but the same MSE value as the input image. We can observe from the images that it is easier to differentiate between the images 2.8(e) and 2.8(f) as compared to images 2.8(c) and 2.8(d). Therefore, we can conclude that

SSIM is a better perceptual quality assessment method as compared to MSE. Also, we can observe that the image 2.8(e) has almost perfect perceptual quality but has a very high MSE value. On the other hand, the perceptual quality of the image 2.8(d) is not as bad as the MSE value suggests. According to these observations MSE failed to accurately predict the perceptual quality of the images.

Mathematical Properties

One often faces major difficulties when solving optimization problems based on visual quality assessment measures in image and video processing applications. This is largely due to the lack of desirable mathematical properties in perceptual quality assessment measures. Although the MSE exhibits poor correlation with subjective scores, it is an ideal target for optimization as it is based on a valid distance measure ($\mathcal{L}_2$) that satisfies positive definiteness, triangular inequality, and symmetry. Additionally, the MSE is differentiable, convex, memoryless, energy preserving under orthogonal transforms, and additive for independent sources [162, 166].

$\sqrt{1-\text{SSIM}}$ has been shown to be a valid distance metric (that satisfies the identity and symmetry axioms as well as the triangle inequality) and has a number of useful local and quasi-convexity, and distance preserving properties [11]. Quasi-convexity, a weaker form of convexity, can be useful for the numerical or analytical optimization of the SSIM index and its derivatives. Local convexity implies that there exists a sphere around the location of minimum for which $\sqrt{1-\text{SSIM}}$ is convex.

Computational Complexity

Table 2.2 compares popular perceptual quality assessment algorithms in terms of normalized complexity. We can observe that the MSE is computationally very simple. On the other hand, VQM [104] and MOVIE [124] are computationally extremely expensive, while SSIM and MS-SSIM achieve a much better balance between quality prediction accuracy and computational complexity.

Table 2.2: Computational complexity comparison of popular perceptual quality assessment measures

Model	Computational Complexity (normalized)
MSE	1
SSIM [165]	5.874
MS-SSIM [177]	11.36
VQM [104]	1083
MOVIE [124]	7229

2.3 Perceptual Image and Video Processing

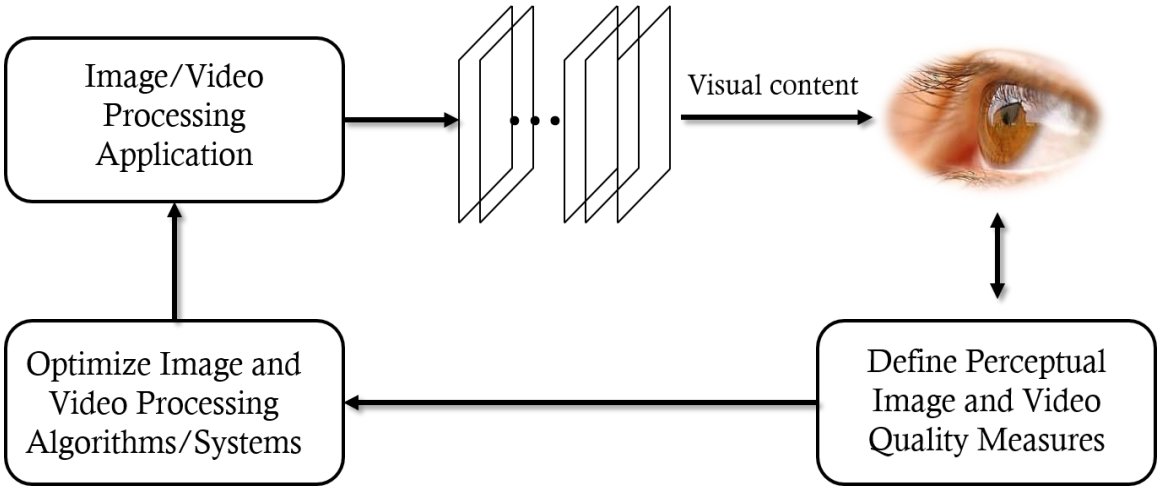


Figure 2.6: General framework for perceptual image and video processing

Objective visual quality assessment algorithms are primarily developed to monitor signal fidelity as part of QoS efforts, and also to benchmark signal processing systems and algorithms in comparative studies [162]. The use of image and video quality assessment measures in the design and optimization of image/video processing algorithms and systems is more desirable, challenging and fruitful but has not been well explored.

Figure 2.6 shows the general framework for perceptual image and video processing. Since the ultimate receiver of visual content is the HVS, the correct optimization goal of image and video processing algorithms should be perceptual quality. A scientific design of a visual processing algorithm always involve certain quality criterion, either explicitly or implicitly. An image/video can only be as good as it is optimized for. If a good perceptual quality measure is available, one may use it for performance assessment of these algorithms and systems and also for their optimization with the objective of producing the optimal image/video under this criterion. There are two capacities in which objective perceptual quality measures can be applied in the optimal design of visual processing algorithms and systems.

- The core visual processing algorithms are used as a black box and the perceptual quality measure is used to provide feedback control signals to update the visual processing algorithm in an iterative manner.
- The perceptual quality measure is used in the core visual processing algorithm which would be practical if the quality measure had good mathematical properties to help solve optimization problems.

An effective objective quality assessment suitable for optimizing image and video processing applications should provide accurate local quality prediction apart from exhibiting high correlation with subjective scores at a low computational cost. Figure 2.9 demonstrates how a perceptual quality assessment could be useful in the context of image coding. The original image (a) is compressed using JPEG compression method. Due to limited resources, the decompressed image (b) has strong blocking and blurring artifacts. Specifically, the blocking artifact in the sky is clearly visible and the boundaries of the building have lost details including the text. Given additional bit budget to improve the image quality, we would allocate the additional rate to the regions that would most improve the perceptual quality of the image. An accurate local quality map can prove to be an excellent guide in such a situation. Figure 2.9(c) shows the absolute error map which serves as the bases for $\mathcal{L}_p$, MSE, and PSNR computation. We can observe that the map provides us with the wrong guidance in suggesting the worst quality areas. The SSIM map,

shown in Figure 2.9(d) provides local quality prediction that is consistent with our visual observations. Since most image and video compression methods are designed based on MSE/PSNR, the dramatic difference between the MSE and SSIM quality maps reveals the great potential of perceptual image and video compression.

Central to the image and video processing algorithms is a constrained optimization problem. The solution to this problem aims to generate an output that is as close as possible to the optimal, by maximizing its similarity with the desired signal in the presence of constraint(s). Depending on the type of application, the constraint term is defined based on factors such as available resources, prior knowledge about the underlying unknown signal, among others. Central to such an optimization problem is the way the similarity is measured. We will briefly discuss the main optimization problem involved in video compression to show the importance of a good perceptual quality assessment algorithm.

Video codecs are primarily characterized in terms of the throughput of the channel and perceived distortion of the reconstructed video. The main task of a video codec is to convey the sequence of images with minimum possible perceived distortion within the available bit rate. Alternatively, it can be posed as a communication problem to convey the sequence with the minimum possible rate while maintaining a specific perceived distortion level. In both versions of the problem, the fundamental issue is to obtain the best trade-off between rate and perceived distortion. The process used to achieve this objective is commonly known as Rate Distortion Optimization (RDO), which can be expressed by minimizing the perceived distortion D with the number of used bits R subjected to a constraint R_c [140]

$$\min\{D\} \quad \text{subject to} \quad R \leq R_c \tag{2.13}$$

This is a typical constrained optimization problem. Lagrangian optimization technique converts this constrained optimization problem to an unconstrained optimization problem [140], which can be expressed as

$$\min\{J\} \quad \text{where} \quad J = D + \lambda \cdot R, \tag{2.14}$$

where J is called the Rate Distortion (RD) cost and the rate R is measured in number of bits per pixel. λ is known as the Lagrange multiplier and controls the trade-off between

R and D. In practice, distortion models such as Sum of Absolute Difference (SAD) and Mean Squared Error (MSE) are used in most actual comparisons [54]. The use of an accurate perceptual quality measure, such as SSIM, instead of MSE can deliver superior performance by offering significant rate reduction, while maintaining the same level of perceptual quality [162].

2.4 Reduced-Reference Image Quality Assessment

2.4.1 Literature Review

A lot of work has been done in the recent past to develop objective quality assessment measures that can automatically measure the perceived distortion in the visual content. The most prominent ones include the structure similarity index (SSIM) [165] and its derivatives [168, 177], visual information fidelity (VIF) [127], visual signal-to-noise Ratio (VSNR) [23], and most apparent distortion (MAD) [73]. Among these methods, SSIM has often been preferred due to its good trade-off between accuracy, simplicity and efficiency [166]. The success of SSIM motivated us to use it for visual communication applications. The difficulty is that SSIM is a Full-Reference IQA (FR-IQA) scheme that requires full availability of the reference image to estimate the quality of the distorted image. This makes it impractical in visual communication applications, where we have no access to the reference image at the receiver side. No-Reference IQA (NR-IQA) is highly desirable as it does not require access to the reference image. In the literature, most NR-IQA algorithms were designed for specific and limited types of distortions [77, 92, 126, 164, 172, 185, 196]. They may not be good choices in modern communication networks, because distortions could be a combination of lossy compression, scaling in the bit-rate and spatial/temporal resolution, network delay and packet loss, and various types of pre- and post-processing filtering (e.g., error concealment, deblocking filtering, sharpening). On the other hand, general purpose NR-IQA is still at an immature stage.

An RR-IQA method requires only a limited number of RR features extracted from the reference for the IQA task [173]. It provides an interesting compromise between FR and

NR approaches in terms of both quality prediction accuracy and the amount of information required to describe the reference. Based on the underlying design philosophy, existing RR-IQA algorithms may be loosely classified into three categories. The first type of methods are primarily built upon models of the image *source*. Since the reference image is not available in a deterministic sense, these models are often statistical in that they capture *a prior* of the low level statistical properties of natural images. The model parameters provide a highly efficient way to summarize the image information, and thus these methods often lead to RR-IQA algorithms with a low RR data rate. In [174, 178], the marginal distribution of wavelet subband coefficients is modeled using a Generalized Gaussian Density (GGD) function, and GGD model parameters are used as RR features and employed to quantify the variations of marginal distributions in the distorted image. The model was further improved in [76] by employing a nonlinear divisive normalization transform (DNT) after the linear wavelet decomposition, resulting in enhanced quality prediction performance, especially when images with different distortion types are mixed together. The second-category RR-IQA methods are oriented to capture image *distortions*. These methods provide useful and straightforward solutions when we have sufficient knowledge about the distortion process that the images have undergone, for example, standard image or video compression [33, 56, 69, 184]. The limitation of such approaches is in their generalization capability. Generally, it is inappropriate to apply these methods beyond the distortions they are designed to capture. The third category of RR-IQA algorithms is based on models of the image *receiver* (i.e., the HVS) [19,20], where computational models from physiological and/or psychophysical vision studies may be employed. These methods have demonstrated good performance for JPEG and JPEG2000 compression [19, 20]. Among the three classes of RR-IQA approaches, the first and third ones, i.e., methods based on modeling image source and receiver, have more potentials to be extended for general-purpose applications because the statistical and perceptual features being used are not restricted to any specific distortion process. There are also interesting conceptual connections between these two types of approaches, because it is a general belief in biological vision science that the HVS is highly tuned for efficient statistical encoding of the natural visual environment [5, 136].

In [145], an interesting RR video quality measure based on SSIM estimation was proposed for quantifying visual degradations caused by channel transmission errors. The

problem with this scheme is that it decomposes the problem of SSIM estimation into many local problems. This requires each component in the SSIM expression to be estimated separately instead of using global statistics to estimate a global SSIM value. As a result in a description of the image content that significantly increases the number of RR features. Also, it assumes that a specific kind of distortion can be applied to assess images with a wide variety of distortion types.

2.4.2 Test Image Databases

The following six publicly available subject-rated image databases are usually used to test the IQA algorithms.

- The LIVE database [130] contains seven data sets of 982 subject-rated images, including 779 distorted images with five types of distortions at different distortion levels. The distortion types are a) JPEG2000 compression (2 sets); b) JPEG compression (2 sets); c) White noise contamination (1 set); d) Gaussian blur (1 set); and e) fast fading channel distortion of JPEG2000 compressed bitstream (1 set). The subjective test was carried out with each data set individually. A cross-comparison set that mixes images from all distortion types was then used to align the subject scores across data sets. The alignment process is rather crude, but the aligned subjective scores (all data) are still useful references for testing general-purpose IQA algorithms, for which cross-distortion comparisons are highly desirable.

- The Cornell-A57 database [23] contains 54 distorted images with 6 types of distortions: a) quantization of the LH subbands of a 5-level discrete wavelet transform, where the subbands were quantized via uniform scalar quantization with step sizes chosen such that the RMS contrast of the distortions was equal; b) additive Gaussian white noise; c) baseline JPEG compression; d) JPEG2000 compression without visual frequency weighting; e) JPEG2000 compression 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 f) blurring by using a Gaussian filter.

- The IVC database [74,97] includes 185 distorted images with four types of distortions: a) JPEG compression; b) JPEG2000 compression; c) Local adaptive resolution (LAR) coding; and d) blurring.

- The Toyama-MICT database [60] contains 196 images, including 168 distorted images generated by JPEG and JPEG2000 compression.

- The Tampere Image Database 2008 (TID2008) [106, 107] includes 1700 distorted images with 17 distortion types at 4 distortion levels. The types of distortions are: a) Additive Gaussian noise; b) Additive noise in color components, which is more intensive than additive noise in the luminance component; c) Spatially correlated noise; d) Masked noise; e) High frequency noise; f) Impulse noise; g) Quantization noise; h) Gaussian blur; i) Image de-noising; j) JPEG compression; k) JPEG2000 compression; l) JPEG transmission errors; m) JPEG2000 transmission errors; n) Non eccentricity pattern noise; o) Local block-wise distortions of different intensity; p) Mean shift (intensity shift); and q) Contrast change.

- The Categorical Image Quality (CSIQ) Database [72] contains 866 distorted images of six types of distortions at four to five distortion levels. The distortion types include JPEG compression, JPEG2000 compression, global contrast decrements, additive pink Gaussian noise, and Gaussian blurring.

2.5 Perceptual Video Coding

Over the past decade, there has been an exponential increase in the demand for digital video services such as high-definition television, web-based television, video conferencing and video-on-demand. To facilitate these services, it demands to significantly reduce the storage space and bandwidth of visual content production, storage and delivery. Therefore, there has been a strong desire of more effective video coding techniques beyond H.264/AVC. The main objective of video coding is to minimize the perceptual distortion D of the reconstructed video with the number of used bits R subjected to a constraint R_c. This can

be expressed as

$$\min\{D\} \quad \text{subject to} \quad R \leq R_c$$

This is a typical constrained optimization problem, and is generally solved using two methods: *Lagrangian optimization* and *dynamic programming*. In practice, the computation complexity of dynamic programming is often too high, and so this method is used only when direct Lagrangian optimization is difficult.

Lagrangian optimization technique converts the constrained optimization problem (2.5) to an unconstrained optimization problem [140], which can be expressed as

$$\min\{J\} \quad \text{where} \quad J = D + \lambda \cdot R, \tag{2.15}$$

where J is called the Rate Distortion (RD) cost and the rate R is measured in number of bits per pixel. λ is known as the Lagrange multiplier and controls the trade-off between R and D.

The distortion introduced by quantization in lossy video coding is content-dependent due to visual masking effects. By exploiting these effects, the design video coding algorithms which are able to reduce the coding bitrate for a given target perceptual quality is desirable. Many perceptual rate allocation techniques are developed based on human visual sensitivity models. The basic idea of these techniques is to allocate fewer bits to the areas or image components that can tolerate more distortions. The perceived distortion, D, is difficult to measure because our knowledge of the Human Visual System (HVS) and statistics of natural images remains limited. In practice, distortion models such as Sum of Absolute Difference (SAD) and Mean Squared Error (MSE) are used in most actual comparisons [54]. Many RDO algorithms have been proposed along this line. The representative work includes rate distortion optimized transform [204], rate distortion optimized quantization [68] and the dependent joint RDO using soft decision quantization [189, 190]. However, the distortion measures such as SAD and MSE are widely criticized for not correlating well with perceived quality [166].

Since the distortion in video coding mainly originates from quantization, many recent methods attempt to incorporate the properties of the HVS into the quantization process [28, 31, 91, 143, 144, 150]. Because HVS has different sensitivities to different frequencies,

the concept of frequency weighting has been incorporated in the quantization process in many picture coding standards from JPEG to H.264/AVC high profile [28,143,144,150]. In [31,170], foveated vision models were employed for optimizing the quantization parameter and Lagrange multiplier. However, these methods are based on near threshold perceptual models, but practical video coding typically works in a suprathreshold range [22,100,198], where the perceptual quality behavior is poorly predicted from the threshold level.

In the literature, significant progress has been made to adapt λ on a frame level when MSE is used as the distortion measure. In [29], Chen *et al.* developed an adaptive λ estimation algorithm by modeling the R and D in the ρ domain, where ρ is defined as the percentage of zero coefficients among quantized transform residuals [58]. In [79], Laplace distribution based rate and distortion models were established to derive λ for each frame dynamically.

Many rate control algorithms such as those in [66,155] showed that better performance and rate control can be achieved by modifying λ on an MB level rather than having the same Lagrange multiplier for all MBs in a frame. In [201] and [200], the authors claimed that fixing the same Lagrange multiplier for the whole frame may not be accurate enough to capture the nature of motion, and therefore a context-adaptive Lagrange multiplier (CALM) selection scheme was introduced. However, all these methods ignore the perceptual aspect in the RDO scheme by adopting SAD/MSE as the measures of perceived distortion.

Recently, a number of video coding methods aiming to incorporate the properties of the HVS have been proposed. Yang *et al.* proposed a Just Noticeable Distortion (JND) model for motion estimation and a residue filtering process in [193,194]. A foveated JND model was employed in [31] for optimizing the quantization parameter and Lagrange multiplier. In [147,148], the authors exploited non-uniform spatial-temporal sensitivity characteristics and developed visual sensitivity models which are based on the visual cues such as motion and textural structures. In [141], motion attention, position, and texture structure models were used in the rate distortion optimization (RDO) process to adapt the Lagrange multiplier based on the content of each MB. To incorporate perceptual information into the MB-based adaptive RDO scheme, three distortion sensitivity models were built into the RDO framework in [141]. Pan *et al.* proposed a content complexity based Lagrange

multiplier selection scheme for scalable video coding [99].

In the literature, significant progress has also been made to incorporate the RDO scheme into the quantization process. In order to achieve an optimal frequency domain bit allocation, the quantization matrix is parameterized and optimized from a RDO point of view for MPEG-2 video coding [75]. In [189, 190], the authors proposed novel soft decision quantization techniques and developed joint RDO algorithms for the hybrid video coding framework. The rate distortion optimized quantization (RDOQ) algorithm for H.264 video coding is also widely accepted because of its simplicity and efficiency [68].

Since SSIM has been proven to be more effective in quantifying the suprathreshold compression artifacts, such as artifacts that distort the structure of an image [9], it was incorporated into motion estimation, mode selection and rate control in hybrid video coding [3, 27, 62, 63, 84–86, 98, 138, 187, 188]. For intra frame coding, new SSIM-based RDO schemes were proposed in [3, 85, 86]. SSIM-based RDO schemes for inter frame prediction and mode selection were developed in [84, 187, 188]. However, following the method proposed in [181], the Lagrange multiplier was determined only by QP values in these schemes. Recently, content-adaptive Lagrange multiplier selection schemes were proposed in [27, 62, 63, 138]. These algorithms employed a rate-SSIM curve to describe the relationship between SSIM and rate, which is given by:

$$D = \zeta R^{\varepsilon} \tag{2.16}$$

where ζ and ε are two fitting parameters that account for the R-D characteristics. Subsequently, the key frames are identified and encoded twice with MSE-based RDO in the sequences to obtain the best parameters ζ and ε. However, two-pass encoding of the key frames brings additional complexities to the encoder. More importantly, this scheme is based on the assumption of constant R-D characteristics in a short time period and uses a periodic refreshment technique to refresh the parameters, which may not be accurate in general.

2.6 Image Restoration using Sparse Representations and Non-Local Means

Image restoration problems are of particular interest to image processing researchers, not only for their practical value, but also because they provide an excellent test bed for image modeling, representation and estimation theories. When addressing general image restoration problems with the help of a Bayesian approach, an image prior model is required. Traditionally, the problem of determining suitable image priors has been based on a close observation of natural images. This tradition leads to simplifying assumptions such as spatial smoothness, low/max-entropy or sparsity in some basis sets. Recently, a new approach has been developed for learning the prior based on sparse representations. Using redundant representations and sparsity as driving forces for de-noising of signals has drawn a lot of research attention in the past decade or so. At first, sparsity of the unitary wavelet coefficients was considered, leading to the celebrated shrinkage algorithm [21, 45, 47, 95, 132]. With the growing realization that regular separable 1-D wavelets are inappropriate for handling images, several new tailored multiscale and directional redundant transforms were introduced, including the curvelet [17], contourlet [41], wedgelet [44] and the steerable wavelet [133]. In parallel, the introduction of the matching pursuit [101] and the basis pursuit de-noising [30] gave rise to the ability to address the image de-noising problem as a direct sparse decomposition technique over redundant dictionaries. All these advances led to what are considered some of the best available image de-noising methods [38, 49, 87].

Recently an example based learning approach has been adapted whereby a dictionary is learned either from the corrupted image or a high-quality set of images with the assumption that it can sparsely represent any natural image. Thus, this learned dictionary encapsulates the prior information about the set of natural images. Such methods have proven to be quite successful in performing image restoration tasks such as image de-noising [49] and image super-resolution [192, 199]. More specifically, an image is divided into overlapping blocks with the help of a sliding window and subsequently each block is sparsely coded with the help of the dictionary. The dictionary, ideally, models the prior of natural images and is therefore free from all kinds of distortions. As a result, the reconstructed blocks, obtained by the linear combination of the atoms of the dictionary, are distortion free. Finally, the

blocks are put back into their places and combined in light of a global constraint for which a minimum MSE solution is reached. The accumulation of many blocks at each pixel location might affect the sharpness of the image. Therefore, the distorted image must also be considered in order to reach the best compromise between sharpness and admissible distortions.

Image self-similarity is an important concept in the image processing literature as pixel-blocks of a natural image can be approximated by other blocks. A de-noising method which makes use of self-similarity of images known as Non-Local Means (NLM) was recently proposed in [13,14]. The NLM de-noising filter estimates a sample of the underlying noise-free "source" by weighted averaging other "target" samples in the noisy image. NLM image de-noising method does not put any restriction on magnitude of the weights assigned to target samples in the close proximity of the source sample as opposed to kernel smoothing schemes. Instead, the de-noising algorithm calculates the weight of each target sample based on its similarity with the source sample. The NLM de-noising algorithm has the ability to outperform classical de-noising methods, including Gaussian smoothing, Wiener filtering, Total Variation filtering [120], wavelet thresholding [24], and anisotropic diffusion [7]. Furthermore, an extension of the algorithm has been developed to address the problem of de-noising image sequences [15].

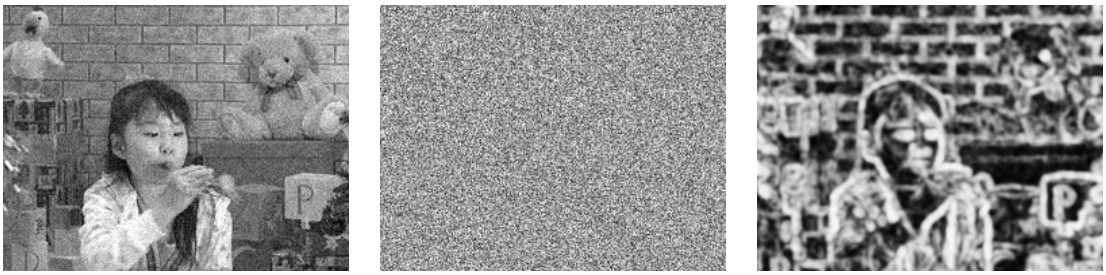
2.7 Time-Varying Subjective Video Quality

Video quality assessment has been an active subject of study in the past decades [163], but how human visual quality-of-experience (QoE) changes with time-varying video quality (in the scale of seconds or longer, rather than frames) is still an unresolved issue. Although quite many video quality databases have been built and subjective experiments conducted to study spatial and temporal video quality, they are not directly applicable in developing and validating computational models of time-varying video quality, because most video sequences in these databases consist of one scene or occasionally a few scenes of similar content and distorted in similar fashion, and thus in the scale of seconds or longer, they have fairly stable quality. Much less has been done in the area of predicting perceptual

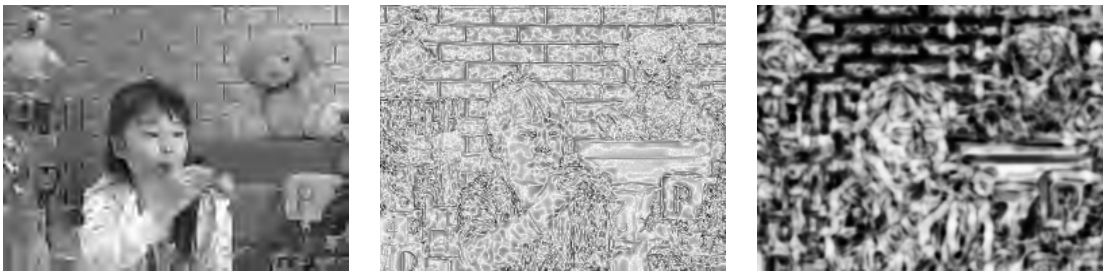
experience of time-varying video quality. Viewer response to time-varying video quality using a single stimulus continuous quality evaluation (SSCQE) in light of forgiveness, recency, and negative-peak and duration-neglect effects were studied in [102]. The findings of this study were applied in the form of an infinite impulse response (IIR) filter model for pooling in [4]. Asymmetric and smooth tracking of time-varying video quality by human subjects was observed and modeled in [146]. Temporal summation based on recursive formulations was used to model the low pass nature of the perceived continuous video quality [93] and hysteresis effect [125]. The historical experiences of the users' satisfaction while consuming a certain video streaming stimulus is modeled and quantified for web QoE in [61] and for VoIP in [112]. These models employ support vector machines and iterative exponential regeression to account for the memory effect. The difference in successive MOS values is exponentially weighted in a symmetric fashion as long as the difference is below a certain threshold. [205] investigates the human perception of variations in layer encoded video resulting in time-varying quality characteristics. Recently, the problem of video quality assessment with dynamically varying distortion on mobile devices was studied in [94].

(a) Global brightness shift (b) Absolute Error map (c) SSIM map

(d) Global constrast strech (e) Absolute Error map (f) SSIM map



(g) Gaussian Noise (h) Absolute Error map (i) SSIM map



(j) JPEG2000 compression (k) Absolute Error map (l) SSIM map

Figure 2.7: Comparison between MSE and SSIM local quality maps. In both maps, brighter indicates better local quality (or lower distortion)

(a) MSE = 0, SSIM = 1

(b) MSE = **128**, SSIM = **0.8416**

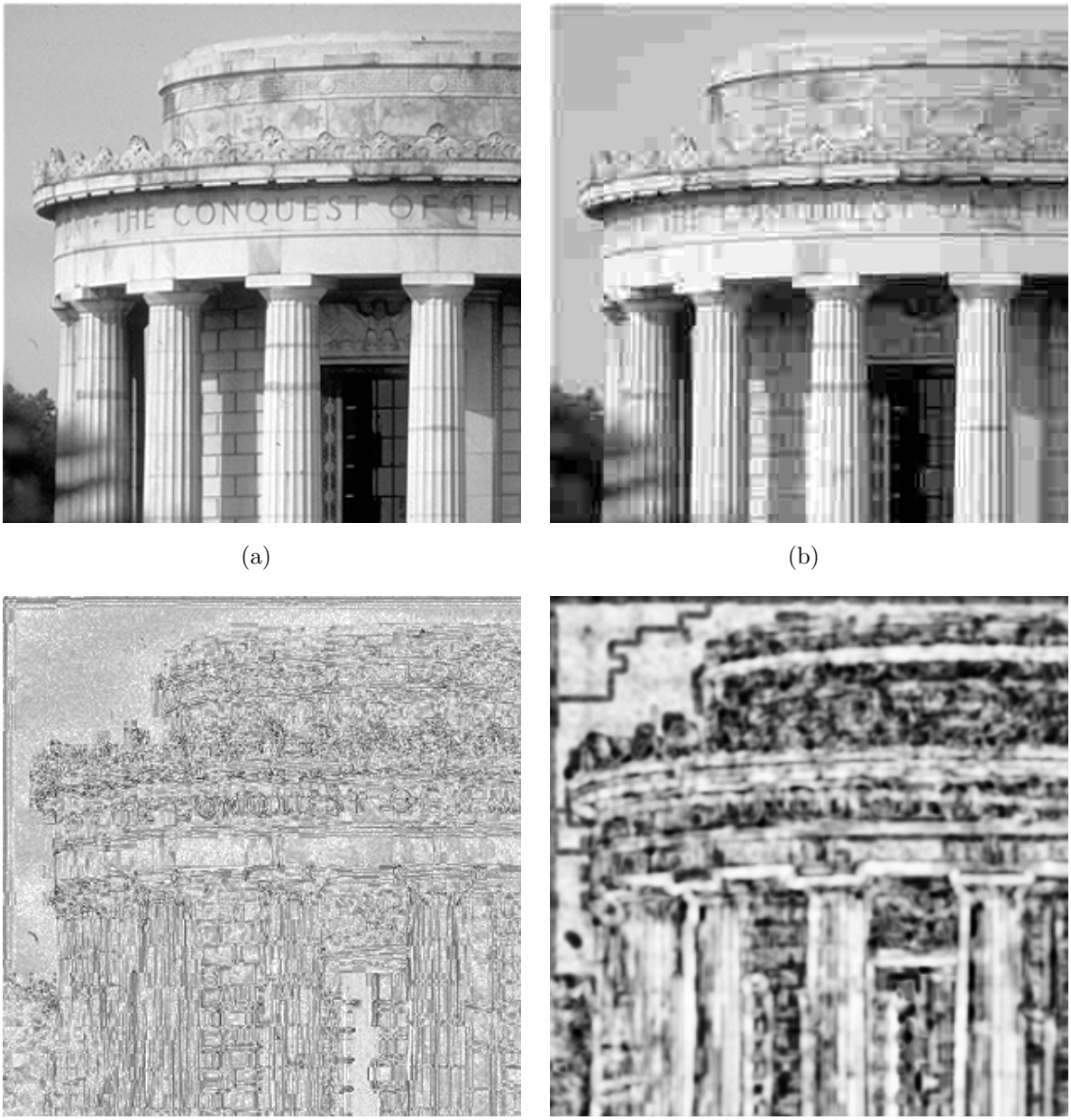
(c) MSE = 72.81, SSIM = **0.8416**

(d) MSE = 6517, SSIM = **0.8416**

(e) MSE = **128**, SSIM = 0.9959

(f) MSE = **128**, SSIM = 0.7316

Figure 2.8: MAD competition between MSE and SSIM as image quality assessment methods



(a) (b)

(c) (d)

Figure 2.9: An original image (a) is compressed by JPEG (b). The absolute error map and the SSIM quality map are shown in (c) and (d), respectively. In both maps, brighter indicates better local quality (or lower distortion).

Chapter 3

Reduced-Reference SSIM Estimation

This chapter presents a Reduced Reference Image Quality Assessment (RR-IQA) algorithm that approximates Full Reference (FR) SSIM by making use of Divisive Normalization Transform (DNT) domain image statistical properties and the design principle of the SSIM approach. We demonstrate the novel concept of image repairing by iteratively matching the DNT-domain statistical properties (available as RR features) of the reference image. The method presented has a fairly low RR data rate (36 scalar features per image in the current implementation) and has good potentials to be employed in visual communications applications for quality monitoring, streaming, and image repairing tasks.

3.1 Introduction

RR-IQA method only requires a limited number of RR features extracted from the reference for the IQA task [173] and provides an interesting compromise between FR and NR approaches in terms of both quality prediction accuracy and the amount of information required to describe the reference. A general framework for the use of RR-IQA in visual communications along with image-repairing capability is shown in Fig. 3.1. An image x is transmitted to the receiver via a transmission channel, which introduces distortions in the received image y. Meanwhile, RR features X extracted at the transmitter side are sent

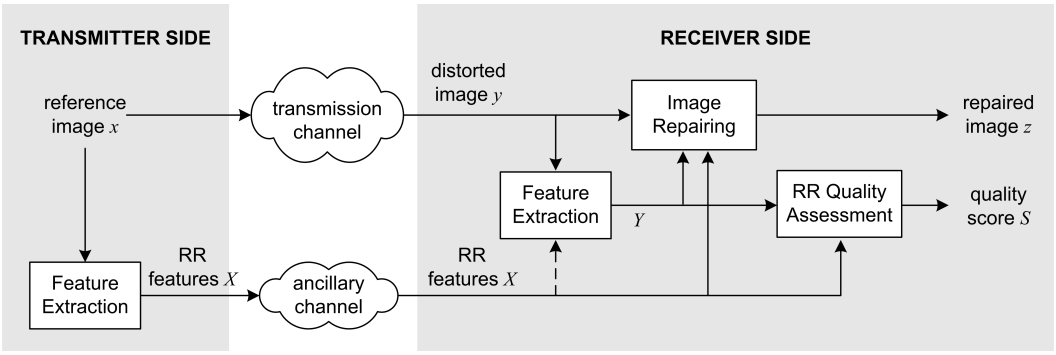


Figure 3.1: General framework for the deployment of RR-IQA systems with image repairing capability.

to the receiver through an ancillary channel. The feature extraction unit at the receiver side calculates the features Y from the received image y in a similar fashion as in the transmitter side. The receiver can use Y as the side information to decode X, which would further reduce the data rate required to transmit X. This option is depicted by a dotted line connecting received RR features with feature extraction algorithm at the receiver side. X and Y are compared at the quality assessment unit, which creates a quality score S of the distorted image y. A good RR-IQA approach should achieve a good trade-off between rate and accuracy. In general, the larger the rate of the RR features, the more accurate the RR-IQA measure can achieve. In the extreme, when the rate is enough to fully reconstruct the reference, RR-IQA converges to FR-IQA. The performance gap between RR- and FR-IQA may be reduced by selecting RR features that are efficient, perceptually relevant, and sensitive to various kinds of distortions. In addition, since the RR features provide information about what the "correct" image is supposed to look like, they may also be used as side information to repair the received distorted image, as illustrated in Fig. 3.1.

Our work here focuses on general-purpose RR-IQA based on *natural image statistics* modeling [136]. In addition, motivated by the success of the FR SSIM index, we develop our method as an attempt to estimate SSIM rather than directly predicting subjective quality. The benefits of this approach are twofold. First, the successful design principle in the construction of SSIM can be naturally incorporated into the development of the RR algorithm. Second, when the algorithm design involves a supervised learning stage, it is

much easier to obtain training data, because SSIM can be readily computed, as opposed to the expensive and time-consuming subjective evaluations. The advantages of our methods are threefold. First, our method is based on natural image statistical modeling and makes use of the perceptually and statistically motivated DNT transform. Second, instead of decomposing the problem of SSIM estimation into many local problems and estimating each component in SSIM expression separately [145], our method uses global statistics to estimate global SSIM value. This allows for a much more efficient description of the image content, and thus significantly lowers the number of RR features. Third, our approach aims for general-purpose RR-IQA that can be applied to assess images with a wide variety of distortion types.

The value of RR-IQA measures is beyond quality evaluations. As illustrated in Fig. 3.1, they may also be employed to partially "repair" the distorted image. Here, we attempt to repair an image by matching the subband statistical properties of the distorted image with those of the reference and use deblurring as an example to demonstrate the idea. The interesting feature of this method is that it requires no knowledge about the blur kernel. Instead, the same repairing procedure is successful to correct images of not only homogeneous blur (e.g., out-of-focus blur), but also directional blur (e.g., motion blur).

3.2 RR-SSIM Estimation

The proposed RR-SSIM estimation algorithm starts from a feature extraction process of the reference image based on a multi-scale multi-orientation divisive normalization transform (DNT). Divisive normalization was found to be an effective mechanism to account for many neuronal behaviors in biological perceptual systems [59, 135, 153]. It also provides a useful model to describe the psychophysical visual masking effect [51, 179]. DNT is typically applied after a multi-scale linear transform (loosely referred to as wavelet transform) that decomposes the image into transform coefficients representing localized structures in space, frequency (scale) and orientation. The DNT-domain representation of the image is then calculated by dividing each coefficient by a local energy measure based on its neighboring coefficients. It was found that the histogram of DNT coefficients within a wavelet subband

can often be well fitted with a zero-mean Gaussian density function [76, 154], which is a one-parameter function that allows for efficient summarization of the statistics of the reference image. In [76], the effect of image distortions on the statistics of DNT coefficients was studied. It was found that different types of distortions modify the statistics of the reference image in different ways, and the levels of statistical differences may be used to quantify image distortions. In order to estimate FR SSIM, we desire the variations of the statistics of the DNT coefficients with respect to different types and levels of distortions to be coherent with the corresponding effects on FR SSIM.

The Gaussian scale mixture (GSM) model provides a convenient framework to define a DNT [154]. A vector Y of length N is regarded as a GSM if it can be represented as the product of two independent components: $Y \doteq zU$, where z is a scalar random variable called mixing multiplier, and U is a zero-mean Gaussian-distributed random vector with covariance C_U. In image processing applications, GSM may be used to model a cluster of wavelet coefficients that are neighbors in space, scale and orientation. If we assume that z takes a fixed value for each cluster but varies across the image, then putting all z values together constitutes a variance field. The DNT can then be accomplished by $\nu = Y/z$, which produces a random vector that is Gaussian. This had been observed in empirical studies in [154], where z is replaced by a local estimation $\hat{z}$ using a maximum-likelihood estimator [154]:

$$\hat{z} = \arg\max_{z} \{\log p(Y|z)\} = \sqrt{Y^T C_U^{-1} Y / N} \tag{3.1}$$

The Gaussianization produced by the DNT process largely reduces the complication in describing the distribution of the subband coefficient x:

$$p_m(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{x^2}{2\sigma^2}) \tag{3.2}$$

where only a single parameter σ needs to be recorded for each subband.

In addition to σ, the Kullback-Leibler divergence (KLD) [36] between model Gaussian distribution, $p_m(x)$, and the true probability distribution of the DNT-domain coefficients,

$p(x)$, denoted by $d(p_m||p)$ is extracted as the second feature for each subband:

$$d(p_m||p) = \int p_m(x) \log \frac{p_m(x)}{p(x)} dx \tag{3.3}$$

This improves model accuracy when the probability distribution is not exactly Gaussian.

The subband distortion of the distorted image can be evaluated by the KLD between the probability distribution of the original image, $p(x)$, and that of the distorted image, $q(x)$:

$$d(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \,. \tag{3.4}$$

Direct computation of this quantity requires full access to $p(x)$, which would require a large number of RR features to describe. Fortunately, the Gaussian model of the DNT coefficients (3.2) provides a good approximation. Therefore, we can estimate $p(x)$ by

$$\hat{d}(p||q) = \int p_m(x) \log \frac{p(x)}{q(x)} dx \tag{3.5}$$

$$= d(p_m||q) - d(p_m||p) \,, \tag{3.6}$$

where $d(p_m||q)$ is the KLD between the model Gaussian distribution and the distribution computed from the distorted image. Although different types of distortions affect the statistics of the reference image in different manners, they are all summarized in (3.6) to a single distortion measure. An added nice feature of this measure is that it equals zero when the two distributions $p(x)$ and $q(x)$ are identical.

At the receiver side, the KLD between the subband coefficient probability distributions of the original and distorted images is calculated as in (3.6). By assuming independence between subbands, the subband-level distortion measure of (3.6) can be combined to provide an overall distortion assessment of the whole image by

$$D = \log \left(1 + \frac{1}{D_0} \sum_{k=1}^{K} \left| \hat{d}^k(p^k||q^k) \right| \right) , \tag{3.7}$$

where K is the total number of subbands, p^k and q^k are the probability distributions of the k-th subband of the reference and distorted images, respectively, $\hat{d}^k$ represents the KLD between p^k and q^k, and D_0 is a constant to control the scale of the distortion measure.

The limitation of the measure in (3.7) is that it does not take into account the relationship (or structures) between the distortions across different subbands. Such distortion structure is a critical issue behind the philosophy of the SSIM approach [165], which attempts to distinguish structural and non-structural distortions. To understand this better, we need to look at the FR SSIM algorithm [165] given by (2.6).

Here we borrow the design philosophy of FR SSIM, but apply it to a completely different domain of image representation. In particular, we attempt to distinguish structural and non-structural changes of the cluster of statistical features extracted from the DNT coefficients from different subbands. This is intuitively sensible because the distortion that is consistent with the underlying signal in the feature vector space needs to be treated differently as compared to non-structural distortions. For example, in the case that the distorted image is a globally contrast scaled (contrast reduction or enhancement) version of the reference image, then the standard deviations of all subbands should scale by the same factor, which is considered consistent non-structural distortion and is less objectional than the case that the subband standard deviations change in different ways.

Let $\boldsymbol{\sigma}_r$ and $\boldsymbol{\sigma}_d$ represent the vectors containing the standard deviation σ values of the DNT coefficients from each subband in the reference and distorted images, respectively. We define a new RR distortion measure as

$$D_n = g(\boldsymbol{\sigma}_r, \boldsymbol{\sigma}_d) \log \left(1 + \frac{1}{D_0} \sum_{k=1}^{K} \left| \hat{d}^k(p^k || q^k) \right| \right) . \tag{3.8}$$

Compared with (3.7), the key difference here is the added function $g(\boldsymbol{\sigma}_r, \boldsymbol{\sigma}_d)$ in the front. This function should serve the purpose of differentiating non-structural from structural distortion directions in the feature vector space of subband σ values, so as to scale the distortion measure D in a way that penalizes more on structural than non-structural distortions. Motivated by the successful normalized correlation formulation in SSIM [165], we define

$$g(\boldsymbol{\sigma}_r, \boldsymbol{\sigma}_d) = \frac{\|\boldsymbol{\sigma}_r\|^2 + \|\boldsymbol{\sigma}_d\|^2 + C}{2(\boldsymbol{\sigma}_r \cdot \boldsymbol{\sigma}_d) + C} , \tag{3.9}$$

where a positive constant C is included to avoid instability when the dot product $\boldsymbol{\sigma}_r \cdot \boldsymbol{\sigma}_d$ is close to 0. This function is lower-bounded by 1, when $\boldsymbol{\sigma}_r$ and $\boldsymbol{\sigma}_d$ are fully correlated,

or in other words, when their directions in the feature vector space are completely aligned (corresponding to non-structural distortions). With the decrease of correlation, $g(\boldsymbol{\sigma}_r, \boldsymbol{\sigma}_d)$ increases, and thus gives more penalty to structural distortions.

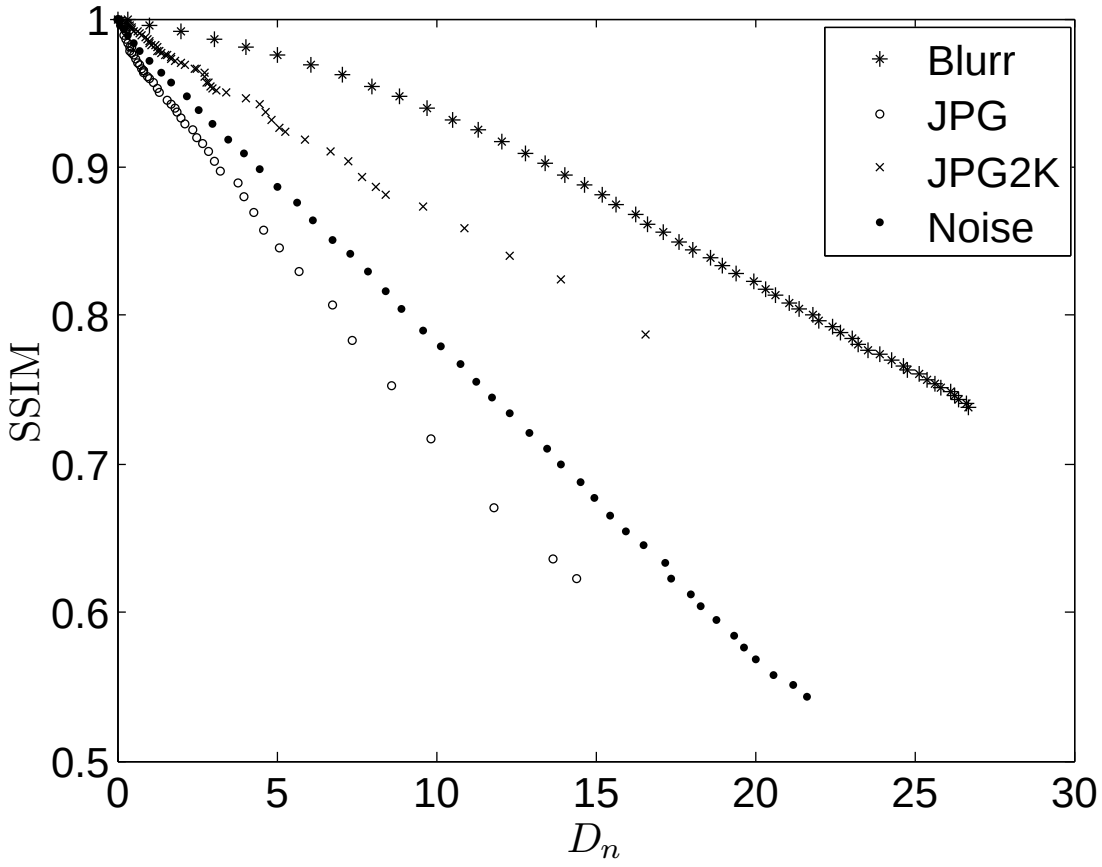


Figure 3.2: Relationship between D_n and SSIM for blur, JPEG compression, JPEG2000 compression, and noise contamination distortions for *Lena* image.

Figure 3.2 plots the D_n values computed using distorted images from the LIVE database [130] for four common distortion types at different distortion levels, and compares them with the corresponding FR SSIM values. Interestingly, for each fixed distortion type, D_n exhibits a nearly perfect linear relationship with SSIM. We regard this as a consequence of the similarity between their design principle, even though the principle is applied to

completely different domains of signal representation. The clean linear relationship helps reduce the SSIM estimation problem to the estimation of the slope factor. Once the slope is determined, we can then use the following straight-line relationship to estimate SSIM:

$$\hat{S} = 1 - \alpha D_n \,. \tag{3.10}$$

The slope factor α in (3.10) varies across distortion types and needs to be learned from examples. Specifically, we adopt a regression-by-discretization approach [180], which is a regression scheme that employs a classifier on a copy of the data that has the class attribute discretized, and the predicted value is the expected value of the mean class value for each discretized interval. The training images were obtained from six image databases described in Section 3.3. The classification is performed using random forests [8] which are built using $|\sigma_r - \sigma_d|$ and $|k_r - k_d|$ values in each subband as the attributes, where k_r and k_d are the kurtosis values of the DNT coefficients computed from the reference and distorted images, respectively. It has been observed with the help of the ground truth data that the values of α tend to lie in various closely packed clusters. Each cluster may contain images belonging to one distortion type. It provides a natural order to the distortion types and therefore does not require an undesirable distortion classification stage which limits the generalization capability of the proposed method. Therefore, the proposed method has the potential to extrapolate to extended distortion types that may not be included in the training samples.

The specification of our implementation is as follows: To extract RR features, the reference image is first decomposed into 12 subbands using a three-scale four-orientation steerable pyramid decomposition [134], a type of redundant wavelet transform that avoids aliasing in subbands. DNT is then performed using 13 neighboring coefficients, including 9 spatial neighbors from the same subband, 1 from parent subband, and 3 from the same spatial location in the other orientation bands at the same scale. The value of the constant C in (3.9) is set to 0.1, which is found to be an insensitive parameter in terms of the performance of the proposed IQA measure. Three features, σ_r, k_r and $d(p_m||p)$, are extracted for each subband, resulting in a total of 36 scalar RR features for a reference image.

3.3 Validation of RR-IQA Algorithm

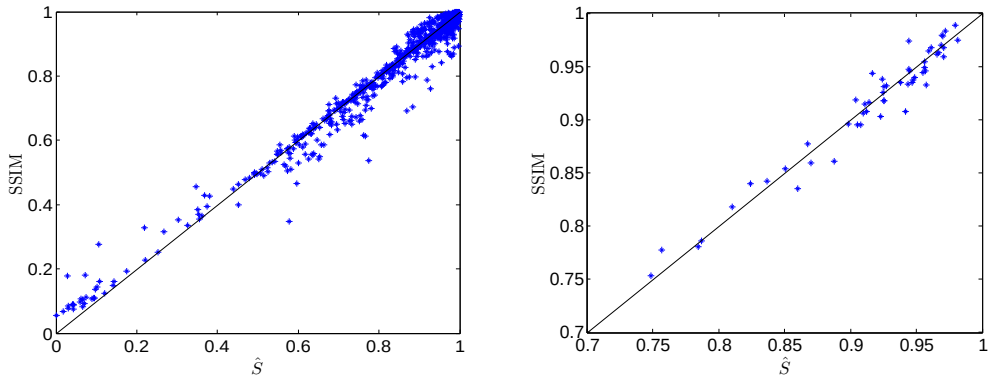
To validate the proposed RR-SSIM algorithm, we first test how well it predicts FR SSIM. Figure 3.3 shows the scatter plots obtained using all six databases, where each point in the plots represent one test image, and the vertical and horizontal axes are FR-SSIM and RR-SSIM, respectively. If the prediction is perfect, then the point should lie on the diagonal line. To provide a quantitative measure, Table 3.1 computes the mean absolute error (MAE) and Pearson linear correlation coefficient between FR SSIM and our RR-SSIM estimate. It can be observed that for all databases, the points are scattered close to the diagonal lines in Fig. 3.3 and the correlation coefficients are above 0.9, indicating good prediction accuracy of the proposed method. The breakdown prediction performance for individual distortion types in different databases are provided in Tables 3.2.

Table 3.1: MAE and PLCC comparisons between SSIM and RR SSIM estimation for six databases

Database	MAE	PLCC
LIVE [130]	0.0317	0.9432
Cornell A57 [23]	0.0266	0.9299
IVC [74], [97]	0.0244	0.9211
Toyama-MICT [60]	0.0119	0.9405
TID2008 [107], [106]	0.0303	0.9004
CSIQ [72]	0.0339	0.9243

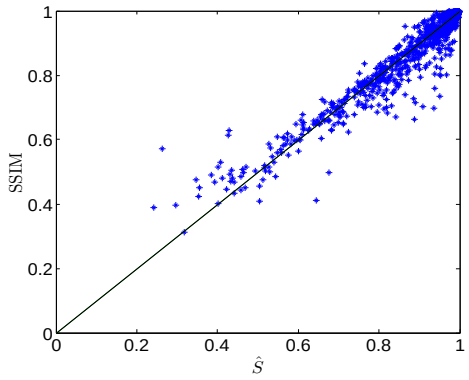
The ultimate goal of RR-IQA algorithms is to predict subjective quality evaluation of images. Therefore, the more important test is to evaluate how well they predict subjective scores. For this purpose, we use five evaluation metrics to assess the performance of IQA measures:

- Pearson linear correlation coefficient (PLCC) after a nonlinear mapping between the subjective and objective scores. For the i-th image in an image database of size N, given its subjective score o_i (mean opinion score (MOS) or difference of MOS (DMOS) between reference and distorted images) and its raw objective score r_i, we

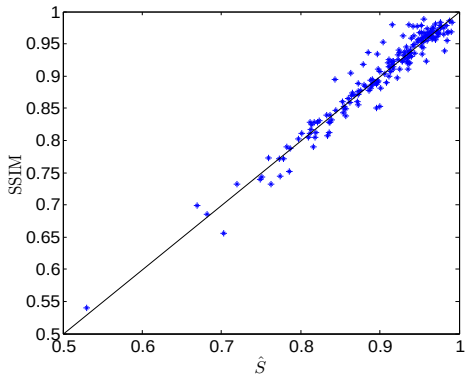


(a) LIVE Image Database

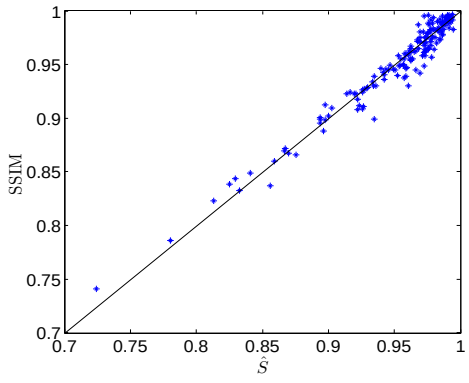
(b) Cornell A57 Database



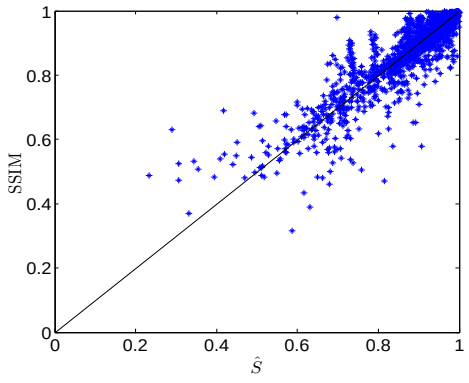
(c) CSIQ Database



(d) IVC Database



(e) Toyama-MICT Database



(f) TID 2008 Database

Figure 3.3: Scatter plots of SSIM versus RR-SSIM estimation $\hat{S}$ for six test databases.

