Visual signal coding and quality evaluation

Liu, Anmin

2011

Liu, A. M. (2011). Visual signal coding and quality evaluation . Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/47587>

<https://doi.org/10.32657/10356/47587>

Visual Signal Coding and Quality Evaluation

A thesis submitted to
School of Computer Engineering
Nanyang Technological University

by

LIU ANMIN

in Partial Fulfillment of the Requirement for the Degree of
Doctor of Philosophy (Ph.D)

2011

Acknowledgments

During my time as a Ph.D student, I received help, advice and support from many people around me. I would like to take this opportunity to thank them for the help that I received over the past four years.

First of all, I am deeply grateful to Prof. Weisi Lin for offering me the opportunity to pursue my doctoral studies under his supervision. He is a considerate supervisor and is always willing to discuss and share his knowledge and skill with me. During my research, I have learned a lot from his professional work attitude, expertise and broad mind. I really appreciate his continuous guidance and encouragement.

I would like to acknowledge the supports to my Ph.D study by the scholarship from the School of Computer Engineering, Nanyang Technological University.

I am also thankful to Dr. Zhenzhong Chen for his comments and suggestions regard to perception based video coding.

I enjoyed working with the labmates at the CeMNet and CeMNet Annex Labs. I wish to thank them for their valuable suggestions in the discussions: Fan Zhang, Manoranjan Paul, Chenwei Deng, Narwaria Manish, Zhouye Gu, Yuming Fang, Xiangang Chen, Feng Zhong, Lu Dong, Huan Yang, Nevrez Imamoglu, Wei Liu, Guangtao Zhai, Wei Zhao and other students and staff in the lab.

I am thankful to all the people who participated in my subjective experiments for sparing their valuable time to help.

I would like to thank in advance the thesis examiners for accepting to be part of the committee, and for their comments and suggestions to improve this thesis.

Last but not the least, I would like to thank all my friends and relatives who help and support me these years during the whole candidature period.

Abstract

Visual signal (i.e., images and videos) coding is to compress digital visual data to be as small in size as possible in order to make use of limited bandwidth of networks and cater for compact storage, by exploring various data redundancy. It exploits the redundancy in signal itself (statistical redundancy, i.e., spatial-temporal redundancy and spectral/color redundancy). Since the human visual system (HVS) is the ultimate receiver and appreciator of most processed visual signal, we should also consider the redundancy due to the human vision properties (i.e., perceptual/psycho-visual redundancy) in the course of coding. The effectiveness of image and video coding methods is traditionally evaluated with their rate-distortion (RD) performance where rate is the number of bits required for the compressed visual signal (or its variants such as bits per pixel (bpp) and bits per second) and distortion is usually measured as peak signal to noise ratio (PSNR). However, it has been found that PSNR is not always in accordance with the human judgment and therefore the measurement for perceptual distortion is an active research area.

Firstly, in this work, we discuss the statistical redundancy of video and then propose a novel optimal compression plane (OCP) based video coding scheme. In the sense of data structure, video is nothing more than a three dimensional data matrix, and the distinction among X (a spatial dimension), Y (the other spatial dimension), and T (the temporal dimension) is not absolutely necessary. We ignore the physical meaning of X, Y, and T axes for a video during the video coding process; frames are allowed to be formed in the

TX (or TY) plane rather than the traditional XY plane to exploit the redundancy more effectively, and therefore better coder performance is achieved.

Secondly, the model reflecting the masking characteristics of the HVS is studied as it is fundamental for perceptual redundancy exploring and visual distortion (quality) measurement. Just noticeable difference (JND) accounts for various masking effects of the HVS. We improve the pixel domain JND model by better contrast masking (CM) evaluation and appropriately accounting for the difference of CM for textural and edge regions. We also investigate into the application of the perceptual models (i.e., visual attention model and JND model) in the context of adaptive sampling based low-bit-rate image coding and JND based histogram adjustment for visually lossless image coding.

Lastly, an effective and efficient metric of visual quality/distortion evaluation is proposed. The metric is based on the similarity between the gradient profiles of the reference and distorted signals which accounts for both the high level premise of the HVS (i.e., high sensitivity to image edges and structure) and the masking property. This new metric is with simple calculation and high accuracy (verified with extensive cross-database tests); it is robust to various distortion types and can be easily embedded in coding systems (as well as other visual signal processing algorithms).

Contents

Acknowledgments	i
Abstract.....	ii
Contents	iv
List of Figures.....	vii
List of Tables	ix
List of Abbreviations	x
Chapter 1 Introduction.....	1
1.1 Background and Motivation	1
1.2 Objective and Scope of This Work.....	5
1.3 Major Technical Contributions	5
1.4 Organization of the Thesis	7
Chapter 2 Literature Survey	9
2.1 Visual Signal Coding Techniques.....	9
2.1.1 Image Coding and Video Coding.....	9
2.1.2 H.264 Standard for Lossy Video Coding	10
2.1.3 Lossless Video Coding.....	12
2.1.4 Other Visual Signal Coding Methods	14
2.2 Perceptual Visual Modeling and Processing.....	17
2.2.1 Human Visual Attention Modeling	17
2.2.2 Just Noticeable Difference Model.....	19
2.2.3 Perception Based Visual Signal Coding.....	23
2.2.4 Visual Quality Evaluation Schemes.....	25
Chapter 3 Video Coding with Adaptive Optimal Compression Plane	
Determination.....	27
3.1 Introduction.....	28
3.2 Video Redundancy Analysis.....	30
3.2.1 Temporal Redundancy	32

3.2.2	Comparison of Redundancy along Different Axes	32
3.2.3	C_d Calculation using Sampled Frames	33
3.3	Proposed Framework	36
3.3.1	Optimal Coding Plane and C_d	36
3.3.2	Overall Scheme	37
3.3.3	Impact of PPU Size	39
3.3.4	Computational Complexity	39
3.4	OCP without Inter-frame Prediction	40
3.5	OCP with Inter-frame Prediction	43
3.5.1	Brute-force OCP Determination.....	43
3.5.2	Efficient OCP Prediction.....	45
3.6	Experimental Results and Discussions	46
3.6.1	OCP under Different Conditions.....	47
3.6.2	Performance Comparison with Motion JPEG-LS	50
3.6.3	Performance Comparison with Motion JPEG, Motion JPEG 2000 and H.264 Intra-only Profile.....	54
3.6.4	Performance Comparison with H.264.....	57
3.7	Concluding Remarks.....	60
Chapter 4	JND Model with Separation of Edge and Textural Regions.....	62
4.1	Introduction.....	63
4.2	The Proposed JND Model.....	64
4.2.1	Image Decomposition	64
4.2.2	Contrast Masking in JND Model	67
4.3	Experimental Results and Discussions	71
4.3.1	Model Validation with Noise Shaping.....	71
4.3.2	Further Validation of the Model.....	73
4.4	Concluding Remarks.....	76
Chapter 5	Perceptual Image Coding.....	77
5.1	Perception based Down-sampling.....	78
5.1.1	System Overview and Visual Attention Determination.....	79
5.1.2	QP Determination.....	81
5.1.3	Sampling Mode Determination	84
5.1.4	Experimental Results and Discussions.....	85
5.2	Visually Lossless Coding.....	88

5.2.1	Image Histogram Adjustment based on JND	89
5.2.2	Iterative Implementation	92
5.2.3	Experimental Results and Discussions.....	94
5.3	Concluding Remarks.....	96
Chapter 6	A New Method for Visual Quality Evaluation	97
6.1	Introduction.....	98
6.2	Structural Similarity (SSIM) Index and Related Schemes	100
6.3	Proposed Gradient Similarity Scheme	102
6.3.1	Gradient Similarity.....	102
6.3.2	Further Analysis for Proposed Scheme and SSIM.....	108
6.3.3	Modified Gradient Similarity	109
6.4	Integration for Overall Image Quality	111
6.4.1	Measurement for Luminance Distortion	111
6.4.2	Adaptive Distortion Integration	112
6.5	Experimental Results and Discussions	113
6.5.1	Databases and Evaluation Criteria	114
6.5.2	Accuracy and Monotonicity Evaluation.....	115
6.5.3	Robustness Evaluation	116
6.5.4	Efficiency Evaluation.....	119
6.5.5	Impact of the Parameter Values	119
6.5.6	Impact of Each Component of the Scheme	124
6.6	Concluding Remarks.....	125
Chapter 7	Summary and Future Work	126
7.1	Summary	126
7.1.1	Statistical Redundancy Reduction.....	126
7.1.2	Perceptual Modeling and Redundancy Removal	128
7.1.3	Quality Evaluation for Visual Signal	129
7.2	Future work.....	131
References		134
Publications		150
	Journal Papers	150
	Conference Papers	151

List of Figures

Figure 1.1: An example for perceptual redundancy.....	4
Figure 1.2: Two images with same PSNR (30.4dB).....	4
Figure 1.3: Major contents and organization of the thesis.	6
Figure 2.1: An example of hierarchical coding structure [15].	11
Figure 2.2: An example of down-sampling based image coding (bpp=0.169).....	16
Figure 2.3: Block diagram of the typical sampling based coding scheme [30].	16
Figure 2.4: An example of visual attention [33].	17
Figure 2.5: The architecture of Itti <i>et al.</i> 's bottom-up attention model [40].	20
Figure 2.6: Operators for calculating the gradient value.....	22
Figure 3.1: Four video sequences with different typical motion characteristics.....	31
Figure 3.2: Inter-frame correlation coefficients of four typical sequences.	31
Figure 3.3: Average inter-frame correlation coefficients along T, Y and X axes.	35
Figure 3.4: Rate-Distortion performance for sequences with different frame formation (without inter-frame prediction).....	38
Figure 3.5: Block diagram of the proposed scheme (illustrated with XY and non-XY frames of “Mobile” video sequence for better visual impression).....	38
Figure 3.6: Distribution of the intra-frame prediction (JPEG-LS) residues.....	42
Figure 3.7: Distribution of the DCT (Motion JPEG) coefficients.....	42
Figure 3.8: Relative frequency vs. quantization parameter (Q_p) for various values of the Lagrange multiplier λ_{RDO}	44
Figure 3.9: Percentage of intra modes for H.264 coding in XY, TX and TY planes.....	45
Figure 3.10: (a) Average saving of bits and (b) overhead bit rate vs. Pre-Processing Unit (PPU) size N_{PPU}	48
Figure 3.11: Results for Motion JPEG.....	51
Figure 3.12: Results for Motion JPEG 2000.....	52
Figure 3.13: Results for H.264 intra-only profile.	53
Figure 3.14: Results for the comparison of the OCP and XY plane coding (i.e. H.264).....	58

Figure 3.15: Simulation result for the sequence “Tempete” (a 720x486 sequence with flowers, falling leaves, and stones).....	58
Figure 4.1: Determination of λ for three images and the average result over the ten test images in the first column of Table 4.1.	65
Figure 4.2: Structure-texture decomposition results.	66
Figure 4.3: Block diagram of the proposed direct pixel domain JND model.	66
Figure 4.4: Detected edge information (binary image, with black pixels representing edges) for the proposed and Yang <i>et al.</i> ’s models, with different threshold (t_e).	68
Figure 4.5: Contrast masking (CM) in different JND models (scaled to [0 255] and higher brightness means a larger masking value).	70
Figure 4.6: JND maps from the proposed model for two images (scaled to [0 255]).	71
Figure 5.1: Block diagram of the down-sampling based coding method (the parts enclosed with dash lines) and the inclusion of the proposed perception-based module.....	79
Figure 5.2: QP vs. average bpp	82
Figure 5.3: Comparison of Different Models in terms of PSNR vs. bpp.	87
Figure 5.4: Reconstructed images by using the method in [30] and the proposed method, under a same bit rate (at 0.105 bpp).	88
Figure 5.5: An example for the proposed scheme.....	93
Figure 6.1: Block diagram of the proposed scheme.....	102
Figure 6.2: An illustration of the difference between the SSIM and the proposed scheme.	103
Figure 6.3: The predicted value from schemes under consideration (X-axis) and the subjective DMOS (Y-axis; with DMOS>50) for the LIVE database.	107
Figure 6.4: A simple example to demonstrate the benefit of the modification for K	110
Figure 6.5: Scatter plots of subjective scores vs. scores from the proposed scheme q on IQA databases.....	117
Figure 6.6: Plot of SROCC as a function of K' for IQA databases.	121
Figure 6.7: Plot of (a) SROCC, (b) CC, and (c) RMSE, as a function of p for the proposed integration approach and for the TID dataset [135].	122
Figure 6.8: Plot of SROCC as a function of p for IQA databases.	123

List of Tables

Table 3.1: Names and indices of the video sequences.	35
Table 3.2: Relationship among C_d values.	35
Table 3.3: Relationship of C_d with different s conditions (shaded: s conditions with different calculated relationship of CCs compared with that of $s=1$).	36
Table 3.4: Bits per pixel (bpp) for lossless compressed videos under different frame formation.	41
Table 3.5: OCPs without (with) inter-frame prediction.	49
Table 3.6: Results of dP and saving of bits for Motion JPEG-LS (near-lossless) for OCP.	49
Table 3.7: PSNR gain of OCP at 0.8 bpp against Motion JPEG, Motion J2K, and H.264 intra- only profile (I264).	55
Table 3.8: Comparison of RD performance of the proposed scheme against H.264 under the IP configuration (IP) and the configurations with B pictures (IBP) and that with two reference frames (S2R) (for 0.05~2 bpp).....	61
Table 4.1: The subjective quality evaluation results (the proposed model against each of those in [47] and [49]) and PSNRs for 10 images with different visual content.	72
Table 4.2: Scores for subjective quality evaluation.	73
Table 4.3: Prediction performance for different approaches.....	76
Table 5.1: Different down-sampling modes (indexed by k).	82
Table 5.2: Candidate QP list.	83
Table 5.3: Subjective viewing results.	87
Table 5.4: Required bits for different coding schemes.	93
Table 5.5: Subjective viewing results.	95
Table 6.1: Gradient and standard deviation for different image blocks in Figure 6.2.	107
Table 6.2: Performance comparison for IQA schemes on six databases.	118
Table 6.3: Average performance over six databases.....	118
Table 6.4: SROCC comparisons for individual distortion types.....	118
Table 6.5: Execution time (in second/image) for different schemes.....	118
Table 6.6: SROCC comparisons for each component of the proposed scheme.....	123

List of Abbreviations

3G/4G	Third/Fourth Generation mobile telecommunications
BME	Block-based Motion Estimation
bpp	bits per pixel
CC	(Pearson) Correlation Coefficient
CI	Confidence Interval
CM	Contrast Masking
DCT	Discrete Cosine Transform
DMOS	Difference Mean Opinion Score
EM	Contrast Masking around Edges
GOP	Group of Pictures
HVS	Human Visual System
IBP	H.264 configuration with Bi-directional predicted frames
IP	H.264 configuration with only Intra- and Inter- predicted frames
IQA	Image Quality Assessment
ITU-T	International Telecommunication Union-Telecommunication Standardization
J2K	JPEG 2000
JND	Justice Noticeable Difference
JPEG	Joint Photographic Experts Group
JPEG-LS	Lossless JPEG
KROCC	Kendall Rank-order Correlation Coefficient
LA	Luminance Adaptation
MAD	Most Apparent Distortion
MB	Macro-block
MOS	Mean Opinion Score
MPEG	Motion Picture Expert Group
MSE	Mean Squared Error
NAMM	Nonlinear Additivity Model for Masking
OCP	Optimal Compression Plane
PPU	Pre-Processing Unit
PSNR	Peak Signal Noise Ratio
QCIF	Quarter Common Intermediate Format
QP	Quantization Parameter
RD	Rate-Distortion
RMSE	Root Mean Squared Error
ROI	Region of Interest
S2R	H.264 configuration with two reference frames
SROCC	Spearman Rank-order Correlation Coefficient
SSIM	Structural SIMilarity
TM	Contrast Masking in Texture Regions
TV	Total Variation
VIF	Visual Information Fidelity
VSNR	Visual Signal-to-Noise Ratio

Chapter 1

Introduction

1.1 Background and Motivation

The explosion of the number of computers and digital systems connected by networks such as the Internet has brought a flow of instant information into a large and increasing number of homes and businesses. Most of the information is in the form of digital visual signals (i.e., images and videos) as intuitive and faithful depiction of things in life and work. A picture is worth a thousand words, and people in different parts of the world are able to perceive the same image/video despite that they speak differently. As a result, products (e.g., phone cameras) and services (e.g., windows media players, YouTube) based upon images and videos, as well as the related delivery (e.g., via 3G/4G networks), have grown at an explosive rate.

Digital visual signals in uncompressed formats require excessive storage capacity and a huge transmission bit rate. For example, a single digital television signal in Consultative Committee of International Radio 601 format [1] requires a transmission rate of 216 Mega-bits per second. This is unacceptably high in bit rate for most practical purposes, and therefore, there is a need to reduce the data rate via coding, before digital television and video can be fed into the storage systems and communication networks. The goal of

visual signal coding is to ensure good signal quality within the provision of transmission and storage specifications. In general, coding quality, compression ratio (bit rate) and computational complexity are factors that measure success of a coding scheme. These factors are usually measured by MSE/PSNR (Mean Squared Error/Peak Signal Noise Ratio), bpp (bits per pixel) and computational time, respectively.

There are a number of existing video coding standards, i.e., MPEG-2, MPEG-4, H.261, H.263, and H.264. These standards have used a few video coding techniques which exploit some of the inherent statistical redundancy within a frame and among frames in order to provide significant visual data compression. Although they have made it practical to store, transmit and manipulate digital image and video information using currently available storage systems and data networks, rooms for further performance improvement are still to be explored, in order to make the related products and services more cost effective, as well as enabling more new functionalities. In this thesis, three aspects of limitation for the existing visual signal coding schemes are addressed.

First, all the existing video coding standards can only explore limited statistical redundancy since they are under the constraint of encoding video one natural spatial frame by one natural spatial frame. Although it is the way a video is captured by sensors and displayed for viewing, encoding in such a way is not absolutely necessary (e.g., in applications such as transmission and storage). A video sequence can be treated as a three-dimensional data matrix in terms of data structure, and from this viewpoint, the physical meaning of spatial natural frames can be ignored and this provides a way for better statistical redundancy reduction.

Second, besides statistical redundancy, visual signal also has perceptual redundancy, which can be exploited since the ultimate receiver is the human visual system (HVS) for the coded visual signals. The capability of information processing of the HVS is limited

and not all the visual information is noticed, processed, or utilized. The un-noticed or un-utilized information is redundant, and therefore can be explored and reduced in the process of coding to save the required bits. One example is given in Figure 1.1, where (a) is the “Lena” image, and (c) is its DCT (discrete cosine transform) result; if we discard the DCT coefficients of the highest frequencies in (c) (as shown in the bottom right corner in (d)), the corresponding image is shown in (b); both (c) and (d) are plotted in logarithm scale to bring out the higher-frequency coefficients for visual display as in [2]. As can be seen, (a) and (b) are visually the same although (b) is with less information than (a). The example demonstrated that the HVS is not sensitive to very high frequency information and discarding this information in coding would not affect the perceived quality significantly. The existing coding schemes have considered the treatment of high frequency components for quantization. In this study, we further investigate into perceptual modeling and redundancy reduction.

Third, image and video coding is an optimization process and the improvement of the optimization criterion would provide better coding. For example, there are many candidate coding modes in the H.264 video coding standard and the finally chosen one for coding is the one with the optimal/best quality with the criterion of MSE/PSNR for a given bit rate. However, the often-used criterion (i.e., PSNR) is not always in accordance with the HVS perception [148], [149]. In Figure 1.2, (a) and (b) have equivalent quality under the PSNR criterion (both with PSNR of 30.4dB) but (a) looks much better than (b) to viewers (especially in the shoulder region). The subjective viewing tests cannot be used as the optimization criterion for on-line optimization. Therefore, better video coding and evaluation can be achieved with the investigation into HVS-oriented objective evaluation criterion for quality and replacement of the current widely-used PSNR (or its relatives) in visual signal coding and quality evaluation.

(a) Lena image

(b) The reconstructed image for (d)

(c) DCT result of (a)

(d) DCT coefficients with the highest frequencies in (c) being discarded (as shown in the bottom right corner)

Figure 1.1: An example for perceptual redundancy.

(a) JPEG 2000 image

(b) JPEG image

Figure 1.2: Two images with same PSNR (30.4dB).

1.2 Objective and Scope of This Work

The objective of this study is to explore new methods for visual signal coding and quality evaluation better than the existing ones, in the three aspects mentioned in Section 1.1 above, by investigating further into appropriate statistical data of typical video sequences as well as the relevant property of the HVS perception.

In particular, we try to address the following three problems of visual signal coding and quality evaluation. Firstly, how to explore the statistical redundancy more effectively without the traditional constraint of natural frames? Secondly, how to accurately model relevant masking characteristics of the HVS and how to design an appropriate coding scheme which incorporates the HVS model seamlessly to reduce the perceptual redundancy as much as possible? Thirdly, how to assess the quality of visual signal (in accordance with the mean opinion of observers)?

1.3 Major Technical Contributions

To achieve a better perceived quality within a given bit rate, this thesis has presented new coding methods and perceptual models which can improve the effectiveness of visual signal coding and quality evaluation in the three identified directions: statistical redundancy reduction, perceptual redundancy removal, and visual quality evaluation. The major technical contributions can be summarized as follows:

- Proposed a pre-processing step with low computation complexity prior to actual video coding, called optimal coding plane (OCP) selection.

The OCP concept is first demonstrated with JPEG-LS (Lossless JPEG; JPEG is the standard from Joint Photographic Experts Group) video coding and then extended to H.264 video coding.

- Modeled and applied the visibility threshold of the HVS.

We first demonstrated that the existing pixel domain JND (just noticeable difference, which accounts for the visibility threshold of the HVS) model can be improved by more appropriate distinguishing masking effect in edge regions from that in texture regions.

We then discussed how the perceptual models can be used in the image coding process such as down-sampling based image coding (by incorporation visual attention model) and visually lossless image coding (by JND based histogram adjustment).

- Designed an HVS-oriented objective image/video quality assessment metric based on gradient similarity, which is of high accuracy, good robustness and low complexity.

Such a metric can be used as a standalone visual quality estimator or a control module in video coding (or other visual processing, e.g., watermarking, and post-processing).

Figure 1.3: Major contents and organization of the thesis.

1.4 Organization of the Thesis

Figure 1.3 illustrates the major contents and organization of this thesis, for easy reference to the reader, and the whole thesis is divided into seven chapters.

Chapter 1 (this chapter) gives an introduction about the thesis, including the background and motivation, objective and scope, technical contribution and thesis organization.

Chapter 2 describes the major related existing work, encompassing the basics of the lossless and lossy video coding, the sampling based image coding framework, and the typical perceptual models (i.e., visual attention and JND models). The state of the art image quality assessment methods are also reviewed in this chapter. More specific literature survey to each proposed technique in this thesis will be further introduced whenever appropriate in Chapters 3-6.

Chapter 3 discusses the benefits of allowing frames to be formed in a plane other than the traditional spatial plane. Statistical redundancy can be explored to a fuller extent and better coding performance is therefore achieved although the frames in a non-spatial plane that does not have any physical meaning.

Chapter 4 describes the proposed JND model to account for the masking effects of the HVS and the estimation of the visibility threshold for the visual signal. The model is designed in image pixel domain and with appropriate distinction between contrast masking (CM, which denotes the visibility reduction of one visual signal at the presence of another one [44]) around edge regions and that for textural regions.

Chapter 5 addresses the use of perception-based models for video coding. By means of quantization parameter and histogram adjustment, the perceptual aspect of down-sampling based coding and lossless coding is explored.

Chapter 6 presents a simple but effective approach for visual quality assessment by using the similarity of gradient information and taking account of the masking property of the HVS. With luminance distortion (contrast and structure invariant), contrast and structure variant distortion is emphasized properly. The approach is demonstrated extensively with images with various visual content and distortion types.

Chapter 7 closes the thesis with a summary of the main research work performed and several directions for further studies.

Chapter 2

Literature Survey

In this chapter, we give a brief overview of the major existing work relevant to visual signal coding, perceptual modeling, and visual quality evaluation, since these topics are the closest to our research in this thesis and our research work is grounded on the surveyed literature.

2.1 Visual Signal Coding Techniques

2.1.1 Image Coding and Video Coding

Video is a collection of natural frames and temporal redundancy exists between these frames besides the spatial redundancy within each natural frame. Video coding and image coding are closely related since each frame of the video can be deemed as an image. To be more specific, video coding is an extension of image coding by dealing with the temporal redundancy, and usually there are two types of techniques for such an extension:

- With inter-frame prediction: this type of extension is usually by reducing the temporal redundancy among successive natural frames prior to intra-frame coding with image coding techniques. The most commonly used technique for inter-frame prediction is Block-based Motion Estimation (BME) [3]-[6]. In BME, each frame is

divided into 8×8 blocks (or 16×16 macro-blocks (MBs)), and each block in the current frame is predicted from a block of equal size in the reference frame. The offset between the two blocks is known as a motion vector. The error between the current block and the similar block in the reference frame is encoded and transmitted along with the motion vector for the block. To exploit the redundancy between neighboring block vectors (e.g., for a single moving object occupying multiple blocks), it is common to encode only the difference between the current and previous motion vectors into the bit stream.

- Without inter-frame prediction: in some cases, to save the power of the processing cell or when the processor's computational resource is limited, BME is impossible due to its high complexity (the computational complexity of BME varies from 50% to 90% of a typical video coding system [7], [8]). Therefore, each frame would be coded independently by using image coding techniques [9]-[14]. The lack of use of inter-frame prediction results in reduction of compression capability, but robustness to error due to transmission. It is often used in mobile appliances also because it is with low processing requirement, ease of implementation, and broad compatibility; it is also used in the case when the zero-delay feature is required.

2.1.2 H.264 Standard for Lossy Video Coding

A video coding standard is the language with which a video encoder and a decoder communicate. The development of international video coding standards has evolved through ITU-T H.261, H.262/MPEG-2, H.263/MPEG-4(part 2), and H.264/MPEG-4(part 10) which are mainly designed for lossy video coding; ITU-T and MPEG are the International Telecommunication Union-Telecommunication Standardization and the Motion Picture Expert Group, respectively. As we know, H.264 video compression has

employed more complicated techniques to achieve higher coding efficiency than its predecessors did. Nevertheless, the fundamental technology behind these standards remains similar: motion compensated prediction to remove inter-frame redundancy, e.g., via BME, followed by transform coding for energy compaction that allows effective quantization and has been proven to be exceptionally effective to compress visual data.

H.264 includes many profiles. Each profile is designed for specific applications and there is no need to support all applications in one profile. For example, Baseline Profile is primarily for low-cost applications, and this profile is used in video conferencing and mobile applications; High Profile is primary for broadcast and disc storage applications such as Blu-ray Disc storage format. The standard also contains four additional all-intra profiles, which are defined as simple subsets of other corresponding profiles. These are mostly for camcorders, editing, and professional applications.

In contrast to the previous video coding standards, the coding and display order of pictures in H.264 is completely decoupled, and such flexibility of H.264 is one of the main reasons for its improved coding efficiency. H.264 coding with hierarchical B pictures (also using B pictures as reference) is presented in Figure 2.1 [15]. In comparison to the widely-used traditional IBBP... structure (not using B pictures as reference), coding gain can be achieved although there is increased coding delay (in the scale of the number of pictures in a GOP (Group of Pictures) minus 1).

Figure 2.1: An example of hierarchical coding structure [15].

Another reason for the improved coding efficiency of H.264 is that a more extensive search/optimization for the best coding mode is used. The optimization process is usually referred as rate-distortion (RD) optimization, which compares different coding modes in terms of coding efficiency (measured by the RD cost). To make an RD optimized decision during encoding a block of video data, the block has to be encoded a number of times before the encoder can arrive at the best mode decision. In H.264, the RD cost for a candidate coding mode (denoted as M_{RDO}) is mathematically described as [16], [17]:

$$J_{RDO}(Q_p, M_{RDO}) = D_{RDO}(Q_p, M_{RDO}) + \lambda_{RDO} \cdot R_{RDO}(Q_p, M_{RDO}) \quad (2.1)$$

with

$$\lambda_{RDO} = 0.85 \cdot (2^{(Q_p-12)/6})^2 \quad (2.2)$$

where Q_p is the quantization parameter (QP); $J_{RDO}(\cdot, \cdot)$, $D_{RDO}(\cdot, \cdot)$, and $R_{RDO}(\cdot, \cdot)$ are the RD cost function, distortion function, and rate function, respectively; λ_{RDO} is Lagrange multiplier. The smaller the RD cost $J_{RDO}(\cdot)$ is, the higher the RD performance becomes.

