Perceptual Quality Assessment and Processing for Visual Signals

MA, Lin

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Electronic Engineering

The Chinese University of Hong Kong
December 2012

題獻/Dedication

獻給我的父母

To my parents

致謝

首先也是最重要的，我要感謝我的父母，在我整個的生命中，他們給予我無私的愛跟無條件的支持。無論何時我遇到什麼困難，他們都能夠給予我最大的鼓勵以及支持。感謝他們給予我生命，撫養我長大成人。他們的愛是我生活前進和追尋夢想的動力。

我要衷心感謝我的論文的指導老師顏慶義教授，顏教授學識淵博，對前沿的研究方向有著敏銳的洞察力。在我攻讀博士學位期間，在學術研究上，他給予了我很大的幫助以及指導。而且他支持我參加了很多國際性的學術會議以及海外交流研究。這些經歷拓寬了我的研究視野。

非常感謝新加坡南洋理工大學的林偉思教授。他是我海外交流研究的指導教授。我在他的研究組里面學習了三個月。其間，他在研究項目以及生活上給予我很大的幫助以及指導。感謝圖像與視頻處理實驗室的湛偉權教授，Thierry Blu教授，王曉剛教授和徐空達教授。他們對問題的真知灼見給予了我很多的建議和幫助。

感謝圖像與視頻處理實驗室的伙伴：陳震中，劉宇，董潔，李杰，魏振宇，麥振文，馮志強，張偉，崔春暉，張茜，歐陽萬里，張任奇，劉強，趙從，李松南，張帆，李超，程邦盛，潘漢杰，趙瑞，薛峰，王妙輝，盛律，韓慶龍。我們在一起度過了非常歡樂的時光。這些永恒的回憶我會永遠的珍藏。另外我要感謝圖像與視頻處理實驗室技術管理員王玉忠和系辦公室的各位員工為我的工作所創造的便利工作環境。另外我還要感謝南洋理工大學的同事們：鄧晨偉，劉安民，方玉明，楊環，董路，付圓圓，張超。跟他們相處的時光是我永遠美好的回憶。

最后我要感謝我的好朋友們：吳國盛，王沖，張洋，王寶勛，于博。他們非常的友好。在我最需要幫助的時候給了我很多的支持和鼓勵。

Acknowledgements

First and foremost, I would like to thank my parents for their unlimited love and unconditional support though my whole life. Whenever I have difficult time, I can always turn to them for the strongest support from them. Thanks a lot for giving birth to me, and raising me up healthily. Their love motivates me to pursue my life and my dreams.

I would like to thank my thesis advisor Prof. King Ngi Ngan for his supervision. Prof. Ngan is a learned scholar with penetrating insight of the research frontiers. He gave me a lot of help and great supervision on my research through my Ph.D study. And he supported me to attend international academic conferences and took part in overseas exchanges, which greatly opened my research horizon.

And it is my great pleasure to express my gratitude towards Prof. Weisi Lin from Nanyang Technological University. He is the host supervisor when I did the overseas research. I spent three months in his group, where he gave me great help and guidances on doing the research project. I would also like to thank Prof. Wai Kuen Cham, Prof. Thierry Blu, Prof. Xiaogang Wang and Prof. Hung Tat Tsui, who are faculty members of the IVP Lab, for their useful suggestions on my research work.

I must express my appreciations to my colleagues in IVP Lab, *i.e.*, Zhenzhong Chen, Yu Liu, Jie Dong, Jie Li, Zhenyu Wei, Chun Man Mak, Chi Keung Fong, Wei Zhang, Chunhui Cui, Qian Zhang, Wanli Ouyang, Renqi Zhang, Qiang Liu, Cong Zhao, Songnan Li, Fan Zhang, Chao Li, Bangsheng Cheng, Hanjie Pan, Rui Zhao, Feng Xue, Miaohui Wang, Lv Sheng, Qinglong Han, and etc. We had happy times together which I will treasure forever. Moreover, I would like to express my gratitude to the colleagues in Nanyang Technological University, *i.e.*, Chenwei Deng, Anmin Liu, Yuming Fang, Huan Yang, Lu Dong, Yuanyuan Fu, and Chao Zhang. The happy times we spent together will be in my memories forever.

Last but not least, I would like to thank my sincere friends, *i.e.*, Guosheng Wu,

Chong Wang, Yang Zhang, Baoxun Wang, and Bo Yu. They are very nice, who gave me many suggestions and supports, when I have difficulties.

摘要

視覺信號，包括圖像，視頻等，在采集，壓縮，存儲，傳輸，重新生成的過程中都會被各種各樣的噪聲所影響，因此他們的主觀質量也就會降低。所以，主觀視覺質量在現今的視覺信號處理跟通訊系統中起到了很大的作用。這篇畢業論文主要討論質量評價的算法設計，以及這些衡量標準在視覺信號處理上的應用。這篇論文的工作主要包括以下五個方面。

第一部分主要集中在具有完全參考原始圖像的圖像質量評價。首先我們研究人類視覺系統的特征。具體說來，視覺在結構化失真上面的水平特性和顯著特征會被建模然后應用到結構相似度(SSIM)這個衡量標準上。實驗顯示我們的方法明顯的提高了衡量標準與主觀評價的相似度。由這個質量衡量標準的啟發，我們設計了一個主觀圖像壓縮的方法。其中我們提出了一個自適應的塊大小的超分辨率算法指導的下采樣的算法。實驗結果證明提出的圖像壓縮算法無論在主觀還是在客觀層面都構建了高質量的圖像。

第二個部分的工作主要討論具有完全參考原始視頻的視頻質量評價。考慮到人類視覺系統的特征，比如時空域的對比敏感函數，眼球的移動，紋理的遮掩特性，空間域的一致性，時間域的協調性，不同塊變換的特性，我們設計了一個自適應塊大小的失真閾值的模型。實驗證明，我們提出的失真閾值模型能夠更精確的描述人類視覺系統的特性。基于這個自適應塊大小的失真閾值模型，我們設計了一個簡單的主觀質量評價標準。在公共的圖像以及視頻的主觀數據庫上的測試結果證明了這個簡單的評價標準的有效性。因此，我們把這個簡單的質量標準應用于視頻編碼系統中。它可以在同樣的碼率下提供更高主觀質量的視頻。

第三部分我們討論具有部分參考信息的圖像質量評價。我們通過描述重組后的離散余弦變換域的系數的統計分布來衡量圖像的主觀質量。提出的評價標準發掘了相鄰的離散余弦系數的相同統計特性，相鄰的重組離散余弦系數的互信息，以及圖像的能量在不同頻率下的分布。實驗結果證明我們提出的質量標準可以超越其他的具有部分參考信息的質量評價標準，甚至還超過了具有完全參考信息的質量評價標準。而且，提取的特征很容易被編碼以及隱藏到圖像中以便于在圖像通訊中進行質量監控。

第四部分我們討論具有部分參考信息的視頻質量評價。我們提取的特征可以很好的描述空間域的信息丢失，和時間域的相鄰兩幀間的直方圖的統計特性。在視頻主觀質量的數據庫上的實驗結果，也證明了提出的方法可以超越其他代表性的視頻質量評價標準，甚至是具有完全參考信息的質量評價標準，譬如：PSNR以及SSIM。我們的方法只需要很少的特征來描述每一幀視頻圖像。對于每一幀圖像，一個特征用于描述空間域的特點，另外三個特征用于描述時間域的特點。考慮到計算的復雜度以及壓縮特征所需要的碼率，提出的方法可以很簡單的在視頻的傳輸過程中監控視頻的質量。

之前的四部分提到的主觀質量評價標準主要集中在傳統的失真上面，譬如：JPEG圖像壓縮，H.264視頻壓縮。在最后一部分，我們討論在圖像跟視頻的retargeting過程中的失真。現如今，隨著消費者電子的發展，視覺信號需要在不同分辨率的顯示設備上進行通訊交互。因此，retargeting的算法把同一個原始圖像適應于不同的分辨率的顯示設備。這樣的過程就會引入圖像的失真。我們研究了對于retargeting圖像主觀質量的測試者的分數，從三個方面進行討論測試者對于retargeting圖像失真的反應：圖像retargeting的尺度，圖像retargeting的算法，原始圖像的內容特性。通過大量的主觀實驗測試，我們構建了一個關于圖像retargeting的主觀數據庫。基于這個主觀數據庫，我們評價以及分析了幾個具有代表性的質量評價標準。

Abstract

Visual signals, including images, videos, etc., are affected by a wide variety of distortions during acquisition, compression, storage, processing, transmission, and reproduction processes, which result in perceptual quality degradation. As a result, perceptual quality assessment plays a very important role in today's visual signal processing and communication systems. In this thesis, quality assessment algorithms for evaluating the visual signal perceptual quality, as well as the applications on visual signal processing and communications, are investigated. The work consists of five parts as briefly summarized below.

The first part focuses on the full-reference (FR) image quality assessment. The properties of the human visual system (HVS) are firstly investigated. Specifically, the visual horizontal effect (HE) and saliency properties over the structural distortions are modelled and incorporated into the structure similarity index (SSIM). Experimental results show significantly improved performance in matching the subjective ratings. Inspired by the developed FR image metric, a perceptual image compression scheme is developed, where the adaptive block-based super-resolution directed down-sampling is proposed. Experimental results demonstrated that the proposed image compression scheme can produce higher quality images in terms of both objective and subjective qualities, compared with the existing methods.

The second part concerns the FR video quality assessment. The adaptive block-size transform (ABT) based just-noticeable difference (JND) for visual signals is investigated by considering the HVS characteristics, e.g., spatio-temporal contrast sensitivity function (CSF), eye movement, texture masking, spatial coherence, temporal consistency, properties of different block-size transforms, etc. It is verified that the developed ABT based JND can more accurately depict the HVS property, compared with the state-of-the-art JND models. The ABT based JND is thereby utilized to develop a simple perceptual quality metric for visual signals. Validations on the image and video

subjective quality databases proved its effectiveness. As a result, the developed perceptual quality metric is employed for perceptual video coding, which can deliver video sequences of higher perceptual quality at the same bit-rates.

The third part discusses the reduced-reference (RR) image quality assessment, which is developed by statistically modelling the coefficient distribution in the reorganized discrete cosine transform (RDCT) domain. The proposed RR metric exploits the identical statistical nature of the adjacent DCT coefficients, the mutual information (MI) relationship between adjacent RDCT coefficients, and the image energy distribution among different frequency components. Experimental results demonstrate that the proposed metric outperforms the representative RR image quality metrics, and even the FR quality metric, *i.e.*, peak signal to noise ratio (PSNR). Furthermore, the extracted RR features can be easily encoded and embedded into the distorted images for quality monitoring during image communications.

The fourth part investigates the RR video quality assessment. The RR features are extracted to exploit the spatial information loss and the temporal statistical characteristics of the inter-frame histogram. Evaluations on the video subjective quality databases demonstrate that the proposed method outperforms the representative RR video quality metrics, and even the FR metrics, such as PSNR, SSIM in matching the subjective ratings. Furthermore, only a small number of RR features is required to represent the original video sequence (each frame requires only 1 and 3 parameters to depict the spatial and temporal characteristics, respectively). By considering the computational complexity and the bit-rates for extracting and representing the RR features, the proposed RR quality metric can be utilized for quality monitoring during video transmissions, where the RR features for perceptual quality analysis can be easily embedded into the videos or transmitted through an ancillary data channel.

The aforementioned perceptual quality metrics focus on the traditional distortions, such as JPEG image compression noise, H.264 video compression noise, and so on. In the last part, we investigate the distortions introduced during the image and video retargeting process. Nowadays, with the development of the consumer electronics, more and more visual signals have to communicate between different display devices of different resolutions. The retargeting algorithm is employed to adapt a source image of one resolution to be displayed in a device of a different resolution, which may introduce

distortions during the retargeting process. We investigate the subjective responses on the perceptual qualities of the retargeted images, and discuss the subjective results from three perspectives, *i.e.*, retargeting scales, retargeting methods, and source image content attributes. An image retargeting subjective quality database is built by performing a large-scale subjective study of image retargeting quality on a collection of retargeted images. Based on the built database, several representative quality metrics for retargeted images are evaluated and discussed.

Publications

Journal Papers

- **Lin Ma**, Songnan Li, and King Ngi Ngan, "Reduced-reference image quality assessment in reorganized DCT domain," accepted for publication in *Signal Processing: Image Communication, Special Issue on Biologically Inspired Approaches for Visual Information Processing and Analysis*.

- **Lin Ma**, Weisi Lin, Chenwei Deng, and King Ngi Ngan, "Image retargeting quality assessment: a study of subjective scores and objective metrics," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on New Subjective and Objective Methodologies for Audio and Visual Signal Processing*, vol. 6, no. 6, pp. 626-639, Oct. 2012.

- **Lin Ma**, Songnan Li, and King Ngi Ngan, "Reduced-reference quality assessment of compressed videos sequences," *IEEE Transactions on Circuits and System for Video Technology*, vol. 22, no. 10, pp. 1441-1456, Oct. 2012.

- **Lin Ma**, Songnan Li, Fan Zhang, and King Ngi Ngan "Reduced-reference image quality assessment using reorganized DCT-based image representation," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 824-829, Aug. 2011.

- **Lin Ma**, King Ngi Ngan, Fan Zhang, and Songnan Li, "Adaptive block-size transform based just-noticeable difference model for images/vidoes," *Signal Processing: Image Communication*, vol. 26, no. 3, pp. 162-174, Mar. 2011.

- **Lin Ma**, Songnan Li, and King Ngi Ngan "Visual horizontal effect for image quality assessment," *IEEE Signal Processing Letters*, vol. 17, no. 7, pp. 627-630, Jul. 2010.

- **Lin Ma**, Debin Zhao, and Wen Gao, "Learning-based image restoration for compressed images," *Signal Processing: Image Communication*, vol. 27, no. 1, pp.

54-65, Jan. 2012.

- Long Xu, Songnan Li, King Ngi Ngan, and **Lin Ma** "Consistent visual quality control in video coding," accepted for publication in *IEEE Transactions on Circuits and System for Video Technology.*

- Yaqing Niu, Matthew Kyan, **Lin Ma**, Azeddine Beghdadi, and Sridhar Krishnan, "Visual saliency's modulatory effect on just noticeable distortion profile and its application in image watermarking," accepted for publication in *Signal Processing: Image Communication, Special Issue on Biologically Inspired Approaches for Visual Information Processing and Analysis.*

- Songnan Li, **Lin Ma**, and King Ngi Ngan, "Full-reference video quality assessment by decoupling detail losses and additive impairments," *IEEE Transactions on Circuits and System for Video Technology*, vol. 22, no. 7, pp. 1100-1112, Jul. 2012.

- Fan Zhang, **Lin Ma**, Songnan Li, and King Ngi Ngan, "Practical image quality metric applied to image coding," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 615-624, Aug. 2011.

- Songnan Li, Fan Zhang, **Lin Ma**, and King Ngi Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935-949, Oct. 2011.

Conference Papers

- **Lin Ma**, Chenwei Deng, Weisi Lin, and King Ngi Ngan, "Overview of quality assessment for visual signals and newly emerged trends," in *Proceedings of 2013 IEEE International Symposium on Circuits and Systems (ISCAS 2013)*, Beijing, China, May. 19-23, 2013.

- **Lin Ma**, Weisi Lin, Chenwei Deng, and King Ngi Ngan, "Study of subjective and objective quality assessment of retargeted images," in *Proceedings of 2012 IEEE International Symposium on Circuits and Systems (ISCAS 2012)*, Seoul, Korea, May. 20-23, 2012.

- **Lin Ma**, Songnan Li, and King Ngi Ngan, "Reduced-reference image quality assessment via intra- and inter-subband statistical characteristics in reorganized DCT domain," in *Proceedings of 2011 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2011)*, Xi'an, China, Oct. 18-21, 2011.

- **Lin Ma**, Songnan Li, and King Ngi Ngan, "Motion trajectory based visual saliency for video quality assessment," in *Proceedings of 2011 International Conference on Image Processing (ICIP 2011)*, Brussels, Belgium, Sep. 11-14, 2011.

- **Lin Ma**, Songnan Li, and King Ngi Ngan, "Perceptual image compression via adaptive block-based super-resolution directed down-sampling," in *Proceedings of 2011 International Symposium on Circuits and Systems (ISCAS 2011)*, Rio de Janeiro, Brazil, May. 15-18, 2011.

- **Lin Ma**, Fan Zhang, Songnan Li, and King Ngi Ngan, "Video quality assessment based on adaptive block-size transform just-noticeable difference model," in *Proceedings of 2010 International Conference on Image Processing (ICIP 2010)*, Hong Kong, China, Sep. 26-29, 2010.

- **Lin Ma**, and King Ngi Ngan, "Adaptive block-size transform based just-noticeable difference profile for videos," in *Proceedings of 2010 International Symposium on Circuits and Systems (ISCAS 2010)*, Paris, France, May. 30 - Jun. 2, 2010.

- **Lin Ma**, and King Ngi Ngan, "Adaptive block-size transform based just-noticeable difference profile for images," in *Proceedings of 2009 Pacific-Rim Conference on Multimedia (PCM 2009)*, Bangkok, Thailand, Dec. 15-18, 2009.

- **Lin Ma**, Feng Wu, Debin Zhao, Wen Gao, and Siwei Ma, "Learning-based image restoration for compressed image through neighbouring embedding," in *Proceedings of 2008 Pacific-Rim Conference on Multimedia (PCM 2008)*, Tainan, Taiwan, Dec. 9-13, 2008. **(Best Paper Award)**.

- **Lin Ma**, Yonghua Zhang, Yan Lu, Feng Wu, and Debin Zhao, "Three-tiered network model for image hallucination," in *Proceedings of 2008 International Conference on Image Processing (ICIP 2008)*, San Diego, California, U.S.A., Oct. 12-15, 2008.

- Long Xu, King N. Ngan, Songnan Li, and **Lin Ma**, "Video Quality Metric for Consistent Visual Quality Control in Video Coding", in *Proceedings of 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2012)*, Hollywood, California, U.S.A., Dec. 3-6, 2012.

- Songnan Li, **Lin Ma**, and King Ngi Ngan, "Video quality assessment by decoupling additive impairments and detail losses," in *Proceedings of 2011 International Workshop on Quality of Multimedia Experience (QoMEX 2011)*, Mechelen, Belgium, Sep. 7-9, 2011.

- Songnan Li, **Lin Ma**, Fan Zhang, and King Ngi Ngan, "Temporal inconsistency measure for video quality assessment," in *Proceedings of 2010 Picture Coding Symposium (PCS 2010)*, Nagoya, Japan, Dec. 7-10, 2010.

- Yaqing Niu, Matthew Kyan, **Lin Ma**, Azeddine Beghdadi, and Sridhar Krishnan, "A visual saliency modulated just noticeable distortion profile for image watermarking," in *Proceedings of 2011 European Signal Processing Conference (EUSIPCO 2011)*, Barcelona, Spain, Aug. 29 - Sep. 2, 2011.

- Fan Zhang, Songnan Li, **Lin Ma**, and King Ngi Ngan, "Limitation and challenge of image quality measurement," in *Proceedings of 2010 SPIC Visual Communications and Image Processing (VCIP 2010)*, Huangshan, China, Jul. 11-14, 2010. **(Invited paper)**.

- Shaohui Liu, **Lin Ma**, Hongxun Yao, and Debin Zhao, "Universal steganalysis based on statistical models using reorganization of block-based DCT coefficients," in *Proceedings of 2009 International Conference on Information Assurance and Security (IAS 2010)*, Xi'an, China, Aug. 18-20, 2009.

Patents

- King Ngi Ngan, **Lin Ma**, Wai Kuen Cham, and Yu Liu, "Method and apparatus for video coding by ABT-based just noticeable difference model," US Non-Provisional Patent Application No. 12 / 750401, Mar. 2010.

- King Ngi Ngan, **Lin Ma** Wai Kuen Cham, and Yu Liu, "通过基于自适应尺寸

变换ABT的最小可觉差JND模型进行视频编码的方法和装置," Patent Pending. Chinese Patent Application No. 20101074145.5.

- Yonghua Zhang, **Lin Ma**, and Feng Wu, "Image upsampling with training images," US Non-Provisional Patent Application No. 12 / 234572, Sep. 2008.

Nomenclature

Abbreviations

2D	Two-Dimension
3D	Three-Dimension
ABT	Adaptive Block-size Transform
ACR	Absolute Category Rating
BDS	BiDirectional Similarity
ANSI	Ametrican National Standards Institute
BME	Block-based Motion Estimation
CBD	City-Block Distance
CPU	Central Processing Unit
CROP	CROPping
CRT	Cathode Ray Tube
CSF	Contrast Sensitivity Function
DCT	Discrete Cosine Transform
DMOS	Differential Mean Opinion Score
DSCQS	Double-Stimulus Continuous Quality-Scale
DSIS	Double-Stimulus Impairment Scale
DWT	Discrete Wavelet Transform
EH	Edge Histogram
EL	Energy Loss
EMD	Earth Mover's Distance
ENER	ENERgy-based deformation
EPSNR	Edge Peak Signal-to-Noise Ratio
EVD	Energy Variation Descriptor
FF	Fast Fading
FR	Full Reference
FRD	Frequency Ratio Descriptor
FT	Fourier Transform
GGD	Generalized Gaussian Density
HD	High Definition
HE	Horizontal Effect
HF	High Frequency
HR	Hit Rate
HVS	Human Visual System
ICW-SSIM	Information Content-Weighted Structure SIMilarity

IP	Internet Protocol
IQA	Image Quality Assessment
ITU	International Telecommunication Union
JNB	Just-Noticeable Blur
JND	Just-Noticeable Difference
JPEG	Joint Photographic Experts Group
KLD	Kullback-Leibler Distance
LCC	Linear Correlation Coefficient
LDW-SSIM	Local Distortion Weighted Structure SIMilarity
LF	Low Frequency
MCS	Motion Characteristic Similarity
MF	Medium Frequency
MI	Mutual Information
MOS	Mean Opinion Score
MOVIE	MOtion-based Video Integrity Evaluation
MPEG	Motion Picture Expert Group
M-RDO	Modified Rate-Distortion Optimization
MSE	Mean Square Error
MSSIM	Multi-scale Structure SIMilarity index
MULT	MULTi-operator
MV	Motion Vector
NR	No Reference
OR	Outlier Ratio
PDF	Probability Density Function
PE	Prediction Error
PQA	Perceptual Quality Assessment
PS	Phase Spectrum
PSNR	Peak Signal-to-Noise Ratio
OBME	Overlapped Block-based Motion Estimation
QFT	Quaternion Fourier Transform
QoE	Quality of user Experiences
QP	Quantization Parameter
QP	Quaternion Representation
RDCT	Reorganized Discrete Cosine Transform
RDO	Rate-Distortion Optimization
RMSE	Root Mean Squared Error
RR	Reduced Reference
SA	SAliency map

SCAL	SCALling
SCS	Spatial Content Similarity
SCSC	optimized Seam Carving and SCale
SCST	SCale and STretch
SD	Structural Distortion
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation
SEAM	SEAM carving
SHIF	SHIFt-map editing
SIFT-flow	Scale-Invariant Feature Transform flow
SMW-SSIM	SMooth region-Weighted Structure SIMilarity
SNR	Signal to Noise Ratio
SR	Spectral Residual
SRDDS	Super-Resolution Directed Down-Sampling
SROCC	Spearman Rank-Order Correlation Coefficient
SSIM	Structural SIMilarity index
SS	Single Stimulus
STVI	STeaming VIdeo
TCSF	Temporal Contrast Sensitivity Function
VIF	Video Information Fidelity
VQA	Visual Quality Assessment
VQEG	Video Quality Expert Group
VQI	Visual Quality Index
VQ Model	Video Quality Metric
VSNR	Visual Signal-to-Noise Ratio
WARP	non-homogeneous retargeting
WGN	White Gaussian Noise
WNISM	Wavelet-domain Natural Image Statistical Metric

Notations

$\lvert \cdot \rvert$	Absolute value
$\lVert \cdot \rVert_2^2$	l_2 norm
$\mathbf{I}$	Original image
(T)	Distorted image
$\mathbf{S}_{HE}$	HE sensitivity function
erf	Error function
$exp(\cdot)$	Exponential function
$\lVert \cdot \rVert_F$	Frobenius norm
g	Gaussian filter
G_{TC}	Coding gain
$mean(\cdot)$	Average process
V_{sign}	Sign of the coefficient
$\otimes$	Kronecker product
ω_{ij}	Spatial frequency
ω_t	Temporal frequency
φ_{ij}	Directional angle of DCT subband
σ	Variance
ξ	Fourier transform
ξ^{-1}	inverse Fourier transform
Γ	Gamma function

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Objectives

Information is exploding with the progresses of technologies, and the developments of the consumer electronics. Nowadays, most of the information is presented to customers in the form of visual signals, including images, videos, and etc., as intuitive and faithful depiction of things in life and work. Therefore, electronic devices (e.g. phone cameras) and services (e.g. YouTube, and IPTV) based on the visual signals have increasingly emerged, which can capture and provide visual signals with better perceptual quality. Better quality of experience (QoE) [1] for customers is thereby provided and gained more interests of both the research communities and industries.

The objective of visual signal processing is to manipulate visual signals to provide consumers the desirable affects, which can deliver more pleased information. And the objective of visual signal communication is to ensure proper transmission of the visual signals from the server/producer side to the receiver/consumer side, which are of acceptable perceptual quality. However, as the typical multimedia service chain consists of sequential processing stages, e.g., acquisition, editing compression, transmission, reconstruction, restoration, presentation, etc., distortions will be inevitably introduced, which will degrade to the perceptual qualities of the visual signals. For example, during the video sequence transmission process for YouTube, the bandwidth of the transmission networks may be limited. Some frames may be dropped or skipped, especially for the video sequences with high spatial resolutions. This in turn will make the latency time of the requested video sequence intolerable. As a result, the satisfaction and enjoyment level of the viewers/customers for whom these visual signals provide are scarified. Perceptual quality assessment plays an important role for visual signal processing and communication.

Given that the ultimate receiver of the visual signals are human eyes, the human subjective opinion is the most reliable value for indicating the perceptual quality of the visual signal. The subjective opinions are obtained through the subjective testing, where a large number of viewers participate in the evaluation process and provide their personal opinions on the perceptual quality of the visual signal according to some pre-defined scales. After processing these subjective scores across the human viewers, a quality score, e.g. mean opinion score (MOS), differential mean opinion score (DMOS), etc., is finally obtained to indicate the perceptual quality of the visual signal. Moreover, in order to ensure repeatable and statistically meaningful results, subjective testing methods should precisely follow the standards [5]- [11] to set up the testing environment, and should recruit sufficient subjects to account for individual differences. The obtained subjective rating value can be regarded as the ground truth of the visual signal perceptual quality. Therefore, they can be employed to reliably evaluate the performances of the algorithms or methods of the visual signal processing. As a result, more and more attentions of the research communities and industries have been paid to the subjective testing methods.

Nowadays, video cameras for capturing high resolution video sequences, e.g. 720P and 1080P, 3D video cameras, depth cameras, eye tracker devices, and Kinect have been invented and developed for real-life applications. Better QoE can be provided due to the developments of these technologies and consumer electronics. However, new challenges are also issued meantime. In order to provide better perceptual quality of visual signals for the customers, the subjective responses need to be further studied and researched. As such, many subjective studies on the perceptual qualities of the emerged visual signals have been presented. In [16]- [24], [56], perceptual qualities of video sequences of different distortions have been studied, which consider not only the standard definition (SD) video sequences [16]- [19], but also high definition (HD) video sequences [23] [24]. Moreover, the perceptual quality of the scalable video sequences [21] [22] and 3D video sequences [20] have been recently discussed. In [25]- [34], the perceptual qualities of images are discussed and researched. The subjective responses of the images distorted by the traditional distortions are studied in [25]- [29]. The affects of wireless transmission on the image perceptual quality are discussed in [30]. Also the subjective opinions on the art image qualities are further researched [32]. Moreover, the subjective opinions on

the stereo images are discussed in [31]. Nowadays, the visual signals communicate between different display devices more and more frequently. Therefore, one source visual signal needs to be displayed on different devices. Retargeting algorithms are thereby developed to adapt the same visual signal to different display devices of arbitrary resolutions. The newly encountered distortions will be inevitably introduced. Therefore, the subjective responses of retargeted images are studied [33] [34]. Furthermore, with the development of the eye tracker devices, the visual attention maps of viewing visual signals are recorded during the subjective testing processes [35] [36], which can help to more accurately depict the human visual system (HVS) properties.

Although many benefits are provided by the subjective evaluation process, the lengthy processing time and high cost make it impractical for the visual signal processing and communication. Therefore, accurate objective perceptual quality assessment (PQA) methods are desired and becomes more and more important, which are expected to replace the subjective testing process for visual signal applications. However, many difficulties need to be overcome for deriving an accurate objective PQA [39]. Firstly, the visual signals are of diverse contents, e.g. sports, animations, cartoons, which produce different visual attentions for different viewers. Secondly, the visual signals go through a life cycle from the server/producer side to the receiver/customer side, such as acquisition, compression, transmission, presentation, and so on. During the pipeline, many types of distortions may be introduced. For example, noises can be introduced by the CMOS image sensors during the acquisition process, and blocking and ringing artifacts are brought in during the compression processes. Various noises introduced in different processing stages present great challenges in the design of accurate PQA methods. Thirdly, viewing conditions for the visual experiences are greatly different. For example, the lightness conditions as well as the types of the display devices will seriously affect the visibility of the distortion. Fourthly, the perceptual quality judgements are viewer-dependent. Different viewers have different interests in the visual signals, which make it as an unpredictable factor during the visual signal quality assessment. All the aforementioned aspects make the design of an accurate PQA method extremely difficult and challenging.

In order to handle the problems introduced above, the objective of this thesis is to develop new methodologies for quality evaluation of visual signals, including image and

videos. To that end, in order to evaluate the visual signals degraded by the traditional distortions, such as compression, blurring, white Gaussian noise (GWN), and so on, we focus on the two crucial aspects in perceptual quality metric design, namely, the HVS properties and the visual signal statistical properties. The HVS, as the ultimate receiver of the visual signals, should be considered to develop an effective quality assessment method. However, the HVS is extremely complex and seems impossible to be completely modelled in the near future. Therefore, only the low-level vision of the HVS perception is depicted and modelled, specifically the visual horizontal effect (HE) and the just-noticeable distortion (JND) model [46]. Secondly, visual signals present strong correlations in both the spatial and temporal domains, which can be clearly depicted by the statistics in the pixel domain or transform domain. Some statistics for depicting the visual signals are very sensitive to distortions. In this respect, these statistics can be employed to depict the distortion level. Intuitively, the distortion level explicitly affects the perceptual quality of the visual signal. Therefore, the statistics of the visual signals are expected to help indicate the visual signal perceptual quality. Furthermore, the subjective responses to newly encountered distortions introduced during the retargeting process, are studied through a subjective testing process. Based upon the reliable subjective rating values, PQA methods for the retargeted images can be evaluated and developed.

In the remaining part of this chapter, some background knowledge related to this thesis is introduced. In Section 1.2, the subjective PQAs are briefly described, including different standardized testing methods, and rating scales. In Section 1.3, an overview of the objective PQAs is presented. Based on different approaches, the objective PQAs can be classified into two categories: visual modelling approaches and engineering approaches. The visual modelling approaches usually consider various low-level characteristics of the HVS properties, such as luminance adaptation, contrast sensitivity function (CSF), contrast masking, etc., which are derived from physiological or psychophysical studies. Engineering approaches, on the other hand, are mostly developed based on the assumptions and prior knowledge, *i.e.*, assumptions of features that are closely related to perceptual quality and prior knowledge of the distortion types. Moreover, based on the available information of the reference visual signal, the objective PQAs can also be classified into 3 categories: (1) full reference (FR) quality

assessment which employs the complete information of the reference image, (2) reduced reference (RR) quality assessment which employs only partial information from the reference image, (3) no reference (NR) quality assessment where no information of the reference image is available. In this section, how to evaluate the performance of the PQA is also discussed. The thesis contributions are highlighted in Section 1.4. And the organization of this thesis is given in Section 1.5.

1.2 Subjective Perceptual Quality Assessment

Objective PQAs aim at depicting the HVS perception accurately. Till now, however, no quality metrics can accurately and completely characterize the subjective responses in perceptual quality assessment [2]- [4] [64]. Therefore, subjective testing process is still the only way to verify the model designing, tuning, and optimization. To this end, standardized testing methods and procedures need to be defined and strictly followed, which can ensure reliable results.

Several international standards [5] - [11] are proposed for subjective image/video quality evaluation for different applications. ITU-R BT.1129-2 [9] was proposed for evaluating the perceptual quality of the standard definition (SD) video sequences. ITU-R BT.500 [5] specified by international telecommunication union (ITU) introduces different methodologies for subjective quality assessment of television pictures. ITU-R BT.710 [6] is an extension of ITU-R BT.500 dedicated for high definition (HD) TV. ITU-T P.910 [7] is another standard which defines the standard procedure of digital video quality assessment with transmission rate below 1.5 Mbit/s. ITU-R BT.814-1 [8] is proposed to set the brightness and contrast of the displays. Among these international standards, ITU-R BT.500-11 is most commonly utilized, which provides the viewing conditions, selection of test material, instructions of subjective assessment, presentation of the subjective rating results, and so on. Also the video quality expert group (VQEG) proposed several subjective evaluation procedures to evaluate the performances of different objective quality metrics [13] - [15], which are similar to the standards defined in ITU-R BT.500 and P.910.

A wide variety of basic test methods have been used in television assessment. In practise, however, particular methods should be used to address particular assessment problems. In the following part, we will briefly introduce different standardized testing

methods, as well as the subjective rating scales, which is detailed in [5].

(1) Double stimulus impairment scale (DSIS) method.