Table 3.2: Distortion type breakdown for MAE and PLCC comparisons between SSIM and RR-SSIM estimation

Distortion Type	Database	MAE	PLCC
Additive Gaussian noise	LIVE	0.0340	0.9903
	TID2008	0.0185	0.9522
	CSIQ	0.0274	0.9771
Noise in color comp.	TID2008	0.0080	0.8978
Spatially corr. noise	TID2008	0.0331	0.9580
Masked noise	TID2008	0.0057	0.5982
High frequency noise	TID2008	0.0227	0.9621
Additive pink noise	CSIQ	0.0212	0.9712
Impulse noise	TID2008	0.0222	0.9667
Quantization noise	TID2008	0.0316	0.7584
Gaussian blur	LIVE	0.0412	0.8973
	IVC	0.0342	0.9288
	TID2008	0.0416	0.8892
	CSIQ	0.0260	0.9783
Image de-noising	TID2008	0.0444	0.8721
JPEG compression	LIVE (Set 1)	0.0214	0.9867
	LIVE (Set 2)	0.0235	0.9840
	IVC	0.0141	0.9476
	Toyama-MICT	0.0144	0.9007
	TID2008	0.0253	0.9325
	CSIQ	0.0490	0.8895
JPEG2000 compression	LIVE (Set 1)	0.0197	0.9820
	LIVE (Set 2)	0.0229	0.9792
	IVC	0.0296	0.9321
	Toyama-MICT	0.0093	0.9472
	TID2008	0.0482	0.9009
	CSIQ	0.0452	0.9223
LAR compression	IVC	0.0227	0.9426
JPEG trans. error	TID2008	0.0420	0.8990
Non ecc. patt. noise	TID2008	0.0149	0.8863
Local block-wise dist.	TID2008	0.0117	0.8837
Mean shift	TID2008	0.0367	0.8205
Contrast change	TID2008	0.0485	0.7085
	CSIQ	0.0372	0.9486

first apply a nonlinear function to r_i [129]

$$q(r) = a_1 \left\{ \frac{1}{2} - \frac{1}{1 + \exp\left[a_2(r - a_3)\right]} \right\} + a_4 r + a_5 \tag{3.11}$$

where a_1 to a_5 are model parameters found numerically using a nonlinear regression process in MATLAB optimization toolbox to maximize the correlations between subjective and objective scores. The PLCC value can then be computed as

$$\text{PLCC} = \frac{\sum_i (q_i - \bar{q}) * (o_i - \bar{o})}{\sqrt{\sum_i (q_i - \bar{q})^2 * \sum_i (o_i - \bar{o})^2}}. \tag{3.12}$$

- Mean absolute error (MAE) is calculated using the converted objective scores after the nonlinear mapping described above:

$$\text{MAE} = \frac{1}{N} \sum |q_i - o_i|. \tag{3.13}$$

- Root mean-squared (RMS) error is computed similarly as

$$\text{RMS} = \sqrt{\frac{1}{N} \sum (q_i - o_i)^2}. \tag{3.14}$$

- Spearman's rank correlation coefficient (SRCC) is defined as:

$$\text{SRCC} = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}, \tag{3.15}$$

where d_i is the difference between the i-th image's ranks in subjective and objective evaluations. SRCC is a nonparametric rank-based correlation metric, independent of any monotonic nonlinear mapping between subjective and objective scores.

- Kendall's rank correlation coefficient (KRCC) is another non-parametric rank correlation metric given by

$$\text{KRCC} = \frac{N_c - N_d}{\frac{1}{2} N(N - 1)}, \tag{3.16}$$

where N_c and N_d are the numbers of concordant and discordant pairs in the data set, respectively.

Among the above metrics, PLCC, MAE and RMS are adopted to evaluate prediction accuracy [152], and SRCC and KRCC are employed to assess prediction monotonicity [152]. A better objective IQA measure should have higher PLCC, SRCC and KRCC while lower MAE and RMS values. All of these evaluation metrics are adopted from previous IQA studies [105, 129, 152]. Only the distorted images in the six databases described in Section 2.4.2 were employed in our tests (i.e., reference images are excluded). This avoids several difficulties in computing the evaluation metrics. Specifically, the reference images have infinite PSNR value, making it difficult to perform nonlinear regression and compute PLCC, MAE and MSE values. In addition, since all reference images are assumed to have perfect quality, there are no natural relative ranks between them, resulting in ambiguities when computing SRCC and KRCC metrics.

The test results are given in Tables 3.3, 3.4 and 3.5. To provide background comparisons, we have also included in the tables four other objective IQA algorithms, among which two are FR-IQA measures namely peak signal-to-noise-ratio (PSNR) and SSIM, and three are RR-IQA measures, which are wavelet marginal-based method [174] and DNT marginal-based method [76]. Other RR-IQA methods are not included in the comparison because they are not designed and tested for general-purpose applications. Although it is unfair to compare RR- with FR-IQA measures, the PSNR and SSIM results supply useful references on the current status of RR approaches. To provide an overall evaluation of the IQA algorithms, we also calculate the direct and weighted average of PLCC, SRCC and KRCC values across all six databases (where the weight assigned to a database is determined by the number of test images in a database). The average results are given in Table 3.5. It can be seen that in general the proposed RR-SSIM method performs moderately inferior to SSIM (which is as expected) but significantly outperforms PSNR and the other RR-IQA methods under comparison.

Statistical significant analysis has been carried out based on variance-based hypothesis testing, which follows the approach introduced in [128] and subsequently adopted by many later papers in the literature. Specifically, the residual difference, between the DMOS and the predicted quality given by each objective IQA algorithm, is assumed to be Gaussian distributed and F-statistic is employed to compare the variances of two sets of sample points. With such a test, we can make a statistically sound judgement of superiority

Table 3.3: Performance comparisons of IQA measures using LIVE, IVC and TID 2008 databases

		LIVE Database (779 Images) [130]				
IQA measure	Type	PLCC	MAE	RMS	SRCC	KRCC
PSNR	FR	0.8721	10.5248	13.3683	0.8755	0.6863
SSIM [165]	FR	0.9448	6.9324	8.9455	0.9479	0.7962
Wavelet Marginal [178]	RR	0.8226	10.5248	13.3683	0.8755	0.6863
DNT Marginal [76]	RR	0.9173	9.7321	11.7862	0.8973	0.7126
RR-SSIM	RR	0.9194	9.1889	11.3026	0.9129	0.7349
		IVC Database (185 Images) [74], [97]				
IQA measure	Type	PLCC	MAE	RMS	SRCC	KRCC
PSNR	FR	0.6719	0.7190	0.9023	0.6884	0.5217
SSIM [165]	FR	0.9119	0.3776	0.4999	0.9018	0.7223
Wavelet Marginal [178]	RR	0.5311	0.8550	1.0322	0.4114	0.2907
DNT Marginal [76]	RR	0.6294	0.7876	0.9466	0.5928	0.4210
RR-SSIM	RR	0.8177	0.5619	0.7014	0.8154	0.6164
		TID 2008 Database (1700 Images) [106, 107]				
IQA measure	Type	PLCC	MAE	RMS	SRCC	KRCC
PSNR	FR	0.5232	0.8683	1.1435	0.5530	0.4027
SSIM [165]	FR	0.7731	0.6546	0.8510	0.7749	0.5767
Wavelet Marginal [178]	RR	0.5891	0.8666	1.0843	0.5119	0.3589
DNT Marginal [76]	RR	0.5746	0.8473	1.0982	0.5597	0.4093
RR-SSIM	RR	0.7231	0.7190	0.9270	0.7210	0.5236

Table 3.4: Performance comparisons of IQA measures using Cornell A57, Toyama-MICT and CSIQ databases

		Cornell A57 Database (54 Images) [23]				
IQA measure	Type	PLCC	MAE	RMS	SRCC	KRCC
PSNR	FR	0.6346	0.1606	0.1899	0.6188	0.4309
SSIM [165]	FR	0.8017	0.1209	0.14688	0.8066	0.6058
Wavelet Marginal [178]	RR	0.5125	0.1971	0.2317	0.31398	0.2210
DNT Marginal [76]	RR	0.6635	0.1655	0.2094	0.5079	0.3623
RR-SSIM	RR	0.7044	0.1433	0.1744	0.7301	0.5345
		Toyama-MICT Database (168 Images) [60]				
IQA measure	Type	PLCC	MAE	RMS	SRCC	KRCC
PSNR	FR	0.6329	0.7817	0.9688	0.6131	0.4442
SSIM [165]	FR	0.8886	0.4385	0.5738	0.8793	0.6939
Wavelet Marginal [178]	RR	0.6542	0.7742	0.9464	0.6322	0.4570
DNT Marginal [76]	RR	0.6671	0.7548	0.9322	0.6518	0.4723
RR-SSIM	RR	0.8051	0.5648	0.7423	0.8003	0.6090
		CSIQ Database (866 Images) [72]				
IQA measure	Type	PLCC	MAE	RMS	SRCC	KRCC
PSNR	FR	0.7512	0.1366	0.1732	0.8058	0.6083
SSIM [165]	FR	0.8612	0.0991	0.1334	0.8756	0.6906
Wavelet Marginal [178]	RR	0.7124	0.1492	0.1842	0.7431	0.5457
DNT Marginal [76]	RR	0.6571	0.1642	0.1978	0.6744	0.4961
RR-SSIM	RR	0.8426	0.1092	0.1413	0.8527	0.6540

Table 3.5: Average performance of IQA measures over six databases

		Direct Average			Database-size Weighted		
IQA measure	Type	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
PSNR	FR	0.6811	0.6924	0.5157	0.6622	0.6887	0.5172
SSIM [165]	FR	0.8636	0.8643	0.6809	0.8416	0.8455	0.6615
Wavelet Marginal [178]	RR	0.6371	0.5813	0.4266	0.6651	0.6383	0.4691
DNT Marginal [76]	RR	0.6848	0.6473	0.4789	0.6729	0.6613	0.4952
RR-SSIM	RR	0.8021	0.8054	0.6121	0.7995	0.7996	0.6061

Table 3.6: Gaussianity of IQA − DMOS residuals

	LIVE	A57	CSIQ	IVC	Toyama	TID 2008
PSNR	1	1	1	1	1	1
SSIM [165]	1	1	0	0	1	1
Wavelet Marginal [174]	1	1	1	1	1	1
DNT Marginal [76]	1	1	1	1	1	1
RR-SSIM	1	1	1	0	1	1

Table 3.7: Statistical Significance matrix based on IQA − DMOS residuals

Model	PSNR	SSIM	Wavelet Marginal [174]	DNT Marginal [76]	RR-SSIM
PSNR	- - - - - -	0 - 0 0 0 0	1 - - - - 0	0 - 1 - - 0	0 - 0 0 0 0
SSIM	1 - 1 1 1 1	- - - - - -	1 1 1 1 1 1	1 1 1 1 1 1	1 - - 1 - -
Wavelet Marginal [174]	0 - - - - 1	0 0 0 0 0 0	- - - - - -	0 - - - - -	0 0 0 0 0 0
DNT Marginal [76]	1 - 0 - - 1	0 0 0 0 0 0	1 - - - - -	- - - - - -	0 - 0 0 0 0
RR-SSIM	1 - 1 1 1 1	0 - - 0 - -	1 1 1 1 1 1	1 - 1 1 1 1	- - - - - -

or inferiority of one IQA algorithm against another. A statistical significance matrix is calculated and given by Table 3.7. Each entry in the table consists of six characters which correspond to the six publicly available databases in the order of {LIVE, A57, CSIQ, IVC, Toyama, TID2008}. The symbol '-' denotes that the two IQA methods are statistically indistinguishable, '1' denotes the IQA method of the row is statistically better than that of the column, and '0' denotes that the IQA method of the column is better than that of the row. It can be observed that full-reference SSIM performs the best among the IQA algorithms under comparison and the performance of the proposed RR-SSIM algorithm is quite close to that of SSIM and is superior to all other IQA methods being compared.

The assumption of Gaussianity is verified with the help of kurtosis values obtained from the prediction residuals. As in [128], the residual values are considered to be Gaussian distributed if the kurtosis value lies between 2 and 4. The results of Gaussianity tests are given in Table 3.6, where '1' means the distribution is considered Gaussian and '0' otherwise. It can be observed that the assumption is met in most cases with a few exceptions.

To examine how the proposed RR-SSIM method performs for different distortion types, we compare it with five other recently proposed RR-IQA algorithms using individual distortion types as well as the "all data" case of the LIVE database. The results are given in Table 3.8 where the best results for each distortion type are highlighted in bold. It can be observed that the proposed method exhibits highly competitive performance in most cases.

Finally, we compare the computational complexity of the proposed RR-SSIM method with five other RR-IQA algorithms. The results are reported in Table 3.9, where we present the average time taken per image, over all the images in the LIVE database, using a computer with Intel i7 processor at 2.67 GHz (the only exception is the method

Table 3.8: Performance comparison of RR-IQA algorithms using LIVE database

Distortion	JP2(1)	JP2(2)	JPG(1)	JPG(2)	Noise	Blur	FF	All Data
	PLCC							
Wavelet Marginal [174]	0.9339	0.9488	0.8278	0.9566	0.8769	0.8395	0.9230	0.8284
DNT Marginal [76]	0.9470	0.9625	0.8228	0.9627	0.9598	**0.9523**	**0.9438**	0.8949
βW-SCM [186]	0.9514	0.9569	0.8673	0.9568	0.9755	0.9454	0.9243	0.8353
Zhang et al. [202]	0.9087	0.9511	0.9094	**0.9777**	0.8623	0.9234	0.9392	0.8744
Ma et al. [83]	0.8065	0.8819	0.8180	0.9663	0.8769	0.9092	0.9178	0.8841
RR-SSIM	**0.9597**	**0.9632**	**0.9448**	0.9761	**0.9772**	0.9154	0.9315	**0.9194**
	SRCC							
Wavelet Marginal [174]	0.9370	0.9419	0.8109	0.8936	0.8600	0.8757	0.9212	0.8270
DNT Marginal [76]	0.9439	**0.9556**	0.8246	0.8853	0.9508	**0.9599**	**0.9431**	0.8882
βW-SCM [186]	0.9495	0.9517	0.8535	0.8705	0.9715	0.9371	0.9258	0.8391
Zhang et al. [202]	0.9134	0.9495	0.9105	**0.9294**	0.8417	0.9265	0.9365	0.8832
Ma et al. [83]	0.7945	0.8717	0.8042	0.9100	0.8619	0.9214	0.8866	0.8807
RR-SSIM	**0.9555**	0.9539	**0.9493**	0.8978	**0.9642**	0.8692	0.9137	**0.9129**

by Ma et. al. [83], which was tested on a slightly faster computer). This measurement provides a rough estimate of the relative computational complexity between different RR-IQA algorithms, as no code optimization has been done. It can be seen that the proposed method takes moderately more time than most of the other methods under comparison, mainly due to the computation of the divisive normalization transform. The additional computational cost is compensated by the improved quality prediction performance, as shown in Table 3.8.

Table 3.9: Comparison of computation time using LIVE database (seconds/image)

Model	Wavelet Marginal [174]	DNT Marginal [76]	βW-SCM [186]	Zhang et. al. [202]	Ma et. al. [83]	RR-SSIM
Time	6.3719	10.3843	6.6258	3.4937	18	11.2309

3.4 Image Repairing Using RR Features

Since the RR features reflect certain properties about the reference image and these properties may be altered in the distorted image, they may be employed to partially repair the distorted image. The proposed method is different from the traditional image restoration methods we have partial knowledge of the reference image in the form RR features. Therefore, we use the term "repair" in order to differentiate the proposed method from

traditional image restoration methods. Here we provide an example that uses DNT-domain RR features to correct blurred images without any knowledge about the blur kernel.

Since blur reduces energy at mid- and high-frequencies, the subband standard deviation σ_d of DNT coefficients in the distorted image is smaller than that of the reference image σ_r. A straightforward way to enforce a "corrected" image to have the same statistical properties as the reference image is to scale up all DNT coefficients in the subband of the distorted image by a fixed scale factor, followed by an inverse DNT to create a reconstructed image. In practice, however, inverting a DNT transform is a non-trivial issue that requires specific conditions of the coefficients and may involve computationally expensive algorithms [90].

Here we propose a different approach that attempts to match DNT-domain statistics but avoids direct inversion of DNT. The idea is to use the DNT-domain statistics to estimate the scale factors and then apply them in the wavelet- rather than DNT-domain. As a result, only inverse wavelet transform is necessary, and the remaining question becomes whether the desired scale ratio in the DNT domain can be well matched by scaling in the wavelet domain. To ensure this, we apply our approach in an iterative manner, and the resulting algorithm is given by Algorithm 1. In our experiment, we find that this iterative algorithm converges quickly and typically three iterations are enough to reconstruct a stable repaired image (and thus $J = 3$ in Algorithm 1) that matches DNT domain statistics quite well. This is demonstrated in Fig. 3.4, which compares the subband histograms of the reference, distorted, and repaired DNT coefficients. It can be observed that the histogram of the scaled DNT coefficients very well approximate that of the reference image. Similar design philosophy of iteratively synthesizing images by matching desirable statistical features have been used before in the literature of texture synthesis, e.g., [108].

An interesting feature of the above image deblurring process is that it does not require any information about the blur kernel. Depending on the nature of the blur process, the energy reductions at different subbands are different. For example, out-of-focus blur may lead to uniform energy reduction in all orientation subbands while motion blur could result in more significant energy reduction along one orientation against another. Since the scale factor s in our algorithm is computed for individual subband independently, it could automatically adapt the energy correction factors based on the energy reduction occurred in individual subbands. Figure 6 provides an example, where the homogeneously Gaussian

Algorithm 1: Iterative image repairing algorithm

1. *Initialization:* Let $j = 0$, $\hat{x}_{(0)} = y$, where y is the distorted image
2. *Repeat J times*
   - *Wavelet transform:* Compute wavelet transform of $\hat{x}_{(j)}$, resulting in wavelet coefficients ω
   - *DNT stage:* Compute DNT from ω, resulting in DNT coefficients ν; For all i, in the i-th subband, calculate std of DNT coefficients σ_ν^i
   - *Scaling factor calculation:* For all i, in the i-th subband, compute the scale factor $s^i = \sigma_r^i / \sigma_\nu^i$, where σ_r^i is the std of DNT coefficients of the reference image (obtained as RR features)
   - *Wavelet coefficient scaling:* For all i, in the i-th subband, let $\omega_{new} = s^i \omega$
   - *Image reconstruction:* Compute inverse wavelet transform of ω_{new}, resulting in $\hat{x}_{(j+1)}$
   - Increase j by 1
3. *Report reconstructed image:* $\hat{x} = \hat{x}_{(J)}$

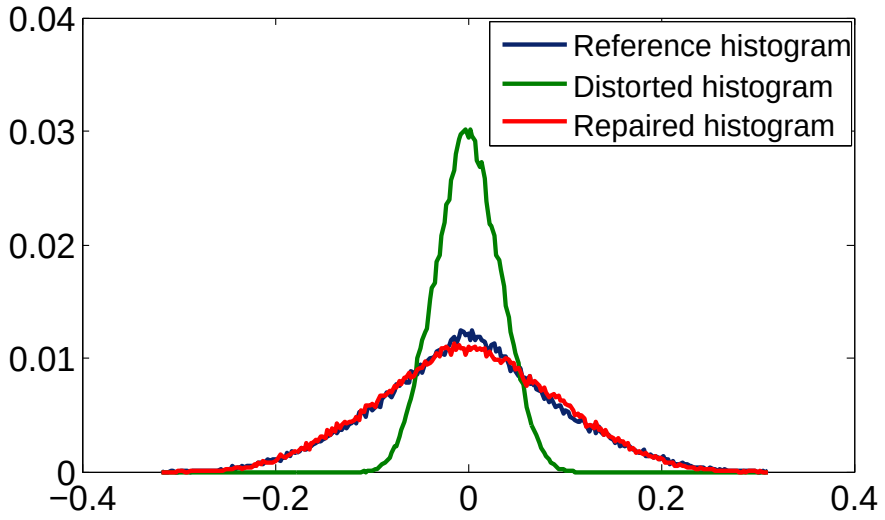
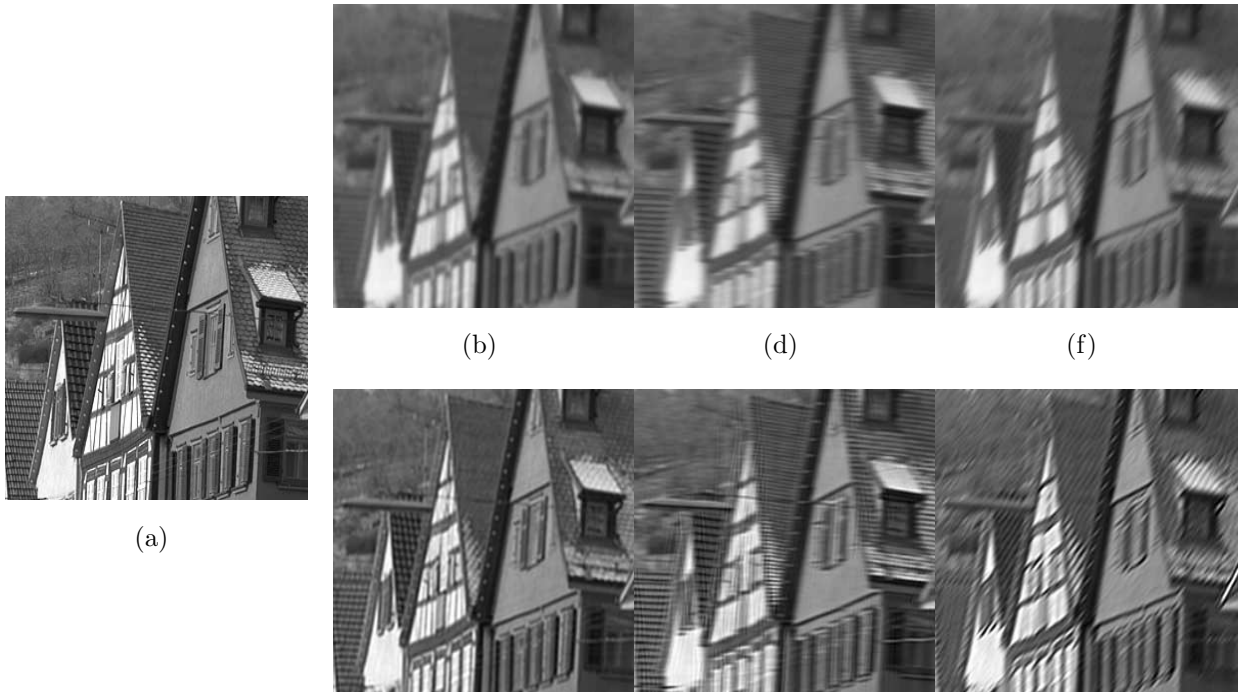


Figure 3.4: DNT coefficient histograms of original, distorted and repaired images.

blurred and directionally motion blurred images at different angles are deblurred using exactly the same image repairing algorithm described above. All repaired images appear to be much sharper and have higher contrast than their blurred versions. The visual effect is also reflected by both FR SSIM and the proposed RR-SSIM evaluations.

RR features only provide limited amount of additional information about the reference image and such information is global in the current implementation (due to the nature of the extracted RR features), thus the same repairing process may or may not work as effectively as we observe in Fig. 3.5 for the types of image distortions other than linear blur.



(a) (b) (c) (d) (e) (f) (g)

Figure 3.5: Repairing homogeneously and directionally blurred images using RR features. (a) Original "building" image (cropped for visibility); (b) Homogeneously blurred image, SSIM = 0.7389, $\hat{S} = 0.7118$; (c) Repaired image SSIM = 0.9142, $\hat{S} = 0.9327$; (d) Directionally blurred image (0 degree), SSIM = 0.6734, $\hat{S} = 0.6821$; (e) Repaired image SSIM = 0.7991, $\hat{S} = 0.8063$; (f) Directionally blurred image (45 degree), SSIM = 0.6612, $\hat{S} = 0.6324$; (g) Repaired image SSIM = 0.7896, $\hat{S} = 0.8135$.

Chapter 4

SSIM-Inspired Image Restoration Using Sparse Representation

The purpose of image restoration is to "compensate for" or "undo" distortions which affect the perceptual quality of an image. This chapter presents an image restoration algorithm that combines perceptual image fidelity measurement with optimal sparse signal representation. The objective is to use sparsity prior of the underlying signal in terms of some dictionary and achieve optimal performance in terms of SSIM. The solution for the optimal coefficients for sparse and redundant dictionary in maximal SSIM sense is presented along with a gradient descent approach to achieve the best compromise between the distorted image and the image reconstructed using sparse representation.

4.1 Introduction

MSE is usually employed as the optimization criterion for image restoration, the resulting output image might not have the best perceptual quality. This motivated us to replace the role of MSE with SSIM in the framework. The solution of this novel optimization problem is not trivial because SSIM is non-convex in nature. There are two key problems that have

to be resolved before effective SSIM-based optimization can be performed. First, how to optimally decompose an image as a linear combination of basis functions in maximal SSIM, as opposed to minimal MSE sense. Second, how to estimate the best compromise between the distorted and sparse dictionary reconstructed images for maximal SSIM. In this thesis, we provide solutions to these problems and use image de-noising and image super-resolution as applications to demonstrate the proposed framework for image restoration problems.

4.2 The Proposed Method

In this section we will incorporate SSIM as our quality measure, particularly for sparse representation. In contrast to what we may expect, it is shown that sparse representation in minimal $\mathcal{L}_2$ norm sense can be easily converted to maximal SSIM sense. We will also use a gradient descend approach to solve a global optimization problem in maximal SSIM sense. Our framework can be applied to a wide class of problems dealing with sparse representation to improve visual quality.

4.2.1 Image Restoration from Sparsity

The classic formulation of image restoration problem is as following:

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{n} \tag{4.1}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{n} \in \mathbb{R}^m$, and $\Phi \in \mathbb{R}^{m \times n}$. Here we assume $\mathbf{x}$ and $\mathbf{y}$ are vectorized versions, by column stacking, of original 2-D original and distorted images, respectively. $\mathbf{n}$ is the noise term, which is mostly assumed to be zero mean, additive, and independent Gaussian. Generally $m \leq n$ and thus the problem is ill-posed. To solve the problem assertion of a prior on the original image is necessary. The early approaches used least square (LS) [122] and Tikhonov regularization [149] as priors. Later minimal total variation (TV) solution [120] and sparse priors [49] were used successfully on this problem. Our focus in the current work is to improve algorithms, in terms of visual quality, that assert sparsity prior on the solution in term of a dictionary domain.

Sparsity prior has been used successfully to solve different inverse problems in image processing [49,88,110,192]. If our desired signal, $\mathbf{x}$, is sparse enough then it has been shown that the solution to (4.1) is the one with maximum sparsity which is unique (within some $\epsilon-$ball around $\mathbf{x}$) [18,46]. It can be easily found by solving a linear programming problem or by orthogonal matching pursuit (OMP). Not all natural signals are sparse but a wide range of natural signals can be represented sparsely in terms of a dictionary and this makes it possible to use sparsity prior on a wide range of inverse problems. One major problem is that the image signals are considered to be high dimensional data and thus, solving (4.1) directly is computationally expensive. To tackle this problem we assume local sparsity on image patches. Here, it is assumed that all the image patches have sparse representation in terms of a dictionary. This dictionary can be trained over some patches [1].

Central to the process of image restoration, using local sparse and redundant representations, is the solution to the following optimization problem [49, 192],

$$\{\hat{\boldsymbol{\alpha}}_{ij}, \hat{\mathbf{X}}\} = \underset{\boldsymbol{\alpha}_{ij}, \mathbf{X}}{\operatorname{argmin}} \ \lambda||\mathbf{X}-\mathbf{Y}||_2^2 + \sum_{ij} \mu_{ij}||\boldsymbol{\alpha}_{ij}||_0 + \sum_{ij} ||\Psi\boldsymbol{\alpha}_{ij} - \mathbf{R}_{ij}\mathbf{X}||_2^2 \tag{4.2}$$

where $\mathbf{Y}$ is the observed distorted image, $\mathbf{X}$ is the unknown output restored image, $\mathbf{R}_{ij}$ is a matrix that extracts the (ij) block from the image, $\boldsymbol{\Psi} \in \mathbb{R}^{n\times k}$ is the dictionary with $k > n$, $\boldsymbol{\alpha}_{ij}$ is the sparse vector of coefficients corresponding to the (ij) block of the image, $\hat{\mathbf{X}}$ is the estimated image, λ is the regularization parameter. In Equation 4.2, the first term is the log-likelihood global force that demands the proximity between the measured image, $\mathbf{Y}$,and its de-noised version $\mathbf{X}$. The second and the third terms are the image prior that makes sure that in the constructed image, $\mathbf{X}$, every patch in the image $\mathbf{X}$ has a sparse representation with bounded error. Since dictionary learning is limited in handling small image patches, we divide the above optimization problem into a local sparsity based model and a global constraint:

$$\hat{\boldsymbol{\alpha}}_{ij} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \ \mu_{ij}||\boldsymbol{\alpha}||_0 + ||\boldsymbol{\Psi}\boldsymbol{\alpha} - \mathbf{R}_{ij}\mathbf{X}||_2^2, \tag{4.3}$$

$$\hat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmin}} \ ||\mathbf{X}-\mathbf{W}||_2^2 + \lambda||DH\mathbf{X}-\mathbf{Y}||_2^2, \tag{4.4}$$

where $\mathbf{W}$ is the image obtained by averaging the blocks obtained using the sparse coefficients vectors $\hat{\boldsymbol{\alpha}}_{ij}$. Equation (4.3) is a local sparsity based method that divides the whole image into blocks and represents each block sparsely using some trained dictionary. Among other advantages, one major advantage of such a method is the ease to train a small dictionary as compared to one large global dictionary. This is achieved with the help of (4.3) which is equivalent to (4.5). As to the coefficients μ_{ij}, those must be location dependent, so as to comply with a set of constraints of the form $||\boldsymbol{\Psi\alpha} - \mathbf{R}_{ij}\mathbf{X}||_2^2 \leq T$. Solving this using the orthonormal matching pursuit [101] is easy, gathering one atom at a time, and stopping when the error $||\boldsymbol{\Psi\alpha} - \mathbf{R}_{ij}\mathbf{X}||_2^2$ goes below T. This way, the choice of μ_{ij} has been handled implicitly. Equation (4.4) applies a global constraint on the reconstructed image and uses the local patches and the noisy image as input in order to construct the output that complies with local-sparsity and also lies within the proximity of the distorted image which is defined by amount and type of distortion.

$$\hat{\boldsymbol{\alpha}}_{ij} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \, ||\boldsymbol{\alpha}||_0 \text{ subject to } ||\boldsymbol{\Psi\alpha} - \mathbf{R}_{ij}\mathbf{X}||_2^2 \leq T. \tag{4.5}$$

In (4.4), we have assumed that the distortion operator Φ in (4.1) may be represented by the product DH, where H is a blurring filter and D the downsampling operator. Here we have assumed each non-overlapping patch of the images can be represented sparsely in the domain of $\boldsymbol{\Psi}$. Assuming this prior on each patch (4.3) refers to the sparse coding of local image patches with bounded prior, hence building a local model from sparse representations. This enables us to de-noise individual patches by solving (4.3) for each patch. By doing so, we face the problem of blockiness at the patch boundaries when de-noised non-overlapping patches are placed back in the image. To remove these artifacts from the de-noised images overlapping patches are extracted from the noisy image which are combined together with the help of (4.4). The solution of (4.4) demands the proximity between the noisy image, $\mathbf{Y}$, and the output image $\mathbf{X}$, thus enforcing the global reconstruction constraint. The $\mathcal{L}_2$ optimal solution suggests to take the average of the overlapping patches [49], thus eliminating the problem of blockiness in the de-noised image.

As stated earlier, we propose a modified restoration method which incorporates SSIM

into the procedure defined by (4.3) and (4.4). It is defined as follows,

$$\hat{\boldsymbol{\alpha}}_{ij} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\ \mu_{ij}||\boldsymbol{\alpha}||_0 + (1 - S(\boldsymbol{\Psi}\boldsymbol{\alpha}, \mathbf{R}_{ij}\mathbf{X})), \tag{4.6}$$

$$\hat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmax}}\ \ S(\mathbf{W}, \mathbf{X}) + \lambda S(DH\mathbf{X}, \mathbf{Y}), \tag{4.7}$$

where $S(\cdot, \cdot)$ defines the SSIM measure. The expression for SSIM index is

$$S(\mathbf{a}, \mathbf{y}) = \frac{2\mu_{\mathbf{a}}\mu_{\mathbf{y}} + C_1}{\mu_{\mathbf{a}}^2 + \mu_{\mathbf{y}}^2 + C_1} \frac{2\sigma_{\mathbf{a},\mathbf{y}} + C_2}{\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2}, \tag{4.8}$$

with $\mu_{\mathbf{a}}$ and $\mu_{\mathbf{y}}$ the means of $\mathbf{a}$ and $\mathbf{y}$ respectively, $\sigma_{\mathbf{a}}^2$ and $\sigma_{\mathbf{y}}^2$ the sample variances of $\mathbf{a}$ and $\mathbf{y}$ respectively, and $\sigma_{\mathbf{ay}}$ the covariance between $\mathbf{a}$ and $\mathbf{y}$. The constants C_1 and C_2 are stabilizing constants and account for the saturation effect of the HVS.

Equation (4.6) aims to provide the best approximation of a local patch in SSIM-sense with the help of minimum possible number of atoms. The process is performed locally for each block in the image which are then combined together by simple averaging to construct $\mathbf{W}$. Equation (4.7) applies a global constraint and outputs the image that is the best compromise between the noisy image, $\mathbf{Y}$, and $\mathbf{W}$ in SSIM-sense. This step is very vital because it has been observed that the image $\mathbf{W}$ lacks the sharpness in the structures present in the image. Due to masking effect of the HVS, same level of noise does not distort different visual content equally. Therefore, the noisy image is used to borrow the content from its regions which are not convoluted severely by noise. Use of SSIM is very well-suited for such a task, as compared to MSE, because it accounts for the masking effect of HVS and allows us to capture improve structural details with the help of the noisy image. Note the use of $1 - S(\cdot, \cdot)$ in (4.6). This is motivated by the fact that $1 - S(\cdot, \cdot)$ is a squared variance-normalized $\mathcal{L}_2$ distance [10]. Solutions to the optimization problems in (4.6) and (4.7) are given in Sections 4.2.2 and 4.2.3, respectively.

4.2.2 SSIM-optimal Local Model from Sparse Representation

This section discusses the solution to the optimization problem in (4.6). Equation (4.3) can be solved approximately using Orthogonal Matching Pursuit (OMP) [101] by including one

atom at a time and stopping when the error $||\mathbf{\Psi}\boldsymbol{\alpha}_{ij} - \mathbf{R}_{ij}\mathbf{X}||_2^2$ goes below $T_{mse} = (C\sigma)^2$. C is the noise gain and σ is the standard deviation of the noise. We solve the optimization problem in (4.6) based on the same philosophy. We gather one atom at a time and stop when $S(\mathbf{\Psi}\boldsymbol{\alpha}, \mathbf{x}_{ij})$ goes above T_{ssim}, threshold defined in terms of SSIM. In order to obtain T_{ssim}, we need to consider the relationship between MSE and SSIM. For the mean reduced $\mathbf{a}$ and $\mathbf{y}$, the expression of SSIM reduces to the following equation

$$S(\mathbf{a}, \mathbf{y}) = \frac{2\sigma_{\mathbf{a},\mathbf{y}} + C_2}{\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2}, \tag{4.9}$$

Subtracting both sides of (4.9) from 1 yields

$$\begin{aligned} 1 - S(\mathbf{a}, \mathbf{y}) &= 1 - \frac{2\sigma_{\mathbf{a},\mathbf{y}} + C_2}{\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2} \\ &= \frac{\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 - 2\sigma_{\mathbf{a},\mathbf{y}}}{\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2} \\ &= \frac{||\mathbf{a} - \mathbf{y}||_2^2}{\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2}, \end{aligned} \tag{4.10}$$

Equation (4.10) can be re-arranged to arrive at the following result

$$S(\mathbf{a}, \mathbf{y}) = 1 - \frac{||\mathbf{a} - \mathbf{y}||_2^2}{\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2} \tag{4.11}$$

With the help of the equation above, we can calculate the value of T_{ssim} as follows

$$T_{ssim} = 1 - \frac{T_{mse}}{\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2}, \tag{4.12}$$

where C_2 is the constant originally used in SSIM index expression [165] and $\sigma_{\mathbf{a}}^2$ is calculated based on current approximation of the block given by $\mathbf{a} := \mathbf{\Psi}\boldsymbol{\alpha}$.

The main difference between SSIM and MSE is the divisive normalization [10,158]. This normalization is conceptually consistent with the light adaptation (also called luminance masking) and contrast masking effect of HVS. It has been recognized as an efficient perceptually and statistically non-linear image representation model [81,153]. It is shown to be a

useful framework that accounts for the masking effect in human visual system, which refers to the reduction of the visibility of an image component in the presence of large neighboring components [51, 179]. It has also been found to be powerful in modeling the neuronal responses in the visual cortex [59, 135]. Divisive normalization has been successfully applied in image quality assessment [76, 115], image coding [90], video coding [118, 158] and image de-noising [109].

Equation (4.12) suggests that the threshold is chosen adaptively for each patch. The set of coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_k)$ should be calculated such that we get the best approximation $\mathbf{a}$ in terms of SSIM. We search for the stationary points of the partial derivatives of S with respect to $\boldsymbol{\alpha}$. The solution to this problem for orthogonal set of basis is discussed in [10]. Here we aim to solve a more general case of linearly independent atoms. The $\mathcal{L}_2$-based optimal coefficients, $\{c_i\}_{i=1}^k$, can be calculated by solving the following system of equations

$$\sum_{j=1}^{k} c_j \langle \psi_i, \psi_j \rangle = \langle \mathbf{y}, \psi_i \rangle, \qquad 1 \leq i \leq k, \tag{4.13}$$

We denote the inner product of a signal with the constant signal $(1/n, 1/n, \ldots, 1/n)$ of length n by $< \psi >:=< \psi, 1/n >$, where $< \cdot, \cdot >$ represents the inner product.

First, we write the mean, the variance and the covariance of $\mathbf{a}$ in terms of α with n the size of the current block:

$$\mu_{\mathbf{a}} = \langle \sum_{i=1}^{k} \alpha_i \psi_i \rangle = \sum_{i=1}^{k} \alpha_i \langle \psi_i \rangle, \tag{4.14}$$

$$\begin{aligned} (n-1)\sigma_{\mathbf{a}}^2 &= \langle \mathbf{a}, \mathbf{a} \rangle - n \langle \mathbf{a} \rangle^2 \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j \langle \psi_i, \psi_j \rangle - n \mu_{\boldsymbol{a}}^2, \end{aligned} \tag{4.15}$$

$$\begin{aligned} (n-1)\sigma_{\mathbf{ay}} &= \langle \mathbf{a}, \mathbf{y} \rangle - n \langle \mathbf{a} \rangle \langle \mathbf{y} \rangle \\ &= \sum_{i=1}^{k} \alpha_i \langle \mathbf{y}, \psi_i \rangle - n \mu_{\mathbf{a}} \mu_{\mathbf{y}}, \end{aligned} \tag{4.16}$$

where $< \cdot >$ represents the sample mean. The partial derivatives are given as follows

$$\frac{\partial \mu_{\mathbf{a}}}{\partial \alpha_i} = \langle \psi_i \rangle, \tag{4.17}$$

$$(n-1)\frac{\partial \sigma_{\mathbf{a}}^2}{\partial \alpha_i} = 2\sum_{j=1}^{k} \alpha_j \langle \psi_i, \psi_j \rangle - 2n\mu_{\mathbf{a}} \langle \psi_i \rangle, \tag{4.18}$$

$$(n-1)\frac{\partial \sigma_{\mathbf{ay}}}{\partial \alpha_i} = \langle \mathbf{y}, \psi_i \rangle - n\mu_{\mathbf{y}} \langle \psi_i \rangle, \tag{4.19}$$

From logarithmic differentiation of (5.2) combined with (4.17)-(4.19), we have

$$\begin{aligned} \frac{1}{S}\frac{\partial S}{\partial \alpha_i} = & \frac{2\mu_{\mathbf{y}}\langle \psi_i \rangle}{2\mu_{\mathbf{a}}\mu_{\mathbf{y}} + C_1} - \frac{2\mu_{\mathbf{a}}\langle \psi_i \rangle}{\mu_{\mathbf{a}}^2 + \mu_{\mathbf{y}}^2 + C_1} + \\ & \frac{2\left[\langle \mathbf{y}, \psi_i \rangle - n\mu_{\mathbf{y}}\langle \psi_i \rangle\right]}{(n-1)\left[2\sigma_{\mathbf{a},\mathbf{y}} + C_2\right]} - \frac{2\left[\sum_{j=1}^{k} \alpha_j \langle \psi_i, \psi_j \rangle - n\mu_{\mathbf{a}}\langle \psi_i \rangle\right]}{(n-1)\left[\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2\right]} \end{aligned} \tag{4.20}$$

After subtracting the corresponding DC values from all the blocks in the image, we are interested only in the particular case where the atoms are made of oscillatory functions, i.e. when $\langle \psi_i \rangle = 0$ for $1 \leq i \leq k$, thus reducing (4.20) to

$$\frac{1}{S}\frac{\partial S}{\partial \alpha_i} = \frac{2\langle \mathbf{y}, \psi_i \rangle}{(n-1)2\sigma_{\mathbf{a},\mathbf{y}} + C_2} - \frac{2\left(\sum_{j=1}^{k} \alpha_j \langle \psi_i, \psi_j \rangle\right)}{(n-1)\left(\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2\right)}. \tag{4.21}$$

We equate (4.21) to zero in order to find the stationary points. The result is the following linear system of equations

$$\sum_{j=1}^{k} \alpha_j \langle \psi_i, \psi_j \rangle = \beta \langle \mathbf{y}, \psi_i \rangle, \qquad 1 \leq i \leq k, \tag{4.22}$$

where

$$\beta = \frac{\sigma_{\mathbf{a}}^2 + \sigma_{\mathbf{y}}^2 + C_2}{2\sigma_{\mathbf{ay}} + C_2}. \tag{4.23}$$

where β is an unknown constant dependent on the statistics of the unknown image block $\mathbf{a}$. Comparing $\boldsymbol{\alpha}$ with the optimal coefficients in $\mathcal{L}_2$ sense denoted by $\mathbf{c}$ and given by (4.13) results in the following solution:

$$\alpha_i = \beta c_i, \qquad 1 \leq i \leq k, \tag{4.24}$$

which implies that the optimal SSIM-based solution is just a scaling of the optimal $\mathcal{L}_2$-based solution. The last step is to find β. It is important to note that the value of β varies over the image and is therefore content dependent. Also, the scaling factor, β, may lead to selection of a different set of atoms from the dictionary, as compared to $\mathcal{L}_2$ where $\beta = 1$, which are better suited to providing a closer and sparser approximation of the patch in SSIM-sense. After substituting (4.24) in the expression (4.23) for β via (4.14), (4.15) and (4.16) and then isolating for β gives us the following quadratic equation

$$\beta^2(B - A) + \beta C_2 - \sigma_{\mathbf{y}}^2 - C_2 = 0. \tag{4.25}$$

where

$$A = \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{k} c_i c_j \langle \psi_i, \psi_j \rangle, \tag{4.26}$$

$$B = \frac{2}{n-1} \sum_{j=1}^{k} c_j \langle \mathbf{y}, \psi_j \rangle. \tag{4.27}$$

Solving for β and picking a positive value for maximal SSIM gives us

$$\beta = \frac{-C_2 + \sqrt{C_2^2 + 4(B - A)(\sigma_{\mathbf{y}}^2 + C_2)}}{2(B - A)}. \tag{4.28}$$

Now we have all the tools required for an OMP algorithm that perform the sparse coding stage in optimal SSIM sense. The modified OMP pursuit algorithm is explained in Algorithm 2. There are two main differences between the OMP algorithm [101] and the one proposed in this work. First, the stopping criterion is based on SSIM. Unlike MSE, SSIM is adaptive according to the reference image. In particular, if the distortion is consistent with the underlying reference e.g. contract enhancement, the distortion is non-structural and is much less objectional than structural distortions. Defining the stopping

Algorithm 2: SSIM-inspired Orthogonal Matching Pursuit

Initialize: $\mathbf{D} = \{\}$ set of selected atoms, $S_{opt} = 0$, $\mathbf{r} = \mathbf{Y}$

while $S_{opt} < T_{ssim}$

- Add the next best atom in $\mathcal{L}_2$ sense to $\mathbf{D}$
- Find the optimal $\mathcal{L}_2$-based coefficient(s) using (4.13)
- Find the optimal SSIM-based coefficient(s) using (4.24) and (4.28)
- Update the residual $\mathbf{r}$
- Find SSIM-based approximation $\mathbf{a}$
- Calculate $S_{opt} = S(\mathbf{a}, \mathbf{y})$

end

criterion according to SSIM essentially means that we are modifying the set of *accepted* points (image patches) around the noisy image patch which can be represented as the linear combination of dictionary atoms. This way, in the space of image patches, we are omitting image patches in the direction of structural distortion and including the ones which are in the same direction as the original image patch in the set of *acceptable* image patches. Therefore, we can expect to see more structures in the image constructed using sparsity as a prior. Second, we calculate the SSIM-optimal coefficients from the optimal coefficients in $\mathcal{L}_2$-sense using the derivation in Section 4.2.2, which are scalar multiple of the optimal $\mathcal{L}_2$-based coefficients.

4.2.3 SSIM-based Global Reconstruction

The solution to this optimization problem defined in Equation (4.7) is the image that is the best compromise between the distorted image and the one obtained using sparse representation in the maximal SSIM sense. With the assumption of known dictionary, the

only other thing the optimization problem in (4.7) requires is the coefficients $\boldsymbol{\alpha}_{ij}$ which can be obtained by solving optimization problem in (4.6). SSIM is a local quality measure when it is applied using a sliding window, it provides us with a quality map that reflects the variation of local quality over the whole image. The global SSIM is computed by pooling (averaging) the local SSIM map. The global SSIM for an image, $\mathbf{Y}$, with respect to the reference image, $\mathbf{X}$, is given by the following equation

$$S(\mathbf{X}, \mathbf{Y}) = \frac{1}{N_l} \sum_{ij} S(\mathbf{x}_{ij}, \mathbf{y}_{ij}), \tag{4.29}$$

where $\mathbf{x}_{ij} = \mathbf{R}_{ij}\mathbf{X}$ and $\mathbf{y}_{ij} = \mathbf{R}_{ij}\mathbf{Y}$ where $\mathbf{R}_{ij}$ is an $N_w \times N$ matrix that extracts the (ij) block from the image. The expression for local SSIM, $S(\mathbf{x}_{ij}, \mathbf{y}_{ij})$, is given by (5.2). N_l is the total number of local windows and can be calculated as

$$N_l = \frac{1}{N_w} \text{tr} \left(\sum_{ij} \mathbf{R}_{ij}^T \mathbf{R}_{ij} \right). \tag{4.30}$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix.

We use a gradient-descent approach to solve the optimization problem given by (4.7). The update equation is given by

$$\begin{aligned} \hat{\mathbf{X}}_{k+1} &= \hat{\mathbf{X}}_k + \lambda \vec{\nabla}_{\mathbf{Y}} S(\mathbf{X}, \mathbf{Y}) \\ &= \hat{\mathbf{X}}_k + \lambda \frac{1}{N_l} \vec{\nabla}_{\mathbf{Y}} \sum_{ij} S(\mathbf{x}_{ij}, \mathbf{y}_{ij}) \\ &= \hat{\mathbf{X}}_k + \lambda \frac{1}{N_l} \sum_{ij} \mathbf{R}_{ij}^T \vec{\nabla}_{\mathbf{y}} S(\mathbf{x}_{ij}, \mathbf{y}_{ij}) \end{aligned} \tag{4.31}$$

where

$$\vec{\nabla}_{\mathbf{y}} S(\mathbf{x}, \mathbf{y}) = \frac{2}{N_w B_1^2 B_2^2} [A_1 B_1 (B_2 \mathbf{x} - A_2 \mathbf{y} + B_1 B_2 (A_2 - A_1) \mu_{\mathbf{x}} \mathbf{1} + A_1 A_2 (B_1 - B_2) \mu_{\mathbf{y}} \mathbf{1}], \tag{4.32}$$

$$A_1 = 2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + C_1, \qquad A_2 = 2\sigma_{\mathbf{xy}} + C_2,$$
$$B_1 = \mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + C_1, \qquad B_2 = \sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 + C_2,$$

where N_w is the number of pixels in the local image patch, $\mu_{\mathbf{x}}$, $\sigma_{\mathbf{x}}^2$ and $\sigma_{\mathbf{xy}}$ represent the sample mean of $\mathbf{x}$, the sample variance of $\mathbf{x}$, and the sample covariance of $\mathbf{x}$ and $\mathbf{y}$, respectively. Equation (4.31) suggests that averaging of the gradients of local patches is to be calculated in order to obtain the global SSIM gradient, and thus the direction and distance of the kth update in $\hat{\mathbf{X}}$. More details regarding the computation of SSIM gradient can be found in [176]. In our experiment, we found this gradient based approach is well-behaved and it takes only a few iterations for $\hat{X}$ to converge to a stationary point. We initialize $\hat{\mathbf{x}}$ as the best MSE solution. Having the gradient of SSIM we follow an iterative procedure to solve (4.7), assuming the initial value derived from minimal MSE solution.

4.3 Applications

The framework we proposed provides a general approach that can be used for different applications. To show the effectiveness of our method we will provide two applications: image de-noising and super-resolution.

4.3.1 Image De-noising

We use the SSIM-based sparse representations framework developed in Sections 4.2.2 and 4.2.3 to perform the task of image de-noising. The noise-contaminated image is obtained using the following equation

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}, \tag{4.33}$$

where $\mathbf{Y}$ is the observed distorted image, $\mathbf{X}$ is the noise-free image and $\mathbf{N}$ is additive Gaussian noise. Our goal is to remove the noise from distorted image. Here we train a dictionary, $\mathbf{\Psi}$, for which the original image can be represented sparsely in its domain. We use K-SVD method [1] to train the dictionary. In this method the dictionary, which is trained directly over the noisy image and de-noising is done in parallel. For a fixed number of iterations, J, we initialize the dictionary by discrete cosine transform (DCT) dictionary. In each step we update the image and then the dictionary. First, based on the current

dictionary, sparse coding is done for each patch, and then K-SVD is used to update the dictionary (interested reader can refer to [1] for details of dictionary updating). Finally, after doing this procedure J times we execute a global construction stage, following the gradient descend procedure. The proposed image de-noising algorithm is summarized in Algorithm 2.

Algorithm 3: SSIM-inspired image de-noising

1. *Initialize:* $\mathbf{X} = \mathbf{Y}$, $\mathbf{\Psi}$ = overcomplete DCT dictionary
2. *Repeat J times*
   - *Sparse coding stage:* use SSIM-optimal OMP to compute the representation vectors α_{ij} for each patch
   - *Dictionary update stage:* Use K-SVD [1] to calculate the updated dictionary and coefficients. Calculate SSIM-optimal coefficients using (4.24) and (4.28)
3. *Global Reconstruction:* Use gradient descent algorithm to optimize (4.7), where the SSIM gradient is given by (4.32).

The proposed image de-noising scheme is tested on various images with different amount of noise. In all the experiments, the dictionary used was of size 64×256, designed to handle patches of 8×8 pixels. The value of noise gain, C , is selected to be 1.15 and $\lambda = 30/\sigma$ [49]. Table 4.1 shows the results for images *Barbara, Lena, Peppers, House*. We compare the proposed method mainly with the K-SVD method [49] as the implementation is inspired by the K-SVD approach. The proposed scheme is a proof of concept for future SSIM-Inspired image processing algorithms as we expect similar gains on top of other state-of-the-art image de-noising algorithms. It can be observed that the proposed de-noising method achieves better performance in terms of SSIM which is expected to imply better perceptual quality of the de-noised image. Figure 4.1 and 4.2 shows that the de-noised images using K-SVD [49] and the proposed methods along with corresponding SSIM maps. It can be observed that SSIM based method outperforms specially in the texture region which confirms that the proposed de-noising scheme preserves the structures better

and therefore has better perceptual image quality.