For a given Q_p , the optimal coding mode (denoted as M_{H264}) in H.264 is searched among all possible coding modes, to achieve the minimal RD cost:

$$M_{H264} = \arg \min_{M_{RDO}} J_{RDO}(Q_p, M_{RDO}) \quad (2.3)$$

2.1.3 Lossless Video Coding

Lossy video coding is widely used for its high compression ratio; however, lossy video coding is not applicable for applications where no loss of pixel values is tolerable since it discards some of the original image information that cannot be later recovered. In some applications, lossless video coding is required where only the statistically-redundant spatial and temporal information is allowed to be removed, and the process is reversible to guarantee that the reconstructed signal is mathematically the same as the original one.

Examples of lossless coding applications are medical imaging (i.e., alteration of the original data are not allowed in order to make sure that physicians analyze pristine diagnostic images), satellite remote sensing (since every piece of information is acquired with high cost and therefore it is better to keep all acquired visual signals), and film archiving and studio applications (where the genuineness of the original images should be preserved).

The image lossless coding standards include lossless JPEG image compression (termed as JPEG-LS) and JPEG 2000 lossless coding, and the lossless video compression schemes proposed in the literature are mostly the extensions of the 2D framework for image coding as discussed in Subsection 2.1.1 (i.e., either encoding each frame independently or exploring the temporal redundancy by motion estimation).

Memon *et al.* [18] were among the pioneers to consider the problem of lossless video coding. A hybrid compression approach exploiting temporal, spatial and spectral redundancy in 3D color signal was investigated based on JPEG-LS. Yang *et al.* [19] suggested a simple scheme, where intra- or inter-frame coding is selected on the basis of temporal and spatial variations, and coding is performed according to the JPEG-LS standard. A 3D version of CALIC (context-based adaptive lossless image codec) [20] has been used to exploit either temporal or spatial redundancy on the pixel basis. Note that, these methods adaptively explore either spatial redundancy or temporal redundancy.

In [21], [22], motion vectors are used to improve the efficiency of temporal prediction and the obtained vectors themselves must be encoded with bits. Aiming to reduce these bit overheads of motion vectors, Park *et al.* in [23] presented an algorithm using backward temporal prediction, in which the motion vector is determined according to neighboring blocks and the same search effort must be performed at the decoder to restore the motion vector. This scheme has been refined in [24] where a pixel based

(instead of a block based) backward prediction is adopted. In spite of the prediction effectiveness (since both spatial redundancy and temporal redundancy are explored) for these lossless video compression methods based upon motion estimation, the computational complexity is high because of the block matching for every candidate reference block.

Recently, lossless coding profiles have also been included in H.264 coding standard [25], where the similar architecture as H.264 lossy video coding is adopted but the prediction residues are entropy encoded directly rather than undergoing transform and quantization first. Similar to H.264 lossy coding, H.264 lossless coding can also be used in all-intra profiles or inter-coding profiles, and it can be used as a lossless image encoder by deeming the image as a video with only one natural frame. Improvements for H.264 lossless coding are also proposed in the aspects of spatial prediction [154], scan order [155], and entropy coding [156].

2.1.4 Other Visual Signal Coding Methods

Many other methods have been derived to cater for different situations and requirements in visual signal coding. For example, it is known that at low bit rates a down-sampled image when JPEG compressed visually beats the compressed high-resolution image with the same number of bits, as illustrated in Figure 2.2, where (a) is by using JPEG compression and decompression, and (b) is down-sampling based, where the down-sampling factor is 0.5 for each direction. The compressed Lena images in both cases use 0.169 bpp. The reason for the better performance in Figure 2.2 (b) over Figure 2.2 (a) lies in that high spatial correlation exists among neighboring pixels in a natural image; in fact, most images are obtained via interpolation from sparse pixel data yielded by a single-sensor camera [26]; therefore, some of the pixels in an image may be omitted

(i.e., the image is down-sampled) before compression and restored from the available data (e.g., interpolated by the neighboring pixels) at the decoding end. In this way, the scarce bandwidth can be better utilized in very low bit rate situations.

In [27], Bruckstein *et al.* exploited the theoretical model of down-sampling and compared it with experimental results. The key point of sampling based coding methods is how to determine the sampling mode (the down-sampling ratio/direction) and the corresponding QP. Methods in [28], [29] manually set (i.e., preset by users) the down-sampling ratio, and the quality of the compressed image is improved when compared with a JPEG compressed one under the same bit rate. However, the encoder has to be switched between a down-sampling scheme and the standard JPEG scheme, in a variable bit-rate application for different images and if good coding quality is sought. In addition, the decimation factor is fixed and this does not reflect local visual significance of different regions of the visual content. In view of these, in [30], an adaptive sampling method is proposed to adaptively decide the appropriate down-sampling mode and the QP for every MB in an image, based upon the local visual significance of the signal. As a consequence, an image independent and larger critical bit rate (the maximum bit rate for a down-sampling based scheme to outperform the JPEG) can be obtained, and the coder switching also becomes automatic and adaptive to the image under processing. The parts enclosed with dash lines in Figure 2.3 shows the typical block diagram of the down-sampling based coding methods. Mode selection and the corresponding QP determination are the core components of the block diagram.

(a) Without down-sampling

(b) With down-sampling

Figure 2.2: An example of down-sampling based image coding (bpp=0.169).

Figure 2.3: Block diagram of the typical sampling based coding scheme [30].

Basically, in a sampling based coding method, a down-sampling filter (e.g., 2x2 average operator [28]) can be applied to reduce the resolution of the content to be coded. The encoded bit stream is stored or transmitted over the bandwidth constrained network. At the decoder side, the bit stream is decoded and up sampled (e.g., via replication filter

and then a 5x5 Gaussian filter [28]) to the original resolution. Alternatively, the full-resolution DCT coefficients can be estimated from the available DCT coefficients resulting from the down-sampled sub image, without the need of a spatial interpolation stage in the decoder [30].

The methods of visual signal coding based upon perceptual models will be surveyed and discussed in Subsection 2.2.3.

2.2 Perceptual Visual Modeling and Processing

2.2.1 Human Visual Attention Modeling

The human visual attention is the result of several millions of years of evolution [31], [32] by which we can rapidly direct our gaze toward objects of interest in the visual field. An example is shown in Figure 2.4 [33], where a lone vertical object in a horizontal field pops-out, and immediately attracts our attention. There are many applications for visual attention models, such as automatic image cropping, adaptive image retargeting, image compression, image retrieval, and video skimming.

Figure 2.4: An example of visual attention [33].

Two major attention mechanisms include top-down (knowledge/task-driven) [34] and bottom-up (stimulus-driven) [33]. In the former mechanism, attention is under the control of the subject and related to cognition processing in the human brain; it is voluntary and effortful. In the latter mechanism, attention is driven by external stimuli to determine which location is sufficiently different from its surroundings to be worthy of one's attention; it is automatic and has a transient time course. Generally, the stimuli involved in top-down control are pattern, shape, and other cognitive features, while the features involved in bottom-up control include luminance, color, orientation and motion contrast. Moreover, audition, touching, and other sensory features also affect visual attention [35].

Face is one of the main top-down visual attention features and face regions usually lie within the ROIs (region of interests) to human observers. The implementation of face detection in OpenCV [36] can be adopted in visual attention modeling to generate the face map, and the outputs are square regions which contain human faces. This face detection method is an implementation based on [37]. It uses the integral image, which allows the features to be evaluated very quickly. Besides, it is based a machine learning algorithm by constructing classifiers corresponding to small number of visual features based on AdaBoost [38]. This method combines different classifiers in cascade so that the background is soon discarded, and the efficiency is improved.

The first explicit bottom-up computational architecture was proposed by Koch *et al.* [39], with the result as a two-dimensional topographic map that represents the stimulus conspicuity, or salience, at every location in the visual scene. This general architecture has been further developed and implemented, yielding the computational model depicted in Figure 2.5 [40]. In this model, the early stages of visual processing decompose the incoming visual input into feature maps of colors, intensities, and orientations. The “center-surround” operation is then implemented on multi-scaled feature images, which

are obtained by using dyadic Gaussian pyramids. All obtained feature maps are then linearly combined into a saliency map to detect attended regions by using a winner-take-all neural network.

There are many other bottom-up visual attention models. Harel *et al.* proposed a Graph-based Visual Saliency (GBVS) model [161] by using graph theory to form saliency maps from low-level features. Bruce *et al.* described a visual attention model based on Shannon's self-information measure [162]. Liu *et al.* used machine learning to achieve the saliency map for images [163], based on the features of multi-scale contrast, center-surround histogram and color spatial distribution. Recently, some computational models for visual attention have been proposed based on Fourier Transform [164], [41]. The model in [41] achieves the final saliency map by Inverse Fourier Transform on a constant amplitude spectrum and the original phase spectrum from the images.

2.2.2 Just Noticeable Difference Model

It is well known that the human visual system (HVS) cannot sense all changes in an image/video due to its underlying physiological and psychological mechanisms [42]. JND can serve as a perceptual threshold to guide an image/video processing task. Methods of automatic JND threshold derivation have been utilized in many visual processing, e.g., image/video compression, watermarking, signal synthesis, and multimedia streaming and transmission. The JND is mainly based upon temporal/spatial contrast sensitivity function (CSF, which describes the sensitivity of the HVS for each frequency component [43], as determined by psychophysical experiments), background luminance adaptation (LA, which refers to the masking effect of the HVS toward background luminance) and CM (contrast masking [44], as defined in Section 1.4), and can be determined for either pixel domain or sub-band domain.

Figure 2.5: The architecture of Itti *et al.*'s bottom-up attention model [40].

Pixel based JND models are often used in motion estimation, visual quality evaluation and video replenishment to avoid extra decomposition. In principle, JND in pixel domain can be viewed as the compound effect of all sub-bands. However, in a practical point of view, it is better to estimate the pixel domain JND directly, for the sake of operating efficiency.

To consider the CSF factor of the HVS, sub-band decomposition is required. Ramasubramanian *et al.*'s model [45] formulates contrast sensitivity and CM in 6 band-pass sub-bands, based on a Laplacian pyramid decomposition of images. However, this model only reflects the spatial CSF roughly because of the wide frequency range in each sub-band. Zhang *et al.* [46] proposed a model of incorporating CSF in pixel domain by summing the effects of the visual thresholds in DCT sub-bands.

A number of pixel based JND models have been developed [45]-[49]. To avoid the sub-band decomposition, many of these pixel based JND models (e.g., the ones proposed by

Chou *et al.* [25], [47] and Chiu *et al.* [48]) only consist of the two remaining components to calculate JND values: the LA and the CM. In both of the models above, LA is calculated based on a parabola-shape function of local background luminance, has a minimum value at mid-range grey level (around 128), and becomes high in the regions with either a very low or a very high grey level; CM is measured based upon the variance between the central pixel and its neighboring pixels. In Chou *et al.*'s model, the CM estimator selects the maximum output from the four edge detectors with horizontal, vertical, 45° and 135° orientations; while in Chiu *et al.*'s model, CM is simply taken from the maximum grey level difference between the central pixel and its four neighboring pixels in horizontal and vertical directions.

The existing models suffer from inaccurate estimation for CM, since both edge regions and texture regions exhibit strong variation (and therefore have a large masking value in the abovementioned models) but edge regions can tolerate less noise (i.e., smaller masking value) than textural regions do. Yang *et al.*, [49] improved Chou *et al.*'s model by accounting for the difference between edge regions and texture regions to estimate the threshold at these regions more properly, and the Canny operator is used for the edge detection. The model in [49] also introduced a formula [50] termed as nonlinear additivity model for masking (NAMM) to integrate LA and CM, for more aggressive JND threshold estimation that matches the HVS' characteristics. The NAMM combines LA and CM by the sum of individual masking components minus the overlapping effect.

Mathematically, JND value in the position (i, j) for an image f is evaluated as (2.4)~(2.6) [49]:

$$T_{JND}(i, j) = T_{LA}(i, j) + T_{CM}(i, j) - C^{lc} \times \min\{T_{LA}(i, j), T_{CM}(i, j)\} \quad (2.4)$$

$$M_1 \qquad M_2 \qquad M_3 \qquad M_4$$

Figure 2.6: Operators for calculating the gradient value.

$$T_{LA}(i, j) = \begin{cases} 17 \times (1 - \sqrt{f(i, j) / 127}) + 3, & \text{if } f(i, j) \leq 127 \\ 3 \times (f(i, j) - 127) / 128 + 3, & \text{otherwise} \end{cases} \quad (2.5)$$

$$T_{CM}(i, j) = \beta_{CM} \cdot g_f(i, j) \cdot W_f(i, j) \quad (2.6)$$

where T_{JND} , T_{LA} , and T_{CM} in (2.4) are the JND value, LA value and CM value, respectively; C^{lc} is the gain reduction factor to address the overlapping between two masking factors, and with a value of 0.3 [49]; $f(i, j)$ in (2.5) is the pixel value at position (i, j) for image f ; β_{CM} in (2.6) is a control parameter, and with a value of 0.117 [49]; w_f is computed by edge detection, and with element values of 0.1 and 1 for edge and non-edge pixels respectively, followed with a Gaussian low-pass filter (to smooth w_f and therefore avoid too dramatic changes in a small neighbourhood); g_f denotes the maximal weighted average of gradients around the pixel, and is calculated as:

$$g_f = \max_{k=2,3,4} \{|f \otimes M_k|/16\} \quad (2.7)$$

with the gradient operators M_k ($k = 1, 2, 3, 4$) as shown in Figure 2.6, where the weighting coefficient decreases as the distance from the central pixel increases; and $\otimes$ is the convolution operator.

The performance of a JND model can be evaluated by its effectiveness in noise shaping in image or video [25], [47], [49]. JND guided noise injection can be made via:

$$C^T = C + \varepsilon \cdot S^{random} \cdot T_{IND} \quad (2.8)$$

where c^T is either a pixel value or a DCT coefficient value of a sub-band in the noise contaminated image (or video frame), c is either a pixel value or a DCT coefficient value of a sub-band in the original undistorted image (or video frame), s^{random} takes +1 or -1 randomly and T_{JND} is the JND value for the pixel in the image or for the DCT sub-band in the image (or video frame). This process is done for every single pixel or DCT sub-band in the image (or video frame).

In (2.8), $\varepsilon(>0)$ regulates the total noise energy to be injected. If $\varepsilon < 1$ and there is no overestimation in the JND model adopted, the noise injection is visually/perceptually lossless. Perceptual visual quality of the resultant noise-injected images can be compared and evaluated with subjective viewing tests. The resultant mean opinion score (MOS) is regarded as an indicator of perceptual quality for each image if a sufficient number of observers are involved.

Under the same level of total error energy (e.g., a same MSE or PSNR), the better perceptual quality the noise-injected image/video has, the more accurate a JND model is; alternatively, with a same level of perceptual visual quality, a more accurate JND model is able to shape more noise (i.e., resulting in lower MSE or PSNR) in an image.

2.2.3 Perception Based Visual Signal Coding

Since the human being is the final receiver/appreciator of most processed images and videos, incorporation of the characteristics of the human perception would not only make the system more customer-oriented but also bring about tremendous benefits for the system, such as performance improvement (e.g., in perceived visual quality, traffic congestion reduction, new functionalities, size of device, price of service) and/or resource saving (e.g., for bandwidth allocation, computing requirements or power dissipation in handheld devices).

In the literature, some approaches for perception based video coding have been proposed, usually based on the existing coding standards, and modifications are made to explore the perceptual aspect of video coding. In [51], visual signal is smoothed within the constraint of JND for better coding performance. In [52], [53], the QPs are adjusted according to the visual impact of the signal for the DCT based coding systems such as JPEG and H.264. In the scheme developed by Tan *et al.* [54], perceived error is approximated by a vision model based perceptual distortion metric for RD optimization in order to maximize the visual quality of JPEG 2000 coded images.

Visually lossless compression is a special type of perception based coding, and usually, the compression is said to be visually lossless when a compressed visual signal cannot be distinguished from its original. In [55], different bit streams are generated by using a standard encoder with different given bit rates, and the one with a resultant visual quality (obtained from a quality measurement, e.g., multi-scale SSIM [127]) close to a predefined threshold (e.g., 0.995 for the multi-scale SSIM) is selected as the bit stream for visually lossless. However, the criterion with which the quality score is close to the predefined threshold is not sufficient for visually lossless, because the quality score can also be high for the case where the image with visible distortion on only a small portion of the image. Therefore, most of the existing visually lossless coding methods ([56]-[61]) are based on the concept of JND (or similar concepts). The JND accounts for the maximum sensory distortion that the HVS does not perceive, and it can be served as a perceptual threshold to guide an image/video processing task. These methods modified one of the standard (e.g., JPEG, or JPEG 2000) encoder to account for the perceptual redundancy, where the distortion related parameters in the encoder are adjusted according to the JND model to guarantee that the reconstructed signal is visually lossless. In [56]-[58], JND models are incorporated into the JPEG 2000 encoder and the encoded bit

stream can be decoded by a JPEG 2000 decoder; in [59], [60], the visually lossless coding is realized by the DCT based encoder with an JND associated QP and the resultant bit stream is not standard compliant. However, these encoder manipulation based methods are embedded in the specific encoder (JPEG, or JPEG 2000), and therefore cannot be used directly in the other coding framework such as the recently proposed H.264 lossless coding [13].

2.2.4 Visual Quality Evaluation Schemes

Image quality assessment (IQA) provides quality criterion for images and videos, and also finds applications in many related algorithms and systems, such as the RD optimization process for video coding. Aimed at accurate and automatic evaluation of image quality in a manner that agrees with subjective human judgments, regardless of the type of distortion corrupting the image, the content of the image, or the strength of the distortion, substantial research effort has been directed towards developing IQA schemes over the years [61], [62]. The well-known schemes proposed in recent ten years include SSIM (structural similarity) [63], PSNR-HVS-M [64], VIF (visual information fidelity) [65], VSNR (visual signal-to-noise ratio) [66]) and the most recently proposed MAD (Most Apparent Distortion) [67].

In PSNR-HVS-M [64], MSE/PSNR in DCT domain is modified so that errors are weighted by the corresponding visibility threshold (which accounts for the masking effects of the HVS). However, as pointed out in [63], there is no clear psycho-visual evidence that the error visibility threshold based scheme is applicable to supra-threshold distortion.

The schemes proposed in [63], [65] are based on the high level property of the images (e.g., structure information [63] or statistical information [65]). They have demonstrated

success for images containing supra-threshold distortions [67], and as a tradeoff, these schemes generally perform less well on images containing near-threshold distortions since such schemes are lack of comprehensive consideration of the HVS' masking property. In [63], the SSIM assumes that the HVS is highly adapted for extracting structural information from a scene, and the structural similarity is measured as the correlation between the two image blocks. The VIF [65] views the IQA problem as an information fidelity problem, and images are modeled using Gaussian scale mixtures to measure the amount of information.

In [66], the VSNR deals with both detectability of distortions (low level vision) and structural degradation based on the global precedence (mid-level visual property), and a better tradeoff for the performance on near-threshold and supra-threshold distortions is achieved. The MAD proposed in [67] yields two quality scores: visibility-weighted error and the differences in log-Gabor sub-bands statistics. The two scores are then combined adaptively to obtain the final quality score. Although it achieves better correlation with the human judgment, it has higher computational complexity.

Chapter 3

Video Coding with Adaptive Optimal Compression Plane Determination

All existing video coding standards developed so far deem video as a sequence of natural frames (formed in the XY plane), and treat spatial redundancy (redundancy along X and Y directions) and temporal redundancy (redundancy along T direction) differently and separately. In this chapter, we investigate into a new compression (redundancy reduction) method for video in which the frames are allowed to be formed in a non-XY plane. We are to exploit fuller extent of video redundancy, and propose an adaptive optimal compression plane (OCP) determination process to be used as a pre-processing step prior to any standard video coding scheme. The essence of the scheme is to form the frames in the plane determined by two axes (among X, Y and T) according to signal correlation evaluation, and this enables better prediction (therefore better compression). In spite of the simplicity of the proposed method, it can be used for both lossless and lossy compression, and with and without inter-frame prediction. Extensive experimental results show that the new coding method improves the performance of the video coding, for a number of coding methods (inclusive of Motion JPEG-LS, Motion JPEG, Motion

JPEG 2000, H.264 intra-only profile and H.264) and videos with different visual content.

3.1 Introduction

Besides the computational complexity consideration, the goal of video coding is to ensure good video quality within the provision of transmission and storage. Therefore, complexity, distortion (or coding quality) and bit rate are factors that measure success of a video coding scheme. These factors are usually measured by computational time, PSNR and bpp, respectively, in the current video coding schemes.

If V_{XYT} represents a video sequence with axes of X, Y and T, successive natural image frames are formed as:

$$V_{XYT} = \{ I_{XY}(t), t=1, 2, 3 \dots \} \quad (3.1)$$

where I_{XY} represents a natural image with axes of X and Y. Therefore, video coding is an extension of image coding. There are two types of techniques for such an extension as discussed in Subsection 2.1.1, i.e., extension with and without inter-frame prediction.

Both of the two types of extension techniques mentioned above have considered the physical meaning of each axis, and they treat X and Y axes equally with each other (as spatial axes) and differently with T axis (as temporal axis). In this chapter, we propose a novel framework of pre-processing for video coding which is different from the existing paradigms by exploring the information redundancy in a fuller extent. Rather than explicitly distinguishing T axis as temporal axis, our scheme ignores the physical meaning of X, Y and T axes (somewhat similar with 3D transform [68], [69]) and focuses on the amount of video redundancy (e.g., measured by the Pearson linear correlation coefficient (CC)) along each axis.

The key part of the proposed framework is an adaptive Optimal Compression Plane

(OCP) determination as a pre-processing module before actual video coding. Different from the traditional XY compression plane, we form frames in the adaptively determined OCP; then, a standard coding scheme is used to better remove the redundancy. In the rest of this chapter, we will justify that OCP can be used with JPEG-LS for lossless video coding and then extend the concept of OCP to many lossy video coding techniques. We also study the distribution of the prediction error in different compression planes and for different compression techniques.

The major characteristics of the research reported in this chapter are: 1) a new coding concept based upon the automatic OCP decision as the pre-processing module is demonstrated; 2) consistent improvement of RD performance can be achieved by using OCP; 3) the proposed scheme can be used with the existing standard video compression codecs (encoders and decoders) since the required pre-processing is independent of the video coding scheme to be used; 4) the additional computational complexity of this scheme is minor because the required operation is only the calculation of the correlation coefficients (CCs), and as shown later, CCs can be calculated with sampled frames to save computation. With experimental results, we confirm that the proposed framework can achieve better RD performance. Moreover, evaluations on the performance of various coding techniques provide more insight into the behavior and improvement of the proposed framework.

Although there have been some pre-processing methods [51], [70], [71] for image/video coding, none of these existing methods has explored the similar problem of this work. In [70], [71] down-sampling and pre-filtering methods for image coding were proposed based on the correlation/characteristics (e.g., direction of edge) within the image frame. In [51], the perceptual redundancy of videos has been explored by modifying the motion compensated residuals to reduce their variation (for better

compression) under the constraint of JND.

The rest of this chapter is organized as follows. In Section 3.2, we analyze video redundancy in terms of CCs. We compare the redundancy along different axes (X, Y and T), for different visual content. Section 3.3 describes the details of the proposed OCP based coding scheme. In Sections 3.4 and 3.5, we discuss how to determine the OCP with and without inter-frame prediction. The experimental results with further discussion are given in Section 3.6. We validate our OCP framework with different visual content, and consistent improvement is achieved compared with standard Motion JPEG-LS [72]-[75], Motion JPEG [76]-[78], Motion JPEG 2000 [79]-[83], H.264 intra-only profile [13], [14] and H.264 [84]-[87] (inclusive of the cases with B pictures and two reference frames). Finally, conclusions for the work presented in this chapter are drawn in Section 3.7.

3.2 Video Redundancy Analysis

A video sequence can be described as a 3D cube with axes of X, Y, and T. The amount of statistical redundancy along one axis can be estimated by the average of CCs between frames formed by the two remainder axes. The bigger the average CC is, the more the statistical redundancy exists. Let us use $f_t(i, j)$ to represent the pixel to be encoded, located at pixel position (i, j) in the t^{th} frame, then the inter-frame CC between the t^{th} and the $(t+L)^{th}$ frames can be calculated as [24]:

$$CC_{t,t+L} = \frac{\sum_{i,j} ((f_t(i, j) - \overline{f_t}) \cdot (f_{t+L}(i, j) - \overline{f_{t+L}}))}{\sqrt{\sum_{i,j} (f_t(i, j) - \overline{f_t})^2 \cdot \sum_{i,j} (f_{t+L}(i, j) - \overline{f_{t+L}})^2}} \quad (3.2)$$

where $L=0, 1, 2, \dots$, and $\overline{f_t}$ and $\overline{f_{t+L}}$ are the average values of the pixels in the t^{th} and the $(t+L)^{th}$ frames, respectively.

Figure 3.1: Four video sequences with different typical motion characteristics.

Figure 3.2: Inter-frame correlation coefficients of four typical sequences.

3.2.1 Temporal Redundancy

Experiments showed that temporal redundancy usually decreases slowly with time (i.e., redundancy along T axis), being still significant on average even between frames separated by ten or more frames [88] (except for the occurrence of scene change). Figure 3.1 has shown four publicly available video sequences with different typical motion characteristics: “Akiyo” (a talking head), “Football” (fast motion), “Mobile” (horizontal, vertical and rotational object motion coupled with camera motion), and “Tempete” (zooming out). All of them are with luminance only and QCIF (Quarter Common Intermediate Format) resolution (176×144), and from the 1st to the 128th frames. Figure 3.2 illustrates the inter-frame CCs for them. Figure 3.2 (a) shows the CCs between the t^{th} and the $(t+1)^{th}$ frames, and Figure 3.2 (b) denotes the CCs between the 1st and the t^{th} frames. It is observed from Figure 3.2 (a) that all the sequences except for the “Football” sequence have very high CC (close to 1) between two adjacent frames while the “Football” sequence has relatively lower (around 0.5 for most of the frames) inter-frame CCs. It can be noticed from Figure 3.2 (b) that CC between frames separated by ten frames (i.e., the 1st and the 11st frames) are still very high (i.e., higher than 0.6 for all the sequences except for the “Football” sequence). It is apparent that there is a potential gain if more than one past reference frames are used for prediction, and this is the ground for video coding with multiple reference frames [89]-[91].

3.2.2 Comparison of Redundancy along Different Axes

We have discussed the redundancy along the T axis (i.e., temporal redundancy) in the previous subsection, and now we consider the redundancy along X and Y axes. As shown in (3.1), pixels in a video sequence are traditionally grouped into frames in the XY plane

as I_{XY} at first, and then frames are grouped into a 3D matrix along the T axis. However, a video sequence can also be grouped as frames in the TX plane (as in (3.3) below) or the TY plane (as in (3.4) below):

$$\{ I_{TX}(h), h=1, 2, 3 \dots \} \quad (3.3)$$

$$\{ I_{TY}(w), w=1, 2, 3 \dots \} \quad (3.4)$$

C_T , C_X and C_Y are used to represent the amount of correlation (and therefore redundancy) measured along T, X and Y axes; they can be estimated by averaging inter-frame (formed in the other axes) CCs, and mathematically described as:

$$C_d = \sum_{t=2,3,\dots,L^d} CC_{(t-1),t}^d / L^d \quad (3.5a)$$

where $d \in \{T, X, Y\}$, and $CC_{(t-1),t}^d$ and L^d are the corresponding $CC_{(t-1),t}$ (as defined in (3.2)) and the number of frames when formed in the axes other than d .

Figure 3.3 has shown C_d for 18 sequences with QCIF resolution (their names and indices are shown in Table 3.1). As already mentioned, all the sequences are with luminance only, and from the 1st to the 128th frames. Note that Figure 3.2 (a) is the CCs between successive frames for four of the sequences when $d=T$, and Figure 3.3 gives the average CCs along different axes for each sequence. It can be observed from Figure 3.3 that for all the sequences except for “Football” (with Index 6) the redundancy along T axis is much more than that along X and Y axes.

3.2.3 C_d Calculation using Sampled Frames

The calculation of C_d (as given in (3.2) and (3.5a)) takes some time since all the pixels in the video are involved in the calculation, and to reduce the computational complexity, we explore the possibility of using sampled frames to calculate C_d (C_T , C_X , and C_Y).

There are 6 types of relationship among C_d , and R_k ($k=1, 2 \dots 6$) is used to represent different relationship of C_d , as shown in Table 3.2.

C_d for sampled frames is calculated as:

$$C_d = \sum_{k=\frac{1}{S}, \frac{2}{S}, \dots, \frac{L_S^d}{S}} CC_{(k-1),k}^d / L_S^d \quad (3.5b)$$

where the frame sampling ratio s is defined as:

$$s = L_S^d / L^d \quad (3.6)$$

where L_S^d is the number of sampled frames within a Pre-Processing Unit (PPU, which is the collection of frames being considered; more details and discussion to be presented in Subsection 3.3.2) when formed in the axes other than d . Note that it is also possible to down-sample the pixels in a selected/sampled frame to further reduce the complexity.