This method is employed to measure the robustness of systems. The reference visual signal (unimpaired image/video) and the test stimulus (impaired image/video) should be presented in a pseudo-random sequence, where the reference one is presented before the impaired one. In any case, the same test stimulus should be presented on two successive occasions with the same or different levels of impairment. The range of the impairments should be chosen so that all grades are used by the majority of observers; a grand mean score (averaged overall judgements made in the experiment) close to three should be aimed at. A session should not last more than roughly half an hour, including the explanations and preliminaries. The subjective testing procedure should begin with a few images/videos indicative of the range of impairments. And judgements of these images/videos need not be taken into account in the final results. The five-grade subjective rating scale utilized is shown in Table 1.1. Assessors should use a form which gives the scale very clearly, and has numbered boxes or some other means to record the gradings.

Quality	Impairment
5: Ecellent	5: Imperceptible
4: Good	4: Perceptibly, but not annoying
3: Fair	3: Slightly annoying
2: Poor	2: Annoying
1: Bad	1: Very annoying

Table 1.1*: Five-grade subjective rating scale*

(2) Double-stimulus continuous quality-scale (DSCQS) method.

This method aims at measuring the perceptual quality of stereoscopic image coding. This double-stimulus method is thought to be especially useful when it is not possible to provide test stimulus test conditions that exhibit the full range of quality. The method is cyclic in that the assessor is asked to view a pair of visual signals, each from the same source, but one via the process under examination, and the the other one directly from the source. The qualities of both visual signals need to be subjectively assessed and

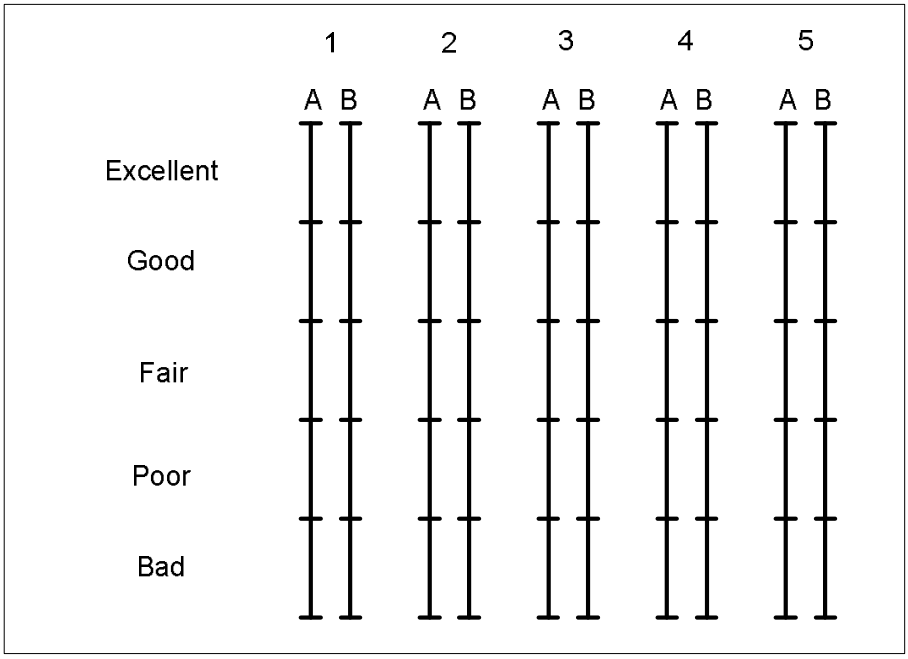


Figure 1.1: *Portion of quality-rating form using continuous scales.*

rated by the viewers. One test session comprises a number of presentations. If there is only one observer, the assessor is allowed to switch between the two visual signals until he or she has the mental measure of the quality associated with each visual signal. Otherwise, if a number of observers are evaluating simultaneously, the pair of visual signals is shown one or more times for an equal length of time to allow the assessor to gain the mental measure of the qualities associated with them.

The method requires the assessment of two versions of each test visual signals. One of each pair of test visual signals is unimpaired while the other presentation might or might not be impairment free. The unimpaired visual signal is included to serve as a reference. However, the observers are not told which one is the reference signal. In the series of tests, the position of the reference visual signal is changed in pseudo-random fashion. The observers are simply asked to assess the overall quality of each presentation by inserting a mark on a vertical scale, as shown in Figure 1.1. The vertical scales are printed in pairs to accommodate the double presentation of each test visual signal. The scales provide a continuous rating system to avoid quantization errors, but they are divided into five equal lengths which correspond to the normal ITU-R five-grade quality scale.

(3) Single stimulus (SS) Method.

A visual signal is presented and the assessor provides an index of the entire presentation. The subjective test session consists of a series of assessment trails. These should be presented in random order, and preferably, in a different random sequence for each observer. A typical assessment trial consists of two displays: a stimulus and a mid-grey post-exposure field. The duration of these displays may vary with viewer task, materials, and the opinions or factors considered. But 10s and 5s are suggested, respectively. The viewer indices have to be collected during display of the post-exposure filed only.

For SS method, the five-grade rating scale as illustrated in Table 1.1 can be employed to indicate the subjective quality. This method yields a distribution of judgements across scale categories for each condition. The way in which responses are analysed depends upon the judgement and information viewed. Also an SS procedure using an 11-grade numerical categorical scale (SSNCS) was studied and compared to graphic and ratio scales. This study, described in ITU-R BT.1082 [12], indicates a clear preference in terms of sensitivity and stability for the SSNCS method when no reference is available. Furthermore, observers can assign a value to each visual signal in non-categorical judgements.

(4) Stimulus-comparison (SC) methods.

In SC methods, tow visual signals are displayed and viewer provides an index of the relation between the given two presentations. The assessment trial can use either one monitor or two well-aligned monitors and generally process as in SS cases. Stimulus-comparison methods assess the relations among conditions more fully when judgements compare all possible pairs of conditions. However, if this requires too large a number of observations, it may be possible to divide the observations among assessors or to use a sample of all possible pairs.

The adjective categorical judgement is employed for the SC methods, where observers assign the relation between members of a pair to one of a set of categories that, typically, are defined in semantic terms. These categories may report the existence of perceptible differences, the existence and direction of perceptible differences, or judgements of extent and direction. The comparison scale recommended by ITU-R is shown in Table 1.2. This method yields a distribution of judgements across scale categories for each condition pair. The way that responses are analysed depends on the judgement

made and the information required. The SC method together with the comparison scale is suitable for evaluating the just-noticeable distortion (JND) performances.

-3	The right one is much worse than the left one
-2	The right one is worse than the left one
-1	The right one is slightly worse than the left one
0	The right one is the same quality as the left one
+1	The right one is slightly better than the left one
+2	The right one is better than the left one
+3	The right one is much better than the left one

Table 1.2*: Comparison scales of SC method*

(5) Simultaneous double stimulus for continuous evaluation (SDSCE) method.

This method targets at measuring the fidelity between two impaired video sequences and comparing different error resilience tools. When the fidelity of the visual signal needs to be evaluated, the reference conditions must be introduced. The method was proposed to motion picture experts group (MPEG) to evaluate the error robustness at very low bit rate. But it can be suitably applied to all those cases where fidelity of visual information affected by time-varying degradation has to be evaluated. The panel of subjects is watching two visual signals in the same time: one is the reference, the other one is the test condition. If the format of the visual signals is SD or smaller, the two signals can be displayed side by side on the same monitor, otherwise two well-aligned monitors should be employed. Subjects are requested to check the differences between the two visual signals and to judge the fidelity of the video information by moving the slider of a handset-voting device. When the fidelity is prefect, the slider should be at the top of the scale range (coded 100), when the fidelity is null, the slider should be at the bottom of the scale (indicated 0). And subjects are aware of which is the reference and they are requested to express their opinion, while they are viewing the visual signals throughout their whole duration.

With these subjective evaluation methods, such as DSIS, DSCQS, SS, SC, and SDSCE, the perceptual quality of the visual signal can be accurately measured. However, as aforementioned subjective visual quality quality assessment suffers from various

drawbacks that limit its applicabilities. (a) It is time-consuming, laborious and expensive, since the resultant subjective opinions are obtained by many observers through repetitive viewing sessions. (b) Incorporation of subjective viewing tests is not feasible for on-line visual signal manipulations, such as visual signal compression, denoising, transmission, and so on. (c) The subjective testing results rely heavily on the viewers' physical conditions, emotional states, personal experience, and the context of preceding display [38] - [40]. As a result, it is necessary to build computational models to predict the perceptual quality of the visual signal in a consistent and objective manner, where the objective visual quality assessment methods are demanded.

1.3 Objective Perceptual Quality Assessment

The simplest and most widely used objective visual quality metric are the mean square error (MSE) and the related signal-to-noise ratio (SNR), and peak signal-to-noise ratio (PSNR). These measurements are appealing for their simple formulations, clear physical meanings, and friendliness for optimization. However, they perform poorly for the perceptual quality predictions of the visual signals [41] [42]. The major reason for the poor performance of MSE or PSNR is that all of the changes in the visual signal is assigned the same importance, regardless of the perceptual properties of the HVS. Objective evaluation of visual signal quality in line with the human perception is a difficult task [43] [44] due to the complex, multi-disciplinary nature of the problem (related to physiology, psychology, vision research, and computer science) and the limited understanding of the mechanisms behind the HVS.

With regards to developing a visual quality assessment method, two different approaches are employed [85] (i) visual modelling approach and (ii) engineering approach. These two approaches and their advantages/disadvantages are discussed as follows, respectively.

1.3.1 Visual Modelling Approach

The visual modelling approaches, as the name implies, are based on the modelling the components of the HVS, which range from the eye to the visual cortex. Although the anatomy of the eye provides us with detailed physiological evidences about the from-end of the HVS (optics, retina, etc.), a thorough understanding of the latter

stages of the visual pathway (visual cortex, etc.) in charge of higher-level perception is still unachievable, which makes the construction of a complete physiological HVS model impossible. As a result, the visual modelling approaches are mostly based on psychophysical studies and only account for low level perception. The physiological and psychophysical factors incorporated into the visual modelling are listed below.

- Colour perception

  The responses of the cones need to be further processed at a higher stage of the HVS for the purpose of signal decorrelation, where the three commonly encountered color channels R, G, and B are highly correlated with each other. Nowadays, many color spaces are employed for representing the visual signal for difference purposes, e.g., CIELAB, YIQ, YUV color spaces, etc. Most of these color spaces share the common characteristics that they treat the luminance and chrominance components of the visual signal separately. However, according to the performance comparison of different color spaces in the visual quality metrics [45], there is no significant performance difference if the chrominance component is discarded, but on the other hand, the computational complexity can be greatly reduced.

- Luminance adaptation

  HVS perception is sensitive to luminance contrast rather than the luminance intensity. Given an image with a uniform background luminance l and a square at the center with a different luminance $l+dl$, if dl is the threshold value at which the central square can be distinguished from the background, then according to Weber's law the ratio of dl divided by l is a constant for a wide range of luminance l. This implies that HVS sensitivity to luminance varies with the local mean luminance value. In other words, the local mean luminance masks the luminance variation: the higher the local mean luminance, the stronger the masking effect is. Therefore, the luminance adaptation factor experimentally forms a U-shape curve. Nowadays, the luminance adaptation factors are employed for constructing the JND models [46].

- Multi-channel decomposition

Compared with the early perceptual quality metrics [47] [48], where only one channel is employed, multi-channel decomposition has been widely used for visual modelling these days. Multi-channel decomposition is justified by the discovery of the spatial frequency selectivity and orientation selectivity of the simple cells in the primary visual cortex. For spatial multi-channel decomposition, most studies suggest that there exist several octave spaced radial frequency channels, each of which is further tuned by orientations with roughly 30 degree spacing [49]. Many other decomposition algorithms serving this purpose exist, e.g., steerable pyramid transform [50], QMF (Quadrature Mirror Filters) transform [51], wavelet transform [52], DCT transform [53], etc. Some of these aim at accurately modeling the decomposition mechanism, while others are used due to their suitability for particular applications, e.g., compression [53]. A detailed comparison of these decomposition algorithms can be found in [54]. For temporal decomposition, it is generally believed that there exist two channels: one low-pass channel, namely sustained channel, and one band-pass channel, namely transient channel. Since most visual detailed information is carried in sustained channel, HVS models employed by some video quality metrics like those in [55] [56] only use a single low pass temporal filter to isolate the sustained channel, while the transient channel is disregarded. Temporal filters can be implemented as either Finite Impulse Response (FIR) filters [55] or Infinite Impulse Response (IIR) filters [57], either before [55] or after spatial decomposition [58].

- Contrast sensitivity function

Contrast sensitivity is the inverse of the contrast threshold - the minimum contrast value for an observer to detect a stimulus. These contrast thresholds are derived from psychophysical experiments using simple stimuli, like sine-wave gratings or Gabor patches. In these experiments, the stimulus is presented to an observer with its contrast increasing gradually. The contrast threshold is determined at the point where the observer can just detect the stimulus. It has been proved by many psychophysical experiments that the HVS's contrast sensitivity depends on the characteristics of the visual stimulus: its spatial frequency, temporal frequency, color, and orientation, etc. Contrast sensitivity function (CSF) can be used to describe these dependences. And CSF is more complex when the influences of

other factors like temporal frequency or color are considered in conjunction with the spatial frequency [59] [60] . As a very important characteristic of HVS, CSF has been incorporated into the development of JND models [46] [53].

- Contrast masking

  Contrast masking effect refers to the visibility threshold elevation of a target signal (the maskee) caused by the presence of a masker signal. It can be further divided into spatial masking and temporal masking. The target and masker stimuli are sine-waves or Gabor patches, during the most spatial masking experiments. The target stimulus is superposed onto the masker stimuli, and contrast threshold of the target stimulus are recorded, together with the masker information, including its contrast, spatial frequency, orientation, phase, etc. Many of these experiments verify that the threshold contrast of the target depends on the masking contrast, and also the other characteristics of the masker. Generally higher masking contrast and larger similarity between the masker and the target in their spatial frequencies, orientations, and phases will lead to higher masking effect, which is known as the contrast masking effect [61] [62]. Compared with spatial masking, temporal masking has received less attention and is of less variety. In most of its implementations in video quality assessment, temporal masking strength is modelled as a function of temporal discontinuity in intensity: the higher the inter-frame difference, the stronger is the temporal masking effect. Particularly, the masking abilities of scene cut have been investigated in many experiments, with both of its forward and backward masking effects identified [63].

- Visual saliency and attention

  It is well known that HVS processes local regions of visual signals with different visual acuities. The artifacts in the attended regions are better perceived than those present in the non-attended areas. It means that the judgement of the observer's will be prejudiced by the contents in the visual salient regions. The perceptual quality of the visual signal will also be significantly affected by severity of the distortions in the attended area. Therefore, the visual saliency and attention map of the corresponding visual signal needs to be considered for deriving an accurate perceptual quality metric.

- Pooling

  In vision system, pooling refers to the process of integrating information of different channels, which is believed to happen at the latter stages of the visual pathway. In visual quality assessment, pooling is used to term the error summation process which combines errors measured in different channels into a quality map or a single quality score. For image quality assessment, most approaches perform error summation across frequency and orientation channels first to produce a 2-D quality map, and then perform it across spaces to obtain a single score indicating the quality of the entire image. For video quality assessment, one more step is performed to combine quality scores for frames into a quality score for the video sequence.

As these factors are intuitively depicted and can somewhat characterize the HVS properties, the visual modelling approaches achieve better results in matching the subjective rating values, compared with MSE, SNR, and PSNR. However, they may suffer from some drawbacks, even though the visual modelling metrics are attractive in theory. The HVS comprises of many complex processes which work in conjunction rather than independently, to produce visual perception. However, the vision modelling based metrics generally utilize results from psychophysical experiments which are typically designed to explore a single dimension of the HVS at a time. In addition, these experiments usually use simple patterns such as spots, bars, and sinusoidal gratings which are much simpler than those occurring in real images. For instance, psychophysical experiments characterize the CSF and masking phenomenon of the HVS by superposing a few simple patterns. Moreover these metrics generally depend on the modelling of the HVS characteristics which are not yet fully understood. Although our knowledge about the HVS has been improving over the years, we are still far from a complete understanding of the HVS and its intricate mechanisms. Furthermore, due to the complex and highly non-linear nature of the HVS, these metrics can be complicated and time-consuming to be used in practice. The complexity of these models usually leads to high computational cost and memory requirement, even for images of a moderate size. In addition, the psychophysical experiments that underlie many error sensitivity models are specifically designed to estimate the threshold at which a stimulus is just barely visible. These measured threshold values are then used to define visual error

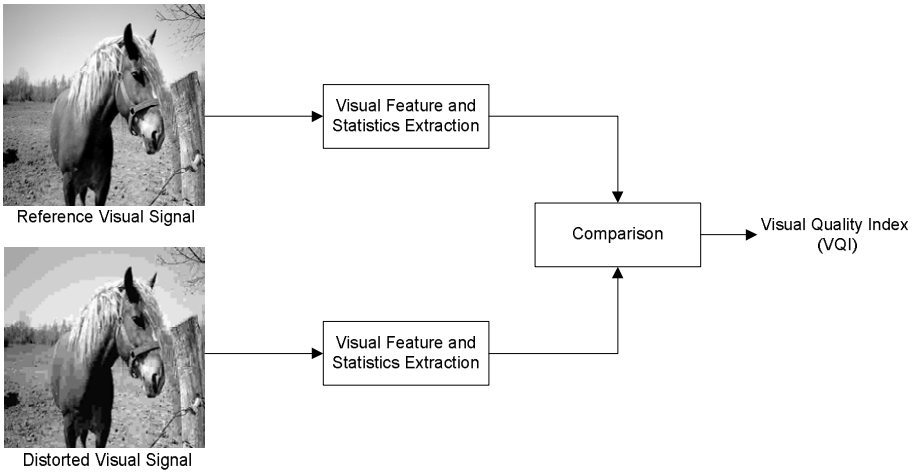


Figure 1.2: *Representative diagram of the engineering approach.*

sensitivity measures, such as the CSF and various masking effects. However, very few psychophysical studies indicate whether such near-threshold models can be generalized to characterize perceptual distortions significantly larger than threshold levels, as is the case in a majority of image processing situations. As it turns out, many of the image quality assessment (IQA) metrics based on vision modelling approach are less effective for suprathreshold distortions [38]- [40]. Owing to these limitations, the second approach namely the engineering approach has gained popularity during recent years and is described next.

1.3.2 Engineering Modelling Approach

To overcome these shortcomings of the vision models, recently many new visual quality metrics were designed by engineering approaches. Instead of founding on accurate experimental data from subjective viewing tests, these engineering-based quality metrics are based on (i) assumptions about visual features that are closely related to visual quality; (ii) prior knowledge about the distortion properties or the statistics of the natural scenes. Since these features and prior knowledge are considered to be higher-level perceptual factors compared with lower-level ones used in the vision model, engineering-based quality metrics are also referred to as top-down quality metrics, and are considered to have the potential to better deal with supra-threshold distortions. In ITU-R BT.1683 [64], four video quality assessment (VQA) methods are recommended after VQEG's FRTV Phase II tests [65], all of which belong to this category. A

representative diagram of the engineering approach is shown in Figure 1.2. It can be observed that the engineering approaches consist of two distinct stages. The first stage is detect and extract the visual features and statistics of the visual signals, which may include image structural elements, such as contours and edges, gradient information, and statistics of specific distortions introduced by a particular process step, *i.e.*, compression, transmission, and so on. The second one is to compare the extracted features and pool their differences together to generate the visual quality index (VQI) for the corresponding distorted visual signal.

Based on the amount of the visual feature or statistics extracted from reference visual signal, PQAs can be classified into three categories: full reference (FR), reduced reference (RR), and no reference (NR). The whole reference visual signal is required to derive the FR metrics, in order to evaluate the perceptual quality of the distorted image, where the reference visual signal is is assumed to be artifact free and of perfect quality. These developed metrics can only be applied to the applications where the reference visual signal is available, such as image compression [66], watermarking [67], and so on. The simplest FR metrics are MSE, SNR, and PSNR, which are widely adopted. However, these PQAs only consider the differences in visual signal pixel level, which are not related to HVS perception properties as aforementioned. Consequently, they are not reliable for evaluating or even controlling the perceptual quality of the distorted image during the processing stages. Nowadays, many FR quality metrics have been developed, among which structure similarity index (SSIM) [68] [69] is the most famous one. SSIM is derived mainly based on the idea of equating the perceived image distortion to the measurement of structural distortion. The metric known as MSVD [71] evaluates the perceptual quality of each image block based on the error in singular values. Another representative FR image quality metric is visual information fidelity (VIF) [72], which is based on the assumption that visual quality is related to the amount of information that the HVS can extract from an image. Briefly, VIF works in the wavelet domain and uses three models to model the original natural image, the distortions, and the HVS, respectively.

Picture quality scale (PQS) [73] is a hybrid image quality metric employing both the HVS model and the engineering design approaches. Among the five distortion factors measured, three of them are obtained basically by using HVS models. Perceptual

factors employed include luminance adaptation, CSF, and texture masking. The other two engineering-based distortion factors measure blockiness and error correlations. To fit in the three processing steps introduced above, in PQS, non-linear mapped luminance values (to account for the luminance adaptation effect) are used as features, and feature comparison is implemented by direct subtraction. These feature differences are further processed by the CSF and by using prior knowledge about the locations of the distortions to produce two distortion maps measuring blockiness and local error correlations, respectively. In the last step, spatial pooling is performed separately on each of the two distortion maps, generating two engineering-based distortion factors. Together with the three HVS-model-based distortion factors, they are de-correlated by singular value decomposition and linearly combined to generate the PQS quality score.

In many real-world applications, we cannot access the reference visual for the quality evaluation, such as image/video denoising, restoration, etc., where only the distorted visual signal is available for analysis. Therefore, the NR PQAs [70]- [79] are thus needed to evaluate and control the perceptual quality of the processed image. Many researchers employ the behaviours of specific distortions for the NR quality assessment, such as the blocking artifact of JPEG coded images, ringing artifact of the JPEG 2000 coded images, and so on. As JPEG 2000 employs the wavelet transform to compress the image, the wavelet statistical model is utilized to capture the compression distortion [74]. Liang *et al.* [75] combined the sharpness, blurring, and ringing measurements together to depict the perceptual quality of the JPEG 2000 coded image. The distribution of the DCT coefficient after quantization is modeled in [76] to predict the PSNR value of the JPEG coded image. Furthermore, Ferzli *et al.* [78] did the psychophysical experiment to test the blurring tolerance ability of the HVS, based on which the just-noticeable blur (JNB) model is developed. These methods employ the behaviors of specific distortions to predict the degradation level. Therefore, if a new distortion is introduced, these methods can hardly evaluate the perceptual quality of the distorted image. In order to compromise between the FR and NR PQAs, RR PQAs are developed. It is expected that the RR methods can effectively evaluate the image perceptual quality based on a limited number of features extracted from the reference image. Only a small number of bits is required for representing the extracted features, which can be efficiently encoded and transmitted for the quality analysis. Consequently, it will be very useful for the

quality monitoring during the image transmission and communication. The image perceptual quality can be easily analysed by referring to the extracted features from the reference image. Therefore, a better quality of user experience can be further provided for the consumers.

RR quality metrics are the tradeoff between FR and NR PQAs. For designing an effective RR quality metric, we need to consider not only its performance but also its RR data rate for representing the extracted features. Firstly, the extracted features should be sensitive to a variety of image distortions and relevant to the HVS perception of the image quality. Secondly, the RR data rates should not be large, as the extracted features need to be embedded or transmitted to the receiver side for the quality analysis. For a larger RR data rate, one may include more information about the reference image. Then a good performance can be obtained. However, it will introduce a heavy burden to the RR feature transmission. The FR PQA can be regarded as an extreme case of RR PQA, with the RR data rate is the whole reference image. For a smaller RR data rate, only a little information of the reference image is available for quality analysis. Therefore, the performance is hard to be ensured. The NR PQA is another extreme case of RR PQA, with no information from the reference image. Therefore, how to balance the RR data rate and the performance is the essential for the RR quality metric development. VQ Model [80] is one of the best proponents of the VQEG FRTV Phase II tests [65]. For a video sequence, VQ Model generates seven distortion factors to measure the perceptual effects of a wide range of impairments, such as blurring, blockiness, jerky motion, noise and error blocks, etc. Viewed conceptually, VQ Model's distortion factors are all calculated in the same steps. Firstly, the video streams are divided into 3D Spatial-Temporal (S-T) sub-regions typically sized by 8 pixel $\times$ 8 lines $\times$ 0.2 second; then feature values are extracted from each of these 3D S-T regions by using statistics (mean, standard deviation, etc.) of the gradients obtained by a 13-coefficient spatial filter, and these feature values are clipped to prevent them from measuring unperceivable distortions; finally these feature values are compared and their differences combined together for quality prediction. Three feature comparison methods used by VQ Model are Euclidean distance, ratio comparison, and log comparison.

1.3.3 Perceptual Subjective Quality Databases

In order to evaluate the performances of the proposed PQAs, many subjective quality databases have been built, such as image databases [25]- [34], video databases [16]- [24]. Most of these databases are built according to the subjective test settings standardized in ITU-R BT.500-11 [5]. In this subsection, we will briefly introduced the major characteristics of the databases, which are employed for the experimental validation in this thesis. In total, four image databases and one video databases are utilized, which are all subjective rated and publicly available.

The LIVE image subjective quality database [25] includes 29 original 24-bits/pixel color images. Totally it consists of 982 images (779 distorted images and 203 reference images). Five types of distortions were introduced to obtain the distorted images: 1) JPEG2000 compression, 2) JPEG compression, 3) white Gaussian noise (WGN), 4) blurring, and 5) Rayleigh-distributed bit stream errors of a JP2K compressed stream or fast fading distortions (FF). Subjective quality scores for each image are available in the form of DMOS. The image resolution is either 768×512 or 512×768.

The IRCCyN/IVC image subjective quality database [26] consists of 10 original color images with a resolution of 512×512 pixels from which 185 color distorted images have been generated, using 4 different processes: JPEG compression, JPEG2000 compression, LAR coding, and blurring. Subjective evaluations have been performed in a normalized room with lighting conditions and display settings adjusted according to ITU-R BT.500-11 [5]. The viewing distance was set to six times the picture's height. A DSIS method, as illustrated in Section 1.2, has been used. Both distorted and original pictures were displayed sequentially.

The MICT subjective quality database [27] contains 182 images of 768 × 512 pixels. 14 were original images (24 bit/pixel RGB) in each group. The rest of the images were JPEG and JPEG2000 coded images (*i.e.*, 84 compressed images for each type of distortion). Six quality scales and six compression ratios were respectively selected for the JPEG and JPEG2000 encoders. Subjective experiments were conducted in a normalized room with low lighting conditions and display settings adjusted according to ITU-R BT.500-11 [5]. The viewing distance was set to four times the picture's height. SS method (illustrated in Section 1.2) together with five-grade rating scales, as shown in Table 1.1, was used during the subjective experiments. The subjects were asked to

provide their opinions on the perceptual quality of the compressed images.

In the A57 database [37], 3 original images of size 512×512 are distorted with 6 types of distortions and 3 contrasts. These result in 54 distorted images (3 images × 6 distortion types × 3 contrasts). The distortion types used are: 1) quantization of the LH subbands of a 5-level DWT of the image using the 9/7 filters, 2) additive WGN, 3) baseline JPEG compression, 4) JPEG2000 compression, 5) JPEG2000 compression with the dynamic contrast-based quantization algorithm in which greater quantization is applied to the fine spatial scales relative to the coarse scales in an attempt to preserve global precedence, and 6) blurring. The subjective scores have been made available in the form of DMOS.

The video database we used is the LIVE video subjective quality database [17] [18]. It contains 150 distorted videos (generated from 10 uncompressed reference videos of natural scenes) with spatial resolution being 768×432. The frame rate is either 25fps or 50fps. The distorted videos have been obtained by using four distortion processes: (a) simulated transmission of H.264 compressed bit streams through error-prone wireless networks, (b) through error-prone internet protocol (IP) networks, (c) H.264 compression, and (d) MPEG-2 compression. Each video was assessed by 38 human subjects and the subjective scores have been made available as DMOS. Table 1.3 lists the major characteristics of the visual subjective quality metric used for validation in this thesis.

	Image				Video
Databases	LIVE	IRCCyN/IVC	MICT	A57	LIVE
NOR	29	10	14	3	10
NDT	5	4	2	6	4
NOD	779	185	168	54	150
RES	768×512 or 512×768	512×512	512×512	512×512	768×432
SSF	DMOS	MOS	MOS	DMOS	DMOS
RNG	(0-100)	(1-5)	(1-5)	(0-1)	(0-100)

Table 1.3: *Major characteristics of the visual subjective quality databases. NOR denotes the No. of the reference images/videos; NDT indicates the No. of the distortion types; NOD is the No. of the distorted images/videos; RES means the resolution of the images/videos; SSF indicates the subjective score format; and RNG indicates the range of the subjective scores of each database*

1.3.4 Performance Evaluation

To evaluate the performance of a visual quality metric, visual subjective quality databases are employed, which are briefly introduced in Section 1.3.3. For each distorted visual signal, *i.e.*, image or video, MOS or DMOS is assigned, which is obtained from subjective viewing tests which follow the standardized procedures as introduced in Section 1.2. To evaluate the predictive performance of a visual quality metric, these subjective scores are used as the ground truths to be compared against the metric's quality predictions (objective quality scores). High correlation between the subjective scores and the objective scores indicates a good performance of the visual quality metric.

Taking into account the non-linearity of the subjective scores introduced during the subjective tests, it is customary to perform a non-linear mapping on the objective scores before the correlation measurement. Following the existing work [72] [81], the following non-linear mapping function is used to map x_j to V_j:

$$V_j = \beta_1 \times \left(0.5 - \frac{1}{1 + exp(\beta_2 \times (x_j - \beta_3))}\right) + \beta_4 \times x_j + \beta_5 \tag{1.1}$$

where x_j represents the objective quality score of the j-th distorted visual signal obtained by the corresponding PQA, and V_j indicates the non-linearly mapped score. The fitting parameters $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ are determined by minimizing the sum of squared differences between the mapped objective scores V_j and the subjective scores, *i.e.*, MOS and DMOS values.

After the non-linear mapping, several performance measurements can be applied, such as the Linear Correlation Coefficient (LCC), the Root Mean Squared Error (RMSE), the Spearman Rank-Order Correlation Coefficients (SROCC), the Outlier Ratio (OR), etc. The mapped scores V_j and the subjective scores serve as their inputs. LCC between two data sets, X and Y, is defined as:

$$LCC(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1.2}$$

measures the correlation. x_i and y_i are the sample value, while $\overline{x}$ and $\overline{y}$ are the corresponding mean value. SROCC assesses how well the metric predicts the ordering of the distorted images, and can be defined as the LCC of the ranks of X and Y, which

is defined as:

$$SROCC(X,Y) = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1.3}$$

where d_i is the difference between the i-th image's rank in subjective and the perceptual quality index. n denotes the total number of samples. RMSE between X and Y is calculated during the fitting process given by:

$$RMSE(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1.4}$$

OR is defined as:

$$OR = \frac{n_{outlier}}{n} \tag{1.5}$$

where $n_{outlier}$ is the number of predictions outside two standard deviations[1] of the subjective scores, and n is also the total number of samples. As can be observed from their definitions, larger LCC/SROCC or smaller RMSE/OR indicates better performance of the visual quality metric. The detailed information and comparison between these performance measurements can be found in [65] [81] [82].

1.4 Thesis Contributions

As aforementioned, according to the availability of the reference signal, PQAs can be categorized into FR, RR, and NR methods. My work in this thesis focuses on the FR and RR PQAs, as well as their corresponding applications. Moreover, perceptual quality of the retargeted images are studied. The key contributions of this thesis are summarized into three parts, which are briefly introduced as follows.

(I) FR quality assessment. The HVS properties are firstly studied, specifically, the visual horizontal effect (HE) to model the orientation sensitivity of the HVS perception, the just-noticeable distortion (JND) over adaptive block-size transform (ABT) to depict the visual tolerance property, the visual saliency to characterize the HVS attention property to the visual signals. The HE is modelled by a simple cubic polynomial function, which can be easily incorporated with the recently developed IQAs, such as SSIM, VIF, and so on. Inspired by HVS HE property,

[1]The standard deviation indicates the variation of individual subjective ratings around the mean value.

a new image compression algorithm is developed by the super-resolution directed down-sampling (SRDDS). ABT-based JND is originally studied for capturing different HVS properties over different block-size discrete cosine transforms (DCT). Simple PQAs for image and video are developed, which are utilized for perceptual video coding. And motion trajectory is firstly considered to more accurately depict the HVS saliency property, which are then utilized to improve the performances of different PQAs.

(II) RR quality assessment. Statistics of the visual signals are studied, specifically the spatial statistics of image, and the spatio-temporal statistics of video sequences. For spatial statistics, the block-based DCT coefficients are firstly grouped into reorganized DCT (RDCT) subbands. The DCT statistics are characterized in the RDCT domain. The coefficient distribution in each domain are depicted by the generalized Gaussian density (GGD) function. The frequency variation is captured by the frequency ratio descriptor (FRD). The relationship between different RDCT subbands are measured by mutual information (MI). These statistics are believed to be sensitive to the distortion introduced in the spatial domain. As a result, they are sensitive to the visual signal degradation level and reliable for depicting the perceptual quality. As for the spatio-temporal statistics, the temporal relationship between adjacent frames needs to be characterized besides the spatial FRD. The frame relationship is captured by the histogram of the difference image, which is also modelled by GGD. With the combination of spatial FRD and temporal GGD, the RR video quality assessment (VQA) for video sequences is developed, which is not only effective but also efficient. The RR features extracted from both spatial and temporal aspects can be easily coded and transmitted to monitor the video quality during transmission process.

(III) Retargeted image quality assessment. The perceptual quality of the retargeted image is studied through a large-scale subjective study of image retargeting quality, where an image regargeting subjective quality database is constructed. Each retargeted image in the database is viewed and subjectively rated by numbers of viewers to generate the final MOS value for indicating its true perceptual quality. The built database is analysed from the perspectives of the retargeting

scale, the retargeting method, and the source image content. Moreover, several publicly available quality metrics for retargeted images are evaluated on the built database. The discussion on how to develop an effective quality metric for retargeted images are provided through a specific designed subjective testing process.