Table 4.1: SSIM and PSNR comparisons of image de-noising results

Image	Barbara				Lena				Peppers				House			
Noise std	20	25	50	100	20	25	50	100	20	25	50	100	20	25	50	100
	PSNR comparison (in dB)															
Noisy	22.11	20.17	14.15	8.13	22.11	20.17	14.15	8.13	22.11	20.17	14.15	8.13	22.11	20.17	14.15	8.13
K-SVD	30.85	29.55	25.44	21.65	32.38	31.32	27.79	24.46	30.80	29.72	26.10	21.84	33.16	32.12	28.08	23.54
Proposed	30.88	29.53	25.50	21.74	32.26	31.28	27.80	24.53	30.84	29.84	26.25	21.98	33.04	32.09	28.13	23.59
	SSIM comparison															
Noisy	0.593	0.503	0.241	0.084	0.531	0.443	0.204	0.074	0.529	0.442	0.212	0.076	0.452	0.368	0.166	0.057
K-SVD	0.894	0.859	0.708	0.519	0.903	0.877	0.733	0.550	0.905	0.883	0.782	0.601	0.909	0.890	0.779	0.549
Proposed	0.906	0.875	0.733	0.526	0.913	0.888	0.754	0.573	0.913	0.894	0.797	0.627	0.915	0.901	0.795	0.574

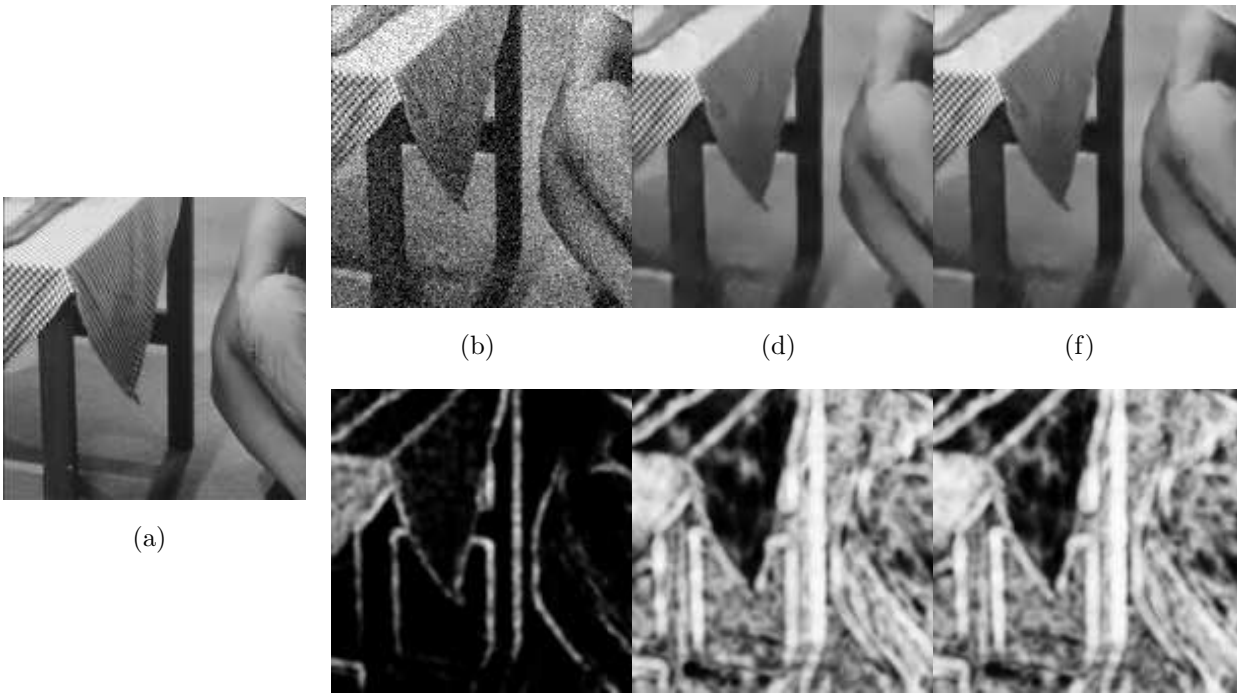
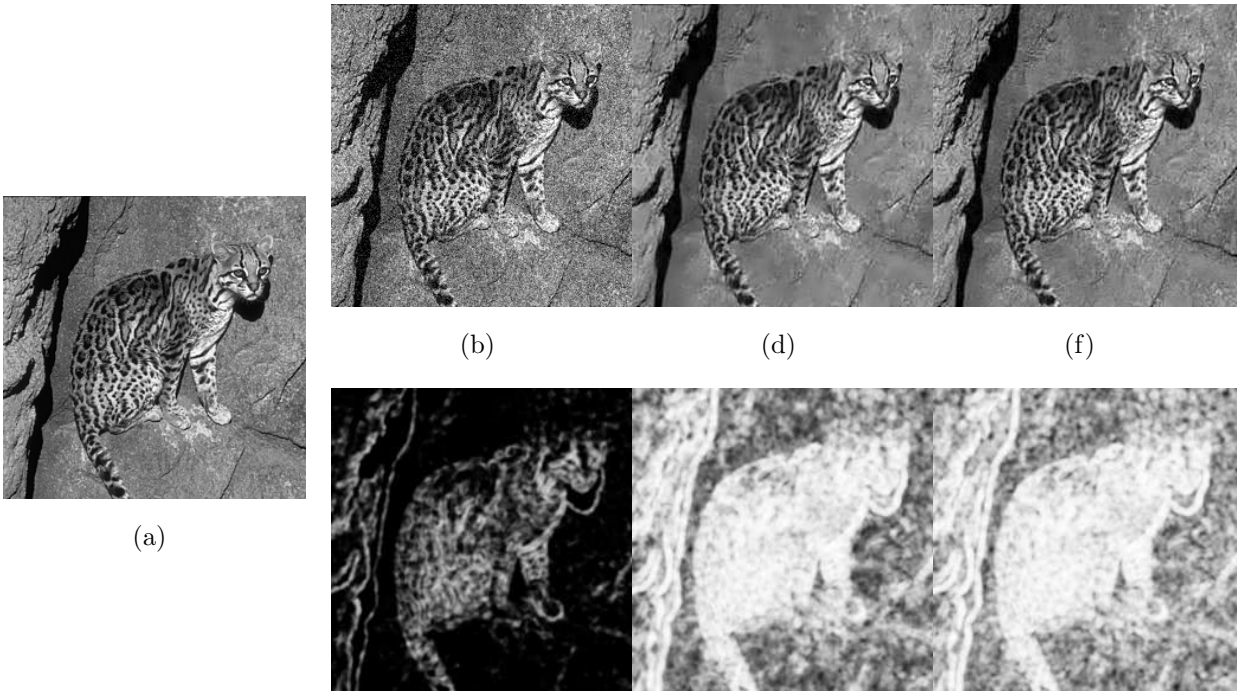


Figure 4.1: Visual comparison of de-noising results. (a) Original image. (b) Noisy image. (c) SSIM-map of noisy image. (d) KSVD-MSE image. (e) SSIM-map of KSVD-MSE image. (f) KSVD-SSIM image. (g) SSIM map of KSVD-SSIM image.



(a) (b) (d) (f)

(c) (e) (g)

Figure 4.2: Visual comparison of de-noising results. (a) Original image. (b) Noisy image. (c) SSIM-map of noisy image. (d) KSVD-MSE image. (e) SSIM-map of KSVD-MSE image. (f) KSVD-SSIM image. (g) SSIM map of KSVD-SSIM image.

4.3.2 Image Super-resolution

In this section we demonstrate the performance of the SSIM-based sparse representations when used for image super-resolution. In this problem, a low resolution image, $\mathbf{Y}$, is given and a high resolution version of the image, $\mathbf{X}$, is required as output. We assume that the low resolution image is produced from high resolution image based on the following equation:

$$\mathbf{Y} = DH\mathbf{X}, \tag{4.34}$$

where H represents a blurring matrix, and D is a downsampling matrix. We use local sparsity model as prior to regularize this problem that has infinite many solutions which satisfy (4.34). Our approach is motivated by recent results in sparse signal representation, which suggests that the linear relationships among high-resolution signals can be accurately recovered from their low-dimensional projections. Here, we work with two coupled dictionaries, $\mathbf{\Psi}_h$ for high-resolution patches, and $\mathbf{\Psi}_l$ for low-resolution ones. The sparse representation of a low-resolution patch in terms of $\mathbf{\Psi}_l$ will be directly used to recover the corresponding high resolution patch from $\mathbf{\Psi}_h$ [199]. Given these two dictionaries, each corresponding patch of low resolution image, $\mathbf{y}$, and high resolution image, $\mathbf{x}$, can be represented sparsely with the same coefficient vector, α in Algorithm 2:

$$y = \mathbf{\Psi}_l \alpha \tag{4.35}$$

$$x = \mathbf{\Psi}_h \alpha \tag{4.36}$$

The patch from each location of the low-resolution image, that needs to be scaled up, is extracted and sparsely coded with the help of SSIM-optimal Algorithm 2. Once the sparse coefficients, α, are obtained, high resolution patches, $\mathbf{y}$, are computed using (4.36) which are finally merged by averaging in the overlap area to create the resulting image. The proposed image super-resolution algorithm is summarized in Algorithm 3:

The proposed image super resolution scheme is tested on various images. To be consistent with [199] patches of 5×5 pixels were used on the low resolution image. Each patch is converted to a vector of length 25. The dictionaries are trained using K-SVD [49] with the sizes of 25×1024 and 100×1024 for the low and the high resolution dictionaries, respectively. 66 natural images are used for dictionary training, which are also used in [191] for

Algorithm 4: SSIM-inspired image super resolution

1. *Dictionary Training Phase:* trained high and low resolution dictionaries $\mathbf{\Psi}_l$, $\mathbf{\Psi}_h$, [199]
2. *Reconstruction Phase*
   - *Sparse coding stage:* use SSIM-optimal OMP to compute the representation vectors α_{ij} for all the patches of low resolution image
   - *High resolution patches reconstruction:* Reconstruct high resolution patches by $\mathbf{\Psi}_h \alpha_{ij}$
3. *Global Reconstruction:* merge high-resolution patches by averaging over the overlapped region to create the high resolution image.

similar purpose. To remove artifacts on the patch edges we set overlap of one pixel during patch extraction from the image. Fixed number of atoms (3) has been used by [199] in the sparse coding stage. However SSIM-OMP determines the number of atoms adaptively from patch to patch based on its importance considering SSIM measure. In order to calculate the threshold, T_{ssim}, defined in (4.12), T_{mse} is calculated using MSE based sparse coding stage in [199]. After calculating sparse representation for all the low resolution patches, we use them to reconstruct the patches and then the difference with the original patch is calculated. We set T_{mse} to the average of these differences. The performance comparison with state-of-the-art image super-resolution methods is given in Table 4.2.

We can observe that the proposed algorithm outperforms the other methods consistently in terms of SSIM evaluations. It is also interesting to observe PSNR improvements in some cases, though PSNR is not the optimization goal of the proposed approach. The improvements are not always consistent (for example, PSNR drops in some cases in Table 1, while SSIM always improves). There are complicated reasons behind these results. It needs to be aware that the so-called "MSE-optimal" algorithms include many suboptimal and heuristic steps and thus have potential to be improved even in the MSE sense. Our methods are different from the "MSE-optimal" methods in multiple stages. Although the differences are made to improve SSIM, they may have positive impact on improving MSE

Table 4.2: SSIM and PSNR comparisons of image super resolution results

Image	Barbara	Lena	Baboon	House	Raccoon	Zebra	Parthenon	Desk	Aeroplane	Man	Moon	Bridge
	PSNR comparison (in dB)											
Yang et al. [191]	30.3	33.4	25.3	34.1	34.0	24.6	28.4	31.9	34.2	33.2	32.2	28.0
Wang et al. [161]	31.6	34.2	25.6	35.4	36.8	25.3	29.0	33.6	36.3	34.5	33.6	28.1
Dong et al. [43]	31.1	34.0	25.5	33.6	35.6	25.5	28.9	33.1	36.2	34.3	33.1	27.8
Zeyde et al. [199]	31.3	33.8	25.5	35.4	36.5	25.0	28.8	33.8	36.1	34.4	33.3	28.5
Proposed	31.4	33.9	25.6	35.5	37.0	25.1	28.9	33.9	36.4	34.6	33.4	28.6
	SSIM comparison											
Yang et al. [191]	0.843	0.888	0.680	0.876	0.880	0.760	0.773	0.871	0.829	0.857	0.746	0.754
Wang et al. [161]	0.871	0.911	0.722	0.901	0.938	0.791	0.819	0.918	0.862	0.894	0.805	0.790
Dong et al. [43]	0.821	0.897	0.688	0.893	0.887	0.802	0.792	0.911	0.859	0.893	0.806	0.793
Zeyde et al. [199]	0.874	0.909	0.710	0.904	0.934	0.789	0.811	0.918	0.860	0.896	0.803	0.783
Proposed	0.877	0.912	0.720	0.906	0.942	0.794	0.815	0.922	0.862	0.900	0.808	0.792

as well. For example, when using the learned dictionary to reconstruct an image patch, if SSIM is used to replace MSE in selecting the atoms in the dictionary, then essentially the set of accepted atoms in the dictionary have been changed. In particular, since SSIM is variance normalized, the set of acceptable reconstructed patches near the noisy patch may be structurally similar but are significantly different in variance. This may lead to different selections of the atoms in the dictionary, which when appropriately scaled to approximate the noisy patch, may result in better reconstruction result. Although the visual and SSIM improvements are only moderate, these are promising results as an initial attempt of incorporating a perceptually more meaningful measure into the optimization problem of KSVD-based super-resolution method.

Figures 4.3 and 4.4 compare the reconstructed images obtained using [192] and the proposed methods for the Raccoon and the Girl images, respectively. It can be seen that the proposed scheme preserves many local structures better and therefore has better perceptual image quality. The visual quality improvement is also reflected in the corresponding SSIM maps, which provide useful guidance on how local image quality is improved over space. It can be observed from the SSIM maps that the areas which are relatively more structured benefit more from the proposed algorithm as the quality measure used is better at calculating the similarity of structures as compared to MSE.

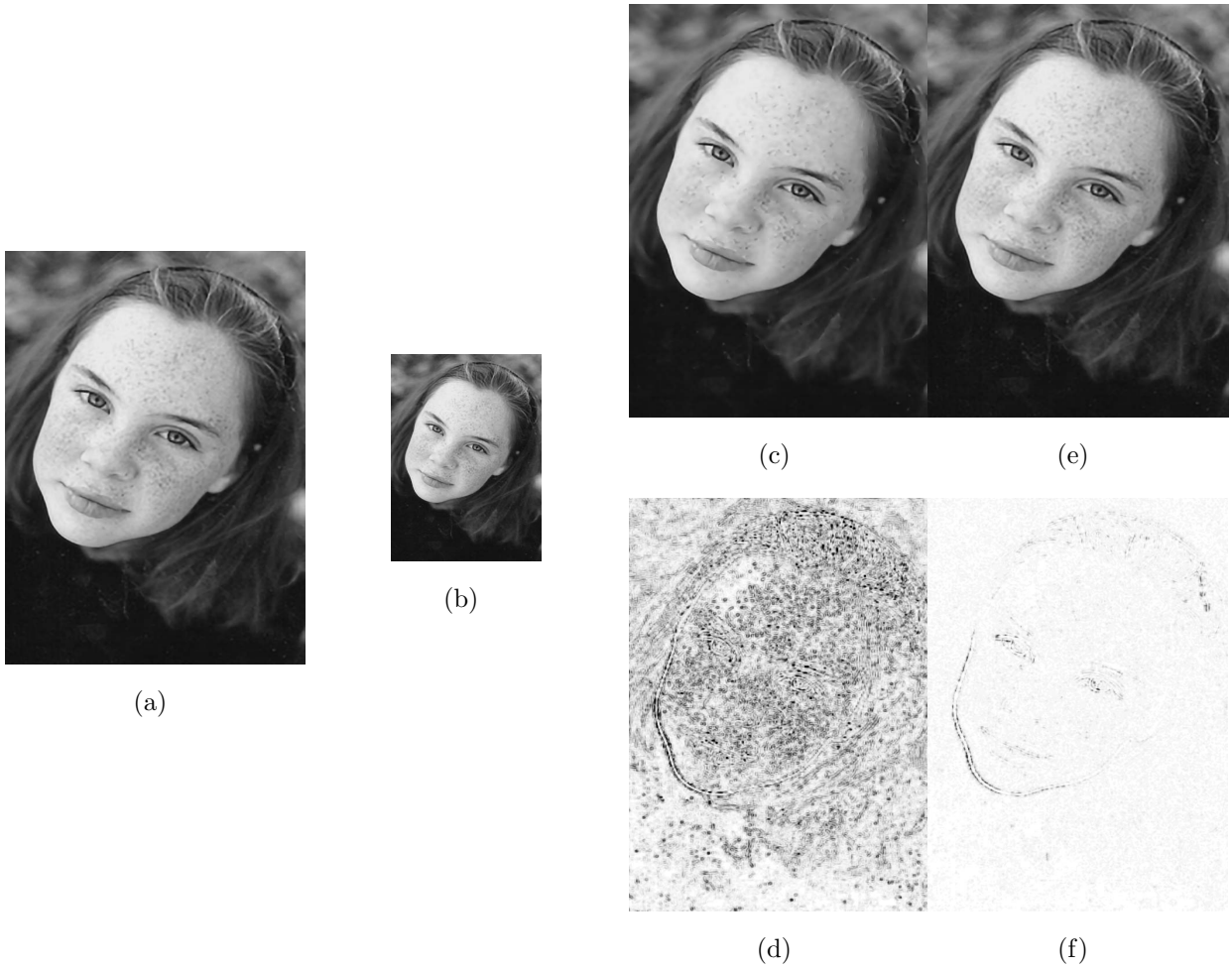
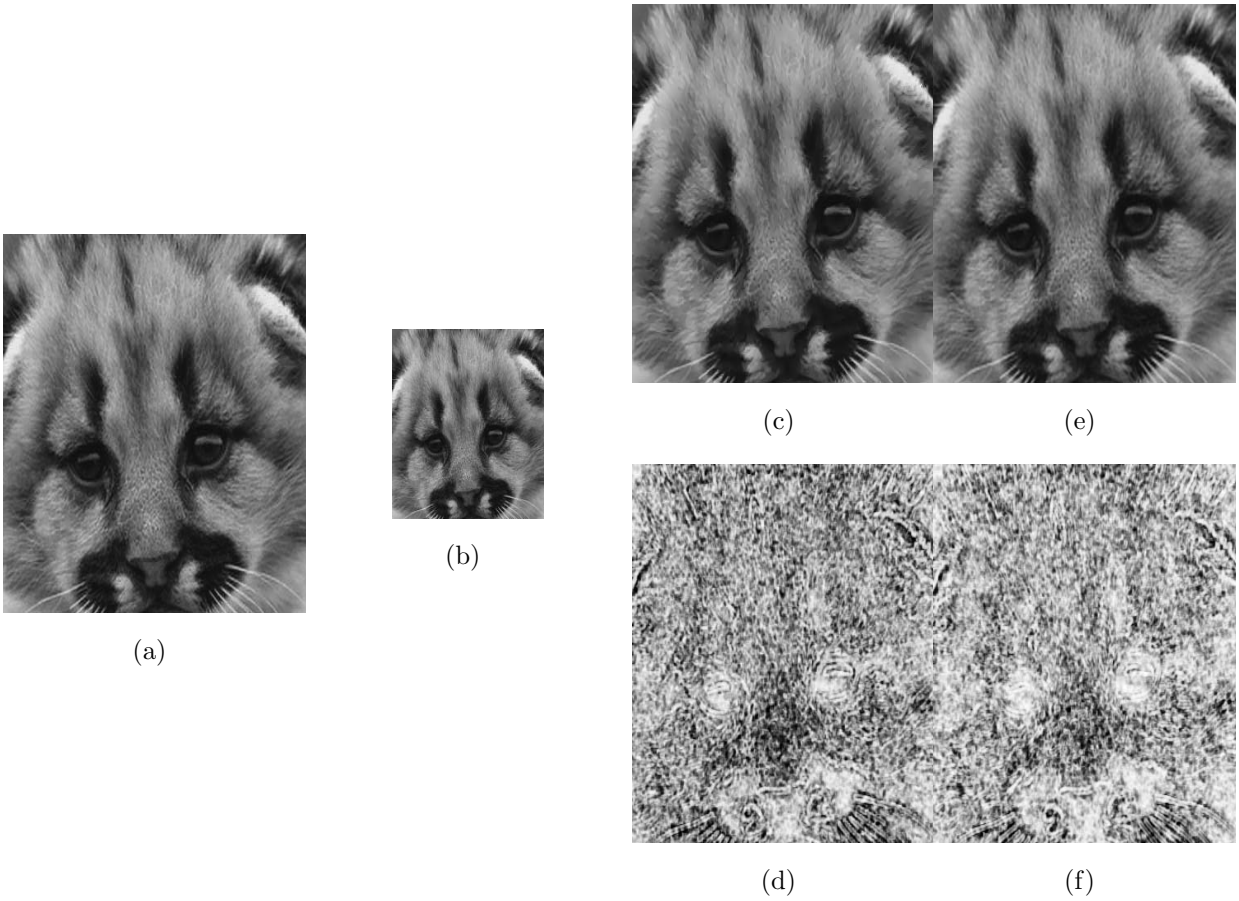


Figure 4.3: Visual comparison of super resolution results. (a) Original image. (b) Low-resolution image. (c) Output of Yang's method. (d) SSIM map of Yang's method. (e) Proposed method. (f) SSIM map of propose method.



(a) (b) (c) (e) (d) (f)

Figure 4.4: Visual comparison of super resolution results. (a) Original image. (b) Low-resolution image. (c) Output of Yang's method. (d) SSIM map of Yang's method. (e) Proposed method. (f) SSIM map of propose method.

Chapter 5

SSIM-based Non-local Means Image De-noising

In this chapter, we propose an SSIM-based NLM method for image de-noising. The key contribution of our approach is to replace the role of MSE with SSIM in measuring patch similarities and in calculating weights. We propose a robust method to estimate SSIM in the presence of noise and adjust the mean and contrast of image patches before using them for weighted averaging. Our simulation results demonstrate the promises of the proposed approach and also indicate the potentials of replacing the ubiquitous PSNR/MSE with SSIM as the optimization criterion in image processing applications.

5.1 Introduction

Recently there has been a great deal of attention paid to the problem of image de-noising, which is not only a practically useful application, but also an ideal test bed for image representation, modeling and estimation theories. One of the most successful image de-noising algorithms is the non-local means (NLM) method [13], which has achieved state-of-the-art performance. NLM de-noising is a nonlocal filtering (or weighted averaging)

technique where the weights are decided based on similarity between the current image patch being de-noised and the other patches in the image within a neighborhood. Since MSE is employed for calculating the weights, the resulting de-noised image might not have the best perceptual quality. This motivates us to replace the role of MSE with SSIM in the framework. There are two issues that need to be resolved before effective SSIM-based approach can be developed. First, we would need to reliably estimate the SSIM value between two original image patches in the presence of noise. Directly using the SSIM value between two noisy patches to define the weight would not lead to good results. This is because SSIM attempts to match the structures of two patches, but when the signal-to-noise ratio is low, the noise submerges the actual structure of the image, and thus SSIM evaluation would favor those patches with the noise pattern best matched. Second, once weights are calculated based on SSIM, it is important to adjust the contrast and mean values of the patches before weighted averaging. This is because SSIM may pick those patches that are structurally similar but with different contrast and mean values, and thus direct averaging these patches (that have different contrast and mean variations) would provoke further undesired distortions. These issues are tackled with the help of proposed two stage de-noising algorithm on similar lines to BM3D [38], a state-of-the-art de-noising method which also uses two stages to perform image de-noising.

5.2 Problem Formulation

NLM algorithm [14] replaces the intensity of each pixel in the noisy image by a weighted average of all the pixel intensities in the image. More generally, the nonlocal filter (NLF) in the continuous space can be represented as follows [78]

$$P_{NLF}(f(x,y)) = \frac{\int_\Omega w(x,y;x',y')f(x',y')dx'dy'}{\int_\Omega w(x,y;x',y')dx'dy'}\,, \tag{5.1}$$

where $w(x,y;x'y')$ is the weighting function related to the similarity between two patches at (x,y) and (x',y'). The weight in NLM de-noising is specified by borrowing ideas from the work of nonparametric sampling-based texture synthesis [48] and is calculated based on $\mathcal{L}_2$ distance between two patches at (x,y) and (x',y').

Table 5.1: Comparisons of NLM de-noising using $\mathcal{L}_2$ and SSIM of original image patches for weight calculation

Test image	Barbara			
Noise std (σ)	15	25	30	50
	PSNR comparison (in dB)			
Noisy image	24.61	20.29	18.81	14.74
$\mathcal{L}_2^*$-NLM	31.42	27.33	25.97	22.49
SSIM*-NLM	32.21	29.59	28.65	25.61
	SSIM comparison			
Noisy image	0.729	0.543	0.474	0.289
$\mathcal{L}_2^*$-NLM	0.925	0.818	0.759	0.543
SSIM*-NLM	0.947	0.902	0.879	0.779

To better reflect the perceptual similarity between two patches and also to give favor to the patches that are structurally similar, we opt to replace the role of $\mathcal{L}_2$ by SSIM in computing the weight function. Let X_1 and X_2 be two image patches extracted from the original noise-free image. The SSIM index between them is defined as

$$S(X_1, X_2) = \frac{(2\mu_{X_1}\mu_{X_2} + C_1)(2\sigma_{X_1X_2} + C_2)}{(\mu_{X_1}^2 + \mu_{X_2}^2 + C_1)(\sigma_{X_1}^2 + \sigma_{X_2}^2 + C_2)}, \tag{5.2}$$

where μ_X, σ_X, and $\sigma_{X_1X_2}$ are the mean, standard deviation, and cross correlation between the two patches, respectively, and C_1 and C_2 are positive stabilizing constants.

To understand the impact of replacing $\mathcal{L}_2$ with SSIM, we carried out an empirical study where all weights were calculated using patches extracted from the original image but computed using $\mathcal{L}_2$ and SSIM, respectively. With these weights, the NLM de-noising results of "Babara" image at different noise levels are shown in Table 5.1, where we observe large gains in both PSNR and SSIM values of the de-noised image when SSIM is employed for weight computation.

The above empirical study, though very instructive, does not provide a working de-noising algorithm, because the original image patches are not accessible. Therefore, the critical problem here is how to predict the SSIM value between X_1 and X_2 from their noisy

observations.

5.3 Proposed Scheme

Let Y_1 and Y_2 be two observed noisy patches that are created from two clean original patches X_1 and X_2 by

$$Y_1 = X_1 + N_1\,, \tag{5.3}$$

$$Y_2 = X_2 + N_2\,, \tag{5.4}$$

where N_1 and N_2 are the corresponding *i.i.d* Gaussian noise patches with standard deviation σ_n. The purpose here is to estimate $S(X_1, X_2)$ using Y_1, Y_2. A simple approximation would be

$$S(X_1, X_2) \approx \frac{(2\mu_{Y_1}\mu_{Y_2} + C_1)(2\sigma_{Y_1Y_2} + C_2)}{(\mu_{Y_1}^2 + \mu_{Y_2}^2 + C_1)(\sigma_{Y_1}^2 + \sigma_{Y_2}^2 - 2\sigma_n^2 + C_2)} \tag{5.5}$$

Here we have made use of the assumptions that the noise N_1 and N_2 are zero-mean, the signal X_1 and X_2 are uncorrelated with noise, and the noise N_1 and N_2 added at different locations are also uncorrelated. Our studies suggest that the approximation in Eq. ((5.5)) does not achieve desired accuracy in estimating $S(X_1, X_2)$ because the assumptions does not hold for small patches. Also, when the variance of noise is significant as compared to that of the image patch, SSIM is in favor of similar noise patterns rather than image structures.

To overcome the problem above, we propose a two-stage method. In the first stage, we compute a local estimate of the noise using the method proposed in [14]. As mentioned in [13], NLM de-noising is based on the "method noise" and the residual image obtained after subtracting the de-noised image from the noise-free image looks like random noise and does not contain structures similar to those contained in the original image. We believe that the noise estimated by NLM de-noising can be used to provide a better estimate of $S(X_1, X_2)$ because more accurate information about the noise pattern at the local patch is available. Suppose the estimated noise is given by $\hat{N}_1$ and $\hat{N}_2$, respectively. It enables us to estimate X_1 and X_2 by

$$\hat{X}_1 = Y_1 - \hat{N}_1\,, \tag{5.6}$$

$$\hat{X}_2 = Y_2 - \hat{N}_2 \,. \tag{5.7}$$

We can then use $\hat{X}_1$ and $\hat{X}_2$ in the second step to estimate $S(X_1, X_2)$ and define our SSIM-based weight as

$$w_{\text{SSIM}} = S(\hat{X}_1, \hat{X}_2) \,. \tag{5.8}$$

Before computing the weighted averaging for each patch, we perform further adjustment on the mean and contrast of each patch Y by

$$Y' = \frac{\sigma_{\hat{X}_c} + c}{\sigma_Y + c} (Y - \mu_Y) + \mu_{\hat{X}_c} \tag{5.9}$$

where μ_Y, σ_Y and $\mu_{\hat{X}_c}$, $\sigma_{\hat{X}_c}$ are the mean and contrast values of the current patch and the patch to be de-noised (estimated using Eq. (5.6)), respectively and c is the stabilizing constant. This adjustment is motivated by the ideas behind SSIM, which separates the measurement of mean, contrast and structure. Indeed, SSIM-based weight calculation may help collect those image patches that are structurally similar to the patch being de-noised but with different contrast and mean values. To avoid creating bias in mean or contrast, it is useful to normalize the patch first, such that only the structural part of the patch contributes to the de-noising task.

Finally, we create our final de-noised patch at location i by

$$\tilde{X}(i) = \frac{\sum_{j \in \mathcal{N}_i} w_{\text{SSIM}}(i,j) Y'(j)}{\sum_{j \in \mathcal{N}_i} w_{\text{SSIM}}(i,j)} \,, \tag{5.10}$$

where $\mathcal{N}_i$ denotes the union of the neighbors around i and $w_{\text{SSIM}}(i,j)$ is the SSIM weight computed between the patches located at i and j.

5.4 Simulation Results

We test image de-noising algorithms on various images with noise standard deviation σ ranging from 15 to 50. The $\mathcal{L}_2$ and SSIM based NLM methods are denoted as $\mathcal{L}_2$-NLM [13] and SSIM-NLM, respectively. All $\mathcal{L}_2$-NLM results are obtained using the code provided by Buades *et. al.* at [12]. The search ranges for both algorithms are fixed at 7×7 in

Table 5.2: SSIM and PSNR comparisons of image de-noising results

Test image	Barbara				Lena				Boat			
Noise std (σ)	15	25	30	50	15	25	30	50	15	25	30	50
	PSNR comparison (in dB)											
Noisy image	24.61	20.29	18.81	14.74	24.62	20.27	18.78	14.71	24.65	20.27	18.76	14.61
$\mathcal{L}_2$-NLM	31.44	28.69	27.55	23.85	32.71	29.94	28.57	25.52	30.87	28.09	27.05	23.91
SSIM-NLM	32.10	29.28	28.21	24.85	33.11	30.52	29.42	25.92	31.26	28.54	27.58	24.39
	SSIM comparison											
Noisy image	0.729	0.543	0.474	0.289	0.489	0.402	0.338	0.192	0.676	0.475	0.406	0.239
$\mathcal{L}_2$-NLM	0.934	0.871	0.832	0.651	0.869	0.832	0.781	0.619	0.889	0.796	0.748	0.565
SSIM-NLM	0.944	0.889	0.858	0.721	0.893	0.858	0.818	0.645	0.900	0.815	0.779	0.599

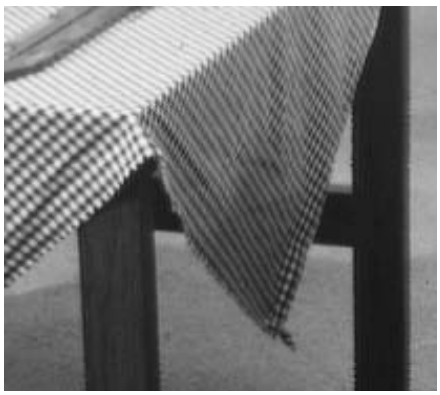
order to limit the complexity of the algorithm. The added computational complexity of SSIM-NLM over $\mathcal{L}_2$-NLM mostly lies in estimating the SSIM values between patches. In our experiment, we found it negligible compared with the overall computational cost of the NLM algorithm.

Table 5.2 shows the results for images "Barbara", "Lena" and "Boat". It can be observed that the proposed SSIM-NLM method achieves better performance than $\mathcal{L}_2$-NLM in terms of not only SSIM, but also PSNR. This may be due to SSIM's capability of collecting those image patches that have similar structure but with different mean and/or contrast. We also observe in our experiment that the performance gap between the two methods increases further when the search range is increased.

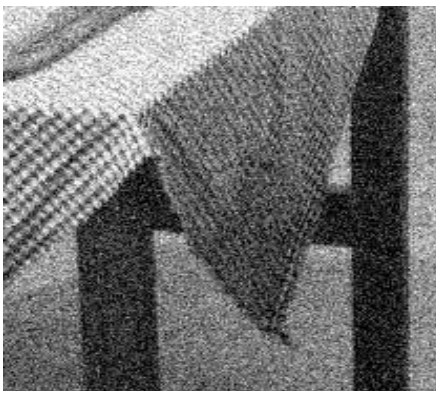
Comparison of the de-noising results of "Babara" image in Table 5.2 with those in Table 5.1 is very interesting. It can be observed that when similarity values are calculated by using the original noise-free image, SSIM-NLM performs significantly better than $\mathcal{L}_2$-NLM in terms of both SSIM and PSNR. Another observation is that the de-noising performance of $\mathcal{L}_2$-NLM degrades when the original image is used to compute the weights. This is likely because of the weight mapping function and thresholds used in the implementation in [12, 14]. When the original image is used, many more patches with lower $\mathcal{L}_2$ distances also make significant impact on de-noising. This often results in blur of the de-noised image. By contrast, the SSIM-NLM method does not suffer from such a problem, implying that SSIM is probably a better measure to select similar patches.

To provide visual comparisons of the de-noising algorithms, Figures 5.1 and 5.2 shows two image areas cropped from the "Babara" image de-noised by $\mathcal{L}_2$-NLM [13] and SSIM-

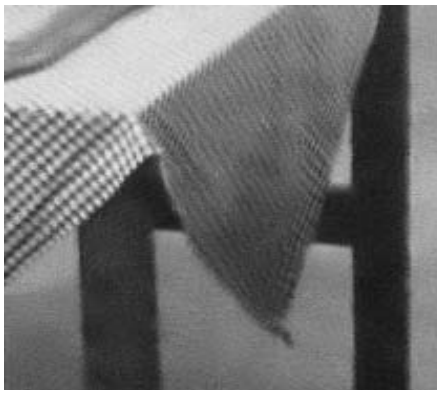
NLM, respectively. It can be seen that the proposed SSIM-NLM scheme preserves many local structures better and therefore has better perceptual image quality. The visual quality improvement is also reflected in the corresponding SSIM maps, which provide useful guidance on how local image quality is improved over space. It can be observed from the SSIM maps that the areas which are relatively more structured benefit more from the proposed de-noising algorithm as the quality measure used is better at calculating the similarity of structures as compared to MSE.



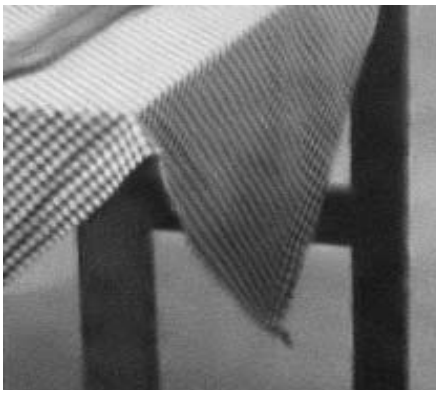
(a) Original image



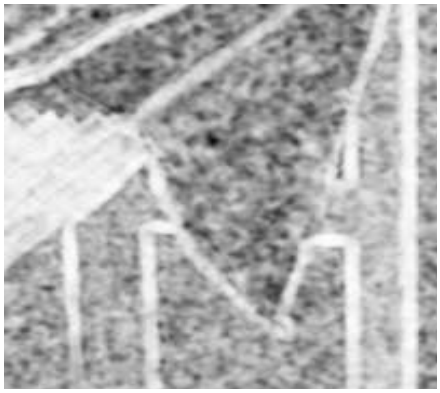
(b) Noisy image ($\sigma = 30$)



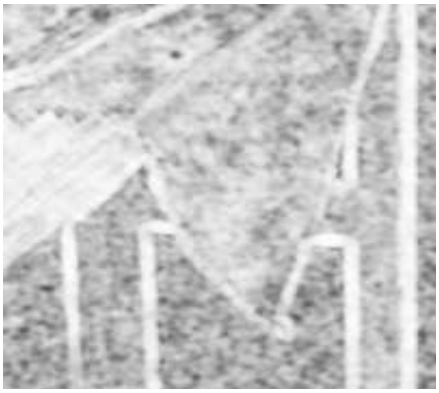
(c) $\mathcal{L}_2$-NLM de-noised



(d) SSIM-NLM de-noised

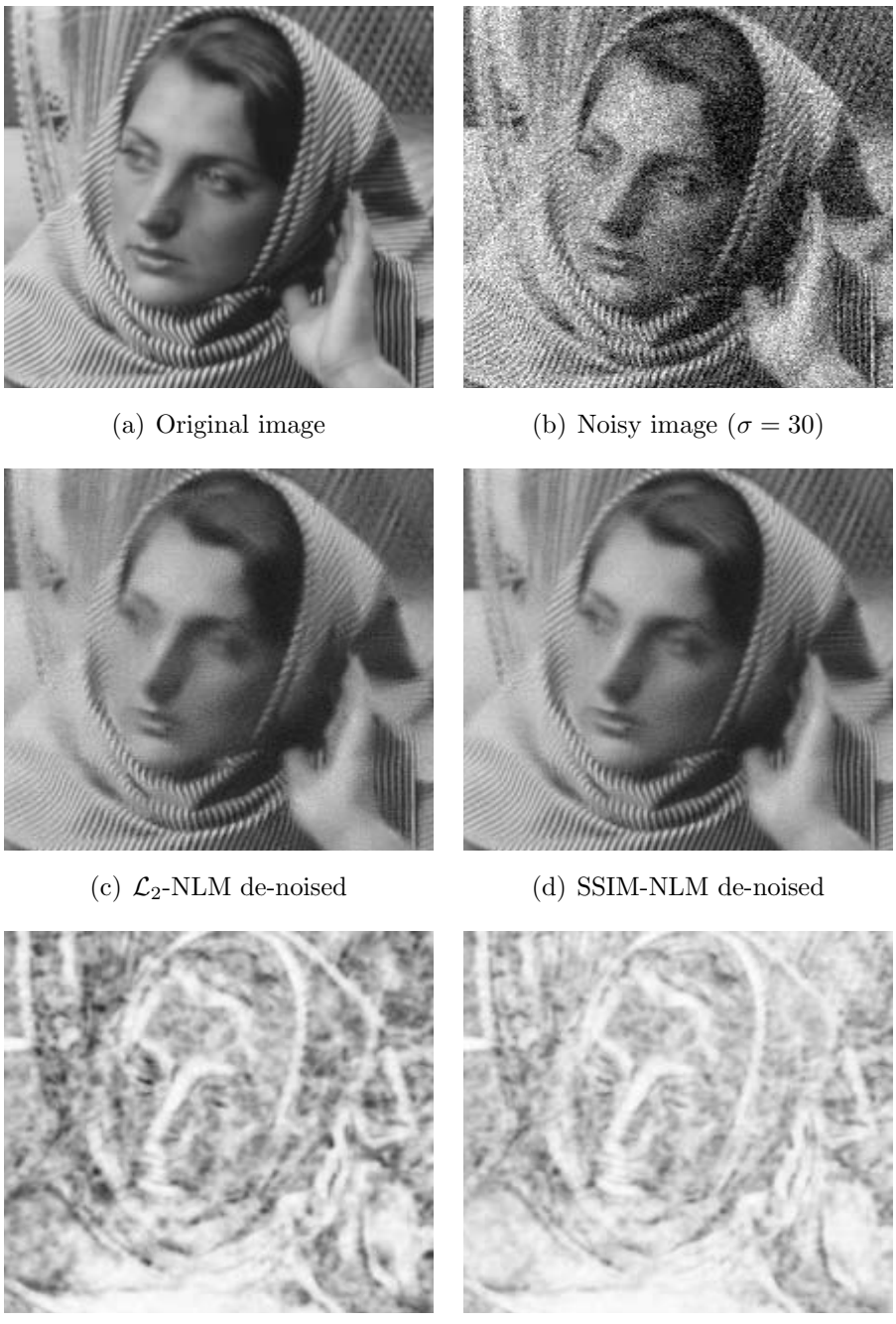


(e) SSIM map of (c)



(f) SSIM map of (d)

Figure 5.1: Visual and SSIM quality map comparisons of de-noising results. Brighter indicates better SSIM value.



(a) Original image

(b) Noisy image ($\sigma = 30$)

(c) $\mathcal{L}_2$-NLM de-noised

(d) SSIM-NLM de-noised

(e) SSIM map of (c)

(f) SSIM map of (d)

Figure 5.2: Visual and SSIM quality map comparisons of de-noising results. Brighter indicates better SSIM value.

Chapter 6

Rate-SSIM Optimization for Video Coding

This chapter presents a Rate Distortion Optimization (RDO) scheme based on the SSIM index[1], which was found to be a better indicator of perceived visual quality than mean-squared-error, but has not been fully exploited in the context of image and video coding. This is achieved by adaptive selection of Lagrange multiplier at the frame level and its further adjustment on macroblock (MB) level which is discussed in detail. At the end, the experimental results are presented and evaluated which show that the proposed scheme can achieve significantly better rate-SSIM performance and provide better visual quality than conventional RDO coding schemes.

6.1 Introduction

In this work, we focus on solving (2.15), where SSIM is used to define the measure of perceived distortion and λ is adapted at both frame and MB levels by taking the properties

[1]proposed in collaboration with S. Wang, a visiting Ph.D. student from Peking University, Beijing.

of the input sequences (statistical properties of residuals, structural information, motion information, etc.) into consideration.

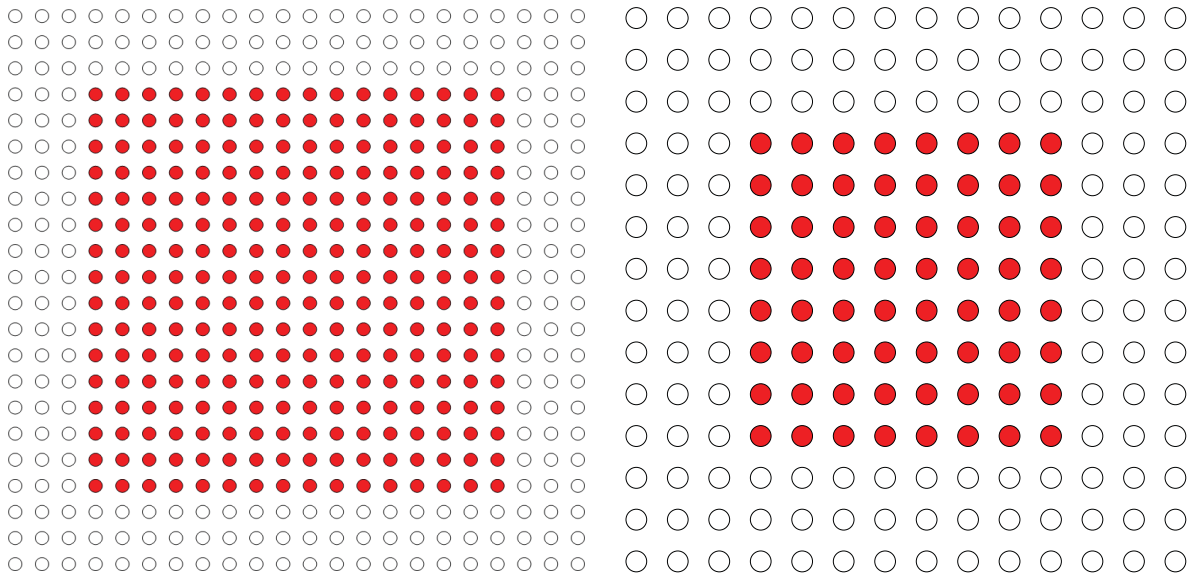
In order to achieve optimal RD performance, it is very important to carefully choose λ and the best coding mode. In the current video coding standards such as H.264/AVC, the coding modes can vary in the mode sets {Intra16x16, Intra8x8, Intra4x4, Inter16x16, Inter16x8, Inter8x16, Inter8x8, Inter8x4, Inter4x8, Inter4x4, SKIP, DIRECT} [57]. During the mode decision process, all available candidate modes are evaluated by the RD cost expression given in (2.15), and the one with the minimum RD cost is selected as the best mode. To achieve a good balance between R and D, in the H.264/AVC coding environment, the Lagrange multiplier is suggested to be [182]

$$\lambda = 0.85 \cdot 2^{\frac{Q_{H.264}-12}{3}}, \tag{6.1}$$

where $Q_{H.264}$ is the quantization parameter (QP). This suggestion was proposed based on empirical results and typical rate-distortion models [140, 181]. It also suggests that λ is a function of QP only and therefore is independent of the frame properties, which simplifies the problem but may not result in optimal λ as some MBs could be more important compared to the others [167]. This motivated us to adapt λ according to the video sequences at both frame and MB levels.

Here, we use SSIM as the distortion measure and propose an adaptive RDO scheme for mode selection. The three main contributions of our work are as follows. First, We employ SSIM as the distortion measure in the proposed mode selection scheme, where both the current MB to be coded and neighboring pixels are taken into account to fully exploit the properties of SSIM. Second, At the frame level, we present an adaptive Lagrange multiplier selection scheme based on a novel statistical reduced-reference SSIM model and a source-side information combined rate model. Third, At the MB level, we present a Lagrange multiplier adjustment scheme, where the scale factor for each MB is determined by an information theoretical approach based on the motion information content and perceptual uncertainty of visual speed perception.

6.2 SSIM Based Rate Distortion Optimization



(a) (b)

Figure 6.1: Illustration of using surrounding pixels to calculate the SSIM index. Solid pixels: To be encoded. Hollow pixels: Surrounding pixels from the input frame. (a) Y Component. (b) Cb, Cr Components.

Analogous to (2.15), the SSIM motivated RDO problem can be defined as

$$\min\{J\} \quad \text{where} \quad J = (1 - SSIM) + \lambda \cdot R. \tag{6.2}$$

In the conventional mode selection process, the final coding mode is determined by the number of entropy coding bits and the distortion of the residuals, while the properties of the reference image are ignored. Unlike MSE, the SSIM index is totally adaptive according to the reference signal [165]. Therefore, the properties of video sequences can also be exploited when using SSIM to define the distortion model.

In H.264/AVC, the encoder processes a frame of video in units of non-overlapping MBs. However, SSIM index is meant to be calculated with the help of overlapping sliding

windows, which are separated by one pixel. To bridge this gap, we calculate the SSIM index between the reconstructed MB and the original MB using an extended MB, which includes the current MB to be coded and the surrounding pixels, as illustrated in Fig. 6.1. Within this extended MB we use a small sliding window which moves pixel by pixel to calculate the SSIM index. Since the smallest size of modes in H.264/AVC is 4×4 (e.g., I4MB), in order to be consistent with the current video coding standards, the size of the sliding window used to calculate SSIM is set to be 4×4 for luminance components. To compute the SSIM index of the chrominance components at the same scale, the sliding window size for Cb and Cr is also set to be 4×4. Therefore, we extend the MB boundaries for three pixels in each direction. For Y component, the SSIM index of the current 16×16 MB to be encoded is calculated within a 22×22 extended MB by using the sliding window. In case of 4:2:0 format, for Cb and Cr components the SSIM index is calculated within a 14×14 extended block. Additional benefit of this approach is that it helps us to alleviate the problem of discontinuities at the MB boundaries. When the MB is on the frame boundaries, we ignore the surrounding pixels in the distortion calculation and only use the MB to be coded for comparison.

Finally, SSIM indices of Y, Cb and Cr components are weighted averaged to obtain a single measure of structural similarity.

$$SSIM = W_Y \cdot SSIM_Y + W_{Cb} \cdot SSIM_{Cb} + W_{Cr} \cdot SSIM_{Cr}, \tag{6.3}$$

where W_Y, W_{Cb} and W_{Cr} are the weights of Y, Cb and Cr components, respectively and are defined as $W_Y = 0.8$ and $W_{Cb} = W_{Cr} = 0.1$, respectively [171].

6.3 Frame Level Lagrange Multiplier Selection

From (6.2), the Lagrange parameter is obtained by calculating the derivative of J with respect to R, then setting it to zero, and finally solving for λ,

$$\frac{dJ}{dR} = -\frac{dSSIM}{dR} + \lambda = 0, \tag{6.4}$$

which yields:

$$\lambda = \frac{dSSIM}{dR} = \frac{\frac{dSSIM}{dQ}}{\frac{dR}{dQ}}, \tag{6.5}$$

where Q is the quantization step. This implies that, in order to estimate λ before actually encoding the current frame, we need to establish accurate SSIM and rate models.

In video coding the most common models for the distribution of transformed residuals are Laplace distribution [79], generalized Gaussian distribution (GGD) [142] and Cauchy distribution [2]. Although GGD is a good statistical model to describe the DCT coefficients, it has more control parameters and closed-form expression of the distortion model can not be obtained [142]. For Cauchy distribution, the mean and variance are not defined, which makes it inappropriate for this framework [79]. The Laplace distribution, which is a special case of GGD, does not suffer from these problems and achieves a good trade-off between model fidelity and the complexity. Therefore, we model the transformed residuals x with the Laplace distribution given by

$$f_{Lap}(x) = \frac{\Lambda}{2} \cdot e^{-\Lambda \cdot |x|}, \tag{6.6}$$

where Λ is called the Laplace parameter.

6.3.1 Reduced Reference SSIM Model

SSIM is a full-reference (FR) measure that requires both the reference and distorted frames to compute. It can not be directly applied in this framework because the distorted frame is not available. Therefore, we develop a reduced-reference (RR) quality assessment algorithm which requires a set of RR features extracted from the reference frame for SSIM estimation. The RR-SSIM estimation method based on a multi-scale multi-orientation Divisive Normalization Transform (DNT) is proposed in Chapter 3, [117] and achieves high SSIM estimation accuracy. However, it can not be directly employed due to the high computational complexity of DNT. We use a similar approach here, but extract features from DCT coefficients instead.

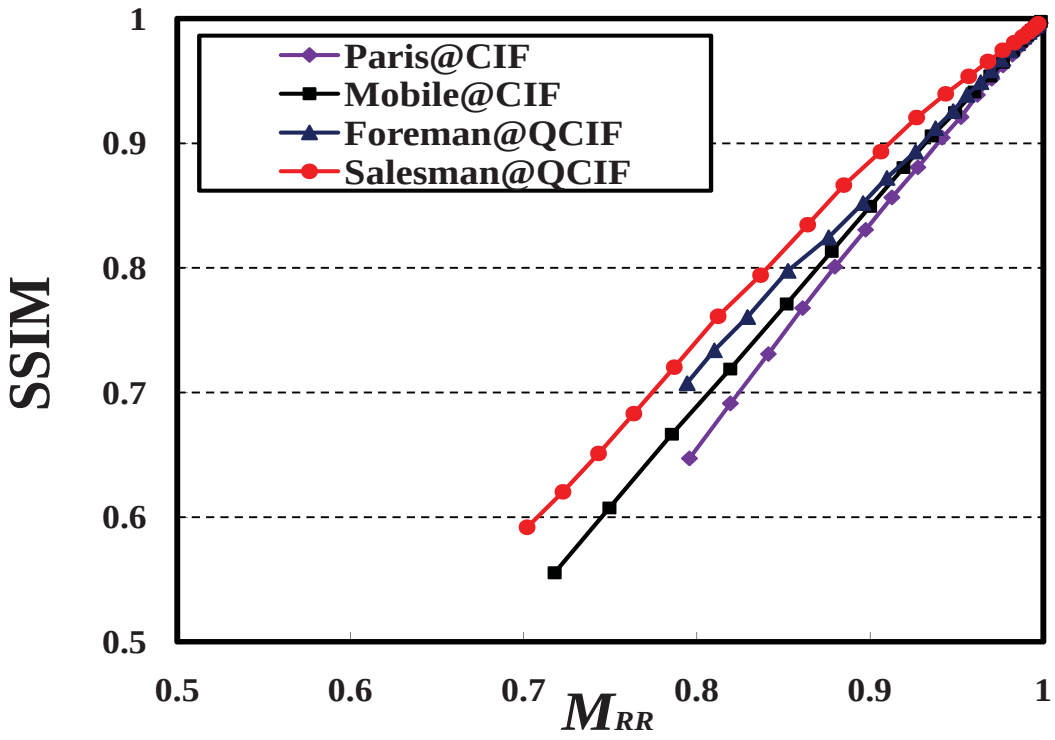


Figure 6.2: Relationship between SSIM and M_{RR} for different sequences.

FR DCT domain SSIM index was first presented by Channappayya *et al.* [25].

$$SSIM(\mathbf{x},\mathbf{y}) = \{1 - \frac{(X(0)-Y(0))^2}{X(0)^2+Y(0)^2+N\cdot C_1}\} \times \{1 - \frac{\sum_{k=1}^{K-1}(X(k)-Y(k))^2}{\sum_{k=1}^{K-1}(X(k)^2+Y(k)^2)+N\cdot C_2}\}, \tag{6.7}$$

where $X(k)$ and $Y(k)$ represent the DCT coefficients for the input signals $\mathbf{x}$ and $\mathbf{y}$, respectively. This equation implies that the SSIM index is represented by the product of two terms, characterizing the distortions of the DC and AC coefficients, respectively. Moreover, the squared errors of DC and AC coefficients are normalized by their respective energy.