The experimental results of the relationship of C_d under different s conditions are shown in Table 3.3. In the table, the sampling in most cases (13 out of 18) yields the same result with the exhausted evaluation (i.e., when $s=1$). From Table 3.3, we see that sampled frames can be used to determine the relationship of C_d for a sequence, because the results with frame sampling ($s < 1$) are largely the same as those without sampling ($s=1$). Note that even when $s=1/8$ (the last column of Table 3.3), only 4 out of 18 sequences obtain different relationship compared with that of $s=1$ (the second column of Table 3.3), and the four mismatched video sequences (indexed as 1, 11, 14, 16) have very similar C_X and C_Y ; so the mismatch between R_1 and R_2 for these four cases would not affect the performance too much in terms of compression efficiency since the compression performance is closely related to C_d (as to be discussed in Subsection 3.3.1).

Table 3.1: Names and indices of the video sequences.

Index	Sequence name	Index	Sequence name
1	Akiyo	10	Highway
2	Carphone	11	Miss-America
3	Claire	12	Mobile
4	Coastguard	13	Mother-daughter
5	Container	14	News
6	Football	15	Salesman
7	Foreman	16	Silent
8	Grandma	17	Suzie
9	Hall	18	Tempete

Table 3.2: Relationship among c_d values.

Label	Relationship	Label	Relationship
R_1	$c_T \geq c_X \geq c_Y$	R_4	$c_X \geq c_Y \geq c_T$
R_2	$c_T \geq c_Y \geq c_X$	R_5	$c_Y \geq c_X \geq c_T$
R_3	$c_X \geq c_T \geq c_Y$	R_6	$c_Y \geq c_T \geq c_X$

Figure 3.3: Average inter-frame correlation coefficients along T, Y and X axes.

Table 3.3: Relationship of c_d with different s conditions (shaded: s conditions with different calculated relationship of CCs compared with that of $s=1$).

Video sequence	s			
	1	1/2	1/4	1/8
1	R ₁	R ₁		R ₂
2	R ₂	R ₂		
3	R ₂	R ₂		
4	R ₁	R ₁		
5	R ₁	R ₁		
6	R ₄	R ₄		
7	R ₁	R ₁		
8	R ₁	R ₁	R ₂	R ₁
9	R ₁	R ₁		
10	R ₁	R ₁		
11	R ₂	R ₁		
12	R ₁	R ₁		
13	R ₂	R ₁		
14	R ₂	R ₁		
15	R ₁	R ₁		
16	R ₁	R ₁	R ₂	
17	R ₂	R ₂		
18	R ₁	R ₁		

3.3 Proposed Framework

3.3.1 Optimal Coding Plane and C_d

The intra-frame redundancy can be measured by C_A and C_B for a frame formed in AB plane ($A, B = X, Y, T$): the larger the values of C_A and C_B are, the higher the intra-frame redundancy exists; therefore, as to be discussed hereafter, the higher the RD performance will be with AB plane without inter-frame prediction. In Figure 3.4, we have demonstrated the RD curves for video sequences with different visual content and frame formation. Each frame is coded independently by JPEG. From Figure 3.3 and Figure 3.4, we can see that the RD performance is in accordance with the C_d criterion (the larger C_A

and c_B are, the higher the RD performance is achieved when frames are formed in AB plane). For example, for “Mobile” (with Index 12) sequence, the RD performance in the TX plane is better than that of the XY plane since c_T is significantly bigger than c_Y for this sequence (as shown in Figure 3.3).

In summary, we can form frames in a non-XY plane before we compress each frame because c_X and c_Y may not happen to be the largest two among the three values of c_d .

3.3.2 Overall Scheme

As shown in Figure 3.5, our OCP-based video coding framework consists of selection of appropriate Pre-Processing Unit (PPU; its size, i.e., the number of frames in the XY plane is denoted as N_{PPU}), adaptive OCP decision, and video coding with a standard compression method. In each PPU, we may form frames in the XY, TX or TY plane. It is obvious that we need 2 bits of overhead to represent the three possible coding planes for each PPU, and this bit overhead is included in our rate calculation throughout this work.

The insight observation for this proposed framework is that the video data acquisition is as usual but we ignore the physical meaning of a video during video coding for better coder performance. As mentioned earlier, video is nothing more than a three dimensional data matrix, and the distinction among X (a spatial dimension), Y (the other spatial dimension), and T (the temporal dimension) is not absolutely necessary, in the viewpoint of compression. The adaptive plane selection according to video content makes sense and the coding system benefits from it, as will be demonstrated later in this chapter.

Figure 3.4: Rate-Distortion performance for sequences with different frame formation (without inter-frame prediction).

Figure 3.5: Block diagram of the proposed scheme (illustrated with XY and non-XY frames of “Mobile” video sequence for better visual impression).

3.3.3 Impact of PPU Size

With the PPU in our design, there is an initial coding delay since we can only start coding after the first PPU rather than the first frame, and this is somewhat like the method with hierarchical B pictures shown in Figure 2.1 [15] where the encoding delay is the GOP size minus one. Therefore, our OCP based coding method works well for video transmission or storage but may not be used for interactive applications. Larger PPUs could be more efficient for exploring the temporal redundancy (this point will be further demonstrated in Figure 3.10 and discussed toward Figure 3.11~Figure 3.13 in Subsection 3.6.3), while the drawbacks of larger frame memory requirement and longer initial coding delay will surface. A larger PPU is also more likely to encounter scene change among the frames. One possible good solution is to conduct scene detection (as in [92]) for video and set each scene as a PPU.

3.3.4 Computational Complexity

Since the proposed OCP is a pre-processing step prior to the actual coding, some additional computational complexity is needed to do the pre-analysis. However, this computational overhead is very small, since what we need to do is just to calculate c_d by using some sampled frames (as already presented in Subsection 3.2.3). For instance, the computational complexity overhead is only about 2.58% and 0.22% compared with Motion JPEG-LS and H.264 when sampling ratio S is 1/8, on a PC with 2.40 GHz Intel Core2 CPU and 2 GB of RAM.

3.4 OCP without Inter-frame Prediction

When there is no inter-frame prediction (i.e., coding each frame independently when the computational resource is limited), redundancy can be only reduced by intra-frame coding [23-28], the axes with bigger redundancy (i.e., larger C_d) should be the intra-frame directions, and therefore, the OCP is determined as:

$$OCP = \begin{cases} XY & \text{if } \min\{C_T, C_X, C_Y\} = C_T \\ TX & \min\{C_T, C_X, C_Y\} = C_Y \\ TY & \min\{C_T, C_X, C_Y\} = C_X \end{cases} \quad (3.7a)$$

In Table 3.4, we have shown the OCP determined by using (3.7a) for video sequences in Table 3.1. To demonstrate the effectiveness of the OCP determination, we have also shown the bpps for lossless compressed videos with different types of frame formation (XY, TX, and TY); each frame is coded independently by JPEG-LS [72], [75]; P_{min} represents the plane with minimal bpp. The saving of bits is defined as:

$$\text{Saving} = (\text{bpp}(XY) - \text{bpp}(q)) / \text{bpp}(XY) * 100\% \quad (3.8)$$

where $q = \text{OCP}$ or P_{min} for calculating the saving in Table 3.4. Apparently, the ideal OCP is P_{min} . From Table 3.4, we can see that there are only three out of the 18 sequences having OCP different from P_{min} ; note that the difference between $\text{bpp}(P_{min})$ and $\text{bpp}(\text{OCP})$ for these three sequences are less than 2.5%, and the average saving of bits for P_{min} and OCP is very close (being 22.1% and 21.9%, respectively). Therefore, (3.7a) is effective for OCP determination.

Figure 3.6 and Figure 3.7 further demonstrate the effect of compression for the OCP, with the histogram of the intra-frame prediction residues (by JPEG-LS) and the histogram of the DCT coefficients (by JPEG) for the four typical sequences in XY plane and TX one (note that the OCPs for the four typical sequences except for “Football” are

all TX), respectively. We can observe from these two figures that for the test sequences with TX as its OCP, the number of residues/coefficients with large magnitude decreases for TX plane, while the number of residues/coefficients with zero or small magnitude increases (i.e., the histogram for TX plane is sharper than that for the XY plane). For JPEG-LS, smaller prediction residues mean that we need fewer bits to code them; for DCT based image/video coding, smaller magnitude coefficients mean more zeros after quantization (with approximately the same quantization level and quantization error for different compression plane), and fewer bits needed for run length coding and entropy coding.

Table 3.4: Bits per pixel (bpp) for lossless compressed videos under different frame formation.

Index	OCP	bpp			Saving (%)	
		XY	TX	TY	P_{min}	OCP
1	TX	3.205	1.598	1.622	50.1	
2	TY	3.619	3.455	3.537	4.5	2.3
3	TY	2.479	1.926	1.900	23.4	
4	TX	5.058	4.556	5.388	9.9	
5	TX	4.234	2.334	2.507	44.9	
6	XY	5.233	5.561	6.501	0.0	
7	TX	4.492	4.126	4.281	8.2	
8	TX	3.922	2.637	2.608	33.5	32.8
9	TX	3.899	3.136	3.156	19.6	
10	TX	3.518	3.239	3.590	7.9	
11	TY	2.698	2.572	2.550	5.5	
12	TX	5.925	5.040	5.263	14.9	
13	TY	3.490	2.592	2.564	26.5	
14	TY	3.845	2.150	2.149	44.1	
15	TX	4.487	2.567	2.861	42.8	
16	TX	4.413	2.597	2.566	41.9	41.1
17	TY	3.613	3.450	3.265	9.6	
18	TX	6.220	5.550	5.845	10.8	
mean		4.130	3.283	3.453	22.1	21.9

Figure 3.6: Distribution of the intra-frame prediction (JPEG-LS) residues.

Figure 3.7: Distribution of the DCT (Motion JPEG) coefficients.

3.5 OCP with Inter-frame Prediction

3.5.1 Brute-force OCP Determination

The most accurate determination of coding plane should adopt a brute-force method (i.e., to form frames in all three possible coding planes, code the video and select the best coding plane according to a joint RD criterion). However, given a fixed QP, coding in XY, TX, and TY would result in different bit rates and PSNR; it is very hard to compare the RD performance directly. Therefore, we design an RD cost function to measure the RD performance for different coding plane in analogy to the one used in H.264 MB level for mode decision as discussed in (2.1)~(2.3) in Subsection 2.1.2. In H.264, the Lagrange multiplier λ_{RDO} is as (2.2) and it is equivalent to:

$$\lambda_{RDO} = \alpha_1 \cdot 2^{Q_p/\alpha_2} + \alpha_3 \quad (3.9)$$

with $\alpha_1 = 0.0531$, $\alpha_2 = 3$, $\alpha_3 = 0$. In our scheme, we take the same form of RD cost function but with different parameters (to be determined next) in (3.9) since our RD cost function is used in coding plane level rather than MB level.

Hence, there are three parameters α_1 , α_2 , and α_3 to be determined. An experiment similar with the one used to determine parameter in (3.9) for H.264 was performed to test Q_p for some selected values for λ_{RDO} . This experiment is described as follows. For a given λ_{RDO} (denoted as λ_v) and a video, the corresponding value of Q_p (denoted as Q_λ) is searched among all possible values of Q_p (i.e., 0-51), to achieve the minimal RD cost with J_{RDO} as in (2.1):

$$Q_\lambda = \arg \min_{Q_p} J_{RDO}(Q_p) \quad (3.10)$$

Twelve video sequences are used, and then we have twelve Q_λ values for every λ_v .

Figure 3.8 shows the relative frequency of the chosen Q_λ for different values of λ_v , and as can be seen, λ_v varies with seven values: 0.13, 0.35, 1.7, 9, 110, 910, and 8300, producing seven normalized histograms for the chosen Q_λ that are plotted in Figure 3.8. As can be seen from Figure 3.8, the value of Q_λ does not vary much given a fixed value of λ (i.e., each histogram in Figure 3.8 has narrow horizontal span), indicating that the relationship in (3.9) is nearly independent of video content.

Finally, the least mean squared error method is used to get the best fitted values of $(\alpha_1, \alpha_2, \alpha_3)$ as:

$$(\alpha_1, \alpha_2, \alpha_3) = \arg \min_{Q_\lambda} \sum [\lambda_v - \lambda_{RDO}(Q_\lambda)]^2 \quad (3.11)$$

As a result, $\alpha_1 = 0.0226$, $\alpha_2 = 2.4187$, $\alpha_3 = 0.0738$.

Therefore, the OCP with the designed RD cost function (3.9) and (3.11) (denoted as OCP^h) is described as:

$$OCP^h = \arg \min_{P_{RDO}} J_{RDO}(Q_p, P_{RDO}) \quad (3.12)$$

where P_{RDO} is the possible coding plane, i.e., $P_{RDO} \in \{XY, TX, TY\}$; and $J_{RDO}(\cdot)$ is not only a function of QP as in H.264 but also a function of coding plane P_{RDO} since in our scheme, both $D_{RDO}(\cdot)$, and $R_{RDO}(\cdot)$ in (2.1) relate to the actual coding plane.

Figure 3.8: Relative frequency vs. quantization parameter (Q_p) for various values of the Lagrange multiplier λ_{RDO} .

Figure 3.9: Percentage of intra modes for H.264 coding in XY, TX and TY planes.

3.5.2 Efficient OCP Prediction

Although the most accurate determination of coding plane is as discussed in Subsection 3.5.1, its computational complexity is about three times of the standard coding schemes. Therefore, we are to propose effective and more efficient OCP determination methods.

With the XY plane for inter coding (i.e., the standard H.264), BME is very effective and most frames are inter-coded. For a non-XY plane for inter coding, BME still brings about some benefit (although BME does not carry its original physical meaning in this case) but is not as effective as for the XY plane; consequently, most “frames” are intra-coded. To confirm this point, we have shown the percentage of intra modes for P frames when coded in XY, TX and TY planes for “Akiyo” sequence in Figure 3.9. From the figure, we can see that the percentage for intra mode is more than 50% for non-XY planes. The possible reason for lesser intra modes in TX/TY plane at lower bit rate is that the quantized prediction residues are almost zeros at a lower bit rate (i.e., larger quantization value) and most of the given bits are used to represent the overhead

information (e.g., coding modes), and inter modes need less overhead bits and therefore it is more likely to be selected at a lower bit rate.

Therefore, two major factors to determine the overall performance comparison of the two schemes (the proposed scheme and the traditional XY plane coding) are the coding efficiency of BME in the XY plane and the coding efficiency of intra-coding in a non-XY plane. The former is good when motion vectors can be estimated well, while the latter is usually good (as shown in the results of intra-only video coding) and becomes even better with video of low motion; please refer to the comparison of Akiyo against Football in Figure 3.4.

We modified (3.7a) and add $C_T - \max\{C_X, C_Y\} > T_{BME}$ into the criteria of selecting a non-XY plane as OCP. Here T_{BME} (chosen as 3.5×10^{-2} in the current work) is a constant threshold which accounts for the advantage of BME in the XY plane; this makes TX or TY plane only to be chosen when C_T is sufficiently high. Therefore, the OCP with inter-frame prediction is formulated as:

$$OCP = \begin{cases} TX, & \text{if } C_X \geq C_Y \text{ and } C_T - \max\{C_X, C_Y\} > T_{BME} \\ TY, & \text{if } C_X < C_Y \text{ and } C_T - \max\{C_X, C_Y\} > T_{BME} \\ XY, & \text{otherwise} \end{cases} \quad (3.7b)$$

As can be seen from (3.7b), there is a higher chance for the OCP to choose a non-XY plane when T_{BME} is smaller.

3.6 Experimental Results and Discussions

We have already discussed the determination of OCP with and without inter-frame prediction in Section 3.4 and Section 3.5. In this section, we will provide extensive experimental results to evaluate the overall effectiveness of OCP with a number of video coding techniques ranging from lossless to lossy coding.

In Subsection 3.6.1 below, we show the results for OCPs with and without inter-frame prediction. In Subsection 3.6.2, the performance of OCP with Motion JPEG-LS for lossless and near-lossless video coding is shown. For low complexity lossy video coding, the results of Motion JPEG [9], [10], Motion JPEG 2000 [11], [12] and H.264 intra-only profile [13], [14] with OCP are shown in Subsection 3.6.3. Finally, we will present the results of OCP with H.264 in Subsection 3.6.4. The video database with 18 standard sequences introduced in Table 3.1 is used to demonstrate the effectiveness and efficiency of allowing TX and TY plane options, with an additional “Tempete” sequence in SD resolution in Subsection 3.6.4. For all the video sequences to be used in this section, only the first 128 luminance frames (XY plane) in the standard sequences are used.

3.6.1 OCP under Different Conditions

With the aforementioned video sequences, there is only one PPU for each sequence if $N_{PPU}=128$ and there are four PPUs if $N_{PPU}=32$ (PPU1 to PPU4 represent the first to the fourth PPUs, respectively). In Table 3.5, we show the OCP for each PPU of all 18 sequences (as listed in Table 3.1), with and without inter-frame prediction.

As we can see from Table 3.5, the OCP for different PPUs (when $N_{PPU}=32$) is largely the same, and this represents the cases where there is no obvious scene change. The only exception is “Miss-America” (with Index 11), and the slightly different OCP results are caused by different motion in PPU1/PPU2/PPU4 (only with facial movement) and PPU3 (there also being some shoulder movement).

Nearly all the video sequences have a non-XY OCP for the case without inter-frame prediction since temporal redundancy can be explored when coding performs in a non-XY plane. For the case with inter-frame prediction, OCPs for high motion sequences switched to the traditional XY plane since BME is more efficient to reduce the temporal

redundancy than intra-frame prediction in TX (or TY) plane, as already analyzed in Subsection 3.5.2.

(a)

(b)

Figure 3.10: (a) Average saving of bits and (b) overhead bit rate vs. Pre-Processing Unit (PPU) size N_{PPU} .

Table 3.5: OCPs without (with) inter-frame prediction.

Index	$N_{PPU}=32$				$N_{PPU}=128$
	PPU1	PPU2	PPU3	PPU4	
1	TX (TX)				TX (TX)
2	TY (XY)				TY (XY)
3	TY (XY)				TY (XY)
4	TX (XY)				TX (XY)
5	TX (TX)				TX (TX)
6	XY (XY)				XY (XY)
7	TX (XY)				TX (XY)
8	TX (TX)				TX (TX)
9	TX (TX)				TX (TX)
10	TX (XY)				TX (XY)
11	TY(TY)	TY(TY)	TY(TY)	TY(TY)	TY (TY)
12	TX (TX)				TX (TX)
13	TY (TY)				TY (TY)
14	TY (TY)				TY (TY)
15	TX (TX)				TX (TX)
16	TX (TX)				TX (TX)
17	TY (XY)				TY (XY)
18	TX (TX)				TX (TX)

Table 3.6: Results of dP and saving of bits for Motion JPEG-LS (near-lossless) for OCP.

Index	Near=1		Near=2		Near=3	
	dP (%)	saving of bits (%)	dP (%)	saving of bits (%)	dP (%)	saving of bits (%)
1	-0.2	53.3	-0.3	54.1	-0.5	54.7
2	0.2	5.9	0.4	10.4	0.6	13.5
3	1.0	44.2	1.3	49.6	1.2	51.6
4	0.0	13.6	0.0	15.2	0.1	14.4
5	0.5	52.6	1.0	56.2	1.0	57.9
6	0.0	0.0	0.0	0.0	0.0	0.0
7	0.1	10.2	0.3	12.1	0.3	13.4
8	0.6	46.8	1.5	52.1	1.7	54.1
9	0.0	27.9	0.9	39.5	1.7	46.4
10	0.1	11.3	0.4	17.2	1.0	25.4
11	0.3	19.9	1.3	29.9	1.3	32.1
12	0.1	20.2	0.2	24.2	0.2	26.6
13	0.7	40.3	1.3	43.5	1.4	44.5
14	0.1	48.7	0.1	50.7	0.7	51.7
15	0.5	50.0	1.3	53.2	1.5	54.3
16	0.5	46.0	1.0	46.9	1.2	47.1
17	0.5	17.8	0.8	19.3	0.7	19.3
18	0.0	14.8	0.0	19.3	0.1	22.5
mean	0.3	29.1	0.6	33.0	0.8	35.0

3.6.2 Performance Comparison with Motion JPEG-LS

In Section 3.4 we have shown with Table 3.4 the performance of lossless coding in OCP (in terms of bpp saving of the proposed method with respect to Motion JPEG-LS), and it can be seen that the proposed scheme outperforms Motion JPEG-LS by 21.9% on average.

Figure 3.10 (a) shows the average (over the same 18 sequences) saving of bits versus the PPU size. From the figure, we can see that PPU size $N_{PPU}=32$ is a good choice considering the trade-off among compression efficiency, memory requirement and initial coding delay, since the RD performance improvement saturates when $N_{PPU} > 32$. The curve for the saving of bits versus the PPU size is similar for all the 18 sequences except for “football”, the saving of bits for “football” is always zero since the OCP for the “football” is the traditional XY plane. For all the other video sequences, the saving of bits is always saturated around the N_{PPU} size of 32. The reasons are:

(a) Smaller PPUs are usually with poorer RD performance since a smaller PPU size is less efficient for exploring the temporal redundancy. Therefore, the saving of bits is increased with the increase of the PPU size.

(b) With a smaller PPU size, the resolution for a non-XY plane is smaller than that of the XY plane, and this leads to larger overhead bit rate (in terms of bpp) for the frame header: the size of the frame header increases very slowly with the increase of frame resolution; therefore, the smaller the frame resolution is, the larger the overhead bit rate becomes for the frame header. The overhead bit rate for JPEG lossless coding is shown in Figure 3.10 (b), from which we can see that the overhead bit rate tends to be constant when the N_{PPU} size is around 32. As a consequence, the saving of bits is saturated around the N_{PPU} size of 32.

(a) Akiyo

(b) Football

(c) Mobile

(d) Tempete

(e) Average gain (PSNR_{OCP} - PSNR_{XY})

Figure 3.11: Results for Motion JPEG.

(a) Akiyo

(b) Football

(c) Mobile

(d) Tempete

(e) Average gain (PSNR_{OCP} - PSNR_{XY})

Figure 3.12: Results for Motion JPEG 2000.

Figure 3.13: Results for H.264 intra-only profile.

Now we come to near-lossless video coding using Motion JPEG-LS. The results shown in Table 3.6 present the coding performance ($N_{PPU}=32$) over each video sequence. “Near” is one input parameter for JPEG-LS, representing the allowed maximal distortion for each pixel (e.g., “Near=1”: the maximal distortion for each pixel is 1). The definition of the saving of bits is the same as that defined in (3.8), and the PSNR difference dP of the OCP is defined as:

$$dP = (PSNR_{OCP} - PSNR_{XY}) / PSNR_{XY} * 100\% \quad (3.13)$$

From the table, we can see that when coding in OCP, we can save bits on average by 29.1%, 33.0% and 35.0% for the distortion levels “Near=1”, “Near=2” and “Near=3”, respectively, with almost the same PSNR (dP being only 0.3%, 0.6% and 0.8%, respectively). The results demonstrate that the proposed OCP method works well with Motion JPEG-LS for near-lossless video coding.

3.6.3 Performance Comparison with Motion JPEG, Motion JPEG 2000 and H.264 Intra-only Profile

Motion JPEG codes each frame of a digital video sequence separately as a JPEG image, and the Matlab function [imwrite] with appropriate parameters is used for JPEG image compression in the experiments in this study; similar to Motion JPEG, Motion JPEG 2000 (J2K) codes each frame separately as a J2K image, and in our experiments, each frame was coded by Jasper [81], which implements the codec specified in the J2K Part-1 standard (i.e., ISO/IEC 15444-1 [79]); MPEG and ITU-T have recently added intra-only profiles to AVC/H.264 to respond to strong demands for high-quality video applications [13]-[14]. Therefore, we will also show the performance comparison of coding in OCP with the H.264 intra-only configuration.

To discuss the impact of PPU size (as the continuation of Subsection 3.3.3), two sets of PPU size (i.e., $N_{PPU}=32$ and $N_{PPU}=128$) were used, and there is only one PPU for each sequence if $N_{PPU}=128$ and there are four PPUs if $N_{PPU}=32$. Figure 3.11 to Figure 3.13 have given the RD curves (XY; OCP with $N_{PPU}=32$; and OCP with $N_{PPU}=128$) for the four typical sequences and the curve of average PSNR gain (i.e., $PSNR_{OCP}-PSNR_{XY}$), with the three coding techniques (namely Motion JPEG, Motion J2K and H.264 intra-only profile), respectively.

From Figure 3.11 to Figure 3.13, we can see that the RD performance with PPU size being 128 is always better than that of PPU size being 32 for non-XY planes (note that the OCP for “Football” (with Index 6) is the traditional XY plane and therefore has the same RD performance for different PPU sizes); this is expected since more temporal redundancy is explored when we use a larger PPU size.

Table 3.7: PSNR gain of OCP at 0.8 bpp against Motion JPEG, Motion J2K, and H.264 intra-only profile (I264).

Index	Motion JPEG	Motion J2K	I264
1	10.3	13.7	15.8
2	2.7	2.0	2.6
3	10.0	3.7	5.9
4	3.1	2.7	3.6
5	12.9	12.2	13.8
6	0.0	0.0	0.0
7	2.5	2.3	2.2
8	9.1	9.8	11.4
9	8.6	8.3	9.2
10	4.1	1.2	3.1
11	3.7	0.6	2.4
12	4.9	6.2	6.9
13	7.0	3.1	8.6
14	8.8	10.8	14.8
15	8.6	11.0	15.5
16	6.6	8.6	12.8
17	1.8	0.6	2.8
18	4.4	4.9	5.9
mean	6.1	5.7	7.6

From Figure 3.11 (a) (c) and (d) we can also see that when the PPU size of 32 is used, the RD curve of the non-XY OCP may fall below that of the traditional XY plane for the cases of a large QP (i.e., low bpp). The two reasons are as follows.

(a) Smaller PPUs are usually less efficient for exploring the temporal redundancy. Figure 3.11~Figure 3.13 have shown the RD performance for two PPU sizes (i.e., 32 and 128) for different video sequences and coding schemes, and it is demonstrated that a smaller PPU size is with poorer RD performance.

(b) With a smaller PPU size, the resolution for a non-XY plane is smaller than that of the XY plane, and this leads to larger overhead bit rate (in terms of bpp) for the frame header: the size of the frame header increases very slowly with the increase of frame resolution; therefore, the smaller the frame resolution is, the larger the overhead bit rate for the frame header becomes. Consequently, when the PPU size is small and QPs are large (i.e., for low bit rate coding), the overhead bit rate due to the frame header becomes significant in comparison with that for coding the visual content itself, and this is the other reason why the RD curve of a non-XY plane with the PPU size of 32 may fall below that of the traditional XY plane for a large QP. In the case of Figure 3.11, the bpp for the intersection point (for the RD curve in the XY plane and that in a non-XY plane) with PPU size of 32 is just slightly larger than the overhead bit rate for a non-XY plane (i.e., around 0.6 bpp); so in this case, nearly all bits are used for the frame header and low RD performance is obtained for a non-XY plane.

To discuss the impact of video content toward the PSNR gain, we list the detailed PSNR gain for different video content and coding techniques in Table 3.7 ($N_{PPU}=128$, bit budget is 0.8 bpp, and I264 represents the H.264 intra-only profile). It can be noticed that OCP without inter-frame prediction works better for video having slow object moving

speed and less moving areas (i.e., there are more stationary areas), since TX and TY options are adopted in these situations, as expected.

3.6.4 Performance Comparison with H.264

When inter-frame prediction is allowed in H.264, coding in a non-XY plane still has benefits as analyzed in Subsection 3.5.2. By using (3.7b), we find that 11 out of 18 test sequences are with non-XY OCPs. In our simulation, PPU size of 128 is used and the configuration (denoted as IP hereinafter) for H.264 is: the high profile of JM 16.0, encoding in gray scale, GOP size being 16 and with one reference frame, Simplified UMHexagon Search with search range of 16, no B frame, and all the other parameters being the default setting of high profile (e.g., high complexity RDO and CABAC coding).

In Figure 3.14 we have shown the results for four sequences. Figure 3.14 (a) is with low motion, so the non-XY plane coding performs better, while in Figure 3.14 (b) where the video is of high but predictable motion, the non-XY plane coding does not show its advantage since BME performs effectively in the XY plane. The videos in Figure 3.14 (c) and (d) share something in common: high and complex motion (objects move along different directions and the video is with camera zooming); since BME does not perform effectively in the XY plane with these two cases, the non-XY plane coding demonstrates its advantage.