1.5 Organization of the Thesis

This thesis has been divided into 7 chapters as outlined as follows. Chapter 1 (this chapter) gives a brief introduction about the thesis, including the motivation and objectives, subjective and objective quality assessment methodologies, thesis contributions, and organization.

Chapter 2 discusses FR image quality assessment, where the HE and saliency properties of the HVS are depicted and modelled. Inspired by the developed FR IQA, a novel image compression algorithm is proposed via adaptive block-based SRDDS.

Chapter 3 investigates FR PQA, where the JND property over ABT is modelled. ABT-based JND is utilized to develop a simple quality metric, which has been proved to be effective over the visual subjective quality databases. Finally, the developed metric is used for guiding the perceptual video coding. Also in this chapter, the visual saliency for the video sequences is studied by considering the motion trajectory.

Chapter 4 describes our newly developed RR IQA method, where the coefficeint statistic of the RDCT subband is modelled and employed for quality assessment. Furthermore, the MI between different RDCT subbands, and the FRD of the whole image are further employed to improve the performance of the RR image quality metric.

Chapter 5 focuses on developing an RR VQA for the compressed video sequences. The spatio-temporal statistics are employed to depict the degradation level of the compressed video sequence, which are proved to be highly related to the perceptual quality of the video sequences.

In Chapter 6, an image retargeting subjective quality database is built, through a large-scale subjective testing process. The subjective responses on the retargeted image qualities are studied. And several quality metrics are employed to evaluate the retargeted image quality. New directions for deriving more accurate quality metric are further discussed.

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

The Appendix provides the detailed information of the constructed image retargeting subjective quality database.

Part I

Full Reference Quality Assessment

Chapter 2

Full Reference Image Quality Assessment

2.1 Visual Horizontal Effect for Image Quality Assessment

2.1.1 Introduction

Full reference (FR) image quality metric plays a fundamental role for many image processing applications, such as compression, watermarking, and etc. As aforementioned, MSE and the related PSNR are the most wildly employed quality metric for evaluating the perceptual quality of the visual signals. However, they do not correlate well with the HVS perception, because they just focus on the pixel value differences but ignore the image content and human perception property [41] [68]. In order to handle this problem, many image quality metrics have been proposed, which attempt to characterize the features that HVS may associate with loss of quality, such as blurring, blocking, and so on. The IQAs that embody this approach include structure similarity index (SSIM) [68] [69], and visual information fidelity (VIF) [72]. SSIM is derived by capturing the information loss of image structures, while VIF employs the mutual information between the original and test image to evaluate the image quality. In [83] , it has been demonstrated that SSIM and VIF have similar performances. And SSIM treats different oriented distortions and different located distortions equally. However, as the HVS perceives images with local varying saliencies [84], the pooling HVS feature [85] needs to be considered to evaluate the image quality. Also, the HVS horizontal effect (HE) property [86]- [89] of natural scenes has been researched for modeling the visual sensitivities of different distortions over image contents with different orientations, in comparison with the HVS oblique effect property for the simple patterns, such as isolated gratings [87]. Therefore, the HVS HE property needs to be taken into account when evaluating the image quality. In this chapter, the HVS properties over structural distortions are considered to improve the IQA performance. Firstly, SSIM is employed

to obtain the structural distortion map. Secondly, the distortion map is refined by the HVS orientation sensitivity modeled by the HE. Finally, the image quality index is obtained by a saliency pooling strategy over the distortion map.

The reminder of this chapter is organized as follows. Section 2.1.2 introduces the proposed FR IQA, including the visual HE modeling and saliency pooling. Its performance is evaluated and compared with other representative FR IQAs in Section 2.1.3. A summary is given in Section 2.1.4.

2.1.2 Proposed Image Quality Assessment Framework

Structure Similarity Index

SSIM is based on the assumption that the HVS is highly adapted to extract structural information from the viewing field. Three types of similarity together constitute the SSIM, which are luminance similarity $l(\mathbf{I}(i,j),\mathbf{T}(i,j))$, contrast similarity $c(\mathbf{I}(i,j),\mathbf{T}(i,j))$, and structure similarity $s(\mathbf{I}(i,j),\mathbf{T}(i,j))$:

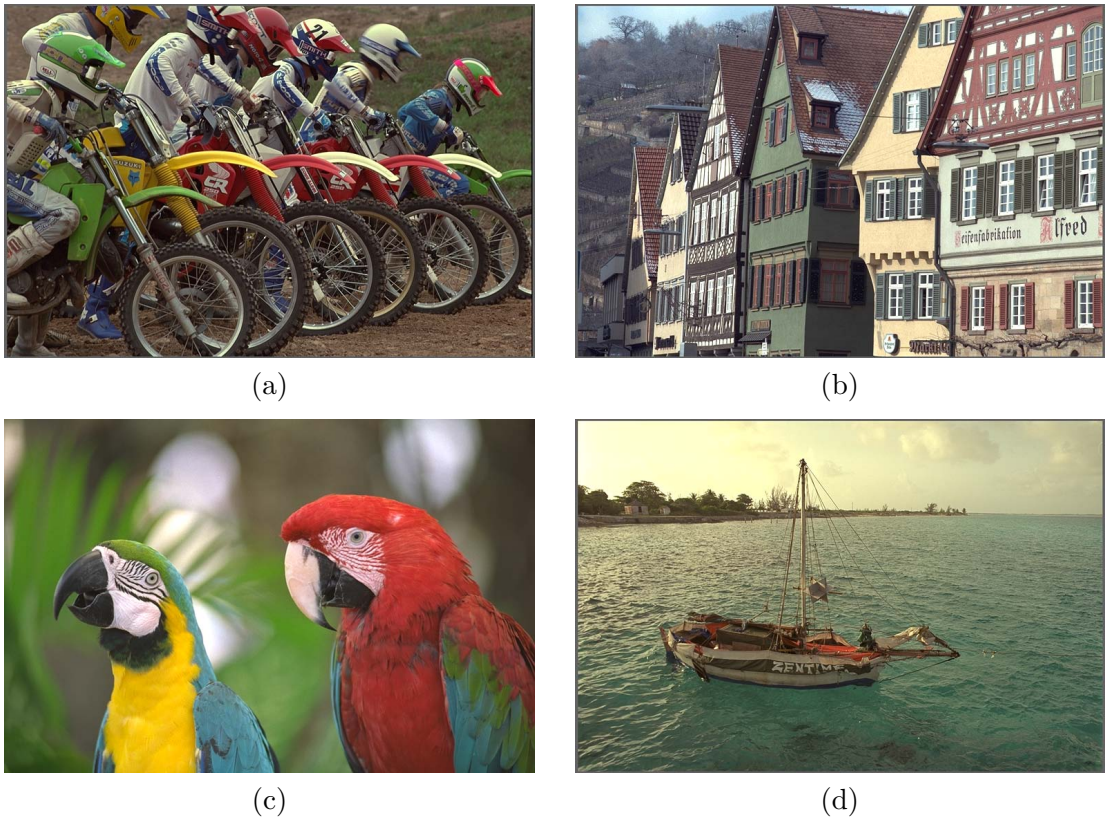
$$\mathbf{SD}(i,j) = [l(\mathbf{I}(i,j),\mathbf{T}(i,j))]^{\alpha} \cdot [c(\mathbf{I}(i,j),\mathbf{T}(i,j))]^{\beta} \cdot [s(\mathbf{I}(i,j),\mathbf{T}(i,j))]^{\gamma} \tag{2.1}$$

where $\mathbf{I}$ and $\mathbf{T}$ denotes the original and distorted images, respectively. (i,j) is the pixel location. $\mathbf{SD}$ is the obtained structural distortion map after performing SSIM. $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are parameters used to adjust the relative importance of the three components. The three components of Eq. (2.1) are relatively independent of each other. In other words, the value change of one component does not necessarily mean that the value of the other components must change accordingly. This is one of the good properties of SSIM, which makes the use of α, β and γ to adjust the importance of the three components reasonable. The formula for $l(\mathbf{I}(i,j),\mathbf{T}(i,j))$, $c(\mathbf{I}(i,j),\mathbf{T}(i,j))$, and $s(\mathbf{I}(i,j),\mathbf{T}(i,j))$ are defined as follows:

$$l(\mathbf{I}(i,j),\mathbf{T}(i,j)) = \frac{2\mu_{I(i,j)}\mu_{T(i,j)} + C_1}{\mu_{I(i,j)}^2 + \mu_{T(i,j)}^2 + C_1} \tag{2.2}$$

$$c(\mathbf{I}(i,j),\mathbf{T}(i,j)) = \frac{2\sigma_{I(i,j)}\sigma_{T(i,j)} + C_2}{\sigma_{I(i,j)}^2 + \sigma_{T(i,j)}^2 + C_2} \tag{2.3}$$

$$s(\mathbf{I}(i,j),\mathbf{T}(i,j)) = \frac{\sigma_{I(i,j)T(i,j)} + C_3}{\sigma_{I(i,j)}\sigma_{T(i,j)} + C_3} \tag{2.4}$$



(a) (b)

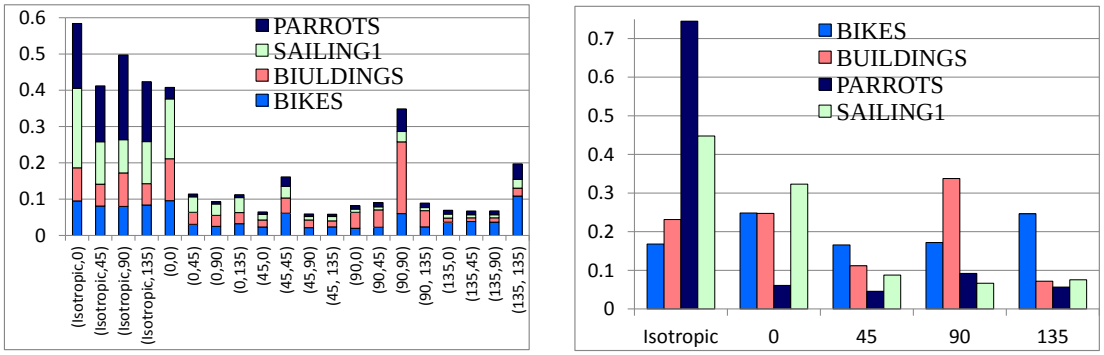
(c) (d)

Figure 2.1: *Four reference images from the LIVE image database [25] for training the visual HE model. (a): bikes; (b): buildings; (c): parrots; (d): sailing1.*

where $\mu_{I(i,j)}$ and $\mu_{T(i,j)}$ are the block means centered at (i, j) of the reference and distorted image, respectively; $\sigma_{I(i,j)}$ and $\sigma_{T(i,j)}$ are the variances of the blocks; $\sigma_{I(i,j)T(i,j)}$ is the covariance of the two blocks describing their structure similarity; C_1, C_2, and C_3 are small constants to avoid instability when the denominator is very close to zero. The classic SSIM [68] [69], or namely the mean SSIM, takes the average of the quality map as the overall score to predict the image visual quality.

Visual Horizontal Effect Modeling

According to Hansen *et al.*'s researches on human vision, [86]- [89], the oblique content is perceived to be the best, whereas the horizontal content is the worst for natural images; also the oblique stimuli are perceived to be the best in naturalistic broad-band stimuli. The phenomenon is known as HE. Hence, we want to model the visual HE sensitivity, which the HVS may associate with the image quality.



(a) (b)

Figure 2.2: *Orientation information of the reference image and their distorted version. (a): content orientation distribution (***x***-axis: content orientation;* ***y****-axis: pixel number probability). (b): content and stimulus orientation joint distribution (***x***-axis: (content orientation, stimulus orientation) pair;* ***y****-axis: pixel number probability).*

In order to model the visual HE sensitivity, first we need to obtain the orientation and energy information for both the content and stimulus. In our approach, the reference image is regarded as the original content, while the difference between the test and reference images is denoted as the stimulus superposed onto the content. As we all know, the kernels of Gabor filters are similar to the 2D receptive field profiles of the mammal cortical simple cells, which exhibit desirable characteristics of spatial locality and orientation selectivity [90]. Therefore, different oriented Gabor filters are employed to filter the original content and stimulus to generate different oriented responses. According to the maximum response, the orientation and energy are determined for depicting the local features of the visual inputs. However, in some local smooth regions, all the filtering responses may appear very small, which are regarded as isotropic for its weak influences over all the orientations.

Four representative images and their distorted versions from the LIVE image database [25] are selected to train the visual HE sensitivities over the image structural distortions. The four original images are illustrated in Figure 2.1. The orientation information of the images and their distorted versions are shown in Figure 2.2. It can be observed that certain oriented contents dominate each selected image, such as: the isotropic contents dominate PARROTS; 135 degree and horizontal contents dominate BIKES and so on. Moreover, we can see that different oriented stimuli are superposed onto different oriented contents with different probabilities. Therefore, we can employ the four representative images (with different dominant oriented contents) and their distorted

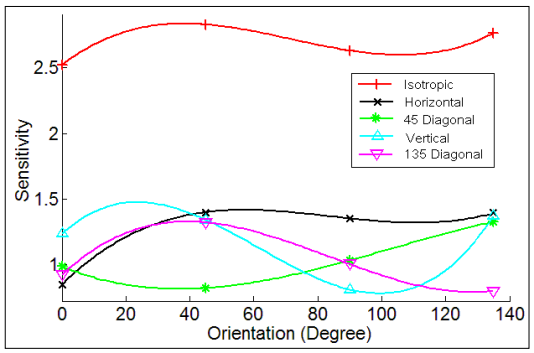


Figure 2.3: *HE sensitivity values of different orientated stimuli over different content bias (each color represents a biased content, and the horizontal axis indicates the stimulus orientation).*

versions (with different oriented stimuli) to train the visual HE sensitivities. During the training process, the following three aspects should be considered.

(I) The content orientation is isotropic. The stimuli presented in these regions are easy for HVS perception, which is modeled by contrast masking in JND models [46]. Therefore, HVS is highly sensitive to this type of content.

(II) The stimulus orientation is the same as the content orientation. It can be viewed as a signal enhancement rather than distortion. Therefore, the lower HVS sensitivity is expected and the distortion is difficult to detect.

(III) The stimulus orientation is perpendicular to the content orientation. The HVS is extremely sensitive and the distortion is very easy to perceive.

The initial HE sensitivity values in [87] are first slightly modified (increased or decreased) by referring to the afore-mentioned three aspects. Based on the structural distortion map SD, if the HE refined SSIM values correlate better with the subjective Differential Mean Opinion Score (DMOS) values, which are provided by the database [25], the HE sensitivity values are tuned by following the same direction. Otherwise, the HE sensitivity values are tuned by following the opposite direction. After several iterations, the optimized HE sensitivity values are obtained. The optimized HE sensitivity values of 4 dominant oriented stimuli over 5 prevailing biased contents are indicated by the spots in Figure 2.3. Based on these sensitivity values, the cubic polynomial functions are fitted to model the sensitivities of oriented stimuli over the same content, illustrated by the curves in Figure 2.3. For the isotropic biased content, the visual sensitivity values are much larger than the other biased contents, which match the HVS contrast

masking properties. For the isotropic and horizontal biased contents, the visual sensitivity values of oblique orientations (45 and 135 degree) are higher than that of the vertical direction, while the horizontal sensitivity value appears the smallest, which matches the experimental results of HE. As for the 45 and 135 degree biased contents, sensitivity values of the orientations around its perpendicular direction are the largest, whereas the sensitivity value of the same orientation appears to be the smallest, which matches the aforementioned aspects. For the vertical biased content, the largest sensitivity value appears around 45 degree according to the HVS HE property and around 0 degree for the perpendicular property. Therefore, the HVS appears to be the most sensitive between 0 and 45 degrees by considering the HVS properties together. The cubic polynomial function for depicting the visual HE sensitivities of orientated stimuli over different oriented contents is expressed as:

$$\mathbf{S}_{HE} = \varphi(\theta_I, \theta_S) = a_{\theta_I}\theta_S^3 + b_{\theta_I}\theta_S^2 + c_{\theta_I}\theta_S + d_{\theta_I} \tag{2.5}$$

where φ is the HE sensitivity function illustrated in Figure 2.3, θ_I and θ_S denote the orientation information of the content and stimulus, respectively, which are determined by the maximum responses of the oriented Gabor filters. a_{θ_I}, b_{θ_I}, c_{θ_I}, and d_{θ_I} are the parameters which relate to the content orientation θ_I. Furthermore, the higher the stimulus energy, the worse is the visual quality of the test image. Therefore, a relationship between stimulus energy and image perceptual quality should be considered. A stimulus energy adaptation factor α_{SE} is used to refine the structural distortion value, which is defined as:

$$\alpha_{SE} = p_1 \cdot erf(p_2 \cdot E_S(i,j) + p_3) + p_4 \tag{2.6}$$

where E_S is the stimulus energy obtained from the Gabor filtering results, erf is the error function, p_1=-0.175, p_2=0.35, p_3=-2.5, and p_4=0.825 are set empirically for adjusting the stimulus energy adaptation factor. Then the refined structural distortion map SM_r is obtained by:

$$SM_r(i,j) = \frac{SD(i,j) \cdot \alpha_{SE}(i,j)}{S_{HE}(\theta_I(i,j), \theta_S(i,j))} \tag{2.7}$$

Moreover, as we have mentioned before, when all the responses of Gabor filtering appear very small, the regions should be regarded as isotropic. It means that the signal

has no inclined orientations. For the stimulus, it means that the distortion obtained is spread over all the orientations. As the stimulus energy is very small, it will have little effect on the image perceptual quality, which can be modeled by JND [46]. In this case, the HE sensitivity and stimulus energy adaptation factor should not be taken into consideration. Therefore, a signal-dependent JND model for the stimulus should be considered by neglecting the influence of the invisible distortion, the magnitude of which is smaller than a threshold Thr. However, according to our experiments, the performance will not be obviously affected as the threshold varies. Thereby, Thr is simply set as 2.2.

Saliency Pooling Strategy

As HVS processes local regions of images with different visual acuities, artifacts that are present in the attended regions are better perceived than those present in the non-attended areas, which means that the observer's assessment of image quality is prejudiced by the perceived structural distortions in salient regions. Therefore, a relative measure of the importance of different regions, indicated by a saliency map, plays an important role in evaluating the image quality. In this study, we employ the spectral residual model [84] to detect the saliency.

Given an image $\mathbf{I}$, Fourier Transform (FT) ξ is firstly applied to obtain the amplitude spectrum $A(f)$ and phase spectrum $P(f)$. The log-spectrum representation of an image is defined as:

$$L(f) = log(A(f)) \tag{2.8}$$

The spectral residual $R(f)$ can be generated based on $L(f)$ according to:

$$R(f) = L(f) - L_a(f) \tag{2.9}$$

where $L_a(f)$ denotes the averaged spectrum, which is derived by convolving the log-spectrum $L(f)$ with an averaging filter. And it is claimed that the spectral residual contains some important information of an image related to the HVS perception [84]. The primary non-trivial part of the scene is constructed by inverse FT ξ^{-1}, which could be interpreted as the unexpected portion of the image. The unexpected portion represents the saliency map SA_M in spatial domain, which indicates the different visual

importances of different locations:

$$SA_M = |\xi^{-1}(exp(R(f) + jP(f)))|^2 \tag{2.10}$$

Based on SA_M, a saliency pooling strategy is proposed to generate the visual quality index (VQI) for evaluating the image perceptual quality:

$$VQI = \frac{\sum SA_M \cdot SM_r}{\sum SA_M} \tag{2.11}$$

2.1.3 Experimental Results

The performance of the VQI is compared with other methods, *i.e.*, PSNR, SSIM [68] [69], and VIF [72]. The IQA methods are evaluated on the LIVE [25] and A57 databases [37], which comprise the most prevailing distortions. The distorted images, excluding the ones generated from the 4 training images, are used for evaluating the IQA performances. The detailed information about the image database can be referred to Section 1.3.3. Also as introduced in Section 1.3.4, three statistical measurements are employed to evaluate the corresponding performances, specifically the LCC, SROCC, and RMSE. detailed information of each measurement can be found in Section 1.3.4.

	LIVE			A57		
	LCC	SROCC	RMSE	LCC	SROCC	RMSE
PSNR	0.891	0.897	12.425	0.644	0.570	0.192
SSIM	0.914	0.922	11.060	0.415	0.407	0.224
LDW-SSIM	0.915	0.919	11.051	0.545	0.495	0.206
ICW-SSIM	0.936	0.942	9.641	0.518	0.455	0.210
SMW-SSIM	0.947	0.953	8.769	0.607	0.557	0.195
VIF	0.961	0.966	7.523	0.614	0.622	0.194
VQI	**0.966**	**0.971**	**7.057**	**0.848**	**0.857**	**0.130**

Table 2.1: *Performance Comparisons of different IQAs on the LIVE and A57 image subjective quality databases*

We compare the performance of VQI with PSNR, SSIM [68] [69], LDW-SSIM (local distortion weighted SSIM) [91], ICW-SSIM (information content weighted SSIM) [92], SMW-SSIM (smooth region weighted SSIM) [93], and VIF [72]. The results are listed in Table 2.1. The performance of our proposed scheme outperforms the other IQAs on the provided two databases with larger SROCC and LCC, and smaller RMSE, which

means that our method demonstrates better performance across a wide range of image distortions. The reason is that the SSIM methods just employ different weights for different locations of the image. However, they do not account for the orientation sensitivity and saliency property of HVS. VIF employ the steerable pyramid to decompose the test image, which extracts the image features at different scales and different orientations. In this way, HVS orientation and saliency properties are included. That is why it can outperform the other IQAs. However, our method can more accurately model the HVS orientation and saliency sensitivities over structural distortions, which outperforms VIF. The scatter-plots of different IQAs are shown in Figure 2.4. It can be observed that the results of our proposed method scatter more closely around the fitted line than other IQAs, which indicates a better performance. Furthermore, it can be observed that VIF performs well on LIVE, but poorly on A57 database. The reason may be that the distortion model embodied in VIF cannot efficiently simulate the two new distortion types in A57, which are (a) quantization of the LH subbands of the image with equal distortion contrast at each scale; (b) JPEG 2000 with dynamic contrast-based quantization compression [37]. However, as the proposed method models the HE and saliency properties of the HVS, it can efficiently capture the distortions in the image which are sensitive to the HVS, no matter what the distortion type is. That is why the proposed metric performs well on both the two databases.

Moreover, we demonstrate the efficiency of each phase (*i.e.*, HE sensitivity and saliency pooling) of our proposed scheme individually by evaluating its performance on the LIVE database. The results are illustrated in Table 2.2. Both the strategies improve the IQA performance. However, the saliency pooling strategy performs better than the HE sensitivity. Intuitively, the results are in accordance with the human perception of a visual input. While perceiving an image, we mainly focus on its interesting or salient portion. If the part appears really interesting and attractive, we will examine it more carefully. Therefore, the saliency pooling is important for image quality assessment. However, the visual HE sensitivity appears to play a lesser but nevertheless an important role in image quality assessment.

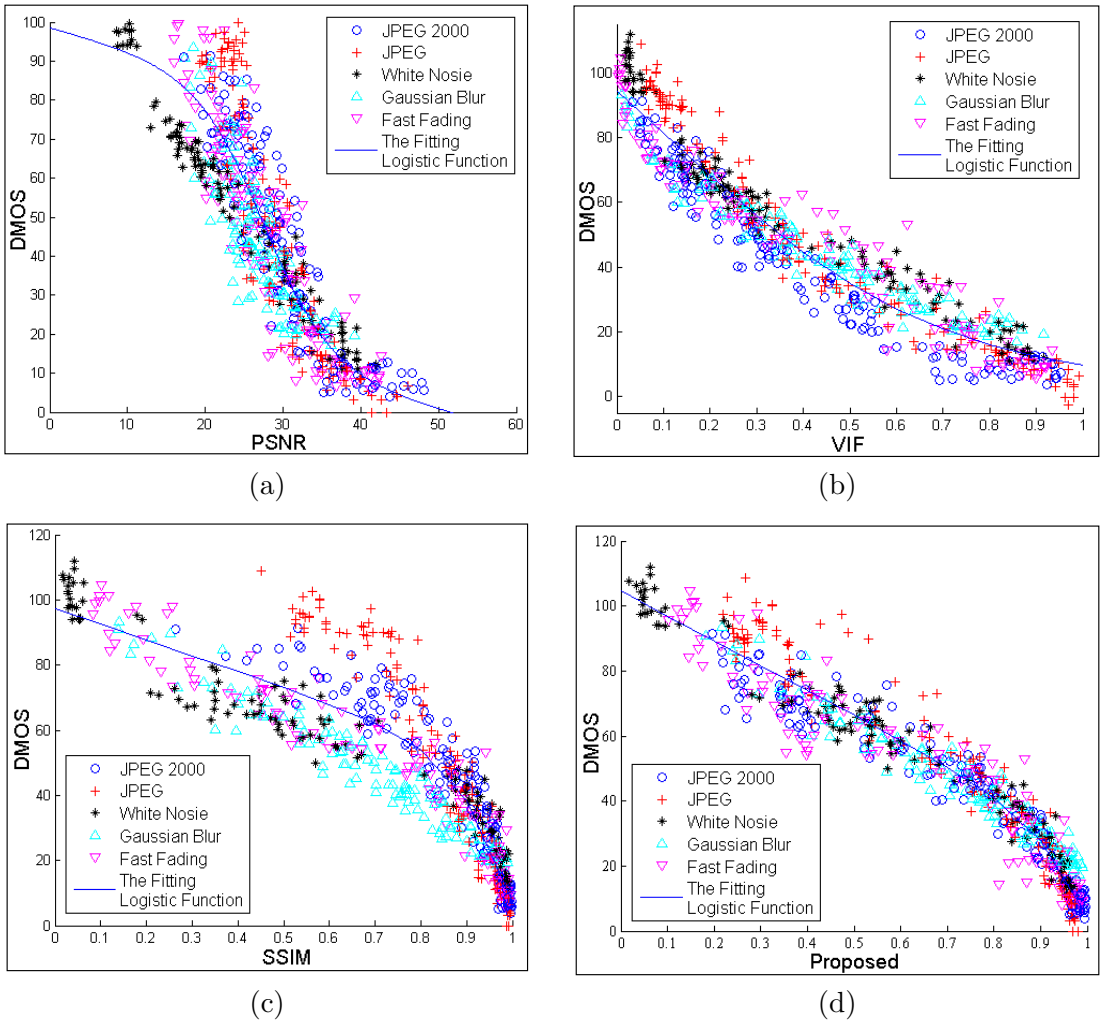


Figure 2.4: *Scatter plots of the DMOS values versus model predictions on the LIVE database. Each sample point represents one test image. (a): PSNR; (b): VIF; (c): SSIM; (d): the proposed method).*

	LCC	SROCC	RMSE
HE Senstivity	0.9331	0.9405	9.820
Saliency Pooling	0.9443	0.9495	8.990

Table 2.2: *Performance of each phase of the proposed scheme on the LIVE image subjective quality database*

2.1.4 Conclusion

In this section, an image quality assessment method is proposed by considering the visual HE sensitivity and saliency properties. The SSIM structural distortion map is refined by the visual HE model. The image quality index is generated by saliently pooling on the refined structural distortion map. Experimental results demonstrate

that the proposed scheme outperforms the other IQAs.

2.2 Image Compression via Adaptive Block-Based Super-Resolution Directed Down-Sampling

2.2.1 Introduction

With the development of the imaging technology, more and more images with high qualities and large spatial resolutions are provided to satisfy people's visual experiences. However, it issues a great challenge to image transmission and storage. Therefore, a more efficient image compression scheme is highly desired, which can ensure a higher image quality with a smaller number of bits for image representation.

Based on the fact that most images can be obtained via interpolation from sparse pixel data yielded by a signal-sensor camera [94], and natural images exhibit high spatial correlations between neighboring pixels [95], many interpolation-based image coding methods [96]- [100] have been proposed. In [98], 2×2 average operator is employed for decimation before JPEG compression. A replication filter and a Gaussian filter are used for restoring the image from the decimated one. The theoretical down-sampling model was studied and compared in [99]. Tsaig *et al.* [100] proposed to code the filter parameters as the side information for better reconstruction at the decoder side. In [96] [97], the authors suggest coding the down-sampled low-resolution image during encoding and recovering the high frequency components during decoding by interpolating the compact image representation generated by sparse sampling in the spatial domain. Although the predominated smooth regions of an image can be satisfactorily recovered by interpolation, the reconstruction of high frequency components of the edge and texture regions still remains a great challenge. In order to overcome the problem, Wu *et al.* [101] employed the piecewise autoregressive model to handle the large phase errors during the interpolation of the image edges. However, there is a heavy computation burden at the decoder side due to the optimal block estimation problem driven by the autoregressive model. Moreover, Lin *et al.* [95] proposed a new image coding method based on the adaptive decision of appropriate down-sampling directions/ratios and quantization steps to achieve higher coding quality. The method tries to avoid down-sampling a macroblock (MB) along the direction of high spatial variations, which signals the existence of edges and other image features with great impact on the perceptual visual

quality. In [95], the down-sampled pixels are obtained by averaging the neighboring pixels of the original resolution image. Although it can somewhat reduce the aliasing artifacts introduced by direct sampling, the blurring artifacts will be introduced. Also as the down-sampling process is independent of the following super-resolution process, the reconstruction errors between the original and the restored MB cannot be ensured to be the smallest. More recently, the JPEG2000 [102] and H.264/AVC [103] have been developed for achieving higher compression performances for images.

In order to tackle the aforementioned problems and inspired by the proposed FR IQA in Section 2.1, we propose a novel perceptual image coding scheme via adaptive block-based super-resolution directed down-sampling (SRDDS). For each MB of a given image, whether down-sampling or not depends on the contents of the visual signal itself, which will be determined by the rate distortion optimization (RDO) process [104]. And the joint method of down-sampling and super-resolution is proposed to minimize the reconstruction errors between the original and the restored MB inferred by the super-resolution method from the down-sampled block. At the decoder side, the super-resolution method performed in DCT domain is employed to recover the full-resolution MB for its simplicity. This section is organized as follows. Section 2.2.2 will introduce the proposed perceptual image compression framework, as well as the super-resolution method and super-resolution directed down-sampling process. Experimental results in Section 2.2.3 will demonstrate the coding efficiency of the proposed method. Finally, Section 2.2.4 concludes the part of this work.

2.2.2 The Proposed Image Compression Framework

The framework of our proposed method is illustrated in Figure 2.5. For each 16×16 MB of a given image, two candidate coding modes are provided. One is the traditional JPEG coding mode. The MB is divided into four 8×8 sub-blocks, each of which is processed by transformation, quantization, de-quantization, and the inverse transformation. The other one is the proposed super-resolution directed down-sampling (SRDDS) mode. Firstly, an 8×8 low-resolution sub-block is obtained by down-sampling the full-resolution 16×16 MB according to the proposed SRDDS. Secondly, the 8×8 sub-block is transformed and quantized (the quantization parameter (QP) parameter

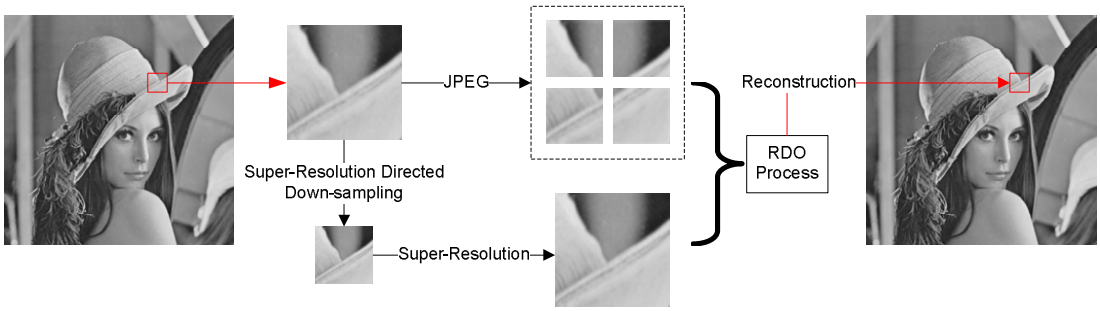


Figure 2.5: *Proposed image compression framework*

is set as half of the one used in the JPEG mode). Then after de-quantization, the corresponding super-resolution method processed in DCT domain is performed together with the inverse transformation. Finally, the RDO process will determine which mode is employed to process the MB. The detailed information of the super-resolution in DCT domain and the SRDDS will be introduced in the following sections.

The proposed method differs with the schemes presented in the prior literatures [95]-[97] [101] [105] [106]. Although the image compression approaches in [96] [97] [101] [105] [106] also employ the interpolation oriented adaptive down-sampling, they are designed to down-sample the whole original image for coding and try to recover the full-resolution image during the decoding process, which makes that the higher frequency components of the local texture and edge regions cannot be faithfully restored. Lin *et al.* presented an adaptive block-based down-sampling method in [95]. Three down-sampling modes with four different QP settings are employed, which results in high complexity of the encoder. Our experimental results reveal that only one mode is sufficient to improve the coding efficiency. Therefore, some down-sampling modes in [95] are not necessary, which just introduce the overhead information for the coded image. Also the down-sampling process in [95] is not optimized that cannot ensure higher quality reconstructed MBs.

Super-Resolution in DCT Domain

In order to reduce the computation complexity for the decoding process, the super-resolution performed in DCT domain [107] [108] is employed for generating the full-resolution MB from the down-sampled low-resolution sub-block.