To develop the RR-SSIM model, we divide each frame into non-overlapping blocks and the size of each block is set to be 4x4. Then DCT transform is performed on each block. In this way, we can obtain the statistical properties of the reference signal, which is consistent with the design philosophy of the SSIM index. Furthermore, we group the DCT coefficients having the same frequency from each 4x4 DCT window into one subband, which results in 16 subbands. Motivated by the DCT domain SSIM index, the new RR distortion measure is defined as

$$M_{RR} = (1 - \frac{D_0}{2\sigma_0^2+C_1})(1 - \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{D_i}{2\sigma_i^2+C_2}), \tag{6.8}$$

where σ_i is the standard deviation of the i^{th} subband and N is the block size. D_i represents the MSE between the original and distorted frames in the i^{th} subband, and is calculated as follows

$$D_i = \int_{-(Q-\gamma Q)}^{(Q-\gamma Q)} x_i^2 f_{Lap}(x_i)dx_i + 2\sum_{n=1}^{\infty}\int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q}(x_i-nQ)^2 f_{Lap}(x_i)dx_i, \tag{6.9}$$

where γ is the rounding offset in the quantization. Fig. 6.2 presents the relationship between the reduced reference distortion measure M_{RR} and the corresponding SSIM index for different sequences. The QP values in Fig. 6.2 cover a wide range from 0 to 50 with an interval of 2. The SSIM index and M_{RR} are calculated by averaging the respective values of individual frames. Interestingly, M_{RR} exhibits a nearly perfect linear relationship with SSIM. We regard this as an outcome of the similarity between their design principles. The

clean linear relationship also helps us to design an SSIM predictor based on M_{RR} because the remaining job is just to estimate the slope and intercept of the straight line. More specifically, an RR-SSIM estimator can be written as

$$\hat{S} = \alpha + \beta \cdot M_{RR}. \tag{6.10}$$

The proposed RR-SSIM model is totally based on the features extracted from the original frames in the DCT domain and the residuals. It can be observed from Fig. 6.2 that the slopes for different video sequences are different. Thus, before coding the current frame we should first estimate the parameters α and β. This requires the knowledge of two points on the straight line relating $\hat{S}$ and M_{RR}. We use $(1, 1)$ as one of the points as it is always located on the line and also because it does not require any computation. This solves half of the problem as we still need $\hat{S}$ and M_{RR} of the second point. The SSIM index $\hat{S}$ and Laplace parameter for each subband Λ_i is not available since we have not encoded the frame yet. Therefore, we estimate them from the previous frames of the same type. The estimation details are provided in Section 6.5. The distortion measure M_{RR} can be calculated by incorporating (6.9) into (6.8), and the standard deviation of the i^{th} subband σ_i is calculated by DCT transform of the original frame. This procedure provides us with the second point required to find out α and β.

6.3.2 Proposed Rate Model

Our rate model is derived based on an entropy model that excludes the bit rate of the skipped blocks [79]:

$$H = (1 - P_s) \cdot [-\frac{P_0 - P_s}{1 - P_s} \cdot \log_2 \frac{P_0 - P_s}{1 - P_s} - 2\sum_{n=1}^{\infty} \frac{P_n}{1 - P_s} \cdot \log_2 \frac{P_n}{1 - P_s}], \tag{6.11}$$

where P_s is the probability of the skipped blocks, P_0 and P_n are the probabilities of transformed residuals quantized to the zero-th and n-th quantization levels, respectively, which can be modeled by the Laplace distribution as follows

$$P_0 = \int_{-(Q-\gamma Q)}^{(Q-\gamma Q)} f_{Lap}(x)dx, \tag{6.12}$$

$$P_n = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} f_{Lap}(x)dx. \tag{6.13}$$

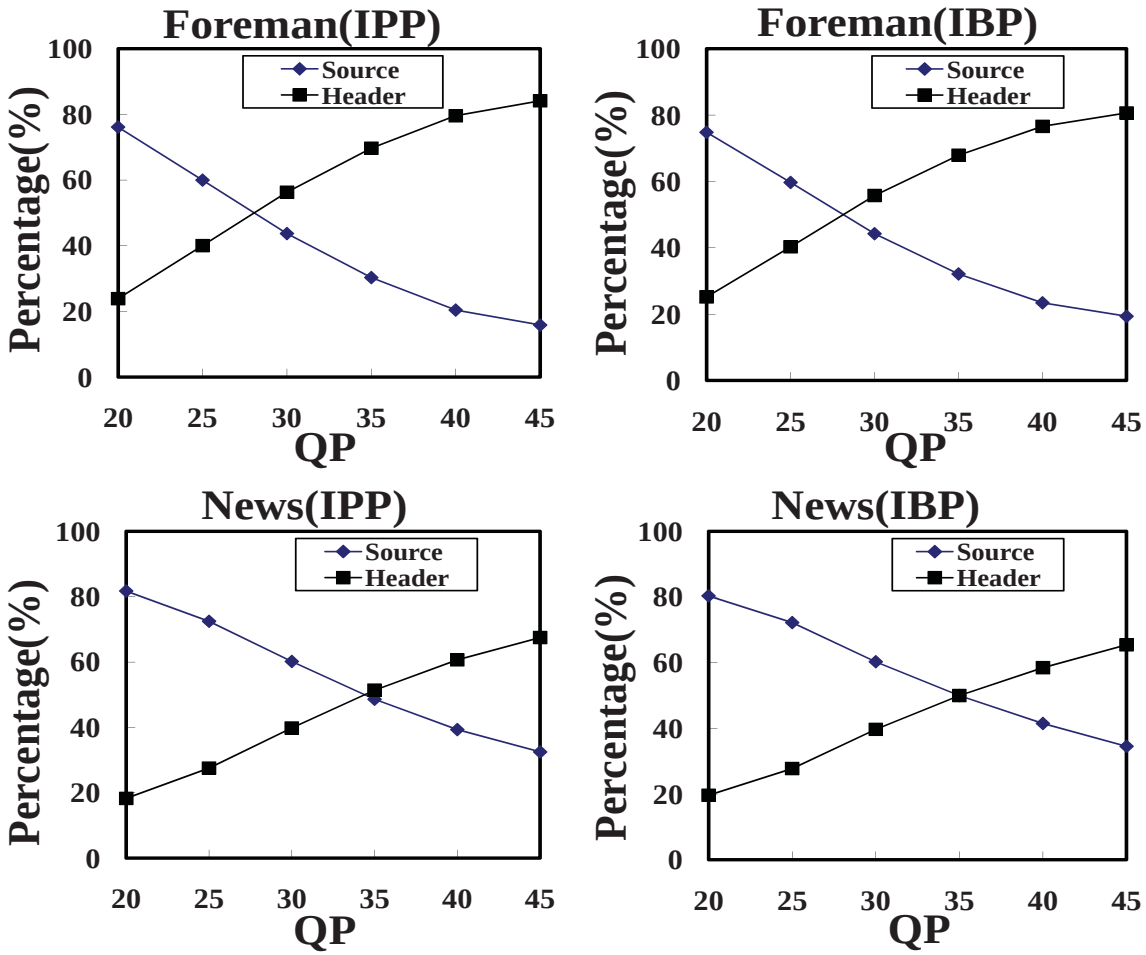


Figure 6.3: Average percentages of header bits and source bits at various QPs.

Subsequently, supposing the rate model in [79] to be R^*, a linear relationship between $ln(R^*/H)$ and $\Lambda \cdot Q$ is observed, where R^* is based on the assumption of negligible side information. However, in H.264/AVC the side information (or header bits) may take a large portion of the total bit rate, especially in low bit rate video coding scenario [70], as illustrated in Fig. 6.3. Therefore, in our rate model, the side information is also taken into consideration. Notice that for the same quantization step, a larger Λ indicates small

residuals, leading to a larger proportion of the side information. For total bit rate R, there is also an approximately linear relationship between $ln(R/H)$ and $\Lambda \cdot Q$, as can be seen in Figs. 6.4 and 6.5. Also, the relationship is totally consistent with the effect of dependent entropy coding and side information. In high bit rate video coding scenario, the effect of dependent entropy coding compensates the side information and $ln(R/H)$ approaches zero; while for low bit rate $ln(R/H)$ becomes larger because of the dominating effect of side information, as illustrated in Figs. 6.4 and 6.5.

Fig. 6.6 shows that the header bits change monotonically with the source bits. Consequently, the final rate model R can be approximated by

$$R = H \cdot e^{\xi \Lambda Q + \psi}, \tag{6.14}$$

where ξ and ψ are two parameters to control the relationship between $ln(R/H)$ and $\Lambda \cdot Q$. We can be observe from Figs. 6.4 and 6.5 that the parameters ξ and ψ are not very sensitive to the video content. Also, for B frames the slope is smaller than that of the I and P frames. It is mainly due to the fact that in case of B frames the residuals are relatively smaller, resulting in a larger value of Λ. Therefore, for both CAVLC and CABAC entropy coding methods, ξ and ψ are set empirically to be

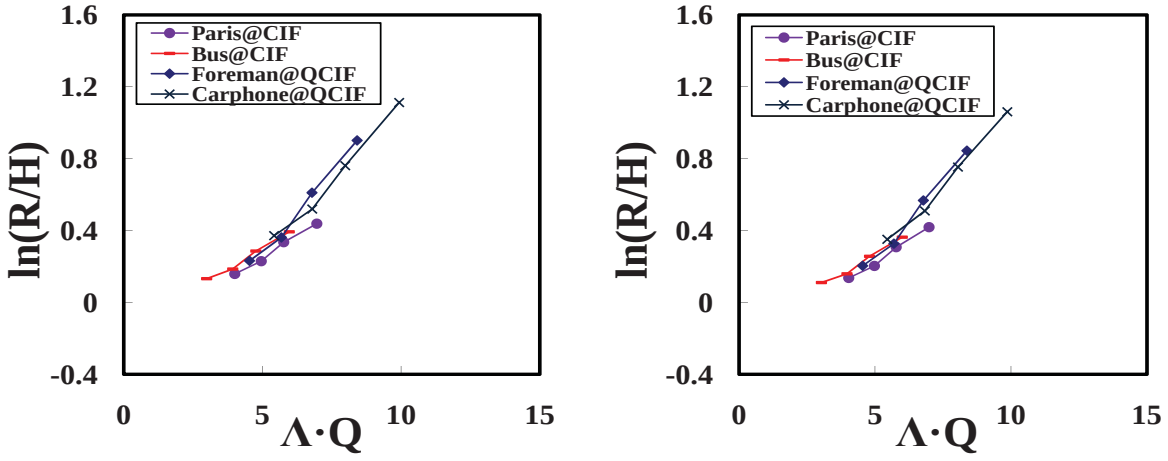
$$\xi = \begin{cases} 0.03 & B \quad frame \\ 0.07 & Otherwise \end{cases} \psi = \begin{cases} -0.07 & B \quad frame \\ -0.1 & Otherwise \end{cases} \tag{6.15}$$

To validate the parameters setting, we use the R^2 [40] metric to examine the accuracy of the fitting. R^2 value lies between 0.0 and 1.0 and a higher value indicates a better fitting. We test the IBP and IPP GOP structures using the parameters in (6.15) with both CAVLC and CABAC entropy coding algorithms. In this test, QP value varies from 15 to 45 with an interval of 5. As shown in Table 6.1, most of the R^2 values are more than 0.9 and some are higher than 0.95. These results suggest that our parameter setting is effective to capture the properties of the R-Q model for video sequences of different video content.

There is one limitation of the proposed rate model. At low bit rate, the skip mode is selected more often and hence the source rate of sequences coded at low bit rate is close to zero. The proposed rate model does not work well in such a situation because the side

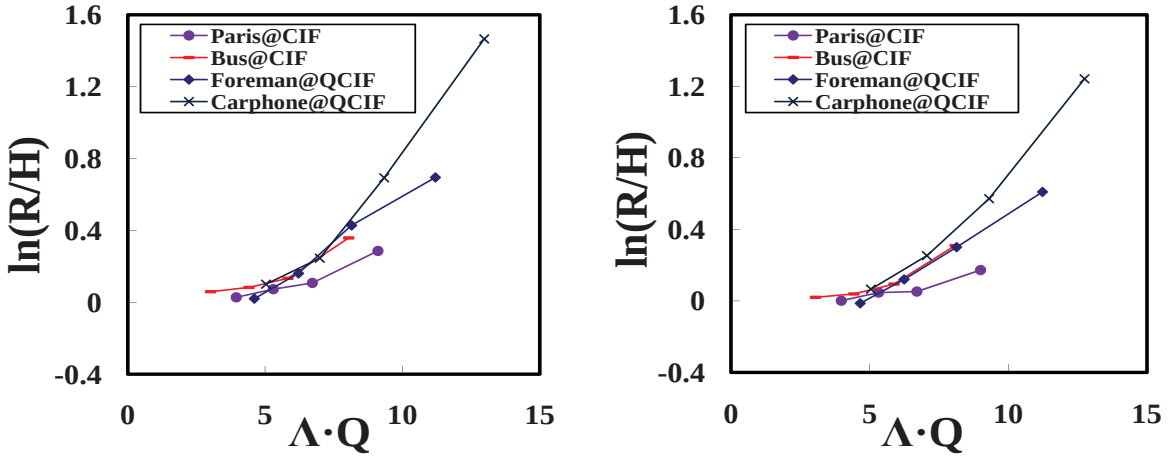
Table 6.1: R^2 Fitting Test for the Proposed Rate-Q Model

Sequences			R^2
Mobile(CIF)	IPP	CABAC	0.9623
		CAVLC	0.9703
	IBP	CABAC	0.9422
		CAVLC	0.9501
Coastguard(CIF)	IPP	CABAC	0.9577
		CAVLC	0.9602
	IBP	CABAC	0.9601
		CAVLC	0.9645
Highway(QCIF)	IPP	CABAC	0.9438
		CAVLC	0.9421
	IBP	CABAC	0.9651
		CAVLC	0.9710
News(QCIF)	IPP	CABAC	0.9682
		CAVLC	0.9527
	IBP	CABAC	0.9095
		CAVLC	0.8826



(a) (b)

Figure 6.4: The relationship between $ln(R/H)$ and $\Lambda \cdot Q$ for different sequences (GOP Structure: IPP). (a) CAVLC entropy coding. (b) CABAC entropy coding.



(a) (b)

Figure 6.5: The relationship between $ln(R/H)$ and $\Lambda \cdot Q$ for B frame of different sequences. (a) CAVLC entropy coding. (b) CABAC entropy coding.

information modeling is based on the source rate. Efficient model of the side information is still an open problem.

Based on the statistical model of the transformed residuals, we obtain the final closed-form solutions of the R and D models. It is observed that the R and D models are functions of two sets of variables: Q and the other variables that describe the inherent properties of the video sequences such as Λ_i and σ_i. When Q varies within a small range, it can be regarded as independent of the other variables [79]. Consequently, before coding the current frame, the frame level Lagrange multiplier can be determined by incorporating the closed-form expressions of R and D into (6.5).

6.4 Macroblock Level Lagrange Multiplier Adjustment

Natural video sequence is not just a stack of independent still images, it also contains critical motion information that relates these images. Therefore, the frames in a natural video can not be considered independently as far as HVS is concerned. Perception of motion information between frames plays an important role towards video quality assessment by HVS. In the conventional video coding framework, motion estimation is performed solely for motion compensation purposes in order to reduce the amount of data to be transmitted. Once the residual frame is calculated, all the MBs are considered equally for bit allocation. This does not conform with HVS, as *perceptual information content* is different in each MB that depends on the *motion information content* and *perceptual uncertainty* in video signals [167]. In [66], the relationship among the Lagrange multiplier λ, the corresponding rate R, and the distortion D was analyzed. A larger λ results in a higher D and a lower R and vice versa, which implies that we can influence the rate and perceptual distortion of each MB by adjusting its Lagrange multiplier. This motivated us to assign more bits to the MBs which are more important as far as *perceptual information content* is concerned. Lagrange multiplier is adjusted with the help of a spatiotemporal weighting factor, η, which increases with the information content and decreases with the perceptual uncertainty.

We employ the scheme proposed in [167] which uses an information communication framework to model the visual perception. We define the relative motion vector, v_r, as the

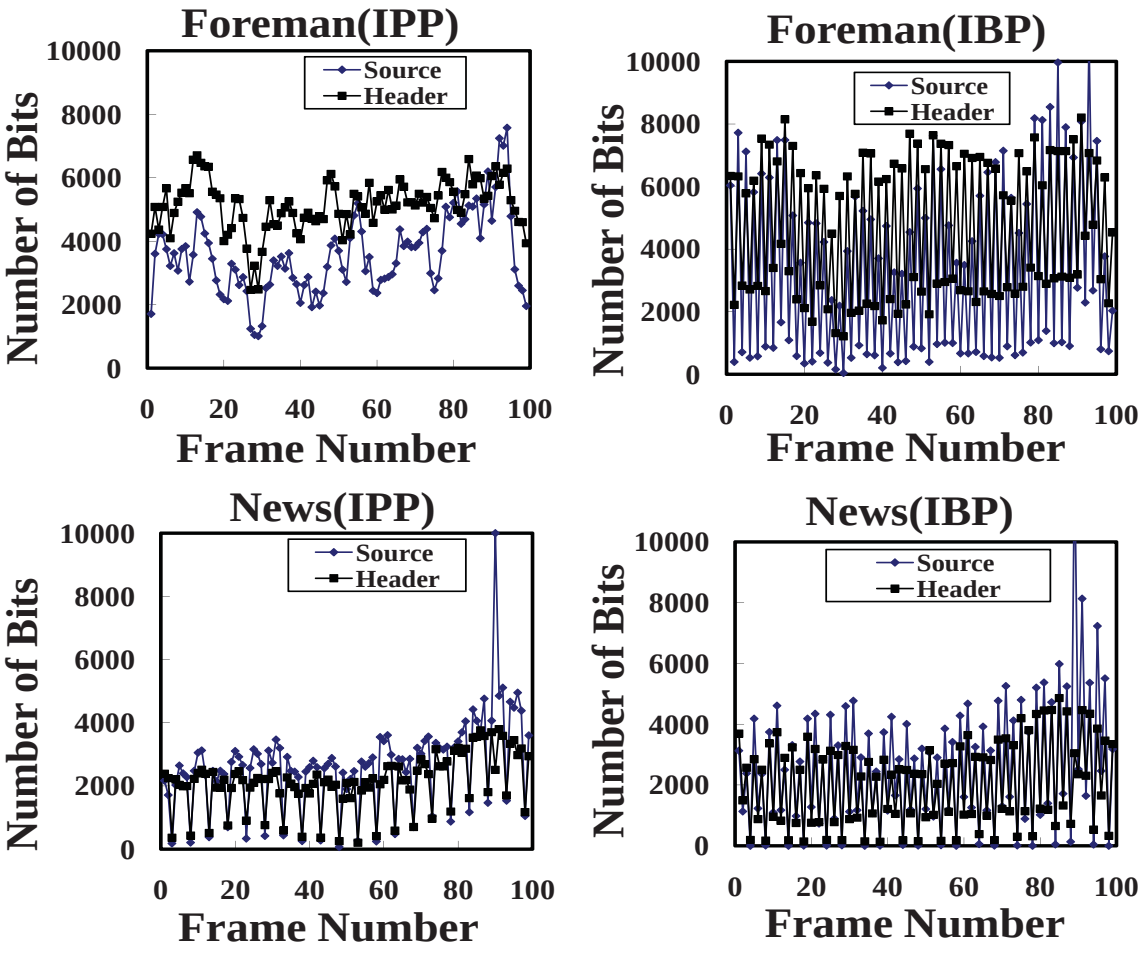


Figure 6.6: The source bits and header bits for each frame at QP=30.

difference between the absolute motion vector, v_a, and global background motion vector v_g:

$$v_r = v_a - v_g. \tag{6.16}$$

In [137], the visual judgment of the speed of motion is modeled by combining some prior knowledge of the visual world and the current noisy measurements. Based on this approach, the motion information content is estimated by the self-information of the relative motion

$$I = \varphi \log v_r + \nu, \tag{6.17}$$

where φ, ν are the parameters of power-law function for the distribution of relative motion and are determined based on psychophysical study conducted in [137].

The perceptual uncertainty is estimated by the entropy of the likelihood function of the noisy measurement, which is given by

$$U = \log v_g - \tau \log c + \delta, \tag{6.18}$$

where τ and δ are the parameters of the log-normal distribution, used to determine perceptual uncertainty, determined based on psychophysical study [136]. The spatio-temporal importance weight function is given by

$$\omega = I - U = \varphi \log v_r + \nu - \{\log v_g - \tau \log c + \delta\}. \tag{6.19}$$

The contrast measure c can be derived by

$$c = 1 - e^{-(c'/\phi)^{\kappa}}, \tag{6.20}$$

$$c' = \frac{\sigma_p}{\mu_p + \mu_0}, \tag{6.21}$$

where σ_p and μ_p are computed within the MB, representing the standard deviation and the mean, respectively. κ and ϕ are constants that control the slope and the position of the functions, respectively [167] and are used to take into account the contrast response

saturation effect at small and large contrast values. It is important to note that the weighting factor is not sensitive to these constants. μ_0 is a constant to avoid instability near 0.

The global motion does not influence the perceptual weight of each MB, thus the weight for each MB is defined as follows

$$\omega = \log(1 + \frac{v_r}{v_0}) + \log(1 + \frac{c}{c_0}), \tag{6.22}$$

where v_0 and c_0 are constants used to avoid unstable evaluation of the weight function when the relative motion v_r and the local contrast c may be close to zero. Note that this weight function increases monotonically with the relative motion and the local contrast, which is in line with the philosophy of visual attention. Consequently, the MBs with higher weights should be allocated more bits and vice versa. This motivated us to adjust the Lagrange multiplier by

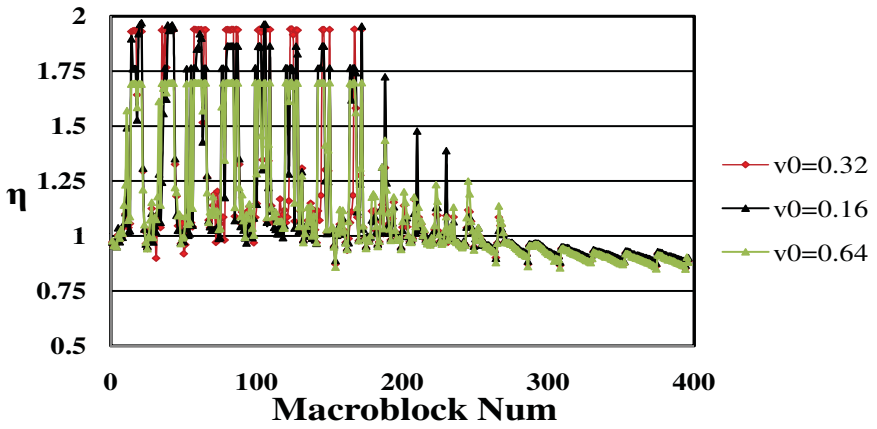


Figure 6.7: The relationship between η and different settings of v_0 for each MB for the Flower sequence.

$$\lambda' = \eta \cdot \lambda. \tag{6.23}$$

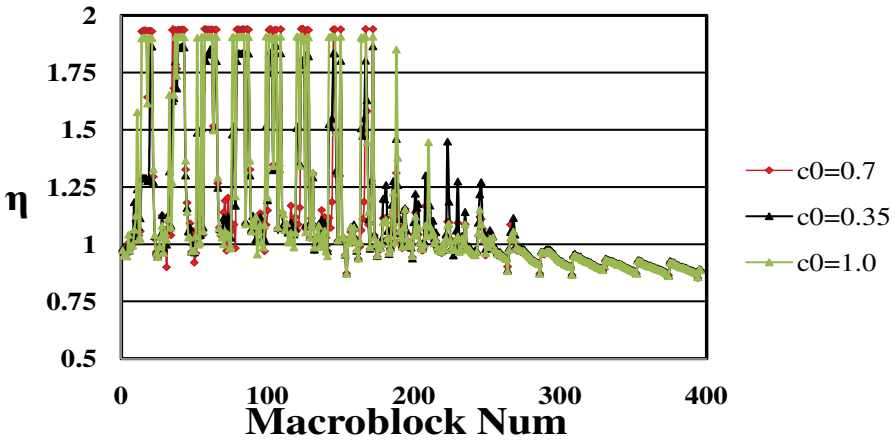


Figure 6.8: The relationship between η and different settings of c_0 for each MB for the Flower sequence.

$$\begin{aligned} D_{dc} &= E[X(0) - Y(0)]^2 & E_{dc} &= E[\frac{1}{2X(0)^2 + N \cdot C_1}] \\ D_{ac} &= E[\sum_{k=1}^{N-1} (X(k) - Y(k))^2] & E_{ac} &= E[\frac{1}{2\sum_{k=1}^{N-1} X(k)^2 + N \cdot C_2}] \end{aligned} \tag{6.25}$$

To determine the adjustment factor η for every MB, we calculate the weight based on the local information, then η is determined in a similar manner as in [155].

$$\eta = (\frac{\omega_{avg}}{\omega})^{\epsilon}. \tag{6.24}$$

The parameter ω_{avg} represents the average weight of the current frame and ϵ is set to be 0.25 as in [155]. Furthermore, as indicated in Fig. 6.7 and 6.8, the final adjustment factor for the Lagrange multiplier is not sensitive to the parameter setting of v_0 and c_0. Therefore, following [167], we set v_0=0.32 and c_0=0.70.

6.5 Implementation Issues

The Lagrange parameter should be determined before coding the current frame in order to perform RDO. However, the parameters Λ_i,$\hat{S}$, Λ, ω_{avg} and v_g can only be calculated after coding the current frame. As shown in Figs. 6.9 and 6.10, the parameters of the frames with the same coding type varies smoothly even for sequences of high motion. This is due to the fact that the inherent properties of the input sequences can be considered unchanged during a short period of time. Therefore, we estimate them by averaging their three previous values from the frames coded in the same manner, i.e.,

$$\hat{\Lambda}_i^j = \frac{1}{3}\sum_{n=1}^{3} \Lambda_i^{j-n} \tag{6.26}$$

where the j indicates the frame number. The global motion vector, v_g, is derived using maximum likelihood estimation which finds the peak of the motion vector histogram [151].

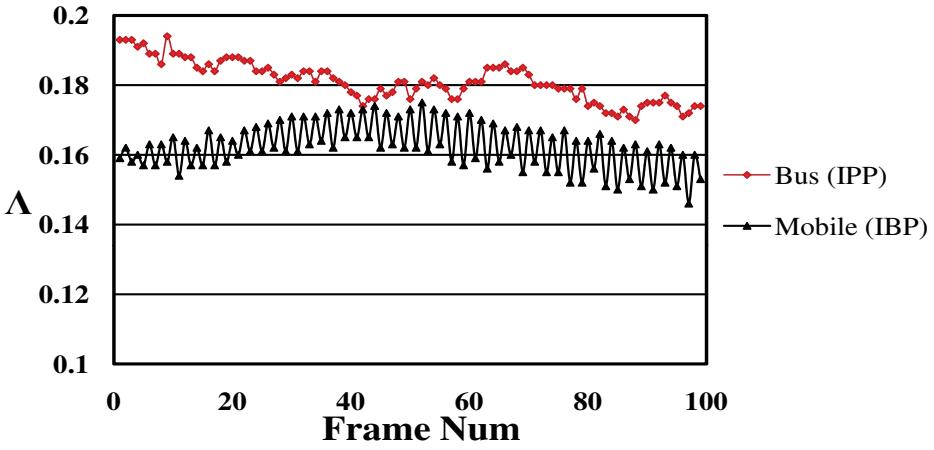


Figure 6.9: Laplace distribution parameter Λ for each frame in Bus (IPP) and Mobile (IBP) with CIF format.

To encode the first few frames, the adaptive Lagrange multiplier selection method is not used since it is difficult to estimate Λ_i, $\hat{S}$, Λ, ω_{avg} and v_g. Motivated by the high rate

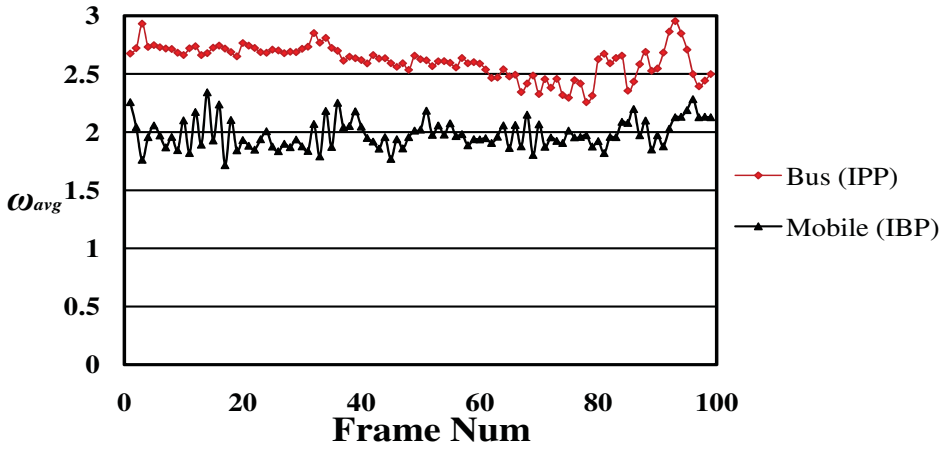


Figure 6.10: Average weight ω_{avg} for each frame in Bus (IPP) and Mobile (IBP) with CIF format.

λ selection method [140, 181], we derive a Lagrange multiplier based on the high bit rate assumption for such a situation.

With the high rate assumption, the SSIM index in the DCT domain can be approximated by the following equation [156]

$$\begin{aligned} E[SSIM(\mathbf{x},\mathbf{y})] \approx \{1 - E[X(0) - Y(0)]^2 \times E[\frac{1}{2X(0)^2 + N \cdot C_1}]\} \\ \times \{1 - E[\sum_{k=1}^{N-1} (X(k) - Y(k))^2] \times E[\frac{1}{2\sum_{k=1}^{N-1} X(k)^2 + N \cdot C_2}]\} \end{aligned} \tag{6.27}$$

where E denotes the mathematical expectation operator. Furthermore, we use D_{dc}, D_{ac}, E_{dc}, E_{ac} to simplify this equation (6.25) and the expectation of SSIM index can be rewritten as:

$$E[SSIM(x,y)] = (1 - E_{dc} \times D_{dc}) \times (1 - E_{ac} \times D_{ac}). \tag{6.35}$$

If the high rate assumption is valid, the source probability distribution can be approximated as uniform distribution and the MSE can be modeled by [55]

$$D = s \cdot Q^2. \tag{6.36}$$

The Lagrange multiplier based on the high rate assumption rate and MSE models is then given by [181]

$$\hat{\lambda}_{HR} = -\frac{dD}{dR} = c \cdot Q^2, \tag{6.37}$$

where c is a constant. Therefore, the general form of λ_{HR} can be derived by calculating the derivative of SSIM with respect to R, which leads to

$$\lambda_{HR} = -\frac{d(E_{ac} \cdot E_{dc} \cdot D_{ac} \cdot D_{dc})}{dR} + \frac{d(E_{dc} \cdot D_{dc})}{dR} + \frac{d(E_{ac} \cdot D_{ac})}{dR}. \tag{6.38}$$

Although E_{ac} and E_{dc} are based on the properties of the frames, to provide a constant solution for SSIM based RDO in the first few frames, we derive a general solution for them. Considering (6.36),(6.37),(6.38), the constant Lagrange multiplier for SSIM based RDO can be expressed by:

$$\lambda_{HR} = a \cdot Q^2 - b \cdot Q^4. \tag{6.39}$$

The values for a and b are determined empirically by experimenting with SSIM and the rate models:

$$a = \begin{cases} 2.1 \times 10^{-4} & B \quad frame \\ 7 \times 10^{-5} & otherwise \end{cases} \tag{6.40}$$

$$b = \begin{cases} 1.5 \times 10^{-9} & B \quad frame \\ 5 \times 10^{-10} & Otherwise \end{cases} \tag{6.41}$$

In our rate model (6.14), the modeling of side information is totally based on the source rate. In the extreme case, e.g., when the source rate is zero, this rate model will fail because the header bit can not be zero in the real video coding scenario. Therefore, we propose an escape method to keep a reasonable performance, where the Lagrange multiplier is given by

$$\lambda = \begin{cases} \lambda_{HR} & H = 0 \\ \frac{\frac{dSSIM}{dQ}}{\frac{dR}{dQ}} & otherwise \end{cases} \tag{6.42}$$

We summarize the whole process of proposed RDO scheme for IPP coding structure in Algorithm 1. Similar process applies to IBP as well. We can be observe that the complexities introduced by the proposed method are only moderate. The additional computations

Algorithm 5: Summary of the proposed RDO. (GOP structure: IPP)

begin

Calculate λ_i for the i^{th} frame

switch *the value of* i **do**

case *0,1,2,3*

$$\lambda_i \leftarrow \lambda_{HR}$$

end

otherwise

1. DCT transform of the input frame.
2. $\lambda_i \leftarrow \begin{cases} \lambda_{HR} & H = 0 \\ \frac{\frac{dSSIM}{dQ}}{\frac{dR}{dQ}} & otherwise \end{cases}$

end

endsw

end

begin

For each MB in the frame

1. Calculate the scale factor at MB level η.
2. Adjust the Lagrange multiplier:

$$\lambda_i' \leftarrow \eta \cdot \lambda_i$$

3. Calculate the RD cost for each Mode k:

$$J_k \leftarrow 1 - SSIM_k + \lambda_i' \cdot R_k$$

4. Select the Mode j with minimal RD cost.
5. Encode the MB with Mode j.

end

begin

Update Λ_i, $\hat{S}$, Λ, ω_{avg} and v_g.

end

are the DCT transform of the original frame, the calculation of the parameters (Λ_i, $\hat{S}$, Λ, ω_{avg} and v_g) and the calculation of SSIM for each mode.

6.6 Validations

To validate the accuracy and efficiency of the proposed perceptual RDO scheme, we integrate our mode selection scheme into the H.264/AVC reference software JM15.1 [67]. All test video sequences are in YCbCr 4:2:0 format. In this section, we present the results of three experiments which are used to validate various aspects of the proposed perceptual RDO algorithm. In the first experiment, we verify the proposed RR-SSIM model by comparing estimated SSIM values with actual SSIM values. In the second experiment, the performance of the proposed perceptual RDO algorithm is evaluated and compared with that of the conventional RDO scheme. In the third experiment, we compare the proposed method with state-of-the-art SSIM and MSE-based RDO schemes.

6.6.1 Comparison between Estimated and Actual SSIM

In this subsection, we compare the estimated (RR) and actual (FR) values of the SSIM index for different sequences with a set of various QP values. The first frame is I-frame while all the rest are inter-coded frames. Equation (6.10) suggests that we first need to calculate the parameters α and β which vary across different video content. Thus, for each frame, we calculate the slope with the help of two points. $(\hat{S}, M_{RR})$ and (1,1), where the point $(\hat{S}, M_{RR})$ is obtained by setting QP=40, the middle point among the quantization steps used for testing the proposed scheme. Once α and β are determined, we can use (6.10) to estimate SSIM for other QP values. Fig. 6.11 plots the estimated and actual values of the SSIM index for various values of QP. It is observed that the proposed SSIM model is robust and accurate for different video contents with different resolutions. Moreover, we have also calculated the Pearson Linear Correlation Coefficient (PLCC) and Mean Absolute Error (MAE) between FR-SSIM and RR-SSIM which are given in Table 6.2 for ten different sequences. The values suggest that the proposed RR-SSIM model achieves high accuracy for different sequences.

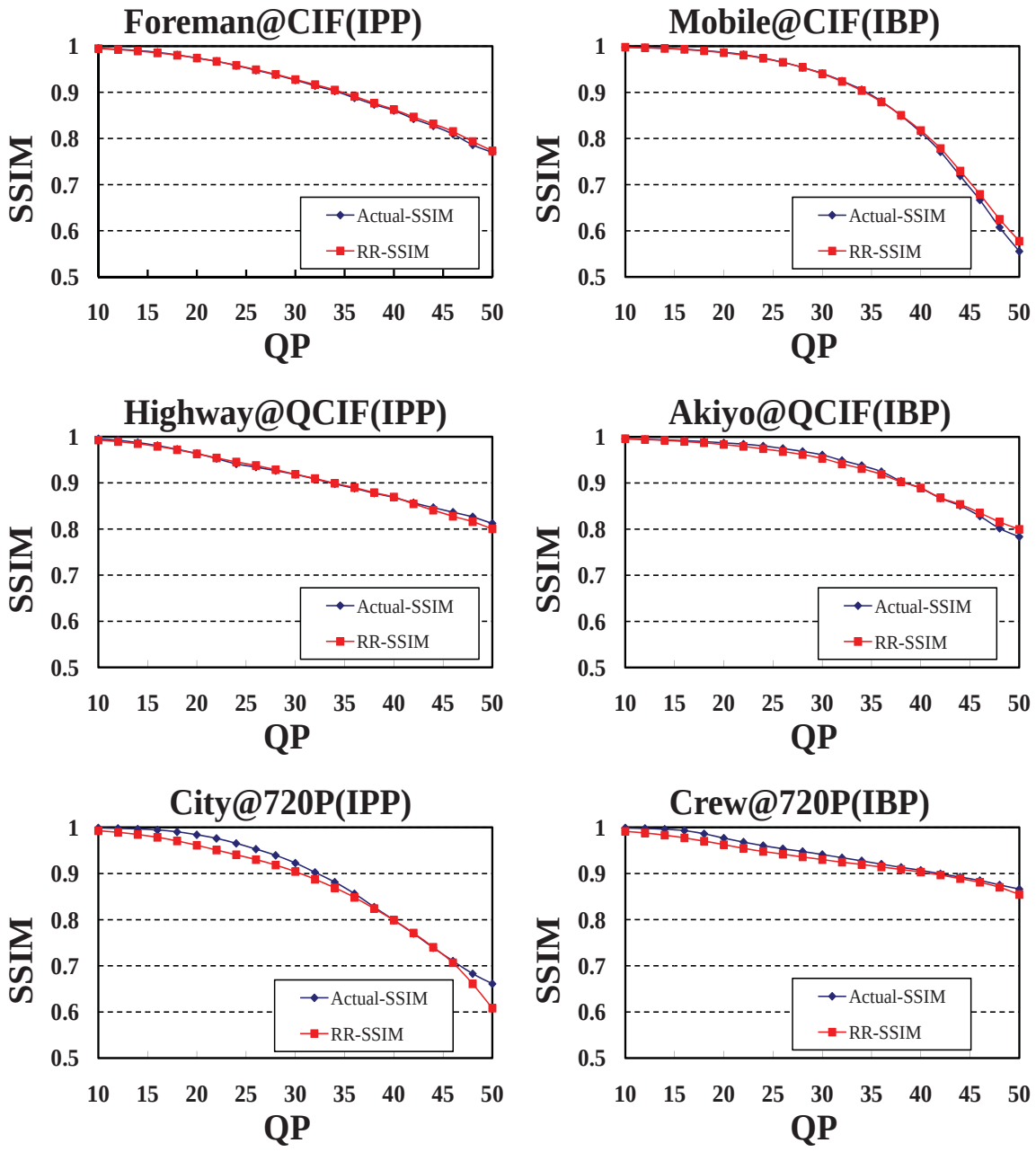


Figure 6.11: Comparison between the actual FR-SSIM and estimated RR-SSIM values.

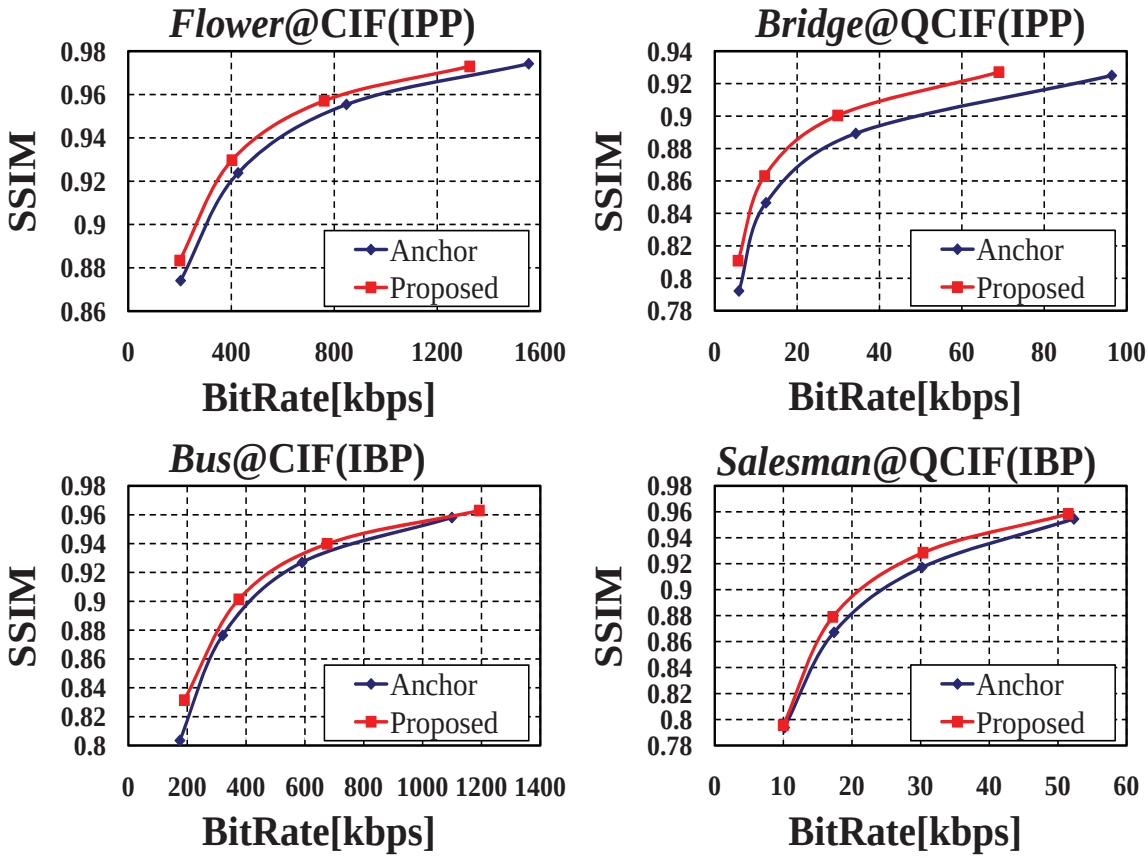


Figure 6.12: Performance comparisons of different RDO algorithms for sequences with CABAC entropy coding method.

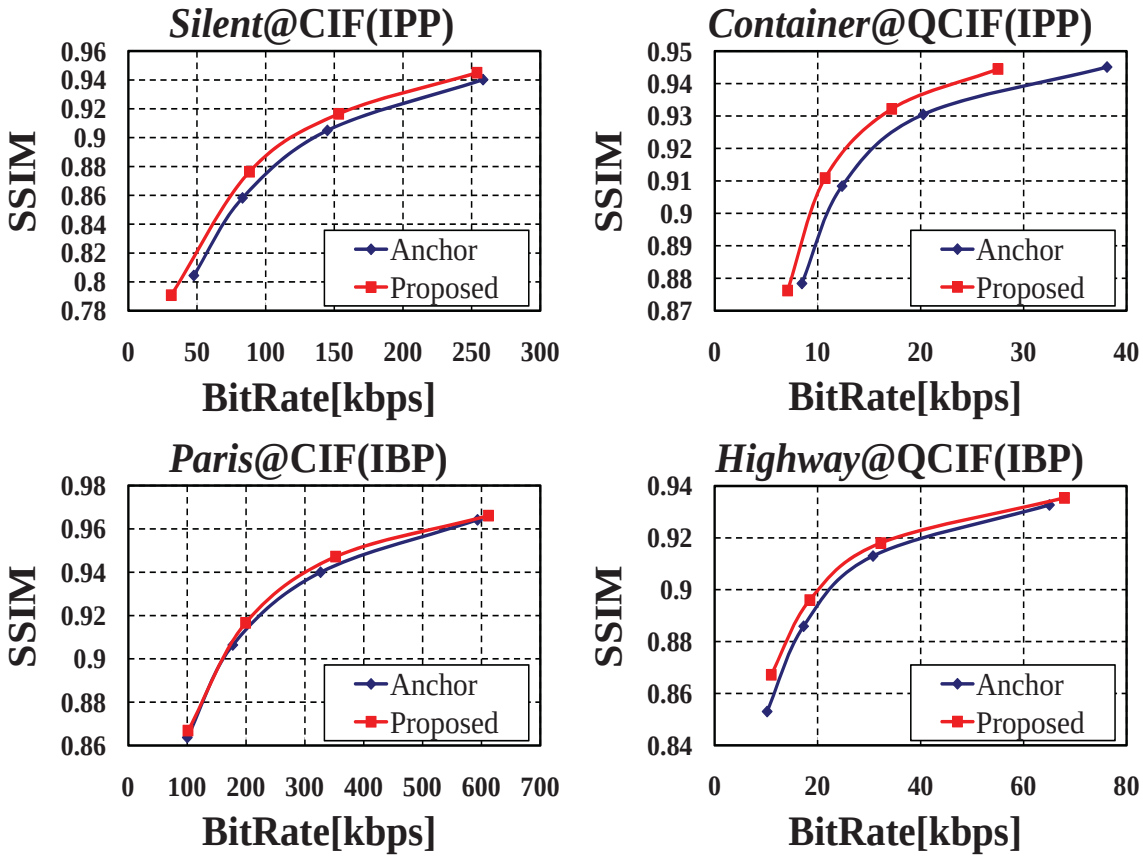


Figure 6.13: Performance comparisons of different RDO algorithms for sequences with CAVLC entropy coding method.

Table 6.2: MAE and PLCC between FR-SSIM and RR-SSIM Estimation for Different Sequences

Sequences	GOP Structure	PLCC	MAE
Foreman(CIF)	IPP	0.999	0.002
News(CIF)	IPP	0.999	0.002
Mobile(CIF)	IBP	0.999	0.004
Paris(CIF)	IBP	0.999	0.003
Highway(QCIF)	IPP	0.998	0.003
Suize(QCIF)	IPP	0.998	0.004
Carphone(QCIF)	IBP	0.997	0.006
Akiyo(QCIF)	IBP	0.998	0.005
City(720P)	IPP	0.994	0.015
Crew(720P)	IBP	0.997	0.009
All		0.996	0.005

6.6.2 Performance Evaluation of the Proposed Algorithms

We compare the RD performance of our proposed perceptual RDO algorithm and the conventional RDO with distortion measured in terms of SSIM, weighted SSIM and PSNR. The three quantities for the whole video sequence are obtained by simply averaging the respective values of individual frames. The size of sliding window to calculate the SSIM index is set to be 8×8. In this experiment, we employ the method proposed in [6] to calculate the differences between two RD curves[1]. Furthermore, the weighted SSIM index is defined as [167]

$$SSIM_{\omega} = \frac{\sum_x \sum_y \omega(x,y) SSIM(x,y)}{\sum_x \sum_y \omega(x,y)} \tag{6.43}$$

where $\omega(x,y)$ indicates the weight value for (x,y) as defined in (6.22). The SSIM indices of Y, Cb and Cr components are combined according to (6.3). Since the $SSIM_{\omega}$ takes the motion information into account, it is more accurate for perceptual video quality assessment

[1]Since R-SSIM curve exhibits a similar shape as R-PSNR curve, we use the same tool proposed in [6] to calculate the average of SSIM differences.

Table 6.3: Performance of the Proposed Algorithms (Compared with Original Rate-Distortion Optimization Technique) for QCIF Sequences at 30 Frames/s

Sequence		CABAC					CAVLC				
		$\Delta SSIM$	ΔR^*	$\Delta SSIM_\omega$	ΔR^{**}	$\Delta PSNR$	$\Delta SSIM$	ΔR^*	$\Delta SSIM_\omega$	ΔR^{**}	$\Delta PSNR$
Akiyo	IPP..	0.0116	-17.85%	0.0142	-19.83%	0.13 dB	0.0123	-19.33%	0.0151	-21.09%	0.21 dB
	IBP..	0.0075	-5.77%	0.0100	-8.93%	-0.06 dB	0.0091	-9.64%	0.0116	-11.17%	0.06 dB
Bridge-close	IPP..	0.0171	-30.65%	0.0192	-34.20%	-0.02 dB	0.0194	-35.64%	0.0228	-41.12%	0.01 dB
	IBP..	0.0148	-29.11%	0.0168	-32.77%	-0.15 dB	0.0150	-30.90%	0.0177	-35.98%	-0.17 dB
Highway	IPP..	0.0108	-21.00%	0.0127	-20.70%	-0.26 dB	0.0109	-21.78%	0.0144	-23.09%	-0.42 dB
	IBP..	0.0043	-7.80%	0.0057	-9.40%	-0.49 dB	0.0046	-10.91%	0.0064	-12.82%	-0.46 dB
Grandma	IPP..	0.0188	-23.03%	0.0219	-25.38%	0.25 dB	0.0192	-22.70%	0.0220	-24.47%	0.28 dB
	IBP..	0.0158	-19.44%	0.0192	-21.74%	0.13 dB	0.0164	-19.68%	0.0198	-21.59%	0.14 dB
Container	IPP..	0.0088	-18.06%	0.0088	-17.12%	-0.10 dB	0.0091	-17.63%	0.0096	-17.01%	-0.10 dB
	IBP..	0.0048	-12.30%	0.0054	-13.11%	-0.47 dB	0.0055	-11.04%	0.0058	-10.72%	-0.47 dB
Salesman	IPP..	0.0189	-17.72%	0.0199	-18.11%	0.11 dB	0.0200	-18.14%	0.0210	-18.28%	0.12 dB
	IBP..	0.0103	-9.44%	0.0125	-11.24%	-0.21 dB	0.0101	-9.25%	0.0118	-10.39%	-0.26 dB
News	IPP..	0.0082	-12.76%	0.0098	-11.82%	-0.15 dB	0.0078	-12.71%	0.0096	-12.96%	-0.19 dB
	IBP..	0.0052	-7.36%	0.0071	-8.56%	-0.35 dB	0.0046	-6.50%	0.0061	-8.21%	-0.38 dB
Carphone	IPP..	0.0035	-6.29%	0.0042	-7.21%	-0.52 dB	0.0034	-5.59%	0.0042	-6.62%	-0.45 dB
	IBP..	0.0010	-2.45%	0.0015	-3.55%	-0.56 dB	0.0010	-2.36%	0.0019	-4.42%	-0.56 dB
Average	IPP..	0.0122	-18.42%	0.0138	-19.30%	-0.07 dB	0.0128	-19.19%	0.0148	-20.58%	-0.07 dB
	IBP..	0.0080	-11.71%	0.0098	-13.66%	-0.27 dB	0.0082	-12.54%	0.0101	-14.41%	-0.26 dB

* Rate reduction while maintaining SSIM.

** Rate reduction while maintaining weighted SSIM.

Table 6.4: Performance of the Proposed Algorithms (Compared with Original Rate-Distortion Optimization Technique) for CIF Sequences at 30 Frames/s

Sequence		CABAC					CAVLC				
		$\Delta SSIM$	ΔR^*	$\Delta SSIM_\omega$	ΔR^{**}	$\Delta PSNR$	$\Delta SSIM$	ΔR^*	$\Delta SSIM_\omega$	ΔR^{**}	$\Delta PSNR$
Silent	IPP..	0.0109	-13.98%	0.0118	-14.69%	-0.18 dB	0.0114	-14.13%	0.0123	-14.85%	-0.21 dB
	IBP..	0.006	-7.79%	0.0077	-9.96%	-0.34 dB	0.0063	-7.84%	0.0074	-9.10%	-0.37 dB
Bus	IPP..	0.0134	-14.85%	0.0122	-13.88%	-0.57 dB	0.0148	-15.61%	0.0136	-14.89%	-0.62 dB
	IBP..	0.0083	-9.39%	0.0087	-9.51%	-0.66 dB	0.0080	-8.63%	0.0081	-8.49%	-0.73 dB
Mobile	IPP..	0.0047	-8.52%	0.0053	-10.50%	-0.58 dB	0.0051	-9.52%	0.0059	-11.76%	-0.63 dB
	IBP..	0.0017	-3.23%	0.0026	-5.52%	-0.64 dB	0.0009	-1.77%	0.0019	-4.35%	-0.68 dB
Paris	IPP..	0.0080	-12.07%	0.0096	-14.35%	-0.38 dB	0.0076	-11.30%	0.0090	-13.69%	-0.43 dB
	IBP..	0.0036	-5.17%	0.0050	-7.36%	-0.62 dB	0.0029	-4.02%	0.0043	-6.55%	-0.36 dB
Flower	IPP..	0.0076	-14.19%	0.0068	-11.69%	-0.57 dB	0.0070	-13.31%	0.0063	-10.86%	-0.71 dB
	IBP..	0.0035	-6.92%	0.0029	-4.65%	-0.47 dB	0.0021	-4.01%	0.0014	-1.78%	-0.71 dB
Foreman	IPP..	0.0023	-4.80%	0.0020	-4.26%	-0.55 dB	0.0028	-5.72%	0.0027	-5.11%	-0.58 dB
	IBP..	0.0008	-1.89%	0.0008	-1.97%	-0.55 dB	0.0009	-1.66%	0.0008	-1.65%	-0.70 dB
Tempete	IPP..	0.0072	-10.28%	0.0083	-11.70%	-0.35 dB	0.0078	-11.27%	0.0088	-12.48%	-0.36 dB
	IBP..	0.0031	-4.13%	0.0040	-5.51%	-0.41 dB	0.0029	-4.26%	0.0038	-5.56%	-0.58 dB
Waterfall	IPP..	0.0207	-15.51%	0.0193	-14.22%	-0.27 dB	0.0237	-17.20%	0.0226	-16.39%	-0.22 dB
	IBP..	0.0097	-9.37%	0.0099	-9.98%	-0.47 dB	0.0092	-8.80%	0.0093	-9.35%	-0.46 dB
Average	IPP..	0.0094	-11.78 %	0.0094	-11.91%	-0.43 dB	0.0100	-12.26%	0.0102	-12.50%	-0.47 dB
	IBP..	0.0046	-5.99%	0.0052	-6.81%	-0.52 dB	0.0042	-5.12%	0.0046	-5.85%	-0.57 dB

* Rate reduction while maintaining SSIM.