The PSNR difference between the proposed scheme and H.264 is a function of bit rate since PSNR itself is a function of bit rate. For a comprehensive evaluation of the proposed scheme, we need to measure the average PSNR (or bit rate) difference over the whole bit rate range. Therefore, BD-PSNR (or BD-Bitrate) is used as the measurement of average PSNR (or bit rate) differences, and it is calculated as follows [93].

Figure 3.14: Results for the comparison of the OCP and XY plane coding (i.e. H.264).

Figure 3.15: Simulation result for the sequence "Tempete" (a 720x486 sequence with flowers, falling leaves, and stones).

(a) Fit a curve through 4 data points in RD plot (PSNR/bitrate are obtained for QP =

16, 20, 24, 28) using a 3-order polynomial function;

(b) Based on this, find an expression for the integral of the curve/function;

(c) The average difference is the difference between the integrals divided by the integration interval.

Note that the bit rate is plotted using logarithmic scale in the calculation; otherwise the difference between the two RD curves is dominated by the high bitrates (e.g., the bit rate range (1500-2000) bits per second gets 4 times the weight of the range (375-500) even if they both represent a bit rate variation of 33%). In Step (a), we can either plot PSNR as a function of bit rate or plot bit rate as a function of PSNR, and the obtained values in Step (c) are BD-PSNR, and BD-Bitrate for the two cases, respectively.

The BD-PSNR (i.e., average PSNR increment) and BD-Bitrate (i.e., average bit rate increase rate) of the proposed scheme against H.264 (with IP configuration) are calculated for the eighteen videos and given in the second and third columns of Table 3.8. A negative BD-Bitrate indicates bit rate saving for the proposed scheme against its competitors. It has shown that coding in OCP has better or equal RD performance in comparison with the traditional XY plane except for the sequence “Miss-America”. The degradation of RD performance for “Miss-America” is due to the prediction error by using (3.7b). The average PSNR improvement (i.e., BD-PSNR) for all 18 videos is 0.89 dB (equivalently, 11.01% in average bit rate saving).

Figure 3.15 has shown the three RD curves for XY, TX, and TY coding planes for the sequence “Tempete” (a sequence with flowers, falling leaves, and stones) of SD (standard definition; 720x486 in frame size). The figure has confirmed that the traditional XY coding plane does not always achieve the best RD performance for SD video sequence, and the proposed coding method (coding in a non-XY plane) can be used for SD resolution for performance improvement.

The simulation results for other configurations of H.264 are also provided as in the last two columns of Table 3.8, where IBP represents the configuration with B-pictures and S2R represents the configuration with two reference frames, and the other parameter settings in both configurations are the same as those in IP (given in the first paragraph of this subsection). From the table, we can see that there are 0.47dB and 0.79dB BD-PSNR improvement on average compared with H.264 for the configuration with B-pictures and that with two reference frames, respectively.

3.7 Concluding Remarks

Since the compressibility of visual signal depends upon the associated signal redundancy, the study presented in this chapter has firstly investigated into the signal correlation with different compression planes. The analysis shows that coding in the traditional XY plane are not always optimal, and therefore we presented a novel framework for video coding via adaptive optimal compression plane (OCP) selection. An OCP is auto-detected by comparing the correlation coefficients (CCs) along different axes, and we have also shown that sparse frame samples and proper fast plane decision are sufficient for an efficient process.

The proposed scheme is used as a preprocessor with any existing video coding standards and with low computational complexity overhead. Extensive experiments have been conducted with different visual content, and the effectiveness of the proposed OCP-based scheme has been confirmed. The proposed method improves the RD performance significantly for various coding schemes for both lossless and lossy video coding, such as Motion JPEG-LS, Motion JPEG, Motion JPEG, Motion J2K, H.264 intra-only profile, and H.264-inter scheme.

Table 3.8: Comparison of RD performance of the proposed scheme against H.264 under the IP configuration (IP) and the configurations with B pictures (IBP) and that with two reference frames (S2R) (for 0.05~2 bpp).

Video	BD-Bitrate (%) (for IP)	BD-PSNR (dB)		
		IP	IBP	S2R
Akiyo	-19.17	2.32	1.66	2.11
Carphone	0.00	0.00	0.00	0.00
Claire	0.00	0.00	0.00	0.00
Coastguard	0.00	0.00	0.00	0.00
Container	-35.40	2.54	1.12	2.31
Football	0.00	0.00	0.00	0.00
Foreman	0.00	0.00	0.00	0.00
Grandma	-20.21	1.24	0.76	1.12
Hall	-29.17	1.69	1.37	1.39
Highway	0.00	0.00	0.00	0.00
Miss-America	15.39	-0.68	-0.87	-0.80
Mobile	-15.52	1.11	-0.95	0.74
Mother-daughter	0.00	0.00	0.00	0.00
News	-16.74	1.72	1.19	1.66
Salesman	-30.73	2.44	2.00	2.07
Silent	-9.68	0.75	0.75	0.61
Suzie	0.00	0.00	0.00	0.00
Tempete	-37.02	2.92	1.43	3.03
Average	-11.01	0.89	0.47	0.79

Chapter 4

JND Model with Separation of Edge and Textural Regions

As introduced in Subsection 2.2.2, the JND accounts for the maximum sensory distortion that the typical HVS does not perceive, and it can be adopted in perceptual image/video processing algorithms and systems. In Section 5.2, the application of the JND model in the proposed visually lossless image coding will be demonstrated. Basically, a JND estimator in pixel domain takes into account two factors: LA and CM.

However, CM has not been estimated accurately in the existing pixel domain JND models since they fail to accurately distinguish TM (CM in texture regions) from EM (CM around edges). In this chapter, we propose an enhanced pixel domain JND model with a new algorithm for CM estimation. In our model, total variation (TV) based image decomposition [108] is used to decompose an image into structural image (i.e., cartoon like, piecewise smooth regions with sharp edges) and textural image for estimation of EM and TM, respectively. Compared with the existing models, the proposed one shows its advantages brought by the better EM and TM estimation, as be demonstrated by experiments on noise shaping and visual quality evaluation.

4.1 Introduction

Edge structure is simpler and more predictable than a texture one, and a typical observer has prior knowledge (i.e., related to a specific object) about what edges should look like [106], [107]. Therefore, distinguishing between TM and EM is very important for JND estimation. One experiment in [106] has shown that three times as much quantization noise can be masked/hidden in the textural image than in the edge image without compromising perceptual quality even if both images have a similar power spectrum and background luminance level in the local regions.

TM is usually underestimated (e.g., in Yang *et al.*'s model [49]) since many textural regions tend to be classified as edge ones (to be detailed in Figure 4.4 below). In this chapter, we propose a new pixel domain JND model to alleviate the aforementioned difficulty in EM and TM estimation, caused by no or not effective separation of edge and texture. In our model, TV based image decomposition is used to decompose an image into structural image (being piecewise smooth with sharp edge) for EM estimation and textural image for TM estimation, respectively; therefore, EM and TM can be effectively distinguished. With the experimental results to be presented later, we confirm that the proposed model correlates better with the HVS perception, evidenced by the average additional PSNR redundancy of 1.15 dB over the latest relevant existing model in [47], as well as the improvement demonstrated for the visual quality evaluation application.

The rest of this chapter is organized as follows. In Section 4.2, we firstly describe the problem of image decomposition modeling. Then we present the details of the proposed JND model. We discuss how to estimate TM and EM, and why the proposed method is better than the existing ones. Experimental results are given and discussed in Section 4.3 to validate the proposed model, while Section 4.4 concludes this chapter.

4.2 The Proposed JND Model

4.2.1 Image Decomposition

Image decomposition is to split an image into two or more component images. An image f can be regarded as the sum of the structural image u (being piecewise smooth and with sharp edge along the contour) and the textural image v (only containing fine-scale details, usually with some oscillatory nature), i.e., $f = u + v$. Decomposition is important for many image processing applications, e.g., image coding, texture discrimination, image denoising, image inpainting, image registration, etc. A general approach for such decomposition is to solve the problem formulated as [108]:

$$\min_u \{ \|I_s(u)\|_{L_A} \mid \|I_t(u, f)\|_{L_B} \leq I_\sigma \} \quad (4.1)$$

where $I_s(\cdot)$ and $I_t(\cdot, \cdot)$ are two functions to be chosen, and $\|\cdot\|_{L_A}$ and $\|\cdot\|_{L_B}$ are certain norms (or semi-norms); the fidelity term $\|I_t(u, f)\|_{L_B} \leq I_\sigma$ forces u to be close to f , and $I_t(u, f)$ is often chosen to be $(f - u)$ which is ideally the texture image v . The constrained minimization problem (4.1) is equivalent to its Lagrangian relaxed form “ $\min \{ \|I_s(u)\|_{L_A} + \lambda \|I_t(u, f)\|_{L_B} \}$ over u ” since $\|I_s(u)\|_{L_A}$ and $\|I_t(u, f)\|_{L_B}$ are convex in u , where λ is the Lagrange multiplier.

The TV based method has been widely used in image structure-texture decomposition models. In such a model, $\|I_s(u)\|_{L_A}$ is chosen as $TV(u)$, where $TV(u) = \int |\nabla u|$ and ∇u denotes the generalized derivative (i.e., gradient) [108] of u . Minimizing $TV(u)$ allows u to have discontinuities; hence edges in the original image are preserved [108]. In [109], different TV based image decomposition models are considered and the model of minimizing total variation with an L^1 -norm fidelity term is shown to achieve better results; we adopt this

(TV-L¹) model in our work for image decomposition, and then (4.1) becomes:

$$\min_u \int_{\Omega} |\nabla u| + \lambda \int_{\Omega} |f - u| \quad (4.2)$$

where Ω is the domain of f .

In (4.2), the larger the λ is, the smaller the fidelity term becomes; this means that u is less smooth and closer to f . Aujol *et al.* [110] proposed the use of the correlation between u and v to estimate λ by finding a minimum of this correlation, and the typical value of λ is between 0.2 to 2 [108], [109] for most natural images. Figure 4.1 shows the said correlation with λ values for three images. As can be seen, the optimal λ is different for different images, and it is computationally expensive to find the optimal value for each image. In this work, a reasonable λ value of 0.8 (see the average curve in Figure 4.1) has been used for the sake of computational efficiency (for non-critical timing applications, the iterative method in [110] can be used).

In Figure 4.2, we have shown the experimental results of decomposition for both a one-dimensional signal and two real-world images (Barbara and Mandrill images, as shown in Figure 4.5 (a1) and (a2), respectively), with the TV-L¹ model; the value of 128 was added to the textural images for clear viewing in Figure 4.2 (b2) and (c2). From Figure 4.2 (a1)~(a3), we can see that the structural component (the boundary of change) and the textural component (the small variation) can be well decomposed for a one-dimensional signal; from Figure 4.2 (b1)~(c2), we can see that structural and the textural component images are reasonably decomposed for the real-world images.

Figure 4.1: Determination of λ for three images and the average result over the ten test images in the first column of Table 4.1.

(b1) structural Barbara image

(b2) textural Barbara image

(c1) structural Mandrill image

(c2) textural Mandrill image

Figure 4.2: Structure-texture decomposition results.

Figure 4.3: Block diagram of the proposed direct pixel domain JND model.

4.2.2 Contrast Masking in JND Model

CM is an important phenomenon in the HVS perception and is referred to the reduction in the visibility of one visual component at the presence of another [44]. Usually noise becomes less visible in the regions with high spatial variation, and more visible in smooth areas. CM in pixel domain includes TM and EM since both texture and edge exhibit high spatial variation; the value of CM on a smooth region is nearly zero since little spatial variation appears there.

Watson *et al.* [111] introduced the concept of entropy masking, which is due to unfamiliarity of the observer to visual content. Entropy masking is due to the limitation of the human brain in processing simultaneous complex phenomena [112]. As we know, texture is less predictable and more difficult to learn than edges [106]. Therefore, the effect of TM should be larger than that of EM if the corresponding regions have the same level of spatial variation.

The proposed JND model is shown in Figure 4.3. The parts enclosed with dash lines represent the modules for CM which is our contribution in this research. The other two modules (i.e., the LA module and the NAMM module) are the same with those in [49] and have been already discussed in Subsection 2.2.2 as (2.5) and (2.4), respectively. Towards CM estimation, we evaluate EM and TM on u and v , respectively, rather than on the whole input image as in the relevant existing models in the literature.

In our model, CM can be represented as:

$$T_{CM}(i, j) = T_{EM}(i, j) + T_{TM}(i, j) \quad (4.3a)$$

$$T_{EM}(i, j) = \beta_{CM} \cdot g_u(i, j) \cdot W_{EM} \quad (4.3b)$$

$$T_{TM}(i, j) = \beta_{CM} \cdot g_v(i, j) \cdot W_{TM} \quad (4.3c)$$

Figure 4.4: Detected edge information (binary image, with black pixels representing edges) for the proposed and Yang *et al.*'s models, with different threshold (t_e).

where $T_{EM}(i, j)$ and $T_{TM}(i, j)$ are EM and TM in u and v , respectively; g_u and g_v are spatial contrasts for u and v , respectively and the spatial contrast is calculated as described with (2.7) and Figure 2.6 in Subsection 2.2.2; w_{EM} and w_{TM} are used to distinguish EM from TM, $w_{EM} < w_{TM}$ since TM results in a higher JND value than EM for a same extent of variation, and their values are set as 1 and 3 [106], respectively; the constant β_{CM} is the same as in [49].

Figure 4.4 shows the detected edge for Barbara image (as shown in Figure 4.5 (a1)), with Yang *et al.*'s model (i.e., via Canny detector [113]) and the proposed one (i.e., decomposition + Canny operator), as in the first row and the second row, respectively. Let Ω_{e-p}^f and Ω_{e-Y}^f be the edge pixels in the proposed model and Yang *et al.*'s model. In the proposed model, $\Omega_{e-p}^f = \Omega_{e-Y}^u$; In Yang *et al.*'s model, Canny operator with a threshold

(denoted as t_e) is used on f . The larger the t_e value is, the less the detected edge becomes. For fair comparison, we use the same Canny operator (and the same t_e) on u in the proposed model.

From the figure, by using the proposed model, we distinguish between EM and TM better. We can see that true edges can be detected in both Yang *et al.*'s method and the proposed method. However, with Yang *et al.*'s method, some texture is detected alongside the edge with various values for t_e (e.g., the middle left part of the Barbara image in the top row of Figure 4.4). Canny operator finds intensity variation over local neighborhoods [114], and is not designed for distinguishing edge and texture. However, by separating firstly the textural image from the image we detect objects' true edges more reliably. Therefore, the proposed method (i.e., decomposition + Canny) outperforms Yang *et al.*'s model (i.e., with only Canny operator).

To further illustrate the advantage of the proposed CM model, we have shown the CM maps for Yang *et al.*'s model and the proposed one in Figure 4.5. Two typical images (Barbara, and Mandrill) are used. From Figure 4.5 we see that there is underestimation on TM in Yang *et al.*'s model (the second row in the figure), especially with the bottom-left of the desk cloth of Barbara image, and the cheek of Mandrill image because the merely Canny-based method can hardly distinguish between edge and texture as we discussed above; on the contrary, the proposed model produced reasonable CM values.

The final JND maps from the proposed model for the two images are shown in Figure 4.6. For Mandrill image, the JND map is closer to the CM map than that for Barbara image, and this indicates that CM has a larger impact toward the JND value for Mandrill image. This is expected since the Mandrill image is with lots of texture and therefore has larger CM values.

(a1) Original Barbara image

(a2) Original Mandrill image

(b1) Model [49] with Barbara

(b2) Model [49] with Mandrill

(c1) Proposed model with Barbara

(c2) Proposed model with Mandrill

Figure 4.5: Contrast masking (CM) in different JND models (scaled to [0 255] and higher brightness means a larger masking value).

Figure 4.6: JND maps from the proposed model for two images (scaled to [0 255]).

4.3 Experimental Results and Discussions

4.3.1 Model Validation with Noise Shaping

In this section, we will compare the proposed scheme's overall performance against the two existing direct pixel domain models (i.e., Yang *et al.*'s model [49] and Chou *et al.*'s model [47]) with different images. A better JND model should be able to guide shaping (i.e., hiding) more noise into an image at a certain level of resultant perceived images quality. The noise can be regulated according to the JND profile as (2.8) in Subsection 2.2.2. At the same perceptual quality, the higher the injected-noise energy (measured by PSNR) is, the higher accuracy the JND model exhibits. Ten images (with different visual content and spatial complexity) were used in this experiment, as listed in the first column of Table 4.1.

For a comprehensive evaluation, the subjective viewing tests were conducted based on "Adjectival categorical judgement methods" recommended by ITU-R BT.500-11 standard [96]. In each test, two images of the same scene (i.e., an image processed by the proposed model and that processed by one of the other models) were juxtaposed on the screen randomly at left or right. The viewing distance was about 50 cm. Twenty one subjects who have either normal or corrected-to-normal eye sight were asked which

image (left or right) is better and how much better it is by using quantitative scores listed as Table 4.2, for all the image pairs.

The mean and standard deviation of the subjective scores of each model under comparison (i.e., Chou *et al.*'s model [47], and Yang *et al.*'s model [49]) against the proposed one are given in the second to fifth columns of Table 4.1: a negative (or positive) subjective score indicates that the image with the proposed model has better (or worse) perceptual quality than that with the model under comparison (either [47] or [49]), and its magnitude represents the extent of quality improvement (degradation). From Table 4.1, we can see that the overall subjective quality tends to be of a near-zero negative mean of -0.014 and -0.010, respectively. That is, the subjective quality for the images noised with the proposed JND model is very close (in fact, being slightly better) when compared to that with other models. From the table, we can also see that images noised with the model in [49] are slightly better than those noised with the one in [47].

Table 4.1: The subjective quality evaluation results (the proposed model against each of those in [47] and [49]) and PSNRs for 10 images with different visual content.

Image	MOS (against the image processed by the proposed model)				PSNR (dB) (as indication of perceptual redundancy)		
	Mean		Standard deviation		Proposed	Model [47]	Model [49]
	Model [47]	Model [49]	Model [47]	Model [49]			
Airplane	0.048	0.286	1.071	0.845	32.65	32.74	34.80
Barbara	0.143	0.333	1.062	0.856	29.93	31.64	31.35
Boat	0.429	-0.095	0.870	0.831	31.37	32.55	32.62
Couple	0.190	-0.048	0.750	0.590	32.43	33.57	32.71
Gold	-0.619	0.000	0.669	1.049	31.16	32.32	31.97
Lena	-0.286	-0.048	0.784	0.669	31.89	32.80	32.72
Mandrill	-0.190	-0.333	0.602	0.856	29.77	32.99	32.54
Peppers	-0.190	0.000	0.512	0.837	30.12	30.78	30.79
Splash	0.190	-0.143	0.928	1.014	30.96	31.43	31.35
Tank	0.143	-0.048	0.854	1.117	34.22	35.22	34.88
average	-0.014	-0.010	0.810	0.866	31.45	32.60	32.58
average additional redundancy (in dB) of our model						1.15	1.13

Table 4.2: Scores for subjective quality evaluation.

Subjective score	0	1	2	3
Description	same quality	only slightly better	better	much better

The PSNR of noised images is then used to measure the amount of noise being added by the JND models under comparison. The sixth to eighth columns of Table 4.1 show the PSNRs with the three JND profiles. The proposed model is better than the existing relevant JND models (and the model in [49] is better than the one in [47]), with the evidence of an average additional PSNR redundancy of 1.15 dB and 1.13 dB from the model in [47] and the one in [49], respectively, without jeopardy of visual quality (in conjunction with the subject results that we have just discussed). Since the images used in Table 4.1 are with both textual and not-so-textural visual content, the results confirm that the proposed scheme works better for both situations over the existing relevant schemes.

4.3.2 Further Validation of the Model

To further demonstrate the proposed JND model, we apply it to a JND based visual quality gauging metric defined in [115]. The metric in [115] is denoted as D (as (4.4)~(4.6) below). D evaluates and distinguishes the JND normalized noticeable contrast changes (C_{dif}) for visual distortion assessment, and it is designed for image and video with the provision of existence of enhanced edges.

If $f(i, j)$ denotes the image pixel value at position (i, j) , $\overline{f(i, j)}$ is the average of $f(i, j)$, and T_{JND} is the JND from a JND model, we can calculate the noticeable contrast as:

$$C_f^D(i, j) = \begin{cases} \frac{|f(i, j) - \overline{f(i, j)}|}{T_{JND}(i, j)} & \text{if } |f(i, j) - \overline{f(i, j)}| > T_{JND}(i, j) \\ 0 & \text{otherwise} \end{cases} \quad (4.4a)$$

We use $C_{f_1}^D$ and $C_{f_2}^D$ to represent the noticeable contrasts calculated by (4.4a) for an original image f_1 and its processed version f_2 , and the noticeable contrast change between f_1 and f_2 is:

$$C_{dif}(i, j) = C_{f_2}^D(i, j) - C_{f_1}^D(i, j) \quad (4.4b)$$

Let C_e^- , C_e^+ and C_{ne} denote the local noticeable contrast decrease for an edge pixel, the local noticeable contrast increase for an edge pixel, the local noticeable contrast change (regardless of decrease and increase) for non-edge pixels, respectively:

$$C_e^-(i, j) = \begin{cases} |C_{dif}(i, j)|, & \text{if } C_{dif}(i, j) < 0 \text{ and } (i, j) \text{ belongs to edge} \\ 0, & \text{otherwise} \end{cases} \quad (4.5a)$$

$$C_e^+(i, j) = \begin{cases} |C_{dif}(i, j)|, & \text{if } C_{dif}(i, j) > 0 \text{ and } (i, j) \text{ belongs to edge} \\ 0, & \text{otherwise} \end{cases} \quad (4.5b)$$

$$C_{ne}(i, j) = \begin{cases} |C_{dif}(i, j)|, & \text{if } (i, j) \text{ belongs to non-edge} \\ 0, & \text{otherwise} \end{cases} \quad (4.5c)$$

Finally, D is formulated as:

$$D = \alpha_{D1} \cdot \frac{\sum_i \sum_j C_e^-(i, j)}{n_e} - \alpha_{D2} \cdot \frac{\sum_i \sum_j C_e^+(i, j)}{n_e} + \alpha_{D3} \cdot \frac{\sum_i \sum_j C_{ne}(i, j)}{n_{ne}} \quad (4.6)$$

where n_e and n_{ne} are the numbers of edge pixels and non-edge pixels in an image, respectively; $\alpha_{D1-D3} > 0$, being model parameters. For each pixel, only one of the three local contrast-change indicators (C_e^- , C_e^+ , and C_{ne}) is possibly nonzero. The indicator C_e^- detects blurring on edges, discontinuous edges and other damage on edges, and therefore has the biggest negative impact (i.e., α_{D1} is bigger than the other two). The indicator C_e^+ represents edge contrast enhancement that usually brings about positive impact on visual quality (therefore, the second term in (4.6) is negative). The third indicator C_{ne} includes the loss of image details, luminance fluctuations, in non-edge regions, and has negative impact toward visual quality. The relative valuation of α_{D1-D3} determines the significance

of the three contributing factors in the metric. In [115], $\alpha_{D1}=3.5$, $\alpha_{D2}=0.5$, $\alpha_{D3}=1.0$, and the same values are used in this work, however, it is possible to customize/optimize $\alpha_{D1\sim D3}$ for specific applications. For instance, $\alpha_{D1\sim D3}$ can be determined/tuned via training with some known visual data and the related subjective ratings, as in [150].

The image quality database in [26] is adopted (as in [115]), which gives subjective distortion scores with ten original images and four processed (i.e., demosaicked) versions for each image. Some of the demosaicked images have better perceived quality than the original images.

Since the performance of D depends upon the accuracy of JND estimation, a better JND model will lead to better image quality gauging in line with subjective viewing scores given in [26]. Denote D_P , D_Y , and D_C as the metric D with the proposed JND model, Yang *et al.*'s JND model and Chou *et al.*'s JND model, respectively. The comparison is in Table 4.3, where the fit of goodness with the subjective scores [26] is quantitatively calculated using Pearson correlation coefficient (CC, for prediction accuracy), Spearman rank order correlation coefficient (SROCC, for monotonicity), and root mean squared error (RMSE) between the subjective distortion scores and the predictive ones (CC, SROCC and RMSE will be introduced in more detail in Subsection 6.5.1). For a perfect match between the objective and subjective scores, $CC=SROCC=1$ and $RMSE=0$. A better quality metric has higher CC and SROCC, and lower RMSE.

From Table 4.3 (with 95% *confidence interval* (CI) [160] indicated for CC and SROCC), it has been found that the metric based on the proposed JND model performs best among all the approaches in terms of CC (prediction accuracy) and RMSE; and SROCC (monotonicity) with the proposed model is better than Chou *et al.*'s one and only slightly lower than that of the best one (i.e., Yang *et al.*'s model).

Overall, we have demonstrated again that the proposed JND model yields better overall performance than the existing relevant JND models.

4.4 Concluding Remarks

In this chapter, a new direct pixel domain JND estimator has been proposed. After an in-depth analysis on the existing relevant approaches, the TV based image decomposition model is then used to improve the accuracy of CM evaluation (via decomposition of an image into structural and textural images), as an attempt for overcoming the major drawback of the existing models. Proper analysis and justification have been given for the proposed model, and subjective viewing tests (both carefully-designed and publicly accessible ones) have confirmed the effectiveness of the proposed schemes.

At the same level of perceptual quality, the proposed model allows more data redundancy (lower PSNR), and therefore, allowing higher QP (equivalent to fewer bits assigned) in image and video coding [51], [53]. It can be applied to other image manipulation tasks as well, like motion estimation [49], watermarking [157], image enhancement [158] and video retargeting [159].

Table 4.3: Prediction performance for different approaches.

Metric	CC (with CI)	SROCC (with CI)	RMSE
D_C	0.8950 ([0.8091, 0.9435])	0.8161 ([0.6766, 0.8991])	17.13
D_Y	0.8885 ([0.7978, 0.9399])	0.8238 ([0.6891, 0.9035])	18.97
D_P	0.9028 ([0.8226, 0.9478])	0.8173 ([0.6785, 0.8997])	13.21

Chapter 5

Perceptual Image Coding

Subsection 2.1.4 and Subsection 2.1.3 have introduced down-sampling based coding and lossless coding in the literature. In this Chapter, the application of perceptual models for better coding performance is to be demonstrated in perception based down-sampling (for low bit rate image coding) and JND based histogram adjustment (for visually lossless image coding), in Section 5.1 and Section 5.2, respectively.

In Section 5.1, the proposed down-sampling scheme (to be more specific, the sampling directions/ratios and the corresponding QPs) is adaptive to not only the local visual content but also the human visual attention. Therefore, it has yielded more desirable compressed visual content than the other sampling based coding methods (and also in comparison with coding methods without down-sampling). In addition, the proposed scheme also automate the switching of the sampling modes for variable bit rate applications.

In Section 5.2, the proposed visually lossless coding scheme is focused on visually lossless image pre-processing, and then any existing or emerging lossless coding method can be adopted. To reduce the perceptual redundancy, the pre-processing step is modeled as an optimization problem and the zero order image entropy is selected as the minimization objective function. This minimization problem is also under the constraint

that limits the resultant error to be not larger than the JND threshold as introduced in Chapter 4 to guarantee that such a process is visually lossless.

5.1 Perception based Down-sampling

As introduced in Subsection 2.1.4, a down-sampling based coding scheme can be used for low bit rate image coding (e.g., a down-sampled image when JPEG compressed visually beats the compressed high-resolution image with the same number of bits as illustrated in Figure 2.2). In [30], an adaptive sampling method is proposed which adaptively decide the appropriate down-sampling mode and the QP for every MB in an image, based upon the local visual content of the signal. However, the method in [30] is based on only the local visual content to decide the sampling mode and does not consider other characteristics of the HVS. Aiming at aligning better with the HVS perception (especially in terms of the global aspects within an image) to achieve better perceived quality for the compressed visual content, in this section, we propose an improved scheme for determination of the down-sampling mode and the corresponding QP, with a computational visual attention model (as discussed in Subsection 2.2.1) being considered. The attention map, obtained from a bottom-up human attention model and the human face detection (which provides higher-level cognitive information), is used in the proposed method.

In Subsections 5.1.1 to 5.1.3 that follow, the detail of the proposed perception based model for sampling mode and QP determination will be described, and the experimental results with further discussion are given in Subsection 5.1.4.

5.1.1 System Overview and Visual Attention Determination

Figure 5.1: Block diagram of the down-sampling based coding method (the parts enclosed with dash lines) and the inclusion of the proposed perception-based module.

The block diagram of the proposed scheme is shown in Figure 5.1, where the parts enclosed with dash lines are the typical block diagram of a down-sampling based coding method which has already been shown in Figure 2.3. In Figure 2.3, the mode selection part is only affected by the quantization results (to be more specific, the RD of the quantization residues) while in Figure 5.1 this part also controlled by the visual attention (or Region of Interest (ROI)) map which can be used to improve the efficiency and effectiveness of the mode selection and the corresponding QP determination processes.