In the decoding process, the de-quantization process results in $N \times N$ DCT coefficients $Coef_{N\times N}$ (N is equal to 8). The DCT coefficients are firstly extended into

$2N \times 2N$ by inserting the remained positions of the $2N \times 2N$ matrix $Coef_{2N\times 2N}$ with 0, which is defined as:

$$Coef_{2N\times 2N} = \begin{bmatrix} Coef_{N\times N} & 0_{N\times N} \\ 0_{N\times N} & 0_{N\times N} \end{bmatrix} \tag{2.12}$$

where $0_{N\times N}$ is $N \times N$ zeros matrix. Then the inverse DCT is applied $Coef_{2N\times 2N}$ to reconstruct the full-resolution MB by:

$$P_{2N\times 2N} = D^T_{2N\times 2N} \times (Coef_{2N\times 2N}) \times D_{2N\times 2N} \tag{2.13}$$

where $P_{2N\times 2N}$ is the full-resolution MB obtained from the super-resolution method, $D_{2N\times 2N}$ denotes the DCT kernel for $2N$ samples the superscript T denotes the tanspose of the matrix. Therefore, Eq. 2.13 can be further expressed as:

$$P_{2N\times 2N}(m,n) = \sum_{p=0}^{2N-1}\sum_{q=0}^{2N-1} \alpha_p\alpha_q \cdot Coef(p,q) \cdot cos\Big(\frac{\pi(2m+1)p}{4N}\Big)cos\Big(\frac{\pi(2n+1)q}{4N}\Big)$$

$$where \quad 0 \le m \le 2N-1, 0 \le n \le 2N-1, \tag{2.14}$$

$$\alpha_\Delta = \begin{cases} 1/\sqrt{2N}, & \Delta = 0 \\ 1/\sqrt{N}, & 1 \le \Delta \le 2N-1 \end{cases}$$

and Δ represents p or q. For the super-resolution method in DCT domain, the full-resolution MB can be reconstructed during the inverse transformation, which can significantly reduce the complexity of the decoder. Moreover, a fast algorithm of the super-resolution method is presented in [107], which only requires 3.1874 multiplications for each pixel.

Proposed Super-Resolution Directed Down-Sampling (SRDDS)

As aforementioned, the decoder employs the simple super-resolution method performed in DCT domain for up-sampling the low-resolution sub-block to the full-resolution MB. Therefore, in order to minimize the reconstruction error, an optimized low-resolution sub-block needs to be generated from the original block by considering the super-resolution process. It can be formulated as:

$$\hat{b}_{N\times N} = \mathbf{argmin}_b \left\{ \| B_{2N\times 2N} = SR(b_{N\times N}) \|_2^2 \right\} \tag{2.15}$$

where $SR(b_{N\times N})$ is the enlarged $2N \times 2N$ block by the super-resolution method presented in previous sub-section, $B_{2N\times 2N}$ is the original full resolution $2N \times 2N$ MB. The solution of Eq. 2.15 is the optimized down-sampled low-resolution sub-block $\hat{b}_{N\times N}$, which yields the smallest reconstruction error.

The super-resolution process in Eq. 2.13 can be expressed as:

$$
\begin{aligned}
P_{2N\times 2N} &= D^T_{2N\times 2N} \times \begin{bmatrix} Coef_{N\times N} & 0_{N\times N} \\ 0_{N\times N} & 0_{N\times N} \end{bmatrix} \times D_{2N\times 2N} \\
&= D^T_{2N\times 2N} \times \begin{bmatrix} D_{N\times N} \times b_{N\times N} \times D^T_{N\times N} & 0_{N\times N} \\ 0_{N\times N} & 0_{N\times N} \end{bmatrix} \times D_{2N\times 2N} \\
&= (D^T_{N\times N} \times D_{N\times 2N})^T \times b_{N\times N} \times (D^T_{N\times N} \times D_{N\times 2N}) \\
&= V_{2N\times N} \times b_{N\times N} \times H_{N\times 2N}
\end{aligned}
\tag{2.16}
$$

where $D_{N\times N}$ denotes the DCT kernel for N samples, $D_{N\times 2N}$ represents the upper most N rows of $D_{2N\times N}$, $V_{2N\times N}$ and $H_{N\times 2N}$ indicate the vertical and horizontal super-resolution kernels, respectively. And the transpose relationship between their kernels reflects $V^T_{2N\times N} = H_{N\times 2N}$. The vertical super-resolution kernel is defined as:

$$
V_{2N\times N}(m,n) = \sum_{k=0}^{N-1} \alpha_k \cdot cos\Big(\frac{\pi(2m+1)k}{4N}\Big) cos\Big(\frac{\pi(2n+1)k}{2N}\Big)
$$

$$
where \quad 0 \le m \le 2N-1, 0 \le n \le N-1, \tag{2.17}
$$

$$
\alpha_k = \begin{cases} 1/N, & k = 0 \\ 2/N, & 1 \le k \le 2N-1 \end{cases}
$$

Therefore, the super-resolution process in Eq. 2.13 in DCT domain can be further interpreted as the corresponding up-sampling in spatial domain, as shown in Eq. 2.16. Then the up-sampling can be implemented separately by multiplying the vertical kernel followed by multiplying the horizontal kernel. In the following, $V_{2N\times N}$ and $H_{N\times 2N}$ are denoted as V and H, respectively, for simplicity. The *Frobenius* norm of the matrix A, with $a_{i,j}$ as its component, is employed as the objective function, which is defined according to:

$$
\| A \|_F \stackrel{def}{=} \Big(\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \| a_{ij} \|_2^2 \Big)^{1/2} \tag{2.18}
$$

Then the optimized low-resolution sub-block $\hat{b}$ from the full-resolution MB B can be obtained by the minimization problem:

$$\hat{b}_v = \mathbf{argmin}_b \left\{ \| \, (V \otimes H^T) b_v - B_v \, \|_2^2 \right\} \tag{2.19}$$

where b_v, $\hat{b}_v$, and B_v are vectors obtained from the corresponding matrices, and $\otimes$ is the *Kronecker* product between two matrices. The optimized $\hat{b}_v$ can be obtained according to:

$$\hat{b}_v = (M^T M)^{-1} M^T B_v \tag{2.20}$$

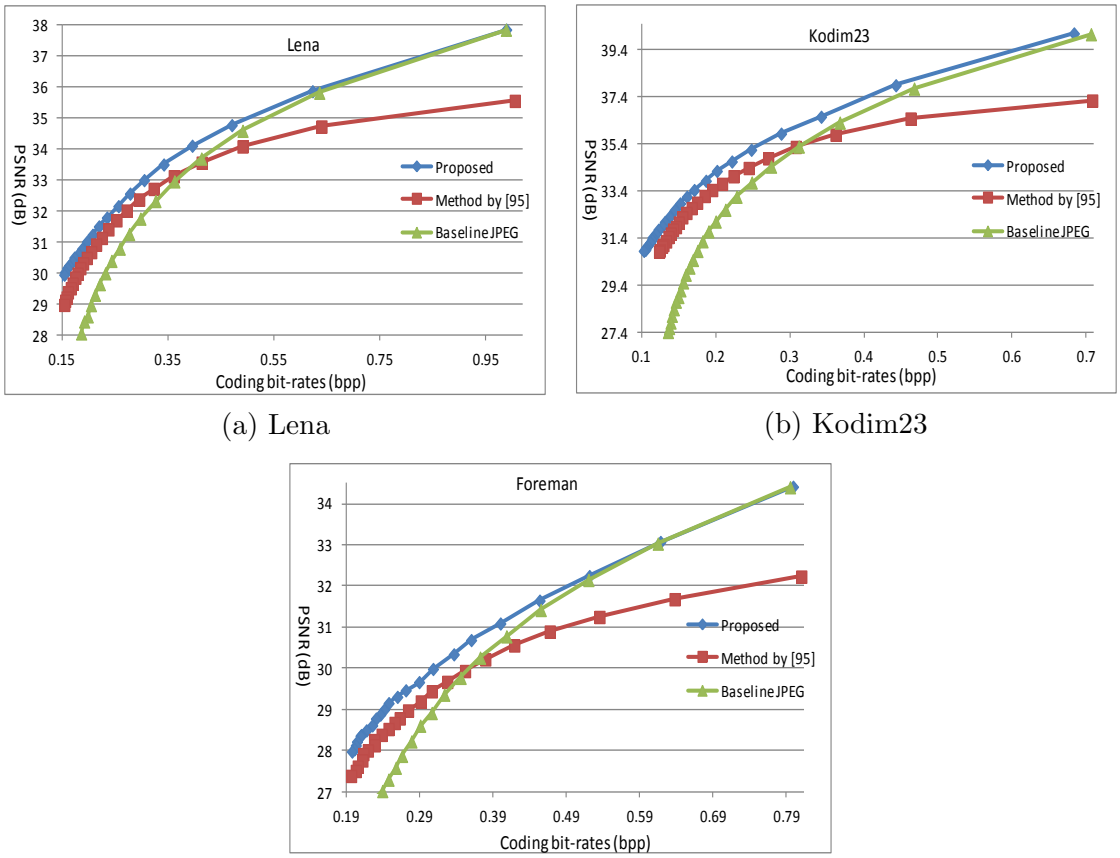
where $M = (V \otimes H^T)$. Then after inverse vectorization, the optimized low-resolution sub-block $\hat{b}$ can be obtained. The sub-block $\hat{b}$ will be processed by transformation, quantization. Finally the RDO process [104] will determine whether the MB is coded by the traditional JPEG mode or the SRDDS mode. Therefore, 1-bit flag for each MB is encoded and transmitted to indicate which mode is employed.

2.2.3 Experimental Results

In order to demonstrate the coding efficiency of the proposed scheme, four typical grey scale images are employed for experiments: Lena (512×512), Goldhill (512×512), Foreman (352×288), and Kodim23 (768×512) [109]. The baseline JPEG coding method and the down-sampling based image coding scheme [95] are compared with the proposed method.

The images are coded by different coding schemes, with the bit rates ranging from 0.1 bpp to 0.7 bpp. The objective quality of the coded image is evaluated by PSNR. The higher the PSNR, the smaller the difference between the reconstructed image and the original one. Detailed information of PSNR comparisons is illustrated in Figure 2.6. From the results, it can be observed that the PSNR of the image inferred from our method is significantly higher than the baseline JPEG coded image and the image coded by [95], especially at the low bit-rates. Furthermore, the method by [95] degrades the performance at high bit-rates. The reason is that it employs three down-sampling methods and four QP settings, which results in too many overhead bits to be en-coded and transmitted.

Furthermore, in order to demonstrate the perceptual gain of our proposed method,



(a) Lena (b) Kodim23

(c) Foreman

Figure 2.6: *PSNR comparisons of the proposed scheme, baseline JPEG, and the method by [95]*

in Figure 2.7, we have illustrated some images decoded from the baseline JPEG method, and the proposed method, respectively. It can be observed that the baseline JPEG decodes images with severe blocking artifacts, which greatly degrades the visual quality. However, the pro-posed method reconstructs images with better visual quality. In order to further evaluate the image quality, SSIM [68] [69], which is believed to be more consistent to the HVS perception than PSNR, is employed to evaluate the perceptual quality of each reconstructed image. According to its definition, the larger the SSIM value, the better the visual quality of the image. As illustrated from the experimental results in Figure 2.7, our proposed method generates better visual quality images with higher SSIM values.

bpp=0.1315;PSNR=26.94dB;SSIM=0.7551

bpp=0.1308;PSNR=32.09dB;SSIM=0.8798

bpp=0.2285;PSNR=27.78dB;SSIM=0.6564

bpp=0.2264;PSNR=29.02dB;SSIM=0.7498

bpp=0.2122;PSNR=29.31dB;SSIM=0.7901

bpp=0.2115;PSNR=31.33dB;SSIM=0.8497

Figure 2.7: *Subjective quality comparisons. Left: baseline JPEG images; right: images generated by the proposed scheme.*

2.2.4 Conclusion

A novel perceptual image coding scheme via adaptive block-based super-resolution directed down-sampling is proposed. The down-sampling method in the encoder is directed by the super-resolution method, which ensures the minimal reconstruction error. In the decoder, the super-resolution method is implemented in the DCT domain, which can be integrated with the inverse DCT transform process. Therefore, it can significantly reduce the computational complexity. The experimental results have demonstrated that our methods can improve the decoded image quality in terms of both objective and subjective measurements. For future works, the proposed SRDDS method will be implemented into JPEG2000 or even H.264 to achieve higher compression performances.

Chapter 3

Full Reference Video Quality Assessment

3.1 Adaptive Block-size Transform based Just-Noticeable Difference Model for Visual Signals

3.1.1 Introduction

Just-noticeable difference (JND) accounts for the smallest detectable difference between a starting and a secondary level of a particular sensory stimulus in psychophysics [110], which is also known as the difference limen or differential threshold. JND model has given a promising way to model the properties of the Human Visual System (HVS) accurately and efficiently in many image/video processing research fields, such as perceptual image/video compression [53] [111]- [114], image/video perceptual quality evaluation [58] [115]- [117] , watermarking [118] and so on.

Generally automatic JND model for images can be determined in the spatial domain or the transform domain, such as DCT and Discrete Wavelet Transform (DWT), or the combination of the two schemes [119]. JND models generated in the spatial domain [120] [121], named as the pixel-based JND, mainly focus on the background luminance adaptation and the spatial contrast masking. In [113] [114], Yang *et al.* deduce the overlapping effect of luminance adaptation and spatial contrast masking to refine the JND model in [120]. However pixel-based JND models do not consider the human vision sensitivities of different frequency components. Therefore it cannot describe the HVS properties accurately. JND models generated in the transform domain, namely the subband-based JND, usually incorporate all the major effecting factors, such as Contrast Sensitivity Function (CSF), luminance adaptation and contrast masking. In [111], the JND model is developed from the spatial CSF. Then the DCTune JND model [53] is developed by considering the contrast masking. Hontsch *et al.* [112] modify the DCTune model by replacing a single pixel with a foveal region, and Zhang

et al. [122] refine the JND model by formulating the luminance adaptation adjustment and contrast masking. More recently, Wei *et al.* [124] incorporate new formulae of luminance adaptation, contrast masking and Gamma correction to estimate the JND threshold in the DCT domain. Zhang *et al.* [119] propose to estimate the JND profile by summing the effects in DCT and spatial domain together.

In order to extend the JND profile from spatial to temporal, temporal characteristics of the HVS are considered. The previous works mostly focus on the perceptual differences between an original video sequence and its processed version [58] [117]. Actually, the temporal HVS properties are highly correlated with the video signals, and can be approximated by a computational model. In [120] [113] [114], an empirical function based on the luminance difference between adjacent frames is proposed to model the temporal masking property. In [125], Kelly proposes to measure the spatio-temporal CSF model at a constant retinal velocity, which is tuned to a particular spatial frequency. Daly [126] refines the model by taking the retina movement compensation into consideration. Jia *et al.* [127] estimate the JND for video sequences by considering both the spatio-temporal CSF and eye movements. And Wei *et al.* [123] [124] take the directionality of the motion into consideration to generate the temporal modulation factor.

However all the existing DCT-based JND models are calculated based on the 8×8 DCT, which do not consider the perceptual properties of the HVS over transforms of different block sizes. Recently adaptive block-size transform (ABT) has attracted researchers' attention for its coding efficiency in image and video compression [128]- [130]. It cannot only improve the coding efficiency but also provide subjective benefits, especially for high definition (HD) movie sequences from the viewpoint of subtle texture preservation [131] [132]. Specifically, transforms of larger blocks can better exploit the correlation within the block, while the smaller block size is more suitable for adapting to the local structures of the image [133]. Therefore by incorporating ABT into the JND, an adaptive JND model is obtained which can more precisely model the spatio-temporal HVS properties. Furthermore, since ABT has been adopted in current video coding standards, the ABT-based JND model for images/videos should be considered for applications such as video compression, image/video quality assessment, watermarking, and so on.

In this work, extension from 8×8 DCT-based JND to 16×16 DCT-based JND is performed by conducting a psychophysical experiment to parameterize the CSF for the 16×16 DCT. For still images or the intra video frames, a new spatial selection strategy based on the spatial content similarity (SCS) is utilized to yield the JND map. For the inter video frames, a temporal selection strategy based on the motion characteristic similarity (MCS) is employed to determine the transform size for generating the JND map. The rest of the chapter is organized as follows. Section 3.1.2 briefly introduces the extension procedure from the 8×8 JND to 16×16 JND. The proposed spatial and temporal selection strategies are presented in Section 3.1.3. The experimental performances are demonstrated and compared with the existing relevant models in Section 3.1.4. Finally, Section 3.1.5 concludes the chapter.

3.1.2 JND Model based on Transforms of Different Block Sizes

JND model in the DCT domain is determined by a basic visibility threshold T_{basic}, the spatial and temporal modulation factors [123]. It can be expressed as:

$$T(k,m,n,i,j) = T_{spatio}(m,n,i,j) \times \alpha_{tempo}(k,m,n,i,j) \tag{3.1}$$

$$T_{spatio}(m,n,i,j) = T_{basic}(i,j) \times \alpha_{lum}(m,n) \times \alpha_{cm}(m,n,i,j) \tag{3.2}$$

where k denotes the frame index of the video sequence, (m,n) is the position of DCT block in the current frame, (i,j) indicates the DCT coefficient position, α_{lum} and α_{cm}, denoting the luminance adaptation and contrast masking, constitute the spatial modulation factor. The video JND model T is obtained by modulating spatial JND model T_{spatio} with the temporal modulation factor α_{tempo}.

Extension From 8×8 to 16×16 DCT based JND

Based on the band-pass property of the HVS in the spatial frequency domain, the HVS sensitivity characteristics are modeled in [134] [135] as:

$$H(\omega) = (a + b\omega) \cdot exp(-c\omega) \tag{3.3}$$

where ω is the specified spatial frequency. JND is defined as the reciprocal of the HVS sensitivity characteristics given by Eq. 3.3. Hence the basic JND threshold can be

modeled as [124]:

$$T_{basic}(i,j) = \frac{s}{\phi_i \phi_j} \frac{exp(c\omega_{ij})/(a + b\omega_{i,j})}{\gamma + (1-\gamma)cos^2\varphi_{ij}} \tag{3.4}$$

where $s = 0.25$ denotes the summation effect factor, γ is set as 0.6, ϕ_i and ϕ_j are the DCT normalization factors, and $\varphi_{ij} = aresin(2\omega_{i0}\omega_{0j}/\omega_{ij}^2)$ indicates the directional angle of the corresponding DCT subband. ω_{ij} is the spatial frequency of the (i,j) subband. As claimed and verified in [130], 4×4 DCT does not contribute much to the efficiency of HD video coding. Since the proposed JND model aims at improving the performance of the perceptual HD video coding, only the 8×8 and 16×16 DCTs are considered to constitute the ABT-based JND model.

In order to extend the 8×8 JND to 16×16, the DCT block dimension N is set to 16. And a psychophysical experiment is carried out to parameterize the three parameters a, b, and c in Eq. 3.4. For a 512×512 image, with all pixel intensities are set as 128, noises are injected into several selected 16×16 DCT subbands to decide whether it is visible. The following two aspects need to be considered for the DCT subbands selection.

(i) The selected DCT subbands should cover the low, middle and high frequency components. We select at least one DCT subband located on each row and each column. Consequently, the selected spatial frequencies are uniformly distributed within the HVS sensitivity frequency range.

(ii) At least one selected DCT subband should be located on each diagonal. Therefore, the spatial frequencies with all directions are covered, with which the HVS directional sensitivities are taken into account.

Furthermore, we consider the oblique effect [136], where human eyes are more sensitive to the horizontal and vertical frequency components than the diagonal ones. The sensitivities of horizontal and vertical components appear to be nearly symmetrical. Consequently, only the DCT subbands of the upper-right portion (as shown in Figure 3.1) are chosen by considering the two aforementioned aspects. For the selected DCT subbands, several amplitude levels of the noises are pre-defined. The initial amplitude of the noise for each selected DCT subband is obtained by referring to the spatial CSF presented in [134] [135]. Then the noise amplitude is tuned into several levels, which make the noise rang from invisible to obviously visible based on the preliminary measure of the authors. During the tuning process, according to the CSF, larger magnitude

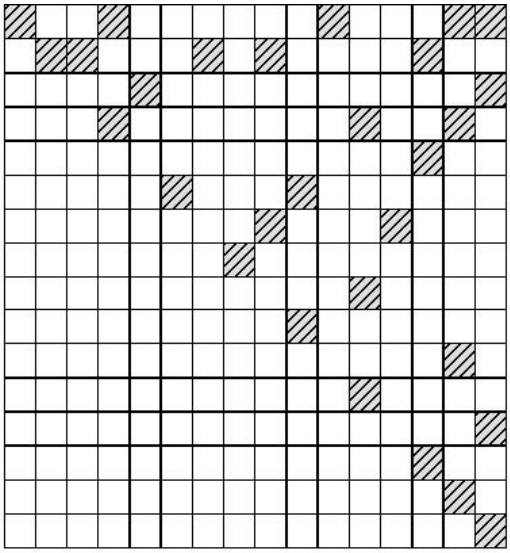


Figure 3.1: *Selected 16×16 DCT subbands for the psychophysical experiment (the shaded cells denote the selected DCT subbands)*

alternations of the noises are performed in the subbands with lower sensitivities. Also the oblique effect [136] results in lower HVS sensitivities for the subbands with larger directional angles. Therefore, the noise amplitude alternations in the subbands with larger directional angles should be larger. Then the noise, with its amplitude as one of the pre-defined levels, is inserted into the selected DCT subbands of the image. The original image and the processed one (with noise insertion) are juxtaposed on the screen. Ten viewers vote on whether the noise is visible. If half of them choose "yes", the noise amplitude is recognized as above the JND threshold. A smaller amplitude noise will be inserted. Otherwise, a larger one will be chosen for injection. Finally, the obtained thresholds of the selected DCT subbands are employed to minimize the least squared error as given in Eq. 3.5 to parameterize (a, b, c):

$$(a, b, c) = \mathbf{argmin} \sum_{\omega_{ij}} \left(T_{\omega_{ij}} - T_{basic}(i, j)\right)^2 \tag{3.5}$$

where $T_{\omega_{ij}}$ is the JND threshold obtained from the psychophysical experiment. The above procedure yields the parameters, $a = 0.183$, $b = 0.165$, and $c = 0.16$ for the 16×16 JND model.

JND is influenced by the intensity scale of the digital image. It is reported that higher visibility threshold occurs in either dark or bright regions compared with the medium brightness regions. The luminance adaptation factor α_{lum} forms a **U**-shape

curve [115] [122] [119] [137] [138]. Therefore, an empirical formula [124] is employed to depict the α_{lum}:

$$\alpha_{lum} = \begin{cases} (60 - I_{ave})/150 + 1, & I_{ave} \leq 60 \\ 1, & 60 < I_{ave} < 170 \\ (I_{ave} - 170)/425 + 1, & I_{ave} \geq 170 \end{cases} \tag{3.6}$$

where I_{ave} denotes the average intensity of the DCT block.

For the contrast masking factor, a block-based method [113] [114] is utilized to accurately describe the different masking properties of different block categories. These methods categorize the blocks into different block types according to the DCT subband energy [119] [122] [127] or image spatial characteristics [123] [124]. As in [124], we categorize the image block into three types, namely $PLANE$, $EDGE$, and $TEXTURE$, based on the proportion of the edge pixels in the 16×16 MB. The MB categorization is defined according to:

$$Categ_{16} = \begin{cases} PLANE, & \sum_{EP} < 16 \\ EDGE, & 16 \leq \sum_{EP} \leq 52 \\ TEXTURE, & \sum_{EP} > 52 \end{cases} \tag{3.7}$$

where $\sum_{EP}$ denotes the number of edge pixels in a given MB. Considering the block category and the intra-band masking effect [122] [124] [119], the contrast masking factor α_{cm} for 16×16 JND is obtained. Detailed information about the contrast masking scheme can be found in [139].

For the temporal modulation factor α_{tempo}, Robson [140] has shown that the form of the sensitivity fall-off at high spatial frequencies is independent of the temporal frequency and vice versa, while a sensitivity fall-off at low spatial frequencies occurs only when the temporal frequency is also low and vice versa. In [123] [124], it demonstrates that the logarithms of the temporal contrast sensitivity values follow approximately the same slope (nearly -0.03) for different spatial frequencies. By further considering the band-pass characteristic at the lower spatial frequencies [125], the temporal modulation

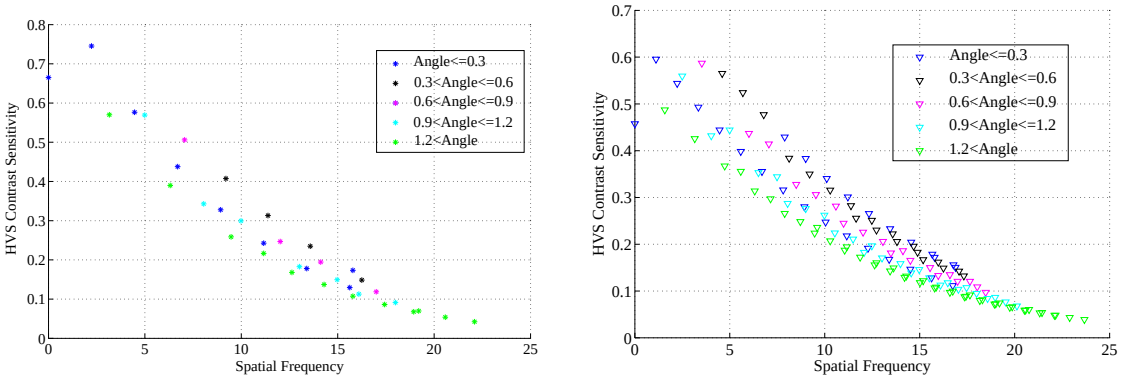


Figure 3.2: *Modeled HVS sensitivities over transforms of different block sizes by Eq. 3.4. (a): the HVS sensitivity over 8×8 DCT in [124]; (b): the HVS sensitivity over 16×16 DCT.*

factor is derived as:

$$\alpha_{tempo} = \begin{cases} 1, & \omega_s < 5cpd \text{ and } \omega_t < 10Hz \\ 10^{-0.03(\omega_t - 10)}, & \omega_s < 5cpd \text{ and } \omega_t \geq 10Hz \\ 10^{-0.03(\omega_t)}, & \omega_s \geq 5cpd \end{cases} \tag{3.8}$$

where ω_s and ω_t denote the spatial and temporal frequency, respectively. ω_s is determined by the transform size and the viewing distance, while ω_t relies on both the spatial frequency ω_s and the motion information, which is approximated by the block-based motion estimation [123] [124].

Why introduce ABT into JND?

The HVS sensitivities over transforms of different block sizes are illustrated in Figure 3.2. Firstly, as explained before, the HVS sensitivities are constrained within a spatial frequency range, which is approximately from 0 to 25 cpd. Therefore, the HVS sensitivities can be modeled more accurately by using a larger number of frequency bases. As shown in Figure 3.2, the HVS sensitivities for the 8×8 DCT are very sparse compared with the ones for the 16×16 DCT. The HVS sensitivity properties cannot be accurately modeled by only employing the 8×8 DCT based sensitivity function. Secondly, the HVS directional sensitivities need to be considered. From Figure 3.2, many points of the 16×16 sensitivities, which have nearly the same spatial frequency but different angle information, demonstrate different HVS sensitivities. The higher the angle

information, the lower the HVS contrast sensitivities, which matches the HVS oblique effect [136]. However for the sensitivity values of 8×8, there are very few points with different angle information. It cannot accurately represent the HVS directional properties. Considering the two aforementioned aspects, the sensitivities of 16×16 can more accurately model the HVS properties. It can help to find more accurate parameters a, b, and c in Eq. 3.4 for depicting the HVS sensitivities.

From the viewpoint of energy compaction, a larger block size transform takes advantage of exploiting the correlation within a block. On the other hand, the smaller one is more adaptive to the local structural changes. Therefore, transforms of different block sizes adapting to the image content play a very important role in image/video processing tasks, especially in image/video compression. And it has been claimed [141] that ABT can provide subjective benefits, especially for HD movie sequences from the viewpoint of subtle texture preservation, such as keeping film details and grain noises which are crucial to the subjective quality [142]. We believe that ABT-based JND model will make the HVS properties modeling more accurate, and benefit the perceptual-related image/video applications.

As ABT has been adopted into the current video coding schemes such as H.264, it is therefore necessary to develop the ABT-based JND model. It can be easily incorporated into the current coding standards. In [113] [114] perceptual video coding schemes employing the 8×8 DCT JND have been proposed. With the proposed ABT-based JND model, a more efficient perceptual coding scheme can be developed.

3.1.3 Selection Strategy Between Transforms of Different Block Sizes

The formulations of the JND models for the 8×8 and 16×16 DCT transforms are described in the previous section. Decision method for the proper transform size, *i.e.*, 8×8 or 16×16, will be discussed in this section.

Spatial Selection Strategy for Transforms of Different Block Sizes

As the selection strategy is designed for each MB, the image is firstly divided into 16×16 MBs. For each MB, two JND models based on 8×8 and 16×16 DCT are obtained. For the still images or intra video frames, where there is no motion information, we propose the spatial content similarity (SCS) to measure the image content homogeneity between

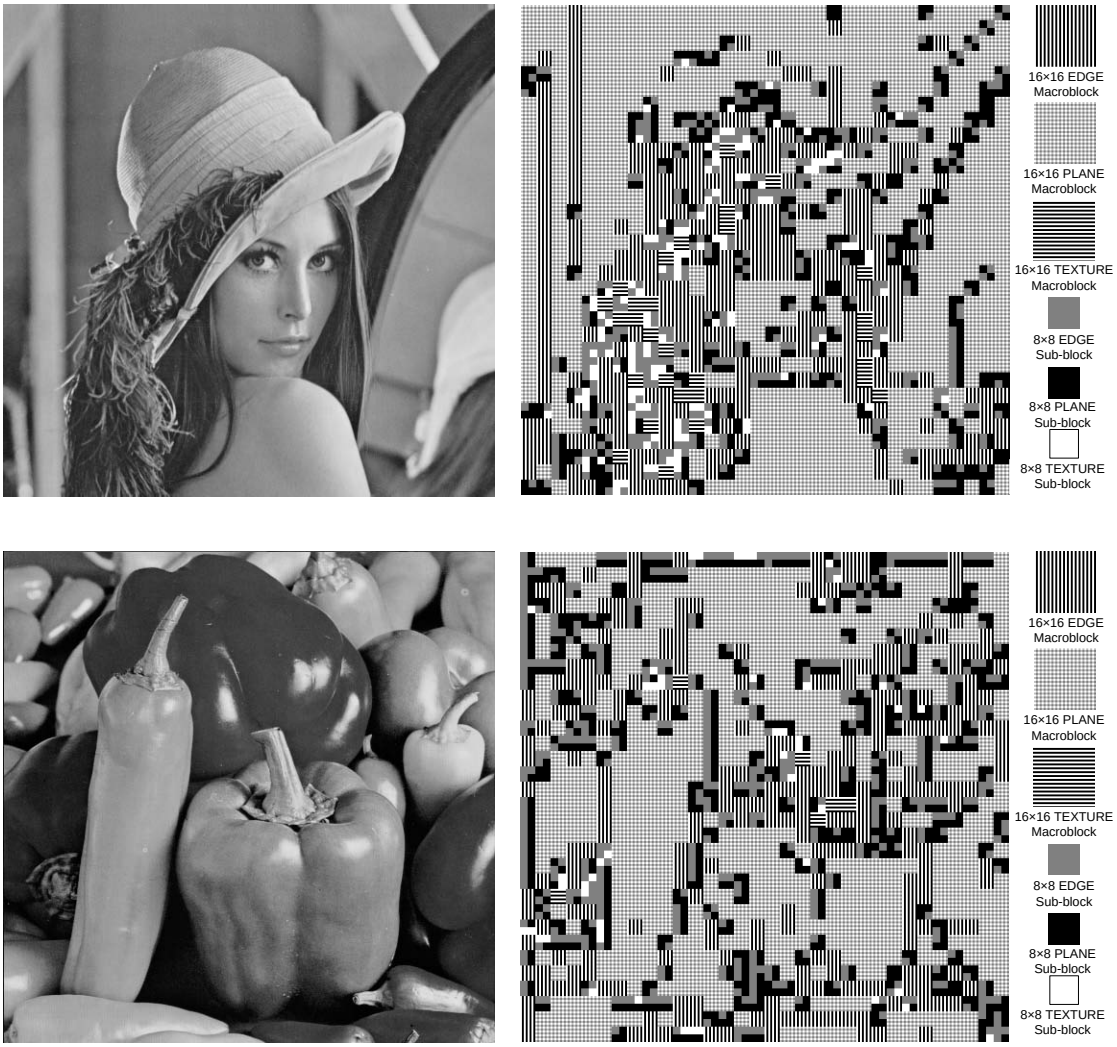


Figure 3.3: *Spatial selection results of Lena and Peppers. Left: the original image; right: spatial selection results in terms of block category and transform block size.*

an MB and its sub-blocks:

$$SCS = \sum_{i=1}^{4} (Categ_{16} \equiv Categ_8^i) \tag{3.9}$$

where $Categ_{16}$ and $Categ_8^i$ denotes the categories of the MB and the i-th 8×8 sub-block, respectively. SCS indicates the number of 8×8 sub-blocks with the same categorization as the MB which they belong to. If SCS is equal to 4, referring to the homogeneous content within the MB, the JND model based on 16×16 DCT will be utilized to yield the resulting JND model. On the contrary, if SCS is smaller than 4, the 8×8 JND model will be employed for adapting the local structures within the sub-blocks. The results of spatial selection strategy for Lena and Peppers are shown in Figure 3.3. Most of the

$PLANE$ regions employ the 16×16 JND model, while the areas with local structure changes utilize 8×8 JND model. The results are consistent with the energy compaction capabilities of the 8×8 and 16×16 DCTs.

Temporal Selection Strategy for Transforms of Different Block Sizes

For inter video frames, the JND model needs consider not only the spatial but also the temporal information. Therefore, we should include the temporal motion characteristics, which are depicted by motion vectors of different size blocks.

Based on the motion vectors of different size blocks, we propose a motion characteristics similarity (MCS) to measure the motion consistency between a MB and its sub-blocks, which is expressed as:

$$MCS = \sum_{i=1}^{4} \parallel Mv_8^i - Mv_{16} \parallel_2^2 \Big/ 4 \tag{3.10}$$

where Mv_8^i denotes the motion vector of the i-th 8×8 sub-block, Mv_{16} is the motion vector of the 16×16 MB, and $\| \cdot \|_2^2$ denotes the Euclidean distance between the two motion vectors. Considering the spatial SCS and temporal MCS, we can make decision on which transform block size to use for the resulting JND.