** Rate reduction while maintaining weighted SSIM.

[167].

For coding complexity overhead evaluation, we calculate ΔT as follows

$$\Delta T = \frac{T_{pro\_RDO} - T_{org\_RDO}}{T_{org\_RDO}} \times 100\% \tag{6.44}$$

where T_{org_RDO} and T_{pro_RDO} indicate the total coding time with the conventional and the proposed SSIM-based RDO schemes, respectively.

To verify the efficiency of the proposed perceptual RDO method, extensive experiments are conducted on standard sequences in QCIF and CIF formats. In these experiments, RD performance of the conventional RDO coding strategy and the proposed SSIM motivated perceptual RDO coding strategy is compared. The common coding configurations are set as follows: all available inter and intra modes are enabled; five reference frames; one I frames followed by 99 inter frames; high complexity RDO and the fixed quantization parameters are set from 28 to 40. The results of the experiments are shown in Tables 6.3 and 6.4, and the RD performances are compared in Figs. 6.12, 6.13 and 6.14.

For IPP GOP structure, on average 15% rate reduction for fixed SSIM and 16% rate reduction while fixing weighted SSIM are achieved for both QCIF and CIF sequences. When the GOP structure is IBP, the rate reductions are 9% on average for fixed SSIM and 10% on average for fixed weighted SSIM. In general, there are three main reasons behind the improved performance. First, we use SSIM for RDO purposes, which is a better predictor of perceived quality by HVS as compared to ubiquitous MSE. Second, the proposed RR-SSIM model and Rate model are more accurate compared to the ones already existing in the literature. Third, we consider the motion between the frames, which is a crucial information for perception of quality by HVS, to further improve the rate distribution among the MBs considering the HVS. The lower gain of IBP coding scheme may be explained by two reasons. First, the B frame is usually coded at relatively low bit rate while our proposed scheme achieves superior performance at high bit rate compared to low bit rate, as can be observed from Fig. 6.12. Second, the parameters estimation scheme proposed in Section 6.5 is not as accurate for this GOP structure because the frames of the same coding types are not adjacent to each other.

Rate reduction peaks for sequences with slow motion such as *Bridge*, in which case 35% of the bits can be saved for the same SSIM value of the received video. It is observed

that for these sequences with larger Λ, the superior performance is mainly due to the selection of the MB mode with less bits. A similar phenomenon has also been observed in [79] and [66]. Another interesting observation is that the performance gain of the proposed method decreases at very low bit rate, such as the *Bridge* and *Salesman* in Fig. 6.12. This is due to the fact that at low bit rate a large percentage of MBs have already been coded with the best mode in the conventional RDO scheme, such as SKIP mode. Also, the limitation of the proposed rate model as stated in Section 6.3 also brings the limited performance gain at low bit rate. We have also compared the performance in terms of PSNR of the luminance component, which is shown in Tables 6.3, 6.4 and Fig. 6.15. Because our scheme is totally adaptive to the video sequences, for some sequences such as *Akiyo* and *Container*, PSNR increases. However, on average PSNR decreases because our optimization objective is SSIM rather than PSNR.

To show the advantage of our frame-MB joint RDO scheme, the performance comparisons of the frame level perceptual RDO (FP-RDO) and the Frame-MB level percpetual RDO (FMP-RDO) are also listed in Table 6.5. As can be observed from Table 6.5, the weighted SSIM increases for sequences with high motion, such as *Flower* and *Bus*. However, the weighted SSIM decreases for constant sequences, such as *Silent*. This performance degradation mainly comes from the inter prediction technique used in video coding. For instance, the MB with higher weight in the current frame may get the prediction pixels from an unimportant MB in the pervious frame, which can cause more quantization errors. Our current work focuses on RDO frame by frame. The interrelationship between frames and the rate control at the GOP level will be studied in the future.

Fig. 6.16 shows the original frame, H.264/AVC coded frame with the conventional RDO and H.264/AVC coded frame with the proposed RDO method. Note that the bit rates for the two coding methods are almost the same. However, since our proposed RDO scheme is based on SSIM index optimization, higher SSIM and lower PSNR are achieved. Furthermore, the quality of the reconstructed frame has been obviously improved by the proposed scheme. We can be observe that more information and details have been preserved, such as the branches on the roof. The visual quality improvement is due to the fact that we can select the best mode from perceptual point of view, resulting in more bits allocated to the areas which are more sensitive to our visual systems.

Table 6.5: Performance comparison of the Proposed FPRDO and FMPRDO Coding (Anchor: Conventional Rate-Distortion Optimization Technique)

Sequence		CABAC				CAVLC			
		IPPPP		IBPBP		IPPPP		IBPBP	
		ΔR^{*}	ΔR^{**}	ΔR^{*}	ΔR^{**}	ΔR^{*}	ΔR^{**}	ΔR^{*}	ΔR^{**}
Flower(CIF)	FMP-RDO	-14.19%	-11.69%	-6.92%	-4.65%	-13.31%	-10.86%	-4.01%	-1.78%
	FP-RDO	-14.34%	-11.43%	-6.73%	-4.05%	-12.73%	-9.75%	-2.04%	0.38%
Waterfall(CIF)	FMP-RDO	-15.51%	-14.22%	-9.37%	-9.98%	-17.20%	-16.39%	-8.80%	-9.35%
	FP-RDO	-15.45%	-14.43%	-8.79%	-9.47%	-16.13%	-15.48%	-7.98%	-8.62%
Bus(CIF)	FMP-RDO	-14.85%	-13.88%	-9.39%	-9.51%	-15.61%	-14.89%	-8.63%	-8.49%
	FP-RDO	-14.71%	-13.72%	-8.95%	-8.84%	-16.05%	-14.96%	-8.72%	-8.63%
Silent(CIF)	FMP-RDO	-13.98%	-14.69%	-7.79%	-9.96%	-14.13%	-14.85%	-7.84%	-9.10%
	FP-RDO	-14.62%	-15.28%	-8.07%	-9.79%	-15.23%	-15.59%	-8.53%	-9.85%
Salesman(QCIF)	FMP-RDO	-17.72%	-18.11%	-9.44%	-11.24%	-18.14%	-18.28%	-9.25%	-10.39%
	FP-RDO	-17.09%	-17.48%	-8.44%	-10.43%	-18.17%	-19.06%	-8.28%	-9.75%
Carphone(QCIF)	FMP-RDO	-6.29%	-7.21%	-2.45%	-3.55%	-5.59%	-6.62%	-2.36%	-4.42%
	FP-RDO	-6.89%	-7.31%	-2.11%	-3.43%	-4.40%	-5.86%	-2.61%	-4.85%
Container(QCIF)	FMP-RDO	-18.06%	-17.12%	-12.30%	-13.11%	-17.63%	-17.01%	-11.04%	-10.72%
	FP-RDO	-17.23%	-16.21%	-12.41%	-13.16%	-18.20%	-17.90%	-11.89%	-11.71%
Bridge(QCIF)	FMP-RDO	-30.65%	-34.20%	-29.11%	-32.77%	-35.64%	-41.12%	-30.90%	-35.98%
	FP-RDO	-30.93%	-34.24%	-30.16%	-33.88%	-33.78%	-39.32%	-30.40%	-35.48%

* Rate reduction while maintaining of SSIM.

** Rate reduction while maintaining weighted SSIM.

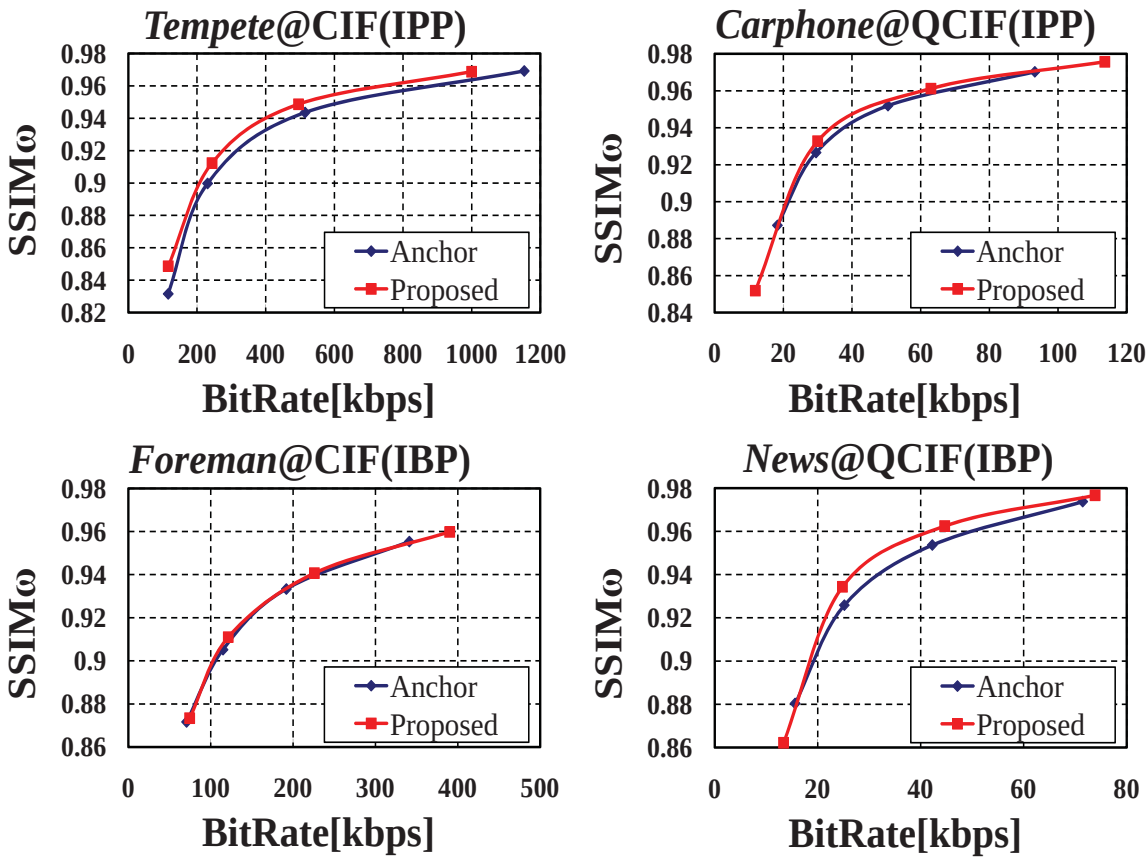


Figure 6.14: Performance comparisons in terms of the weighted SSIM index for sequences with CABAC entropy coding method.

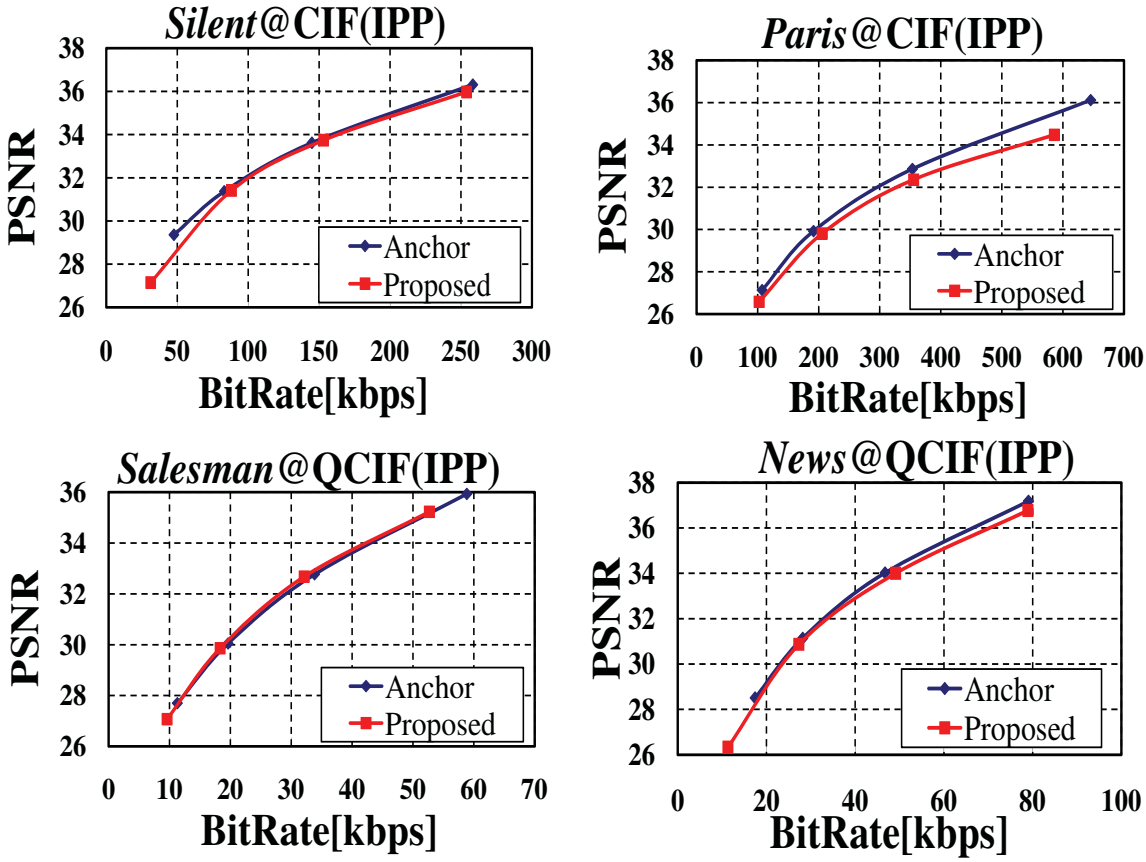


Figure 6.15: Performance comparisons in terms of PSNR for sequences with CAVLC entropy coding method.

(a)

(b)

(c)

Figure 6.16: Visual quality comparison between the conventional RDO and proposed RDO scheme, where the fortieth frame (cropped for visualization) of the *Flower* sequence is shown. (a) Original. (b) H.264/AVC coded with conventional RDO; Bit rate: 203.5 kbit/s, SSIM: 0.8710, PSNR: 25.14dB. (c) H.264/AVC coded with proposed RDO; Bit rate: 199.82 kbit/s, SSIM: 0.8805, PSNR: 24.57dB.

(a)

(b)

(c)

Figure 6.17: Visual quality comparison between the FP-RDO and FMP-RDO scheme, where the thirty fifth frame (cropped for visualization) of the *Paris* sequence is shown. (a) Original. (b) H.264/AVC coded with FP-RDO; Bit rate: 101.5 kbit/s, SSIM: 0.8667, PSNR: 26.69dB. (c) H.264/AVC coded with FMP-RDO; Bit rate: 102.5 kbit/s, SSIM: 0.8690, PSNR: 26.91dB.

Fig. 6.17 exhibits the visual performance of the FP-RDO and the FMP-RDO in the low bit rate video coding environment. The bit rate of FMP-RDO is 102.5 kbit/s while that of FP-RDO is 101.5 kbit/s. For FMP-RDO, the moving objects are allocated more bits, such as the face of the man; while the background MBs are allocated less bits. Therefore, the quality of the moving regions which attract more attention in the whole frame is improved.

Table 6.6: SSIM Indices and Bit Rates of Testing Sequences Used in the Subjective Test

Sequences		Conventional RDO		Proposed RDO	
		SSIM	Bit rate	SSIM	Bit rate
1	*Bus*	0.996	6032.68 kbit/s	0.9955	5807.44 kbit/s
2	*Hall*	0.9899	4976.36 kbit/s	0.99	4745.04 kbit/s
3	*Container*	0.9745	994.04 kbit/s	0.9754	883.72 kbit/s
4	*Tempete*	0.9726	1248.4 kbit/s	0.9707	1044.72 kbit/s
5	*Akiyo*	0.9711	97.81 kbit/s	0.9722	75.68 kbit/s
6	*Silent*	0.9655	457.68 kbit/s	0.9669	423.02 kbit/s
7	*Mobile*	0.9577	728.87 kbit/s	0.9572	703.34 kbit/s
8	*Stefan*	0.8956	179.42 kbit/s	0.8973	174.33 kbit/s

To further validate our scheme, we carried out a subjective quality evaluation test based on a two-alternative-forced-choice (2AFC) process that is widely used in psychophysical studies, where in each trial, a subject is shown a pair of video sequences and is asked (forced) to choose the one he/she thinks to have better quality. In our experiment, we selected eight pairs of sequences of CIF format that were coded by the conventional and the proposed RDO schemes to achieve the same SSIM levels (where the proposed scheme uses much lower bit rates). Table V lists all the test sequences as well as their SSIM values and bit rates. In the 2AFC test, each pair is repeated six times with random order. As a result, we obtained 48 2AFC results for each subject. Ten subjects participated in this experiment.

The subjective test results are reported in Figs. 6.18 and 6.19, which show the percentage ϖ by which the subjects are in favor of the conventional RDO against the proposed RDO schemes. As can be observed in the figures, the overall percentage (the rightmost bar

in the figures) is very close to 50% (52.5%), meaning that there is no significant perceptual difference of visual quality between the video sequences coded by the two schemes (though the proposed scheme uses much lower bit rates). In the figures, we also plot the variations of the percentage over the ten subjects and over the eight sequences, together with the error bars ($\pm$ one standard deviation between the measurements). Error bars of right most data points are calculated based on standard deviation of average values. It turns out that for almost all cases the value of ϖ is close to 50% and all error bars cross the 50% line, showing the robustness of the measurement. These results provide useful evidence that the proposed method achieves the same level of quality with lower bit rates.

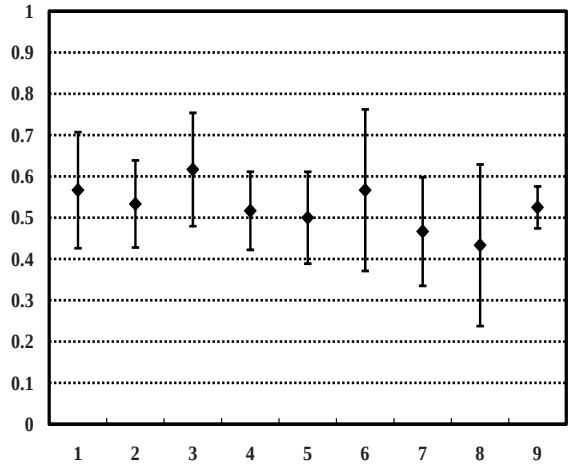


Figure 6.18: Error-bar plot with in units of ϖ and standard deviation for each test sequence (1$\sim$8: sequence number, 9: average).

Table 6.7 reports the computation overhead of the proposed scheme with both CABAC and CAVLC entropy coding methods, where ΔT is calculated according to (7.40). The coding time is obtained by encoding 100 frames of IPPP GOP structure with Intel 2.83 GHz Core processor and 4GB random access memory. On average the computation overhead is 6.3% for our scheme. As already indicated in [63] that the computation of SSIM index in the mode selection process causes about 5% overhead. Therefore, in our method the computation overhead is mainly due to the calculation of the SSIM index for each mode. We also observe that the overhead is stable for different video sequences.

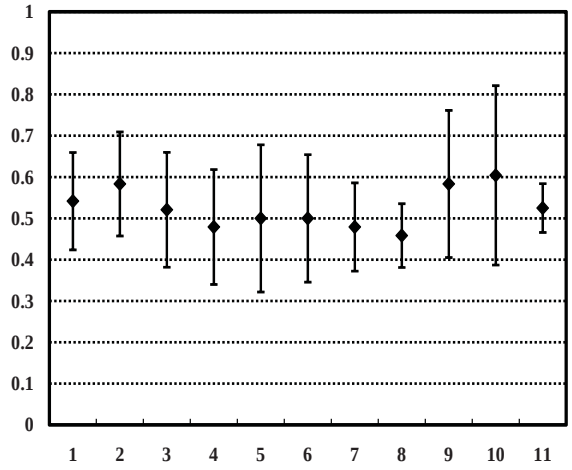


Figure 6.19: Error-bar plot with in units of ϖ and standard deviation for each subject (1∼10: subject number, 11: average).

Table 6.7: Encoding Complexity Overhead of the Proposed Scheme

Sequences	ΔT with CABAC	ΔT with CAVLC
Akiyo(QCIF)	5.21%	5.72%
News(QCIF)	5.18%	5.60%
Mobile(QCIF)	5.82%	6.14%
Silent(CIF)	7.04%	7.46%
Foreman(CIF)	6.79%	7.03%
Tempete(CIF)	7.04%	7.13%
Average	6.18%	6.51%

Table 6.8: Performance comparison of Using Different Previous Frames for Parameter Estimation

Sequences			three previous frames		five previous frames		seven previous frames	
			$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR
Akiyo(QCIF)	IPP	CABAC	0.0116	-17.85%	0.0115	-16.91%	0.0116	-18.57%
		CAVLC	0.0123	-19.33%	0.0120	-17.64%	0.0118	-16.80%
	IBP	CABAC	0.0075	-5.77%	0.0078	-6.83%	0.0069	-5.10%
		CAVLC	0.0091	-9.64%	0.0085	-8.41%	0.0090	-9.26%
Highway(QCIF)	IPP	CABAC	0.0108	-21.00%	0.0103	-20.51%	0.0102	20.33%
		CAVLC	0.0109	-21.78%	0.0107	-20.41%	0.0105	-19.70%
	IBP	CABAC	0.0043	-7.80%	0.0045	-8.13%	0.0045	-8.24%
		CAVLC	0.0046	-10.91%	0.0048	-11.72%	0.0045	-10.10%
Mobile(CIF)	IPP	CABAC	0.0047	-8.52%	0.0051	-9.22%	0.0045	-8.01%
		CAVLC	0.0051	-9.52%	0.0047	-8.41%	0.0053	-10.09%
	IBP	CABAC	0.0017	-3.23%	0.0015	-2.81%	0.0015	-3.03%
		CAVLC	0.0009	-1.77%	0.0010	-1.89%	0.0010	-2.01%
Flower(CIF)	IPP	CABAC	0.0076	-14.19%	0.0074	-13.87%	0.0075	-13.90%
		CAVLC	0.0070	-13.31%	0.0068	-12.88%	0.0072	-14.60%
	IBP	CABAC	0.0035	-6.92%	0.0032	-5.74%	0.0033	-6.04%
		CAVLC	0.0021	-4.01%	0.0022	-4.58%	0.0023	-4.60%

Table 6.8 lists the experimental results of using three, five and seven previous frames to estimate the parameters in Section 6.5, respectively. Both IPP and IBP GOP structures are tested and both CAVLC and CABAC entropy coding algorithms are employed. As indicated in Table 6.8, the final performance is not sensitive to the number of pervious frames used in the estimation. This can be explained by the stable properties of video sequences during a short period of time, as shown in Figs. 6.9 and 6.10. This suggests us to use three previous frames, as they are enough to capture the properties of the video sequences and to obtain an accurate estimation of the required parameters.

6.6.3 Comparisons With State-of-the-Art RDO Algorithms

In this experiment, the proposed scheme is compared with state-of-the-art RDO algorithms, including Huang *et al.*'s SSIM-based RDO algorithm [63], Yang *et al.*'s SSIM-based RDO

Table 6.9: Performance comparison with the State of the Art RDO Coding Algorithms for IPP GOP Structure (Anchor: Conventional RDO Technique)

Sequences		Proposed		Huang *et al.*'s		Yang *et al.*'s		CALM		RDOQ	
		$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR
Akiyo(CIF)	QP_1	**0.0026**	**-26.11%**	0.0020	-19.40%	0.0004	-4.28%	0	0.46%	0.0001	-1.08%
	QP_2	**0.0078**	**-28.06%**	0.0056	-15.78%	0.0024	-13.60%	0	0.25%	0	0.11%
Bus(CIF)	QP_1	**0.0016**	**-7.77%**	0.0011	-5.95%	0.0015	-7.12%	0	-0.04%	0.0006	-2.20%
	QP_2	**0.0099**	**-14.87%**	0.0086	-13.25%	0.0038	-6.03%	0	-0.07%	0.0007	-1.36%
Coastguard(CIF)	QP_1	**0.0013**	**-4.77%**	0.0004	-2.28%	0.0005	-2.16%	0	-0.06%	0.0006	-1.54%
	QP_2	**0.0076**	**-8.91%**	0.0038	-5.04%	0.0036	-3.97%	-0.0002	0.3%	0.0005	-0.80%
Silent(CIF)	QP_1	**0.0026**	**-9.64%**	0.0013	-5.28%	-0.0002	0.04%	0	-0.14%	0.0012	-4.15%
	QP_2	**0.0091**	**-12.43%**	0.0046	-6.83%	-0.0008	0.58%	0	-0.05%	0	-0.08%
Hall(CIF)	QP_1	0.0034	-25.89%	**0.0035**	**-26.41%**	0.0013	-10.01%	0	0.27%	0.0005	-3.78%
	QP_2	**0.0062**	**-25.46%**	0.0059	-22.84%	0.0003	-1.51%	0	0.11%	0.0002	-2.80%
Mother_Dau(CIF)	QP_1	**0.0008**	**-6.43%**	0.0004	-2.76%	0	0.56%	0	0.03%	0.0003	-1.49%
	QP_2	**0.0049**	**-8.94%**	0.0022	-4.69%	0.0015	-2.84%	0	-0.3%	0	-0.19%
Spincalendar(720P)	QP_1	0.0028	-11.89%	**0.0030**	**-12.78%**	0.0021	-8.29%	0	0.02%	0.0022	-9.13%
	QP_2	**0.0042**	**-15.57%**	0.0040	-12.81%	0.0006	-2.16%	0	-0.43%	0.0011	-2.50%
Night(720P)	QP_1	**0.0019**	**-6.65%**	0.0011	-3.45%	-0.0002	0.85%	0	0.14%	0.0009	-4.70%
	QP_2	**0.0062**	**-16.02%**	0.0029	-11.38%	0.0002	-0.96%	0	0.09%	0.0010	-2.05%
Average	QP_1	**0.0021**	**-12.39%**	0.0016	-9.79%	0.0007	-3.8%	0	0.09%	0.0008	-3.51%
	QP_2	**0.0070**	**-16.28%**	0.0047	-11.58%	0.0015	-3.81%	0	-0.01%	0.0004	-1.21%

algorithm [188], the context adaptive Lagrange multiplier (CALM) selection scheme [200] and the rate distortion optimized quantization (RDOQ) scheme [68]. For this experiment, both IPP and IBP GOP structures are employed and CAVLC entropy coding method is used. We use two different sets of QP values in the experiments: $QP_1 = \{16, 20, 24, 28\}$ and $QP_2 = \{24, 28, 32, 36\}$, where QP_1 indicates a high bit rate coding configuration. For each scheme, the improvement of the SSIM index as well as the rate reduction compared to the conventional RDO coding schemes are tabulated in Table 6.9 and 6.10.

From Tables 6.9 and 6.10, we can be observe that over a wide range of bit rates, for most of the cases our scheme achieves better performance than state-of-the-art SSIM-based RDO methods. Specifically, when compared to Huang *et al.*'s method, on average the proposed scheme achieves better rate reduction of 12.39% vs 9.79% for QP_1 and 16.28% vs 11.58% for QP_2 while maintaining the same SSIM values for IPP GOP structure. For IBP GOP structure, the performance gain is 6.66% vs 4.85% for QP_1 and 7.74% vs 3.85% for QP_2. We believe that there are three main factors that are responsible for the performance improvement. Firstly, the proposed scheme uses more accurate statistical SSIM and rate models which are derived from the inherent properties of SSIM index and the video se-

Table 6.10: Performance comparison with the State of the Art RDO Coding Algorithms for IBP GOP Structure (Anchor: Conventional RDO Technique)

Sequences		Proposed		Huang *et al.*'s		Yang *et al.*'s		CALM		RDOQ	
		$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR	$\Delta SSIM$	ΔR
Akiyo(CIF)	QP_1	**0.0014**	**-17.39%**	0.0007	-9.72%	0.0003	-5.01%	0	-0.49%	0	-0.19%
	QP_2	**0.0030**	**-8.56%**	0.0022	-6.41%	0.0015	-4.60%	0	0.32%	-0.0005	2.01%
Bus(CIF)	QP_1	0.0004	-2.04%	**0.0006**	**-3.95%**	0.0003	-1.12%	0	0.15%	0.0002	-1.20%
	QP_2	**0.0048**	**-7.58%**	0.0036	-5.25%	0.0038	-6.05%	0	0.12%	0.0021	-3.36%
Coastguard(CIF)	QP_1	**0.0007**	**-3.41%**	0.0003	-1.96%	0.0005	-2.59%	0	0.46%	0.0006	-2.86%
	QP_2	**0.0027**	**-3.31%**	0.0011	-2.04%	0.0009	-1.67%	0	0.25%	0.0014	-1.89%
Silent(CIF)	QP_1	**0.0014**	**-4.64%**	0.0013	-4.28%	0	-0.03%	0	0.06%	0.0006	-2.75%
	QP_2	**0.0050**	**-6.76%**	0.0036	-4.60%	0.0018	-2.11%	0	0%	0.0012	-1.73%
Hall(CIF)	QP_1	**0.0009**	**-7.60%**	0.0003	-2.41%	0.0003	-2.72%	0	0.21%	0.0003	-2.09%
	QP_2	**0.0031**	**-19.42%**	0.0007	-4.87%	0.0005	-3.27%	0	0.43%	0.0003	-2.51%
Mother_Dau(CIF)	QP_1	**0.0009**	**-7.43%**	0.0006	-5.80%	0.0001	-1.23%	0	-0.59%	0.0003	-2.28%
	QP_2	**0.0041**	**-5.94%**	0.0007	-1.69%	0.0015	-2.91%	0.0001	-0.16%	0.0003	-0.51%
Spincalendar(720P)	QP_1	0.0006	-5.79%	**0.0010**	**-7.18%**	0.0004	-4.10%	0	0.15%	0.0005	-5.60%
	QP_2	**0.0037**	**-4.59%**	0.0021	-3.81%	0.0009	-1.16%	0	-0.53%	0.0013	-2.57%
Night(720P)	QP_1	**0.0013**	**-4.94%**	0.0010	-3.51%	0.0002	-0.91%	0	-0.15%	0.0007	-3.61%
	QP_2	**0.0019**	**-5.73%**	0.0006	-2.11%	0.0004	-1.96%	0	-0.23%	0.0016	-3.33%
Average	QP_1	**0.0010**	**-6.66%**	0.0007	-4.85%	0.0003	-2.21%	0	0.03%	0.0004	-2.57%
	QP_2	**0.0035**	**-7.74%**	0.0018	-3.85%	0.0014	-2.17%	0	0.03%	0.0010	-1.74%

quences. Secondly, in this scheme, the Lagrange multiplier is derived adaptively for each frame. Finally, in the mode selection process, the surrounding pixels are employed to accurately obtain the SSIM index for each mode. The performances of the MSE based RDO coding schemes are also given in Table 6.9 and 6.10. Since their optimization objective is MSE rather than SSIM, there is no significant change of SSIM values in these schemes. Enlarged R-D curves that cover both low and high bit rates are shown in Fig. 6.20 and 6.21. We can observe that the proposed method achieves better performance than the other methods under comparison.

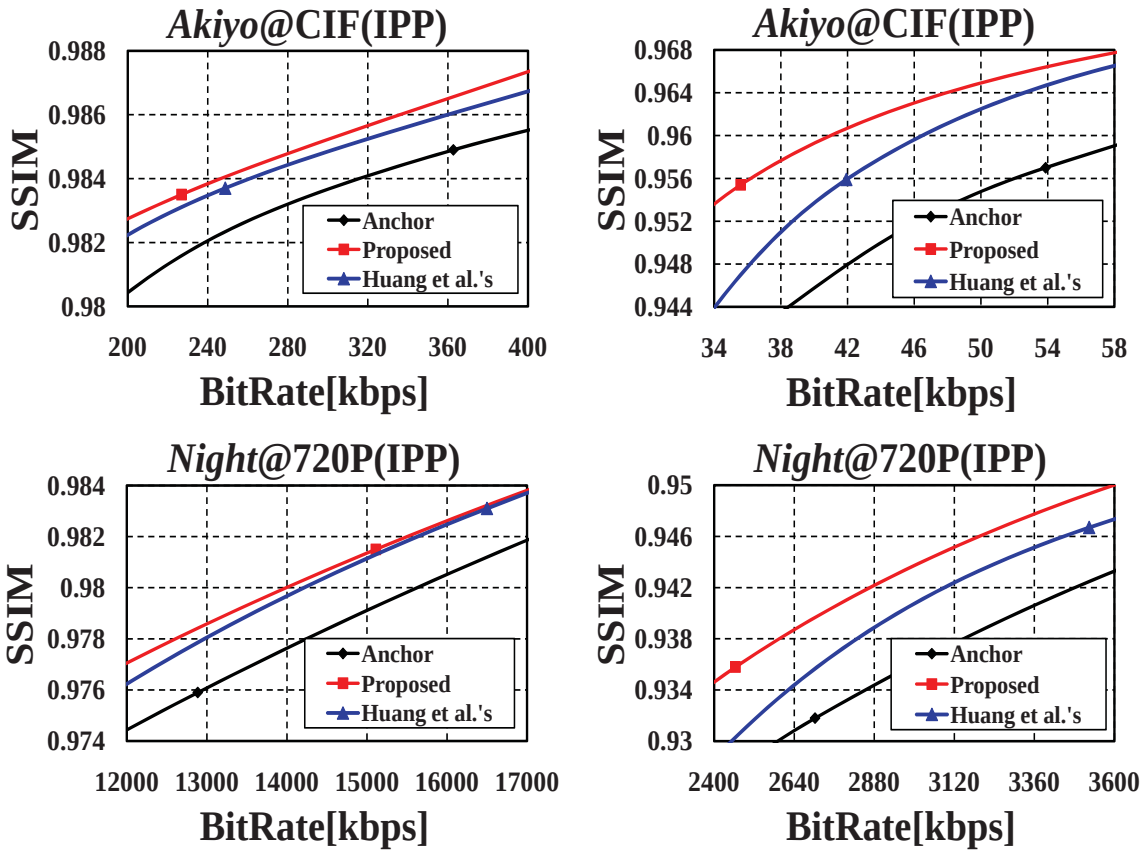


Figure 6.20: Enlarged R-D curves at both low and high bit rates for different RDO schemes (IPP GOP structure).

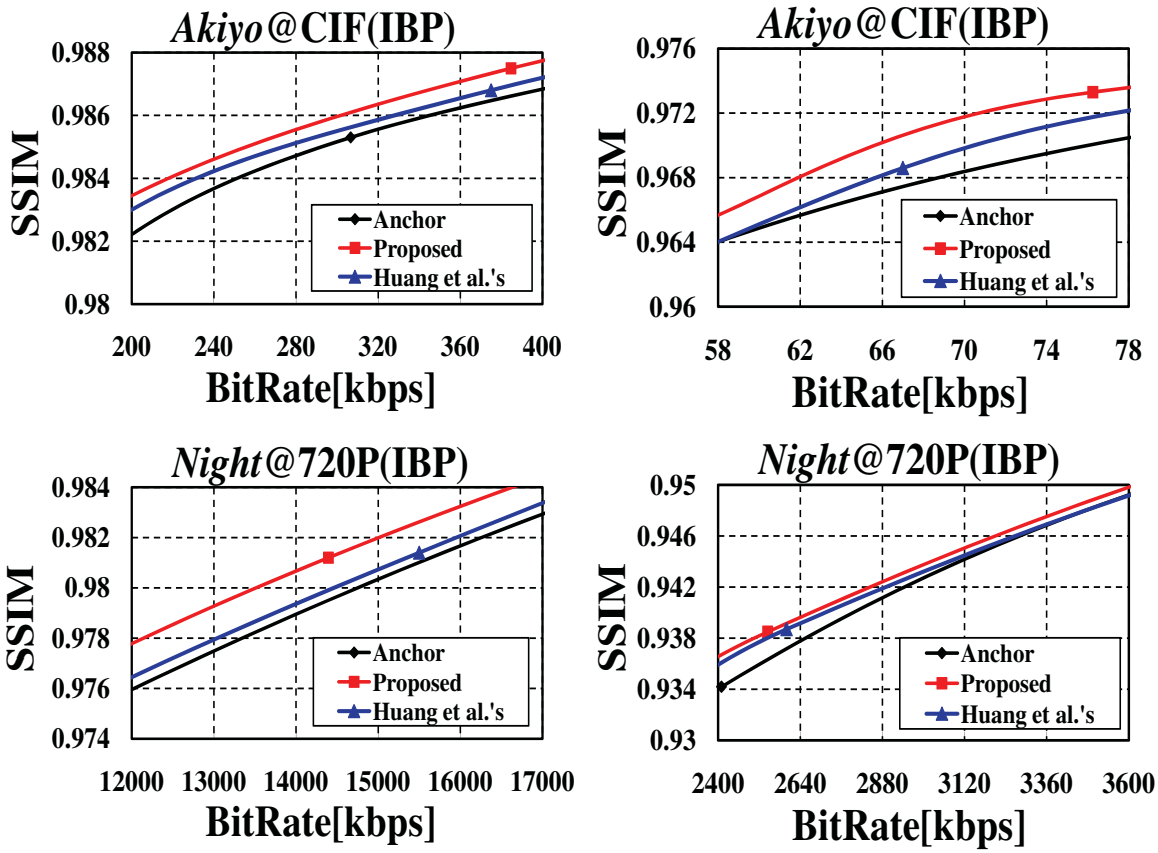


Figure 6.21: Enlarged R-D curves at both low and high bit rates for different RDO schemes (IBP GOP structure).

Chapter 7

Residual Divisive Normalization Based Perceptual Video Coding

In this chapter, we propose a perceptual video coding framework based on the divisive normalization scheme[1], which was found to be an effective approach to model the perceptual sensitivity of biological vision, but has not been fully exploited in the context of video coding. At the macroblock (MB) level, we derive the normalization factors based on the structural similarity (SSIM) index as an attempt to transform the DCT domain frame residuals to a perceptually uniform space. We further develop an MB level perceptual mode selection scheme and a frame level global quantization matrix optimization method. Extensive simulations and subjective tests verify that, compared with the H.264/AVC video coding standard and HEVC test Model (HM), the proposed method can achieve significant gain in terms of rate-SSIM performance and provides better visual quality.

A block-based Adaptive Quantization (AQ) algorithm is also proposed based on SSIM-inspired divisive normalization. The AQ scheme is implemented at the encoder so that decoding can be performed using a standard decoder. Simulation results show that the AQ scheme shows similar performance to the SSIM-Inspired divisive normalization method.

[1]proposed in collaboration with S. Wang, a visiting Ph.D. student from Peking University, Beijing.

7.1 Introduction

The main objective of video coding is to minimize the perceptual distortion D of the reconstructed video with the number of used bits R subjected to a constraint R_c. This can be expressed as

$$\min\{D\} \quad \text{subject to} \quad R \leq R_c. \tag{7.1}$$

Central to such an optimization problem is the way in which the distortion D is defined because the quality of video can only be as good as it is optimized for. Since the ultimate receiver of video is the Human Visual System (HVS), the correct optimization goal should be perceptual quality. However, existing video coding techniques typically use the sum of absolute difference (SAD) or sum of square difference (SSD) as the model for distortion, which have been widely criticized in the literature for the lack of correspondence with perceptual quality [54, 163, 166]. For many years, there have been numerous efforts in developing subjective-equivalent quality models in an attempt to generate quality scores close to the opinions of human viewers. The more accurate the model is, the more distortion can be allowed without generating perceivable artifact, and the better compression can be achieved.

In this work, we aim to transform the optimization process in Equation (7.1) into a perceptually uniform domain by incorporating the divisive normalization framework. It has already been shown that the main difference between SSIM and MSE lies in the locally adaptive divisive normalization process [10]. In general, divisive normalization transform is recognized as a perceptually and statistically motivated non-linear image representation [81, 154]. It is shown to be a useful framework that accounts for the masking effect in the HVS, which refers to the reduction of the visibility of an image component in the presence of neighboring components [51, 179]. It has also been found to be powerful in modeling the neuronal responses in the human perceptual systems [59, 123, 135]. Divisive normalization has been successfully applied in image quality assessment [76, 117], image coding [90], video coding [118, 158] and image de-noising [81, 109].

7.2 SSIM-Inspired Divisive Normalization

Block motion compensated inter-prediction technique plays an important role in existing hybrid video codecs. In this work, we follow this framework, where previously coded frames are used to predict the current frame and only residuals after prediction are coded.

7.2.1 Divisive Normalization Scheme

Assume $C(k)$ to be the k^{th} DCT transform coefficient of a residual block, then the normalized coefficient is computed as $C(k)' = C(k)/f(k)$, where $f(k)$ is a positive normalization factor for the k^{th} subband that will be discussed later.

The quantization process of the normalized residuals for a given predefined quantization step Q_s can be formulated as

$$\begin{aligned} Q(k) &= \text{sign}\{C(k)'\}\text{round}\{\frac{|C(k)'|}{Q_s} + p\} \\ &= \text{sign}\{C(k)\}\text{round}\{\frac{|C(k)|}{Q_s \cdot f(k)} + p\} \end{aligned} \tag{7.2}$$

where p is the rounding offset in the quantization.

At the decoder, the de-quantization and reconstruction of $C(k)$ is performed as

$$\begin{aligned} R(k) &= R(k)' \cdot f(k) = Q(k) \cdot Q_s \cdot f(k) \\ &= \text{sign}\{C(k)\}\text{round}\{\frac{|C(k)|}{Q_s \cdot f(k)} + p\} \cdot Q_s \cdot f(k) \end{aligned} \tag{7.3}$$

The purpose of the divisive normalization process is to convert the transform residuals into a perceptually uniform space. Thus the factor $f(k)$ determines the perceptual importance of each of the corresponding transform coefficient. The proposed divisive normalization scheme can be interpreted in two ways. An adaptive normalization factor is applied, followed by quantization with a predefined fixed step Q_s. Alternatively, an adaptive quantization matrix is defined for each MB and thus each coefficient is quantized with a different quantization step.

$$\begin{aligned}\mathrm{SSIM}(\mathbf{x},\mathbf{y}) =&\{1-\frac{((C(0)+P(0))-(R(0)+P(0)))^2}{X(0)^2+Y(0)^2+N\cdot C_1}\}\times\{1-\frac{\frac{\sum_{k=1}^{N-1}((C(k)+P(k))-(R(k)+P(k)))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2}\}\\ =&\{1-\frac{(C(0)-R(0))^2}{X(0)^2+Y(0)^2+N\cdot C_1}\}\times\{1-\frac{\frac{\sum_{k=1}^{N-1}(C(k)-R(k))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2}\}\end{aligned} \tag{7.4}$$

$$\begin{aligned}\mathrm{SSIM}(\mathbf{x},\mathbf{y}) =& \{1-\frac{(C(0)'\cdot f_{dc}-R(0)'\cdot f_{dc})^2}{X(0)^2+Y(0)^2+N\cdot C_1}\}\times\{1-\frac{\frac{\sum_{k=1}^{N-1}(C(k)'\cdot f_{ac}-R(k)'\cdot f_{ac})^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2}\}\\ \approx& \{1-\frac{(C(0)'-R(0)')^2}{\mathbb{E}(\sqrt{X(0)^2+Y(0)^2+N\cdot C_1})^2}\}\times\{1-\frac{\frac{\sum_{k=1}^{N-1}(C(k)'-R(k)')^2}{N-1}}{\mathbb{E}(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2})^2}\}\end{aligned} \tag{7.5}$$

In the context of computational neuro-science as well as still image processing and coding, several different approaches have been used to derive the normalization factor, which may be defined as the sum of the squared neighboring coefficients plus a constant [90], or derived from a local statistical image model [154]. In this work, our objective is to optimize the SSIM index, therefore, we employ a model based on the DCT domain SSIM index.

The DCT domain SSIM index was first presented in [25]:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \left(1-\frac{(X(0)-Y(0))^2}{X(0)^2+Y(0)^2+N\cdot C_1}\right)\times\left(1-\frac{\frac{\sum_{k=1}^{N-1}(X(k)-Y(k))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2}\right), \tag{7.6}$$

where $X(k)$ and $Y(k)$ represent the DCT coefficients of the input signals $\mathbf{x}$ and $\mathbf{y}$, respectively. C_1 and C_2 are used to avoid instability when the means and variances are close to zero and N denotes the block size. The DCT domain SSIM index is composed of the

product of two terms, which are the normalized squared errors of DC and AC coefficients, respectively. Moreover, the normalization is conceptually consistent with the light adaptation (also called luminance masking) and contrast masking effect of the HVS [52, 135, 179]. Equation (7.6) can be re-written as

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \left(1 - \left(\frac{X(0)}{\sqrt{\eta_{dc}}} - \frac{Y(0)}{\sqrt{\eta_{dc}}}\right)^2\right) \times \left(1 - \frac{1}{N-1}\sum_{k=1}^{N-1}\left(\frac{X(k)}{\sqrt{\eta_{ac}}} - \frac{Y(k)}{\sqrt{\eta_{ac}}}\right)^2\right), \quad (7.7)$$

where

$$\eta_{dc} = X(0)^2 + Y(0)^2 + N \cdot C_1, \quad (7.8)$$

$$\eta_{ac} = \frac{\sum_{k=1}^{N-1}(X(k)^2 + Y(k)^2)}{N-1} + C_2. \quad (7.9)$$

Equation (7.7) suggests that the DCT domain SSIM index can be computed from normalized MSE of DC and AC coefficients. This inspires us to use SSIM-based divisive normalization for perceptual video coding.

In the video coding scenario, let $P(k)$ be the prediction signal of the k^{th} subband in DCT domain, then the SSIM index can be rewritten as in (7.4).

Since the local statistics do not change significantly within each MB, we divide each MB into l sub-MBs for DCT transform and $X_i(k)$ denotes the k^{th} DCT coefficient in the i^{th} sub-MB. As the SSIM index differentiates between the DC and AC coefficients, we use separate normalization factors for AC and DC coefficients, which are defined as

$$f_{dc} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{X_i(0)^2 + Y_i(0)^2 + N \cdot C_1}}{\mathbb{E}(\sqrt{X(0)^2 + Y(0)^2 + N \cdot C_1})}, \quad (7.10)$$

$$f_{ac} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(X_i(k)^2 + Y_i(k)^2)}{N-1} + C_2}}{\mathbb{E}(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2 + Y(k)^2)}{N-1} + C_2})}, \quad (7.11)$$

where $\mathbb{E}(\cdot)$ denotes the mathematical expectation operator. The expectations are over the whole frame, and thus do not affect the relative normalization factors across space within the same frame.

As a result of the use of f_{dc} and f_{ac}, the normalized DCT coefficients for residuals can be expressed as

$$C(k)' = \begin{cases} \frac{C(0)}{f_{dc}} & k = 0 \\ \frac{C(k)}{f_{ac}} & otherwise \end{cases} \tag{7.12}$$

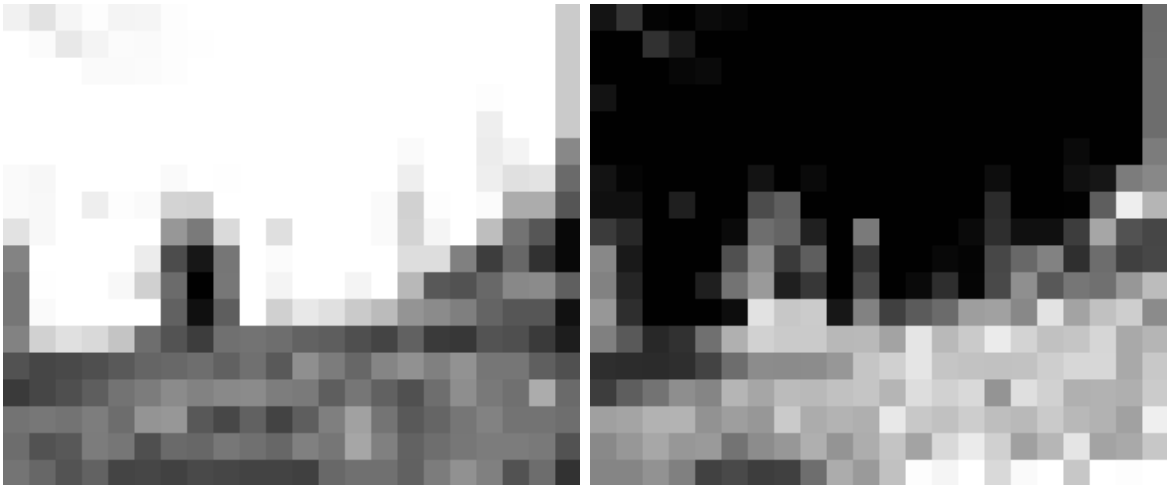
$$R(k)' = \begin{cases} \frac{R(0)}{f_{dc}} & k = 0 \\ \frac{R(k)}{f_{ac}} & otherwise \end{cases} \tag{7.13}$$

Therefore, the SSIM index in the divisive normalization framework can be expressed as in (7.5), which implies that in the divisive normalization space, the SSIM index is dependent on the difference of the normalized signals but not adaptive to the local normalized signals themselves and therefore all the MBs can be treated as perceptually identical. Since the clearly visible distortion regions will be perceptually more apparent [73], transforming all the coefficients into the perceptually uniform domain is also a convenient approach to improve the perceptual quality according to the philosophy behind distortion-based pooling scheme [168].

The divisive normalization factor is spatially adaptive and dependent on the content of the MB and determines the relative perceptual importance of each MB. The MBs which are less important are quantized more coarsely as compared to the more important MBs. The expected values of DC and AC energy are used as the reference point to determine the importance of each MB. The MBs with higher energy than the mean value have effectively larger quantization step and vice versa. By doing so, we are borrowing bits from the regions which are perceptually less important and using them for the regions with more perceptual relevance, as far as SSIM is concerned, so that all the regions in the frame conceptually have uniform perceptual distortion. It is important to note that the reference point, mean AC and DC energies, is highly dependent on the content of the video frame. The frames with significant texture regions are likely to get more perceptual improvement because the texture regions are the main beneficiaries of the spatially adaptive normalization process.

The calculation of divisive normalization factors for DC and AC coefficients are demonstrated in Fig. 7.1, where darker MBs indicate smaller normalization factors. As the flower textures can mask more distortions, we assign larger normalization factors to the AC coefficients in these regions. However, since the luminance values in these regions are relatively

(a)



(b) (c)

Figure 7.1: Visualization of spatially adaptive divisive normalization factors for the Flower sequence. (a) The original frame. (b) Normalization factors for DC coefficients for each MB. (c) Normalization factors for AC coefficients for each MB.

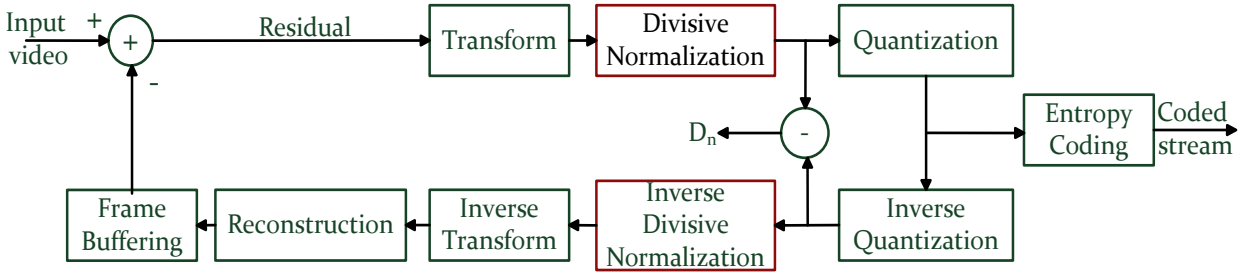


Figure 7.2: Framework of the proposed scheme.

lower, we assign smaller normalization factors to the DC coefficients. These are conceptually consistent with the light adaptation and contrast masking effects of the HVS.

7.2.2 Perceptual Rate Distortion Optimization for Mode Selection

The RDO process in video coding can be expressed by minimizing the perceived distortion D with the number of used bits R subject to a constraint R_c. This can be converted to an unconstrained optimization problem by

$$\min\{J\} \quad \text{where} \quad J = D + \lambda \cdot R, \tag{7.14}$$

where J is called the Rate Distortion (RD) cost and λ is known as the Lagrange multiplier that controls the trade-off between R and D.