In this work, Itti's model [40] was used for the bottom-up visual attention map (denoted as A_{bu}) calculation since it is one of the most well-known (with more than 2000 citations) models and is still popular (adopted in the existing research [165]-[170]). However, other bottom-up visual attention models (e.g., the ones as introduced in Subsection 2.2.1) can also be used in the proposed system.

For an input image f , nine spatial scales are created using dyadic Gaussian pyramids [11], which progressively low-pass filter and sub-sample the input image, yielding horizontal and vertical image-reduction factors ranging from 1:1 (scale zero) to 1:256 (scale eight) in eight octaves. The intensity information $I_{VA}(l_{VA})$ (where $l_{VA} = \{0, 1, \dots, 8\}$ represents the scale) is the image intensity values at each pyramid level. The orientation information $O_{VA}(l_{VA}, \theta_{VA})$ (where $\theta_{VA} = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ represents the used orientation) is calculated by convolving the intensity pyramid with oriented Gabor filters.

The feature maps are computed by a set of linear “center-surround” operations, which are implemented as the difference between fine and coarse scales: The center is a pixel at scale $c_{VA} \in \{2, 3, 4\}$, and the surround is the corresponding pixel at scale $s_{VA} = c_{VA} + \delta_{VA}$, with $\delta_{VA} \in \{3, 4\}$. The across-scale difference between two maps, denoted as “ $\ominus$ ”, is obtained by interpolation to the finer scale and point-by-point subtraction. Using several scales not only for c_{VA} but also for δ_{VA} yields truly multi-scale feature extraction, by including different size ratios between the center and surround regions. Therefore, intensity feature maps $I_{VA}(c_{VA}, s_{VA})$ and orientation feature maps $O_{VA}(c_{VA}, s_{VA}, \theta_{VA})$ are calculated as follows:

$$I_{VA}(c_{VA}, s_{VA}) = |I_{VA}(c_{VA}) \ominus I_{VA}(s_{VA})| \quad (5.1a)$$

$$O_{VA}(c_{VA}, s_{VA}, \theta_{VA}) = |O_{VA}(c_{VA}, \theta_{VA}) \ominus O_{VA}(s_{VA}, \theta_{VA})| \quad (5.1b)$$

At this point, 6 intensity feature maps and 24 orientation feature maps are created. Each group of feature maps is combined into a conspicuity map through across-scale addition “ $\oplus$ ” by down-sampling each map to scale 4 of the pyramid followed by point-by-point addition. Finally, the bottom-up visual attention map A_{bu} is calculated by averaging over the two normalized ($\aleph(\cdot)$) conspicuity maps as below:

$$A_{bu} = \frac{1}{2} \left(\bigoplus_{c_{VA}, s_{VA}} \aleph(I_{VA}(c_{VA}, s_{VA})) + \bigoplus_{c_{VA}, s_{VA}, \theta_{VA}} \aleph(O_{VA}(c_{VA}, s_{VA}, \theta_{VA})) \right) \quad (5.1c)$$

The overall visual attention map is mathematically described as:

$$A_f(i, j) = \beta_A A_{td}(i, j) + (1 - \beta_A) A_{bu}(i, j) \quad (5.1d)$$

where (i, j) represents the pixel position in f ; A_{td} is the attention map from the top down mechanism (the face stimulus is used [36], as already discussed in Subsection 2.2.1); β_A is a weighting parameter for the relative importance of the two contributing factors. In (5.1), a higher A_f value indicates a higher visual attention level for the pixel (as a result of combining the top down and bottom up stimuli), and therefore a higher probability for the pixel to belong to the ROI.

In this study, we treat the first 20% of the MBs with the highest attention levels given by A_f in (5.1) as the ROI and deem the other ones as the non-ROI. Let T_{MB} denote the type of MB. If an MB belongs to the ROI, $T_{MB}=1$; otherwise, $T_{MB}=0$.

In [30], the sampling mode is selected, based upon the local visual content of the signal. Sampling along the direction of high spatial variations (which signal the existence of edges and other image features with great impact on perceptual visual quality) is avoided. Four different down-sampling modes (denoted as k) are used for a 16x16 MB, with sampling ratios varying from 1/4 to 1, as listed in Table 5.1, where N_k represents the ratio of number of pixels in an MB before sampling over that after sampling. With various down-sampling directions/ratios, the quantization step size can be reduced to code the DCT coefficients for the down-sampled pixels more accurately. We will firstly discuss how to estimate the quantization step size for each sampling mode in the next subsection.

5.1.2 QP Determination

The plot between QP and average bpp has been shown as the “original curve” in Figure

5.2, with the JPEG coding for six images (Airplane, Couple, Gold, Mandrill, Peppers and Tank) from USC-SIPI Image Database [95]. The relationship between the average number of the used bits, R_{MB} , and QP, Q_{MB} , for low bit rate coding, illustrated in Figure 5.2, can be fitted using a second-degree polynomial with the form:

$$R_{MB}(Q_{MB}) = \alpha_1 + \alpha_2 * (1/Q_{MB}) + \alpha_3 * (1/Q_{MB})^2 \quad (5.2a)$$

or its inverse function

$$Q_{MB}(R_{MB}) = \alpha_4 + \alpha_5 * (1/R_{MB}) + \alpha_6 * (1/R_{MB})^2 \quad (5.2b)$$

where $\alpha_1 = 0.117$, $\alpha_2 = 0.619$, $\alpha_3 = 0.202$, $\alpha_4 = -0.021$, $\alpha_5 = -0.112$, and $\alpha_6 = 0.396$, being the model parameters, and these parameters were calculated by fitting (using the Matlab function [fminsearch]) the functions (5.2) with the “original curve” in Figure 5.2, where the fitted results are also shown.

Table 5.1: Different down-sampling modes (indexed by k).

k	Description	N_k
0	one 16x16 MB to four 8x8 blocks (no down-sampling)	1
1	one 16x16 MB to two 8x8 blocks (1/2 down-sampling in horizontal direction)	2
2	one 16x16 MB to two 8x8 blocks (1/2 down-sampling in vertical direction)	2
3	one 16x16 MB to one 8x8 blocks (1/2 down-sampling in both directions)	4

Figure 5.2: QP vs. average bpp

Table 5.2: Candidate QP list.

k	candidate QP list: Q_k^e	Number of candidates
0	$Q_0^e = \begin{cases} 1.05Q_0, & \text{if } T_{MB}=0 \\ 0.80Q_0, & \text{if } T_{MB}=1 \end{cases}$	2
1	$\begin{cases} \{Q_1^e, Q_1^e+1, Q_1^e+2\}, & \text{if } T_{MB}=0 \\ \{Q_1^e-1, Q_1^e, Q_1^e+1\}, & \text{if } T_{MB}=1 \end{cases}$	4
2	$\begin{cases} \{Q_2^e, Q_2^e+1, Q_2^e+2\}, & \text{if } T_{MB}=0 \\ \{Q_2^e-1, Q_2^e, Q_2^e+1\}, & \text{if } T_{MB}=1 \end{cases}$	4
3	$\begin{cases} \{Q_3^e, Q_3^e+1, Q_3^e+2, Q_3^e+3, Q_3^e+4\}, & \text{if } T_{MB}=0 \\ \{Q_3^e-1, Q_3^e, Q_3^e+1, Q_3^e+2, Q_3^e+3\}, & \text{if } T_{MB}=1 \end{cases}$	6

The goodness of fit is quantitatively evaluated using Pearson correlation coefficient (CC, for prediction accuracy) and Spearman rank order correlation coefficient (SROCC, for monotonicity) [130] between the actual curve and the estimation by (5.2). The two correlation coefficients should be with the value of 1 when an ideal match is found; in this study, for (5.2a) and (5.2b), respectively, CCs are 0.9817 and 0.9868; SROCCs are 1 for both, and this is expected since the functions (i.e., (5.2a) and (5.2b)) used for fitting are monotonic (for a larger value of QP, the actual bpp value is always smaller and this is also the case for the fitted bpp value).

Thus, if the initial QP value when no down-sampling is performed is Q_0 , the corresponding bit rate is estimated as $R_{MB}(Q_0)$. For Mode k ($k=0,1,2,3$), the allowed bit rate is $N_k * R_{MB}(Q_0) - R_o$ (where R_o is the required overhead bit rate, as to be discussed next), and the corresponding QP (Q_k^e) can be estimated as:

$$Q_k^e = Q_{MB}(N_k * R_{MB}(Q_0) - R_o), (k=0,1,2,3) \quad (5.3)$$

where $R_{MB}(\cdot)$ and $Q_{MB}(\cdot)$ are the functions determined as (5.2a) and (5.2b), respectively; the actual QP (with the target bits associated with initial QP value of Q_0) should be

around that is estimated by (5.3), and therefore, we have chosen the candidate QPs to be around Q_k^e ($k=0,1,2,3$). It can be found that Q_0^e in (5.3) is visual content independent and its value is close to Q_0 (the initial QP used for the standard JPEG) since R_o is small.

Aiming at adapting the QP to the visual content of an MB, we designed the candidate QP list according to the type of the MB as shown in Table 5.2. We have assigned a bigger Q_0^e for a non-ROI block and smaller Q_0^e for an ROI block as in (5.4):

$$Q_0^e = \begin{cases} 1.05Q_0, & \text{if } T_{MB}=0 \\ 0.80Q_0, & \text{if } T_{MB}=1 \end{cases} \quad (5.4a)$$

$$Q_k^e = Q_{MB}(N_k * R_{MB}(Q_0^e) - R_o), (k=1,2,3) \quad (5.4b)$$

where T_{MB} denotes the type of the MB as already discussed in Subsection 5.1.1 according to the visual attention model.

As can be seen from Table 5.2, there are totally 16 different combinations of the sampling modes and the corresponding QPs. So we need 4 more bits to represent the side information for each MB, and therefore, $R_o = 4/16/16 = 0.0156$.

5.1.3 Sampling Mode Determination

For each down-sampling Mode k ($k=0,1,2,3\dots$) and the candidate QPs in Table 5.2, we calculate the coding distortion of the MB in terms of MSE ε_l and count the corresponding number of nonzero quantized DCT coefficients NZC_l (where $l=1,2,3,\dots,16$, representing the different combinations of the sampling modes and the corresponding QPs for an MB). Note that exploiting the linear relationship between the bit rate and nonzero coefficients as presented in [94], instead of using actual number of coded bits, we use NZC_{ini} (the number of nonzero quantized DCT coefficients for Mode 0) as the reference for searching the best combination. The notation $\varepsilon_{\min}$ is used to keep track of

the minimum MSE ever achieved in the course of searching, and its initial value is ε_{ini} .

We evaluate each mode and the corresponding candidate QPs in turn to choose the best mode and QP satisfying the following two conditions:

$$NZC_l < NZC_{ini}, \quad \varepsilon_l < \varepsilon_{ini} \quad (5.5)$$

Inequality (5.5) is used to choose the mode with lower MSE but with no increase the used bits. Once (5.5) is met, the ε_{min} is updated as ε_l . The mode decision arrives at Mode 0 (i.e., without down-sampling, and a new QP) when all the combinations in Tables 5.1 and 5.2 fail to yield smaller NZC_l and ε_l at the same time; this occurs more often when the bit rate is high because down-sampling loses its advantage there. With the aforementioned mode decision, the switching of the sampling modes is automated for variable bit rate applications, with graceful transition to the standard coding method (i.e., without down-sampling).

5.1.4 Experimental Results and Discussions

Other four images (Barbara, Boat, Lena and Splash — different images from those used in Subsection 5.1.2 for parameter fitting) are used for the experiments and the comparison with the down-sampling based method earlier [30]. All of the images are of 512x512 in size; color images were converted to black and white ones using the Matlab function [rgb2gray]. We used two point averaging for 1-D down-sampling (for Modes 1 and 2, since these are one-dimensional down-sampling), and four-point averaging for 2-D down-sampling (for Mode 3, since this is two-dimensional down-sampling). In the experimental results to be presented below, the additional overheads (i.e., 4 bits per MB) for representing the combination of the sampling mode and the corresponding QP are included in the bpp calculation for the proposed method.

Figure 5.3 compares the PSNR over the four images versus bit rate for the standard JPEG method, the model in [30] and the proposed method. We can observe from Figure 5.3 that the curve of the model in [30] and that of the proposed method nearly match. This is expected since the difference of the proposed method over the model in [30] is actually the re-allocation of bits according to the visual attention, so the overall resultant PSNR should be very similar; we will demonstrate the advantage of the proposed method with the subjective quality viewing tests that are to be presented next. From Figure 5.3, we can also see the automated, graceful transition to the standard JPEG coding when the bit rate becomes higher.

In light of the aforementioned difference of the two models under discussion, it is more meaningful to compare the subjective quality of decoded images. Figure 5.4 shows the amplified face and shoulder regions of image Lena, which is coded at 0.105 bpp; the proposed method has perceptual quality improvement.

To comprehensively evaluate if the proposed method have improvement in perceived quality against the method in [30], we have conducted subjective viewing tests following the procedures similar to the Simultaneous for Continuous Evaluation (SDSCE) method in Rec. ITU-R BT.500 [96], with participation of N_T ($N_T=11$) observers. The images were coded with the same bpp for each scene and shown with HP 19' LCD. The observers were asked to view the images in a room with the typical lab illumination, and the viewing distance was about 50 cm. In each test, two images of the same scene (i.e., image processed by using method in [30] and that processed by using the proposed method) were juxtaposed on the screen, subjects viewed two images for four seconds before they were asked to indicate which image has better visual quality (or if they have the same quality). The images appeared randomly on the left- or right-hand side of the

screen to avoid the possible bias.

This type of tests is a simplified version of the subjective tests used in Subsection 4.3.1, and it is used because subjects find it is easier to just indicate which image is better in perceived quality than evaluating different quality levels of images.

Figure 5.3: Comparison of Different Models in terms of PSNR vs. bpp.

Table 5.3: Subjective viewing results.

Image	N_B	N_E	N_W	$(N_B + N_E - N_W)/N_T$
Barbara	6	3	2	63.6%
Boat	9	2	0	100%
Lena	9	1	1	81.8%
Splash	7	4	0	100%

(a) Method in [30] (26.54 dB)

(b) Proposed method (26.87 dB)

Figure 5.4: Reconstructed images by using the method in [30] and the proposed method, under a same bit rate (at 0.105 bpp).

Let N_B , N_E and N_W represent the numbers of subjects who rate the images by using the proposed method better, equal, and worse than the ones processed by using [30], respectively; note that $N_B + N_E + N_W = N_T$. The values of them for the four images are listed in Column 2 to 4 of Table 5.3, respectively. We have also calculated the value of $N_B + N_E - N_W$; the higher this value is, the higher reliability the subjects rate that the images processed by using the proposed method are better than or equal to the ones processed by using [30]. The percentages of $(N_B + N_E - N_W)/N_T$ for the tested images are listed in the last column of Table 5.3, and from which it can be seen that the proposed method has better or equal coding performance judged by the majority (larger than 60% for the worst case) of human observers.

5.2 Visually Lossless Coding

As introduced in Subsection 2.1.3, lossless coding finds its applications in many visual signal processing, such as medical imaging, satellite remote sensing, and film archiving and studio applications. However, lossless image coding only considers the spatial redundancy of the image itself, and we may consider also the human vision properties in the course of coding for better coding performance (i.e., reduced bitrates) in the circumstances where the HVS is the ultimate receiver and appreciator of processed visual signal. In the cases where visually lossless coding is needed, the reconstructed signal is required to be visually the same rather than mathematically the same as the original one. Existing methods of visually lossless coding have been discussed in Subsection 2.2.3.

In this section, we will propose a novel visually lossless coding scheme which is applicable to any existing or emerging lossless coding framework. The proposed scheme is used as image pre-processing (i.e., image histogram manipulation under the guidance of the JND model) rather than embedded manipulations in an encoder (like the existing work surveyed in Subsection 2.2.3). In the rest of this section, the detail of the proposed scheme will be introduced in Subsections 5.2.1 and 5.2.2. Subsection 5.2.3 will present the experimental results and the corresponding discussions.

5.2.1 Image Histogram Adjustment based on JND

The idea of the proposed scheme is to reduce the perceptual redundancy prior to coding (rather than in coding). This redundancy reduction step can be modeled as a minimization/maximization process, and the objective function can be image entropy, image smoothness, and so on. Of course, this process should be with the constraint, so that the resultant error should not exceed the associated JND value to guarantee that the

process is visually lossless. In our current implementation, the zero-order image entropy (which is calculated without consideration of spatial correlation in an image) is used for demonstration, since it gives the theoretical limit in terms of information for memoryless signal according to Shannon's source coding theory [97]. Although an image is not memoryless and there exists correlation among neighboring image pixel values, the concept of image entropy is still often used in image/video processing, such as motion estimation [98], image enhancement [99], image alignment [100], feature selection [101], edge detection [102], watermarking [103], and similarity measurement [104].

The zero-order entropy of an image can be adopted in this study as a guide to reduce visual details (so that more compression is achieved) without being noticed, with the help of a JND model; it can be calculated based on the image histogram, which is the occurrence information of all the intensity values in the image. Pun [105] calculated the entropy of an 8-bit gray scale image f as

$$E_p(f) = - \sum_{k=0}^{255} p_k \log p_k \quad (5.6)$$

where p_k is the probability of occurrence of the k^{th} gray level intensity.

Let the size of the image f be $H_f \times W_f$ and the pixel value at (i, j) is $f(i, j)$, where $0 \leq f(i, j) \leq 255$, $1 \leq i \leq H_f$ and $1 \leq j \leq W_f$. Denote a reconstructed image of f to be $\hat{f}$ with pixel value of $\hat{f}(i, j)$; the pixel domain JND profile of f is $T_{JND}(i, j)$ which is determined by the JND model described in Chapter 4.

In lossless coding, $\hat{f}(i, j)$ must be exactly the same as $f(i, j)$. However, this is not the case in visually lossless coding. When $\hat{f}(i, j) \neq f(i, j)$, we consider that the two images have the same perceptual quality as long as the relation

$$|f(i, j) - \hat{f}(i, j)| \leq T_{JND}(i, j) \quad (5.7)$$

is satisfied. This means that the image pixels can vary within the visibility threshold $T_{JND}(i, j)$.

The proposed perceptual redundancy reduction step (i.e., zero-order entropy minimization) is modeled as:

$$\min(E_p(f)), \quad |f(i, j) - f(i, j)| \leq T_{JND}(i, j) \quad (5.8)$$

Suppose that we change one pixel in f , $f(i, j) = a_f$, to $f(i, j) = a_k$ in f . At this pixel point, $f(i, j) = a_f$ being a constant because a_f belongs to the original image so it is fixed in this minimization process, and $f(i, j) = a_k$ being a variable because a_k is any value that the pixel can change to; therefore, we have

$$P_{a_f} = P_{a_f} - \Delta, \quad P_{a_k} = P_{a_k} + \Delta, \quad \Delta = \frac{1}{H_f * W_f} \quad (5.9)$$

$$E_p(f) = E_p(f) + P_{a_f} \log P_{a_f} - P_{a_f} \log P_{a_f} + P_{a_k} \log P_{a_k} - P_{a_k} \log P_{a_k} \quad (5.10)$$

At the right-hand side of Equation (5.10), the first three terms are constant and the last two terms can be rewritten as

$$P_{a_k} \log P_{a_k} - (P_{a_k} + \Delta) \log (P_{a_k} + \Delta) \quad (5.11)$$

The function $H(f)$ monotonically decreases with P_{a_k} since its first derivative, $\log\left(\frac{P_{a_k}}{P_{a_k} + \Delta}\right)$, is smaller than zero. Therefore, to minimize this function, we must choose the maximum P_{a_k} . According to the definition of JND, we cannot sense the difference between f and f if $|a_k - a_f| \leq T_{JND}(i, j)$. In other words, f is visually lossless if (5.7) is satisfied. With this constraint, when we choose a_k as the one which has the largest P_{a_k} , the entropy of f is minimized.

5.2.2 Iterative Implementation

For the convenience for the following description, we reorder the histogram $\{P_k\}$ of f in the ascending order as $\{P_{a_0}, P_{a_1}, \dots, P_{a_{255}}\}$ ($a_0, a_1, \dots, a_{255} \in \{0, 1, \dots, 255\}$ and they are different one another); thus $0 \leq P_{a_0} \leq P_{a_1} \leq \dots \leq P_{a_{255}} \leq 1$. Therefore, the above discussion can be generalized to all the pixels, and the minimization process in (5.8) is implemented as follows:

(a) begin: change all the pixel values that satisfy (5.7) to the value with the largest probability (i.e., a_{255}), then iterate as (b) with $k=2$.

(b) the k^{th} ($k = 2, 3, 4, \dots$) iteration: for the remaining pixels (i.e., pixels with values other than $a_{255}, a_{254}, a_{253}, \dots, a_{255-(k-1)}$), change all the pixel values that satisfy (5.7) to the value a_{255-k} .

(c) $k=k+1$ and go to (b): the iteration repeats, and the number of remaining pixels reduces.

(d) end: the iteration terminates when there is no remaining pixel.

In Figure 5.5, the image Boat is shown in (a), (b) is the histogram-adjusted version of (a), and (c) and (d) are the histograms of (a) and (b), respectively. As can be seen, Figure 5.5 (a) and (b) are visually the same since we have ensured that the differences between (a) and (b) is no larger than the visibility threshold for every pixel. From the figure, we can also see that (d) is easier to be lossless compressed than (c) since the histogram in (d) is sparser than that in (c). That is, the visual detail in the image is reduced so that its compressibility improves, at no price of image quality (thanks to the constraint (5.7) by the JND).

Figure 5.5: An example for the proposed scheme.

Table 5.4: Required bits for different coding schemes.

Image	Lossless JPEG			H.264 Lossless		
	Original	Histogram adjusted	Saving (%)	Original	Histogram adjusted	Saving (%)
Airplane	4.72	3.63	23.07	4.76	3.69	22.57
Barbara	6.20	5.20	16.13	6.17	5.62	8.83
boat	5.69	4.90	13.94	6.26	5.50	12.21
Couple	5.16	4.29	16.78	5.05	4.23	16.21
Gold	5.40	4.20	22.26	6.04	5.09	15.74
Lena	5.17	4.13	19.97	5.01	3.90	22.05
Man	5.57	4.31	22.59	6.16	5.07	17.59
Peppers	5.16	4.06	21.30	5.25	4.29	18.37
Splash	4.28	3.13	26.88	3.84	2.58	32.90
Tank	5.06	4.52	10.66	6.07	5.20	14.25
Average	5.24	4.24	19.36	5.46	4.52	18.07

5.2.3 Experimental Results and Discussions

In order to demonstrate the effectiveness of the proposed histogram adjustment scheme for visually lossless coding, we have conducted experiments using ten gray scale images [21], with different visual content. The JND map of each image was calculated. We show the required bit rate for the original images and that for their histogram-adjusted versions (as described in Subsections 5.2.1 and 5.2.2) in Table 5.4, benchmarked with both Lossless JPEG encoder (implemented with Matlab [imwrite] function) and the H.264 lossless coding mode.

All the images tested require fewer bits for the histogram-adjusted version no matter which coding standard is used as shown in Table 5.4. On average, there are 19.36% and 18.07% decrease in required bits for Lossless JPEG and H.264 Lossless coding, respectively.

To evaluate if the modified images have noticeable changes against the original versions, we have conducted subjective viewing tests for all images shown in Table 5.4, with participation of N_T ($N_T=11$) observers who have either normal or corrected-to-normal eye sight. The testing condition and the procedure are the same as introduced in Subsection 5.1.4. The same notations N_B , N_E and N_W are used to represent the numbers of subjects who rate the images by using the proposed method better, equal, and worse than the original ones. Their values are collected during the subjective viewing tests.

We evaluate the final result of the subjective quality comparison using:

$$d_B = N_B - N_W \quad (5.12a)$$

and

$$D_B = \frac{d_B}{N_T} \times 100\% \quad (5.12b)$$

where d_B indicates the disparity for the subjects toward the perceptual quality of the two versions of the images, while D_B represents the normalized disparity.

When $d_B=0$ (i.e., $N_B=N_W$), there are equal numbers of subjects rating the original image better and worse, respectively, than the modified one, and this indicates that the subjects have no clear preference to any one of the two image versions. A negative value of d_B represents the case in which the modified image has a worse perceptual quality than the original one. The bigger the absolute value of d_B is, the bigger disparity the subjects have toward the perceptual quality of the two versions of the images. The result of the subjective viewing tests is shown in Table 5.5.

From Table 5.5, we can conclude that subjects' disparity in viewing quality is small (less than 20% in D_B for the worst scenario, and very close to zero on average). In other words, statistically, the HVS cannot sense the quality difference of the processed images by the histogram adjustment to minimize the zero order entropy. As can be seen from Table 5.4 and the results of D_B in Table 5.5, the proposed scheme can achieve around 19% for JPEG-LS (or around 18% for H.264 lossless) in saving of required bits without jeopardizing the perceptual quality.

Table 5.5: Subjective viewing results.

Image	d_B	D_B
Airplane	0	0
Barbara	1	9.09%
boat	0	0%
Couple	1	9.09%
Gold	-1	-9.09%
Lena	-2	-18.18%
Mandrill	1	9.09%
Peppers	-1	-9.09%
Splash	-1	-9.09%
Tank	1	9.09%
Average	-0.1	-1%

5.3 Concluding Remarks

Perceptual models can be adopted to reduce perceptual redundancy in visual signal coding. We have demonstrated two cases of possible applications: visual attention based down-sampling and JND based histogram-adjustment, in this chapter.

The sampling based image coding is better to be performed with the incorporation of a computational visual attention model, while most of the previous methods of sampling based coding consider very little about this property of the HVS. In Section 5.1, we have proposed a perception based process for sampling mode selection and the corresponding QP determination; different modes are chosen for different visual content to achieve the highest coding quality under the same bit rate constraint. The experimental results showed consistently that the perceived quality of the compressed visual content has been improved, with the proposed perception based approach.

Visually lossless coding needs to keep the coding error not noticeable to the HVS. In Section 5.2, we have proposed an image pre-processing scheme for codec independent visually lossless image coding, based on image histogram adjustment. The histogram of an image is first adjusted to minimize the zero-order image entropy, within the constraint under which the resultant pixel error is not bigger than the associated value from the JND model, and then a standard lossless image codec can be applied. Experiments with different visual content and image codecs demonstrate that with lossless JPEG and H.264 lossless coding, the required bitrate for images can be reduced (around 19% and 18%, respectively) without subjective viewing quality being affected.

Chapter 6

A New Method for Visual Quality

Evaluation

In this chapter, we propose a new image quality assessment (IQA) scheme, with emphasis on gradient similarity. Gradients convey important visual information and are crucial to scene understanding. Using such information, structural and contrast changes can be effectively captured. Therefore, we use the gradient similarity to measure the change of contrast and structure in images.

Apart from the structural/contrast changes, image quality is also affected by luminance changes which must also be accounted for complete and robust IQA. Hence, luminance changes are also considered in the proposed scheme to effectively assess image quality. Furthermore, we have made our scheme to follow the masking effect and visibility threshold more closely when both masked signal and masking signal are small.

Finally, the effects of the changes in luminance and contrast-structure are integrated via an adaptive method to obtain the overall image quality score. Extensive experiments conducted with six publicly-available subject-rated databases (comprising of diverse images and distortion types) have confirmed the effectiveness, robustness, and efficiency

of the proposed scheme in comparison with the relevant state of the art schemes.

6.1 Introduction

Digital images are usually affected by a wide variety of distortions during acquisition and processing which generally result in loss of visual quality. Therefore, IQA is useful in many applications like image acquisition, watermarking, compression, transmission, restoration, enhancement, reproduction, etc. As explained in Chapter 1, the goal of IQA is to calculate the extent of quality degradation and is thus used to evaluate/compare the performance of processing systems and/or optimize the choice of parameters in processing. For example, the well-cited SSIM index [63] has been used in image and video coding [116], [117].

The HVS is the final receiver of the majority of processed images, and evaluation based on subjective experiments (as formally defined in ITU-R Recommendation BT.500 [96]) is the most reliable way of IQA if there are a sufficient number of subjects. However, subjective evaluation is time-consuming, laborious, expensive, and not easily-repeatable; it cannot be easily and routinely performed for many scenarios, e.g., selection the prediction mode in H.264. These drawbacks have led to the development of objective IQA measures which can be easily embedded in image processing systems.

The simplest and most widely used IQA scheme is the MSE/PSNR. It is popular due to its mathematical simplicity as well as being easy for use in various optimization processes. It is however well known that MSE/PSNR does not always agree with the subjective viewing results especially when the distortion is not additive in nature [118]. This is not surprising given that MSE/PSNR is simply based on an average of the squared pixel differences between the original and distorted images.