If the calculated MCS is smaller than a threshold, it is deemed that the motion characteristics of the MB and its corresponding sub-blocks are nearly the same. In this chapter, we empirically set the threshold as 1.25 pixels. When SCS is equal to 4 and MCS smaller than the threshold, the MB is considered to be a single unit. Therefore, 16×16 DCT based JND is utilized to generate the JND model. On the other hand, if the MCS is larger than the threshold, indicating that motion vectors of the MB and its sub-blocks are quite different, the MB should be separated into 4 sub-blocks because of the smaller SCS and larger MCS. The 8×8 DCT based JND for each sub-block is then employed to obtain the resulting JND model.

In order to further test the consistency between the spatial and temporal selection strategies, the hit ratio (HR) curve is used to demonstrate the hit rate for each inter video frame. Firstly, we record the MB JND types determined by the aforementioned spatial and temporal selection strategies, respectively. The hit rate h of each INTER frame measures the percentage of the MBs (as determined by the spatial and temporal

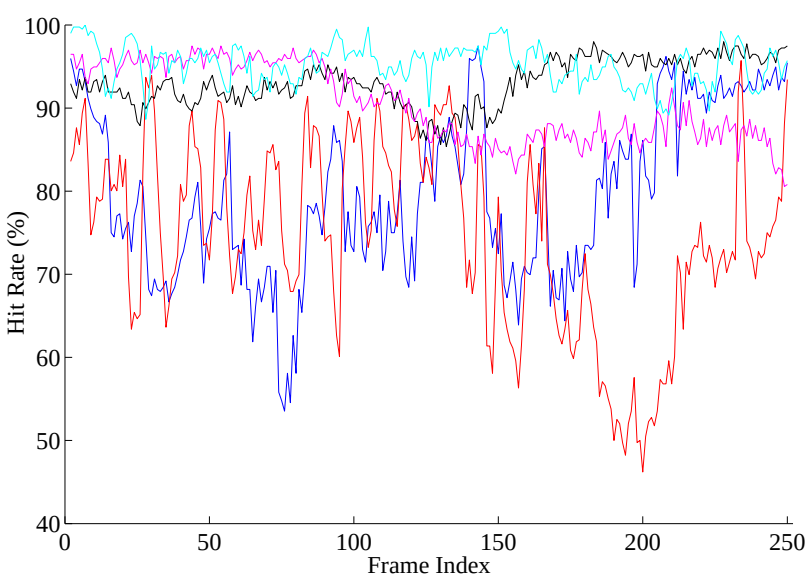


Figure 3.4: *HR curves of the MBs for each inter frame of the test video sequences*

selection strategies) are identical. In this case, the transform of the same block size is selected for a macroblock to generate the resulting JND model. The HR curves for each INTER frame of several typical CIF (352×288) sequences are illustrated in Figure 4. The hit rates h are high, corresponding to the fact that the proposed temporal selection strategy accords well with the spatial selection strategy. The proposed selection strategy is efficient for depicting both spatial image content information and temporal video motion characteristics. Furthermore, the hit rates of Football and Foreman are a bit lower than the other sequences, with the average hit rate as 77%. The reason is that both sequences contain high motion characteristics. Therefore, the consistency between spatial and temporal characteristics tends to be low. On the other hand, as the motion appears slightly in the other sequences, the hit rate become much higher, with the average value as 93%.

3.1.4 JND Model Evaluation

In order to demonstrate the efficiency of the proposed ABT-based JND model, the noise is injected into each DCT coefficient of each image or video frame to evaluate the HVS error tolerance ability:

$$\widetilde{I}_{typ}(k,m,n,i,j) = I_{typ}(k,m,n,i,j) + R \cdot T_{typ}(k,m,n,i,j) \tag{3.11}$$

where $\widetilde{I}_{typ}$ is the noise-contaminated DCT coefficient which is located on the $(i,j)-th$ position of the $(m,n)-th$ block in the $k-th$ frame. For still images, k is set as 0. R takes the value of +1 or -1 randomly to avoid introducing a fixed pattern of changes. I_{typ} is the JND threshold obtained by the proposed ABT-based scheme, and typ denotes the final transform block size to generate the resulting JND model.

Evaluation on Images

We tested the proposed JND model on several typical 512×512 images and 768×512 Kodim images [109]. We compare the proposed method with Yang *et al.*'s method [113], which evaluated the JND in the image domain, and Wei *et al.*'s method [124] which calculates the JND in the DCT domain. Comparisons in terms of PSNR are listed in Table 3.1, which show that our proposed JND method yields smaller PSNR values compared with other JND models. Here if the image visual quality stays the same as the original one, it implies that our JND model can tolerate more distortions.

	Yang (dB)	Wei (dB)	Proposed JND (dB)
Baboon	32.53	28.38	27.46
Barbara	31.35	29.49	29.02
Bridge	30.96	29.01	28.53
Lena	32.72	29.97	29.51
Peppers	30.78	29.99	29.66
Kodim06	32.21	29.02	28.61
Kodim08	31.21	29.11	28.73
Kodim13	30.59	28.75	28.42
Kodim14	30.00	29.41	29.14
Kodim21	32.15	29.43	29.06

Table 3.1: *PSNR comparisons of different JND models*

In order to provide a more convincing evaluation of the proposed JND model, subjective tests are conducted to assess the perceptual qualities of the noise-contaminated images. In the subjective test, two images were juxtaposed on the screen. One is the original image as the reference and the other is the distortion-inserted version, which is regarded as the SC method specified in Section 1.2. In this experiment, the viewing monitor is a Viewsonic professional series P225fb CRT display. The viewing distance is set as 4 times the image height. Ten observers (half of them are experts in

image processing and the other half are not) are asked to offer their opinions on the subjective quality of the images, by following the quality comparison scale shown in Table 1.2. Their average subjective values are calculated to indicate the image visual quality, which is illustrated in Table 3.2. Also the mean and variance values of the subjective scores are calculated. According to the quality comparison scale in Table 1.1, the smaller the subjective scores, the better quality of the noise contaminated images. The proposed method has the smallest mean value (only 0.37), demonstrating the best performance. From the subjective results, Yang *et al.*'s method can generate higher quality images, such as Baboon, Kodim13, and Kodim14. These images exhibit more texture information. For the images with much plain or edge information, such as Peppers and Kodim21, the visual quality will degrade significantly. Our method generates smaller variance compared with the other methods, indicating that the proposed scheme performs more consistently over images of different types. The noise-inserted images generated by our method can be found in [143].

	Yang	Wei	Proposed JND
Baboon	0.5	0.2	0.2
Barbara	1.2	0.4	0.5
Bridge	0.7	0.3	0.3
Lena	0.8	0.3	0.4
Peppers	1.0	0.4	0.4
Kodim06	1.0	0.6	0.4
Kodim08	1.2	0.5	0.6
Kodim13	0.4	0.4	0.3
Kodim14	0.5	0.3	0.2
Kodim21	1.5	0.6	0.4
Average	0.88	0.40	0.37
Variance	0.362	0.133	0.125

Table 3.2: *Subjective evaluation results. Left: noise-contaminated image by different JND model; right: the original image*

Evaluation on Videos

The proposed JND model was evaluated on several typical CIF (352×288) video sequences, with a frame rate of 30fps. In our experiments, 250 frames of each sequence are tested, with the first frame as intra and the rest as inter frames. We also compare

the proposed method with Yang *et al.*'s [113], and Wei *et al.*'s JND models [124]. Since we have evaluated the efficiency of ABT-based JND model for images, here only the average PSNR of the inter frames is calculated. Comparisons in terms of PNSR are listed in Table 3.3. It is observed that the proposed JND model yields smaller PSNR values compared with other JNDs. It shows that the ABT-based JND model can tolerate more distortions.

	Yang (dB)	Wei (dB)	Proposed JND (dB)
Tempete	31.68	27.42	27.04
Football	34.43	28.39	28.17
Foreman	35.29	28.29	28.02
Mobile	33.10	27.48	26.93
Silence	34.43	28.26	27.93
Table	36.37	27.81	27.33
Stefan	35.20	27.83	27.38
Paris	33.56	27.60	27.07
Flower	35.57	27.18	26.80
Waterfall	33.88	27.83	27.52

Table 3.3: *PSNR comparisons of different JND models*

The subjective test was conducted to further assess the perceptual quality of the noise-contaminated videos. DSCQS method, as specified in Section 1.2, is employed to evaluate the perceptual quality. Two sequences were presented to viewers, one of which is original and the other is processed. Ten viewers (half of them are experts in image/video processing and the other half are not) were asked to offer their opinions. Five-grade subjective rating scale as illustrated in Table 1.1 were employed by the viewers for the rating process. The difference between subjective scores of the original and noise-injected video sequence is calculated as the DMOS. Hence, the smaller the DMOS, the higher is the quality of the noise-contaminated video. The testing conditions are the same as the image evaluation process. Detailed subjective test results are depicted in Table 3.4. The mean DMOS value of the proposed scheme is 6.89, which is smaller than Yang *et al.*'s and Wei *et al.*'s methods. It reflects that our proposed method can generate similar quality videos with the original ones. Also it can be found that variance of the DMOS value is the smallest. Compared with the other methods, our approach delivers more consistent results for both the fast-moving video sequences,

e.g., Football and Stefan, and the slightly-moving video sequences, e.g., Silence and Paris.

	Yang	Wei	Proposed JND
Tempete	7.3	6.6	6.4
Football	7.6	6.2	5.6
Foreman	13.2	9.2	8.3
Mobile	9.7	7.0	7.1
Silence	13.9	9.7	8.5
Table	6.9	6.2	5.2
Stefan	7.2	6.0	5.4
Paris	14.2	9.4	9.2
Flower	13.2	8.2	7.4
Waterfall	6.5	5.6	5.8
Average	9.97	7.41	6.89
Variance	3.269	4.565	1.429

Table 3.4: *Subjective evaluation results. DMOS for noise-contaminated video sequences.*

3.1.5 Conclusion

In this section, a novel ABT-based JND profile for visual signals is proposed by exploiting the HVS properties over different transform sizes. New selection strategies are proposed for each MB to decide which transform block-size is to be employed by considering not only spatial SCS but also temporal MCS. The developed JND profile can tolerate more distortions with the same visual quality compared with other JND models.

3.2 Perceptual Quality Assessment

Traditional error measures for images/videos, such as MSE and PSNR, do not correlate well with the HVS for evaluating the image/video perceptual quality [41] [68] [72] [80] [144] [145]. In this section, we design a very simple visual quality metric based on the proposed ABT-based JND model (introduced in Section 3.1, which is defined as:

$$Diff_{typ}(k,m,n,i,j) = \begin{cases} 0, & if\ \mathbf{D}(k,m,n,i,j) \le T_{typ}(k,m,n,i,j) \\ \mathbf{D}(k,m,n,i,j) - T_{typ}(k,m,n,i,j), & otherwise \end{cases} \tag{3.12}$$

$$where\ \ \mathbf{D}(k,m,n,i,j) = \left| I_{typ}(k,m,n,i,j) - I^{D}_{typ}(k,m,n,i,j) \right|$$

$$P_{dist}(k,m,n,i,j) = \tau_{typ}\frac{Diff_{typ}(k,m,n,i,j)}{T_{typ}(k,n,n,i,j)} \tag{3.13}$$

$$V_Q = 10\log_{10}\left(\mathbf{mean}_{(k,m,n,i,j)}\left(P^2_{dist}(k,m,n,i,j)\right)\right) \tag{3.14}$$

where T_{typ} is the ABT-based JND, typ denotes the transform block size for generating the JND, I_{typ} is the DCT coefficient of the reference image/frame, I^D_{typ} denotes the DCT coefficient of the distorted image/frame, and $Diff_{typ}$ denotes the DCT coefficient difference between the reference image/frame and the distorted one by considering the HVS error tolerance ability. Since the JND denotes the threshold for detecting the perceptual difference (as demonstrated in Section 3.1.4), the distortions below the JND thresholds cannot be perceived by the human eyes. They need not be accounted in measuring the visual quality, where the visual difference is set as 0. In Eq. 3.12 above, only the distortions larger than the JND thresholds are calculated for measuring the visual quality. The adjustable parameter τ_{typ} is introduced according to the different energy compaction properties, which are determined by the coding gains of different block transforms. The coding gain [146] for the block transform is defined as:

$$G_{TC} = 10\log_{10}\left[\frac{\frac{1}{N}\sum_{i=0}^{N-1}\sigma_i^2}{\left(\Pi_{i=0}^{N-1}\sigma_i^2\right)}\right]^{\frac{1}{N}} \tag{3.15}$$

where N is the number of the transform subbands, σ_i^2 is the variance of each subband i, for $0 \le i \le N-1$. Then τ_{typ} is defined according to:

$$\tau_{typ} = \begin{cases} G^8_{TC}/G^{16}_{TC}, & typ\ is\ 16\times 16 \\ 1, & typ\ is\ 8\times 8 \end{cases} \tag{3.16}$$

where G^8_{TC} and G^{16}_{TC} denote the coding gains of 8×8 and 16×16 DCT, respectively. After testing on the reference images of the LIVE subjective image quality database [25], the coding gain ratio appears to be nearly the same. Therefore, we simply set it as 0.95. P_{dist} is the distortion masked by the proposed ABT-based JND model. The visual quality metric V_Q is obtained by aggregating the P_{dist} of all the transform blocks in one frame. If we evaluate the visual quality metric of an image, only the spatial JND model is employed and k is set as 0. If the video quality is assessed, the proposed metric employs the spatio-temporal JND model. In our approach, the visual quality of each frame is measured individually. Hence the visual quality of the whole video

sequence is given by the mean quality value of all the frames.

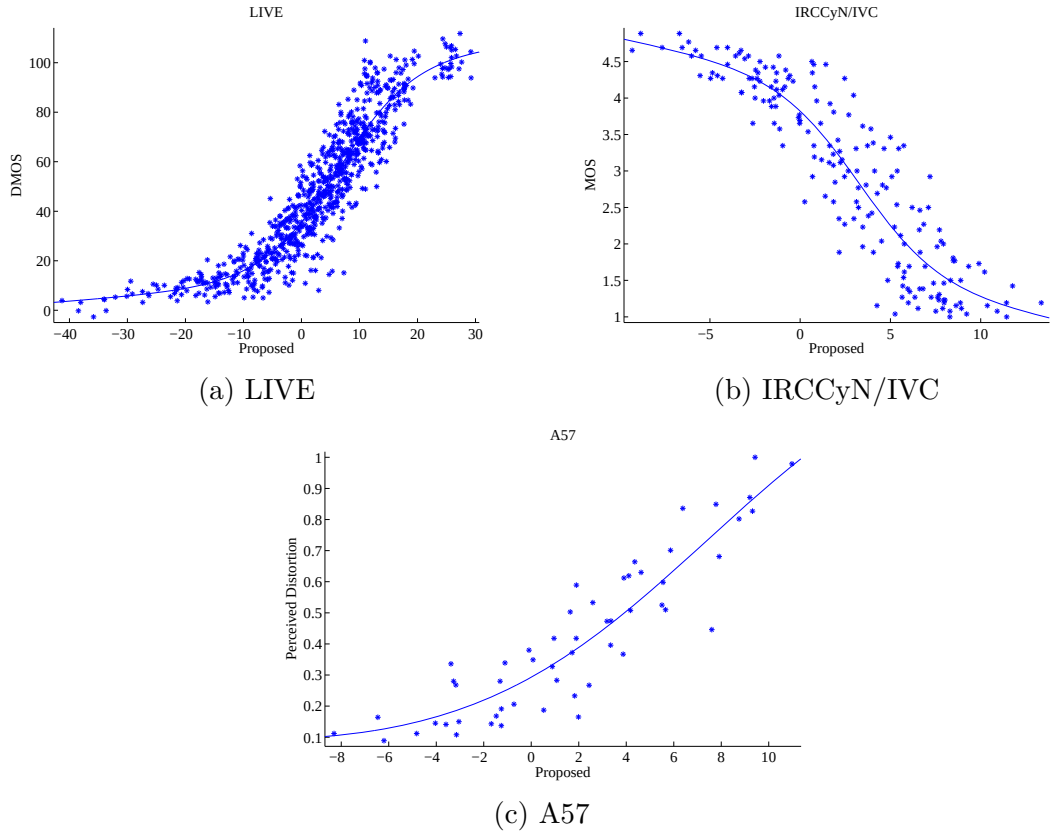
3.2.1 Experimental Results

We have tested the performance of the proposed metric, as well as the state-of-the-art image quality metrics, such as SSIM [68] [69], VIF [72], and VSNR [144] over the LIVE [25], A57 [37], and IRCCyN/IVC [26] image subjective quality database. Table 1.3 lists some major characteristics of these image databases. They contain the most prevailing distortions, such as JPEG, JPEG 2000, blurring, additive Gaussian noise, and so on. Each distorted image in these subjective quality databases is assigned a subjective score, e.g., DMOS for LIVE image/video database, MOS for the IRCCyN/IVC database, and perceived distortion for the A57 database. These subjective scores were obtained from subjective viewing tests where many observers participated and provided their opinions on the visual quality of each distorted image. These subjective scores are regarded as the ground truths for evaluating the performances of different visual quality metrics. As introduced in Section 1.3.4, SROCC, LCC, and RMSE are employed to evaluate the performances, which are illustrated in Table 3.5. And the scatter-plots of different quality metrics are illustrated in Figure 3.5 and [143]. It can be observed that our proposed method scatter closely around the fitted curve, which indicates a good performance.

Database		PSNR	SSIM	VSNR	VIF	Proposed
LIVE	LCC	0.8716	0.904	0.637	0.956	0.933
	SROCC	0.8765	0.910	0.648	0.958	0.934
	RMSE	13.392	11.68	21.13	7.99	9.881
IRCCyN/IVC	LCC	0.704	0.776	0.800	0.903	0.913
	SROCC	0.679	0.778	0.798	0.896	0.909
	RMSE	0.866	0.769	0.731	0.524	0.498
A57	LCC	0.644	0.415	0.942	0.618	0.913
	SROCC	0.570	0.407	0.936	0.622	0.901
	RMSE	0.192	0.224	0.083	0.193	0.101

Table 3.5: *Performance Comparisons of different image quality metrics*

Furthermore, we tested the proposed visual quality metric on the LIVE video subjective quality database [17], whose major characteristics are also listed in Table 1.3. The video subjective quality index is obtained by averaging the frame VQ scores, the



(a) LIVE (b) IRCCyN/IVC

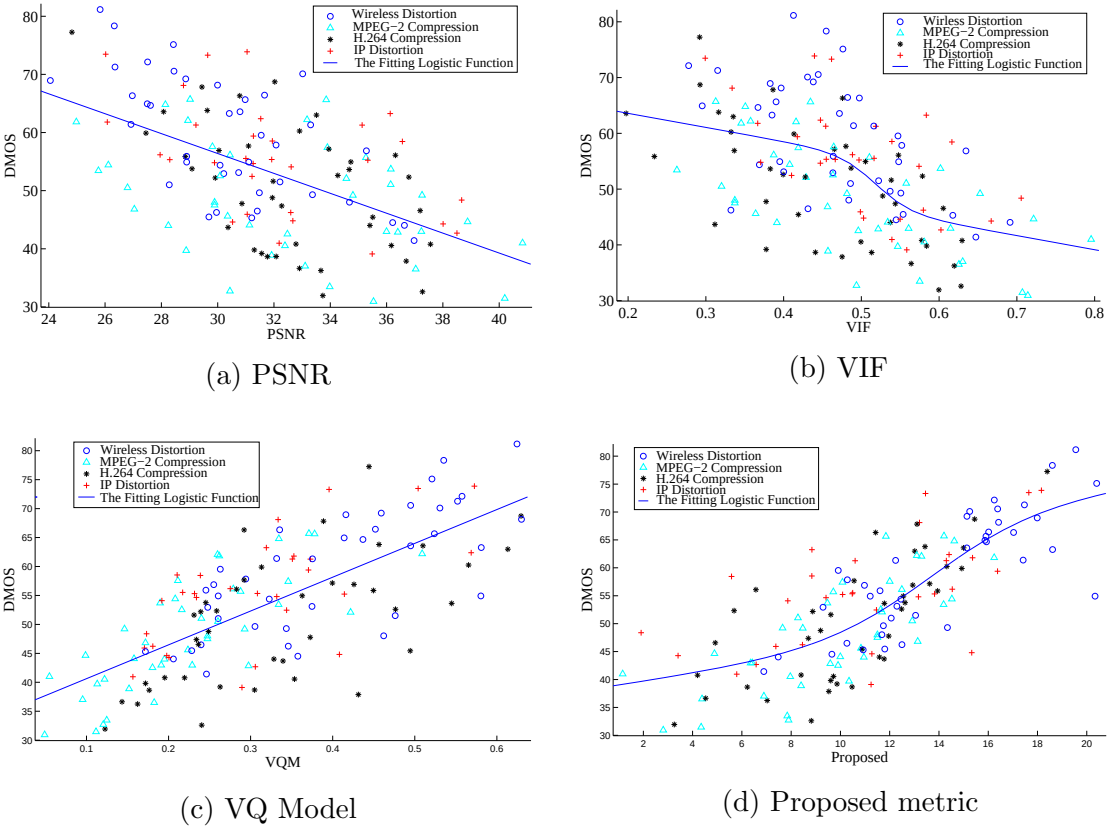
(c) A57

Figure 3.5: *Scatter plots of the subjective values versus model prediction on the image subjective databases.*

same as PSNR, SSIM, and VIF. And we also compared with the most popular video quality metrics VQ Model [80] and MOVIE [145]. As usual, after non-linear mapping, CC, SROCC and RMSE are employed for evaluating the performances, as shown in Table 3.6. It is observed that the proposed method outperforms other video quality metrics, while slightly inferior to MOVIE. The scatter-plots are provided in Figure 3.6 and [143]. The results of our proposed method scatter closely around the fitted curve, indicating a good performance.

Database		PSNR	SSIM	VIF	VQ Model	MOVIE ⋆	Proposed
LIVE	LCC	0.5398	0.4999	0.5735	0.7160	0.8116	0.780
	SROCC	0.5234	0.5247	0.5564	0.7029	0.7890	0.761
	RMSE	9.241	9.507	8.992	7.664	-	6.935

Table 3.6: *Performance Comparisons of different video quality metrics. (⋆ LCC and SROCC value of MOVIE are obtained directly from [145], which does not provide the RMSE value.)*



(a) PSNR

(b) VIF

(c) VQ Model

(d) Proposed metric

Figure 3.6: *Scatter plots of the DMOS values versus model prediction on the LIVE video subjective quality database.*

3.2.2 Conclusion

From the test results, our proposed visual quality metric performs comparably with the stat-of-the-art quality metrics. It clearly demonstrates that the proposed ABT-based JND model can incorporate the HVS properties into the context of perceptual quality assessment. It is found that SSIM and VIF perform very well on image quality evaluation. But they fail in assessing the video subjective quality. The reason is that SSIM and VIF succeed to depict the spatial distortions, but fail to capture the temporal distortions. That is the reason why VQM outperforms SSIM and VIF, for it has considered the temporal effect. However, the temporal effect in VQ Model is simply modeled by the fame differences. It cannot efficiently depict the temporal distortions, resulting in a slightly better performance. MOVIE is developed by considering the complex temporal and spatial distortion modeling, leading to the best performance. However, it is very complex and time-consuming, hence cannot be easily applied in practical applications.

As the proposed visual quality metric has modeled both the spatial and temporal HVS properties, it performs equally well with VIF and MOVIE. It maintains a very simple formulation in DCT domain. Therefore, the proposed visual quality metric can be easily applied to image/video applications, especially the perceptual video coding.

3.3 Motion Trajectory Based Visual Saliency for Video Quality Assessment

Since the HVS is the ultimate receiver of the images/videos, it is very important and advantageous to incorporate HVS properties into the PQAs. In Section 3.1, the HVS perceptual property is depicted by JND, which has demonstrated good performances while incorporating with PQAs in Section 3.2. In Section 2.1, the HVS orientation property is modeled by the visual HE, which can help improve the IQA performances as shown in Section 2.1.3. Among the HVS properties, the visual saliency is straightforward and extremely important for PQAs. Nowadays, many computational models [84] [147] [148] have been proposed to simulate human's visual attention. Itti *et al.* propose a bottom-up model and build a system named Neuromorphic Vision C++ Toolkit [148]. Hou *et al.* propose a spectral residual (SR) approach [84], which is proved to be useful for IQA in Section 2.1. However, SR only considers the spatial information for images. Guo *et al.* propose phase spectrum (PS) [147] for detecting the video saliency. Its temporal information is simply modeled by the frame differences. As claimed and verified in [149], the performances of VQAs can be improved by considering the distortions along the temporal trajectories. Therefore, we propose to incorporate the motion trajectory for efficiently detecting the visual saliency of video sequences. A quaternion representation (QR) for each frame is constructed, which comprises the spatial image content, the motion trajectories, and the temporal residuals. Based on the QR, the quaternion Fourier transform (QFT) is employed to construct the visual saliency. Finally, the visual saliency is incorporated with several video quality metrics for evaluating its efficiency. The rest of the section is organized as follows. In Section 3.3.1 - Section 3.3.4, the proposed visual saliency model and its application on VQAs are introduced. Experimental results are demonstrated in Section 3.3.5. Finally, Section 3.3.6 concludes the section.

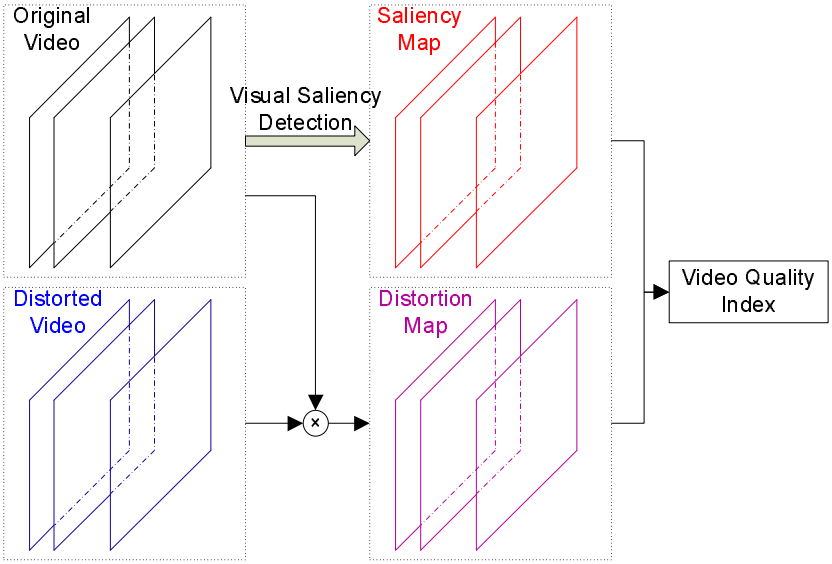


Figure 3.7: *VQA framework based on the proposed visual saliency*

3.3.1 Motion Trajectory based Visual Saliency for VQA

As illustrated in Figure 3.7, the proposed visual saliency model is applied on the original video sequences by considering both the image spatial content and the temporal motion trajectory. The distortion map is obtained by performing different VQAs, e.g. MSE, SSIM, on the original and distorted videos. Finally, by incorporating the saliency map with the distortion map, the video quality index of the distorted video is generated.

3.3.2 New Quaternion Representation (QR) for Each frame

In order to apply the proposed visual saliency model, each frame of the original video sequence needs to be represented as a quaternion image [150]. It consists of four components, each of which captures the useful information from one certain aspect. As we only perform VQAs on the luminance part of the distorted videos, the chroma information is not required to construct the quaternion image. Define the video sequence as $V(t), t = 1, 2, \ldots, N$, where N is the total frame number. $l(t)$ denotes the luminance part of $V(t)$.

The overlapped block-based motion estimation (OBME) scheme is employed to depict the temporal motion trajectory. After OBME, 3 temporal components of each frame are obtained. $MV_x(i,j)$ and $MV_y(i,j)$ denote the horizontal and vertical motion vector of the block centered at $(i,j) - th$ pixel, respectively. $PE(i,j)$ indicates the

corresponding motion prediction error. Together with the luminance $l(t)$, we have obtained the four components of the quaternion image. $l(t)$ represents the spatial image content. $MV_x(t)$ and $MV_y(t)$ describe the motion trajectory. $PE(t)$ depicts the temporal residual information, which compensates the inaccurate OBME. Each frame can be represented as the new quaternion image $q_i(t)$ [150] according to:

$$\begin{aligned} &q_i(t) = l(t) + PE(t)\mu_1 + MV_x(t)\mu_2 + MV_y(t)\mu_3 \\ &\mu_i^2 = -1, \quad i = 1, 2, 3 \\ &\mu_1 \perp \mu_2,\ \mu_2 \perp \mu_3,\ \mu_3 \perp \mu_1 \\ &\mu_3 = \mu_1\mu_2 \end{aligned} \tag{3.17}$$

We can further represent $q_i(t)$ in a symplectic form:

$$\begin{aligned} q_i(t) &= f_1(t) + f_2(t)\mu_2 \\ f_1(t) &= l(t) + PE(t)\mu_1 \\ f_2(t) &= MV_x(t) + MV_y(t)\mu_1 \end{aligned} \tag{3.18}$$

In [147], the quaternion image comprises one intensity channel, two color channels, and one motion channel. However, the motion channel is simply described by the adjacent frame difference. On the contrary, our new quaternion image consists of one luminance channel, two motion vector channels depicting the temporal trajectory, and one temporal residual channel. With the consideration of the temporal trajectory, the visual saliency map can be faithfully reconstructed, which will benefit the VQAs.

3.3.3 Saliency Map Construction by QR

As clarified in [147], only the phase spectrum is sufficient to represent the saliency information of each frame. Given an image $I(x, y)$,

$$\begin{aligned} f(x, y) &= \xi\big(I(x, y)\big) \\ p(x, y) &= P\big(f(x, y)\big) \\ SA_M(x, y) &= g(x, y) \times \parallel \xi^{-1}\Big(exp\big(i \cdot p(x, y)\big)\Big) \parallel_2^2 \end{aligned} \tag{3.19}$$

where ξ and ξ^{-1} denote the Fourier transform and inverse Fourier transform, respectively. $P(f)$ represents the phase spectrum of the image. $g(x, y)$ is a Gaussian filter. After the process in Eq. 3.19, the saliency map $SA_M(x, y)$ of $I(x, y)$ is generated.

For a quaternion image, the quaternion Fourier transform (QFT) [150] is employed to generate the visual saliency map. The QFT of a quaternion image $q(n,m)$ can be expressed as:

$$\begin{aligned} Q(\mu,\nu) &= F_1(\mu,\nu) + F_2(\mu,\nu)\mu_2 \\ F_i(\mu,\nu) &= \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} exp\big(- \mu_1 2\pi(\frac{m\nu}{M} + \frac{n\mu}{N})\big) f_i(n,m) \end{aligned} \tag{3.20}$$

where (n,m) and (μ,ν) are the locations of each pixel in time and frequency domain. N and M are the image height and width. $f_i,\ i \in \{1,2\}$ is obtained from Eq. 3.18.

The inverse QFT is defined as:

$$f_i(m,n) = \frac{1}{\sqrt{MN}} \sum_{\mu=0}^{M-1} \sum_{\nu=0}^{N-1} exp\big(- \mu_1 2\pi(\frac{m\nu}{M} + \frac{n\mu}{N})\big) F_i(\mu,\nu) \tag{3.21}$$

By applying Eq. 3.20, the frequency response $Q_i(t)$ of $q_i(t)$ can be obtained in the polar form as:

$$Q_i(t) = \parallel Q_i(t) \parallel exp(\mu \cdot p_i(t)) \tag{3.22}$$

where $p_i(t)$ is the phase spectrum of $Q_i(t)$ and μ is a unit pure quaternion.

As shown in Eq. 3.19, only the phase spectrum is sufficient to construct the visual saliency map. Therefore, $\parallel Q_i(t) \parallel$ is set as 1. Then by applying the inverse QFT in Eq. 3.21, the reconstructed quaternion image $\acute{q}_i$ is generated. Finally, the visual saliency map is constructed by the Gaussian filtering:

$$SA_M(t) = g * \parallel \acute{q}_i \parallel^2 \tag{3.23}$$

3.3.4 Incorporating Visual Saliency with VQAs

Several VQAs, such as MSE, SSIM [68], MSSIM [151], and ABT-based JND metric introduced in Section 3.2, incorporate the detected visual saliency for improving their performances. For MSE and ABT-based JND metric, the visual saliency map is employed to weigh the calculated differences:

$$sDiff(t) = |Diff(t)| \cdot SA_M(t) \tag{3.24}$$

where $Diff(t)$ denotes the differences between the original frame $O(t)$ and the distorted frame $D(t)$. For MSE, $Diff(t) = O(t) - D(t)$. As to ABT-based JND metric, $Diff(t)$

indicates the difference after the JND masking process, as illustrated in Eq. 3.13. The quality index for each frame is obtained by summing $sDiff(t)$ together:

$$Index(t) = 10\log_{10}\Big(mean\big(sDiff^2(t)\big)\Big) \tag{3.25}$$

As to SSIM, the visual saliency pooling strategy is performed over the structural distortion map, as defined in Eq. 2.11. MSSIM tries to apply SSIM over different scales of the image and sum the quality indexes together to evaluate the image quality. Adapting to this scheme, we down-sample the visual saliency map to different scales. By saliency pooling over different scales, the quality index of each frame is generated.

For each VQA, the quality index of each frame has been generated by considering the visual saliency map. Then the indexes are finally averaged to yield the video quality value (VQI):

$$VQI = \frac{\sum_{t=1}^{N} Index(t)}{N} \tag{3.26}$$

where N is the total frame number of the video sequence.