Here we replace the conventional SAD and SSD with a new distortion model that is consistent with the residual normalization process. As illustrated in Fig. 7.2, for each MB, the distortion model is defined as the SSD between the normalized DCT coefficients, which is expressed as

$$\begin{aligned} D &= \sum_{i=1}^{l} \sum_{k=0}^{N-1} (C_i(k)' - R_i(k)')^2 \\ &= \sum_{i=1}^{l} \frac{(X_i(0) - Y_i(0))^2}{f_{dc_i}^2} + \frac{\sum_{k=1}^{N-1} (X_i(k) - Y_i(k))^2}{f_{ac_i}^2} \end{aligned} \tag{7.15}$$

Based on (7.14), the RDO problem is given by

$$\min\{J\} \quad \text{where} \quad J = \sum_{i=1}^{l} \sum_{k=0}^{N-1} (C_i(k)' - R_i(k)')^2 + \lambda_{\mathcal{H}} \cdot R, \tag{7.16}$$

where $\lambda_{\mathcal{H}}$ indicates the Lagrange multiplier defined in H.264/AVC or HEVC coding standard with the predefined quantization step Q_s.

From the residual normalization point of view, the distortion model calculates the SSD between the normalized original and distorted DCT coefficients, as shown in Fig. 7.2. Therefore, we can still use the respective Lagrange multiplier defined in H.264 or HEVC, $\lambda_{\mathcal{H}}$, in this perceptual RDO scheme.

7.2.3 Sub-band Level Normalization Factor Computation

In this sub-section, we show that the proposed method in section 7.2.1 can be improved further by fine tuning the DCT normalization matrix so that each AC coefficient has a different normalization factor. Motivated by the fact that the normalized DCT coefficients of residuals of different frequencies have different statistical distributions, we propose a frame level quantization matrix selection algorithm considering the perceptual quality of the reconstructed video. To begin with, we model the normalized transform coefficients x with Laplace distribution, which has been proved to achieve a good trade-off between model fidelity and complexity [79]:

$$f_{Lap}(x) = \frac{\Lambda}{2} \cdot e^{-\Lambda \cdot |x|}, \tag{7.17}$$

where Λ is called the Laplace parameter.

From (7.14), the Lagrange parameter is obtained by calculating the derivative of J with respect to R, then setting it to zero, and finally solving for λ,

$$\frac{dJ}{dR} = \frac{dD}{dR} + \lambda = 0, \tag{7.18}$$

which yields

$$\lambda = -\frac{dD}{dR} = -\frac{\frac{dD}{dQ_s}}{\frac{dR}{dQ_s}}. \tag{7.19}$$

In [79], Laplace distribution based rate and distortion models were established to derive λ for each frame dynamically. However, all the transform coefficients were modeled with a single distribution and the variation in the distribution between DCT sub-bands was ignored. Here we model the distortion and rate in a similar way as in [79], where D is obtained by summing the perceptual distortion in each quantization interval and R is calculated with the help of the entropy of the normalized coefficients. Let $c_{i,j}^m$ be the DCT coefficient in the $(i,j)^{th}$ sub-band of the m^{th} block and $\hat{c}_{i,j}^m$ the reconstructed coefficient of the same position in the decoder, the perceptual distortion for this sub-band $D_{i,j}$ is defined as

$$\begin{aligned} D_{i,j} =& \frac{1}{N_B}\sum_{m=1}^{N_B}\left(\frac{c_{i,j}^m}{f_{i,j}^m}-\frac{\hat{c}_{i,j}^m}{f_{i,j}^m}\right)^2 \\ =& \frac{1}{N_B}\sum_{m=1}^{N_B}\left(c_{i,j}^{m'}-\hat{c}_{i,j}^{m'}\right)^2 \end{aligned} \tag{7.20}$$

where N_B is the number of DCT blocks in each frame and $f_{i,j}^m$ represents the normalization factor for the $(i,j)^{th}$ sub-band of the m^{th} block; $c_{i,j}^{m'}$ and $\hat{c}_{i,j}^{m'}$ are the normalized coefficients of $c_{i,j}^m$ and $\hat{c}_{i,j}^m$, respectively.

More specifically, the perceptual distortion defined in (7.20) is equivalent to the MSE in the divisive normalization domain. If $x_{i,j}$ denotes the normalized coefficient in the $(i,j)^{th}$ sub-band, then $D_{i,j}$ can be modeled in the divisive normalization domain according to the quantization process in H.264/AVC, which is given by

$$D_{i,j} \approx \int_{-(Q_s-\gamma Q_s)}^{(Q_s-\gamma Q_s)} x_{i,j}^2 f_{Lap}(x_{i,j})dx_{i,j} + 2\sum_{n=1}^{\infty}\int_{nQ_s-\gamma Q_s}^{(n+1)Q_s-\gamma Q_s}(x_{i,j}-nQ_s)^2 f_{Lap}(x_{i,j})dx_{i,j}, \tag{7.21}$$

where γ is the rounding offset. Subsequently, we model the rate of the $(i,j)^{th}$ sub-band by calculating its entropy [203]:

$$R_{i,j} = -P_0\cdot\log_2 P_0 - 2\sum_{n=1}^{\infty}P_n\cdot\log_2 P_n, \tag{7.22}$$

where P_0 and P_n are the probabilities of the transformed residuals quantized to the zero-th and n-th quantization levels, respectively, which can be modeled by the Laplace distribution as

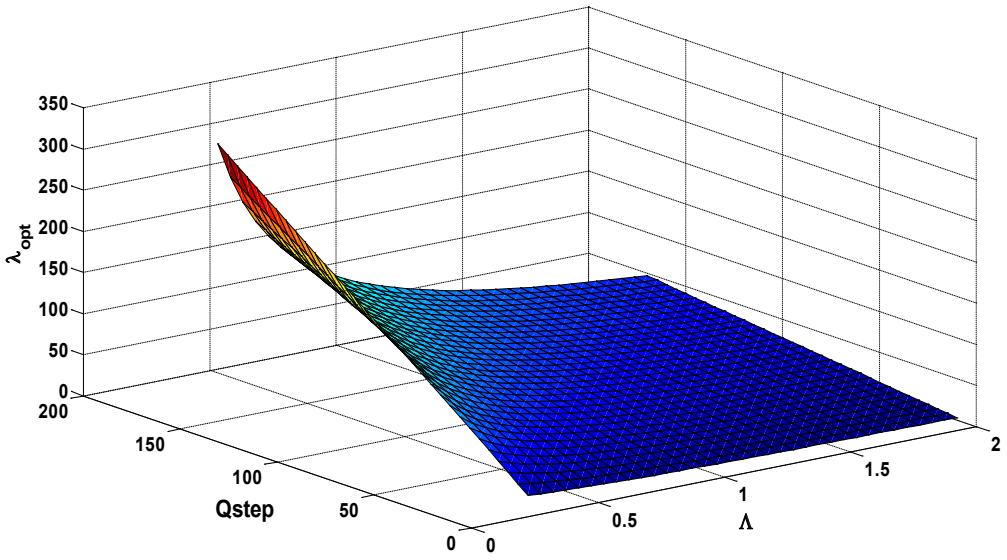


Figure 7.3: Relationship between the optimal λ and $(\Lambda, Qstep)$.

$$P_0 = \int_{-(Q_s-\gamma Q_s)}^{(Q_s-\gamma Q_s)} f_{Lap}(x_{i,j})dx, \tag{7.23}$$

$$P_n = \int_{nQ_s-\gamma Q_s}^{(n+1)Q_s-\gamma Q_s} f_{Lap}(x_{i,j})dx. \tag{7.24}$$

Since the rounding offset can be regarded as a constant value for each frame, by incorporating (7.22) into (7.19), we conclude that the optimal Lagrange multiplier which controls the trade-off between R and D is a function of the Laplace parameter and the quantization step only, which is given by

$$\lambda_{opt} = f(\Lambda, Q_s). \tag{7.25}$$

The λ_{opt} for each (Λ, Q_s) is shown in Fig. 7.3, which confirms the idea that λ_{opt} increases monotonically with Q_s but decreases monotonically with Λ. It suggests that for the same λ_{opt} but different Λ, we will have different Q_s values.

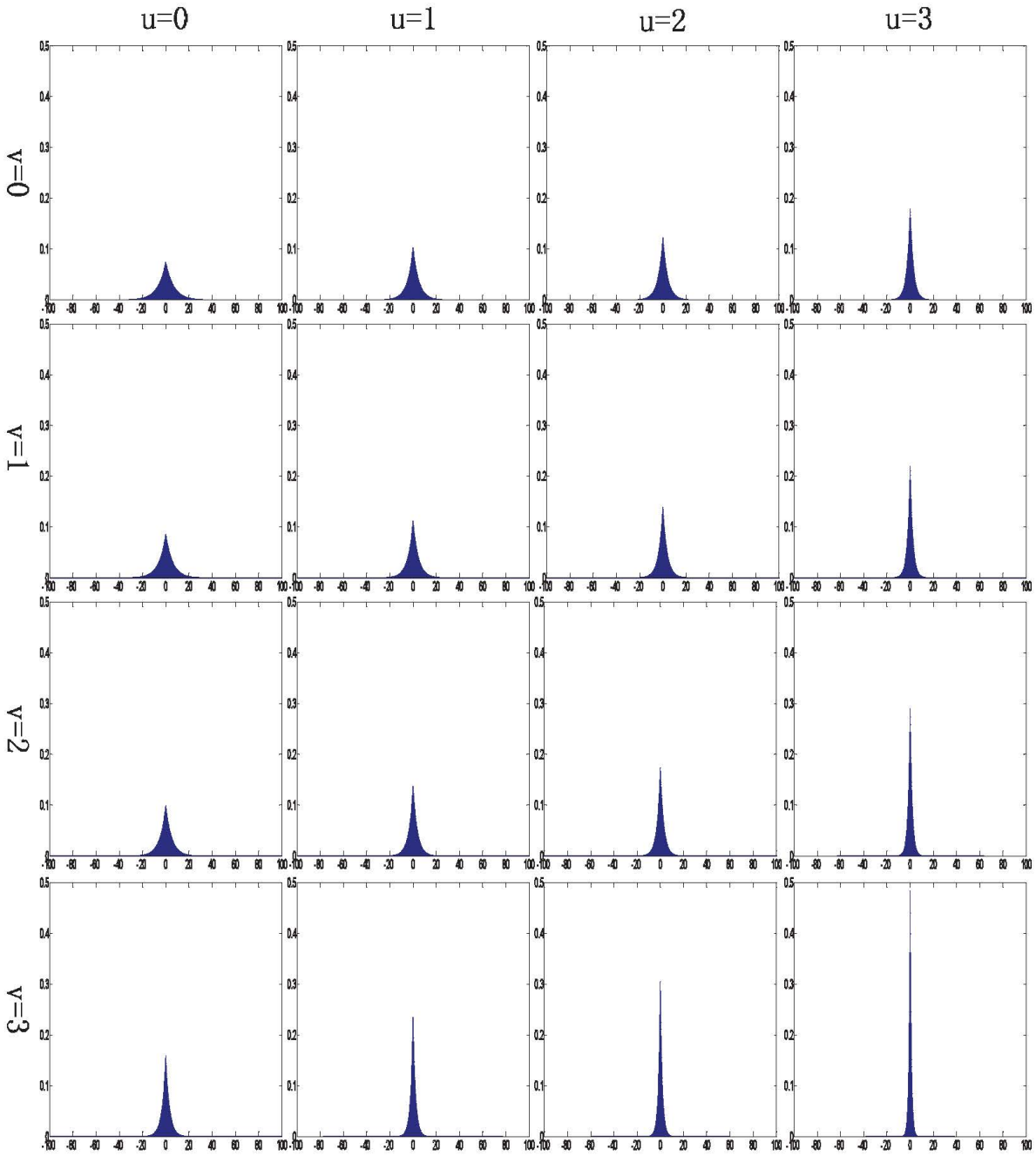
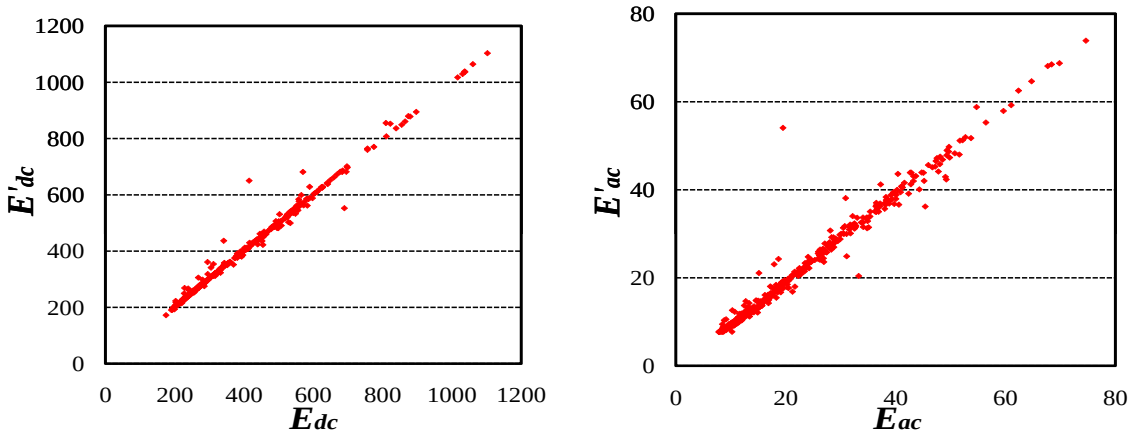


Figure 7.4: Laplace distributions for DCT subband coefficients (Bus sequence).



(a) (b)

Figure 7.5: (a) Relationship between E_{dc} and E'_{dc} at QP=30 for the Bus sequence. (b) Relationship between E_{ac} and E'_{ac} at QP=30 for the Bus sequence.

Fig. 7.4 shows that the distribution of the normalized transform coefficients in different sub-bands have similar shape but different widths [71, 203], thus their optimal Lagrange multipliers should be different. However, in the current hybrid video coding framework, directly adjusting λ_{opt} for each subband is impractical because the Lagrange multiplier needs to be uniform across the whole frame in RD optimization. To overcome this, we generate a uniform λ_{opt} for each subband by modifying Q_s values. Given the optimal λ_{opt}, the optimal quantization step for the $(i,j)^{th}$ sub-band is calculated as

$$Q_{i,j} = g(\lambda_{opt}, \Lambda_{i,j}) \tag{7.26}$$

In our implementation, we keep λ of the DC coefficients unaltered and modify Q_s of the AC coefficients. To obtain the optimal $Q_{i,j}$, we build a look-up table based on Fig. 7.3.

7.3 H.264/AVC Implementation

In video coding, the normalization factors defined in (7.10) and (7.11) need to be computed at both the encoder and the decoder. However, before coding the current frame, the distorted MBs are not available, which creates a chicken or egg causality dilemma. Moreover,

at the decoder side, the original MB is not accessible either. Therefore, the normalization factors defined in (7.10) and (7.11) cannot be directly applied in practice. To overcome this problem, we propose to make use of the predicted MB, which is available at both the encoder and the decoder for the calculation of the normalization factors. As such, we do not need to transmit any additional overhead information to the decoder.

The relationship between E_{dc} and E'_{dc} as well as E_{ac} and E'_{ac} are illustrated in Fig. 7.5, where E_{dc}, E'_{dc}, E_{ac} and E'_{ac} are defined in (7.27). In these equations, $Z_i(k)$ is the k^{th} DCT coefficient of the i^{th} prediction sub-MB for each mode. We can observe dependency between the DC and AC energy values of the original and predicted MBs. Therefore, the DC and AC energy of the original MB can be approximated with the help of the corresponding energy of the prediction MB. Consequently, the approximation of the normalization factors can be determined by

$$f'_{dc} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{2Z_i(0)^2 + N \cdot C_1}}{\mathbb{E}(\sqrt{2Z(0)^2 + N \cdot C_1})}, \tag{7.28}$$

$$f'_{ac} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(Z_i(k)^2 + s \cdot Z_i(k)^2)}{N-1} + C_2}}{\mathbb{E}(\sqrt{\frac{\sum_{k=1}^{N-1}(Z(k)^2 + s \cdot Z(k)^2)}{N-1} + C_2})}. \tag{7.29}$$

For intra mode, we use the MB at the same position in the previously coded frames.

In order to compensate for the loss of AC energy, we use a factor s to bridge the difference between the energy of AC coefficients in the prediction MB and the original MB, which can be defined as

$$E_{dc} = \frac{1}{l}\sum_{i=1}^{l}\sqrt{X_i(0)^2 + Y_i(0)^2 + N \cdot C_1} \qquad E'_{dc} = \frac{1}{l}\sum_{i=1}^{l}\sqrt{2Z_i(0)^2 + N \cdot C_1}$$

$$E_{ac} = \frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(X_i(k)^2 + Y_i(k)^2)}{N-1} + C_2} \qquad E'_{ac} = \frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(2 \cdot Z_i(k)^2)}{N-1} + C_2} \tag{7.27}$$

$$s = \frac{\mathbb{E}(\sum_{k=1}^{N-1} X(k)^2)}{\mathbb{E}(\sum_{k=1}^{N-1} Z(k)^2)}. \tag{7.30}$$

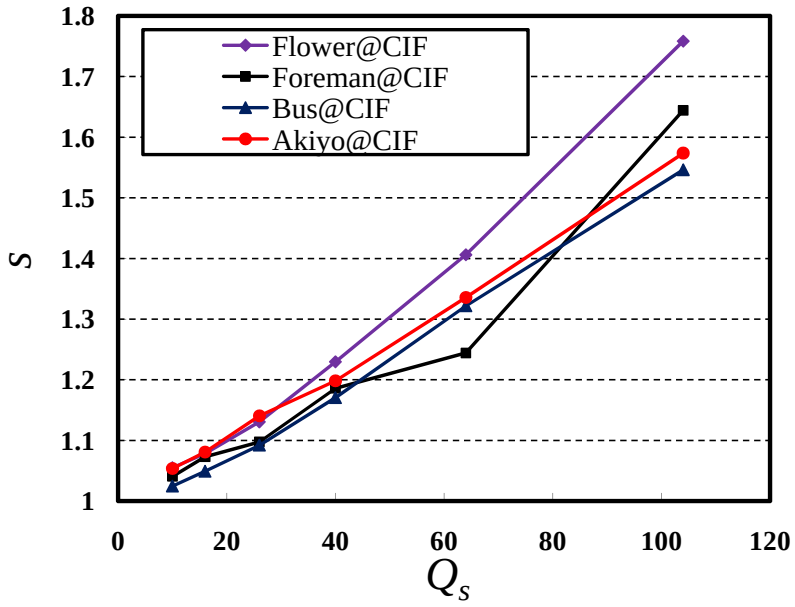


Figure 7.6: Relationship between s and Q_s for different sequences.

As depicted in Fig. 7.6, we can approximate s by a linear relationship with Q_s, which can be modeled empirically as

$$s = 1 + 0.005 \cdot Q_s. \tag{7.31}$$

In order to compute the normalization factors for DC and AC coefficients, as defined in (28) and (29), the DC and AC energy of the prediction MB should firstly be calculated. The DCT is an orthogonal transform that obeys Parseval's theorem. Thus we will have the following relations between the DCT coefficients and the spatial domain mean and variance:

$$\mu_x = \frac{\sum_{i=0}^{N-1} x(i)}{N} = \frac{X(0)}{\sqrt{N}}, \tag{7.32}$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{N-1} X(i)^2}{N-1}. \tag{7.33}$$

Therefore, to calculate the normalization factors in (28) and (29), in the actual implementation for both the encoder and decoder, it is not necessary to perform the actual DCT transform. Instead, we only need to compute the mean and variance of the prediction block in spatial domain.

In our implementation, we combine the frame-level quantization matrix selection and divisive normalization together and employ one quantization matrix to achieve two goals. Analogous to [150], the quantization matrix for 4×4 DCT transform is defined as

$$WS_{ij} = 16 \cdot \begin{bmatrix} f'_{dc} \cdot \omega_{0,0} & f'_{ac} \cdot \omega_{0,1} & f'_{ac} \cdot \omega_{0,2} & f'_{ac} \cdot \omega_{0,3} \\ f'_{ac} \cdot \omega_{1,0} & f'_{ac} \cdot \omega_{1,1} & f'_{ac} \cdot \omega_{1,2} & f'_{ac} \cdot \omega_{1,3} \\ f'_{ac} \cdot \omega_{2,0} & f'_{ac} \cdot \omega_{2,1} & f'_{ac} \cdot \omega_{2,2} & f'_{ac} \cdot \omega_{2,3} \\ f'_{ac} \cdot \omega_{3,0} & f'_{ac} \cdot \omega_{3,1} & f'_{ac} \cdot \omega_{3,2} & f'_{ac} \cdot \omega_{3,3} \end{bmatrix}, \tag{7.34}$$

where

$$\omega_{i,j} = Q_{i,j}/Q_s. \tag{7.35}$$

The Laplace parameter $\Lambda_{i,j}$ and the expectation of the energy (as indicated in (7.11)) should be available before coding the current frame. However, these quantities can only be obtained after coding it. As they are approximately constants during a very short period of time, we estimate them by averaging their corresponding values from previous frames coded in the same manner, i.e.,

$$\hat{\Lambda}^t_{i,j} = \frac{1}{N_f} \sum_{n=1}^{N_f} \Lambda^{t-n}_{i,j}, \tag{7.36}$$

where t indicates the frame number and N_f represents the number of previous frames used. Practically, N_f is set to be 3 in this work.

At the decoder, the Laplace distribution parameters of the normalized coefficients in each sub-band are not available. To address this issue, we transmit the frame-level quantization matrix to the decoder. As the statistics of frames in a short time do not change considerably, we empirically define a threshold to determine whether to refresh the quan-

tization matrix, which is expressed as

$$\omega^t = \begin{cases} \omega^{t-1} & \sum(\omega_{i,j}^t - \omega_{i,j}^{t-1})^2 < T_r \\ \omega^t & otherwise \end{cases} \tag{7.37}$$

where

$$\omega = 16 \cdot \begin{bmatrix} \omega_{0,0} & \omega_{0,1} & \omega_{0,2} & \omega_{0,3} \\ \omega_{1,0} & \omega_{1,1} & \omega_{1,2} & \omega_{1,3} \\ \omega_{2,0} & \omega_{2,1} & \omega_{2,2} & \omega_{2,3} \\ \omega_{3,0} & \omega_{3,1} & \omega_{3,2} & \omega_{3,3} \end{bmatrix}. \tag{7.38}$$

We set the threshold T_r to be 100 to balance the transmitted bits and the accuracy of the matrix. Empirically, we find this to be a non-sensitive parameter as the quantization matrix of each frame is very stable and the transmission of the matrix takes only a small number of bits.

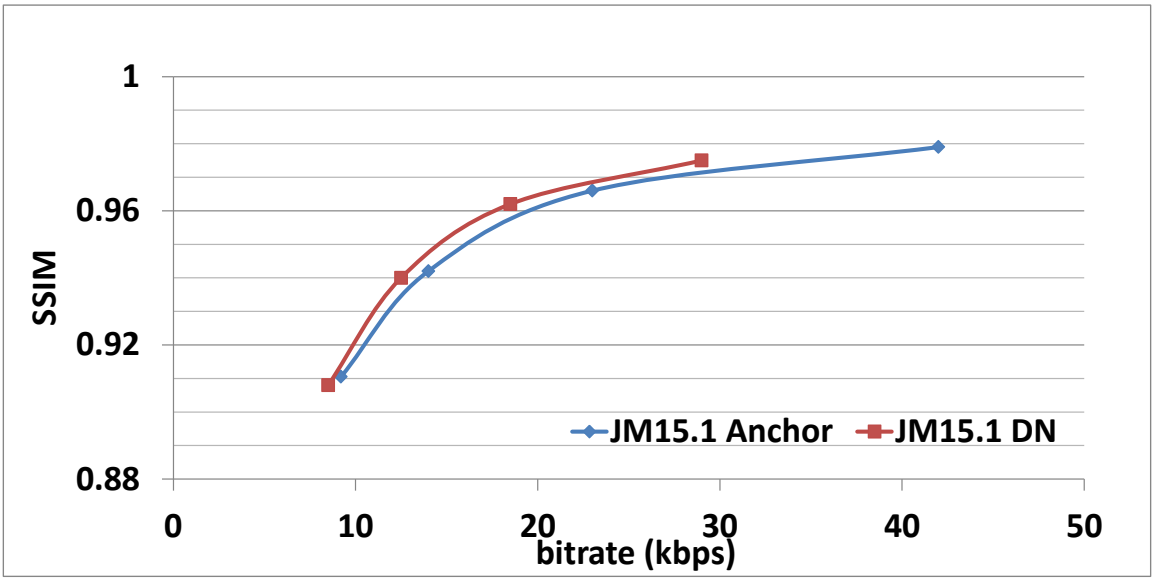
7.3.1 Objective Performance Evaluation

To validate the proposed scheme, we integrate it into H.264/AVC reference software JM15.1 [67]. All test video sequences are in YUV 4:2:0 format. The common coding configurations are set as follows: all available inter and intra modes are enabled; five reference frames; one I frame followed by all P frames; high complexity RDO and fixed quantization parameters (QP).

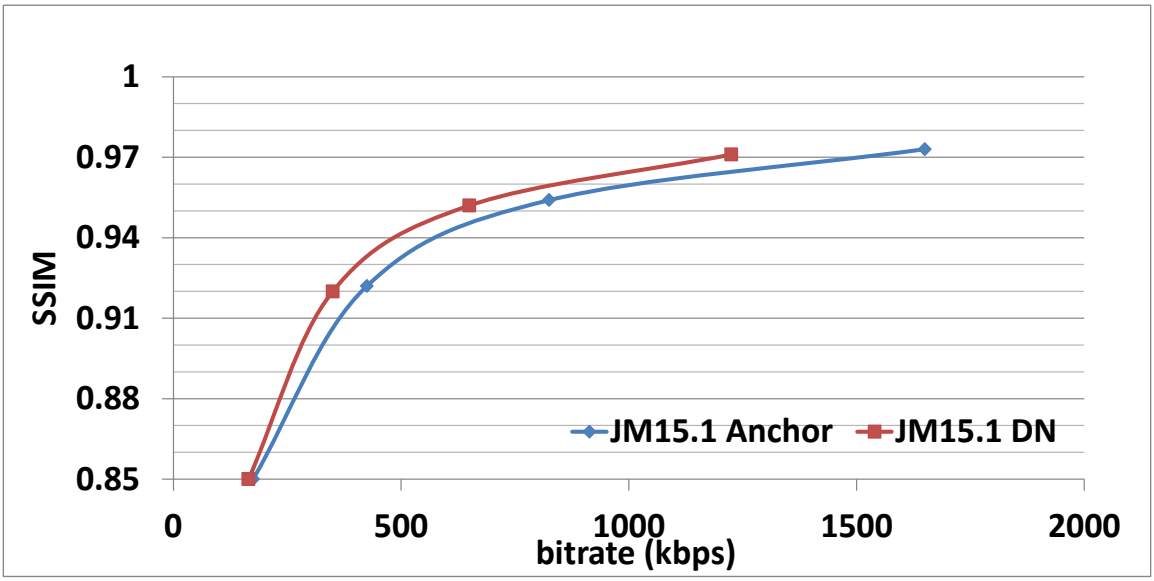
The RD performance is measured in two cases: SSIM of Y component only and SSIM of Y, Cb and Cr components, respectively. To apply SSIM to all three color components, we combine the SSIM indices of these components by [171]

$$SSIM_\omega = W_Y \cdot SSIM_Y + W_{Cb} \cdot SSIM_{Cb} + W_{Cr} \cdot SSIM_{Cr}, \tag{7.39}$$

where $W_Y = 0.8$, $W_{Cb} = 0.1$ and $W_{Cr} = 0.1$ are the weights assigned to Y, Cb and Cr components, respectively. These quantities for the whole video sequence are obtained by simply averaging the respective values of individual frames. The method proposed in [6] is used to calculate the differences between two RD curves.



(a) Akiyo



(b) Tempete

Figure 7.7: Rate-SSIM Performance comparisons (Anchor: H.264/AVC).

We use two different sets of QP values in the experiments: QP_1={22, 26, 30, 34} and QP_2={26, 30, 34, 38}, which represent high bit-rate and low bit-rate coding configurations, respectively. From Table 7.2, it can be observed that over a wide range of test sequences with resolutions from QCIF to 720p, the proposed scheme achieves average rate reduction of 15.0% for QP_1 and 16.0% for QP_2 for fixed SSIM values and the maximum coding gain is 42.5%. It can also be observed that our scheme performs better when there exist significant statistical differences between different regions in the same frame, for example, in the cases of *Bus* and *Flower*. This is likely because these frames allow us to borrow bits more aggressively from the regions with complex texture or high contrast (high normalization factor) and allocating them to the regions with relatively simple textures (low normalization factor).

The R-D performances for two of the test sequences with different resolutions are shown in Fig. 7.7. It can be observed that the proposed scheme achieves better R-D performance over the full range of QP values. Moreover, the gains become more significant at middle bit-rates. The reason may be that at high bit rate, the quantization step is small and thus the differences of quantization steps among the MBs are not significant, while at low bit rate, since the AC coefficients are severely distorted, the normalization factors derived from the prediction frame do not precisely represent the properties of the original frame.

When evaluating the coding complexity overhead, we calculate ΔT as

$$\Delta T = \frac{T_{pro} - T_{H.264}}{T_{H.264}} \times 100\%, \tag{7.40}$$

where $T_{H.264}$ and T_{pro} indicate the total coding time for the sequence with H.264/AVC and the proposed schemes, respectively. Table 7.3 shows the computational overhead for both encoding and decoding. The coding time is obtained by encoding 100 frames of IPPP GOP structure with Intel 2.83 GHz Core processor and 4GB random access memory. As indicated in Section 7.3, we do not need to perform DCT transform at either the encoder or the decoder. Therefore, it is observed that the encoding overhead is negligible (1.16% on average). The complexity of the decoder is increased by 8.48% on average.

To show the advantage of our divisive normalization scheme, the performance comparisons of the proposed scheme, the state of the art SSIM based RDO scheme [159]

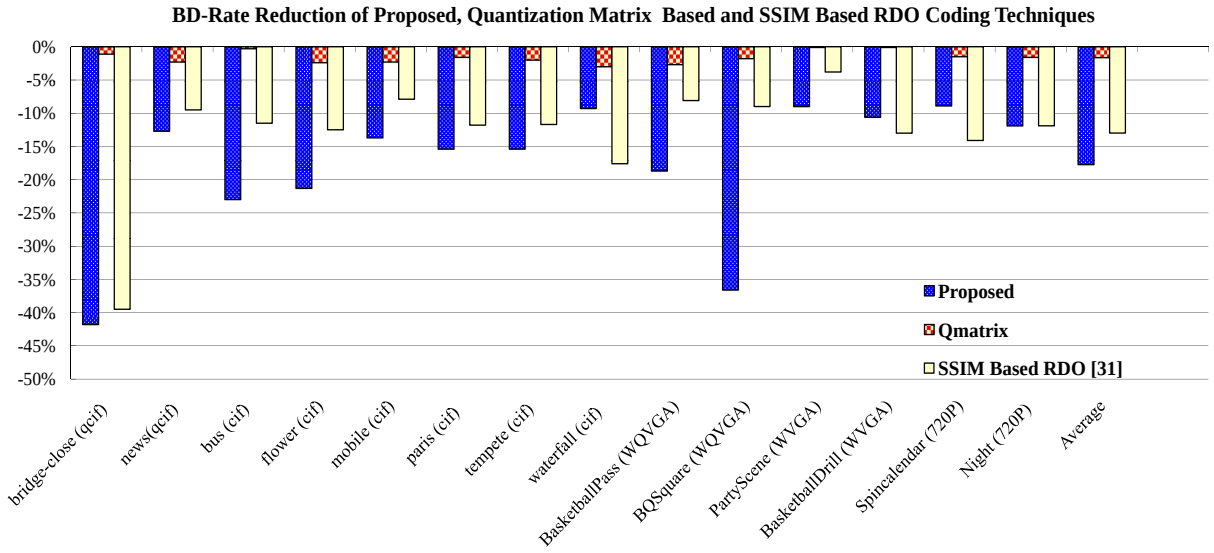


Figure 7.8: Performance comparisons of the proposed, quantization matrix and the SSIM based RDO coding techniques. (Anchor: conventional H.264/AVC)

and standard quantization matrix based video coding scheme in H.264/AVC are shown in Fig. 7.8. In this experiment, IPP GOP structure and CABAC coding techniques are used. The QP values range from 23 to 38 with an interval of 5. For most of the sequences, the proposed divisive normalization scheme achieves better coding performance. As discussed before, our scheme performs better especially for the sequences with significant statistical differences in the same frame, such as *Flower* and *Bus*. On average, compared with the SSIM based RDO scheme [159], the proposed scheme achieves better rate reduction of -17.7% vs -13.0%.

7.3.2 Subjective Performance Evaluation

To further validate our scheme, we carried out two subjective quality evaluation tests based on a two-alternative-forced-choice (2AFC) method. This method is widely used in psychophysical studies [64, 183], where in each trial, a subject is shown a pair of video sequences and is asked (forced) to choose the one he/she thinks to have better quality.

Table 7.1: Performance comparison of the proposed algorithm with H.264/AVC Anchor using HEVC standard testing sequences

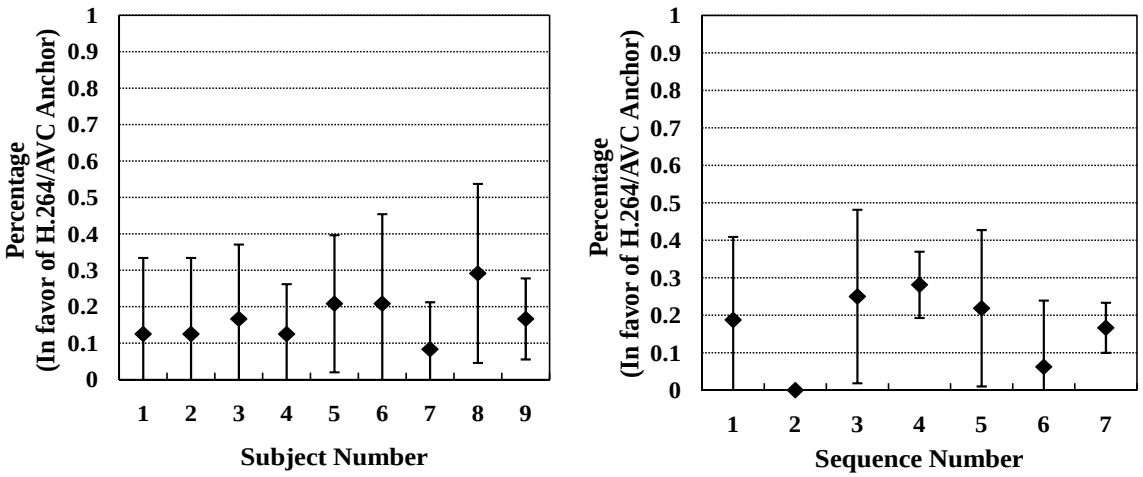
	Sequence	Resolution	Bit-Rate Savings	Average
Class A	Traffic	2560 × 1600	-19.48%	-18.37%
	PeopleOnStreet	2560 × 1600	-17.26%	
Class B	Kimono	1920 × 1080	-5.90%	-15.23%
	ParkScene	1920 × 1080	-12.55%	
	Cactus	1920 × 1080	-13.22%	
	BasketballDrive	1920 × 1080	-14.71%	
	BQTerrace	1920 × 1080	-29.81%	
Class C	BasketballDrill	832 × 480	-15.08%	-14.24%
	BQMall	832 × 480	-15.01%	
	PartyScene	832 × 480	-15.03%	
	RaceHorses	832 × 480	-11.86%	
Class D	BasketballPass	416 × 240	-16.17%	-22.02%
	BQSquare	416 × 240	-46.54%	
	BlowingBubbles	416 × 240	-14.79%	
	RaceHorses	416 × 240	-10.56%	
Class E	FourPeople	1280 × 720	-13.87%	-14.10%
	Johnny	1280 × 720	-14.89%	
	KristenAndSara	1280 × 720	-13.55%	
Class F	BasketballDrillText	832 × 480	-18.79%	-13.66%
	ChinaSpeed	1024 × 768	-15.70%	
	SlideEditing	1280 × 720	-6.48%	
	SlideShow	1280 × 720	-24.02%	
	Average			**−16.26%**

For each subjective test, we selected six pairs of sequences with different resolutions. In the first test, the sequences were compressed by H.264/AVC and the proposed method at the same bit rate but with different SSIM levels. In the second test, the sequences were coded to achieve the same SSIM levels (where the proposed scheme uses much lower bit rates). Tables 7.4 and 7.5 list all the test sequences as well as their SSIM values and bit rates. In the 2AFC test, each pair is repeated four times with random order. As a result, in each test we obtained 24 2AFC results for each subject. Eight subjects participated in the experiments.

The results of the two subjective tests are reported in Figs. 7.9 and 7.10, respectively. In each figure, the percentage by which the subjects are in favor of the H.264/AVC against the proposed scheme are shown. We also plot the error bars ($\pm$ one standard deviation between the measurements) over the eight subjects and over the six sequences. Error bars of right most data points are calculated based on standard deviation of average values. As can be observed in Fig. 6.18, the subjects are inclined to select the proposed method for better video quality. On the contrary, for the second test in Fig. 6.19, it turns out that for almost all cases the percentage is close to 50% and nearly all error bars cross the 50% line. These results provide useful evidence that the proposed method achieves the same level of quality with lower bit rates or creates better quality video at the same bit rates.

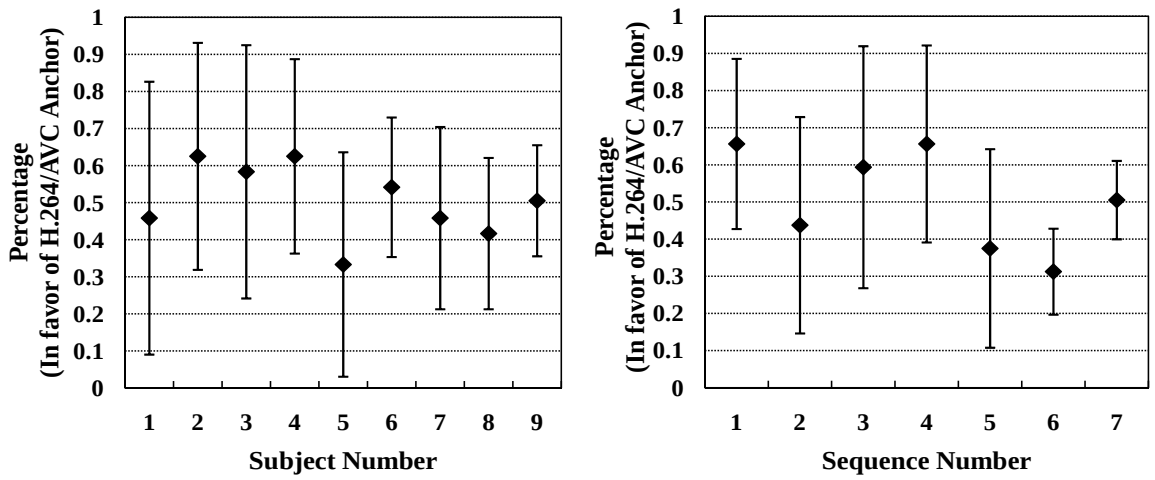
7.4 HEVC Implementation

Recent advances in video capturing and display technologies, along with the exponentially increasing demand of video services, challenge the video coding research community to design new algorithms able to significantly improve the compression performance of the current H.264/AVC standard. This target is currently gaining evidence with the standardization activities in the High Efficiency Video Coding (HEVC) project. The distortion models used in HEVC are mean squared error (MSE) and sum of absolute difference (SAD). However, they are widely criticized for not correlating well with perceptual image quality. The structural similarity (SSIM) index has been found to be a good indicator of perceived image quality. Meanwhile, it is computationally simple compared with other



(a) (b)

Figure 7.9: Subjective test 1: Similar bit rate with different SSIM values. (a) Mean and standard deviation (shown as error-bar) of preference for individual subject (1∼8: subject number, 9: average).(b) Mean and standard deviation (shown as error-bar) of preference for individual sequence (1∼6: sequence number, 7: average).



(a) (b)

Figure 7.10: Subjective test 2: Similar SSIM with different bit rates. (a) Mean and standard deviation (shown as error-bar) of preference for individual subject (1∼8: subject number, 9: average).(b) Mean and standard deviation (shown as error-bar) of preference for individual sequence (1∼6: sequence number, 7: average).

Table 7.2: Performance of the Proposed Algorithms (Compared with H.264/AVC Video Coding)

Sequence	QP_1={18,22,26,30}				QP_2={26,30,34,38}			
	$\Delta SSIM$	ΔR	$\Delta SSIM_\omega$	ΔR_ω	$\Delta SSIM$	ΔR	$\Delta SSIM_\omega$	ΔR_ω
Akiyo(QCIF)	0.0038	-20.5%	0.0044	-23.0%	0.0091	-14.0%	0.0084	-14.6%
Bridge-close(QCIF)	0.0066	-33.1%	0.0069	-28.3%	0.0289	-42.5%	0.0241	-42.6%
Carphone(QCIF)	0.0022	-12.9%	0.0027	-14.1%	0.0040	-8.2%	0.0042	-9.2%
Coastguard(QCIF)	0.0034	-7.0%	0.0027	-6.6%	0.0094	-9.0%	0.0075	-8.7%
Container(QCIF)	0.0024	-10.5%	0.0007	-3.9%	0.0046	-12.3%	0.0034	-10.9%
Grandma(QCIF)	0.0063	-20.0%	0.0066	-21.5%	0.0131	-14.6%	0.0119	-15.0%
News(QCIF)	0.0033	-15.7%	0.0034	-15.1%	0.0078	-13.2%	0.0077	-13.4%
Salesman(QCIF)	0.0041	-12.6%	0.0050	-14.3%	0.0136	-12.2%	0.0127	-12.7%
Akiyo(CIF)	0.0029	-20.5%	0.0032	-23.4%	0.0043	-12.5%	0.0042	-13.4%
Bus(CIF)	0.0048	-17.1%	0.0041	-14.6%	0.0205	-23.7%	0.0170	-23.2%
Coastguard(CIF)	0.0033	-7.4%	0.0028	-7.4%	0.0119	-11.7%	0.0097	-11.7%
Flower(CIF)	0.0036	-23.0%	0.0052	-24.7%	0.0092	-19.2%	0.0111	-22.1%
Mobile(CIF)	0.0014	-9.2%	0.0020	-9.7%	0.0056	-14.0%	0.0058	-13.8%
Paris(CIF)	0.0036	-15.0%	0.0025	-10.1%	0.0109	-17.9%	0.0091	-15.9%
Tempete(CIF)	0.0023	-13.4%	0.0035	-15.9%	0.0084	-14.7%	0.0084	-15.2%
Waterfall(CIF)	0.0038	-13.1%	0.0042	-12.7%	0.0132	-10.5%	0.0118	-10.5%
BigShip(720P)	0.0040	-11.8%	0.0036	-12.10%	0.0051	-7.3%	0.0044	-7.5%
Night(720P)	0.0030	-13.0%	0.0031	-14.1%	0.0064	-11.5%	0.0060	-12.0%
Spincalendar(720P)	0.0046	-19.9%	0.0024	-11.60%	0.0035	-13.8%	0.0017	-9.1%
Parkrun(720P)	0.0084	-3.9%	0.0066	-15.2%	0.0317	-36.5%	0.0257	-35.4%
Average	0.0039	-15.0%	0.0038	-14.9%	0.0111	-16.0%	0.0097	-15.8%

Table 7.3: Complexity Overhead of the Proposed Scheme

Sequences	ΔT in Encoder	ΔT in Decoder
Akiyo(QCIF)	1.20%	8.97%
News(QCIF)	1.17%	11.30%
Mobile(QCIF)	1.34%	5.3%
Bus(CIF)	1.16%	9.16%
Flower(CIF)	1.11%	8.75%
Tempete(CIF)	0.96%	7.38%
Average	1.16%	8.48%

Table 7.4: SSIM Indices and Bit Rates of Testing Sequences Used in the Subjective Test I. (Similar Bit Rate but Different SSIM Values)

Sequences	H.264/AVC		Proposed	
	SSIM	Bit rate	SSIM	Bit rate
Bridge-close(QCIF)	0.8892	29.56	0.9216	29.07
Bus(CIF)	0.8259	273.7	0.8531	262.03
Flower(CIF)	0.9121	317.8	0.9170	296.43
Mobile(CIF)	0.9462	631.89	0.9532	630.69
Paris(CIF)	0.8825	144.2	0.8902	142.59
Parkrun(720P)	0.7921	4311.6	0.8527	3768.34

state-of-the-art perceptual quality measures and has a number of desirable mathematical properties for optimization tasks. We propose a perceptual video coding method to improve upon the current HEVC based on an SSIM-inspired divisive normalization scheme as an attempt to transform the DCT domain frame prediction residuals to a perceptually uniform space before encoding. Based on the residual divisive normalization process, we define a distortion model for mode selection and show that such a divisive normalization strategy largely simplifies the subsequent perceptual rate-distortion optimization procedure. Experiments show that the proposed scheme can achieve significant gain in terms of rate-SSIM performance when compared with HEVC.

Table 7.5: SSIM Indices and Bit Rates of Testing Sequences Used in the Subjective Test II. (Similar SSIM Values but Different Bit Rate)

Sequences	H.264/AVC		Proposed	
	SSIM	Bit rate	SSIM	Bit rate
Bridge-close(QCIF)	0.8777	23.35	0.8764	12.76
News(QCIF)	0.9784	102.51	0.9786	86.18
Waterfall(CIF)	0.9619	474.09	0.962	408.79
Mobile(CIF)	0.9462	631.89	0.9467	537.78
Night(720P)	0.9845	18706.85	0.9839	15671.46
Bigship(720P)	0.9018	1552.8	0.9015	1390.08

The proposed scheme is completely compatible with any frame type supported by HEVC, as well as any size or shape choices of CTB, PU and TU, which create significant complications as opposed to the macroblock (MB) structure defined in previous video coding standards such as H.264/AVC. First, the expected values of local divisive normalization factors (the denominator in (7.10) and (7.11)) are obtained by first dividing the predicted current frame into 4×4 blocks (the greatest common divisor size for CTB, PU and TU) and then averaged over the whole frame. This avoids the problem of variable sizes of TU that create uneven number of DCT coefficients, and thus reduces the difficulty in estimating the expected values of the divisive normalization factor. Second, the divisive normalization factor for each 4×4 block is computed in pixel domain rather than DCT transform domain. Since DCT is a unitary transform that obeys Parseval's theorem, we have

$$\mu_x = \frac{\sum_{i=0}^{N-1} x(i)}{N} = \frac{X(0)}{\sqrt{N}}, \tag{7.41}$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{N-1} X(i)^2}{N-1}, \quad \sigma_{xy} = \frac{\sum_{i=1}^{N-1} X(i)Y(i)}{N-1}. \tag{7.42}$$

As a result, although our algorithm is derived in DCT domain, it is not necessary to perform actual DCT transform for each block in order to perform residual normalization. It allows us to calculate the energy values in pixel domain instead of DCT domain. Since the pixel values used to calculate the energy values are available at the decoder as well,

(15) and (16) can also be employed at the decoder. Third, the divisive normalization factor is spatially adaptive but coincides with individual TU. In other words, every TU is associated with a single set of divisive normalization factors but different from other TUs. The normalization matrix thus varies based on the size of TU. However, only two divisive normalization factors are used, one for the DC coefficient and the other for all AC coefficients. Since each TU may contain multiple 4×4 blocks, the divisive normalization factor for each TU is estimated by averaging the divisive normalization factors computed for all 4×4 blocks contained in the TU.

7.4.1 Objective Performance Evaluation

To validate the accuracy and efficiency of the proposed divisive normalization representation based perceptual video coding scheme, we integrated our scheme into the HEVC reference software HM 8.0. All test video sequences are in YUV 4:2:0 format. We use the standard configuration files to compare our method with the HEVC coding scheme in various aspects, including the R-D performance, the coding and decoding complexities and the visual performance. The SSIM index for the whole video sequence are obtained by simply averaging the respective values of individual frames. We employ the method proposed in [6] to calculate the differences between two RD curves which is also used by JCT-VC to compare the performance of various algorithms.[1] The QP values used to obtain the RD curves are 22, 27, 32 and 37, respectively.

Tables 7.6, 7.7, and 7.8 show the rate savings achieved using proposed scheme for various standard test sequences using All-Intra, Low-Delay P, and Random Access configurations, respectively. Over a wide range of test sequences with resolutions from WQVGA to 4K, our proposed scheme achieves an average rate reduction of about 6% for the same SSIM value and the maximum coding gain of 32%. The divisive normalization based video compression can substantially improve the rate-distortion performance of HEVC as rate-saving of more than 1% is considered very significant in the HEVC community. The performance improvement varies substantially, depending on the content of the video frame

[1] Since R-SSIM curve exhibits a similar shape as R-PSNR curve, we use the same tool proposed in [6] to calculate the average of SSIM differences.

Table 7.6: Performance comparison of the proposed algorithm with HEVC Anchor (HM 8.0) for All-Intra configuration

	Sequence	Resolution	Bit-Rate Savings	Average
Class A	Traffic	2560 × 1600	-4.28%	-4.85%
	PeopleOnStreet	2560 × 1600	-9.73%	
	Nebuta	2560 × 1600	-1.99%	
	SteamLocomotive	2560 × 1600	-3.41%	
Class B	Kimono	1920 × 1080	-1.27%	-4.50%
	ParkScene	1920 × 1080	0.56%	
	Cactus	1920 × 1080	-3.54%	
	BasketballDrive	1920 × 1080	-7.41%	
	BQTerrace	1920 × 1080	-10.84%	
Class C	BasketballDrill	832 × 480	-9.81%	-3.81%
	BQMall	832 × 480	-3.85%	
	PartyScene	832 × 480	1.25%	
	RaceHorses	832 × 480	-2.82%	
Class D	BasketballPass	416 × 240	-7.20%	-6.76%
	BQSquare	416 × 240	-18.28%	
	BlowingBubbles	416 × 240	2.06%	
	RaceHorses	416 × 240	-3.62%	
Class E	FourPeople	1280 × 720	-4.95%	-7.19%
	Johnny	1280 × 720	-5.28%	
	KristenAndSara	1280 × 720	-11.35%	
Class F	BasketballDrillText	832 × 480	-12.03%	-9.51%
	ChinaSpeed	1024 × 768	-13.65%	
	SlideEditing	1280 × 720	4.16%	
	SlideShow	1280 × 720	-16.50%	
	Average			**−6.10%**

Table 7.7: Performance comparison of the proposed algorithm with HEVC Anchor (HM 8.0) for Low Delay P configuration

	Sequence	Resolution	Bit-Rate Savings	Average
Class B	Kimono	1920 × 1080	-2.24%	-7.05%
	ParkScene	1920 × 1080	-4.83%	
	Cactus	1920 × 1080	-5.53%	
	BasketballDrive	1920 × 1080	-9.64%	
	BQTerrace	1920 × 1080	-13.03%	
Class C	BasketballDrill	832 × 480	-11.97%	-5.85%
	BQMall	832 × 480	-1.85%	
	PartyScene	832 × 480	-3.49%	
	RaceHorses	832 × 480	-6.09%	
Class D	BasketballPass	416 × 240	-10.52%	-12.72%
	BQSquare	416 × 240	-32.42%	
	BlowingBubbles	416 × 240	-0.56%	
	RaceHorses	416 × 240	-7.41%	
Class E	FourPeople	1280 × 720	2.31%	2.36%
	Johnny	1280 × 720	4.18%	
	KristenAndSara	1280 × 720	0.59%	
Class F	BasketballDrillText	832 × 480	-13.74%	-3.83%
	ChinaSpeed	1024 × 768	-15.21%	
	SlideEditing	1280 × 720	25.47%	
	SlideShow	1280 × 720	-11.85%	
	Average			**−5.42%**

Table 7.8: Performance comparison of the proposed algorithm with HEVC Anchor (HM 8.0) for Random Access configuration

	Sequence	Resolution	Bit-Rate Savings	Average
Class A	Traffic	2560 × 1600	-3.91%	-6.99%
	PeopleOnStreet	2560 × 1600	-9.29%	
	Nebuta	2560 × 1600	-10.18%	
	SteamLocomotive	2560 × 1600	-4.59%	
Class B	Kimono	1920 × 1080	-1.45%	-3.56%
	ParkScene	1920 × 1080	-1.44%	
	Cactus	1920 × 1080	-1.81%	
	BasketballDrive	1920 × 1080	-7.22%	
	BQTerrace	1920 × 1080	-5.87%	
Class C	BasketballDrill	832 × 480	-7.44%	-2.35%
	BQMall	832 × 480	-0.84%	
	PartyScene	832 × 480	2.20%	
	RaceHorses	832 × 480	-3.31%	
Class D	BasketballPass	416 × 240	-7.14%	-7.17%
	BQSquare	416 × 240	-19.64%	
	BlowingBubbles	416 × 240	2.59%	
	RaceHorses	416 × 240	-4.47%	
Class F	BasketballDrillText	832 × 480	-7.64%	-3.46%
	ChinaSpeed	1024 × 768	-11.13%	
	SlideEditing	1280 × 720	15.16%	
	SlideShow	1280 × 720	-10.22%	
	Average			**−4.70**%

being encoded. In general, the video frames that have large variations in terms of the texture content often result in more performance gain.