As introduced in Subsection 2.2.4, there are many IQA schemes proposed in the recent ten years and they have been demonstrated to outperform MSE/PSNR. Among them, the SSIM [63] is the most widely accepted due to its reasonably good evaluation accuracy [119]-[120], pixel-wise quality measurement, and simple mathematical formulation which facilitates analysis and optimization. However, as pointed in some existing works [121]-[123], the SSIM is less effective for badly blurred images (actually the SSIM is less capable of predicting the relative quality of blurred image and image with white noise; more details are to be discussed in Subsection 6.3.2) since it underestimates the effect of edge damage and treats every region in image equally. It is well known that edges are crucial for visual perception and play a major role in the recognition of image content [124]-[126]. In [115], it has also been demonstrated that edge information and differentiated distortion at the edges are important for perceptual quality gauging. Another example for the significance of edges comes from the fact that a mere sketch image can convey most information in the scene [124].

Therefore, in this chapter, we propose an IQA scheme which is based on the edge/gradient similarity and demonstrate its effectiveness by careful analysis and extensive experiments. The proposed approach is expected to work well with various distortion levels since it has considered both the high level property of the HVS (i.e., the HVS is highly adapted to the structure information and this mainly accounts for the supra-threshold distortion) and the low level masking property of the HVS (which mainly accounts for the near threshold distortion). The main contributions of the research work in this chapter in comparison to the relevant existing work are as follows: 1) we demonstrate that gradient information is able to capture both contrast and structure of the image, and by using gradient similarity we allows more emphasis on distortions around edge regions in the proposed IQA scheme, leading to more accurate assessment of image

quality; 2) our scheme matches better with the masking effect of the HVS, especially for the cases when both masked signal and masking signal are small; 3) we devise an adaptive approach to integrate the different components (i.e., luminance and contrast-structure) of distortion.

The rest of this chapter is organized as follows. In Section 6.2, we provide a brief introduction of the SSIM, as well as some SSIM based schemes including the existing gradient based methods which are modified from the SSIM, as the ground for the following analysis, discussion, and comparison. Section 6.3 describes the details of the gradient similarity in the proposed scheme. In Section 6.4, we discuss how to combine the luminance change and contrast-structural change. The experimental results with further discussion are given in Section 6.5. We will validate our IQA scheme with six IQA databases for benchmarking (with totally 3752 images and more than 20 distortion types). Performance improvement is achieved, compared with the state of the art schemes (e.g., the SSIM, multi-scale SSIM, VIF, VSNR, MAD, and IW-SSIM). Finally, concluding remarks out of this work are drawn in Section 6.6.

6.2 Structural Similarity (SSIM) Index and Related Schemes

As proposed in [63], the SSIM assumes that natural images are highly structured, and the human vision is sensitive to structural distortion. The structural information in an image is defined as those attributes that represent the structure of objects in the scene, independent of the average luminance and contrast [63].

The SSIM is calculated for each overlapped image block by using pixel-by-pixel

sliding window, and therefore, it can provide the distortion/similarity map in pixel domain. It has also been extended using the multi-scale analysis [127], complex wavelets [128], and discrete wavelets [129]. For any two image blocks x and y , the SSIM models the distortion/similarity between them as three components: luminance similarity, contrast similarity, and structural similarity, and these three components are mathematically described as (6.1)~(6.3) below, respectively:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (6.1)$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (6.2)$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (6.3)$$

where μ_x , μ_y , σ_x^2 , σ_y^2 , and σ_{xy} is the mean of x , the mean of y , the variance of x , the variance of y , and the covariance of x and y , respectively; C_1 , C_2 , and C_3 are claimed as small constants to avoid the denominator being zero.

The SSIM for the image blocks is given as

$$SSIM(x, y) = [l(x, y)]^{\beta_l} \cdot [c(x, y)]^{\beta_c} \cdot [s(x, y)]^{\beta_s} \quad (6.4)$$

where β_l , β_c and β_s are positive constants used to adjust the relative importance of the three components. The higher the value of $SSIM(x, y)$ is, the more similar the image blocks x and y are. Obviously if y (assumed to the image block under evaluation) and x (assumed to the reference block) are the same, (6.1)~(6.4) will all give the value of 1, indicating no change in the block y with respect to the block x . The overall image quality score is determined using the mean of the SSIM for each image block [63] or calculated as the information content weighted average of the local SSIM [130].

Figure 6.1: Block diagram of the proposed scheme.

The schemes in [121]-[123], [131] are also based on the SSIM and take into account the importance of edge. In these existing schemes, one or more components of the SSIM are changed to calculate the value in the edge domain (note that (6.1)~(6.3) are calculated in the pixel domain); for example, the structure comparison component is changed in [121], [131] to the edge domain, while both contrast and structure comparison components are changed in [122]; in [123], [131], the luminance comparison component is not included. As simple variations from the SSIM, these schemes lack due consideration of the HVS' masking and visibility characteristics.

6.3 Proposed Gradient Similarity Scheme

The block diagram of the proposed scheme is shown in Figure 6.1, where the parts enclosed with dash line are discussed in this section and the remaining parts are discussed in Section 6.4.

6.3.1 Gradient Similarity

The proposed gradient similarity is defined as:

$$g(x, y) = \frac{2g_x g_y + C_4}{g_x^2 + g_y^2 + C_4} \quad (6.5a)$$

(a0) Original image from LIVE

(a1) JPEG image from LIVE (a2) white Gaussian noise image from LIVE

(b1) magnified version of block in (a1)

(b2) magnified version of block in (a2)

DMOS=83.6, SSIM=0.769, $g=0.787$

DMOS=51.4, SSIM=0.674, $g=0.795$

(c1) SSIM map for (b1)

(c2) SSIM map for (b2)

(d1) quality map from (6.5a) for (b1)

(d2) quality map from (6.5a) for (b2)

Figure 6.2: An illustration of the difference between the SSIM and the proposed scheme.

where g_x and g_y are the gradient value for the central pixel of the image blocks x and y , respectively, and c_4 is the small constant as c_2 in (6.2) to avoid the denominator being zero (e.g., $c_4=10^{-5}$); $g(x,y)$ is the gradient similarity between x and y , and its value lies in $[0\ 1]$. The initial form of the proposed scheme in (6.5a) is mathematically similar to the luminance/contrast term of the SSIM, and as will be demonstrated, it is more effective than that in the SSIM.

An example is shown in Figure 6.2. The original, JPEG and white Gaussian noise images have been taken from the LIVE database [132]. In this figure, (a0) is the original image; (a1) is the JPEG compressed image; and (a2) is the images with white noise; (b1)~(b2) are the amplified versions of the blocks corresponding to the same region, as highlighted in (a1)~(a2), respectively: (b1) is an edge (to be exact, due to a false edge caused by JPEG compression) block, and (b2) is a noisy block; (c1) and (c2) show the SSIM map (darker grey levels represent larger distortion) for the image blocks whose central pixels are within (b1) and (b2), respectively; (d1) and (d2) are the quality map from (6.5a) for the image blocks whose central pixels are within (b1) and (b2), respectively. Note that (c1) is almost white except for edges, and this means that the error caused by JPEG compression is mainly due to the introduced false edges; on the other hand, (c2) is with low brightness, this represents that the SSIM is quite sensitive to white noise in smooth region (note that the image block indicated by the block in (a0) is smooth), and it is a case where SSIM overestimates the white Gaussian distortion.

As we can see from Figure 6.2, blocking in (a1) is much more annoying than white noise in (a2). This is also confirmed by the subjective DMOS (Difference Mean Opinion Score) values which are 83.6 for (a1) and 51.4 for (a2) in the LIVE database for the whole images. Note that a smaller DMOS indicates higher quality. However, Figure 6.2

(c1) and (c2) have similar brightness (i.e., the SSIM value) for the distortion regions, and this is not consistent with the HVS perception (i.e., the DMOS values). The quality values for (b1) and (b2) are given in the images for (c1) and (c2), respectively, and they are similar. On the contrary, as shown in (d1) and (d2), the proposed $g(x, y)$ by (6.5a) can measure the relative quality loss (mainly the quality loss around edge and that in non-edge regions) better than the SSIM, with the evidence that the given value for the block (b1) by (6.5a) is much lower than that for (b2). This is the reason why the quality scores from (6.5a) for (a1) and (a2) are more consistent with the DMOS values compared with the SSIM scores. The effectiveness of the proposed scheme will be further demonstrated through extensive experiments described in Section 6.5.

The gradient value g_x (same for g_y) is calculated as the maximum weighted average difference for the block as described with (2.7) and Figure 2.6 in Subsection 2.2.2. The Sobel or Prewitt operator is not used since the kernels are too small (3 by 3) to include sufficient neighboring information and only two directions (horizontal and vertical) are considered.

$g(x, y)$ can be interpreted either as a block wise version (i.e., gradient similarity for image blocks x and y), or a pixel wise version (i.e., the gradient similarity for the central pixels of image blocks x and y). In the pixel version, we define

$$g(x_c, y_c) = g(x, y) \quad (6.5b)$$

where x_c and y_c are the central pixels of image blocks x and y , respectively. Note that the blocks overlap in the pixel version, in contrast with the block version.

The proposed formulation for $g(x, y)$ is able to measure both image contrast (the degree of signal variation) change and image structure (structure of objects in the scene) change since the gradient value (i.e., g_x , g_y) is a contrast and structure variant feature. To

illustrate this, firstly consider $y = \alpha \cdot x$ with α is a constant and $\alpha \neq 1$, and then we have $g_y = \alpha \cdot g_x$; therefore, $g_y \neq g_x$ since $\alpha \neq 1$, and it can be seen that g_x is contrast variant.

To further demonstrate that g_x is structure variant, we have shown in Table 6.1 the gradient values for (b1)~(b2) in Figure 6.2. The standard deviation values (i.e., σ_x , as used in the SSIM to measure the contrast change) for the image blocks are also given for a comparison. From Table 6.1, we can see that the values of σ_x for (b1) and (b2) are very close; however, (b1) and (b2) have very different structure: one is edge and the other is weak texture. As expected, the gradient values for (b1) and (b2) are different. Although both standard deviation and gradient can be used to measure the contrast, their difference is as follows. Given a group of pixel values, the standard deviation for the pixels is a constant, no matter how these pixels are positioned (i.e., independent of the pixel positioning), while the gradient values for the same group of pixels change according to the positioning of these pixels.

Note that $g(x, y)$ can be rewritten as

$$\begin{aligned}
 g(x, y) &= \frac{\max(g_x, g_y)^2 \cdot \left[2 \frac{\min(g_x, g_y)}{\max(g_x, g_y)} + \frac{C_4}{\max(g_x, g_y)^2} \right]}{\max(g_x, g_y)^2 \cdot \left[1 + \left(\frac{\min(g_x, g_y)}{\max(g_x, g_y)} \right)^2 + \frac{C_4}{\max(g_x, g_y)^2} \right]} \\
 &= \frac{2 \frac{\min(g_x, g_y)}{\max(g_x, g_y)} + K}{1 + \left(\frac{\min(g_x, g_y)}{\max(g_x, g_y)} \right)^2 + K} = \frac{2 \left[1 - \frac{|g_x - g_y|}{\max(g_x, g_y)} \right] + K}{1 + \left(1 - \frac{|g_x - g_y|}{\max(g_x, g_y)} \right)^2 + K} \\
 &= \frac{2(1 - R) + K}{1 + (1 - R)^2 + K}
 \end{aligned} \tag{6.6}$$

With the masked gradient change defined as

$$R = \frac{|g_x - g_y|}{\max(g_x, g_y)} \tag{6.7}$$

$K = C_4 / \max(g_x, g_y)^2$ and it is close to zero (when C_4 is small enough, e.g., $C_4 = 10^{-5}$ as

aforementioned); R is the gradient change (i.e., $|g_x - g_y|$) relative to the masking gradient (note that either g_x or g_y is masking signal, and here we use the larger one, i.e., $\max(g_x, g_y)$ as the masking gradient) and lies in the range of $[0, 1]$. As can be seen from (6.7), $R=0$ when $g_x = g_y$, i.e., the gradient of the test image block y is the same as that of the reference image block x ; $R=1$ when either g_x or g_y is zero (therefore $|g_x - g_y| = \max(g_x, g_y)$), i.e., when either false gradient is created from an originally smooth block or an originally non-smooth block is changed into a complete smooth one.

It is revealed from (6.6) and (6.7) that with the same amount of gradient change, the proposed $g(x, y)$ is less sensitive to the case of higher masking contrast than that of lower masking contrast, and this is consistent with the masking effect of the HVS for high masking contrast, as will be discussed further in Subsection 6.3.3.

Table 6.1: Gradient and standard deviation for different image blocks in Figure 6.2.

type	(b1)	(b2)
g_x	32	14.7
σ_x	15.42	12.08

Figure 6.3: The predicted value from schemes under consideration (X-axis) and the subjective DMOS (Y-axis; with $DMOS > 50$) for the LIVE database.

6.3.2 Further Analysis for Proposed Scheme and SSIM

Consider Figure 6.3 in which the X-axis represents the predicted value from the scheme under consideration and the Y-axis represents the subjective DMOS. We show the 5-parameter logistic mapping curve between the objective outputs and the subjective DMOS (the logistic mapping is to be discussed in Subsection 6.5.1 as (6.12)) as the middle curve. The upper curve denotes the $T_{DMOS}\%$ (here, $T_{DMOS}=20$ as an example in the figure) upper bound (i.e., with DMOS value being $T_{DMOS}\%$ higher than the middle curve); the lower curve denotes the lower bound (i.e., with DMOS value being $T_{DMOS}\%$ lower than the middle curve). In other words, the DMOS value of the images corresponding to the data points lying between the upper and lower curves can be “correctly” predicted by the IQA scheme if $T_{DMOS}\%$ measurement error is allowed; the DMOS value of the remaining images cannot be “correctly” predicted from the IQA scheme. The more the data points between the bounding curves are, the better the scheme is. If a data point lies below the lower bound, it means that the visual distortion is overestimated (like Figure 6.2 (c2)) since the scheme predicts a higher score (i.e., lower image quality) than the actual DMOS value; the distortion is underestimated for the data points lying above the upper bound curve.

Figure 6.3 (a) shows the plot of DMOS vs. SSIM for high distortion images (i.e. with DMOS values bigger than 50) from the LIVE database, where J2K, JPG, WN, GB and FF represent the distortion types of JPEG 2000 compression, JPEG compression, white Gaussian noise, Gaussian blur, and fast fading, respectively. In Figure 6.3 (a), some (more than 60 in number) data points lie outside the bounding curves; note that these data points mainly correspond to the images with WN distortion and therefore, we have demonstrated statistically that the SSIM shows higher sensitivity to white noise, i.e., it

overestimates the visual distortion with WN. It can also be observed from Figure 6.3 (a) that the SSIM tends to underestimate the visual distortion caused by JPEG compression.

The similar plot with the data given by (6.5a) is shown in Figure 6.3 (b), and only 29 data points lie outside the bounding curves. The effect of WN is also overestimated but to lesser extent; in addition, there is no underestimation for JPG error due to the proper emphasis on the edge information.

6.3.3 Modified Gradient Similarity

In this subsection, we examine more closely on how well (6.6) and (6.7) work. With a test image block y and its reference block x , the gradient (or contrast) of y can be regarded as the gradient of x being combined with an error signal, $g_y - g_x$; alternatively, the gradient of x can be regarded as that of y being combined with an error signal $g_x - g_y$. When two visual signals combine, there is masking of one signal over the other, and a visibility threshold, can be determined for the masked signal against the masking signal, as introduced in Subsection 2.2.2 and further discussed in Chapter 4.

As can be seen from (2.4)~(2.6), the visibility threshold T_{JND} for the masked signal increases with the masking contrast (i.e., g_f in (2.7), which is selected as $\max(g_x, g_y)$ as aforementioned). The masked gradient change R defined in (6.7) is a largely reasonable choice since when the masking signal is higher (i.e., $\max(g_x, g_y)$ is higher), the perceptual effect of the masked signal is lower for the same amount of $|g_x - g_y|$ (resulting in a lower R) due to the higher T_{JND} . As discussed in Subsection 6.3.1, the lower value of R , the higher value for $g(x, y)$ (therefore the higher predicted quality of the test block y).

200	201	200	200	200
200	201	200	200	200
200	201	200	200	200
200	201	200	200	200
200	201	200	200	200

(a)

200	204	200	200	200
200	204	200	200	200
200	204	200	200	200
200	204	200	200	200
200	204	200	200	200

(b)

Figure 6.4: A simple example to demonstrate the benefit of the modification for K .

The only problem for the formulation with (6.6) and (6.7) is the case with both the following conditions being met: (i) $|g_x - g_y|$ causes distortion that is smaller than or close to T_{JND} ; (ii) $\max(g_x, g_y)$ is comparable to $|g_x - g_y|$ in value. With the case (i), R should be 0 and therefore $g(x, y)$ should be 1, because the difference between x and y is not visible; however, if this situation coincides with the case (ii), (6.7) can give a value to R that is significantly bigger than 0, and this overestimates the image distortion.

As a general scheme, we need to make the necessary provision for overcoming the aforementioned overestimation. K in (6.6) is modified to satisfy the conditions: (A) K does not cause too much change to $g(x, y)$ when the cases (i) and (ii) do not happen simultaneously; (B) $g(x, y)$ is close to 1 for the cases with both (i) and (ii). For (A), K should be comparable with (several times of) $2(1-R)$ and $1+(1-R)^2$ in (6.6); for (B) K should be much larger than (e.g., more than 20 times of) $2(1-R)$ and $1+(1-R)^2$. Note that both $2(1-R)$ and $1+(1-R)^2$ lie in the range of $[0, 2]$, and therefore a reasonable K value is: $K < 20$ for (A) and $K > 40$ for (B). Obviously, K must be adaptive to $\max(g_x, g_y)$.

In this work, we have chosen K as $K = K' / \max(g_x, g_y)$, where K' is a positive constant and is called masking parameter. Typically, $\max(g_x, g_y)$ is around 50 and 5, for (A) and (B), respectively; so the choice of K' is between 200 and 1000 (we choose $K' = 200$ in the current stage of the work). Substituting the fine-tuned K into (6.6), we get

$$g(x, y) = \frac{2(1-R) + K' / \max(g_x, g_y) + 10^{-5}}{1 + (1-R)^2 + K' / \max(g_x, g_y) + 10^{-5}} \quad (6.8)$$

where $g(x, y)$ in (6.8) is still in the range of $[0 \ 1]$.

To illustrate the effect of the fine-tuning for K , Figure 6.4 (a) and (b) show two simple image blocks that have gradient values of 1 and 4, respectively; the visibility threshold T_{JND} for (a) is 4.7 when calculated using the method in Chapter 4. From the figure, we can see that (a) and (b) are very similar and this represents a case for the condition (B). However, the quality score given by (6.5a) is 0.471 (which means that (a) and (b) are not similar). On contrary, the score given by (6.8) is 0.989 which is more reasonable.

6.4 Integration for Overall Image Quality

In the previous section, we have proposed the gradient similarity based scheme considering better (in comparison with the SSIM) the masking effect and visibility threshold. For the parts within the dash box in the block diagram in Figure 6.1, the Masking Processing module is used to determine the value of K and the Contrast&Structure Comparison module is used to get the value of $g(x, y)$; (6.8) is the combination of these two modules. In (6.8), the gradient information of the reference and distorted images is used for evaluating the contrast and structural similarity. We also need to account for luminance similarity. In the rest of this section we discuss how to tackle the luminance similarity and the integration of such feature with the gradient similarity measurement.

6.4.1 Measurement for Luminance Distortion

Besides contrast/structure changes, luminance changes can also cause the visible distortion although they are not as annoying as contrast/structure changes. However, (6.8)

cannot account for the luminance change/distortion since it uses only the gradient information as the input and the gradient information is not affected by the non-contrast/structure changes. For example, the gradient information is the same if a constant value is added or subtracted from the image. Therefore, we have used the squared pixel error to measure the luminance changes in this work, after normalization by the dynamic range of the pixel values; the luminance similarity can be defined as

$$e(x_k, y_k) = 1 - \left(\frac{x_k - y_k}{L_e} \right)^2 \quad (6.9)$$

where x_k and y_k are the pixels at position k ($k \in \{(i, j)\}$) in image blocks x and y , respectively; and L_e is the dynamic range of the pixel values (255 for 8-bit gray scale images); $e(x_k, y_k)$ is the luminance similarity between image pixels x_k and y_k , and also with the range of $[0 \ 1]$.

6.4.2 Adaptive Distortion Integration

The existing gradient based schemes either fail to account for luminance distortion [123], [131] or use simple multiplication for the integration of the contrast-structure and luminance distortions [121], [122]. However, such integration method ignores the fact that contrast-structure distortion has more impact than luminance distortion toward visual quality estimation, and multiplication luminance distortion with contrast-structure distortion would degrade the overall performance of the scheme as demonstrated in [133].

A general form of integration for gradient similarity and luminance similarity to derive the overall quality indicator $q(x_k, y_k)$ for the image pixel pair x_k , and y_k can be given as

$$q = (1 - W_q(g, e)) \cdot g + W_q(g, e) \cdot e \quad (6.10)$$

where q , g , and e is the abbreviated form of $q(x_k, y_k)$, $g(x_k, y_k)$, and $e(x_k, y_k)$, respectively; $W_q(g, e)$ is the weighting function used to adjust the relative importance of the two

components. $w_q(g, e)$ should be significantly smaller than 0.5 since contrast-structural distortion (i.e., g) plays a much more significant role toward final quality evaluation than the luminance distortion (i.e., e).

In addition, the simultaneous existence of multiple distortion components in a neighborhood will mask the perception of each other. Since g is more important as aforementioned and e is more likely to be masked by the contrast-structural distortion if the contrast-structural distortion is high. Therefore, the weight $w_q(g, e)$ should be adaptive with respect to g (i.e. $w_q(g, e)$ should be small when g is small), i.e., we have

$$W_q(g, e) = p \cdot g \quad (6.11)$$

where p is a positive weighting parameter. Since g is in the range of $[0, 1]$, p has to also be significantly smaller than 0.5. We chose p to be 0.1 in this study, in order to balance: 1) the need of accounting for luminance change for a general solution; 2) the fact that luminance change contributes less than structural one. We will analyze the impact of this parameter choice in Subsection 6.5.5, with data from different images and distortion types, in various databases publicly available.

6.5 Experimental Results and Discussions

In this section, we provide extensive experimental results to evaluate the overall accuracy, robustness and efficiency with six publicly-available subject-rated benchmark IQA databases. We also compare the performance of the proposed scheme (q as defined in (6.10)) with the state-of-the-art IQA schemes including the SSIM [63], multi-scale SSIM [127], VIF [65] VSNR [66], MAD [67], and IW-SSIM [130]; and the results for PSNR and PSNR-HVS-M [64] are also included for comparison.

We first introduce the databases and the evaluation criteria adopted and then present

the experimental results. We will also discuss the computational complexity of the proposed scheme.

6.5.1 Databases and Evaluation Criteria

Six publicly-available and subject-rated image databases in the IQA community are used, namely, LIVE [132], TID [134], [135], Toyama [136], A57 [66], IVC [137], [138] and CSIQ [67], [139]. The number of distorted images is 779 for LIVE, 1700 for TID2008, 168 for Toyama, 54 for A57, 185 for IVC, and 866 for CSIQ, and these images and their corresponding subjective ratings are used as the ground truths to be compared against the IQA scheme outputs.

A 5-parameter logistic mapping between the objective outputs (s_o) and the subjective scores is employed, following the Video Quality Experts Group (VQEG) Phase-I/II tests and validation method [140], to remove any nonlinearity due to the subjective rating process and to facilitate the comparison of the schemes in a common analysis space. The logistic function used has the following form:

$$s_m = f(s_o) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2(s_o - \beta_3))} \right) + \beta_4 s_o + \beta_5 \quad (6.12)$$

where $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ are the parameters to be fitted by minimizing the sum of squared differences between the mapped values s_m and subjective scores. The subjective scores are the MOS or DMOS between reference and distorted images which have been provided by the IQA databases.

We used four evaluation criteria to compare the performance of IQA schemes, i.e., 1) Spearman rank-order correlation coefficient (SROCC), 2) Kendall rank-order correlation coefficient (KROCC), 3) Pearson linear correlation coefficient (CC), and 4) root mean squared error (RMSE), between the objective scores after nonlinear regression (i.e., S_m)

and the subject scores. 1)~3) can be calculated with the Matlab function [corr] with parameter “Pearson”, “Spearman”, “Kendall”, respectively. Among these four criteria, SROCC and KROCC are employed to assess prediction monotonicity [140], and these two criteria operate only on the rank of the data points (i.e. they are independent of any monotonic nonlinear mapping) and ignore the relative distance between data points; while CC and RMSE are adopted to evaluate prediction accuracy [140]. A better objective IQA measure has a higher CC, SROCC and KROCC, but lower RMSE values; as we have stated in Subsection 4.3.2, for a perfect match between the mapped objective scores (i.e., S_m) and the subjective scores, $CC=SROCC=KROCC=1$ and $RMSE=0$. More details, comparison and discussion of these performance measures can be found in [141], [142].

6.5.2 Accuracy and Monotonicity Evaluation

Figure 6.5 shows the scatter plots of the proposed IQA scheme on the six databases. The values of SROCC, KROCC, CC, and RMSE are listed in Table 6.2. We can see that the proposed scheme performs consistently well across all the databases. Even though the performance of the proposed scheme and MAD is similar across databases, the proposed scheme has a clear advantage over MAD in terms of computational effort and time required (execution time will be compared in Subsection 6.5.4). With regards to IW-SSIM, it is a good IQA scheme because it basically weighs SSIM (which itself is reasonably effective quality measure) values using weights derived via image information content [130]. Similar weighting can be employed for the proposed scheme to further improve its performance, as the next step of the work since the weighting strategy is general to different schemes.

To provide an overall indication of the comparative performance of the different

schemes, Table 6.3 gives the average SROCC, KROCC and CC results over six databases, where the average values are computed in two cases, in the same way as done in [130]. In the first case, the correlation scores are directly averaged, while in the second case, different weights are assigned to the databases depending on the number of distorted images in each database (refer to Subsection 6.5.1 for such numbers). From Table 6.3, we can see that the proposed scheme performs the best or close to the best on average no matter what kind of averaging is used and what the evaluation criterion is.

6.5.3 Robustness Evaluation

The robustness of the proposed scheme on different databases is also demonstrated in Table 6.2. Note that, the SSIM performs well on IVC database but not on A57 database; VIF is not good on A57 although it is the best for LIVE; on the other hand, VSNR is the best for A57 but is relatively poor on TID database. Overall, MAD, IW-SSIM, and the proposed scheme give more consistent and stable performance across all the six databases in comparison to the other schemes. Among the three schemes, MAD performs slightly less well on TID; IW-SSIM performs slightly less well on A57; while the proposed one performs slightly less well on CSIQ.

To further examine the robustness of the IQA schemes, the performance on each distortion type in TID database is shown in Table 6.4, where the two best IQA schemes have been highlighted in boldface for each distortion type. The TID database is used since it is the largest database and contains the most distortion types, which are listed in the first column of Table 6.4. We include only the SROCC values since other performance criteria lead to similar conclusions. From the table, we can see that for noise contaminated images, traditional PSNR and its variation (i.e., PSNR-HVS-M) are still the best IQA schemes; the proposed scheme, the SSIM, multi-scale SSIM, and IW-SSIM

have similar performance. For other distortion types, the proposed scheme performs quite well (i.e., the best or the second best). Such results are expected since for the cases where noise would not severely damage the edge or create false edge, the proposed scheme emphasizing edge information maintains the performance; when the distortion causes error both around edges and non-edges (e.g., block-wise distortion), performance of the proposed scheme is good. It is also noted that the MAD has some difficulty to predict the quality of the images with distortions caused by mean value shift or contrast change.

Figure 6.5: Scatter plots of subjective scores vs. scores from the proposed scheme q on IQA databases.