3.3.5 Experimental Results

We first provide the processing results during the visual saliency detection, which is illustrated in Figure 3.8. As we have discussed in Section 3.3.2, each frame will be represented as a quaternion image, comprising luminance $l(t)$, horizontal and vertical motion vector $MV_x(t)$ and $MV_y(t)$, and prediction error $PE(t)$. For better visualization, MV is rescaled by $5 \times MV + 128$; PE is rescaled by $PE + 128$. It can be observed that the entire object generates nearly the same motion information, such as the ball, the boat, and the players in the video sequences. After performing the OBME, the prediction error is obtained. By incorporating the motion trajectory information (depicted by the motion vectors) and the temporal residual information, the visual saliency map for the corresponding frame is constructed using QFT, as shown in Figure 3.8. It can be observed that the visual saliency can significantly detect the motion object (highlighted white) in the saliency map. By considering the accurate visual saliency map, the VQA performances can be significantly improved.

We incorporated the detected saliency map with MSE, SSIM [68], MSSIM [151], and ABT-based JND metric introduced in Section 3.2. All of these VQAs were tested on the

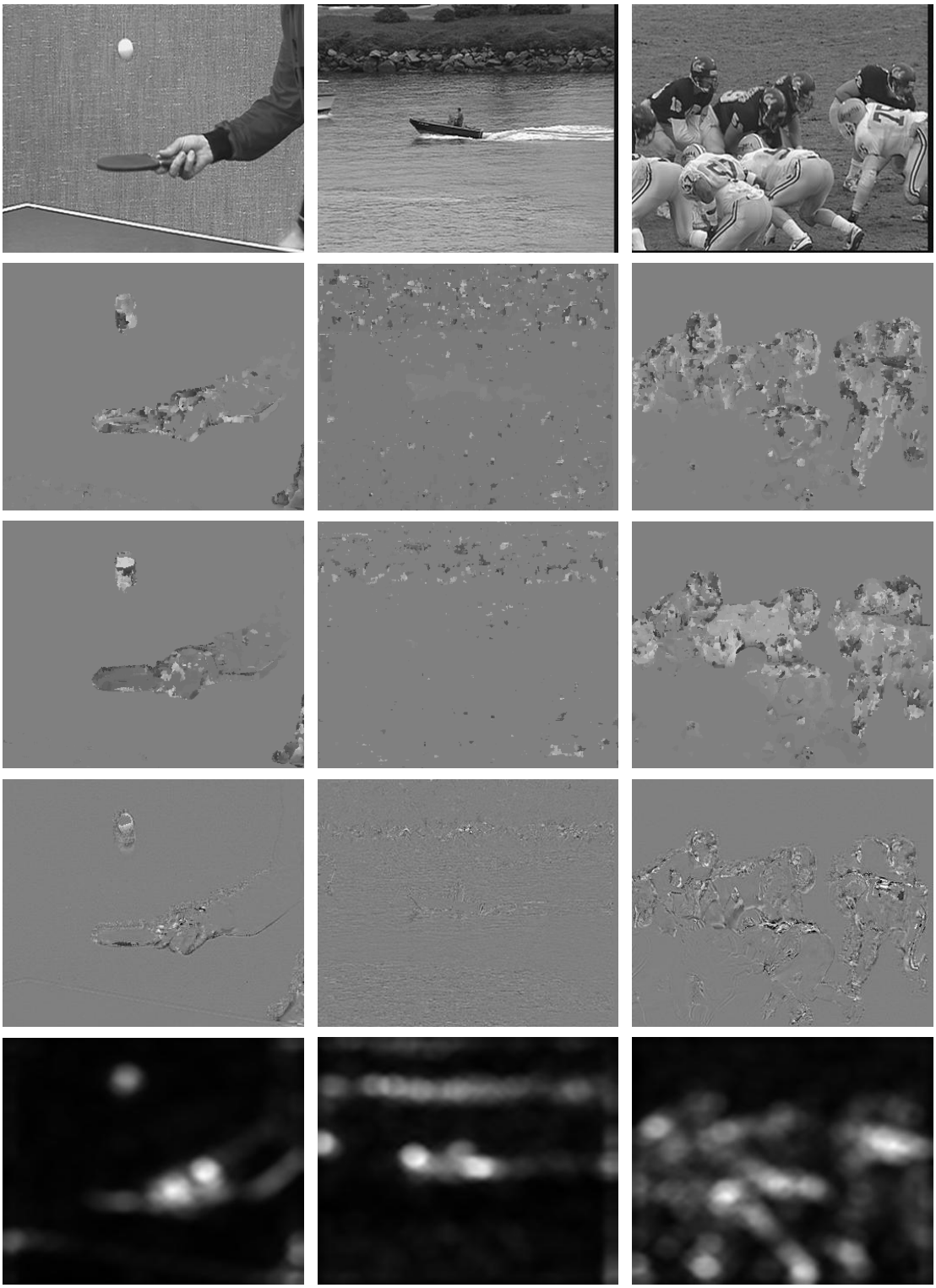


Figure 3.8: *Quaternion Representation (QR) of each frame and the visual saliency map. From top to bottom: luminance $l(t)$, horizontal motion vector $MV_x(t)$, vertical motion vector $MV_y(t)$, motion prediction error $PE(t)$, and the visual saliency map.*

LIVE video subjective quality database [17]. Detailed information about the database was introduced in Section 1.3.3. After the nonlinearly mapping, the LCC, SROCC, and RMSE are employed to evaluate different VQA performances. The performances of VQAs incorporating different visual saliency models are shown in Table I, where SR denotes the saliency model in [84]; PS is the saliency model in [147]; VS is our proposed method. It can be observed that all the saliency weighted metrics can outperform the non-weighted metrics. It means that the visual saliency is important to HVS and helpful for the VQAs. Furthermore, VS weighted VQAs outperform the other saliency weighted VQAs. The reason is that the proposed method considers the motion trajectory, which is useful to improve the VQA performances, as demonstrated in [149]. However, the saliency weighted methods still perform inferiorly to MOVIE [145]. The reason is that MOVIE has employed complex HVS model for depicting the temporal and spatial distortions, compared to the proposed saliency weighting method. Another observation is that the improvement of ABT-based JND metric is not so significant, compared with the other metrics. The reason is that ABT-based JND metric has considered some HVS properties, such as contrast masking, which has somehow modeled the HVS saliency property.

VQA methods	LCC	SROCC	RMSE
MSE	0.5398	0.5234	9.241
SSIM	0.4999	0.5247	9.507
MSSIM	0.6754	0.7329	8.095
ABT-based JND metric	0.7627	0.7372	7.099
SR-MSE	0.6164	0.6104	8.644
SR-SSIM	0.6215	0.6012	8.600
SR-MSSIM	0.7472	0.7360	7.296
SR-ABT-based JND metric	0.7623	0.7322	7.105
PS-MSE	0.6230	0.6191	8.588
PS-SSIM	0.6051	0.5909	8.740
PS-MSSIM	0.7371	0.7245	7.419
PS-ABT-based JND metric	0.7685	0.7338	7.023
VS-MSE	0.6295	0.6268	8.531
VS-SSIM	0.6308	0.6187	8.518
VS-MSSIM	0.7583	0.7468	7.157
VS-ABT-based JND metric	0.7768	0.7484	6.913

Table 3.7: *Performance comparisons of different VQAs*

3.3.6 Conclusion

In this section, we propose a new quaternion representation for each frame of the video sequence. Then the quaternion image is employed to generate the corresponding visual saliency map. By incorporating the visual saliency map with different VQAs, the metric performances can be significantly improved, which further confirms that the proposed method can accurately model the HVS saliency property.

3.4 Perceptual Video Coding

In this section, the ABT-based JND is incorporated into the video coding scheme for pursuing higher visual quality with the same bit-rates according to:

$$Re_{typ}(k,m,n,i,j) = DCT_{typ}\{I(k,m,i,j,k) - I_{pre}(k_{ref},m,n,i,j)\} \tag{3.27}$$

$$\acute{R}e_{typ}(k,m,n,i,j) = \begin{cases} 0, & if|Re_{typ}(k,m,n,i,j)| \leq T_{typ}(k,m,n,i,j) \\ V_{sign} \cdot \big(|Re_{typ}(k,m,n,i,j)| - T_{typ}(k,m,n,i,j)\big), & otherwise \end{cases}$$

where I is the MB to be encoded, I_{pre} is the predicted MB by inter motion estimation or intra prediction, typ denotes the transform size (8×8 or 16×16 DCT), Re_{typ} is the DCT coefficient of the prediction error, T_{typ} is the calculated JND threshold for different transform sizes, V_{sign} denotes the sign of the coefficient $Re_{typ}(k,m,n,i,j)$. According to the definition of JND and the quality metric in Eq. 3.12, the HVS cannot detect the distortions which are smaller than the JND threshold. Therefore, the distortions below the JND threshold need not be accounted. The perceptual redundancies in the video signals are removed according to Eq. 3.27, which will not cause any visual degradation. Then the resulting DCT coefficients $\acute{R}e_{typ}$ without perceptual redundancies are encoded.

For the traditional video coding strategy, MSE is utilized to calculate the distortions in rate distortion optimization (RDO), which is justified to be inconsistent with the HVS perception [41]. Here, P_{dist} is employed for depicting the HVS responses of the distortions, which is defined as:

$$P_{dist}(k,m,n,i,j) = \tau_{typ}\frac{\acute{R}e_{typ}(k,m,n,i,j)}{T_{typ}(k,m,n,i,j)} \tag{3.28}$$

The sum squared error of P_{dist} will be utilized as the distortion measurement for the

modified RDO (M-RDO) process. As demonstrated in Section 3.2.1, P_{dist} correlates better with the HVS than MSE, which is believed to benefit the perceptual video coding. During the encoding process, a suitable λ needs to be determined for the M-RDO process:

$$Cost = D_p + \lambda R \tag{3.29}$$

where D_p is the sum squared error of P_{dist}, and R denotes the bit-rate. In our experiments, four 720P sequences, Crew, Harbor, Sailormen, and Spincalendar are encoded with the H.264 platform provided by [130]. The test conditions are listed in Table 3.8 (only 100 frames), with QP ranging from 28 to 40. Then D_p is used to evaluate the coded sequences. According to the derivation in [104], the optimal λ is set as:

$$\lambda = -\frac{dD_p}{dR} \tag{3.30}$$

In our experiments, the tangent slopes at each identical QP point of the four testing sequences appear to be similar. Therefore, the average value of the tangent slopes is employed as λ in the M-RDO process.

Platform	JM 11(H.264) [130]
Sequence structure	IBBPBBP
Intra period	10 frames
Transform size	8×8, and 16×16
Entropy coding	CABAC
Deblocking filter	On
R-D Optimization	On
Rate control	Off
Reference frame	2
Search range	± 32
Frame rate	30 frames/s
Total frame number	199

Table 3.8: *Test conditions.*

In the encoding process, the M-RDO process is employed to determine the best transform type. We believe that the proposed selection strategy has strong ties with the M-RDO process. For one Mb, if the spatial content is homogenous within its sub-blocks, and the motion vector differences between the MB and its sub-blocks are small, the MB is regarded as a unit. The 16×16 DCT is chosen by the proposed selection

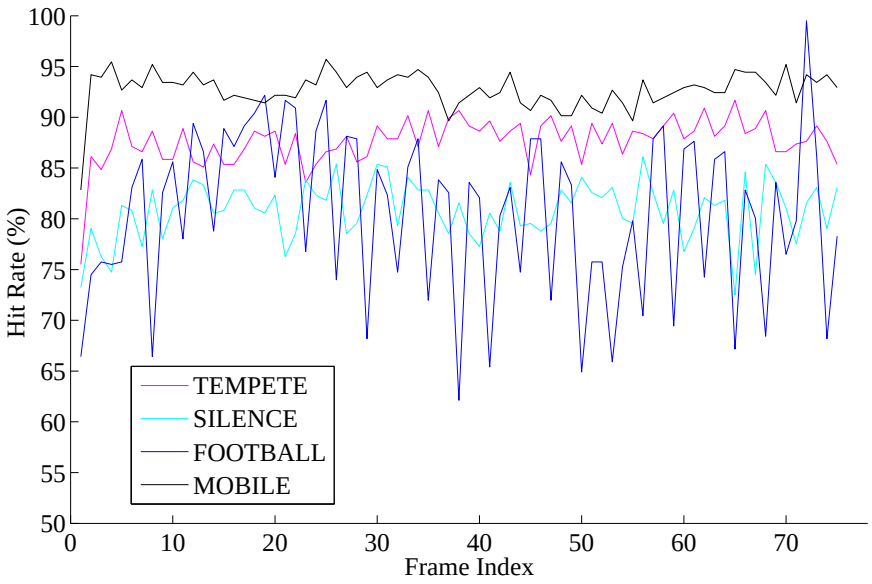


Figure 3.9: *HR curve for each CIF sequence*

strategy. During the encoding process, the MB can be well predicted by the 16×16 MB motion estimation. The prediction error will be very small. The 16×16 DCT thus can efficiently compact the energy, which will be chosen by the M-RDO process. Otherwise, the 8×8 DCT will be determined by both the selection strategy and M-RDO.

In order to demonstrate the relationship between the selection strategy and the M-RDO process, the Hit Ratio (HR) curve is employed to demonstrate the hit rates. The transform type (8×8 or 16×16 DCT) is first determined by the proposed selection strategy for each MB. Then the video sequences are encoded by the proposed perceptual coding scheme. The QP is fixed as 20 and the test conditions are listed in Table 3.8. During the encoding process, the transform type (8×8 or 16×16 DCT) for each MB as determined by the M-RDO process is also recorded. The hit rate h of each video frame measures the percentage of the MBs whose transform types determined by the M-RDO process and the proposed selection strategy are identical. It indicates that the selection strategy and M-RDO choose the same size transform. The HR curves of several typical CIF (352×288) sequences are illustrated in Figure 3.9. The hit rates are high, with the average hit rate higher than 80%. It means that the proposed selection strategy correlates well with the M-RDO process. During the video encoding, the M-RDO process will take the role to determine which size transform to use. For other applications, such as visual quality assessment, watermarking, and so on, where the

M-RDO process is not applicable, the proposed selection strategy will determine which size transform to utilize.

3.4.1 Experimental Results

The 720P test sequences, Crew, Harbor, Sailormen, and Spincalendar, were coded with fixed QP parameters. The H.264/AVC software platform used is the JM 11 with ABT implementation [130]. The test conditions are listed in Table 3.8. With different QP parameters, nearly the same bit-rates are ensured by the traditional ABT codec and the proposed ABT-based JND codec, as shown in Table 3.9. It can be observed that there is a slight PSNR loss. As explained before, the PSNR correlates poorly with the HVS perception, which makes it an improper criterion for visual quality assessment. The proposed visual quality metric V_Q as illustrated in Eq. 3.14 has demonstrated better performances in matching subjective ratings. We calculate the V_Q of the distorted sequences. According to its definition in Eq. 3.14, the smaller the V_Q value, the better is the visual quality. It can be observed from Table 3.9 that the sequences generated by our proposed method possess smaller VQ indexes, compared to the sequences processed by [130].

	Video	Bit-rates (kbit/s)	PSNR (dB)	V_Q	DMOS
ABT codec [130]	Crew	807.79	36.68	2.88	25.0
	Harbor	1068.34	30.05	13.32	37.3
	Sailormen	572.40	30.92	10.09	33.8
	Spincalendar	683.91	31.23	8.40	30.5
The Proposed ABT-based JND codec	Crew	806.28	36.42	2.76	22.3
	Harbor	1056.37	29.83	13.22	32.5
	Sailormen	576.37	30.86	9.78	30.5
	Spincalendar	688.70	31.05	8.23	25.3
Performance differences	Crew	-1.51	-0.26	-0.12	-2.7
	Harbor	-11.97	-0.22	-0.10	-4.8
	Sailormen	+4.11	-0.06	-0.31	-3.3
	Spincalendar	+4.79	-0.18	-0.17	-5.2

Table 3.9: *Performance comparisons between the tradition ABT codec [130] and the proposed ABT-based JND*

In order to demonstrate the perceptual gain of our proposed video codec, the DSC-QS subjective test as introduced in Section 1.2 was conducted to evaluate the visual qualities of the coded video sequences. And the DMOS value for each coded sequence

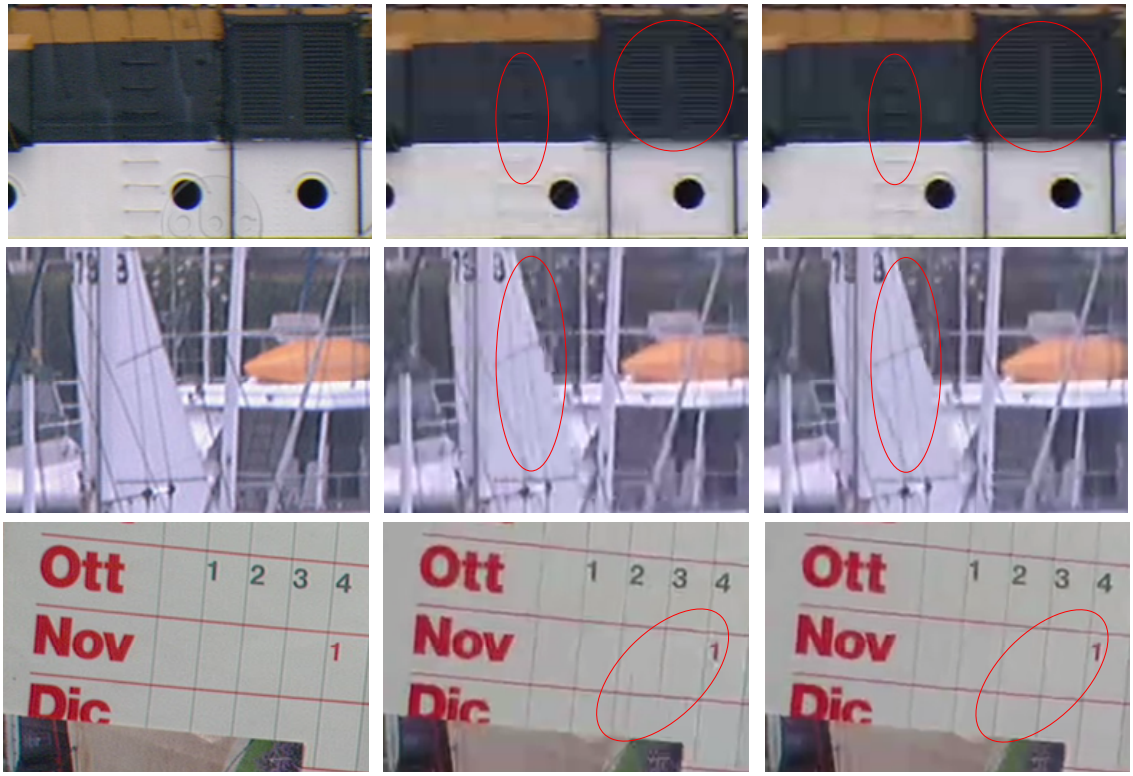


Figure 3.10: *Visual quality comparison of regions of the reconstructed frames generated by different video codec. Left: original frame; middle: reconstructed frame from ABT codec [130]; right: reconstructed frame for the proposed ABT-based JND codec. Top: 113^{th} frame of Sailormen; center: 109^{th} frame of Harbor; bottom: 40^{th} frame of Spincalendar.*

is listed in Table 3.9. As explained before, the smaller the DMOS value, the better the visual quality. Therefore, it can be observed that the proposed method can improve the visual quality of the coded video sequences with the constraint of the same bit-rates. Figure 3.10 shows some pictures of videos coded and decoded with the JM 11 with ABT implementation [130] on one hand and with the proposed method on the other hand. Generally, the proposed method generates frames with higher visual quality, especially the detailed information, such as the lines and edges of Harbor and Spincalendar sequences.

3.4.2 Conclusion

In this section,the ABT-based JND model is employed for guiding perceptual video coding. Experimental results on the proposed ABT-based JND codec demonstrate a better visual quality videos with the same bit-rates. It further confirms the efficiency of our proposed ABT-based JND in modeling the HVS characteristics.

Part II

Reduced Reference Quality Assessment

Chapter 4

Reduced Reference Image Quality Assessment

4.1 Introduction

As mentioned before, reduced reference (RR) quality metric is the compromise between the full reference (FR) and no reference (NR) quality metrics. It is expected that the RR methods can effectively evaluate the image perceptual quality based on a limited number of features extracted from the reference image. Only a small number of bits is required for representing the extracted features, which can be efficiently encoded and transmitted for the quality analysis. Consequently, it will be very useful for the quality monitoring during the image transmission and communication. The image perceptual quality can be easily analyzed by referring to the extracted features from the reference image. Therefore, a better quality of user experience can be further provided for the consumers.

For designing an effective RR quality metric, we need to consider not only its performance but also its RR data rate for representing the extracted features. Firstly, the extracted features should be sensitive to a variety of image distortions and relevant to the HVS perception of the image quality. Secondly, the RR data rates should not be large, as the extracted features need to be embedded or transmitted to the receiver side for the quality analysis. For a larger RR data rate, one may include more information about the reference image. Then a good performance can be obtained. However, it will introduce a heavy burden to the RR feature transmission. The FR IQA can be regarded as an extreme case of RR IQA, with the RR data rate is the whole reference image. For a smaller RR data rate, only a little information of the reference image is available for quality analysis. Therefore, the performance is hard to be ensured. The NR IQA is another extreme case of RR IQA, with no information from the reference image. Therefore, how to balance the RR data rate and the performance is the essential

for the RR quality metric development.

The RR quality metrics aim to monitor the video perceptual quality during the transmission and communication processes. Therefore, many approaches [152]- [157] try to model the distortions of the encoded video sequences, such as the MPEG-2 compressed videos, in the quality monitoring system. For example, Wolf *et al.* [152] [153] extracted a set of spatial and temporal features which are very sensitive to the distortions introduced in the standard video compression framework. In [154], Le Callet *et al.* depicted the blur, blocking and temporal artifacts of the MPEG-2 coded sequences by some representative features. By accounting for differences between these features, the degradation level of the coded videos can be estimated. In [155], Yang employed the ratio information of DCT coefficients to measure the perceptual quality of MPEG-2 coded sequences. In [156], the artifacts of the H.264/AVC coded video sequence, such as blur and blocking, are depicted and measured by the objective features. They are combined together into a single measurement for the overall video quality. Furthermore, Tagliasacchi *et al.* [157] approximated the SSIM value of the videos corrupted by channel errors through employing coding tools provided by the distributed source coding theory.

Furthermore, in order to provide a more accurate performance, the HVS properties [80] [158]- [166] have been considered during the feature extraction. Le Callet *et al.* [158] employed a neural network to train and evaluate the perceptual quality of video sequences, based on the perception related features of the video frames. In [159] [160], the authors extracted perceptual features motivated from the computational models of the low level vision. These features are utilized as the reduced descriptors to represent the visual quality. Tao *et al.* [161] incorporated the merits of the contourlet transform, the contrast sensitivity function, and Weber's law of JND to derive an RR IQA. Engelke *et al.* [162] designed an RR IQA for wireless imaging by accounting for different structural information that is observed in the distortion model of wireless link. Then the structural information from the viewing area is trained for the HVS. In [80], the authors extract several HVS related features to indicate the spatial information losses, edge information changes, contrast information, and color impairments. By combining these different components with different weights, the final video perceptual quality

index is obtained. These HVS related features are compressed for video quality monitoring [163]. It is demonstrated that a compression ratio of more than 30:1 can be achieved with only a small error introduced in the final quality values. Moreover, as the HVS is sensitive to the degradation around the edges, the RR video quality metric proposed in [164] mainly measures the edge degradations. The edge degradation is computed by measuring the mean squared error of the edge pixels. Therefore, this method is named as edge PSNR (EPSNR). In [165], the authors employed discriminative local harmonic strength with motion consideration to evaluate the distorted video quality. The gradient information of each frame is employed for harmonic and discriminative analysis. Furthermore, the authors in [166] derived the RR quality metric for 3D videos. The edge information of depth maps and information from the corresponding color image in the areas in the proximity of edges are extracted for the RR quality metric, which can be utilized for 3D video compression and transmission.

Recently, the statistical modeling of the image signal has been investigated for the image perceptual quality assessment for both RR IQAs [167]- [175]. In [169], the divisive normalization is employed to depict the coefficient distributions of the wavelet subbands. The distribution difference between the reference and distorted images is used to depict the image perceptual quality, which we name as RR-DNT. In [172] [173], the developed RR image quality metric RR-SSIM extracted the statistical features from a multi-scale, multi-orientation divisive normalization transform. By following the philosophy in the construction of SSIM, a distortion measurement is developed to estimate the SSIM index of the distorted image. In [174], the statistics of image gradient magnitude are modeled by the Weibull distribution to develop an RR image quality metric, which is named as RR-Weibull. Also the statistics of the edge [175] are utilized for developing the RR IQA, which we name as RR-Edge. In [170], the authors measure the differences between the entropies of wavelet coefficients of the reference and distorted image to quantify the image information change, which can indicate the image perceptual quality. In [171], the color distribution changes of an image as a consequence of the distortions are employed for depicting the perceptual quality, where the color correlogram is extracted as the RR feature. Wang *et al.* [167] [168] proposed a wavelet-domain natural image statistic metric (WNISM), which models the marginal probability distribution of the wavelet coefficients of a natural image by the generalized

Gaussian density (GGD) function. The Kullback-Leibler distance (KLD) is used to depict the distribution difference. Although WNISM can achieve good performances in image quality assessment, some limitations still exist. Firstly, KLD is asymmetric [176], which is not suitable for the quality analysis. The perceptual quality distance from one image to another should be identical no matter how it is measured. Secondly, as revealed in [169], although WNISM can work quite well on individual distortion types, its performance degrades significantly when image of different distortion types are evaluated together.

In this chapter, a novel RR IQA is developed by depicting the intra and inter subband statistical characteristics in the RDCT domain. It is shown that after performing DCT the statistical dependencies between the DCT subbands still exist. Applying the reorganization strategy, the intra RDCT subband statistical characteristic, specifically the identical natural of the coefficient distribution within the RDCT subband, is exploited by GGD modeling. The inter RDCT subband dependency is captured by the mutual information (MI) between the DCT coefficient pair in corresponding RDCT subbands, such as parent-child pair coefficient, brother-child pair coefficient, and cousin-child pair coefficient. Furthermore, a frequency ratio descriptor (FRD) computed in the RDCT domain is employed to measure the energy distribution among different frequency components. It can be further utilized to simulate the HVS texture masking property. By considering the intra RDCT subband GGD modeling, inter RDCT subband MI values, and the image FRD value, an effective RR IQA is developed. This chapter is organized as follows. The relationships of intra and inter RDCT subbands are presented in Section 4.3. The DCT reorganization strategy is described in Section 4.2. Section 4.4 discusses the RR feature extraction. And the quality analysis in the receiver side is introduced in Section 4.5. Finally, Section 4.7 concludes this chapter.

4.2 Reorganization Strategy of DCT Coefficients

Since the HVS is more sensitive to luminance than chrominance [177], the proposed image quality metric and the others used for comparison work with luminance only. Color inputs will be converted to gray scale before further analysis. As natural images can be viewed as smooth regions delimited by edge discontinuities, after block-based DCT the image energy of smooth regions is compacted into the DC coefficients, and

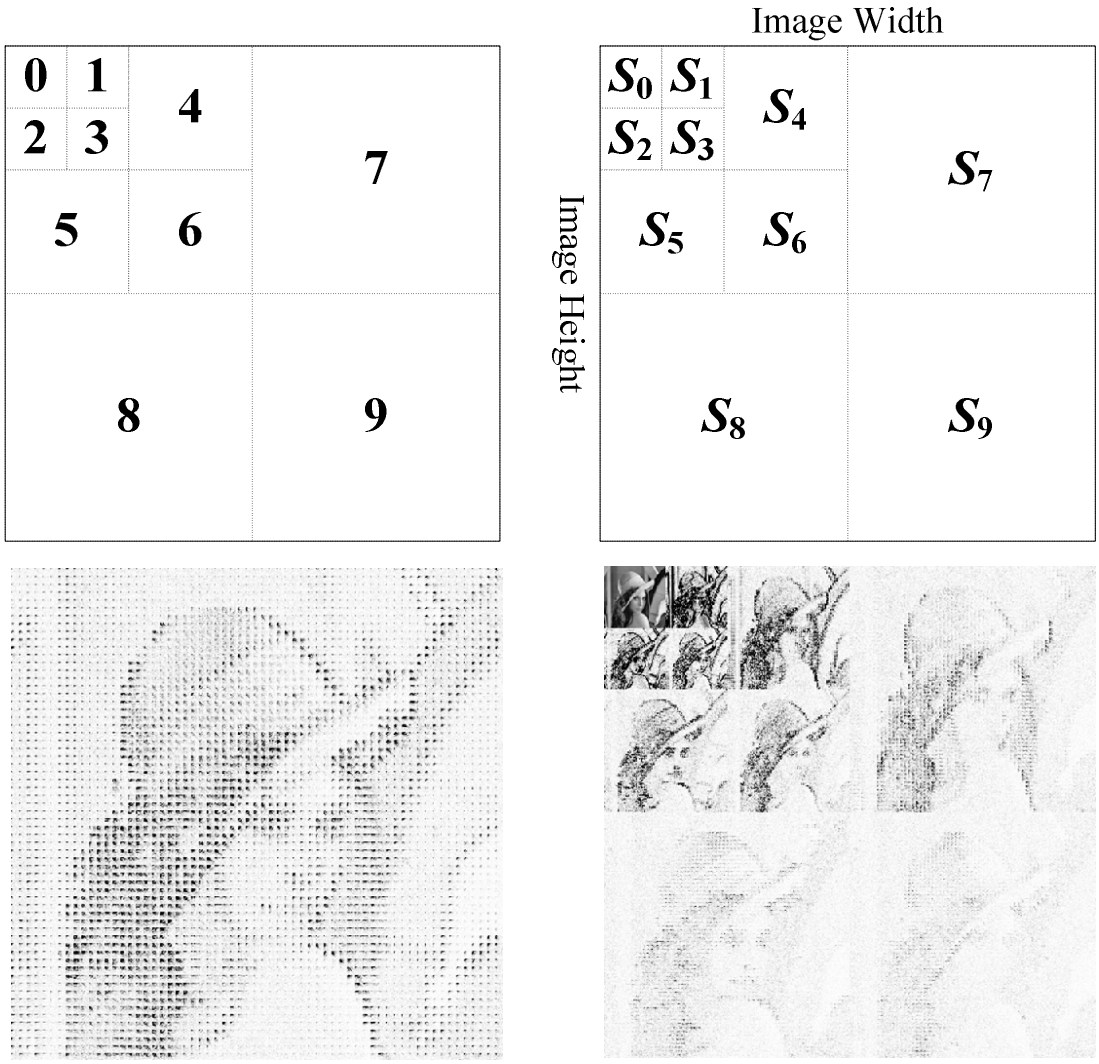


Figure 4.1: *Reorganization strategy of DCT coefficients. Top left: one 8×8 DCT block with ten subband decomposition; top right: the reorganized DCT image representation taken as a three-level coefficient tree; bottom left: 8×8 DCT representation of Lena image; bottom right: the RDCT representation of Lena image. (For better visualization, the DC components are rescaled to integers between 0 and 255, while the AC coefficients are obtained by* $255 - (5 \times |AC|)$*.)*

some high-frequency AC coefficients. For edges, only a small number of high-frequency AC coefficients contribute to its energy. Also the coefficients obtained by the block-based DCT exhibit high correlations, which can be employed for depicting the image degradation level. In [155], after the 8-tap DCT, the second DCT coefficient and the third/fourth DCT coefficient are related to each other as the parent and child bands of the wavelet transform. In [76], the Laplacian probability density function (pdf) is employed to model the coefficient distribution of each DCT subband. The fitted Laplacian pdf parameter λ of one DCT subband can be linearly predicted by

the λ values of the neighboring upper and left DCT subbands. Therefore, although DCT has decomposed the spatial image block into different frequency components, the relationship between the related DCT subbands still exists. In order to utilize the identical nature of the neighboring coefficient distributions, the reorganization strategy [178] [179] is employed to compose the block-based DCT coefficients into a three-level tree structure, as demonstrated in Figure 4.1.

For the subbands **0**, **1**, **2**, and **3**, each subband only contains one DCT coefficient. For the subbands **4**, **5**, and **6**, each subband contains a 2×2 DCT coefficient matrix. For the subbands **7**, **8**, and **9**, each subband contains a 4×4 DCT coefficient matrix. After the decomposition, the same subbands of all the 8×8 DCT blocks are grouped and organized together according to their corresponding positions, as shown in Figure 4.1 (top left). In this manner, the block-based DCT coefficients are reorganized into a three-level coefficient tree. In Figure 4.1 (top right), S_n denotes the grouped subband of all the DCT coefficients lying on the position denoted by n. For example, S_7 is the reorganized subband by grouping the 4×4 DCT coefficient matrix lying on the position **7** of all the 8×8 DCT blocks. An example of the reorganization of the Lena DCT coefficient image is illustrated in Figure 4.1. The 8×8 DCT representation is obtained by applying the non-overlapped 8×8 block based DCT, as shown in Figure 4.1 (bottom left). The reorganized DCT (RDCT) representation is shown in Figure 4.1 (bottom right). It can be observed that the RDCT representation appears like a wavelet representation, *i.e.*, exhibiting structural similarities between subbands, and coefficient magnitude decaying toward high-frequency subbands. Moreover, the RDCT representation is more efficient for the RR quality metric design than the wavelet representation, such as the steerable pyramid [180] [181], even though the wavelet directly has an access to the oriented subbands.