The R-D performance for sequences with various resolutions are shown in Fig. 7.11. In general, the performance gap between the proposed method and the HEVC codec is maximum at the mid-range of QP values because at high bit rate, the quantization step is relatively smaller and thus the differences of quantization steps among the TUs are not significant and at low bit rate, since the AC coefficients are severely distorted, the normalization factors derived from the prediction frame do not precisely represent the properties of the original frame.

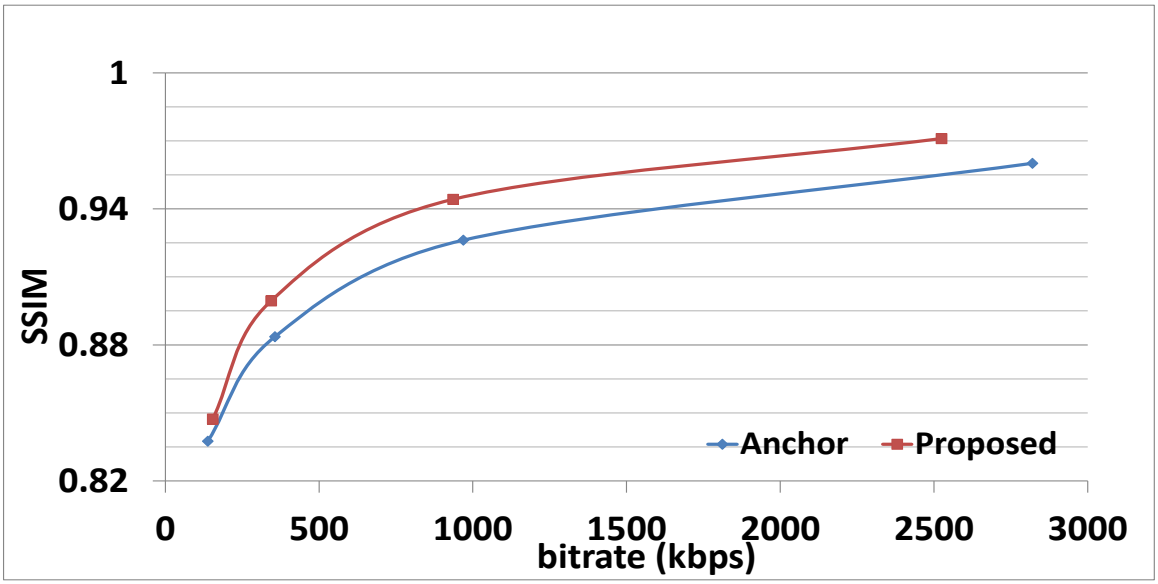
Rate-distortion optimized quantization (RDOQ) is employed in HEVC as a tool for pursuing high coding efficiency. RDOQ requires an exhaustive search over multiple candidates to determine the optimal quantized level by comparing their rate-distortion cost using MSE as the distortion measure. We study the effect of RDOQ on the performance of the proposed algorithm. Specifically, we compare the RD-performance of the divisive normalization scheme (HM-DNT) with the combined performance of divisive normalization scheme and RDOQ, i.e., HM-DNT-RDOQ. Table 7.9 shows that the performance of the proposed algorithm degrades by almost 3% when RDOQ is employed. The loss in performance can possibly be due to the conflict of optimization criterion between divisive normalization and RDOQ, the former aims to maximize SSIM, however, the latter optimizes for minimum MSE.

When evaluating the coding complexity overhead, we calculate ΔT with

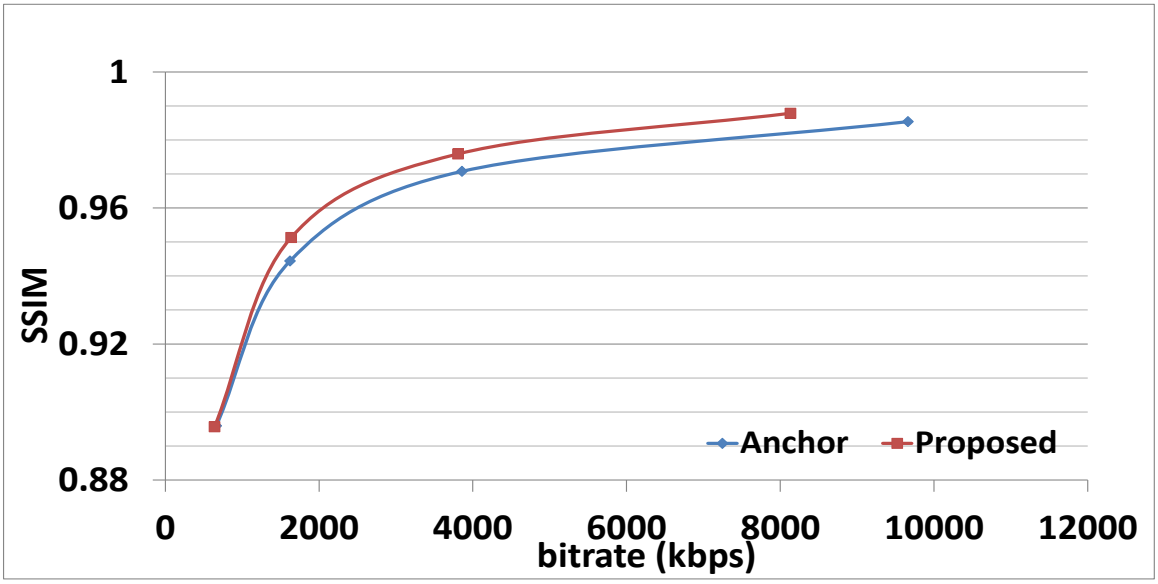
$$\Delta T = \frac{T_{pro} - T_{HEVC}}{T_{HEVC}} \times 100\%, \tag{7.43}$$

where T_{HEVC} and T_{pro} indicate the total coding time for the sequence with the HEVC and the proposed coding schemes, respectively. The average encoding and decoding overhead is 6% and 8%, respectively.

Figure 7.12 visually compares the proposed scheme with HEVC. For a fair comparison, the bit rate for the proposed scheme is lower than that of HEVC. However, since our proposed divisive normalization scheme is based on SSIM index optimization, higher SSIM and lower PSNR values are achieved. We can be observe by visual comparison of the



(a) BQSquare



(b) PartyScene

Figure 7.11: Rate-SSIM performance comparison between HEVC and the proposed video coding scheme using Low Delay P configuration

reconstructed frame with the original frame, the proposed method achieves significantly better visual quality for the same rate. Furthermore, the quality improvement of the reconstructed frame by the proposed scheme is evident from the SSIM maps. The proposed method does a better job in preserving the texture present in the original frame as depicted by the overall brighter SSIM map of the reconstructed frame. The distortion distribution of the proposed scheme is more uniform across space and more information and details have been preserved. The visual quality improvement is due to the fact that we perform coding algorithms in a perceptual uniform space which can result in a better R-D performance from perceptual point of view.

7.5 Adaptive Quantization

Residual divisive normalization based video compression is non-normative as the decoder has to be modified as shown in Figure 7.2. Therefore, we devise the Adaptive Quantization (AQ) algorithm to make use of the divisive normalization approach in case of a standard decoder. The purpose of the divisive normalization process is to convert the transform residuals into an perceptually uniform space. Thus the factor f determines the perceptual importance of each of the corresponding transform coefficient. The proposed divisive normalization scheme can be interpreted in two ways. An adaptive normalization factor is applied, followed by quantization with a predefined fixed step Q_s. Alternatively, an adaptive quantization matrix is defined for each coding unit based on the perceptual information it carries and is subsequently quantized with quantization step Q'_s. The blocks which are less important are quantized coarsely with respect to the more important block. Therefore, divisive normalization based video coding can also be interpreted as video compression with adaptive quantization. Now, we will derive the relationship between the normalization, f, and the change in quantization parameter, ΔQP, as a result of divisive normalization process.

Assume $X(k)$ to be the k^{th} DCT transform coefficient of a residual block, then the normalized coefficient is computed as $X(k)' = X(k)/f$, where f is a positive normalization factor which is calculated as the energy of a cluster of neighboring coefficients.

Table 7.9: Effect of RDOQ on the performance of the proposed algorithm (HM-DNT vs HM-DNT-RDOQ)

	Sequence	Resolution	Bit-Rate Savings	Average
Class B	Kimono	1920 × 1080	4.24%	3.95%
	ParkScene	1920 × 1080	3.05%	
	Cactus	1920 × 1080	4.18%	
	BasketballDrive	1920 × 1080	5.30%	
	BQTerrace	1920 × 1080	3.01%	
Class C	BasketballDrill	832 × 480	1.80%	3.13%
	BQMall	832 × 480	3.48%	
	PartyScene	832 × 480	3.07%	
	RaceHorses	832 × 480	4.17%	
Class D	BasketballPass	416 × 240	3.93%	3.30%
	BQSquare	416 × 240	2.53%	
	BlowingBubbles	416 × 240	3.04%	
	RaceHorses	416 × 240	3.72%	
Class E	FourPeople	1280 × 720	1.44%	1.47%
	Johnny	1280 × 720	0.61%	
	KristenAndSara	1280 × 720	2.36%	
Class F	BasketballDrillText	832 × 480	1.76%	2.84%
	ChinaSpeed	1024 × 768	2.27%	
	SlideEditing	1280 × 720	2.17%	
	SlideShow	1280 × 720	5.17%	
	Average			**3.06**%

The quantization process of the normalized residuals for a given predefined Q_s can be formulated as

$$\begin{aligned} Q(k) &= \text{sign}\{X(k)'\}\text{round}\{\frac{|X(k)'|}{Q_s} + p\} \\ &= \text{sign}\{X(k)\}\text{round}\{\frac{|X(k)|}{Q_s \cdot f} + p\} \\ &= \text{sign}\{X(k)\}\text{round}\{\frac{|X(k)|}{Q'_s} + p\} \end{aligned} \tag{7.44}$$

where p is the rounding offset.

The HEVC test model (HM) and the H.264/AVC standard employ a similar quantization parameter (QP) scaling scheme. The quantization step size applied to each transform coefficient is determined approximately as

$$Q_s = 2^{\frac{\text{QP}-4}{6}}. \tag{7.45}$$

The expression for Q'_s can be written as

$$\begin{aligned} Q'_s &= f \cdot Q_s, \\ &= 2^{\frac{\text{QP}'-4}{6}}, \end{aligned} \tag{7.46}$$

where $\text{QP}' = \text{QP} + \Delta\text{QP}$ is the modified quantization parameter as a result of the divisive normalization process. The corresponding ΔQP as a function of the normalization factor, f, is given by

$$\Delta\text{QP} = 6 \log_2 f. \tag{7.47}$$

Since f is real, ΔQP is not necessarily an integer, which provides fine tuning of the QP value of each coding unit in order to obtain the best perceptual quality. If the decoder the standard compatible then we cannot signal a non-integer ΔQP to the decoder. Therefore we perform the rounding operation to determine the final ΔQP using

$$\Delta\text{QP}_\text{f} = \lfloor 6 \log_2 f + 0.5 \rfloor. \tag{7.48}$$

The effective f_n as a result of the rounding operation can be determined as

$$f_n = 2^{\frac{\Delta \mathrm{QP_f}}{6}}. \tag{7.49}$$

For the purpose of Adaptive Quantization, $\Delta \mathrm{QP_f}$ is calculated for every MB in case of H.264/AVC and for every CU in case of HEVC. The average MB/CU's energy is calculated after dividing it in non-overlapping 4×4 sub-blocks. The perceptual rate distortion optimization for mode selection is also performed based on the method explained in Section 7.2.2.

We use the images in LIVE and TID2008 databases with compression as the distortion type to test the performance of f_n based IQA measure. We also provide its comparison with PSNR and SSIM. Three metrics are employed for evaluation, which include PLCC and MAE after nonlinear mapping between subjective and objective scores and Spearman's rank-order correlation coefficient (SRCC). The results are shown in Table 7.10. The proposed distortion measure significantly outperforms PSNR and mostly performs at least as well as SSIM.

Table 7.10: Performance comparison of IQA measures using the LIVE and TID2008 databases

	PLCC			MAE			SRCC		
	PSNR	SSIM	f_n	PSNR	SSIM	f_n	PSNR	SSIM	f_n
LIVE - JPEG2000 (1)	0.9332	0.9687	0.9656	6.5033	4.7620	4.7941	0.9264	0.9637	0.9588
LIVE - JPEG2000 (2)	0.8740	0.9726	0.9591	9.9693	5.2016	4.913	0.8549	0.9604	0.9611
LIVE - JPEG (1)	0.8867	0.9667	0.9637	8.6817	4.7096	4.731	0.8779	0.9637	0.9594
LIVE - JPEG (2)	0.9168	0.9851	0.9817	10.0107	4.6077	4.4379	0.7699	0.9215	0.9308
TID2008 - JPEG2000	0.8629	0.9667	0.9702	0.8137	0.3969	0.3619	0.8132	0.9625	0.9591
TID2008 - JPEG	0.8666	0.9540	0.9589	0.6858	0.3725	0.3673	0.3590	0.9252	0.9124

The complexity overhead, calculated using Equation (7.43), at the encoder is 1%, on average. There is no change in the complexity of the decoder as the bitstream generated by the encoder can be decoded by a standard decoder.

7.5.1 H.264/AVC

To validate the accuracy and efficiency of the proposed adaptive quantization based perceptual video coding scheme, we integrated our scheme into the H.264/AVC reference

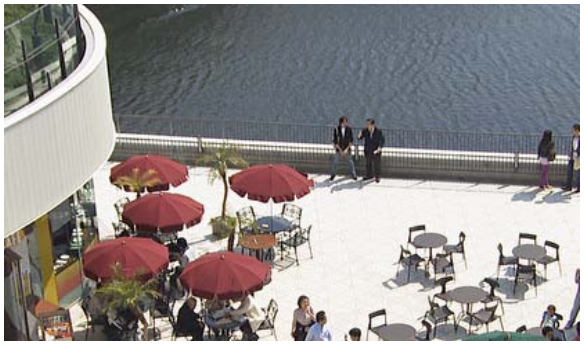
software JM 15.1. All the test video sequences are in YUV 4:2:0 format. We use the IPPP GOP structure and compare our scheme with the H.264/AVC coding schemes in terms of R-D performance. From Table 7.11, we can be observe that over a wide range of test sequences with resolutions from WQVGA to 4K, the proposed scheme achieves an average rate reduction of 14.36% while keeping the SSIM value the same and the maximum coding gain is 41.77%. The RD-curves for two of the sequences used to test the performance of the proposed method are shown in Figure 7.13. The adaptive quantization mechanism substantially improves the rate-SSIM performance of H.264/AVC on average.

7.5.2 HEVC

We integrated our scheme into the HEVC reference software HM 8.0 in order to validate the accuracy and efficiency of the proposed adaptive quantization based perceptual video coding scheme. All the test video sequences are in YUV 4:2:0 format. We use standard configuration settings of All-Intra, Low-Delay P, and Random Access profiles to compare our method with the HEVC coding scheme in terms of R-D performance. From Tables 7.12, 7.13, and 7.14, we can be observe that over a wide range of test sequences with resolutions from WQVGA to 4K, the proposed scheme achieves an average rate reduction of -3.67%, -2.53% and -2.64% respectively at the same quality level in terms of the SSIM index. The RD-curves for two of the sequences used to test the performance of the proposed method are shown in Figure 7.14. The adaptive quantization method can substantially improve the rate-distortion performance of HEVC as rate-saving of more than 1% is considered very significant in the HEVC community.

Table 7.11: Performance comparison of the proposed Adaptive Quantization algorithm with H.264/AVC Anchor (JM 15.1) using HEVC standard testing sequences

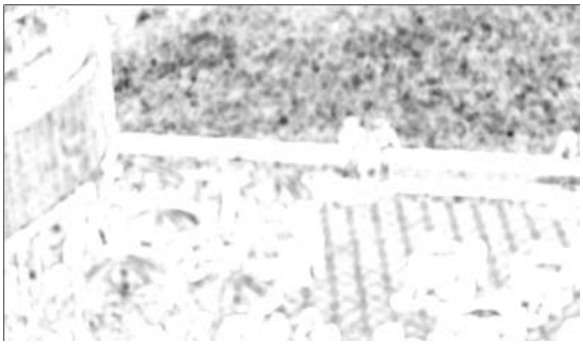
	Sequence	Resolution	Bit-Rate Savings	Average
Class A	Traffic	2560 × 1600	-17.69%	-15.51%
	PeopleOnStreet	2560 × 1600	-13.34%	
Class B	Kimono	1920 × 1080	-2.92%	-12.92%
	ParkScene	1920 × 1080	-10.37%	
	Cactus	1920 × 1080	-9.28%	
	BasketballDrive	1920 × 1080	-15.02%	
	BQTerrace	1920 × 1080	-27.01%	
Class C	BasketballDrill	832 × 480	-13.10%	-17.09%
	BQMall	832 × 480	-13.05%	
	PartyScene	832 × 480	-15.22%	
	RaceHorses	832 × 480	-12.19%	
Class D	BasketballPass	416 × 240	-14.94%	-20.06%
	BQSquare	416 × 240	-41.77%	
	BlowingBubbles	416 × 240	-13.88%	
	RaceHorses	416 × 240	-9.64%	
Class E	FourPeople	1280 × 720	-3.81%	-4.78%
	Johnny	1280 × 720	-4.53%	
	KristenAndSara	1280 × 720	-6.01%	
Class F	BasketballDrillText	832 × 480	-18.24%	-15.78%
	ChinaSpeed	1024 × 768	-13.01%	
	SlideEditing	1280 × 720	-5.79%	
	SlideShow	1280 × 720	-26.11%	
	Average			**−14.36**%



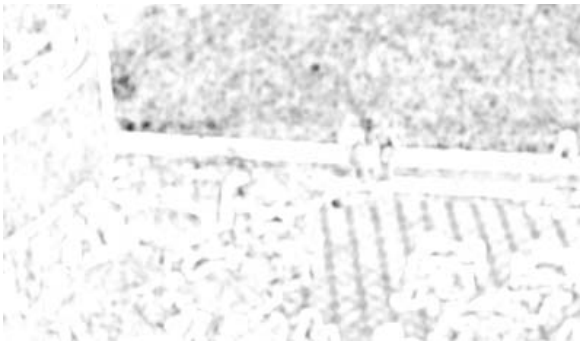
(a)

(b)

(c)

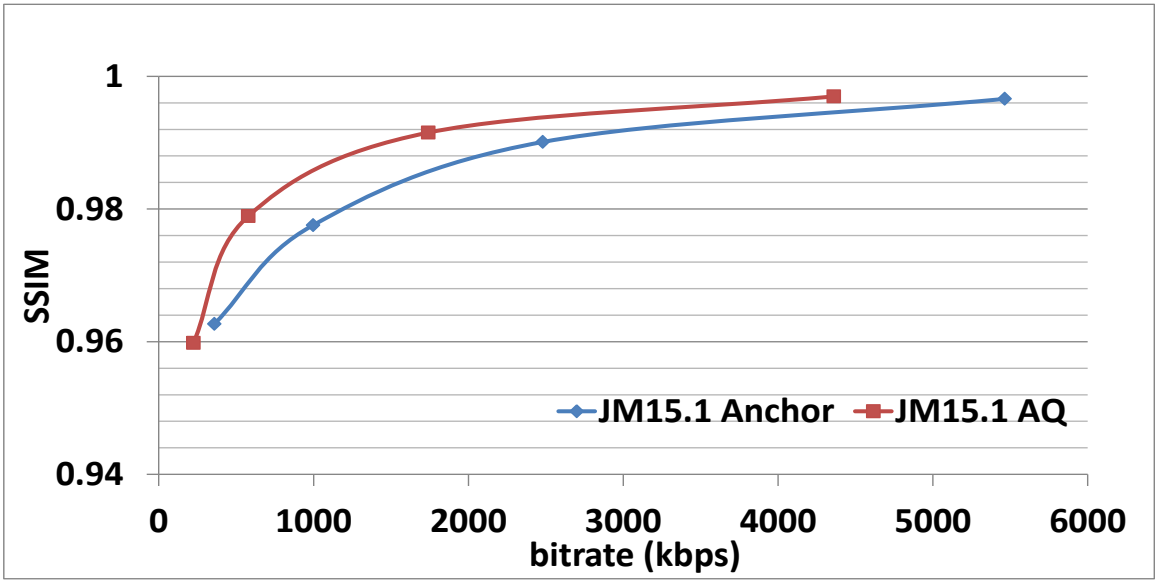


(d)

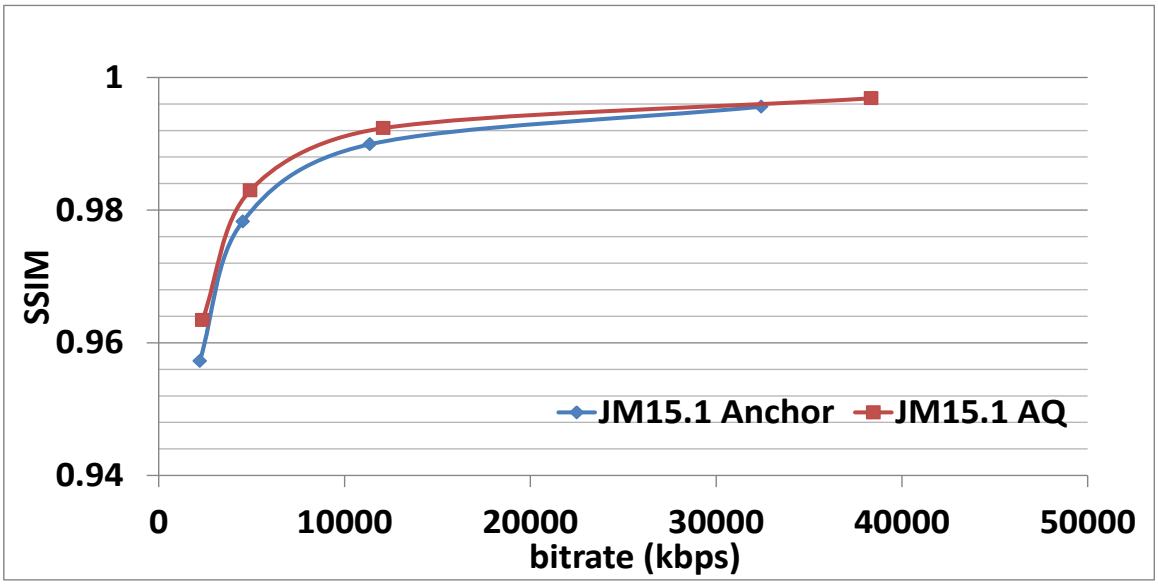


(e)

Figure 7.12: Visual quality comparison between HEVC and the proposed coding scheme: (a) Original frame; (b) HEVC coded; Bit rate: 356.5192 Kbit/s, SSIM = 0.8744, PSNR = 30.949 dB; (c) Proposed scheme; Bit rate: 349. 6576 Kbit/s, SSIM = 0.8936, PSNR = 29.1254 dB; (d) SSIM map of the HEVC coded video; (e) SSIM map of the video coded using the proposed scheme. In SSIM maps, brighter indicates better quality/larger SSIM value.

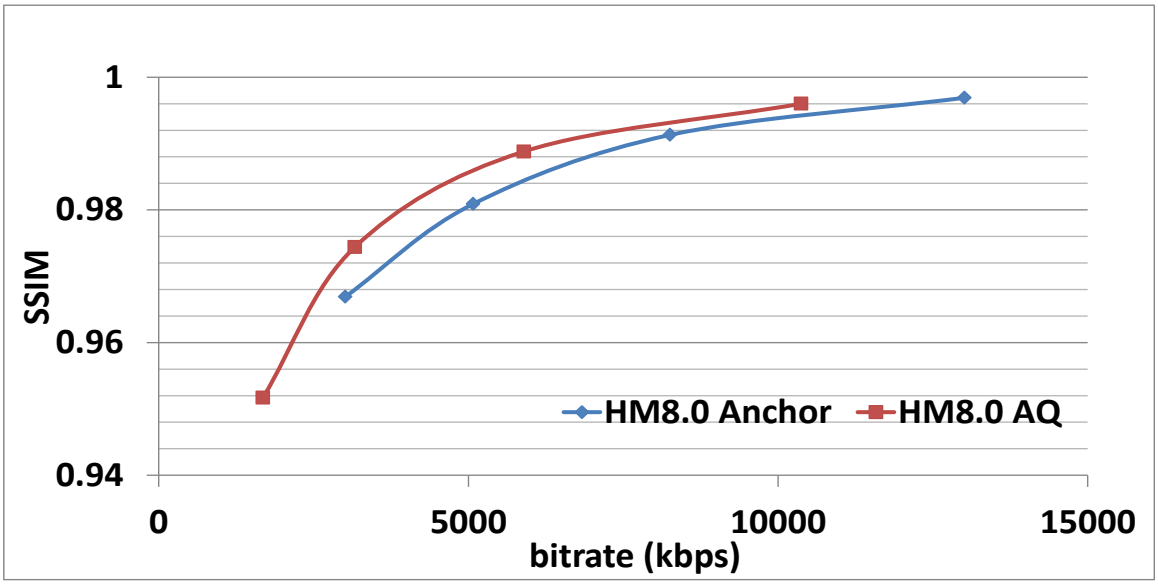


(a) BQSquare

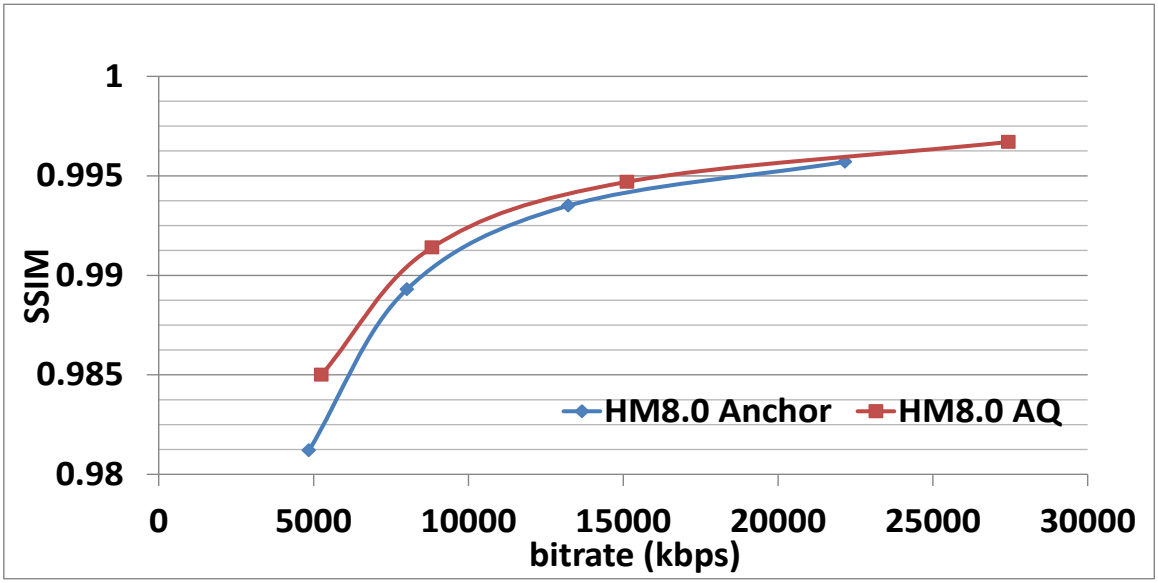


(b) Traffic

Figure 7.13: Rate-SSIM performance comparison between H.264/AVC and the proposed video coding scheme using IPP GOP structure



(a) BQSquare



(b) KristenAndSara

Figure 7.14: Rate-SSIM performance comparison between HEVC and the proposed video coding scheme using All-Intra configuration

Table 7.12: Performance comparison of the proposed Adaptive Quantization algorithm with HEVC Anchor (HM 8.0) for All-Intra configuration

	Sequence	Resolution	Bit-Rate Savings	Average
Class A	Traffic	2560 × 1600	-3.41%	-3.70%
	PeopleOnStreet	2560 × 1600	-4.70%	
	Nebuta	2560 × 1600	-0.70%	
	SteamLocomotive	2560 × 1600	-5.99%	
Class B	Kimono	1920 × 1080	-1.89%	-3.78%
	ParkScene	1920 × 1080	-1.79%	
	Cactus	1920 × 1080	-3.38%	
	BasketballDrive	1920 × 1080	-5.17%	
	BQTerrace	1920 × 1080	-6.68%	
Class C	BasketballDrill	832 × 480	-6.79%	-3.37%
	BQMall	832 × 480	-2.40%	
	PartyScene	832 × 480	-1.25%	
	RaceHorses	832 × 480	-3.04%	
Class D	BasketballPass	416 × 240	-3.20%	-3.22%
	BQSquare	416 × 240	-9.69%	
	BlowingBubbles	416 × 240	1.31%	
	RaceHorses	416 × 240	-1.29%	
Class E	FourPeople	1280 × 720	-2.63%	-3.85%
	Johnny	1280 × 720	-1.50%	
	KristenAndSara	1280 × 720	-7.43%	
Class F	BasketballDrillText	832 × 480	-7.50%	-4.11%
	ChinaSpeed	1024 × 768	-5.60%	
	SlideEditing	1280 × 720	-6.81%	
	SlideShow	1280 × 720	-10.15%	
	Average			**−3.67**%

Table 7.13: Performance comparison of the proposed Adaptive Quantization algorithm with HEVC Anchor (HM 8.0) for Low Delay P configuration

	Sequence	Resolution	Bit-Rate Savings	Average
Class B	Kimono	1920 × 1080	-0.91%	-4.03%
	ParkScene	1920 × 1080	-1.65%	
	Cactus	1920 × 1080	-1.94%	
	BasketballDrive	1920 × 1080	-6.27%	
	BQTerrace	1920 × 1080	-9.37%	
Class C	BasketballDrill	832 × 480	-4.90%	-0.47%
	BQMall	832 × 480	-1.81%	
	PartyScene	832 × 480	4.41%	
	RaceHorses	832 × 480	-0.63%	
Class D	BasketballPass	416 × 240	-3.47%	-3.43%
	BQSquare	416 × 240	-15.27%	
	BlowingBubbles	416 × 240	5.54%	
	RaceHorses	416 × 240	-0.53%	
Class E	FourPeople	1280 × 720	0.39%	-1.02%
	Johnny	1280 × 720	-1.07%	
	KristenAndSara	1280 × 720	-2.39%	
Class F	BasketballDrillText	832 × 480	-6.47%	-3.43%
	ChinaSpeed	1024 × 768	-4.69%	
	SlideEditing	1280 × 720	6.47%	
	SlideShow	1280 × 720	-9.04%	
	Average			**−2.53**%

Table 7.14: Performance comparison of the proposed Adaptive Quantization algorithm with HEVC Anchor (HM 8.0) for Random Access configuration

	Sequence	Resolution	Bit-Rate Savings	Average
Class A	Traffic	2560 × 1600	-5.06%	-2.98%
	PeopleOnStreet	2560 × 1600	-3.98%	
	Nebuta	2560 × 1600	0.49%	
	SteamLocomotive	2560 × 1600	-3.36%	
Class B	Kimono	1920 × 1080	-1.02%	-3.31%
	ParkScene	1920 × 1080	-2.02%	
	Cactus	1920 × 1080	-2.32%	
	BasketballDrive	1920 × 1080	-5.03%	
	BQTerrace	1920 × 1080	-6.16%	
Class C	BasketballDrill	832 × 480	-4.46%	-1.56%
	BQMall	832 × 480	-1.45%	
	PartyScene	832 × 480	1.77%	
	RaceHorses	832 × 480	-2.08%	
Class D	BasketballPass	416 × 240	-2.37%	-3.04%
	BQSquare	416 × 240	-11.61%	
	BlowingBubbles	416 × 240	2.65%	
	RaceHorses	416 × 240	-0.84%	
Class F	BasketballDrillText	832 × 480	-4.95%	-2.30%
	ChinaSpeed	1024 × 768	-5.10%	
	SlideEditing	1280 × 720	8.80%	
	SlideShow	1280 × 720	-7.94%	
	Average			**−2.64%**

Chapter 8

Perceptual Experience of Time-Varying Video Quality

In real-world visual communications, it is a common experience that end-users receive video with significantly time-varying quality due to the variations in video content/complexity, codec configuration, and network conditions. How human visual quality-of-experience (QoE) changes with such time-varying video quality is not yet well-understood. In this Chapter, we present the subjective experiments we designed to examine the quality predictability between individual video segment of relatively constant quality and combined video consisting of multiple segments that have significantly different quality. Based on the subjective data, we propose an asymmetric adaptation (AA) model that leads to improved performance of both subjective and objective quality assessment approaches when using segment-level quality scores to predict multi-segment time-varying video quality.

8.1 Introduction

In practical network digital video communication systems, the source video content is subject to a series of distortions during the compression and transmission processes before

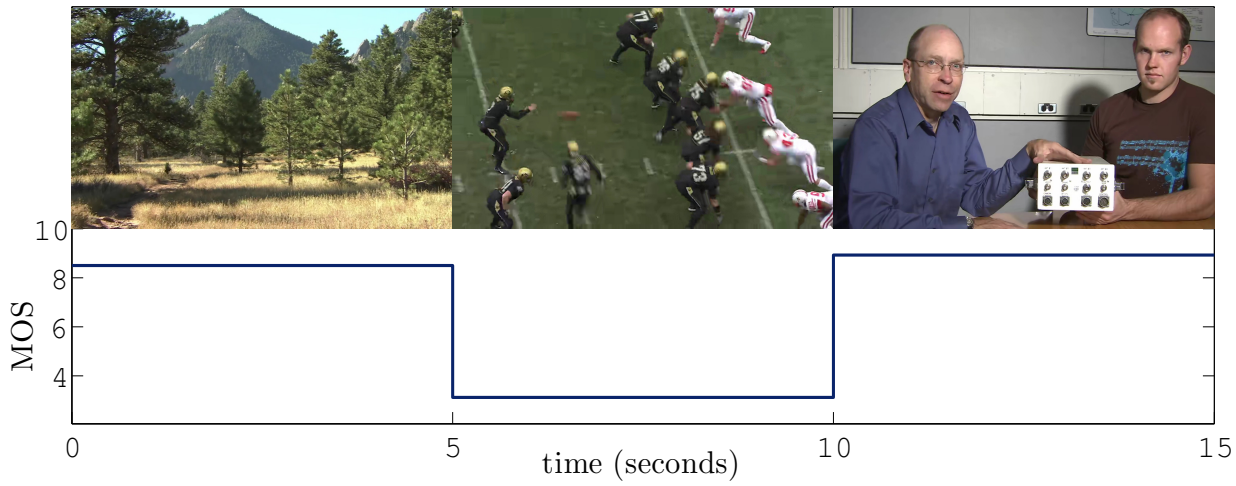


Figure 8.1: A schematic example of a three-scene sequence with time-varying quality in the subjective test.

being delivered to the end receivers. Very often, the quality of the received video varies over time. The source of such time-varying video quality may be at the sender side or within the communication network. At the sender side, video is compressed to meet the bandwidth constraints. Because of the large variations in the spatial/temporal/motion complexity in the video content, it is difficult to maintain constant video quality while making the best use of the communication channels, which often prefer approximately constant bit rate. In the communication network, packet loss and delay occur in somewhat random fashion, which, combined with the complexity of the coded video stream, often result in complicated distortions and quality variations when the video is decoded at the receiver side. Error correction and concealment techniques are commonly applied to partially recover the video but their performance varies as well.

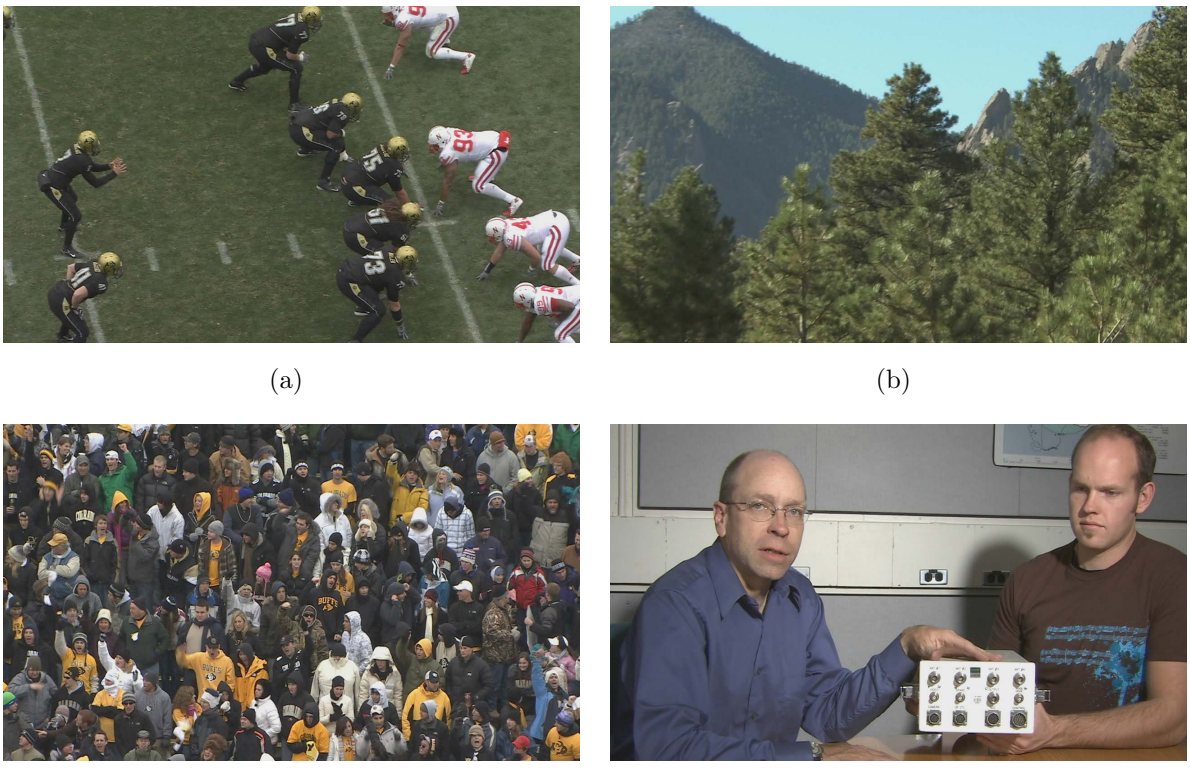
In this work, we attempt to investigate the problem in a more straightforward way. In particular, we carry out subjective test on both individual video segments (each with a single scene) and combined video consisting of multiple segments that have significantly different quality. We then study different approaches that use the quality of the individual segments to predict that of the combined multi-segment video. This study is different from previous works, which typically focused on instantaneous video quality (often measured on

a frame-by-frame basis) and its relationship to the aggregated quality of a video that contains one scene or multiple scenes. In reality, however, human subjects rarely judge video quality at such a high temporal resolution. Instead, based on our observation, they would rather give a single score to a segment of video, often of the same scene (regardless of the instantaneous quality variations between frames within the scene). Further, subjects tend to maintain their opinions until scene cut occurs, especially when adjacent scenes have very different content and quality. Eventually, the overall subjective opinion of the multi-scene video would be a result of pooling the segment-level quality. In this sense, our study better matches real-world scenarios, where a meaningful video content (such as a Youtube video) often contains multiple scenes with different levels of complexity and quality. The data collected from our subjective experiment allows us to study the quality predictability between individual video segments and combined multi-segment video. Our results show that none of the simple models such as linear averaging and weighted-averaging, nonlinear min- and median-filtering, and distortion-weighted averaging, produces impressive performance. We thus propose an asymmetric adaptation model to better account for the data. The model is useful in better understanding the psychological behavior of human subjects in evaluating time-varying video quality. It can also be directly applied to objective VQA algorithms to improve their performance, which is demonstrated using peak signal-to-noise-ratio (PSNR) and the multi-scale structural similarity index (MS-SSIM) [177] as examples.

8.2 Subjective Study

8.2.1 Video Database

We start building our video database by selecting video segments, each of which contains a single scene, thus in the rest of the chapter, the terms "scene" and "segment" are interchangeable. Four reference video segments are selected that contain indoor and outdoor scenes, flat areas and complex patterns, camera zooming/panning and object motion towards different directions. Frames extracted from the reference video segments are shown in Figure 8.2. The video sequences are progressively scanned, with high definition (HD) resolution (1280×800), and in YUV 4:2:0 format. All the videos are five seconds long,



(a) (b)

(c) (d)

Figure 8.2: Frames extracted from the reference video segments used in the subjective test

with a frame rate of 30 frames/second. Every raw video scene is compressed at three quality levels using the recent high efficiency video coding (HEVC) reference software HM 8.0 [139]. The three quality levels are obtained by adjusting the quantization parameter (QP) of the encoder, for which a small-scale initial subjective test was conducted, such that each scene has three compressed versions at high-, medium- and low-quality levels (the distribution of quality levels will be discussed later). In the end, a total of 147 video sequences are included in the database, which are classified into three categories:

- 12 single-scene 5-second-long sequences, created by HEVC compression;
- 27 two-scene 10-second-long sequences, constructed by concatenating two of the single-scene sequences with combinations of varying quality;
- 108 three-scene 15-second-long sequences, constructed by concatenating three single-scene sequences with combinations of varying quality.

Figure 8.1 shows representative frames extracted from a three-scene test sequence, where the time-varying segment-level quality are indicated by the variations of the Difference of Mean Opinion Score (DMOS). A large number of combinations are included in the 2-scene and 3-scene categories to provide precise information necessary to study human behaviors in evaluating time-varying video quality. In addition, single-scene videos are used as prefixes of two-scene videos. Likewise, two-scene videos are used as prefixes of three-scene videos. As a result, by simply asking each subject to score every sequence (1-scene, 2-scene, or 3-scene), we have the chance to monitor, track, and record the changes in quality scores along with the subject.

8.2.2 Subjective Test

Our subjective test generally follows the Absolute Category Rating (ACR) methodology, as suggested by ITU-T recommendation P.910 [65]. Although SSCQE [65] is designed for continuously tracking instantaneous video quality over time, it is not adopted in our experiment for the following reasons. First, as mentioned earlier, in practice human subjects

often opt to judge video quality on per scene or segment basis, discounting the instantaneous quality variations between frames within a scene. Second, in our database, the same coding configuration and parameters are applied to the full duration of each scene, which is also roughly constant in terms of content and complexity. As a result, a single quality score is sufficient to summarize its quality. Third, in SSCQE, there is time delay between the recorded instantaneous quality and the video content, and such delay varies between subjects and is also a function of slider "stiffness". This is an unresolved issue of the general SSCQE methodology, but is avoided when only a single score is acquired. Fourth, we observe that humans tend to keep their opinions unless there is a significant change in video quality that attracts their attention. This is more realistically matched to real-world scenarios when subjects are watching a movie or online video. Compared with SSCQE, ACR is much simpler and provides more reliable and more realistic quality evaluations in our video database.

Thirty naïve subjects (17 males, 13 females) - all university undergraduate and graduate students - took part in the 40-minute subjective test. The viewing distance is set to be four times of the picture height. Instructions were given to the subjects in both written and oral forms. A training session preceded the test where the subject was shown examples of distorted video sequences expected in the test. All the reference video sequences were also shown during the training session. During the main test, the 147 distorted video sequences were ordered randomly irrespective of their categories. Subjects scored the quality of each video sequence according to the eleven-grade $0-10$ numerical quality scale suggested in ITU-T recommendation P.910 [65].

After screening the data, 4 subjects were discovered to be outliers, and the scores given by the remaining 26 subjects were averaged to produce a mean opinion score (MOS) for each test sequence. Figure 8.3 plots the MOS scores versus video indices. Thanks to the initial subjective test before determining the Qp parameters used to create the compressed videos (as mentioned in Section 8.2.1), the resulting MOS values scatter in a wide range of the available scales $[0-10]$, which allows us to study different cases of quality transitions between the scenes.

After each test session, we also discussed with each subject, inquiring about what strategy had been used by the subject to determine the scores. This step did not affect the

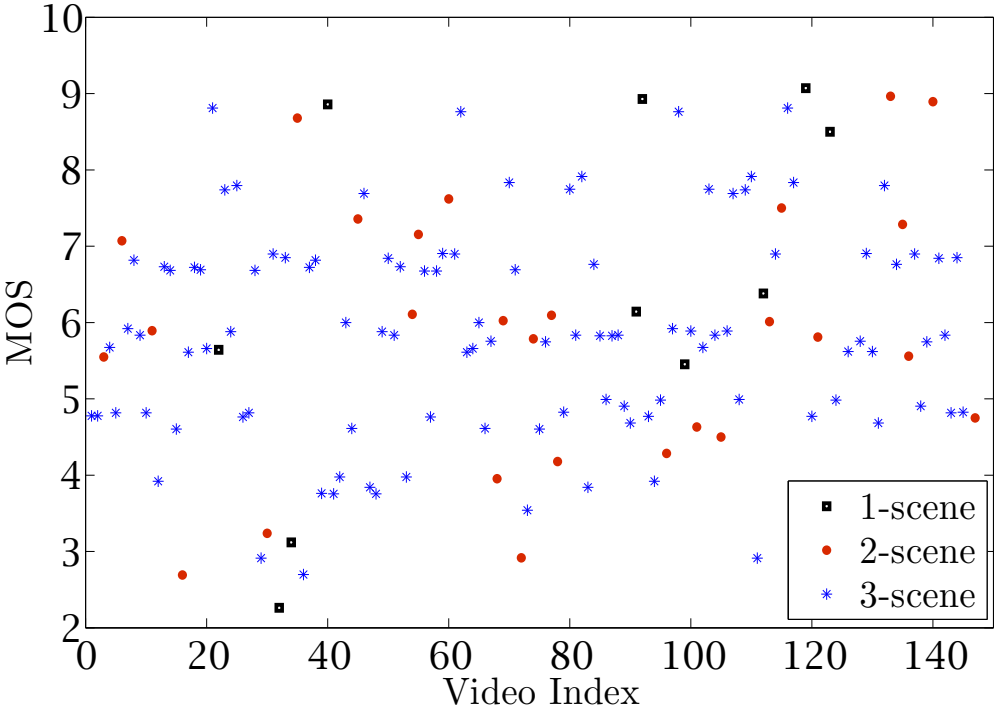


Figure 8.3: MOS scores of all video sequences.

data that had been collected, but helped us understand the data better, and also provided us with intuitive ideas that could be employed in the development of computational models that mimic human behaviors.

8.2.3 Observations

By investigating the subjective data collected and discussing with the subjects regarding their scoring strategies, we have a number of empirical observations. Although these observations are only qualitative, they provide useful insights in understanding the problem and in developing quantitative models that approximate human judgment. These observations are summarized as follows. Generally speaking, when watching a video with time-varying quality,

1. Subjects are *resistant* in updating their opinions. When there is a small quality variation between consecutive scenes, subjects tend to keep their opinions or change their opinions only slightly;

2. Subjects use *asymmetric* strategies in updating their opinions. A significant quality degradation between consecutive scenes results in a large penalty, as compared to the reward obtained by a significant quality improvement between consecutive scenes;

3. Subjects prefer *consistent* quality over time. Maintaining a "reasonable" quality for longer duration results in a small bias towards better subjective experience;

4. Subjects' judgments are not heavily influenced by the quality of the last (or the first) scene, which is in contrast to what was reported in [102]. This observation is also reflected in the numerical test results reported in Section 8.3.2.

8.3 Objective Model

8.3.1 Asymmetric Adaptation (AA) Model

Based on the analysis of the subjective data and the observations described in Section 8.2.3, here we propose a model to better account for the perceptual experience of time-varying video quality. Assume that when subjects are watching a video, they maintain their overall opinions about the video quality until quality changes in consecutive scenes are observed. We can then focus on modeling the human strategy in updating their opinions.

Let n be the number of scenes in a video sequence, q_i be the perceptual quality of the i-th scene in the sequence (i.e., the quality when the single scene is assessed), l_i be the time span of the i-th scene, and Q_i be the perceptual quality experience after the i-th scene (i.e., the quality opinion after the first i scenes are watched). The change in the quality of successive individual scenes can be calculated by

$$\Delta q_i = \begin{cases} q_i, & i = 1 \\ q_i - q_{i-1}, & i = 2, 3 \ldots n \end{cases}. \tag{8.1}$$

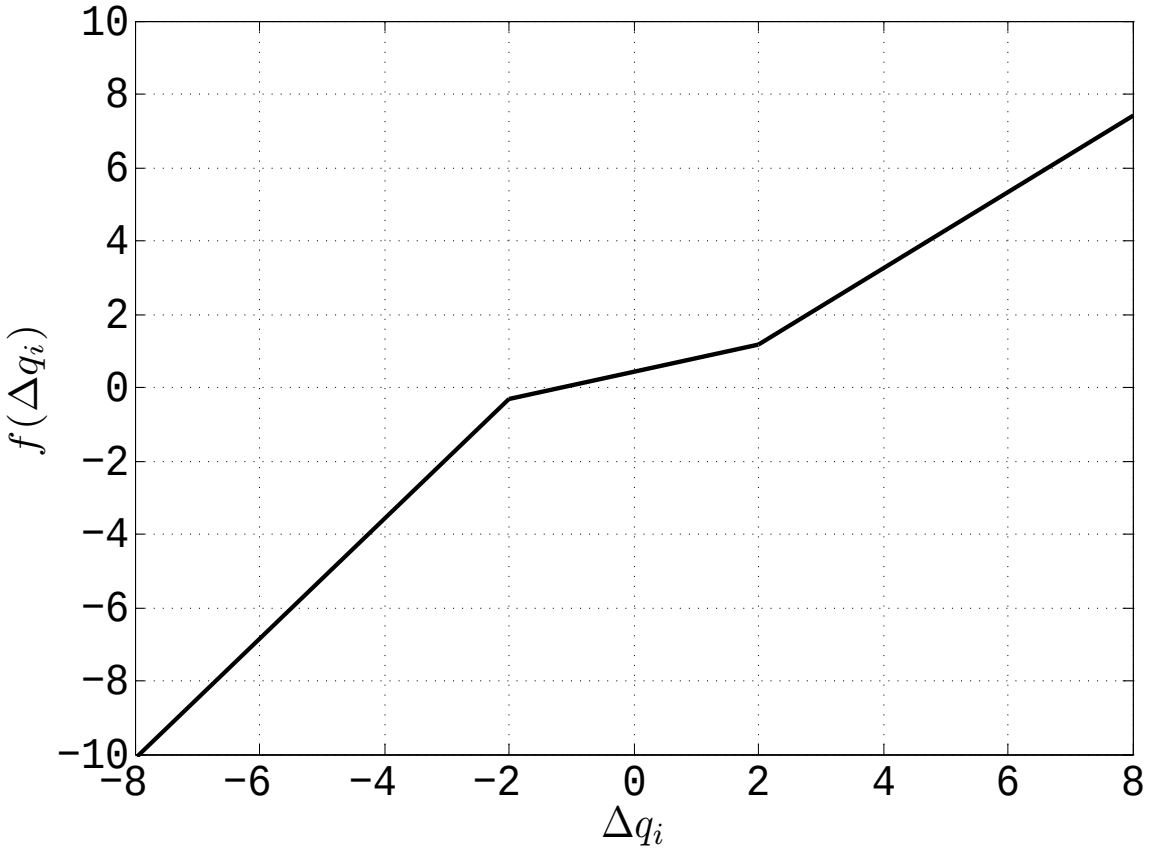


Figure 8.4: Relationship between change in quality of successive scenes and change in perceptual quality experience.

We model the quality opinion update between watching the $(i-1)$-th and the i-th scenes as

$$Q_i = \begin{cases} q_i, & i = 1 \\ \alpha_i f(\Delta q_i) + (1-\alpha_i) Q_{i-1}, & i = 2, 3 \dots n \end{cases}, \tag{8.2}$$

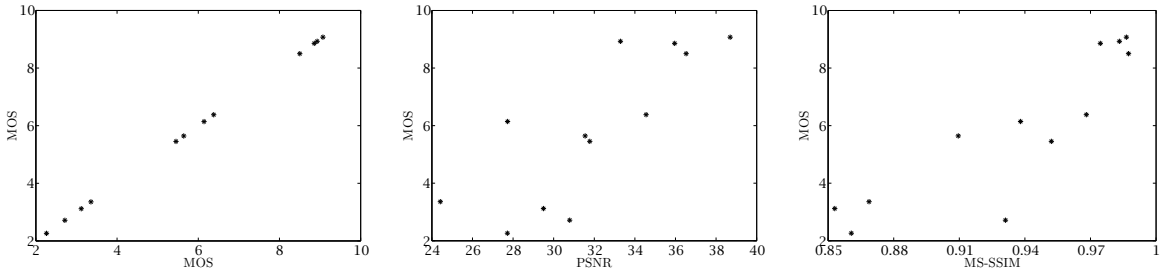
where $\alpha_i = l_i / \sum_{k=1}^{i} l_k$ controls the scale of change that decreases as time progresses, and the function f determines how subjective opinion changes as a function of Δq_i. In a simple special case, when $f(x) = x$, the model corresponds to quality averaging over time. However, the observations discussed in Section 8.2.3 suggest that f should be nonlinear. In

particular, based on Observation 1 in Section 8.2.3, f should change slowly when $|\Delta q_i|$ is small; By Observation 2, f needs to change faster with negative values of Δq_i and slower for positive values of Δq_i; By Observation 3, f should be slightly positive when Δq_i is close to 0. Combining all the desired properties, we use a piecewise linear function to approximate f, which is plotted in Figure 8.4, where the three linear pieces correspond to significantly decreasing Δq_i, small change of Δq_i, and significantly increasing Δq_i, respectively. Because of the asymmetric properties of f, we call our quality updating scheme the asymmetric adaptation (AA) model.