Table 6.2: Performance comparison for IQA schemes on six databases.

database	criteria	PSNR	SSIM	MS-SSIM	VSNR	VIF	PSNR-HVS-M	Proposed	MAD	IW-SSIM
LIVE	SROCC	0.8756	0.9479	0.9513	0.9280	0.9632	0.9295	0.9554	0.9669	0.9567
	KROCC	0.6865	0.7963	0.8044	0.7625	0.8270	0.7659	0.8131	0.8421	0.8175
	CC	0.8723	0.9449	0.9409	0.9237	0.9598	0.9251	0.9437	0.9674	0.9522
	RMSE	13.3597	8.9454	9.2593	10.4694	7.6670	10.3722	9.0376	6.9235	8.347
TID	SROCC	0.5794	0.7749	0.8542	0.7049	0.7496	0.5951	0.8554	0.8340	0.8559
	KROCC	0.4210	0.5768	0.6568	0.5345	0.5863	0.4692	0.6651	0.6445	0.6636
	CC	0.5726	0.7710	0.8451	0.6823	0.8090	0.5784	0.8462	0.8306	0.8579
	RMSE	1.1003	0.8546	0.7173	0.9810	0.7888	1.0947	0.7151	0.7474	0.6895
Toyama	SROCC	0.6132	0.8794	0.8874	0.8608	0.9077	0.8480	0.9233	0.9362	0.9202
	KROCC	0.4443	0.6939	0.7029	0.6745	0.7315	0.6568	0.7541	0.7823	0.7537
	CC	0.6353	0.8876	0.8922	0.8704	0.9138	0.8580	0.9287	0.9405	0.9248
	RMSE	0.9664	0.5764	0.5652	0.6160	0.5084	0.6428	0.4640	0.4251	0.4761
A57	SROCC	0.6189	0.8066	0.8414	0.9355	0.6223	0.8962	0.9002	0.9023	0.8709
	KROCC	0.4309	0.6058	0.6478	0.8031	0.4589	0.7261	0.7205	0.7233	0.6842
	CC	0.6347	0.8014	0.8575	0.9497	0.6157	0.8749	0.8976	0.9059	0.9034
	RMSE	0.1899	0.1470	0.1264	0.0769	0.1937	0.1190	0.1084	0.1041	0.105
IVC	SROCC	0.6884	0.9018	0.8980	0.7993	0.8964	0.8832	0.9294	0.9146	0.9125
	KROCC	0.5218	0.7223	0.7203	0.6053	0.7158	0.6935	0.7626	0.7406	0.7339
	CC	0.7196	0.9119	0.9108	0.8034	0.9028	0.8905	0.9399	0.9210	0.9231
	RMSE	0.8460	0.4999	0.5029	0.7255	0.5239	0.5544	0.4160	0.4747	0.4686
CSIQ	SROCC	0.8005	0.8756	0.9133	0.8104	0.9195	0.8179	0.9126	0.9467	0.9213
	KROCC	0.5984	0.6907	0.7393	0.6237	0.7537	0.6430	0.7403	0.7970	0.7529
	CC	0.7998	0.8612	0.8990	0.7993	0.9277	0.8137	0.8979	0.9502	0.9144
	RMSE	0.1576	0.1334	0.1150	0.1578	0.0980	0.1526	0.1156	0.0818	0.1063

Table 6.3: Average performance over six databases.

database	criteria	PSNR	SSIM	MS-SSIM	VSNR	VIF	PSNR-HVS-M	Proposed	MAD	IW-SSIM
Direct Average	SROCC	0.6960	0.8644	0.8909	0.8398	0.8431	0.8283	0.9127	0.9168	0.9063
	KROCC	0.5171	0.6810	0.7119	0.6673	0.6789	0.6591	0.7426	0.7550	0.7343
	CC	0.7057	0.8630	0.8909	0.8381	0.8548	0.8234	0.9090	0.9193	0.9126
Weighted Average	SROCC	0.6994	0.8455	0.8914	0.7905	0.8456	0.7458	0.8967	0.8971	0.8978
	KROCC	0.5232	0.6615	0.7116	0.6160	0.6860	0.5941	0.7228	0.7328	0.7240
	CC	0.6982	0.8405	0.8830	0.7777	0.8742	0.7369	0.8874	0.8970	0.8974

Table 6.4: SROCC comparisons for individual distortion types.

Distortion type	PSNR	SSIM	MS-SSIM	VSNR	VIF	PSNR-HVS-M	Proposed	MAD	IW-SSIM
WN	0.9070	0.8107	0.8086	0.7734	0.8799	0.9110	0.8577	0.8388	0.8028
WN for color	0.8995	0.8029	0.8054	0.7798	0.8785	0.8479	0.8091	0.8258	0.8015
Spatially Correlated noise	0.9170	0.8145	0.8209	0.7673	0.8703	0.9262	0.8907	0.8678	0.7909
Masked noise	0.8515	0.7795	0.8107	0.7348	0.8698	0.7555	0.7409	0.7336	0.8068
High frequency noise	0.9270	0.8729	0.8694	0.8822	0.9075	0.9136	0.8936	0.8864	0.8732
Impulse noise	0.8724	0.6732	0.6907	0.6438	0.8331	0.8339	0.7229	0.6499	0.6579
Quantization noise	0.8696	0.8531	0.8589	0.8255	0.7956	0.9007	0.8752	0.8160	0.8182
GB	0.8697	0.9544	0.9563	0.9323	0.9546	0.9219	0.9589	0.9197	0.9580
Image denoising	0.9416	0.9530	0.9582	0.9290	0.9189	0.9572	0.9724	0.9434	0.9463
JPG	0.8717	0.9252	0.9322	0.9181	0.9170	0.9338	0.9392	0.9275	0.9181
J2K	0.8132	0.9625	0.9700	0.9517	0.9713	0.9762	0.9759	0.9707	0.9749
JPG transmission errors	0.7516	0.8678	0.8681	0.8052	0.8582	0.8105	0.8835	0.8661	0.8560
J2K transmission errors	0.8309	0.8577	0.8606	0.7914	0.8510	0.9321	0.8925	0.8394	0.8313
Non eccentricity noise	0.5815	0.7107	0.7377	0.5723	0.7608	0.6812	0.7372	0.8287	0.7719
Block-wise distortion	0.6193	0.8462	0.7546	0.1952	0.8320	0.6238	0.8862	0.7970	0.7889
Mean shift	0.6957	0.7231	0.7338	0.3885	0.5132	0.6932	0.7170	0.5161	0.6757
Contrast change	0.5859	0.5246	0.6381	0.4250	0.8190	0.5815	0.6737	0.2723	0.6273

Table 6.5: Execution time (in second/image) for different schemes.

Schemes	PSNR	SSIM	MS-SSIM	VSNR	VIF	PSNR-HVS-M	Proposed	MAD	IW-SSIM
Time	0.0037	0.0454	0.1673	0.4452	3.4829	2.5586	0.0873	34.3821	0.5762

6.5.4 Efficiency Evaluation

To compare the efficiency (i.e., computational complexity) of different schemes, we measured the average execution time required for per image in A57 database on a PC with 2.40 GHz Intel Core2 CPU and 2 GB of RAM. Table 6.5 shows the required time in second per image, with all the codes being implemented with Matlab. The codes for the SSIM, multi-scale SSIM, VSNR, VIF, PSNR-HVS-M, MAD, IW-SSIM and are from [143], [143], [144], [132], [145], [146], [147], respectively. It can be seen from Table 6.5 that the proposed scheme takes more time than the PSNR and the SSIM only, and takes less time than the multi-scale SSIM since no multi-scale image decomposition is involved in the proposed one. MAD and IW-SSIM take much longer processing time than the proposed scheme. To be more precise, the proposed method takes only about 0.25% and 15.2% of the time taken by MAD and IW-SSIM respectively.

6.5.5 Impact of the Parameter Values

The impact of the parameter values is shown with the plots of SROCC as a function of the parameter values. In the proposed scheme, there are two parameters: κ' and p , where κ' is for the masking effect (its valuation has been determined in Subsection 6.3.3) and p is used to integrate contrast-structure distortion and luminance distortion (as discussed in Subsection 6.4).

To confirm that the chosen κ' is a reasonable choice, Figure 6.6 plots the SROCC as a function of κ' for all the databases. As can be observed: when the value of κ' is large (e.g., $\kappa' \geq 200$), the value for the SROCC is nearly constant no matter what the value of κ' is used, the difference of SROCC values for different values of κ' is smaller than 0.01, and this is valid for every database. Although the optimal value of κ' to achieve the

highest SROCC value is different for different databases (i.e., 15 for Toyama; 40 for LIVE, IVC, and CSIQ; 1000 for TID and A57), there are two common trends for different databases as shown in these plots: (a) SROCC drops fast when the value of κ' close to 0, and this is expected since for a small κ' , the masking effect for the case the masking signal is small is not properly reflected; (b) when $\kappa' > 200$, the SROCC is very close to (i.e., no smaller than 97.7% of) the highest possible value when the optimal value for κ' is used. Therefore, our choice of $\kappa' = 200$ is good for the proposed scheme.

To show the impact of p for (6.11), we simulated various values of p on the TID database [135] first since it includes luminance distortion (mean shift distortion, with 25 reference images and four distortion levels for each reference image). In Figure 6.7 we show the absolute values of the SROCC, CC, and RMSE plotted as a function of p for the overall performance (i.e., the images with all the 17 distortion types) and the performance on luminance distortion (i.e., the images with mean shift distortion). We found that $p = 0.1$ to yield good tradeoff between performance on luminance distortion and for the overall cases (i.e., the SROCC value for $p = 0.1$ is over 99.0% of the highest SROCC value, among all the possible SROCC values when different values of p are used). From the figure, we can see that the overall performance decreases with the increase of p (i.e., increase of the contribution of e in q) since e is less capable of measuring the contrast and structural distortions compared to g . On the other hand, p cannot be zero since that will mean that the luminance distortion is completely ignored. Therefore, we use $p = 0.1$ since it gives good performance both for luminance distortion as well as for the overall case.

 Figure 6.6: Plot of $|SROCC|$ as a function of κ' for IQA databases.

In addition, Figure 6.8 shows the SROCC values as a function of p for the other databases. From the figure, we observe the following: (a) for four databases, SROCC drops with the increase of p . This is expected since contrast-structure distortion is the dominant factor in IQA, and according to (6.10) and (6.11), the larger the value of p is,

the larger the impact of the luminance distortion measurement becomes. (b) when the value of p is in the range of $[0, 0.2]$, the values of the SROCC is nearly constant no matter what the value of p is used (the difference of SROCC values for difference values of p is smaller than 0.01), and the values for SROCC is very close to (i.e., no smaller than 99.0% of) the highest possible value. Therefore, a value of p in the range of $[0, 0.2]$ is a good database-independent choice for the overall performance. However, if the value of p is too small (i.e., $p < 0.1$), the performance on luminance distortion would degrade too much as shown and discussed with Figure 6.7. Therefore, the current choice of $p=0.1$ is again a reasonable choice.

Figure 6.7: Plot of (a) SROCC, (b) CC, and (c) RMSE, as a function of p for the proposed integration approach and for the TID dataset [135].

 Figure 6.8: Plot of $|SROCC|$ as a function of p for IQA databases.

Table 6.6: SROCC comparisons for each component of the proposed scheme.

Database	q	g	g_m	g_g	$\Delta(g_m)$	$\Delta(g_g)$
LIVE	0.9554	0.9555	0.8689	0.9057	0.0866	0.0498
TID	0.8554	0.8532	0.6592	0.6231	0.1940	0.2301
Toyama	0.9233	0.9241	0.6799	0.7135	0.2442	0.2106
A57	0.9002	0.9004	0.5624	0.7742	0.3380	0.1262
IVC	0.9294	0.9297	0.8002	0.7725	0.1296	0.1572
CSIQ	0.9126	0.9126	0.8153	0.7616	0.0973	0.1510

6.5.6 Impact of Each Component of the Scheme

To discuss the impact of each component in the proposed scheme towards the performance, the SROCC values for different components are given in Table 6.6: q is the proposed overall scheme as in (6.10); g is the proposed gradient similarity scheme as in (6.8); g_m is the initial form of the proposed gradient similarity scheme defined in (6.5a), without the modification in masking as in Subsection 6.3.3; g_g is the proposed scheme without gradient component (i.e., the gradient information g_x and g_y are substituted with pixel gray level information x and y as the input in (6.8)); $\Delta(g_m)$ and $\Delta(g_g)$ are the differences of g_m , g_g with g , respectively.

From the table, we can see that q and g have similar performance on each database, and this is expected since g accounts for the major effect in the evaluation. The results of $\Delta(g_g)$ have shown the impact of the gradient component, and from the table we can see that the incorporation of gradient component has improved about 0.05 in the SROCC value for LIVE database and 0.1~0.2 for the other databases; the possible reason for the smaller improvement on LIVE database is that: i) this database contains a wider range of distortion strength; and ii) many IQA schemes can achieve good performance on it. It is also shown in $\Delta(g_m)$ that the modification for masking can achieve some performance improvement on every database; the performance improvement on LIVE and CSIQ databases are the smallest (around 0.09) since the considered masking effect has more impact on near-threshold distortion compared with the impact on supra-threshold distortion, and the two databases have fewer images (in percentage) with near-threshold distortion; for A57 database, the performance improvement is the largest (larger than 0.3) since almost all the images in the database are with near-threshold distortion.

6.6 Concluding Remarks

Visual quality evaluation is important by itself and with applications in many (if not all) image manipulations. Since the HVS is more sensitive to edge regions than non-edge ones, edge information should be carefully and explicitly incorporated in designing of an IQA scheme. We have proposed a new IQA scheme based on the concept of gradient similarity to alleviate the shortcoming of the existing relevant schemes in this regard. We have demonstrated that the proposed gradient similarity measure can be used to gauge contrast and structural changes. In addition, we have made our scheme matches better with masking effect and visibility threshold. Finally, luminance similarity is devised and incorporated to form a complete quality scheme.

The effectiveness of the proposed IQA scheme has been demonstrated with six public benchmark IQA databases. Compared with eight other representative and prominent IQA schemes, similar or better performance (i.e., more consistent with the human judgment) is achieved. Robustness across different databases with 3752 images and more than 20 distortion types is also confirmed. Besides its robustness and high accuracy, the proposed scheme also provides the pixel error map and is relatively simple with its mathematical formulation and computational complexity, making it practical to be used or embedded in various optimization processes in image processing tasks, as well as on-line and in-service applications.

The proposed metric performs well on images with various distortion levels since it has considered both the high level property of the HVS (i.e., highly adapted to the structure information and therefore accounting for supra-threshold distortion) and the low level masking property of the HVS (mainly accounting for near threshold distortion).

Chapter 7

Summary and Future Work

7.1 Summary

To improve the quality of a coded visual signal with a given bit rate, this thesis has presented a number of approaches which can better explore the redundancy in the following three aspects: further statistical redundancy reduction for coding on top of the state-of-the-art technology, perceptual visual modeling and redundancy removal toward additional improvement of performance, and user-oriented quality evaluation of visual signal for possible use in coding.

The major technical contributions of this project have been highlighted in Section 1.3, and in this section, we will give a summary of the research that has been performed in this project, with links back into the preceding chapters.

7.1.1 Statistical Redundancy Reduction

Unlike the traditional coding schemes, physical meaning (i.e., spatial and temporal dimension) of a video has been ignored and a novel video pre-processing framework has been therefore proposed in Chapter 3. In our method, the video is still encode in a frame by frame manner; however, the frame is no longer fixed with the spatial plane and is

allowed to be formed in all the possible planes (there are three different planes using any two of the three axes, i.e., XY, TX, and TY if X, Y, and T are one spatial axis, the other spatial axis and the temporal axis, respectively). Therefore, the first step of the proposed method is to re-organize video data in an appropriate frame by frame manner adaptively, and then a standard coding method can be adopted, as illustrated in Subsection 3.3.2. Encoding a video with different frame formation would result in different RD performance and as we discussed in Subsection 3.3.1, RD performance is closely related with the average intra-frame redundancy if no inter-frame prediction is applied: the higher the intra-frame redundancy is, the better the RD performance becomes. Average inter-frame correlation coefficient (CC) along one of the three possible axes when the frames are formed in the remaining two axes is used as the measurement of redundancy along that axis, and hence, the intra-frame (formed in AB plane, where A, B = X, Y, T) redundancy can be indicated by the CCs for A and B axes.

We have formulated the manipulations for the optimal coding plane (OCP, i.e., the plane with the best RD performance) selection and the associated frame formation as a pre-processing step for any existing video coding schemes (e.g., Motion JPEG and H.264 standards, lossless and lossy coding schemes, methods with and without inter-frame prediction, and so on). Intuitively, the OCP needs to be found prior to actual coding. A brute-force method can find the best OCP but is computationally expensive. To determine such the OCP in practice, we presented more efficient selection criteria for coding without and with inter-frame prediction in Sections 3.4 and 3.5, respectively. These criteria are based on the comparison of the CCs along different axes and it has also been demonstrated that sparse frames are sufficient for CCs comparison and the required computational complexity overhead is small. Extensive experiments have been conducted to confirm the effectiveness and benefits of the proposed methodology in

improving the RD performance for video coding, with a large number of coding schemes (Motion JPEG-LS, Motion JPEG, Motion J2K, H.264 intra-only profile, and H.264-inter coding) and video sequences of diversifying nature. We have also provided careful analysis and useful insight information to justify the improvement of the proposed method.

7.1.2 Perceptual Modeling and Redundancy Removal

To further improve the coding performance, perceptual masking effects have been modeled and perceptual redundancy has been explored for visual signal coding, since the human visual system (HVS) is the ultimate perceiver for most coded signals.

Chapter 4 has aimed at better formulating the JND model. In pixel domain, the perceptual effect usually includes luminance masking and contrast masking (CM). We proposed a model to accurately evaluate the CM by better separation of edges from texture, because the CM for edge regions (edge masking) is smaller than that for texture regions (texture masking) even if the two regions have similar spatial contrast. To be more specific, an image is firstly decomposed into a cartoon component (i.e., piecewise smooth and with sharp edges) and a textural component, and then the edge masking and texture masking are evaluated on these two components, respectively. This has alleviated the difficulty in separating edge masking and texture masking in CM estimation.

Experiments with perceptually-driven noise shaping have been performed with the proposed models, proving the improvement over the existing relevant schemes. The noise shaping indicates the image regions allowing higher hidden noise without being noticed and therefore these regions can tolerate a higher quantization step (equivalent to fewer bits assigned) in image and video coding.

In Chapter 5, the benefit of adopting the perceptual/masking model to remove

perceptual redundancy has been demonstrated in the down-sampling based image coding framework (i.e., down-sampling prior to coding for low bit rate coding) and signal pre-processing for visually lossless image coding framework.

Unlike the previous work for down-sampling based image coding, we have adjusted the QP for the MB depending on not only the sampling ratio of the macro-block (MB) but also the MB's visual importance (which is estimated from a computational visual attention model) as described in Section 5.1.

In Section 5.2, the detail of the visually lossless image histogram adjustment method has been presented as a pre-processing step, and therefore, is independent of the codec used. In histogram adjustment, the JND model designed in Chapter 4 was used as the error visibility threshold, and the image zero order entropy is minimized under the constraint that limits the resultant error to be not larger than the JND threshold. With the help of such a method, the required bit rate for the image is reduced while keeping the same visual quality as the original image (i.e., being visually lossless).

7.1.3 Quality Evaluation for Visual Signal

Objective visual quality evaluation is not only essential as a benchmarking tool for comparing different algorithms/systems, as well as monitoring various process, but also central to shaping/ optimizing many visual signal processing (including video coding, of course).

Chapter 6 has further presented a simple but effective method that can deal with images with different visual content and distortion types. The aim of the work is to measure visual quality in a way that is more consistent with the human judgment. We demonstrated that the generated image gradient map contains both contrast information and structure information, and therefore the similarity between the two gradient maps

(i.e., one is from the reference image and the other is from the distorted image) can be used to measure the contrast and structure distortion. More importantly, better performance is achieved over its counterpart in the relevant existing method (i.e., contrast comparison term and structure comparison term in SSIM) since the significance of edge is emphasized, as discussed in Section 6.3.

Given a distorted image, not all the pixel/contrast/structure differences from its reference can be perceived as the distortion. The proposed approach is capable of working well on images with various distortion levels since it has considered both the high level property of the HVS (i.e., the HVS is highly adapted to the structure information and this mainly accounts for the supra-threshold distortion) and the low level masking property of the HVS (which mainly accounts for the near threshold distortion). In the proposed method, luminance distortion is measured as the squared difference and a new content based adaptive formula for integrating contrast-structure distortion and luminance distortion has also been developed as in Section 6.4, based on the well-known knowledge that the contrast-structure distortion has more impact than the luminance distortion toward the perceived visual quality.

Compared to the relevant previous work, apart from the improved accuracy, our method is appealing in practice also because it can provide not only the overall visual quality score but also the local quality map, as well as being computational efficient and robust to various distortion scenarios. Experimental results on six public image databases demonstrate that the proposed method outperforms the existing relevant methods for images with various contents, distortion types and distortion levels. The developed metric can be used as a standalone visual quality estimator or an embedded control module in video coding, watermarking, post-processing, and many other visual processing tasks, due to its operating efficiency.

7.2 Future work

From the current work, the experimental results and the tracking of the latest development of the relevant technology, there are several interesting avenues for further research as the possible extension of the project work performed so far.

The proposed pre-processing method for video coding in Chapter 3 attempts to explore the statistical redundancy in a better and more comprehensive way by allowing the frames to be formed in the all three possible planes (i.e., XY, TX and TY). From Figure 3.9, we can see that there are more MBs take intra-coding modes when encoding in a non-XY plane in comparison with encoding in the traditional XY plane. As we know, an intra-coded video usually is robust to transmission error than the inter coded one since there is less dependency in the intra encoded bit stream (i.e., to reconstruct one intra encoded frame, the information from the other frames are not needed; while the reconstruction of one inter encoded frame needs the information from the other frames). Therefore, it would be possible to extend the proposed scheme into error resilient video coding. Besides, it is worth to investigate into the extension of the proposed method for high resolution videos since the characteristics of high resolution videos need to be studied (in fact, H.264 itself and its coding modes need to be extended for the high resolution cases for better performance [151]-[153]).

The perceptual effects of the HVS have been better modeled in Chapter 4 for image pixel domain. When applied to image and video coding, the designed models can be used to adjust the quantization parameter (e.g., to increase the quantization step for pixels if they associate with strong masking effect and to decrease the quantization step for the pixels with weak masking effect). Overhead bits are usually required to inform the decoder how to change the quantization step since such information is derived from the

original image or video at the encoder. Hence, it would be beneficial to estimate the side information from the decoder side to avoid the excessive need of sending the overhead bits. The needed information can be obtained from the bit stream or estimated from the previous decoded information.

The benefit of perceptual redundancy reduction with this project has been mainly demonstrated in Chapter 5 with two application scenarios: down-sampling based low bit rate image coding and visually lossless coding. As stated in Section 5.2, the perceptual redundancy is explored by histogram adjustment to minimize the zero-order image entropy since the entropy of the visual signal is the lower bit rate bound for the signal to be lossless coded. But the zero-order entropy is an overestimation for the actual entropy of an image since image pixel values are spatially correlated with each other rather than being statistically independent. Therefore, it is desirable to investigate into proper formulation for image entropy and substitute for the zero order entropy as the objective function to be minimized in the process. Then, the concept and calculation of perceptual entropy can be further improved.

As mentioned in Chapter 6, overall image quality score is from the scores of local image quality, and only the simple average (i.e., uniform weighting) is used to obtain the overall quality in our current study. It is demonstrated that image information content based weighting can improve the performance of quality evaluation [130]. Therefore, it would be possible to incorporate the advancement in quality weighting to further increase the accuracy the proposed quality evaluation scheme. Besides, the proposed scheme may not work well in the case where the image is with color distortion that does not follow the distribution of the luminance one, since only the luminance information was utilized in the current modeling. Therefore, we hope to explore how color distortion is perceived and interpreted by the HVS, and extend the proposed method to color images. For

example, we can derive the quality index for each color component (e.g., in RGB or HSV space) of the image, and then obtain the final quality score through appropriate integration. Although the current formulation can be applied to color space in a straightforward manner, the color aspects deserve more careful, dedicated investigation since color perception is important. We would also like to account for more relevant features and extend the ideas in this work to tackle various visual signal processing processes (like video coding, watermarking, graphics, error resilience, post-processing, etc.). The most straightforward way of course is to use the average of frame level quality (quality of each frame computed using the proposed scheme) as the overall video quality measure. Other possibility is to use the proposed scheme to measure the spatial quality and incorporate a temporal factor (e.g. by using similarity between the motion vectors [171] or the variation of quality along the time axis [172]). The spatial and the temporal factors may also be combined via machine learning as in [172]. Another feasible approach is to use the proposed method to measure the quality (or similarity) of motion-compensated blocks (in reference and distorted video sequences), which can serve as the temporal factor (similar approach reported in [173] uses SSIM).

References

- [1] Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios. ITU-R Recommendation BT.601-7, 2011.
- [2] A. Bovik, “Handbook of Image and Video Processing,” *Academic Press*, pp. 67, New York, 2000.
- [3] K. Ng, L. Po, K. Wong, C. Ting, K. Cheung, “A Search Patterns Switching Algorithm for Block Motion Estimation,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19(5), pp. 753–759, May 2009.
- [4] M. Sarwer, Q. Wu, “Adaptive Variable Block-Size Early Motion Estimation Termination Algorithm for H.264/AVC Video Coding Standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19(8), pp. 1196–1201, May 2009.
- [5] L. Po, C. Ting, K. Wong, K. Ng, “Novel Point-Oriented Inner Searches for Fast Block Motion Estimation,” *IEEE Trans. Multimedia*, vol. 9(1), pp. 9–15, Jan. 2007.
- [6] V. Nguyen, Y. Tan, “Efficient block-matching motion estimation based on Integral frame attributes,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16(3), pp. 375–385, Mar. 2006.
- [7] A. Bahari, T. Arslan, A. Erdogan, “Low-Power H.264 Video Compression Architectures for Mobile Communication,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19(9), pp. 1251–1261, Sept. 2009.
- [8] C. Chen, S. Chien, Y. Huang, T. Chen, T. Wang, L. Chen, “Analysis and Architecture Design of Variable Block-Size Motion Estimation for H.264/AVC,” *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 53(3), pp. 578–593, Mar. 2006.
- [9] D. Vo, T. Nguyen, “Quality Enhancement for Motion JPEG Using Temporal Redundancies,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18(5), pp. 609–619, May 2008.

- [10] M. Fujiyoshi, K. Kuroiwa, H. Kiya, "A scrambling method for Motion JPEG videos enabling moving objects detection from scrambled videos," in *Proc. Int. Conf. Image Processing*, pp. 773–776, 2008.
- [11] J. Meessen, C. Parisot, X. Desurmont, J. Delaigle, "Scene analysis for reducing motion JPEG 2000 video surveillance delivery bandwidth and complexity," in *Proc. Int. Conf. Image Processing*, pp. 577–580, 2005.
- [12] T. Tillo, M. Grangetto, G. Olmo, "Multiple description coding with error correction capabilities: an application to motion JPEG 2000," in *Proc. Int. Conf. Image Processing*, pp. 3129–3132, 2004.
- [13] G. Sullivan, T. Wiegand, H. Yu, "New Profiles for Professional Applications," *JVT document JVT-V204*, Jan. 2007.
- [14] H. Yu, "New MPEG-4-AVC/H.264 Profiles for Professional and High-Quality Video Applications," *SMPTE Motion Imaging Journal*, pp. 45–49, Jan. 2009.
- [15] H. Schwarz, D. Marpe, T. Wiegand, "Analysis of Hierarchical B Pictures and MCTF," in *Proc. Int. Conf. Multimedia and Expo*, pp. 1929–1932, 2006.
- [16] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13(7), pp. 688–703, Jul. 2003.
- [17] E. Yang, X. Yu, "Rate Distortion Optimization for H.264 Interframe Coding: A General Framework and Algorithms," *IEEE Trans. Image Process.*, vol. 16(7), pp. 1774–1784, Jul. 2007.
- [18] N. Memon and K. Sayood, "Lossless compression of video sequences," *IEEE Trans. Commun.*, vol. 44(10), pp. 1340–1345, Oct. 1996.
- [19] K. Yang and A. Faryar, "A context-based predictive coder for lossless and near-lossless compression of video," in *Proc. Int. Conf. Image Processing*, pp. 144–147, 2000.
- [20] E. Carotti, J. Martin, "Motion-Compensated Lossless Video Coding in the CALIC Framework," *Proc. Int'l Symp. on Signal Process. and Information Technology*, pp. 600–605, 2005.