4.3 Relationship Analysis of Intra and Inter RDCT subbands

The statistical relationships between RDCT coefficients are examined in the following ways. Firstly, consider the parent-child coefficient pair representing the information at adjacent scale subbands of the same orientation (e.g. S_4 and S_7). Each parent coefficient in the subband S_4 corresponds to four child coefficients in the subband S_7, as illustrated in Figure 4.2. In order to exploit the underlying statistics, the joint histogram

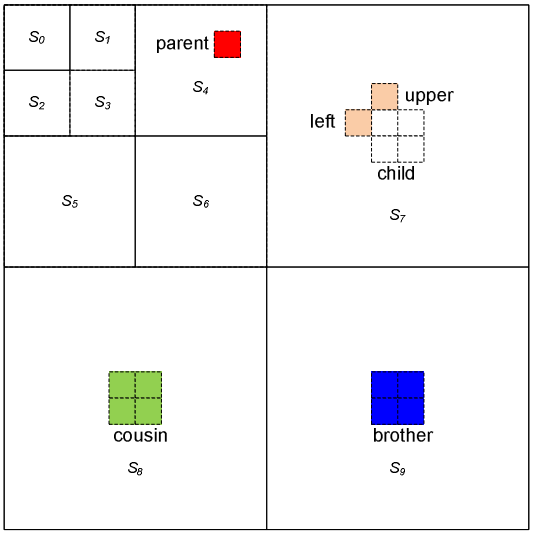


Figure 4.2: *Statistical correlation between inter RDCT subbands. Each parent coefficient in the coarser scale RDCT subband corresponds to four child coefficients in the finer scale subband. Each child coefficient corresponds to one cousin/brother coefficient in the same scale subbands of different orientations.*

of the coefficient pair (parent, child) is built, which is gathered over the spatial extent of the image. Figure 4.3 (a) shows the conditional histogram $h(child|parent)$, which is simply calculated by counting the child coefficients in the subband S_7 conditioned on the coarser-scale subband S_4. Several important aspects can be observed from the conditioned histogram. These coefficients are approximately second-order decorrelated, as the value of the child coefficient is always zero when the values of parent coefficients are not large enough. Moreover, the standard deviation of the child coefficients highly depends on the value of the parent coefficient. The larger the parent coefficient value, the larger the standard deviation of the child coefficients tends to be, as illustrated by the blue curve in Figure 4.3. In [182] [183], it has been demonstrated that the mean and the standard deviation curves of the conditional histogram can be well fitted by a Student's t model of a cluster of coefficients. Furthermore, although they are decorrelated, the statistical dependency can still be observed between the child and parent coefficients. These dependencies also exist in the wavelet coefficient pairs [184], which cannot be eliminated by the linear transformations. This statistical dependency can be more clearly observed by converting the coefficient value into the log-domain as shown in Figure 4.3 (b). The left part of the conditional histogram $h(\log_2(child)|\log_2(parent))$ concentrates on a nearly horizontal line (shown by the green curve), which means that

the value of $\log_2(child)$ is independent of $\log_2(parent)$ in this area. Actually, natural images are composed of smooth regions which are delimited by edge discontinuities. After performing DCT, most of the image energy is compacted to the low-frequency components, which results in a small amount of energy in the high-frequency components. Therefore, the child coefficient values in the finer RDCT subband tend to be small, especially when the parent coefficient values are not large enough. The right part of the conditional histogram in log-domain presents a nearly linear correlation. It implies that the conditional expectation $\varepsilon(\log_2(child)|\log_2(parent))$ is approximately proportional to $\log_2(parent)$.

Figure 4.3 (c) and (e) show the histograms of the child coefficient conditioned on the brother and cousin coefficient, respectively. Compared with the conditional histogram in Figure 4.3 (a), the child coefficient values vary significantly, which do not present a close scattering around the zero value. When the cousin or brother coefficient value becomes larger, the child value fluctuates more dramatically, which can be observed by the standard deviation values (the blue curve of each figure). The brightness of Figure 4.3 corresponds to the probability. The brighter the area, the larger the corresponding probability is. Compared with Figure 4.3 (a), the brightness of Figure 4.3 (c) and (e) is not so significant. It means that the child coefficient value depends on the cousin/brother coefficient less than the parent coefficient. Furthermore, it can be observed that Figure 4.3 (e) is brighter than Figure 4.3 (c). And the standard deviation curves of Figure 4.3 (e) appear to be more regular than those of Figure 4.3 (c). The observations show that the dependency relationship between child-cousin coefficient pair is closer than that between child-brother coefficient pair. After converting the histograms into log-domain, as shown in Figure 4.3 (d) and (f), the correlations appear much looser. Although the mean value in the log-domain concentrates approximately on a line, the child coefficient values are of great differences. The child coefficients do not present a concentrated distribution (with larger standard deviation values), which makes the probability of each coefficient value to be very small. Therefore, the brightness of Figure 4.3 (d) and (f) can hardly be detected. Although the relationship between child and brother/cousin appears to be much looser than that between child and parent, it is admitted that the dependency does exist however in a very complex way, which is very hard to depict.

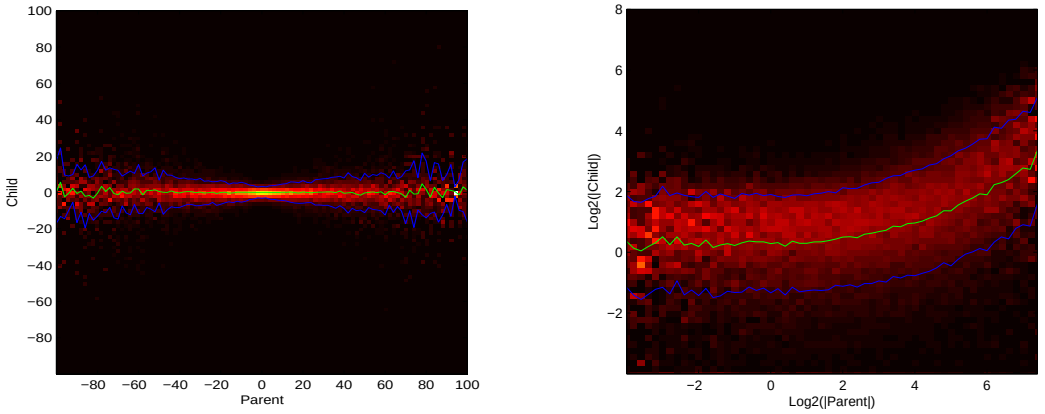


Figure 4.3: *Conditional histogram for the coefficients of the RDCT subbands from the BOAT image. Brightness corresponds to the probability. Each column has been individually rescaled for a better visualization. (a) histogram of the child coefficient conditioned on the parent coefficient; (b) log-domain representation of (a); (c) histogram of the child coefficient conditioned on the brother coefficient; (d) log-domain representation of (c); (e) histogram of the child coefficient conditioned on the cousin coefficient; (f) log-domain representation of (e). The green curve corresponds to $E(child|condition)$, and The blue curves correspond to $E(child|condition) \pm std(child|condition)$, where the condition of each figure is the parent, cousin, and brother, respectively.*

In order to provide a more accurate description about the relationship between RDCT subbands, the mutual information (MI) is employed to describe the dependencies between the child and its condition parent, brother, cousin, upper, and left coefficients, as illustrated in Figure 4.2. As introduced in [176], MI admits the direct data compression and classification interpretations. Let X and Y be two random variables (or vectors) having a joint pdf $p(x, y)$. The MI between X and Y is defined as:

$$\begin{aligned} I(X;Y) &= \int_x \int_y p(x,y) \log \tfrac{p(x,y)}{p(x)p(y)} dxdy \\ &\triangleq E_{XY}\big(\log \tfrac{p(x,y)}{p(x)p(y)}\big) = D\big(p(x,y) \parallel p(x)p(y)\big) \end{aligned} \tag{4.1}$$

where $D(\|)$ is the relative entropy between two distributions, known as the KLD. The MI $I(X;Y)$ indicates how much information Y conveys about X. Therefore, the larger the MI value, the more information is shared by X and Y. Hence, the statistical correlation between X and Y is stronger.

Subband orientation	Inter RDCT subband			Intra RDCT subband	
Horizontal	parent-child (S_4 and S_7)	brother-child (S_9 and S_7)	cousin-child (S_8 and S_7)	upper-child (S_7)	left-child (S_7)
	0.5496	0.2739	0.2908	0.3892	0.3918
Vertical	parent-child (S_5 and S_8)	brother-child (S_9 and S_8)	cousin-child (S_7 and S_8)	upper-child (S_8)	left-child (S_8)
	0.5091	0.2685	0.2908	0.3672	0.3508
Diagonal	parent-child (S_6 and S_9)	brother-child (S_7 and S_9)	cousin-child (S_8 and S_9)	upper-child (S_9)	left-child (S_9)
	0.2974	0.2739	0.2685	0.2165	0.2095

Table 4.1: *Mutual information between the RDCT subbands.*

The MI values between the RDCT subbands are illustrated in Table 4.1. We have provided the MI values of the inter RDCT subbands, such as parent-child S_4 and S_7, brother-child S_9 and S_7, and cousin-child S_8 and S_7, and the intra RDCT subbands, such as upper-child and left-child. In order to provide a more convincing result, each entry gives the average MI value over all the reference images from the LIVE image subjective quality database [25]. Some interesting findings can be observed. Firstly, no matter what the subband orientation is, the MI value of parent-child is the largest. It means that the parent coefficients in coarser subband affect the child coefficients in the

finer subband most, which presents the same property as the wavelet transform. These dependencies have been successfully utilized for the image compression [178] [179]. Secondly, the parent-child MI value of diagonal RDCT subband is much smaller than those of horizontal and vertical ones. The reason is that natural images present much more horizontal and vertical information than the diagonal one. Therefore, most of the DCT coefficients in the diagonal subbands tend to be zero. Furthermore, the MI values somewhat match the HVS property, namely the oblique effect [136], that is, the HVS is more sensitive to the horizontal and vertical frequency components, compared with the diagonal ones. Thirdly, for the horizontal and vertical RDCT subbands, the MI values of intra RDCT subbands are larger than those of brother-child and cousin-child inter RDCT subbands. Therefore, the relationship between neighboring DCT coefficients also exists. This relationship has been further employed for image compression [179] and image quality metric [76]. In [76], the authors employ the neighboring DCT subband relationship to improve the modeling accuracy of the DCT coefficient distribution. Finally, the dependencies between cousin-child and brother-child RDCT subbands can be observed. Although DCT has decomposed the spatial image content into different components with different orientations and frequencies, the dependencies cannot be removed by the linear transformations. Therefore, the correlations between inter RDCT subbands can be exploited for image processing researches, such as compression [184] [185], and so on.

4.4 Reduced Reference Feature Extraction in Sender Side

As discussed above, the RR IQAs aim at evaluating the image perceptual quality based on some RR features extracted from the reference image. In order to design an effective RR IQA, the features extracted should be sensitive to the distortions related to the HVS perception property, and efficient for representation. Therefore, the RR features are critical to the RR IQA performances. Based on the analysis in the above section, the dependencies of intra and inter RDCT subbands do exist, as shown in Table 4.1, which can be depicted and quantified in the receiver side. And these dependencies are expected to be sensitive to the distortions, which can be utilized as the RR features for the quality analysis in the receiver side. Figure 4.4 provides the framework of extracting the RR features from the reference image. For intra RDCT subband,

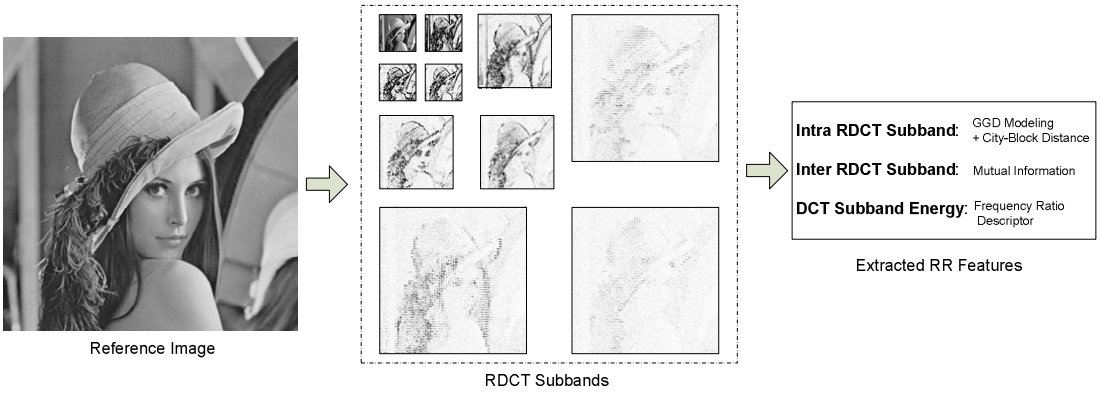


Figure 4.4: *RR feature extraction in the sender side.*

the GGD modeling together with the city-block distance (CBD) is employed to characterize the relationship. For the inter RDCT subband, MI is employed to depict the correlation. And the frequency ratio descriptor (FRD) is used to calculate the RDCT subband energy distribution. Detailed information will be introduced in the following sub-sections.

4.4.1 Intra RDCT Subband Modeling

It has been claimed [167] [168] that the wavelet coefficient distributions of natural images are highly kurtotic (with a sharp peak at zero and a fat-tail distribution). Based on a strict mathematical analysis, Lam *et al.* [186] pointed out that the high-frequency DCT coefficients also follow the kurtotic distribution, which a GGD usually fits well. The probability density function of GGD is defined as:

$$p_{\alpha,\beta}(x) = \frac{\alpha}{2\beta\Gamma\left(\frac{1}{\alpha}\right)} exp\Big\{ - \left(\frac{|x|}{\beta}\right)^{\alpha} \Big\} \tag{4.2}$$

where α models the width of the PDF peak (standard deviation), while β is inversely proportional to the decreasing rate of the peak. α and β are also referred to as the scale and shape parameters, respectively. Γ is the Gamma function given by:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} exp(-t) dt \tag{4.3}$$

Therefore, we can see that only two parameters are needed to completely define each GGD model. However, if the coefficient distributions of all the DCT subbands are to be modeled, too many parameters are needed for the RR IQA. Considering the 8×8

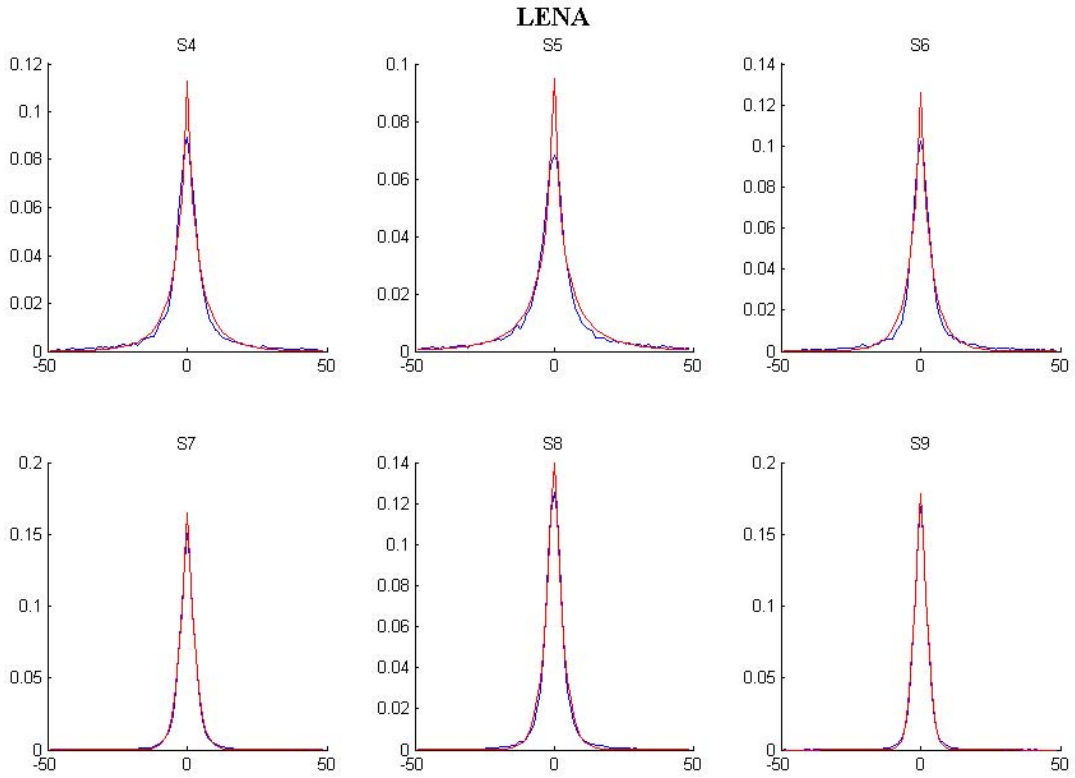


Figure 4.5: *Coefficient distribution (blue line) and the fitted GGD curve (red line) of the RDCT from S_4 to S_9.*

DCT as an example, if all the DCT subbands are to be depicted, there are at least $63 \times 2 = 126$ parameters. It is too large and conflicts with the purpose of RR IQA, which requires less reference information for the quality assessment. In order to reduce the RR data rate and further utilize the identical nature of the coefficient distribution between adjacent DCT subbands, the aforementioned reorganization strategy is employed to group the DCT coefficients into fewer representative RDCT subbands.

After the reorganization process, the number of RDCT subbands containing AC coefficients is reduced to 9, which is more reasonable for RR IQA. The GGD model is employed to model the coefficient distribution of each RDCT subband. The DCT coefficient distribution (blue line) and the fitted GGD curve (red line) of the reorganized subbands S_4-S_9 from the Lena image are illustrated in Figure 4.5. It can be observed that the two curves overlap with each other, which means that the GGD model can efficiently depict the coefficient distributions of the RDCT subbands. Applying this process, we not only model the DCT coefficient distribution, but also exploit the identical nature of the coefficient distributions between adjacent DCT subbands, which will

help improve the RR IQA performance.

It has been shown that the GGD model provides an efficient way to represent the coefficient histogram for each RDCT subband of the reference image. Therefore, for each GGD model, two parameters $\{\alpha, \beta\}$ are needed for the RR IQA. In order to further improve the GGD modeling precision, another parameter denoted as the prediction error is introduced. As we have discussed before, the KLD is asymmetric, which is not suitable for measuring the visual quality distance between the two images. Therefore, the city-block distance (CBD) between two distributions p and $p_{\alpha,\beta}$ is proposed to measure their differences:

$$d_{CBD}(p, p_{\alpha,\beta}) = \sum_{i=1}^{h_L} |p(i) - p_{\alpha,\beta}(i)| \tag{4.4}$$

where p is the histogram distribution of the actual RDCT subband, $p_{\alpha,\beta}$ is the fitted GGD curve, and h_L is the total number of the histogram bins. From the definition, it can be observed that d_{CBD} is symmetric, which means that $d_{CBD}(p, p_{\alpha,\beta}) = d_{CBD}(p_{\alpha,\beta}, p)$. Therefore, compared with KLD, CBD is symmetrical to capture the visual distance between two images, which is reasonable for evaluating image perceptual quality. According to the oblique effect [136] of the HVS, human eyes present similar sensitive values to the horizontal and vertical information, while less sensitive to the diagonal information. Therefore, in order to reduce the RR data rates, only three horizontal RDCT subbands, specifically S_1, S_4, and S_7, are employed for GGD modeling and CBD calculation to extract the RR features.

4.4.2 Inter RDCT Subband Modeling

Referring to Table 4.1, the dependencies between inter RDCT subbands also exist. MI as defined in Eq. 4.1 is employed to capture the corresponding dependencies, which can be further expressed as:

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= E_X\big(-\log_2(p(x))\big) - E_{XY}\big(-\log_2(p(x|y))\big) \end{aligned} \tag{4.5}$$

where $h(X)$ and $h(X|Y)$ denote the entropy of X and X conditioned on Y, respectively. As shown in Eq. 4.5, we can observe that the MI is symmetric and non-negative. If X and Y are independent, the MI is equal to zero. While if X is a function of Y, $I(X;Y) =$

∞. Actually, the MI $I(X;Y)$ indicates how much information Y conveys about X. It admits a well-known data compression interpretation: coding X to a precision ΔX costs $h(X) - \log_2(\Delta X)$ bits, based on the assumption that ΔX is sufficiently small. If Y is known, by considering the same encoding precision ΔX, the total bits cost for encoding X is equal to $h(X) - \log_2(\Delta X) - I(X;Y)$ bits [176]. Therefore, the total saving bits by introducing Y is $I(X;Y)$.

The MI value is introduced in the sender side to describe the essential relationship between inter RDCT subbands, which is changed by the introduced distortion. As we employed three horizontal RDCT subbands (S_1, S_4, and S_7) for GGD modeling and CBD calculation to depict the intra RDCT relationship, the MI values between these horizontal RDCT subbands and other related ones are computed as the RR features to depict the inter RDCT dependencies. These RR features include two MI values to depict the parent-child correlation between the RDCT subband pairs (S_1, S_4) and (S_4, S_7), three MI values to depict the cousin-child correlation between the RDCT subband pairs (S_2, S_1), (S_5, S_4), and (S_8, S_7), and three MI values to depict the brother-child correlation between the RDCT subband pairs, (S_3, S_1), (S_6, S_4), and (S_9, S_7). Therefore, there are 8 MI values in total extracted to capture the inter RDCT subband dependencies.

4.4.3 Image Frequency Feature

Furthermore, in order to accurately represent the reference image characteristic, an image-level feature, specifically the frequency ratio descriptor (FRD), is proposed by considering the HVS properties. For our RR feature extraction, after performing 8×8 DCT, the coefficients are reorganized into several RDCT subbands, as illustrated in Figure 4.1. The frequency ω_{ij} of the $(i,j) - th$ subband for each 8×8 DCT block can be obtained by [124]:

$$\begin{aligned} \omega_{ij} &= \tfrac{1}{2N}\sqrt{(i/\theta_x)^2 + (j/\theta_y)^2} \\ \theta_\Delta &= 2 \times arctan\Big(\tfrac{\S}{2\times l}\Big), \quad (\S = x, y) \end{aligned} \tag{4.6}$$

where N is the dimension of the DCT block (in this study, $N = 8$), θ_x and θ_y are the horizontal and vertical visual angels of a pixel. l is the viewing distance and § stands for the display width/length of a pixel on the monitor. According to the international

standard ITU-R BT.500-11 [5], the ratio of viewing distance to picture height should be a fixed number between 3 and 6. Moreover, for most of the displays, pixel aspect ratio (PAR) is equal to 1. It means that the horizontal and vertical visual angles (θ_x, θ_y) are identical:

$$\theta_x = \theta_y = 2 \times arctan\left(\frac{1}{2 \times R_{vd} \times H_{pic}}\right) \tag{4.7}$$

where R_{vd} is the ratio of viewing distance to picture height. H_{pic} is the number of pixels in picture height. The frequency values obtained by Eq. 4.6 and the spatial contrast sensitivity function (CSF) values [124] of the 8×8 DCT subbands are illustrated in Figure 4.6. It can be observed that the adjacent DCT subbands present similar frequency and CSF values. The lower the frequency component, the larger is the CSF value. After the reorganization process introduced in Section 4.2, the CSF values of the RDCT subbands S_0, S_1, S_2, and S_3 are larger than 0.5, which are the most sensitive components to the HVS. By checking the frequency ω_{ij} value, we can find that the frequency values of these RDCT subbands are smaller than 5. Therefore, these RDCT subbands S_0, S_1, S_2, and S_3 (denoted by the red box) are regarded as the low frequency (LF) components. For the RDCT subbands S_4, S_5, and S_6, the CSF values (except the one of ω_{33}) are larger than 0.2 and smaller than 0.5. And the frequency values (except ω_{33}) are larger than 5 and smaller than 12. These RDCT subbands S_4, S_5, and S_6 (indicated by the blue box) are viewed as the medium frequency (MF) components with medium sensitivity values. The rest of the RDCT subbands S_7, S_8, and S_9 (denoted by the green box) present the lowest sensitivity values and the highest frequency values larger than 12, which are regarded as the high frequency (HF) components.

The introduced distortion will not only change the histogram distribution in each RDCT subband and dependencies between adjacent RDCT subbands, but also alter frequency components of the image. For example, if JPEG is utilized to code the reference image, the blocking and ringing artifacts will appear as a result of frequency coefficient truncation. As the quantization steps of the HF components are higher than the LF ones, the HF components will be degraded more seriously than LF ones. Here, the image-level feature FRD is proposed by considering the ratio information between the LF, MF, and HF components. The FRD can be efficiently computed in the RDCT

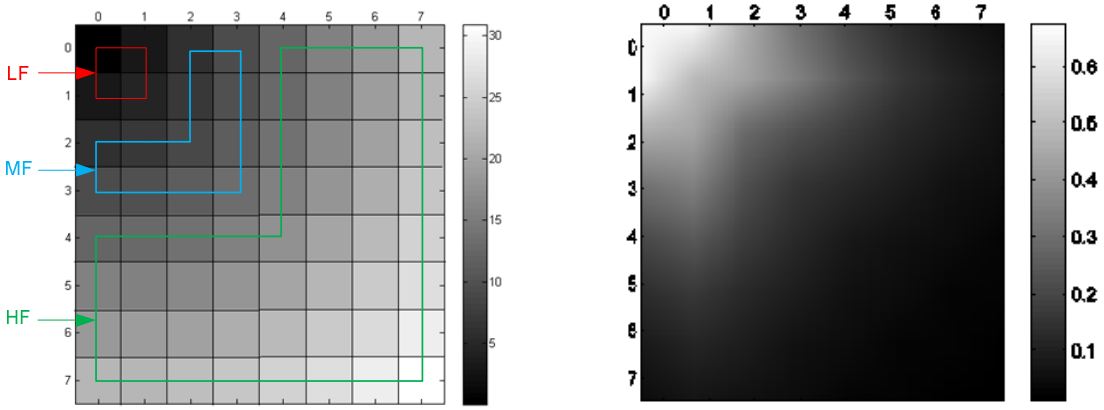


Figure 4.6: *Frequency ω_{ij} and the spatial contrast sensitivity function (CSF) value of each DCT subband. Left: frequency ω_{ij} value; right: spatial CSF value.*

domain, which is defined as:

$$FRD = \frac{M_{value}+H_{value}}{L_{value}} \tag{4.8}$$

where L_{value}, H_{value}, and M_{value} represent the sums of the absolute DCT coefficient values in the LF (S_0, S_1, S_2, and S_3), MF (S_4, S_5, and S_6) and HF (S_7, S_8, and S_9) RDCT subbands, respectively. The FRD can help to capture the proportion of frequency changes caused by the distortions. Furthermore, the larger the value of FRD, the more energy the MF and HF components possess. It means that the DCT block is more likely to contain texture information. For the plain block, the energy mostly concentrates in the LF components. For the edge block, there will be only a small number of DCT coefficients in the HF group. Consequently, the texture block will present higher FRD. As discussed in the JND models [119] [122] [124], the texture block can tolerate more distortions than the plain and edge block, which is interpreted as the texture masking property of the HVS. Therefore, the proposed FRD can be employed to simulate the texture masking property for the derivation of the final image quality metric.

As discussed above, there are total 3 parameters $\left(\alpha, \beta, d_{CBD}(p, p_{\alpha,\beta}\right)$ to depict the histogram distribution of each RDCT subband. Considering the HVS oblique effect, only the 3 horizontal subbands are included, which results in 9 parameters. For the inter RDCT subband relationship, 8 MI values in total are introduced to depict the parent-child, cousin-child, and brother-child relationships. For the frequency distribution, only one parameter named as FRD is extracted from the reference image. Therefore, the

proposed RR method employs $9 + 8 + 1 = 18$ parameters to represent the reference image. By comparing them with the ones extracted from the distorted image, the perceptual quality can be analyzed.

4.5 Perceptual Quality Analysis in the Receiver Side

In the receiver side, we need to compare the extracted features to analyze the perceptual quality of the distorted image. The parameters are extracted from intra RDCT subbands, inter RDCT subbands, and image-level feature FRD.

4.5.1 Intra RDCT Feature Difference Analysis

In the receiver side, for each distorted image, the aim is to compute the CBD between the coefficient distributions of RDCT subbands from the original image p and the distorted image p_d, respectively:

$$d_{CBD}(p, p_d) = \sum_{i=1}^{h_L}(p(i) - p_d(i)) \tag{4.9}$$

However, the coefficient distributions of the original image are unavailable. Therefore, we employ the fitted GGD model and the prediction error to approximate the CBD between p and p_d. The inequality property:

$$\begin{aligned} &\textstyle\sum_{i=1}^{h_L} |p_{\alpha,\beta}(i) - p_d(i)| - \sum_{i=1}^{h_L} |p(i) - p_{\alpha,\beta}(i)| \\ &\leq \textstyle\sum_{i=1}^{h_L} |p(i) - p_d(i)| \\ &\leq \textstyle\sum_{i=1}^{h_L} |p_{\alpha,\beta}(i) - p_d(i)| + \sum_{i=1}^{h_L} |p(i) - p_{\alpha,\beta}(i)| \end{aligned} \tag{4.10}$$

implies that $d_{CBD}(p, p_d)$ is bounded by:

$$d_{CBD}(p_{\alpha,\beta}, p_d) - d_{CBD}(p_{\alpha,\beta}, p) \leq d_{CBD}(p, p_d) \leq d_{CBD}(p_{\alpha,\beta}, p_d) + d_{CBD}(p_{\alpha,\beta}, p) \tag{4.11}$$

Here, we employ the lower bound denoted as $\acute{d}_{CBD}(p, p_d)$ to approximate the CBD between p and p_d:

$$\acute{d}_{CBD}(p, p_d) = d_{CBD}(p_{\alpha,\beta}, p_d) - d_{CBD}(p_{\alpha,\beta}, p) \tag{4.12}$$

For the distorted image, we need not fit p_d to a GGD model, which is not appropriate for the distorted images. What we compute is the distance $d_{CBD}(p, p_d)$ between the

fitted GGD of the reference image and the coefficient distribution of the distorted image according to Eq. 4.9. By considering the prediction error of GGD modeling, we can obtain the approximated distance according to Eq. 4.11 to analyse the intra RDCT relationship changes, which can further help to depict the perceptual quality.

4.5.2 Inter RDCT Feature Difference Analysis

For the inter RDCT subband, the differences between the corresponding MI values of the adjacent RDCT subbands are calculated:

$$d_{MI}(S_m, S_n) = I(S_m, S_n) - \hat{I}(S_m, S_n) \tag{4.13}$$

where $I(S_m, S_n)$ is the MI of the RDCT subband S_m and S_n in the reference image, and $\hat{I}(S_m, S_n)$ is the MI of the S_m and S_n in the distorted image. In this way, the inter RDCT subband relationship is captured, which can help to depict degradation level of the perceptual quality.

4.5.3 Image Frequency Feature Difference Analysis

For the image frequency, as the distortion will degrade the HF, MF, and LF components differently, the FRD distance can effectively represent the frequency component changes:

$$FL = |FRD_{ori} - FRD_{dist}| \tag{4.14}$$

where FRD_{ori} is the original feature, FRD_{dist} is calculated from the distorted image, and FL denotes the frequency information change. As discussed before, FRD can represent how much texture information the image contains. Therefore, it can help to simulate the texture masking property of the HVS. Furthermore, as discussed in [187] [188], for the content of the original image and the artifacts, one's presence will affect the visibility of the other. Therefore, a novel mutual masking strategy is proposed by considering the FRD values of both the original and distorted image:

$$FL_v = \begin{cases} \frac{FL}{FL+FRD_{ori}}, & FRD_{ori} < FRD_{dist} \\ \frac{FL}{FL+FRD_{dist}}, & FRD_{ori} \geq FRD_{dist} \end{cases} \tag{4.15}$$

where FL_v is the final HVS-related features to depict the frequency information change. FL in the denominator is employed to scale FL_v into the range $[0, 1]$. When an image

containing texture information is smoothed by the distortions, such as JPEG compression and blur, the detailed texture information cannot be perceived by the HVS. Therefore, no visual masking effect should occur. Also if a smooth image is distorted to be highly textured by the distortion, such as additive Gaussian noise and fast-fading in the LIVE image subjective quality database [25], only the noise can be perceived from the degraded image. In this case, there should be no visual masking effect either. This phenomenon is named as the mutual masking [189]. In [188], the mutual masking effect is determined by the minimum value of the thresholds calculated from the original and distorted image. In this study, as the computed FRD value can depict the texture information of the image, we employ Eq. 4.15 to depict the mutual masking effect of the HVS perception, where the smaller value of FRD_{ori} and FRD_{dist} is employed to model the masking effect. In this way, only the image is highly textured in both the reference and distorted images (large FRD_{ori} and FRD_{dist} values) can produce a significant masking effect. In other cases, an insignificant masking effect will be introduced, as expressed in Eq. 4.15.

Now we have obtained the CBD values of the intra RDCT subbands, MI difference values of the inter RDCT subbands, and the FL_v value depicting the image frequency information change. How to combine them together for developing an effective RR IQA needs to be considered. Here, a simple linear combination method is employed to obtain the final quality values:

$$Q = par_1 \times \sum_{sub} \acute{d}_{CBD}(p^{sub}, p_d^{sub}) + par_2 \times \sum_{(m,n)} d_{MI}(S_m, S_n) + par_3 \times FL_v \tag{4.16}$$

where (par_1, par_2, par_3) are the three weighting parameters to be determined, Q is the perceptual quality index of the distorted image. Firstly, we sum together the CBD values of intra RDCT subbands, and MI differences of inter RDCT subbands, respectively. Their sum values and the FL_v value are further linearly combined together according to Eq. 4.16. In order to find the optimal parameters (par_1, par_2, par_3), the genetic algorithm [190] is employed to train them on several distorted images. In this study, four reference images and their corresponding distorted images in the LIVE image quality assessment database [25] are employed to obtain the three parameter values. The selected four reference images for parameterization are 'rapids', 'paintedhouse',

'plane', and 'building2'. The correlation between the DMOS values and the calculated Q values in Eq. 4.16 of the training images is maximized to determine the optimized parameters (par_1, par_2, par_3). As there are only 3 parameters to be determined, the number of the genes is equal to 3. Each gene uses 8-bit binary representation. Each gene will be divided by 255 to constrain them within $[0, 1.0]$. The generation gap is set as 0.9, which means that only $30 - 30 \times 0.9 = 3$ best fitted genes will be propagated to the successive generation. Therefore, 27 new genes will be produced at each generation. The crossover for creating new genes is a single-point with probability 0.7. And the mutation for creating new genes is with probability 0.0014. The number of generations is set to 100. The initial population is created randomly and uniformly distributed. The fitness assignment is based on ranking instead of raw performance. Selection method is stochastic universal sampling. Reinsertion is fitness-based (instead of uniform random). After performing the genetic algorithm, the parameterization result is par_1=0.4883, par_2=0.0313, and par_3=0.6719. Furthermore, as in [167] [168], a logarithm process is employed to scale the perceptual quality index Q:

$$VQI = \log_{10}\left(1 + \frac{Q}{D_0}\right) \tag{4.17}$$

where VQI is the final obtained quality score, D_0 is utilized for scaling the distortion measure to avoid the variation of Q being too small. It just helps to depict the perceptual quality index clearly, which will not influence the performance of the proposed RR IQA. In this study, D_0 is set as 0.0001 for simplicity.