8.3.2 Validation

We test the proposed AA model by using it to predict the MOS value of a sequence from the MOS values of individual scenes that compose the sequence. All the MOS values are available in the subjective database described in Section 8.2. In addition to the proposed AA model, a series of other predictive models are also included for comparison. These include the Mean, Min, Max, and Median MOS values of all scenes, the MOS value of the first scene (FS) and the last scene (LS), weighted average MOS with increasing weights (W+), where $w = [\frac{1}{6}\ \frac{2}{6}\ \frac{3}{6}]$ for 3 scenes; decreasing weights (W-), where $w = [\frac{3}{6}\ \frac{2}{6}\ \frac{1}{6}]$, and distortion-based weights (DW), where $w = 1/\text{MOS}$. Correlation between the predicted and actual sequence-level MOS scores is then calculated to provide quantitative evaluation of the performance. The results are reported in Table 8.1, where due to space limit, only Kendall's rank-order correlation coefficient (KRCC) results are given, but other measures give similar results. Furthermore, Figs. 8.5(d) and 8.5(g), and Figs. 8.6(a) and 8.6(d) compare the scatter plots of the actual MOS values versus Mean- and AA-predicted MOS values for 2-scene and 3-scene sequences, respectively. It can be observed that AA provides better predictions then Mean-MOS, which is one of the best in Table 8.1 among all other pooling methods being compared.

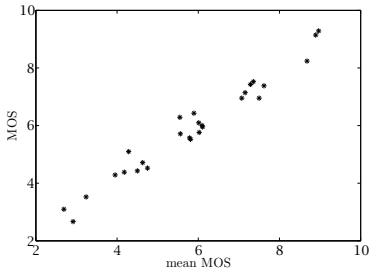
If a pooling scheme is effective at predicting sequence-level quality using the quality of each segment, then it should also be useful in improving objective VQA models in the pooling stage. We use the well-known PSNR and MS-SSIM [177] as examples to verify this. Note that the purpose here is not to find the best objective VQA approach, but to



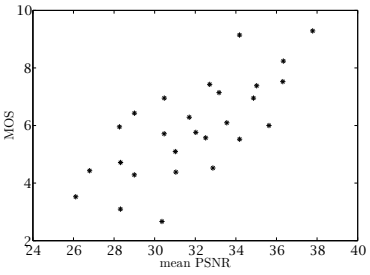
(a) KRCC = 1.0000

(b) KRCC = 0.6061

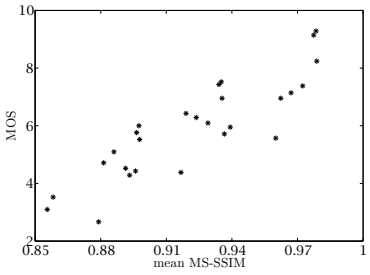
(c) KRCC = 0.7273



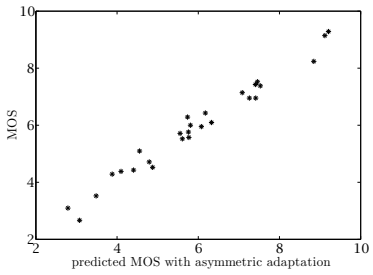
(d) KRCC = 0.8388



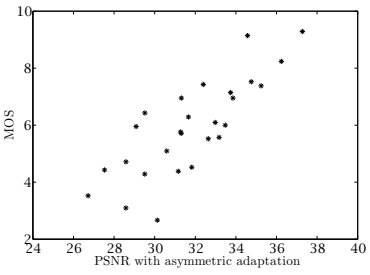
(e) KRCC = 0.6163



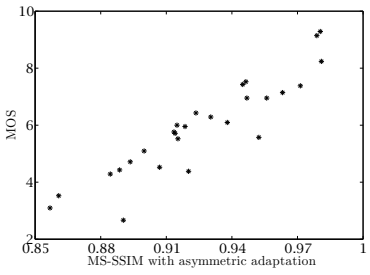
(f) KRCC = 0.7346



(g) KRCC = 0.8902

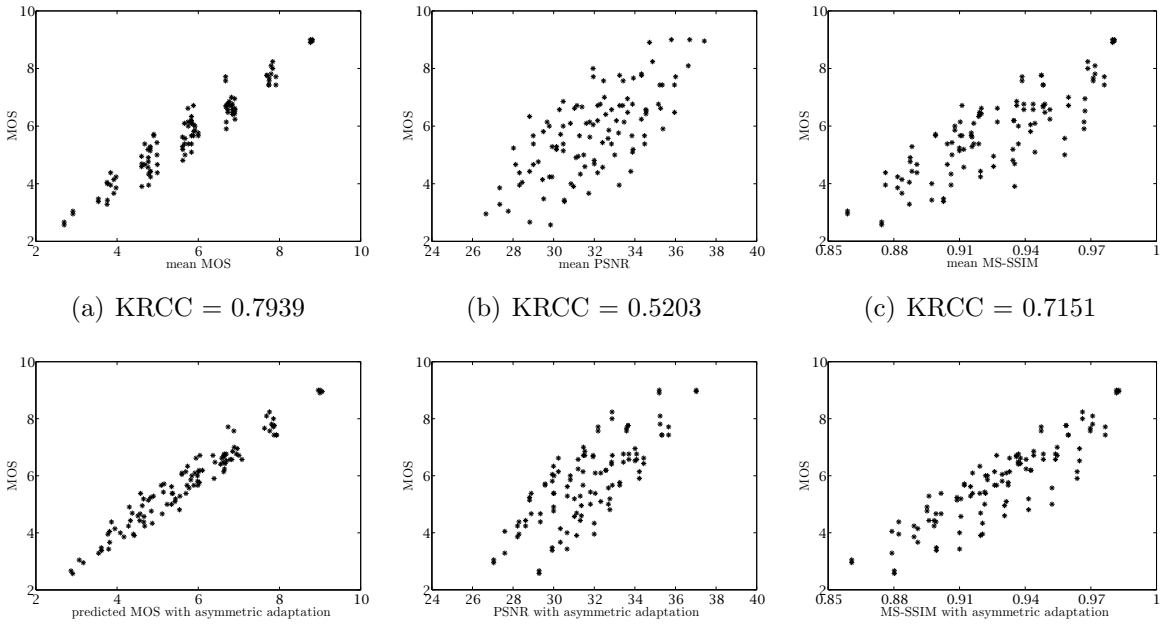


(h) KRCC = 0.6676



(i) KRCC = 0.7703

Figure 8.5: Scatter plots of sequence-level actual MOS (vertical axis) versus predicted MOS (horizontal axis) using different scene-level base quality measures and different pooling strategies. Column 1: predicted by scene-level MOS; Column 2: predicted by scene-level PSNR; Column 3: predicted by scene-level MS-SSIM. Row 1: 1-scene sequence; Rows 2 and 3: 2-scene sequence; Rows 2: Mean prediction; Rows 3: AA prediction.



(a) KRCC = 0.7939 (b) KRCC = 0.5203 (c) KRCC = 0.7151

(d) KRCC = 0.8447 (e) KRCC = 0.5660 (f) KRCC = 0.7517

Figure 8.6: Scatter plots of sequence-level actual MOS (vertical axis) versus predicted MOS (horizontal axis) using different scene-level base quality measures and different pooling strategies for 3-scene sequences. Column 1: predicted by scene-level MOS; Column 2: predicted by scene-level PSNR; Column 3: predicted by scene-level MS-SSIM. Rows 1: Mean prediction; Rows 2: AA prediction.

Table 8.1: KRCC comparison between actual MOS and predicted MOS using different base quality measures (scene-level MOS, PSNR and MS-SSIM) and different pooling strategies

Base measure	MOS			PSNR			MS-SSIM		
Sequence type	1-scene	2-scene	3-scene	1-scene	2-scene	3-scene	1-scene	2-scene	3-scene
Mean	1.0000	0.8388	0.7939	0.6061	0.6163	0.5203	0.7273	0.7346	0.7151
Min	1.0000	0.7274	0.6245	0.6061	0.5722	0.4752	0.7273	0.6477	0.5214
Max	1.0000	0.6546	0.4973	0.6061	0.5477	0.4468	0.7273	0.5928	0.4639
Median	1.0000	0.8388	0.7033	0.6061	0.6163	0.6133	0.7273	0.7346	0.6601
FS	1.0000	0.5553	0.3574	0.6061	0.4365	0.3156	0.7273	0.5078	0.3452
LS	1.0000	0.5292	0.4390	0.6061	0.4763	0.3828	0.7273	0.5075	0.4113
W+	1.0000	0.7475	0.7299	0.6061	0.6562	0.5288	0.7273	0.6733	0.6657
W-	1.0000	0.8103	0.6553	0.6061	0.5307	0.4784	0.7273	0.7247	0.6136
DW	1.0000	0.8445	0.7808	0.6061	0.6220	0.5380	0.7273	0.7232	0.7133
AA	1.0000	0.8902	0.8447	0.6061	0.6676	0.5660	0.7273	0.7703	0.7517

demonstrate the usefulness of the proposed model. The PSNR and MS-SSIM values are computed for each frame and then averaged within each scene, resulting the scene-level PSNR and MS-SSIM measures, which are used as the basis to predict the sequence-level MOS. The quantitative results are shown in Table 8.1 and the corresponding scatter plots for Mean- and AA-prediction are given in Figures 8.5 and Figures 8.6. It can be seen that the pooling schemes being tested generally behave consistently when using MOS, PSNR and MS-SSIM as the basis for scene-level quality measurement, and the proposed AA model generally outperforms the other approaches.

Chapter 9

Conclusion and Future Work

9.1 Conclusion

The goal of this thesis was to propose novel solutions for perceptually optimal visual communications. In the first section we will summarize the contributions to the scientific community that were brought forward in this thesis. In the second section we will discuss different avenues for future research. Our related publications are listed at the end of the chapter.

In Chapter 3 we present a Reduced Reference Image Quality Assessment (RRIQA) method for visual communication by SSIM estimation. The contributions are as follows:

- General-purpose Image Quality Assessment measure;
- Reduced-Reference SSIM estimation with high accuracy;
- Partial "repair" of the received distorted images using reduced-reference features.

We demonstrate in Chapter 4 that SSIM should be used for image restoration tasks so as to obtain better perceptual quality than MSE, and present our SSIM-based image restoration algorithm using sparse and redundant representations. Our contributions in this chapter are summarized below:

- Combination of SSIM with optimal sparse signal representation in the context of image restoration;
- Solution for the optimal coefficients for a sparse and redundant dictionary in a maximal SSIM sense;
- Modification of Orthogonal Matching Pursuit (OMP), keeping in view the SSIM index instead $\mathcal{L}_2$ distance;
- Estimation of the best compromise between the distorted and sparse dictionary reconstructed images for maximal SSIM.

Chapter 5 presents SSIM-Inspired Non-Local means image de-noising algorithm. The main contribution are:

- Use of SSIM as the similarity criterion for non-local means image de-noising;
- A two-stage approach for robust SSIM-estimation in the presence of noise.

Chapter 6 proposed a novel method for Rate Distortion Optimization (RDO) for video coding using SSIM, which aims to achieve optimal perceptual quality for an available rate budget. This chapter makes the following key contributions:

- We employ SSIM as the distortion measure in the proposed mode selection scheme, where both the current MB to be coded and neighboring pixels are taken into account to fully exploit the properties of SSIM;
- At the frame level, we present an adaptive Lagrange multiplier selection scheme based on a novel statistical reduced-reference SSIM model and a source-side information combined rate model;
- At the MB level, we present a Lagrange multiplier adjustment scheme, where the scale factor for each MB is determined by an information theoretical approach based on the motion information content and perceptual uncertainty of visual speed perception.

Chapter 7 presents an SSIM-inspired novel residual divisive normalization scheme for perceptual video coding. The main highlights of the chapter are as follows:

- Divisive normalization scheme to transform the DCT domain residuals which are obtained after prediction to a perceptually uniform space based on a DCT domain SSIM index;
- Following the divisive normalization scheme, we define a new distortion model and propose a novel perceptual RDO scheme for mode selection;
- In the divisive normalized domain, we propose a frame-level quantization matrix selection approach so that the normalized coefficients of different frequencies share the same R-D relationship;
- Adaptive Quantization method inspired residual divisive normalization that generates bit-stream compatible with standard decoders.

Chapter 8 presents a study that helped us better understand perceptual experience of time-varying video quality in more realistic scenarios. Our contributions in this chapter are summarized are:

- We created a video database and carried out subjective test that are designed to directly examine the perceptual experience of time-varying video quality;
- Simple models that pool segment-level quality are limited in predicting the overall human quality assessment of the combined video;
- The proposed asymmetric adaptation (AA) model leads to improved performance of both subjective and objective quality assessment approaches;
- The scheme has the potential to be employed in the optimization of modern video compression technologies and in the optimal allocation of network resources.

9.2 Future Work

The research work presented in this thesis aims to convince the readers that optimization of image and video processing algorithms based on perceptual image and video quality assessment methods yields fruitful results. As this is just the beginning of this exciting direction of research, we expect that many researchers will realize the potential of using perceptual quality assessment measures for image and video processing applications. Some of the possible directions to continue this research work are mentioned as follows.

9.2.1 Video Processing based on Perceptual Video Quality Assessment Methods

PSNR and SSIM are mainly IQA methods as they do not consider inter-frame interactions and as a result fail to capture specific temporal artifacts such as flickering and ghosting in compressed video [197]. There is a strong need to develop novel approaches for video quality assessment that possess the following properties, which are critical for their use in the optimization of video processing algorithms/solutions.

- High correlation with subjective video quality scores
- Low computational complexity so that the algorithm is practically usable for video processing
- Accurate local quality prediction that can help determine varying local quality level based on content
- Good mathematical properties that can help in solving optimization problems i.e. a valid distance metric that satisfies convexity, differentiability, symmetry, etc.

9.2.2 Perceptual Video Compression

Video quality is generally divided into spatial quality and temporal quality of a video. SSIM-Inspired divisive normalization based video compression technique, proposed in Chap-

ter 7, provides an intuitive method to convert transform residuals to a perceptually uniform space by adaptively adjusting the quantizer of each block based on perceptual importance. The framework can easily be scaled to Frame-level and GOP-level using

$$\Delta QP = \Delta QP_b + \Delta QP_f + \Delta QP_g, \tag{9.1}$$

where ΔQP_b, ΔQP_f, ΔQP_g represent the change in QP at the block, frame, and GOP level, respectively.

Perceptual cues such as the variation and complexity of visual content, amount of motion between consecutive frames, etc. can prove useful in the calculation of ΔQP_f. The subjective study presented in Chapter 8 can be applied to adjust the GOP-level quantizer in order to provide an overall better quality of service to viewers.

9.2.3 No-Reference Video Quality Assessment

Existing FR image and video quality assessment measures cannot be used to evaluate the perceptual quality of a video at the receiver or at a network node because the reference video is available only at the transmitter. Therefore, NR image and video quality assessment techniques are highly desirable in visual communications, specially for the purpose of QOS monitoring. Network service providers need to monitor quality degradation in real-time, in order to optimize network resource allocation and achieve the required quality of service within certain cost constraints. General purpose NR-IQA is still in its infancy stage. Also, objective video quality assessment is a greater challenge than objective image quality assessment. Considering that the most of the data transmitted over the networks suffer mainly from lossy compression, solving the problem of NR-VQA of a compressed video would be an excellent start.

9.2.4 SSIM-based Dictionary Learning Algorithm for Sparse Representations and Image Restoration

Sparse representation based algorithms have become a key research topic in signal and image processing with numerous applications, e.g., image de-noising, restoration, compression and more. Sparse representations based algorithms represent most or all information contained in a signal, with a linear combination of a small number of elements or atoms adequately chosen from an overcomplete or redundant bases or dictionary. Formally, such a dictionary is a collection of atoms whose number is much larger than the dimension of the signal space, i.e., than the number of components of the vector representing the signal. Any signal admits then an infinite number of representations and the sparsest such representation has interesting properties for a number of image processing tasks. The crucial question in sparse representations is the choice of dictionary. One can use a variety of predefined bases like DCT, wavelets, or others. However, the sparsity of the representation depends on how well the dictionary is adapted to the data at hand. The problem of dictionary learning, that goes beyond the concatenation of a few off-the-shelf bases, has therefore become a key issue for further progress in this area. Thus, researchers have developed various learning schemes in order to provide adapted dictionaries for the data considered. The popular dictionary learning algorithms include the K-SVD, the Method of Optimal Directions (MOD) and so on which use MSE as the IQA measure. Significant improvement in visual quality can be expected by improving the dictionary learning process based on SSIM, as dictionary encapsulates in itself the prior knowledge about the image to be restored. An SSIM-optimal dictionary will capture structures contained in the image in a better way and the restoration task will result into sharper output image. Further improvement is also expected when some of the advanced mathematical properties of SSIM and normalized metrics are incorporated into the optimization framework.

9.2.5 SSIM-motivated non-local sparse image restoration

State-of-the-art image de-noising performance can be achieved when K-SVD (dictionary learning) and Non-Local image de-noising methods are merged. The idea of joint sparsity

in Leaning Simultaneous Sparsity Coding (LSSC) [87] and image clustering in Clustering-based Sparse Representation (CSR) [42] takes advantage of such an approach. We believe it is fruitful to further explore such connections in the future while keeping SSIM in mind as the optimization criterion.

Publications

Patents

1. Z. Wang and **A. Rehman**, "*Method and system for structural similarity based video coding*", International Patent application, PCT/CA2012/000519, filed 2012.

2. Z. Wang and **A. Rehman**, "*Perceptual high efficiency video coding based on structural similarity quality measure*", US patent application, filed 2011.

Journal Papers

3. **A. Rehman**, Y. Gao, J. Wang, and Z. Wang, "*Image classification based on complex wavelet structural similarity*", Signal Processing: Image Communication, special issue on Biologically Inspired Approaches for Visual Information Processing and Analysis, accepted, to appear 2013.

4. **A. Rehman**, and Z. Wang, "*Reduced-Reference Image Quality Assessment by Structural Similarity Estimation*", IEEE Transactions on Image Processing (TIP),vol. 21, no. 8, pp. 3378-3389, Aug. 2012.

5. S. Wang, **A. Rehman**, Z. Wang, S. Ma, and W. Gao., "*Perceptual Video Coding Based on SSIM-Inspired Divisive Normalization*", IEEE Transactions on Image Processing (TIP),vol. 22, no. 4, pp. 1418-1429, Apr. 2013.

6. **A. Rehman**, M. Rostami, Z. Wang, D. Brunet, and E. R. Vrscay, "*SSIM-inspired image restoration using sparse representation*", EURASIP Journal on Advances in Signal Processing, vol. 2012:16, Jan, 2012.

7. S. Wang, **A. Rehman**, Z. Wang, S. Ma, and W. Gao., "*SSIM-Motivated Rate Distortion Optimization for Video Coding*", IEEE Transactions on Circuits System and Video Technology (CSVT), vol. 22, pp. 516-529, Apr. 2012.

Conference Papers

8. **A. Rehman** and Z. Wang, "*Perceptual experience of time-varying video quality*", International Workshop on Quality of Multimedia Experience (QoMEX), June 2013.

9. K. Zeng, **A. Rehman**, J. Wang, and Z. Wang, "*From H.264 to HEVC: Coding gain predicted by objective video quality assessment models*", International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, AZ, Jan.-Feb. 2013.

10. **A. Rehman** and Z. Wang, "*SSIM-Inspired Perceptual Video Coding for HEVC*", IEEE International Conference on Multimedia and Expo (ICME), pp. 497 - 502, July 2012.

11. **A. Rehman** and Z. Wang, "*Reduced-Reference SSIM Estimation*", IEEE International Conference on Image Processing (ICIP), Sept. 26-29, 2010.

12. **A. Rehman**, Z. Wang, D. Brunet, and E. R. Vrscay., "*SSIM-inspired image denoising using sparse representations*", IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP) , pp. 1121 - 1124, May 2011.

13. S. Wang, **A. Rehman**, Z. Wang, S. Ma, and W. Gao., "*Rate-SSIM optimization for video coding*", IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP) , pp. 833 - 836, May 2011.

14. **A. Rehman** and Z. Wang, "*SSIM-based Non-local Means Image Denoising*", International Conference on Image Processing (ICIP), pp. 221 - 224, Sep. 2011.

15. S. Wang, **A. Rehman**, Z. Wang, S. Ma, and W. Gao., "*SSIM-inspired divisive normalization for perceptual video coding*", International Conference on Image Processing (ICIP) , pp. 1693 - 1696, May 2011.

16. Y. Gao, **A. Rehman**, and Z. Wang, "*CW-SSIM Based Image Classification*", International Conference on Image Processing (ICIP), pp. 1273 - 1276, Sep. 2011.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. 64, 73, 74

[2] Y. Altunbasak and N. Kamaci. An analysis of the DCT coefficient distribution with the H.264 video coder. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 3:III–177–80., 2004. 95

[3] B. Aswathappa and K. R. Rao. Rate-distortion optimization using structural information in H.264 strictly intra-frame encoder. *South Eastern Symposium on System Theory*, pages 367–370, 2010. 33

[4] M. Barkowsky, B. Eskofier, R. Bitto, J. Bialkowski, and A. Kaup. Perceptually motivated spatial and temporal integration of pixel based video quality measures. In *Welcome to Mobile Content Quality of Experience*, pages 1–7, 2007. 36

[5] H. B. Barlow. Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, 1961. 28

[6] G. Bjontegaard. Calculation of average PSNR difference between RD curves. *Proc. ITU-T Q.6/SG16 VCEG 13th Meeting, Austin, TX*, Apr. 2001. 116, 150, 161

[7] M.J. Black, G. Sapiro, D.H. Marimont, and D. Heeger. Robust anisotropic diffusion. *IEEE Transactions on Image Processing*, 7(3):421–432, 1998. 35

[8] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 47

[9] A. C. Brooks, X. Zhao, and T. N. Pappas. Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions. *IEEE Transactions on Image Processing*, 17:121–132, Aug. 2008. 33

[10] D. Brunet, E. R. Vrscay, and Z. Wang. Structural similarity-based approximation of signals and images using orthogonal bases. In M. Kamel and A. Campilho, editors, *Proc. Int. Conf. on Image Analysis and Recognition*, volume 6111 of *LNCS*, pages 11–22. Springer, Heidelberg, 2010. 19, 66, 67, 68, 135

[11] D. Brunet, E. R. Vrscay, and Z. Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing,*, 21(4):1488–1499, 2012. 19, 23

[12] A. Buades, B. Coll, and J.-M. Morel. Non-local means denoising implementation. http://ipol.im/pub/algo/bcm_non_local_means_denoising. 86, 87

[13] A. Buades, B. Coll, and J. M. Morel. Denoising image sequences does not require motion estimation. In *IEEE Conf. on Advanced Video and Signal Based Surveillance*, pages 70–74, September 2005. 35, 82, 85, 86, 87

[14] A. Buades, B. Coll, and J. M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, 4(2):490–530, 2005. 35, 83, 85, 87

[15] A. Buades, B. Coll, and J. M. Morel. Nonlocal image and movie denoising. *Intl. J. Comp. Vis.*, 76(2), 2008. 35

[16] D. Burr. Sensitivity to spatial phase. *Vision Research*, 20(5):391 – 396, 1980. 10

[17] E. Candes and D. Donoho. Recovering Edges in Ill-Posed Inverse Problems: Optimality of Curvelet Frames. *Ann. Statist*, 30(3):784–842, Jun 2002. 34

[18] E. J. Candés, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Info. Theory*, 52(2):489–509, 2006. 64

[19] M. Carnec, P. Le Callet, and D. Barba. An image quality assessment method based on perception of structjjral information. In *Proc. IEEE Int. Conf. Image Proc.*, volume 3, pages 185–188, September 2003. 28

[20] M. Carnec, P. Le Callet, and D. Barba. Visual features for image quality assessment with reduced reference. In *Proc. IEEE Int. Conf. Image Proc.*, volume 1, pages 421–424, September 2005. 28

[21] A. Chambolle, R. DeVore, Y. Lee, and B. Lucier. Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal Through Wavelet Shrinkage. *IEEE Tran. Image Proc.*, 7(3):319–333, 1998. 34

[22] D. Chandler and S. Hemami. Dynamic contrast-based quantization for lossy wavelet image compression. *IEEE Trans. on Image Processing*, 14:397–410, 2005. 32

[23] D. M. Chandler and S. S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16:2284–2298, Sep. 2007. 21, 27, 29, 48, 54

[24] S. Grace Chang, Bin Yu, and Martin Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9(9):1532–1546, 2000. 35

[25] S. Channappayya, A. C. Bovik, and Jr. R. W. Heathh. Rate bounds on SSIM index of quantized images. *IEEE Trans. on Image Processing*, 17:1624–1639, Sep. 2008. 97, 137

[26] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R.W. Heath. Design of linear equalizers optimized for the structural similarity index. *IEEE Transactions on Image Processing*, 17(6):857 –872, jun. 2008. 19

[27] H. H. Chen, Y. H. Huang, P. Y. Su, and T. S. Ou. Improving video coding quality by perceptual rate-distortion optimization. *Proc. IEEE Int. Conf. Multimedia Exp*, pages 1287–1292, Jul. 2010. 33

[28] J. Chen, J. Zheng, and Y. He. Macroblock-level adaptive frequency weighting for perceptual video coding. *IEEE Trans. on Consumer Electronics*, 53:775–781, May. 2007. 31, 32

[29] L. Chen and I. Garbacea. Adaptive Lambda estimation in Lagrangian rate-distortion optimization for video coding. *Proc. SPIE*, 6077:60772B 1–8, 2006. 32

[30] S. Chen, D. Donoho, and M. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999. 34

[31] Z. Chen and C. Guillemot. Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model. *IEEE Trans. on Circuits and Systems for Video Technology*, 20:806–819, Jun. 2010. 31, 32

[32] H. Chipman, E. Kolaczyk, and R. Mcculloch. Adaptive Bayesian wavelet shrinkage. *Journal of The American Statistical Association*, 92:1413–1421, 1997. 10

[33] K. Chono, Yao-Chung Lin, D. Varodayan, Y. Miyamoto, and B. Girod. Reduced-reference image quality assessment using distributed source coding. In *IEEE International Conference on Multimedia and Expo*, pages 609 –612, apr. 2008. 28

[34] CISCO. Cisco visual networking index: Forecast and methodology, 2010 − 2015. http://www.cisco.com/en/US/netsol/ns827/networking_solutions_white_papers_list.html. 1

[35] CISCO. Cisco visual networking index: Global mobile data traffic forecast update, 2010 − 2015. http://www.cisco.com/en/US/netsol/ns827/networking_solutions_sub_solution.html. 1, 2

[36] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991. 43

[37] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Trans. Signal Processing*, 46(4):886–902, April 1998. 10

[38] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16:2080 – 2095, 2007. 34, 83

[39] S. J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. *Digital Images and Human Vision*, pages 179–206, 1993. 11

[40] J. L. Devore and N. R. Farnum. *Applied Statistics for Engineers and Scientists*. New York, Duxbury, 1999. 100

[41] M Do and M. Vetterli. Framing pyramids. *IEEE Transactions on Signal Processing*, 51(9):2329–2342, 2003. 34

[42] W. Dong, X. Li, L. Zhang, and G. Shi. Sparsity-based image denoising vis dictionary learning and structural clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 198

[43] W. Dong, D. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011. 79

[44] D. Donoho. Wedgelets: nearly-minimax estimation of edges. *Ann. Statist*, 27:859–897, 1999. 34

[45] D. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, August 2002. 34

[46] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 64

[47] D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. 34

[48] A. A. Efros and T. K. Leung. Texture synthesis by non-parameteric sampling. In *Proc. Int'l Conference on Computer Vision*, Corfu, 1999. 83

[49] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Processing*, 15(12):3736–3745, December 2006. 34, 63, 64, 65, 74, 77

[50] D. Field. What Is the Goal of Sensory Coding? *Neural Computation*, 6(4):559–601, July 1994. 10

[51] J. Foley. Human luminance pattern mechanisms: masking experiments require a new model. *J. Opt. Soc. Amer.*, 11(6):1710–1719, 1994. 20, 42, 68, 135

[52] J. Foley. Human luminance pattern mechanisms: Masking experiments require a new model. *J. Opt. Soc. Amer.*, 11:1710–1719, 1994. 138

[53] Y. Gao, A. Rehman, and Z. Wang. CW-SSIM based image classification. In *IEEE International Conference on Image Processing ICIP*, pages 1249–1252, Brussels, Belgium, September 2011. 19

[54] B. Girod. What's wrong with mean-squared error. In A. B. Watson, editor, *Digital Images and Human Vision*, pages 207–220. the MIT press, 1993. 27, 31, 135

[55] H. Gish and J. Pierce. Asymptotically efficient quantizing. *IEEE Trans. on Information Theory*, 14:676–683, Oct. 1968. 109

[56] I. P. Gunawan and M. Ghanbari. Reduced reference picture quality estimation by using local harmonic amplitude information. In *Proc. London Communicationa Symposium*, pages 137–140, September 2003. 28

[57] ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 Part 10). JVT, advanced video coding (AVC), 2004. 92

[58] Z. He and S. Mitra. Optimum bit allocation and accurate rate control for video coding via rho-domain source modeling. *IEEE Trans. on Circuits and Systems for Video Technology*, 12:840–849, Oct. 2002. 32

[59] D. J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2):181–197, 1992. 20, 42, 68, 135

[60] Y. Horita, K. Shibata, Y. Kawayoke, and Z. M. Parvez. MICT image quality evaluation database. http://mict.eng.u-toyama.ac.jp/mictdb. 21, 30, 48, 54

[61] T. Hosfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler. The memory effect and its implications on web qoe modeling. In *23rd International Teletraffic Congress (ITC)*, pages 103–110, Sep. 2011. 36

[62] Y. H. Huang, T. S. Ou, and H. H. Chen. Perceptual-based coding mode decision. *Proc. IEEE Int. Symp. Circuits Syst.*, pages 393–396, May. 2010. 33

[63] Y. H. Huang, T. S. Ou, P. Y. Su, and H. H. Chen. Perceptual rate-distortion optimization using structural similarity index as quality metric. *IEEE Trans. on Circuits and Systems for Video Technology*, 20:1614–1624, Nov. 2010. 33, 127, 129

[64] ITU-T Recommendation BT.500-13. Methodology for the subjective assessment of the quality of television pictures. Technical report, International Telecommunication Union, Geneva, Switzerland, January 2012. 153

[65] ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications. Technical report, International Telecommunication Union, Geneva, Switzerland, April 2008. 183, 184

[66] M. Jiang and N. Ling. On Lagrange multiplier and quantizer adjustment for H.264 frame-layer video rate control. *IEEE Trans. on Circuits and Systems for Video Technology*, 16:663–669, May 2006. 32, 103, 120

[67] Joint video team (JVT) reference software [online]. http://iphome.hhi.de/suehring/tml/download/old_jm. 19, 112, 150

[68] M. Karczewicz, Y. Ye, and I. Chong. Rate distortion optimized quantization. *VCEG-AH21*, Jan. 2008. 31, 33, 130

[69] 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*, pages 71–74, October 2003. 28

[70] D. Kwon, M. Shen, and C. Kuo. Rate control for H.264 video with enhanced rate and distortion models. *IEEE Trans. on Circuits and Systems for Video Technology*, 17:517–529, May 2007. 99

[71] E. Y. Lam and J. W. Goodman. A mathematical analysis of the dct coefficient distributions for images. *IEEE Trans. on Image Processing*, 9(10):1661–1666, Oct. 2000. 146

[72] E. C. Larson and D. M. Chandler. Categorical image quality (CSIQ) database. http://vision.okstate.edu/csiq. 21, 30, 48, 54

[73] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006:1–21, Jan.-Mar. 2010. 27, 139

[74] P. Le Callet and F. Autrusseau. Subjective quality assessment IRCCyN/IVC database, 2005. http://www.irccyn.ec-nantes.fr/ivcdb/. 21, 30, 48, 53

[75] J. Lee. Rate-distortion optimization of parameterized quantization matrix for MPEG-2 encoding. *International Conference on Image Processing*, 2:383–386, Oct. 1998. 33

[76] Q. Li and Z. Wang. Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE Journal on Selected Topics in Signal Processing*, 3(2):202–211, 2009. 20, 28, 43, 52, 53, 54, 55, 56, 57, 68, 135

[77] X. Li. Blind image quality assessment. In *Proc. IEEE Int. Conf. Image Proc.*, volume 1, pages 449–452, September 2002. 27

[78] X. Li. Collective sensing: a fixed-point approach in the metric space. In *Visual Communications and Image Processing*, volume 7744, Huangshan, China, 2010. SPIE. 83

[79] X. Li, N. Oertel, A. Hutter, and A. Kaup. Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding. *IEEE Trans. on Circuits and Systems for Video Technology*, 19:193–205, Feb. 2009. 32, 95, 98, 99, 103, 120, 142, 143

[80] J. Lubin. A visual discrimination model for imaging system design and evaluation. *Vision Models for Target Detection and Recognition*, pages 245–283, 1995. 11

[81] S. Lyu and E. P. Simoncelli. Statistically and perceptually motivated nonlinear image representation. *Proc. SPIE Conf. Human Vision Electron. Imaging XII*, 6492:649207–1–649207–15, Jan. 2007. 19, 67, 135

[82] S. Lyu and E. P. Simoncelli. Reducing statistical dependencies in natural signals using radial Gaussianization. In *Adv. Neural Information Processing Systems (NIPS*08)*, volume 21, pages 1009–1016. MIT Press, May 2009. 10

[83] L. Ma, S. Li, F. Zhang, and K. Ngan. Reduced-reference image quality assessment using reorganized dct-based image representation. *IEEE Transactions on Multimedia*, 13(4):824–829, aug. 2011. 57

[84] Z. Y. Mai, C. L. Yang, K. Z. Kuang, and L. M. Po. A novel motion estimation method based on structural similarity for H.264 inter prediction. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:913–916, 2006. 19, 33

[85] Z. Y. Mai, C. L. Yang, L. M. Po, and S. L. Xie. A new rate-distortion optimization using structural information in H.264 I-frame encoder. *Proc. ACIVS*, pages 435–441, 2005. 33

[86] Z. Y. Mai, C. L. Yang, and S. L. Xie. Improved best prediction mode(s) selection methods based on structural similarity in H.264 I-frame encoder. *IEEE International Conference on Systems, Man and Cybernetics*, pages 2673–2678, 2005. 33

[87] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *ICCV09*, pages 2272–2279, 2009. 34, 198

[88] J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *Multiscale Modeling & Simulation*, 7(1):214–241, 2008. 64

[89] S. Mallat. *A Wavelet Tour of Signal Processing*. Wavelet Tour of Signal Processing. Elsevier Science, 1999. 18

[90] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli. Non-linear image representation for efficient perceptual coding. *IEEE Trans. on Image Processing*, 15:68–80, Jan. 2006. 20, 58, 68, 135, 137

[91] J. Malo, J. Gutierrez, I. Epifanio, F. Ferri, and J. M. Artigas. Perceptual feedback in multigrid motion estimation using an improved dct quantization. *IEEE Trans. on Image Processing*, 10:1411–1427, Oct. 2001. 31

[92] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. Perceptual Blur and Ringing Metrics: Application to JPEG2000. *Signal Processing : Image Communication*, 19(2):163–172, 2004. 27

[93] M. A. Masry and S. S. Hemami. A metric for continuous quality evaluation of compressed video with severe distortions. *Signal Processing: Image Comm*, 19:133–146, 2004. 36

[94] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. Veciana. Video quality assessment on mobile devices: Subjective,behavioral and objective studies. *Journal Selected Topics Signal Processing*, 6(6):652–671, 2012. 36

[95] P. Moulin and J. Liu. Analysis of Multiresolution Image Denoising Schemes Using Generalized-Gaussian and Complexity Priors. *IEEE Trans. Info. Theory*, 45:909–919, 1998. 34

[96] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42(5):577–685, 1989. 10

[97] A. Ninassi, P. Le Callet, and F. Autrusseau. Pseudo no reference image quality metric using perceptual data hiding. *Human Vision and Electronic Imaging XI*, 6057(1):60570G, 2006. 30, 48, 53

[98] T. Ou, Y. Huang, and H. Chen. A perceptual-based approach to bit allocation for H.264 encoder. *SPIE Visual Communications and Image Processing*, Jul. 2010. 19, 33

[99] F. Pan, Y. Sun, Z. Lu, and A. Kassim. Complexity-based rate distortion optimization with perceptual tuning for scalable video coding. *International Conference on Image Processing*, 2005. 33

[100] T. N. Pappas, T. A. Michel, and R. O. Hinds. Supra-threshold perceptual image coding. *IEEE International Conference on Image Processing (ICIP)*, pages 237–240, 1996. 32

[101] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. *Twenty Seventh Asilomar Conference on Signals, Systems and Computers*, 1:40–44, 1993. 34, 65, 66, 70

[102] D. E. Pearson. Viewer response to time-varying video quality. *Proc. SPIE Human Vision and Electronic Imaging*, 3299(1):16–25, 1998. 36, 186

[103] G. Piella and H. Heijmans. A new quality metric for image fusion. In *IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 173 – 176, sept. 2003. 19

[104] M. H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, 50(3):312–322, 2004. 13, 23, 24

[105] N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, and V. Lukin. Metrics performance comparison for color image database. In *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, USA, January 2009. 52

[106] N. Ponomarenko and K. Egiazarian. Tampere image database 2008 TID2008. http://www.ponomarenko.info/tid2008. 21, 30, 48, 53

[107] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, 2009. 30, 48, 53

[108] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *Intl. J. Comp. Vis.*, 40(1):49–71, December 2000. 58

[109] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Processing*, 12(11):1338–1351, November 2003. 20, 68, 135

[110] M. Protter and M. Elad. Image sequence denoising via sparse and redundant representations. *IEEE Transactions on Image Processing*, 18(1):27–35, 2009. 64

[111] C. A. Prraga, T. Troscianko, and D. J. Tolhurst. The human visual system is optimised for processing the spatial information in natural visual images. *Current Biology*, 10(1):35 – 38, 2000. 10

[112] A. Raake. Short- and long-term packet loss behavior: Towards speech quality prediction for arbitrary loss distributions. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1957–1968, 2006. 36

[113] A. Rehman, Y. Gao, J. Wang, and Z. Wang. Image classification based on complex wavelet structural similarity. *Signal Processing: Image Communication*, 2013. 19

[114] A. Rehman, M. Rostami, Z. Wang, D. Brunet, and E. Vrscay. SSIM-inspired image restoration using sparse representation. *EURASIP Journal on Advances in Signal Processing*, 2012(1):16, 2012. 19

[115] A. Rehman and Z. Wang. Reduced-reference SSIM estimation. *International Conference on Image Processing*, pages 289–292, Sep. 2010. 20, 68

[116] A. Rehman and Z. Wang. SSIM-based non-local means image denoising. In *IEEE International Conference on Image Processing (ICIP 11)*, pages 217–220, Brussels, Belgium, September 2011. 19

[117] A. Rehman and Z. Wang. Reduced-reference image quality assessment by structural similarity estimation. *IEEE Transactions on Image Processing*, 21(8):3378–3389, 2012. 95, 135

[118] A. Rehman and Z. Wang. Ssim-inspired perceptual video coding for hevc. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 497–502, 2012. 20, 68, 135

[119] A. Rehman, Z. Wang, D. Brunet, and E. R. Vrscay. SSIM-inspired image denoising using sparse representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1121 –1124, may 2011. 19

[120] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60:259–268, November 1992. 35, 63

[121] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE Trans. Image Processing*, 18(11):2385–2401, November 2009. 18, 19

[122] A. Savitzky and M. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36:1627–1639, 1964. 63

[123] O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature: Neuroscience*, 4(8):819–825, August 2001. 135

[124] K. Seshadrinathan and A. C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2):335–350, February 2010. 23, 24

[125] K. Seshadrinathan and A. C. Bovik. Temporal hysteresis model of time varying subjective video quality. In *IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1153–1156, Prague, Czech Republic, May 2011. 36

[126] H. R. Sheikh, A. C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Trans. Image Processing*, 14(11):1918–1927, November 2005. 27

[127] H. R. Sheikh, A. C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Processing*, 14(12):2117–2128, December 2005. 11, 27

[128] H. R. Sheikh, M. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, November 2006. 52, 56

[129] 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*, 15(11):3440–3451, November 2006. 51, 52

[130] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack. Image and video quality assessment research at LIVE. http://live.ece.utexas.edu/research/quality/. 21, 29, 46, 48, 53

[131] E. Simoncelli. Modeling the joint statistics of images in the wavelet domain. In *Proc SPIE, 44th Annual Meeting*, volume 3813, pages 188–195, Denver, July 1999. 10

[132] E. P. Simoncelli and E. H. Adelson. Noise removal via Bayesian wavelet coring. In *Third Int'l Conf on Image Proc*, volume I, pages 379–382, Lausanne, September 1996. IEEE Sig Proc Society. 34

[133] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Trans. Information Theory*, 38:587–607, 1992. 34

[134] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Trans. Information Theory*, 38(2 pt II):587–607, 1992. 47

[135] E P Simoncelli and D J Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761, March 1998. 20, 42, 68, 135, 138

[136] E. P. Simoncelli and B. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, May 2001. 28, 41, 105

[137] A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9:578–585, 2006. 105

[138] P. Y. Su, Y. H. Huang, T. S. Ou, and H. H. Chen. Predictive lagrange multiplier selection for perceptual-based rate-distortion optimization. *Proc. 5th Int. Workshop Video Process. Qual. Metrics Consumer Electron.*, Jan. 2010. 33

[139] G. Sullivan, J. Ohm, W. Han, and T. Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Trans. Circuits Syst. Video Techn.*, 22(12):1649–1668, 2012. 183

[140] G. J. Sullivan and T. Wiegand. Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, 15:74–90, Nov. 1998. 26, 31, 92, 109

[141] C. Sun, H.-J. Wang, and H. Li. Macroblock-level rate-distortion optimization with perceptual adjustment for video coding. *Proc. IEEE DCC*, page 546, 2008. 32

[142] J. Sun, W. Gao, D. Zhao, and Q. Huang. Statistical model, analysis and approximation of rate-distortion function in MPEG-4 FGS videos. *IEEE Trans. on Circuits and Systems for Video Technology*, 16:535–539, Apr. 2006. 95

[143] T. Suzuki, P. Kuhn, and Y. Yagasaki. Quantization tools for high quality video. *Joint Video Team of ISO/IEC MPEG and ITU-T VCEG JVT-B067*, Jan. 2002. 31, 32

[144] T. Suzuki, K. Sato, and Y. Yagasaki. Weighting matrix for jvt codec. *Joint Video Team of ISO/IEC MPEG & ITU-T VCEG JVT-C053*, May. 2002. 31, 32

[145] M. Tagliasacchi, G. Valenzise, M. Naccari, and S. Tubaro. A reduced-reference structural similarity approximation for videos corrupted by channel errors. *Multimedia Tools and Applications*, 48:471–492, 2010. 28, 42

[146] K. T. Tan, M. Ghanbari, and D. E. Pearson. An objective measurement tool for MPEG video quality. *IEEE Trans. Signal Proc.*, 70(3):279–294, November 1998. 36

[147] C.-W. Tang. Spatial temporal visual considerations for effcient video coding. *IEEE Trans. on Multimedia*, 9(2):231–238, Jan. 2007. 32

[148] C.-W. Tang, C.-H. Chen, Y.-H. Yu, and C.-J. Tsai. Visual sensitivity guided bit allocation for video coding. *IEEE Trans. on Multimedia*, 8(1):11–18, Feb. 2006. 32

[149] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problem*. V. H. Winston, Washington, D.C., 1977. 63

[150] Toshiba. Adaptive quantization matrix selection. In *ITU WP3/SC16 Delayed contribution 267, T05-SG16-060403-D-0266*, Geneva, Apr. 2006. 31, 32, 149

[151] T. Vlachos. Simple method for estimation of global motion paramenters using spares translational motion vector fields. *Electronics letters*, 34:90–91, 1998. 108

[152] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment. Technical report, available at http://www.vqeg.org/, Apr 2000. 52

[153] M. J. Wainwright. Visual adaptation as optimal information transmission. *Vision Research*, 39:3960–3974, 1999. 42, 67

[154] M. J. Wainwright and E. P. Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. *Adv. Neural Information Processing Systems*, 12:855–861, 2000. 19, 43, 135, 137

[155] M. Wang and B. Yan. Lagrangian multiplier based joint three-layer rate control for H.264/AVC. *IEEE Signal Process. Lett.*, 16:679–682, Aug. 2009. 32, 107

[156] S. Wang, S. Ma, and W. Gao. SSIM based perceptual distortion rate optimization coding. *Visual Communications and Image Processing(VCIP)*, 7744, July 2010. 109

[157] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao. Rate-SSIM optimization for video coding. In *Proc. ICASSP*, Prague, Czech, May 2011. 19

[158] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao. SSIM-inspired divisive normalization for perceptual video coding. In *IEEE International Conference on Image Processing ICIP*, pages 1657–1660, Brussels, Belgium, September 2011. 19, 20, 67, 68, 135

[159] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao. SSIM-motivated rate distortion optimization for video coding. *IEEE Trans. on Circuits and Systems for Video Technology*, 22:516–529, Apr. 2012. 152, 153

[160] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao. Perceptual video coding based on SSIM-inspired divisive normalization. *IEEE Transactions on Image Processing*, 22(4):1418–1429, 2013. 20

[161] S. Wang, L. Zhang, Yan Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 0:2216–2223, 2012. 79

[162] Z. Wang. Applications of objective image quality assessment methods [applications corner]. *IEEE Signal Processing Magazine*, 28(6):137–142, 2011. 23, 24, 27

[163] Z. Wang and A. C. Bovik. *Modern Image Quality Assessment*. Morgan & Claypool Publishers, March 2006. 15, 35, 135

[164] Z. Wang, A. C. Bovik, and B. L. Evans. Blind measurement of blocking artifacts in images. In *Proc. IEEE Int. Conf. Image Proc.*, volume 3, pages 981–984, September 2000. 27

[165] 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*, 13(4):600–612, April 2004. 4, 11, 17, 18, 21, 24, 27, 45, 53, 54, 55, 67, 93

[166] 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*, 26:98–117, Jan. 2009. 14, 15, 19, 23, 27, 31, 135

[167] Z. Wang and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *Journal of Optical Society of America A*, 24(12):B61–B69, December 2007. 92, 103, 105, 107, 116, 119

[168] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Trans. Image Processing*, 20(5):1185–1198, May 2011. 18, 21, 27, 139

[169] Z. Wang, Q. Li, and X. Shang. Perceptual image coding based on a maximum of minimal structural similarity criterion. *IEEE International Conference on Image Processing*, 2:II –121 –II –124, September 2007. 19

[170] Z. Wang, L. Lu, and A. C. Bovik. Foveation scalable video coding with automatic fixation selection. *IEEE Trans. Image Processing*, 12(2):243–254, February 2003. 32

[171] Z. Wang, L. Lu, and A. C. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication,* special issue on objective video quality metrics, 19(2):121–132, February 2004. 94, 150

[172] Z. Wang, H. R. Sheikh, and A. C. Bovik. No-reference perceptual quality assessment of JPEG compressed images. In *Proc. IEEE Int. Conf. Image Proc.*, Rochester, September 2002. 27

[173] Z. Wang, H. R. Sheikh, and A. C. Bovik. Objective video quality assessment. In Borko Furht and Oge Marques, editors, *The Handbook of Video Databases: Design and Applications*, pages 1041–1078. CRC Press, September 2003. 27, 40

[174] Z. Wang and E. P. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Human Vision and Electronic Imaging X, Proc. SPIE*, volume 5666, San Jose, CA, January 2005. 28, 52, 55, 56, 57

[175] Z. Wang and E. P. Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *Proc. ICASSP*, Philadelphia, PA, March 2005. 18, 19

[176] Z. Wang and E. P. Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):1–13, September 2008. 22, 73

[177] 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*, pages 1398–1402, Pacific Grove, CA, November 2003. 17, 21, 24, 27, 181, 188

[178] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, En-Hui Yang, and A. C. Bovik. Quality-aware images. *IEEE Trans. Image Processing*, 15(6):1680–1689, June 2006. 28, 53, 54, 55

[179] A. B. Watson and J. A. Solomon. Model of visual contrast gain control and pattern masking. *J. Opt. Soc. Am. A*, 14(9):2379–2391, 1997. 20, 42, 68, 135, 138

[180] S. M. Weiss and N. Indurkhya. Rule-based machine learning methods for functional prediction. *Journal of Artificial Intelligence Research*, 3:383–403, 1995. 47

[181] T. Wiegand and B. Girod. Lagrange multiplier selection in hybrid video coder control. *International Conference on Image Processing*, pages 542–545, 2001. 33, 92, 109, 110

[182] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullian. Rate-constrained coder control and comparison of video coding standards. *IEEE Trans. on Circuits and Systems for Video Technology*, 13:688–703, Jul. 2003. 92

[183] S. Winkler. Analysis of public image and video databases for quality assessment. *J. Sel. Topics Signal Processing*, 6(6):616–625, 2012. 153

[184] S. Wolf and M. H. Pinson. Spatio-temporal distortion metrics for in-service quality monitoring of any digital video system. *Proc. SPIE*, 3845:266–277, 1999. 28

[185] H. R. Wu and M. Yuen. A generalized block-edge impairment metric for video coding. *IEEE Signal Processing Letters*, 4(11):317–320, November 1997. 27

[186] W. Xue and X. Mou. Reduced reference image quality assessment based on Weibull statistics. In *Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6, june 2010. 57

[187] C. Yang, H. Wang, and L. Po. Improved inter prediction based on structural similarity in H.264. *IEEE International Conference on Signal Processing and Communications*, 2:340–343, 2007. 19, 33

[188] C. L. Yang, , R. K. Leung, L. M. Po, and Z. Y. Mai. An SSIM-optimal H.264/AVC inter frame encoder. *IEEE International Conference on Intelligent Computing and Intelligent Systems*, 4:291–295, 2009. 33, 130

[189] E. H. Yang and X. Yu. Rate distortion optimization for H.264 inter-frame video coding: A general framework and algorithms. *IEEE Trans. on Image Processing*, 16:1774–1784, Jul. 2007. 31, 33

[190] E. H. Yang and X. Yu. Soft decision quantization for H.264 with main profile compatibility. *IEEE Trans. on Circuits and Systems for Video Technology*, 19:122–127, Jan. 2009. 31, 33

[191] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478, 2012. 77, 79

[192] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861 –2873, nov. 2010. 34, 64, 79

[193] X. K. Yang, W. S. Lin, Z. K. Lu, E. P. Ong, and S. S. Yao. Just noticeable distortion model and its applications in video coding. *Signal Processing: Image Communication*, 22:662–680, Aug. 2005. 32

[194] X. K. Yang, W. S. Lin, Z. K. Lu, E. P. Ong, and S. S. Yao. Motion-compensated residue pre-processing in video coding based on just-noticeable-distortion profile. *IEEE Trans. on Circuits and Systems for Video Technology*, 15:742–752, Jun. 2005. 32

[195] YouTube. YouTube statistics. http://www.youtube.com/t/press_statistics. 1

[196] Z. Yu, H. R. Wu, S. Winkler, and T. Chen. Vision-model-based impairment metric to evaluate blocking artifact in digital video. *Proceedings of the IEEE*, 90(1):154–169, January 2002. 27

[197] K. Zeng, A. Rehman, J. Wang, and Z. Wang. From H.264 to HEVC: Coding gain predicted by objective quality assessment models. In *Seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, February 2013. 195

[198] W. Zeng, S. Daly, and S. Lei. An overview of the visual optimization tools in JPEG 2000. *Signal Processing: Image Communication*, 17:85–104, 2001. 32

[199] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *Curves & Surfaces*, Avignon-France, June 2010. 34, 77, 78, 79

[200] J. Zhang, X. Yi, N. Ling, and W. Shang. Context adaptive lagrange multiplier (CALM) for motion estimation in JM-improvement. *Joint Video Team (JVT) of ISO/IEC MPEG ITU-T VCEG*, Jul. 2006. 32, 130

[201] J. Zhang, X. Yi, N. Ling, and W. Shang. Context adaptive Lagrange multiplier (CALM) for rate-distortion optimal motion estimation in video coding. *IEEE Trans. on Circuits and Systems for Video Technology*, 20:820–828, June. 2010. 32

[202] M. Zhang, W. Xue, and X. Mou. Reduced reference image quality assessment based on statistics of edge. *Digital Photography VII*, 7876(1):787611, 2011. 57

[203] X. Zhao, J. Sun, S. Ma, and W. Gao. Novel statistical modeling, analysis and implementation of rate-distortion estimation for H.264/AVC coders. *IEEE Trans. on Circuits and Systems for Video Technology*, 20:647–660, May. 2010. 143, 146

[204] X. Zhao, L. Zhang, S. Ma, and W. Gao. Rate-distortion optimized transform for intra-frame coding. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1414–1417, Mar. 2010. 31

[205] M. Zink, O. Künzel, J. Schmitt, and R. Steinmetz. Subjective impression of variations in layer encoded videos. In *Proceedings of the 11th international conference on Quality of service*, pages 137–154, 2003. 36