References

- [21] D. Brunello, G. Calvagno, G. Mian, R. Rinaldo, "Lossless compression of video using temporal information," *IEEE Trans. Image Process.*, vol. 12(2), pp. 132–139, 2003.
- [22] M. Zhang, j. Hu, and L. Zhang, "Lossless video compression using combination of temporal and spatial prediction," in *Proc. IEEE. Int. Conf. Neural Networks Signal Processing*, pp. 1193–1196, Dec. 2003.
- [23] S. Park, E. Delp, and H. Yu, "Adaptive lossless video compression using an integer wavelet transform," in *Proc. Int. Conf. Signal Processing*, pp. 1123–1126, 2004.
- [24] Y. Li and K. Sayood, "Lossless video sequence compression using adaptive prediction," *IEEE Trans. Image Process.*, vol. 16(4), pp. 997–1007, 2007.
- [25] C. Chou and Y. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5(6), pp. 467–476, 1995.
- [26] P. Longere, X. Zhang, P. Delahunt, and D. Brainaro, "Perceptual assessment of demosaicing algorithm performance," *Proc. IEEE*, vol. 90(1), pp. 123–132, 2002.
- [27] A. Bruckstein, M. Elad, and R. Kimmel, "Down-scaling for better transform compression," *IEEE Trans. Image Process.*, vol. 12(9), pp. 1132–1144, Sep. 2003.
- [28] B. Zeng and A. Venetsanopoulos, "A JPEG-based interpolative image coding scheme," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, pp. 393–396, 1993.
- [29] Y. Tsaig, M. Elad, G. Golub, and P. Milanfar, "Optimal framework for low bit-rate block coders," in *Proc. Int. Conf. Image Processing*, pp. 219–222, 2003.
- [30] W. Lin and L. Dong, "Adaptive Downsampling to Improve Image Compression at Low Bit Rates," *IEEE Trans. Image Process.*, vol. 15(9), pp. 2513–2521, 2006.
- [31] M. Chun and J. Wolfe, "Visual attention," in *Blackwell Handbook of Perception*, B. Goldstein, Ed. Oxford, UK: Blackwell, pp. 272–310, 2001.
- [32] L. Itti, "Models of bottom-up and top-down visual attention," *Ph.D. dissertation*, Dept. Computat. Neur. Syst., California Inst. Technol., Pasadena, Jan. 2000.
- [33] L. Itti, "Visual salience," *Scholarpedia*, 2(9):3327, 2007. [Online]. Available:

References

- http://www.scholarpedia.org/article/Visual_salience.
- [34] J. Hopfinger, M. Buonocore, G. Mangun, “The neural mechanisms of top-down attentional control,” *Nature Neuroscience*, vol. 3(3), pp.284–291, 2000.
 - [35] D. Lloyd, N. Merat, F. McGlone, and C. Spence, “Crossmodal links between audition and touch in covert endogenous spatial attention,” *Perception Psychophys.*, vol. 65(6), pp.901–924, 2003.
 - [36] D. Bolme, and Pyvision, “computer vision toolbox,” 2008. [Online]. Available: <http://pyvision.sourceforge.net>,
 - [37] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 511–518, 2001.
 - [38] Y. Freund and R. Schapire, “A decision-theoretic generalization of online learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55(1), pp. 119–139, 1997.
 - [39] C. Koch, S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” *Human Neurobiology* 4, pp.219–227, 1985.
 - [40] L. Itti, C. Koch, E. Niebur, “A model of saliency based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Analy. Mach. Intell.*, vol. 20(11), pp.1254–1259, 1998.
 - [41] C. Guo, Q. Ma, and L. Zhang, “Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform,” in *Proc. Int. Conf. Computer Vision*, pp. 1–8, 2008.
 - [42] N. Jayant, J. Johnston, R. Safranek, “Signal compression based on models of human perception,” *Proc. IEEE*, vol. 81, pp. 1385–1422, 1993.
 - [43] D. Kelly, “Motion and vision I: Stabilized images of stationary gratings,” *J. Opt. Soc. Am.*, vol. 69(9), pp. 1266–1274, 1979.
 - [44] G. Legge, J. Foley, “Contrast masking in human vision,” *Journal of the Optical Society of America*, vol. 70, pp. 1458–1471, 1980.
 - [45] M. Ramasubramanian, S. Pattanaik, D. Greenberg, “A perceptual based physical

- error metric for realistic image synthesis,” *Proceedings of Computer Graphics Conference*, vol. 4(33), pp. 73–82, 1999.
- [46] X. Zhang, W. Lin and P. Xue, “Just-Noticeable Difference Estimation With Pixels in Images,” *Journal of Visual Communication and Image Representation*, vol. 19(1), pp. 30–41, 2008.
- [47] C. Chou, C. Chen, “A perceptually optimized 3-D subband image codec for video communication over wireless channels,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6(2), pp. 143–156, 1996.
- [48] Y. Chiu, T. Berger, “A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancement,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9(3), pp. 438–450, 1999.
- [49] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, “Just noticeable distortion model and its applications in video coding,” *Signal Process.:Image Commun.*, vol. 20, pp. 662–680, 2005.
- [50] H. Nothdurft, “Saliency from feature contrast: additivity across dimensions,” *Vision Research*, vol. 40, pp. 1183–1201, 2000.
- [51] X. Yang, W. Lin, Z. Lu, E. Ong and S. Yao, “Motion-compensated Residue Pre-processing in Video Coding Based on Just-noticeable-distortion Profile,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15(6), pp.742–750, June, 2005.
- [52] M. Naccari, F. Pereira, “Decoder side just noticeable distortion model estimation for efficient H.264/AVC based perceptual video coding,” in *Proc. Int. Conf. Image Processing*, pp. 2573–2576, 2010.
- [53] Z. Chen and C. Guillemot, “Perceptually-friendly H.264/AVC video coding based on foveated Just-Noticeable-Distortion model,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20(6), pp. 806–819, June 2010.
- [54] D. Tan, C. Tan, H. Wu, “Perceptual Color Image Coding With JPEG2000,” *IEEE Trans. Image Process.*, vol. 19(2), pp. 374–383, June, 2010.
- [55] V. Lukin, M. Zriakhov, N. Ponomarenko, S. Krivenko, Z. Miao, “Lossy compression of images without visible distortions and its application,” in *Proc. Int. Conf. Signal Processing*, pp. 698–701, 2010.

References

- [56] D. Wu, D. Tan, H. Wu, "Perceptual coding at the threshold level for the Digital Cinema System Specification," in *Proc. Int. Conf. Multimedia and Expo*, pp. 796–801, 2010.
- [57] D. Wu, D. Tan, M. Baird, J. DeCampo, C. White, H. Wu, "Perceptually Lossless Medical Image Coding," *IEEE Trans. Medical Imaging*, vol. 25(3), pp. 335–344, 2006.
- [58] O. Han, A. Bilgin, W. Michael, "Multi-resolution visually lossless image coding using JPEG2000," in *Proc. Int. Conf. Image Processing*, pp. 2581–2584, 2010.
- [59] Z. Wei, K. Ngan, "The perceptually transparent coding for image," in *Proc. Int. Symp. Circuits and Systems*, pp. 1621–1624, 2009.
- [60] E. Tan, W. Gan, M. Wong, "Perceptually lossless coder based on just noticeable distortion profile with subband DCT," in *Proc. Int. Symp. Consumer Electronics*, pp. 253–257, 2005.
- [61] H. Wu and K. Rao, "Digital Video Image Quality and Perceptual Coding," *CRC Press*, 2005.
- [62] W. Lin, C. Kuo, "Perceptual Visual Quality Metrics: A Survey," *J. of Visual Communication and Image Representation*, vol. 22(4), pp. 297–312, May 2011.
- [63] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13(4), pp. 600–612, Apr. 2004.
- [64] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of dct basis functions," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan. 2007.
- [65] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15(2), pp. 430–444, Feb. 2006.
- [66] D. Chandler, S. Hemami, "VSNR: A wavelet-based visual signal-to-noise-ratio for natural images," *IEEE Trans. Image Process.*, vol. 16(9), pp. 2284–2298, Sept. 2007.

References

- [67] E. Larson and D. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19(1), pp. 011006:1–21, 2010.
- [68] Y. Liu, F. Wu, K. Ngan, "3-D Object-Based Scalable Wavelet Video Coding With Boundary Effect Suppression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17(5), pp. 639–644, May 2007.
- [69] V. Seran, L. Kondi, "New scaling coefficients for bi-orthogonal filter to control distortion variation in 3D wavelet based video coding," in *Proc. Int. Conf. Image Processing*, pp. 1873–1876, 2006.
- [70] X. Wu, X. Zhang, X. Wang, "Low Bit-Rate Image Compression via Adaptive Down-Sampling and Constrained Least Squares Upconversion," *IEEE Trans. Image Process.*, vol. 18(3), pp. 552–561, 2009.
- [71] Y. Tanaka, M. Hasegawa, S. Kato, M. Ikehara, T. Nguyen, "Adaptive Directional Wavelet Transform Based on Directional Prefiltering," *IEEE Trans. Image Process.*, vol. 19(4), pp. 934–945, 2010.
- [72] Information Technology-Lossless and near-lossless compression of continuous-tone images-Baseline. International Telecommunication Union (ITU-T Recommendation T.87). ISO/IEC 14495-1, 1998.
- [73] X. Li, X. Chen, X. Xie, G. Li, L. Zhang, C. Zhang, Z. Wang, "A Low Power, Fully Pipelined JPEG-LS Encoder for Lossless Image Compression," in *Proc. Int. Conf. Multimedia and Expo*, pp. 1906–1909, 2007.
- [74] S. Miaou, F. Ke, S. Chen, "A Lossless Compression Method for Medical Image Sequences Using JPEG-LS and Interframe Coding," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13(5), pp. 811–821, Sept. 2009.
- [75] M. Weinberger, and G. Seroussi, "From LOCO-I to the JPEG-LS Standard," *Computer Systems Laboratory*, HP Laboratories Palo Alto, HPL-1999-3, January 1999. [Online]. Available: <http://www.hpl.hp.com/loco/>.
- [76] Information technology - Digital compression and coding of continuous-tone still images - Part 1: Requirements and guidelines - First Edition; Corrigendum 1:2005. ISO/IEC 14495-1, 1994.

References

- [77] J. Byrne, S. Ierodiaconou, D. Bull, D. Redmill, P. Hill, “Unsupervised image compression-by-synthesis within a JPEG framework,” in *Proc. Int. Conf. Image Processing*, pp. 2892–2895, Oct. 2008.
- [78] R. Neelamani, R. Queiroz, Z. Fan, S. Dash, R. Baraniuk, “JPEG compression history estimation for color images,” *IEEE Trans. Image Process.*, vol. 15(6), pp. 1365–1378, June 2006.
- [79] International Organization for Standardization and International Electrotechnical Commission. ISO/IEC 15444-1:2000, Information technology—JPEG 2000 image coding system—Part 1: Core coding system.
- [80] G. Baruffa, P. Micanti, F. Frescura, “Error Protection and Interleaving for Wireless Transmission of JPEG 2000 Images and Video,” *IEEE Trans. Image Process.*, vol. 18(2), pp. 346–356, Feb. 2009.
- [81] M. Adams, R. Ward, “Jasper: A portable flexible open-source software tool kit for image coding/processing,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 241–244, 2004.
- [82] J. Meessen, C. Parisot, X. Desurmont, J. Delaigle, “Scene analysis for reducing motion JPEG 2000 video surveillance delivery bandwidth and complexity,” in *Proc. Int. Conf. Image Processing*, pp. 577–580, 2005.
- [83] T. Tillo, M. Grangetto, G. Olmo, “Multiple description coding with error correction capabilities: an application to motion JPEG 2000,” in *Proc. Int. Conf. Image Processing*, pp. 3129–3132, 2004.
- [84] Advanced video coding for generic audiovisual services. International Telecommunication Union (ITU-T Recommendation H.264). ISO/IEC 14496-10, 2009.
- [85] T. Wiegand, G. Sullivan, G. Bjontegaard, A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13(7), pp. 560–576, July 2003.
- [86] P. Lambert, W. Neve, P. Neve, I. Moerman, P. Demeester, R. Walle, “Rate-distortion performance of H.264/AVC compared to state-of-the-art video codecs,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16(1), pp. 134–140, Jan. 2006.

References

- [87] “JM Software,” [Online]. Available: <http://iphone.hhi.de/suehring/tml/>.
- [88] E. Carotti, J. Martin, A. Meo, “Low-complexity lossless video coding via adaptive spatio-temporal prediction,” in *Proc. Int. Conf. Image Processing*, pp. 197–200, 2003.
- [89] S. Kim, J. Han, J. Kim, “An efficient scheme for motion estimation using multireference frames in H.264/AVC,” *IEEE Trans. Multimedia*, vol. 8(3), pp. 457–466, June 2006.
- [90] C. Chiang, S. Lai, “Fast multi-reference motion estimation via statistical learning for H.264/AVC,” in *Proc. Int. Conf. Multimedia and Expo*, pp. 61–64, 2009.
- [91] P. Wu, C. Xiao, “An adaptive fast multiple reference frames selection algorithm for H.264/AVC,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1017–1020, 2008.
- [92] C. Cai, K. Lam, Z. Tan, “An Efficient Scene-Break Detection Method Based on Linear Prediction With Bayesian Cost Functions,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18(9), pp. 1318–1323, 2008.
- [93] G. Bjontegaard, “Calculation of average psnr differences between rd-curves,” *VCEG Contribution VCEG-M33*, April 2001.
- [94] Z. He and S. Mitra, “Optimum bit allocation and accurate rate control for video coding via ρ domain,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12(10), pp. 840–849, October 2002.
- [95] USC-SIPI Image Database. [Online]. Available: <http://sipi.usc.edu/services/database/database.cgi?volume=misc>
- [96] Methodology for the subjective assessment of the quality of television pictures. ITU-R Recommendation BT.500-11, 2002.
- [97] C. Shannon. “A mathematical theory of communications,” *Bell system tech. journal*, 27:397-423, 623-656, 1948.
- [98] S. Boltz, E. Wolsztynski, E. Debreuve, E. Thierry, M. Barlaud, L. Pronzato, “A Minimum-Entropy Procedure for Robust Motion Estimation,” *IEEE Trans. Pattern Analy. Mach. Intell.*, vol. 21(4), pp. 291–310, 1999.

References

- [99] M. Celik, G. Sharma, A. Tekalp, "Algorithm of image enhancement based on order morphology filtering and image entropy difference," in *Proc. Int. Conf. Image Processing*, pp. 509–512, 2003.
- [100] F. Wang, B. Vemuri, M. Rao and Y. Chen, "Cumulative residual entropy, a new measure of information & its application to image alignment," in *Proc. Int. Conf. Computer Vision*, pp. 548–553, 2003.
- [101] S. Zhu, Y. Wu, and D. Mumford, "Entropy based feature selection scheme for real time simultaneous localization and map building," *Int. J. Comput. Vis.*, vol. 27(2), pp. 107–126, 1998.
- [102] N. Balakrishnan, K. Hariharakrishnan, and D. Schonfeld, "An Image Edge Detection Algorithm Based on Local Entropy," *IEEE Trans. Pattern Analy. Mach. Intell.*, vol. 27(9), pp. 1367–1378, 2005.
- [103] J. Portilla and E. Simoncelli, "An entropy masking model for multimedia content watermarking," *Int. J. Comput. Vis.*, vol. 40(1), pp. 49–71, 2000.
- [104] A. Speis and G. Healey, "Entropy-Based 2D Image Dissimilarity Measure," *IEEE Trans. Image Process.*, vol. 5(3), pp. 447–458, 1996.
- [105] T. Pun, "A new method for gray-level picture thresholding using the entropy of the histogram," *Signal Processing*, vol. 2, pp. 223–237, 1980.
- [106] M. Eckert and A. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Process.*, vol. 70, pp. 177–200, 1998.
- [107] G. Fan, W. Cham, "Model-based edge reconstruction for low bit-rate wavelet compressed images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10(1), pp.120–132, 2000.
- [108] W. Yin, D. Goldfarb, and S. Osher, "Total variation-based image cartoon-texture decomposition," *University CORC Report TR-2005-01*, Columbia, UCLA CAM Report 0527, 2005.
- [109] W. Yin, D. Goldfarb, and S. Osher, "A comparison of three total variation based texture extraction models," *J. Visual Commun. and Image Representation*, vol. 18(3), pp. 240–252, June 2007.

References

- [110] J. Aujol, G. Gilboa, T. Chan, and S. Osher, "Structure-texture image decomposition—modeling, algorithms and parameter selection," *Int. J. Comput. Vis.*, vol. 67(1), pp. 111–136, Feb. 2006.
- [111] A. Watson, R. Borthwick and M. Taylor, "Image quality and entropy masking," *SPIE Conf.*, vol. 3016, 1997.
- [112] A. Houmansadr, H. Pirsiavash, and S. Ghaemmaghami, "Robust Content-based Video Watermarking Exploiting Motion Entropy Masking Effect," *Int. Conf. Signal Process. and Multimedia Appl.*, Aug. 2006.
- [113] J. Canny, "A Computational Approach To Edge Detection," *IEEE Trans. Pattern Analy. Mach. Intell.*, vol. 8, pp. 679–698, 1986.
- [114] X. Kang, C. Han, Y. Yang, and T. Tao, "SAR Image Edge Detection by Ratio-based Harris Method," *Int. Conf. Acoust., Speech, and Signal Process.*, vol. 2, pp. 837–840, 2006.
- [115] W. Lin, L. Dong, and P. Xue, "Visual distortion gauge based on discrimination of noticeable contrast changes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15(7), pp. 900–909, July 2005.
- [116] S. Channappayya, A. Bovik, and R. Heath, "Rate Bounds on SSIM Index of Quantized Images," *IEEE Trans. Image Process.*, vol. 17(9), pp. 1624–1639, 2008.
- [117] Y. Huang, T. Ou, P. Su, and H. Chen, "Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20(11), pp. 1614–1624, July 2010.
- [118] Z. Wang, A. Bovik and L. Lu, "Why is image quality assessment so difficult?" in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, May 2002.
- [119] X. Gao, W. Lu, D. Tao, and X. Li, "Image Quality Assessment Based on Multiscale Geometric Analysis," *IEEE Trans. Image Process.*, vol. 18(7), pp. 1409–1423, 2009.
- [120] Q. Ma, L. Zhang, and B. Wang, "New strategy for image and video quality assessment," *Journal of Electronic Imaging*, vol. 19(1), pp. 011019:1-14, 2010.
- [121] G. Chen, C. Yang, and S. Xie, "Edge-based structural similarity for image quality

- assessment,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 14–19, 2006.
- [122] G. Chen, C. Yang, and S. Xie, “Gradient-based structural similarity for image quality assessment,” in *Proc. Int. Conf. Image Processing*, pp. 2929–2932, 2006.
- [123] D. Kim, H. Han, and R. Park, “Gradient information-based image quality metric,” *IEEE Trans. Consumer Electronics*, vol. 56(2), pp. 930–936, 2010.
- [124] X. Ran and N. Farvardin, “A perceptually motivated three-component image model-Part I: Description of the model,” *IEEE Trans. Image Process.*, vol. 4(4), pp. 401–415, 1995.
- [125] E. Ong, W. Lin, Z. Lu, S. Yao and M. Etoh, “Visual Distortion Assessment with Emphasis on Spatially Transitional Regions,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14(4), PP. 559–566, 2004.
- [126] D. Rouse and S. Hemami, “Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with multi-scale SSIM,” in *Proc. Western New York Image Processing Workshop*, October 2007.
- [127] Z. Wang, E. Simoncelli, and A. Bovik, “Multi-scale structural similarity for image quality assessment,” in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, pp. 1398–1402, Nov. 2003.
- [128] Z. Wang and E. Simoncelli, “Translation insensitive image similarity in complex wavelet domain,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 573–576, 2005.
- [129] C. Yang, W. Gao, and L. Po, “Discrete wavelet transform-based structural similarity for image quality assessment,” in *Proc. Int. Conf. Image Processing*, pp. 377–380, 2008.
- [130] Z. Wang and Q. Li, “Information Content Weighting for Perceptual Image Quality Assessment,” *IEEE Trans. Image Process.*, vol. 20(5), pp. 1185–1198, 2011.
- [131] G. Cheng, J. Huang, C. Zhu, Z. Liu, L. Cheng, “Perceptual image quality assessment using a geometric structural distortion model,” in *Proc. Int. Conf. Image Processing*, pp. 325–328, 2010.

References

- [132] H. Sheikh, K. Seshadrinathan, A. Moorthy, Z. Wang, A. Bovik, and L. Cormack, "Image and video quality assessment research at LIVE." [Online]. Available: <http://live.ece.utexas.edu/research/quality/>.
- [133] D. Rouse, S. Hemami, "Understanding and Simplifying the Structural Similarity Metric," in *Proc. Int. Conf. Image Processing*, pp. 325–328, 2008.
- [134] N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, V. Lukin, "Metrics performance comparison for color image database," in *Fourth International Workshop on Video Process. and Quality Metrics for Consumer Electronics*, 2009.
- [135] N. Ponomarenko and K. Egiazarian, "Tampere Image Database 2008 TID2008." [Online]. Available: <http://www.ponomarenko.info/tid2008.htm>.
- [136] Y. Horita, K. Shibata, Y. Kawayoke, Z. Sazzad, "MICT Image Quality Evaluation Database." [Online]. Available: <http://mict.eng.u-toyama.ac.jp/mictdb.html>.
- [137] A. Ninassi, P. Callet, and F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding," in *Proc. SPIE: Human Vision and Electronic Imaging*, vol. 6057, Jan. 2006.
- [138] A. Ninassi, P. Callet, and F. Autrusseau, "Subjective quality assessment- IVC database." [Online]. Available: <http://www2.irccyn.ec-nantes.fr/ivcdb>.
- [139] E. Larson and D. Chandler, "Categorical Image Quality (CSIQ) Database." [Online]. Available: <http://vision.okstate.edu/csiq>.
- [140] Video Quality Expert Group, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment II," 2003.
- [141] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15(11), pp. 3440–3451, Nov. 2006.
- [142] A. Moorthy and A. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3(2), pp. 193–201, Apr. 2009.
- [143] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli, "The SSIM Index for Image Quality Assessment." [Online]. Available: <http://ece.uwaterloo.ca/~z70wang/research/ssim/>

References

- [144] MeTriX MuX Visual Quality Assessment Package. [Online]. Available: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/
- [145] PSNR-HVS-M download page. [Online]. Available: <http://www.ponomarenko.info/psnrhvs.htm>
- [146] E. Larson and D. Chandler, "Full-Reference Image Quality Assessment and the Role of Strategy: The Most Apparent Distortion." [Online]. Available: <http://vision.okstate.edu/mad/>
- [147] Z. Wang, "IW-SSIM: Information Content Weighted Structural Similarity Index for Image Quality Assessment." [Online]. Available: <https://ece.uwaterloo.ca/~z70wang/research/iwssim/>
- [148] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*, the MIT press, pp. 207–220, 1993.
- [149] Z. Wang and A. Bovik, "Mean squared error: love it or leave it? -A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26(1), pp. 98–117, Jan. 2009.
- [150] Z. Yu, H. Wu, S. Winkler, and T. Chen, "Vision-model-based impairment metric to evaluate blocking artifacts in digital video," *Proc. IEEE*, vol. 90(1), pp. 154–169, 2002.
- [151] J. Dong, K. Ngan, C. Fong and W. Cham, "2D Order-16 Integer Transforms for HD Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19(10), pp. 1462–1474, 2009.
- [152] J. Dong and K. Ngan, "Parametric Interpolation Filter for HD Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20(12), pp. 1892–1897, 2010.
- [153] C. Deng, W. Lin, B. Lee, C. Lau, and M. Sun, "Performance Analysis, Parameter Selection and Extensions to H.264/AVC FRExt for Super-High Definition Video Coding," *J. Visual Commun. and Image Representation*, 2011. [Online]. Available: <http://dx.doi.org/10.1016/j.jvcir.2011.01.004/>
- [154] X. Wu, G. Zhai, X. Yang, W. Zhang, "Adaptive Sequential Prediction of Multidimensional Signals With Applications to Lossless Image Coding," *IEEE Trans. Image Process.*, vol. 20(1), pp. 36–42, 2011.

- [155] Y. Wei and J. Yang, "Adaptive mode-dependent scan for H.264/AVC intracoding," *Journal of Electronic Imaging*, vol. 19(3), pp. 033008 1–12, 2010.
- [156] J. Heo, S. Kim, and Y. Ho, "Improved CAVLC for H.264/AVC Lossless Intra-Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20(2), pp. 213–222, 2010.
- [157] M. Bouchakour, G. Jeannic, F. Autrusseau, "JND mask adaptation for wavelet domain watermarking," in *Proc. Int. Conf. Multimedia and Expo*, pp. 201–204, 2008.
- [158] T. Huang, C. Liang, S. Yeh, H. Chen, "JND-based enhancement of perceptibility for dim images," in *Proc. Int. Conf. Image Processing*, pp. 1752–1755, 2008.
- [159] C. Chiang, S. Wang, Y. Chen, S. Lai, "Fast JND-Based Video Carving With GPU Acceleration for Real-Time Video Retargeting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19(11), pp. 1588–1597, Nov. 2009.
- [160] N. Cliff, *Analyzing Multivariate Data*. San Diego, CA: Harcourt Brace Jovanovich, 1987.
- [161] J. Harel, C. Koch and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems*, pp. 545–552, 2006.
- [162] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in Neural Information Processing Systems*, pp. 155–162, 2006.
- [163] T. Liu, J. Sun, N. Zheng, X. Tang and H. Shum, "Learning to detect a salient object," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 1–8, 2007.
- [164] X. Hou and L. Zhang, "Saliency Detection: A spectral residual approach," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 1–8, 2007.
- [165] W. Cheng, C. Wang, J. Wu, "Video Adaptation for Small Display Based on Content Recomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17(1), pp. 43–58, 2007.
- [166] X. Feng, T. Liu, D. Yang, Y. Wang, "Saliency based objective quality assessment of decoded video affected by packet losses," in *Proc. Int. Conf. Image Processing*, pp. 2560–2563, 2008.

- [167] L. Yang, N. Zheng, J. Yang, M. Chen, H. Chen, “A biased sampling strategy for object categorization,” in *Proc. Int. Conf. Computer Vision*, pp. 1141–1148, 2009.
- [168] Z. Chen; C. Guillemot, “Perceptually-Friendly H.264/AVC Video Coding Based on Foveated Just-Noticeable-Distortion Model,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20(6), pp. 806–819, 2010.
- [169] W. Yin, J. Luo, C. Chen, “Event-Based Semantic Image Adaptation for User-Centric Mobile Display Devices,” *IEEE Trans. Multimedia*, vol. 13(3), pp. 432–442, 2011.
- [170] S. Wang, S. Lai, “Compressibility-Aware Media Retargeting With Structure Preserving,” *IEEE Trans. Image Process.*, vol. 20(3), pp. 855–865, 2011.
- [171] K. Seshadrinathan and A. Bovik, “A structural similarity metric for video based on motion models,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1869–1872, 2007.
- [172] A. Ninassi, O. Meur, P. Callet, and D. Barba, “Considering temporal variations of spatial visual distortions in video quality assessment,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 3(2), pp. 253–265, 2009.
- [173] A. Moorthy and A. Bovik, “Efficient Video Quality Assessment along Temporal Trajectories,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20(11), pp. 1653–1658, 2010.

Publications

Journal Papers

- A. Liu, W. Lin, M. Paul, F. Zhang, C. Deng, “Optimal Compression Plane for Efficient Video Coding,” *IEEE Transactions on Image Processing*, vol. 20(10), pp. 2788-2799, 2011.

(A part of Chapter 3 in this thesis is based upon this paper)

- A. Liu, W. Lin, M. Paul, C. Deng, F. Zhang, “Just Noticeable Difference for Images with Decomposition Model for Separating Edge and Textured Regions,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20(11), pp. 1648-1652, 2010.

(A part of Chapter 4 in this thesis is based upon this paper)

- A. Liu, W. Lin, M. Narwaria, “Image Quality Assessment Based on Gradient Similarity,” *IEEE Transactions on Image Processing*, accepted for publication, 2011.

(A part of Chapter 6 in this thesis is based upon this paper)

- A. Liu, W. Lin, Z. Chen, “Just Noticeable Difference Based Histogram Adjustment for Visually Lossless Image Coding,” *IEEE Signal Processing Letters*, submitted for review.

(A part of Section 5.2 in this thesis is based upon this paper)

Conference Papers

- A. Liu, W. Lin, M. Paul, C. Deng, “Optimal Compression Plane (OCP)-A New Framework for H.264 Video Coding,” in *Proc. Int. Conf. Multimedia and Expo*, pp. 456-461, 2010.
(A part of Chapter 3 in this thesis is based upon this paper)
- A. Liu, W. Lin, F. Zhang, M. Paul, “Enhanced Just Noticeable Difference (JND) Estimation with Image Decomposition,” in *Proc. Int. Conf. Image Processing*, pp. 317-320, 2010.
(A part of Chapter 4 in this thesis is based upon this paper)
- A. Liu, W. Lin, F. Zhang, “Lossless Video Compression with Optimal Compression Plane Determination,” in *Proc. Int. Conf. Multimedia and Expo*, pp. 173-176, 2009.
(A part of Chapter 3 in this thesis is based upon this paper)
- A. Liu, W. Lin, “Perception Based Down Sampling for Low Bit Rate Image Coding,” in *Proc. Int. Conf. Pacific-Rim Conference on Multimedia*, 2009.
(A part of Section 5.1 in this thesis is based upon this paper)
- A. Liu, M. Verma, W. Lin, “Modeling the Masking Effect of the Human Visual System with Visual Attention Model,” in *Proc. Int. Conf. Information, Communications and Signal Processing*, 2009.