4.6 Experimental Results

In this sub-section, we firstly show the efficiency of the reorganized DCT strategy for the proposed method. Subsequently, the performances of different IQAs will be compared to demonstrate the efficiency of the proposed RR IQA for evaluating the image perceptual quality.

4.6.1 Efficiency of the DCT Reorganization Strategy

All the reference images from the LIVE image database [25] are employed to demonstrate the efficiency of the DCT reorganization strategy, compared with the steerable pyramid [167] [168], which has been employed in the FR IQAs, such as VIF [72]. As

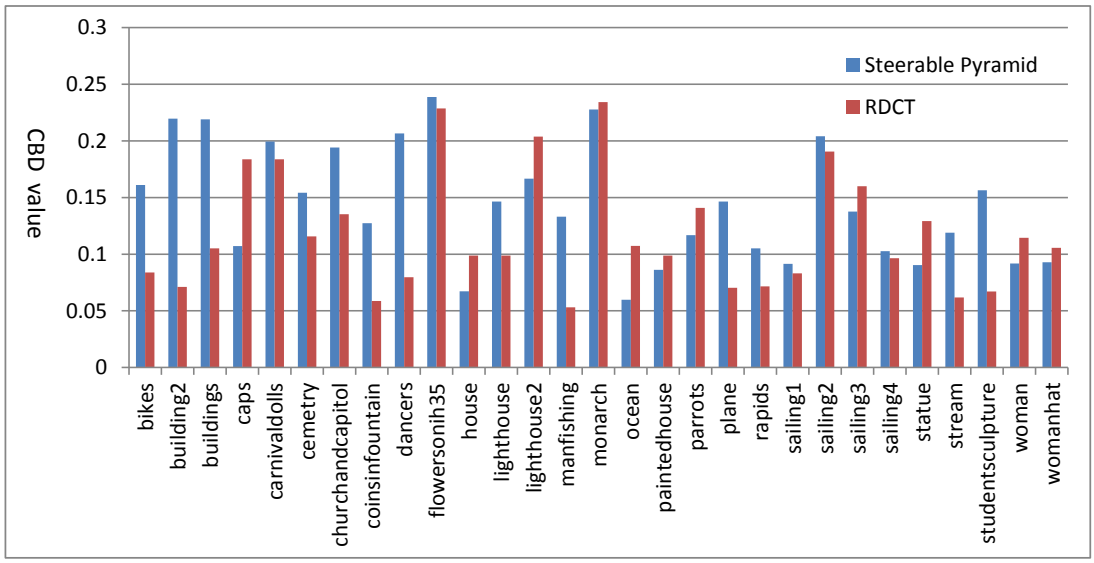


Figure 4.7: *Prediction error of the reference images in the LIVE image database [25].*

illustrated in [167] [168], after the 3-scale, 3-orientation steerable pyramid decomposition, the high-frequency subbands correspond to the reorganized DCT subbands from S_1 to S_9. As described in Section 4.4.1, the average prediction error, specifically the CBD, between the fitted GGD function and the actual coefficient distribution of the 6 subbands (from S_4 to S_9), are employed as the criterion to evaluate the performances of different transforms. According to the definition in Eq. 4.4, the smaller the prediction error, the better fitting is the GGD function, which means that the GGD can more accurately describe the coefficient distribution. The prediction error of each reference image in the LIVE image database is illustrated in Figure 4.7. It can be observed that for most images the prediction errors using the reorganized DCT are smaller than those using the steerable pyramid. The average prediction error using the reorganized DCT of all the images is only 0.1183, compared with 0.1664 using the steerable pyramid. This result means that the coefficient distributions of the reorganized DCT subbands are more suitable for GGD modeling, which will further help improve the RR IQA performance.

4.6.2 Performance of the Proposed RR IQA

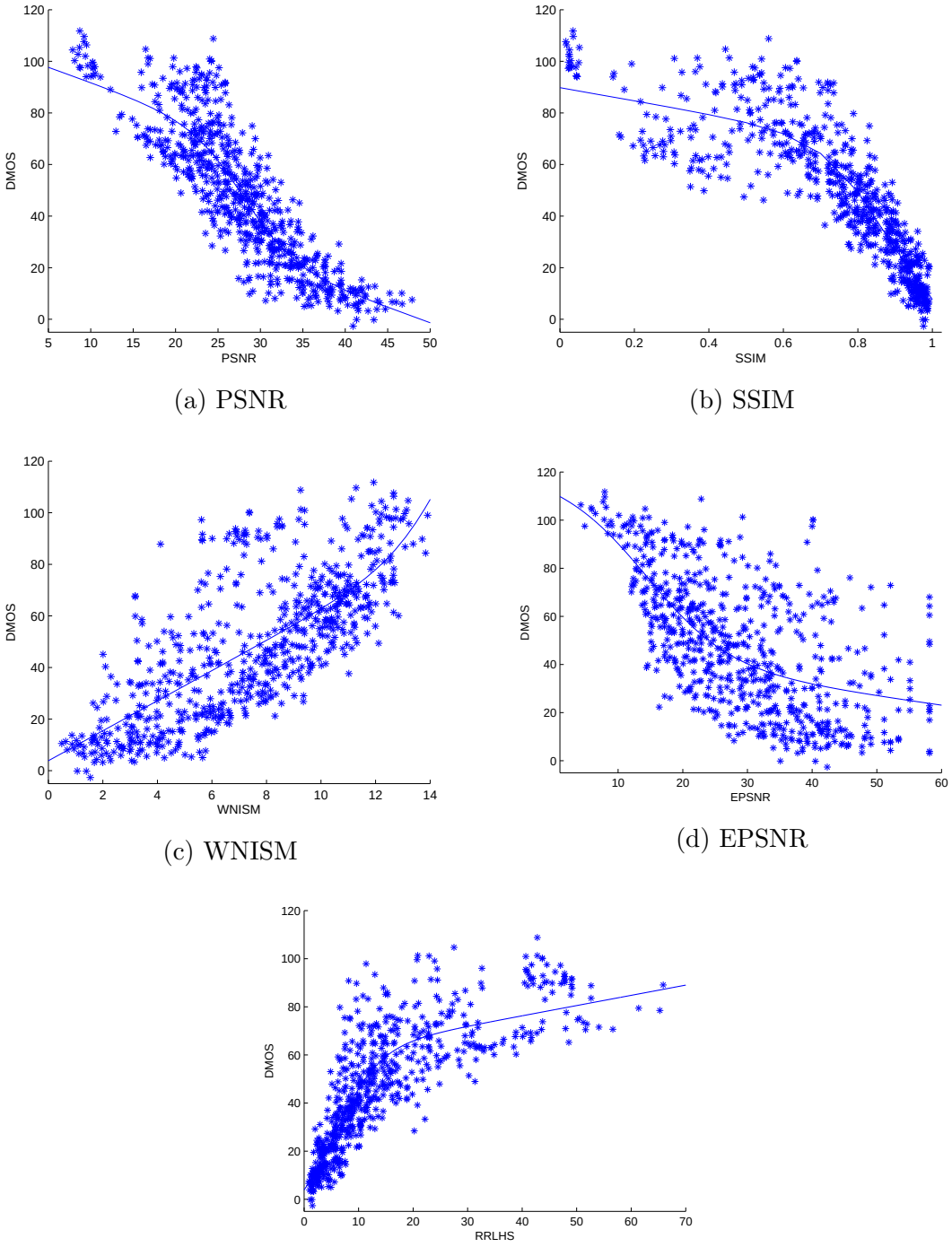
We compare the performance of our proposed RR IQA with the representative RR image quality metric: WNISM [167] [168], recently developed RR-LHS [165], EPSNR [164], RR-DNT [169], RR-SSIM [172] [173], RR-Weibull [174], RR-Edge [175], and the

FR metrics: PSNR, and SSIM [68]. The LIVE image database [25] (excluding the distorted images generated from the four training reference images), the IRCCyN/IVC image database [26], and the MICT image database [27] are employed to compare the performances of these metrics. Detailed information of these image databases can be referred to Section 1.3.3. As usual, three statistical measurements LCC, SROCC, and RMSE are employed to evaluate the corresponding performances of these metrics.

Database	Method	LCC	SROCC	RMSE	RR feature number	RR data rate
LIVE	PSNR	0.8759	0.8813	13.157	-	-
	SSIM	0.9041	0.9112	11.653	-	-
	WNISM	0.7585	0.7709	17.771	18	162 bits
	EPSNR	0.6571	0.6257	20.559	30	270 bits
	RR-LHS	0.8809	0.8831	12.909	320	2560 bits
	RR-Weibull	0.8567	0.8650	14.475	6	*
	RR-Edge	0.8613	0.8908	14.256	12	96 bits
	Proposed	**0.9309**	**0.9279**	**9.965**	**18**	**153 bits**
IRCCyN/IVC	PSNR	0.7037	0.6791	0.866	-	-
	SSIM	0.7758	0.7778	0.769	-	-
	WNISM	0.4525	0.4094	1.087	18	162 bits
	EPSNR	0.3947	0.3958	1.119	30	270 bits
	RR-LHS	0.8078	0.8203	0.718	320	2560 bits
	RR-DNT	0.6316	0.6099	0.9446	48	*
	RR-SSIM	0.8177	0.8156	0.7014	36	*
	Proposed	**0.7712**	**0.7649**	**0.776**	**18**	**153 bits**
MICT	PSNR	0.6154	0.5748	0.987	-	-
	SSIM	0.7174	0.7870	0.872	-	-
	WNISM	0.6568	0.6446	0.944	18	162 bits
	EPSNR	0.4016	0.4059	1.146	30	270 bits
	RR-LHS	0.7623	0.7644	0.810	320	2560 bits
	RR-DNT	0.6733	0.6521	0.9253	48	* bits
	RR-SSIM	0.8051	0.8003	0.7423	36	* bits
	Proposed	**0.8282**	**0.8317**	**0.701**	**18**	**153 bits**

Table 4.2: *Performance comparisons of different RR IQAs over different image subjective quality databases. ("-" means that the IQA is an FR metric, where the RR feature number is the pixel number of the image, and the RR data rate is also viewed as the whole image. "*" means that the RR IQA only calculates the number of the features, while the number of the bits for representing the RR parameters cannot be provided.)*

The performances of different IQAs over different image subjective quality databases are illustrated in Table 4.2, where the RR data rate of each IQA is also illustrated. It can be observed that the proposed method can outperform the other RR and FR



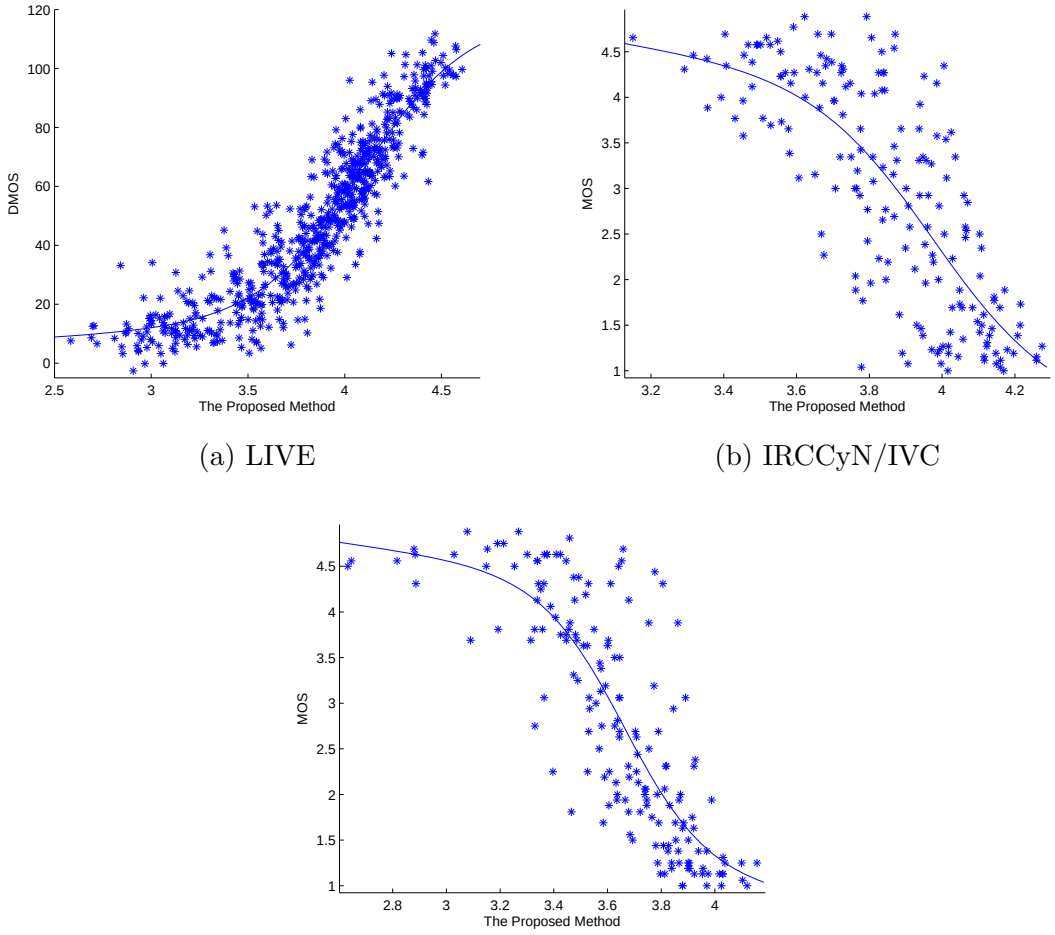
(a) PSNR (b) SSIM

(c) WNISM (d) EPSNR

(e) RR-LHS

Figure 4.8: *Scatter plots of the DMOS values versus model predictions on the LIVE image quality assessment database.*

metrics on the LIVE [25] and MICT [27] image databases, with larger LCC/SROCC and smaller RMSE value. While for the IRCCyN/IVC [26] image database, only the



(a) LIVE

(b) IRCCyN/IVC

(c) MICT

Figure 4.9: *Scatter plots of the DMOS or MOS values versus model predictions on the three image subjective quality databases. Each sample point represents one test image.*

metrics RR-LHS and RR-SSIM can generate better performances. However, these two RR metrics require a much larger bit rate to represent the RR features than the proposed RR metric. From Table 4.2, experimental results demonstrate that PSNR performs badly, although it requires the whole reference image for perceptual quality analysis. The reason is that PSNR only measures the pixel absolute differences, which does not take the HVS property into consideration. For SSIM, the structural distortions are measured rather than the absolute pixel value differences, which are sensitive to the HVS perception. Therefore, SSIM demonstrates a better performance than PSNR. However, SSIM also utilizes the whole reference image for quality analysis, which will introduce a heavy burden for the RR feature transmission. For EPSNR, in order to

reduce the bits to represent the location, the reference and distorted images are firstly cropped to 614×454, where only the central parts are kept. Therefore, as shown in [164], 19 bits are required to encode the location, while 8 bits are needed to represent the pixel value. In our comparisons, 10 pixels together with their locations are employed as the RR features, which require 270 bits in total for representation. The performances of EPSNR over the three databases seem to be the worst. Although the HVS is sensitive to the edges, 10 edge points are not sufficient to accurately represent the image perceptual quality. If more edge pixels are included, the performance will be better. In that case, a heavy burden for transmitting the RR features will be introduced.

For RR-Weibull and RR-Edge, as the authors only provide the performance results on the LIVE image database, their performances results on the IRCCyN/IVC and MICT image databases are not available for comparison. RR-Weibull extracted 6 scalar parameters from each source image to depict the statistics of the image gradient magnitude. It can generate a better performance than WNISM and EPSNR. However, as the number of the RR features is very small, which may not be sufficient to depict the information of the source image, the performance is not good enough, compared with other RR image metrics. RR-Edge further incorporated more RR features to depict the statistics of the edge. In total, 12 RR features are extracted from the source image, which generates a better performance compared with RR-Weibull. However, it is still not good enough. Since RR-LHS considers the motion information to design the RR video quality metric, in our comparisons only the discriminative local harmonic strength in the spatial domain is employed for the RR image quality assessment. Therefore, about 320 elements of each image are extracted as the RR features. If 8 bits are employed to encode each element, the RR data rate is 2560 bits. It is a high burden for the RR data transmission. For RR-DNT and RR-SSIM, the performance results on the three image databases are illustrated in [173].

All the 779 distorted images in the LIVE image database are employed to demonstrate the performances of RR-DNT and RR-SSIM. However, as the proposed method utilized four reference images and the corresponding distorted images for training the parameters, it is not fair to compare RR-DNT and RR-SSIM with the proposed method on the LIVE database. Therefore, only the performance comparisons on IRCCyN/IVC

and MICT image databases are illustrated in Table 4.2. RR-DNT employs the divisive normalization to depict the coefficient distributions of the wavelet subbands. The distribution difference between the reference and distorted images is used to depict the image perceptual quality. However, a training process is utilized to determine the 5 parameters in RR-DNT. And the performances of RR-DNT seem to be sensitive to these parameters. That is the reason why RR-DNT performs very well over the LIVE image database, while performs poorly over the IRCCyN/IVC and MICT image databases. RR-SSIM extracted the statistical features from a multi-scale, multi-orientation divisive normalization transform. By following the philosophy in the construction of SSIM, a distortion measurement is developed to estimate the SSIM index of the distorted image. As a linear relationship between the RR-SSIM and SSIM has been discovered, the performances of RR-SSIM are good, which is comparable with the proposed method. RR-SSIM outperforms the proposed metric on the IRCCyN/IVC database, while its performance is worse than the proposed one on the MICT database. However, RR-SSIM extracted 36 RR features to represent the source image, which is twice of that extracted by the proposed RR metric.

WNISM [29] [30] is proposed in the wavelet domain by depicting the marginal probability distribution of each wavelet subband. The steerable pyramid is firstly employed to decompose the image into several wavelet subbands, whose coefficient distributions are modeled by GGD. As demonstrated in Section 4.6.1, GGD can more accurately model the coefficient distribution of RDCT subband than that of the steerable pyramid. Moreover, KLD is utilized in WNISM to depict the histogram distribution distance. However, KLD is asymmetric, which is not suitable for image quality evaluation, because the visual quality distance from one image to another should be identical no matter how it is measured. Those are the reasons why WNISM performs badly over the three image quality databases, as illustrated in Table 4.2.

For the RR data rate, each wavelet subband needs 3 parameters to describe its distribution, which requires $8+8+8+3=27$ bits for representing these parameters. In total 6 wavelet subbands are considered to construct WNISM, which results in $27\times 6=162$ bits to encode all the RR features of the reference image. For the proposed method, the intra RDCT subband relationship is captured by GGD modeling and CBD; the inter RDCT subband correlation is depicted by the MI values; and the frequency

distribution is captured by FRD, which can further simulate the HVS texture masking property. As the three horizontal RDCT subbands are employed for depicting the intra RDCT subband distribution, $(8 + 8 + 8 + 3) \times 3 = 81$ bits are needed to represent the GGD modeling parameters. For the MI and FRD values, 8-bit representation is employed. Therefore, $8 \times 9 = 72$ bits are needed for representing all the MI and FRD values. In total, $81 + 72 = 153$ bits are required to encode all the RR features extracted for the proposed method. The scatter-plots of different IQAs over the LIVE image database are illustrated in Figure 4.8. And the scatter plots of the proposed RR metric on the three image subjective quality database are illustrated in Figure 4.9. It can be observed that the points of the proposed method scatter more closely to the fitted line, compared with other IQAs. It means that the DMOS or MOS values correlate better with the perceptual quality values obtained by the proposed RR IQA.

4.6.3 Performance of the Proposed RR IQA over Each Individual Distortion Type

Furthermore, we tested the proposed RR IQA over individual distortion types from the LIVE image database, which are illustrated in Table. III. It can be observed that PSNR performs well over JPEG2000 and WGN images, especially for WGN images. However, for the JPEG, Blur, and FF noise images, PSNR performs poorly. EPSNR only employs several edge pixels to measure the corresponding PSNR. Therefore, it presents a performance similar to PSNR. For the WGN images, EPSNR demonstrates a very good performance. It means that the perceptual qualities of WGN images correlate closely with the absolute pixel value differences, in contrast to other noise images. RR-LHS demonstrates good performances on the JPEG2000, JPEG, WGN, and FF noise images. However, its performance over the Blur noise images is very poor, even a very large number of RR features has been employed. It means that the discriminative local harmonic strength is not suitable for depicting the perceptual quality of the Blur noise images.

For WNISM, RR-Weibull, and RR-Edge, the experimental results over the individual distortion types are very good. However, their performances degrade significantly when images with different types of distortions are tested together, as shown in Table 4.2. As revealed by the previous literature [169], it is also the main drawback

of WNISM. The experimental results demonstrate that proposed RR metric outperforms WNISM except for the JPEG 2000 distortion. Actually, JPEG 2000 employed the wavelet transform for compression. Therefore, the steerable pyramid employed in WNISM is more suitable for depicting the coefficient distribution than the DCT. Also as new RR features, specifically the inter RDCT subband MI and image FRD FL_v, have been introduced, the performance has been greatly improved. Therefore, the proposed method not only performs very well over individual distortion types, but also provides a good performance across different distortion types. It means that it performs more robustly for evaluating image visual quality. Furthermore, the proposed metric maintains a smaller RR data rate, compared with WNISM, and RR-LHS. The improvements have demonstrated that the intra and inter RDCT subband dependencies and the image FL_v value are helpful for designing an effective RR IQA. It reflects that the CBD difference, MI differences, and FL_v value can help to depict the levels of the introduced distortions. Therefore, for the proposed RR IQA, the RR features for depicting the vertical RDCT subbands are excluded to save some bit rates for the inter RDCT subband MI values and image FRD value.

As illustrated in Table 4.2 and Table 4.3, the effectiveness of our proposed RR quality metric has been clearly demonstrated compared with the other RR metrics or even FR metrics in terms of both performance and required RR data rate. The computational complexities of RR feature extraction and comparison need to be further evaluated. The processing complexity in the sender side is different from that in the receiver side. In the sender side, as introduced in Section 4.4, 8×8 block-based DCT is firstly performed on the source image. After the reorganization strategy, the DCT subbands are grouped into several representative RDCT subbands. The DCT coefficient distribution of each RDCT subband is modeled by the GGD. MI is employed to depict the relationship between different RDCT subbands. Based on the RDCT subband, the image FRD FL_v is calculated.

We implemented the RR feature extraction in Matlab. During our implementation, we did not perform any optimizations. A speed test was performed on our PC with a 3.0GHz Quad CPU and 1.0GB memory. For each source image of LIVE image database, it only requires 2.94s on average to extract the RR features. In the receiver side, as illustrated in Section 4.5, the 8×8 block-based DCT and reorganization strategy was

Method		JPEG2000	JPEG	WGN	Blur	FF
PSNR	LCC	0.9078	0.8942	0.9857	0.7856	0.8880
	SROCC	0.9042	0.8853	0.9850	0.7894	0.8897
	RMSE	10.546	14.406	4.711	11.214	12.898
WNISM	LCC	0.9270	0.8629	0.8791	0.9234	0.9422
	SROCC	0.9211	0.8539	0.8572	0.9290	0.9350
	RMSE	9.434	16.258	13.346	6.955	9.399
EPSNR	LCC	0.6773	0.6489	0.9700	0.4890	0.6129
	SROCC	0.6816	0.6400	0.9670	0.3086	0.5612
	RMSE	18.500	24.280	6.776	15.807	22.167
RR-LHS	LCC	0.8861	0.9761	0.9345	0.6051	0.8569
	SROCC	0.8792	0.9557	0.9848	0.6250	0.8575
	RMSE	11.654	6.995	9.970	14.427	14.462
RR-Weibull	LCC	0.9422	0.9493	0.9771	0.9471	0.9234
	SROCC	0.9415	0.9402	0.9749	0.9404	0.9261
	RMSE	7.912	10.115	5.954	5.817	10.741
RR-Edge	LCC	0.9404	0.9383	0.8815	0.9152	0.9421
	SROCC	0.9406	0.9408	0.8654	0.9083	0.9329
	RMSE	8.592	11.128	13.224	7.302	9.400
Proposed	LCC	0.8983	0.9528	0.9275	0.9459	0.9437
	SROCC	0.8912	0.9520	0.9093	0.9525	0.9204
	RMSE	11.051	9.766	10.471	5.880	9.277

Table 4.3: *Performances of different IQAs over individual distortion types on the LIVE image database*

also performed. But the fitting process of the GGD does not need to be performed. Only the histogram of each RDCT subband was constructed. And MI values between RDCT subbands, and the image FRD FL_v are calculated. Therefore, the computation is faster. The speed test was performed on the same PC, which indicates that only 1.93s per image on average is needed for the image quality analysis. If further optimization is applied, it is believed that the quality analysis in the receiver side can perform even faster.

4.6.4 Statistical Significance

To assess the statistical significance of the performance difference between two metrics, F-test was conducted on the prediction residuals between the metric outputs (after nonlinear mapping) and the subjective ratings. The residuals are supposed to be Gaussian. Smaller residual variance implies more accurate prediction. Let F denotes the ratio between the residual variances of two different metrics (with the larger variance as the numerator). If F is larger than $F_{critical}$ which is calculated based on the number

of residuals and a given confidence level, then the difference between the two metrics are considered to be significant at the specified confidence level. Table 4.4 lists the residual variance of each metric on the three subjective image databases. Notably due to the differences in employed subjective scales, the residual variance varies a lot across different image databases. The $F_{critical}$ with 95% confidence is also shown in Table IV for each database.

In Table 4.5, the proposed metric is compared with the other metrics regarding the statistical significance. In each entry, the symbol "1", "0", or "=" means that on the image databases indicated by the first row of the table, the proposed metric is statistically (with 95% confidence) better, worse, or indistinguishable, respectively, when compared with its competitors indicated by the first column. "*" means that the comparison cannot be performed due to the unavailable result data. For the RR metrics RR-Weibull and RR-Edge, the metric outputs of the distorted images on the IRCCyN/IVC and MICT image databases are not available. Therefore, we cannot compare the statistical significances of these two metrics with the proposed method on these two databases. By referring to the other entry values shown in Table 4.5, it can be observed that the proposed metric outperforms most of its competitors statistically. Although its performance on IRCCyN/IVC image database seems to be equivalent to other IQAs, overall it demonstrates better performances on the other two image databases.

	LIVE(672 images) $F_{critical}$=1.1355	IRCCyN/IVC(185 images) $F_{critical}$=1.275	MICT(168 images) $F_{critical}$=1.291
PSNR	173.3645	0.7534	0.9804
SSIM	136.0017	0.5942	0.7647
RR-LHS	166.8887	0.5186	0.6600
EPSNR	423.3052	1.2599	1.3213
RR-Weibull	209.8357	-	-
RR-Edge	203.5310	-	-
WNISM	333.7304	1.1869	0.8958
Proposed	99.6236	0.6049	0.4948

Table 4.4: *Residual variances of the IQAs on the three image subjective quality databases*

	LIVE(672 images) $F_{critical}$=1.1355	IRCCyN/IVC(185 images) $F_{critical}$=1.275	MICT(168 images) $F_{critical}$=1.291
PSNR	1	=	1
SSIM	1	=	1
RR-LHS	1	=	1
EPSNR	1	1	1
RR-Weibull	1	*	*
RR-Edge	1	*	*
WNISM	1	1	1

Table 4.5: *Performance comparisons regarding the statistical significance. In each entry, the symbol "1", "0" or "=" means that on the image database the proposed RR metric is statistically (with 95% confidence) better, worse or indistinguishable in comparison to its competitor. "*" means that the comparison cannot be performed due to the unavailable result data.*

4.6.5 Performance Analysis of Each Component

As we have mentioned before, the intra RDCT subband correlation, the inter RDCT subband dependency, and the image frequency distribution are utilized to design the RR IQA. In this part, we will try to figure out the contribution of each component to the final performance.

Table 4.6 illustrates the individual performance of each component of the proposed RR metric over the LIVE image database. For the CBD values of intra RDCT subbands, only three horizontal RDCT subbands are considered. Therefore, as 3 parameters are required to depict the coefficient distribution, $3 \times 3 = 9$ parameters are extracted for the RR features of intra RDCT subband correlation. According to the HVS oblique effect, HVS presents similar sensitivity to the horizontal and vertical information. Therefore, by considering only the horizontal ones, the visual quality of the distorted image can be accurately depicted.

For the MI difference of inter RDCT subbands, 8 MI values are employed to depict the parent-child, cousin-child, and brother-child dependencies. The performance is better than the WNISM, while it only requires far smaller number of RR features (8 parameters *vs.* 18 parameters of WNISM). Additionally, it can be observed that the MI differences perform worse than the CBD values. The reason is that the correlations between inter RDCT subbands have been essentially ensured by the linear transformations. Therefore, compared with the coefficient distribution in each RDCT subband, the MI values between different subbands vary less significantly, thus cannot effectively depict the image distortions. However, the introduced distortion in the image will affect

the MI values between RDCT subbands. Therefore, it is necessary to incorporate the inter RDCT subband dependencies in designing the RR IQA, which plays a less but nevertheless an important role in image quality assessment.

For the FL_v of the image, the performance is very good. Even with only one parameter FRD extracted from the reference image, the performance is comparable with PSNR, and even better than WNISM, RR-Weibull, and RR-Edge, as shown in Table. 4.2. Therefore, if we want to further reduce the RR data rate, we can extract the FRD only and transmit it to the receiver side for perceptual quality analysis. It only requires 8 bits to represent the FRD of the reference image. The good performance may attribute to two reasons. Firstly, the distortions introduced will significantly change the frequency distribution of the image. The larger the FRD changes, the higher the distortion level. For example, the more compression is introduced for JPEG coded image, the more HF and MF components are discarded, compared with the LF ones. The FRD differences as in Eq. 4.14 will become larger, which indicates worse perceptual quality. Therefore, the FRD difference can depict the distortion level. Secondly, the mutual masking strategy is employed as formulated in Eq. 4.15. As discussed in [187], for the content of the original image and the artifacts, one's presence will affect the visibility of the other. Therefore, by using mutual masking, the texture masking effect of the HVS can be more accurately simulated.

	CBD	MI	FL_v
LCC	0.8983	0.7746	0.8770
SROCC	0.8943	0.7697	0.8809
RMSE	11.981	17.248	13.106

Table 4.6: *Performance of each component of the proposed RR metric on LIVE image database.*

In order to further demonstrate the contribution of each component of the proposed RR metric, different combinations of these components are evaluated on the LIVE image database, as well as over each individual distortion type. The experimental results are illustrated in Table 4.7 and Table 4.8. It can be observed that the proposed method can outperform all of these different combinations. It means that each component of our proposed RR metric does contribute to the final performance. Comparing the three combinations, we can see that CBD+FL_v can achieve the best performance. It is also

consistent with the performances illustrated in Table 4.7, where CBD and FL_v perform better than MI. However, CBD+FL_v is still not as good as the proposed RR metric. Therefore, the MI is a necessary component that contributes to the performance improvement of the proposed RR metric. In this case, if a very small RR data rate is required, we can extract FL_v and transmit it to the receiver side for perceptual quality analysis. With the increasing of the required RR data rate, we can further transmit the CBD RR features to the receiver side. Finally, if the RR data rate is sufficient, all the three components, specifically the FL_v, CBD, and MI, will be extracted and transmitted to the receiver side for a better performance.

	CBD+MI	CBD+FL_v	FL_v+MI
LCC	0.9100	0.9206	0.9114
SROCC	0.9050	0.9194	0.9127
RMSE	11.309	10.652	11.220

Table 4.7: *Performances of the combinations of different components of the proposed metric.*

Method		JPEG2000	JPEG	WGN	Blur	FF
CBD+MI	LCC	0.8461	0.9347	0.9141	0.9245	0.9394
	SROCC	0.8400	0.9241	0.9010	0.9331	0.9177
	RMSE	13.403	11.438	11.357	6.907	9.617
CBD+FL_v	LCC	0.8908	0.9556	0.9150	0.9320	0.9383
	SROCC	0.8821	0.9540	0.8988	0.9406	0.9170
	RMSE	11.426	9.479	11.296	6.568	9.703
FL_v+MI	LCC	0.9315	0.9586	0.9294	0.9652	0.9288
	SROCC	0.9223	0.9597	0.9144	0.9666	0.9268
	RMSE	9.149	9.160	10.331	4.736	10.395

Table 4.8: *Performances of the combinations of different components of the proposed metric over individual distortion type.*

4.7 Conclusion

In this chapter, we propose a novel RR IQA by considering the intra and inter subband correlations in the RDCT domain. The CBD and MI values are firstly employed to depict the intra and inter RDCT relationships, respectively. The FRD calculated in RDCT domain depicts the frequency distribution of the images, which can be employed

to simulate the HVS texture masking effect in a mutual masking way. Combining the CBD values, MI differences, and FRD value together, an effective RR IQA is developed. Evaluations on several image quality databases demonstrate that the proposed method outperforms the state-of-the-art RR metrics and even FR metrics PSNR and SSIM. It means that the proposed metric correlates well with the human perception of the image quality. Meanwhile, only a small number of RR features are extracted.

Chapter 5

Reduced Reference Video Quality Assessment

5.1 Introduction

As introduced in Section 4.1, many RR IQAs have been developed by considering the distortions behaviours, HVS properties, and the statistics modeling of visual signals. Nowadays, many RR VQAs are developed by extending the RR IQA by characterizing the distortions in spatio-temporal domain rather than in spatial domain only. Among these RR VQAs, VQ Model [80] is one of the best proponents of the VQEG FRTV Phase II tests [65]. For a video sequence, VQ Model generates seven distortion factors to measure the perceptual effects of a wide range of impairments, such as blurring, blockiness, jerky motion, noise and error blocks, etc. Viewed conceptually, VQ Model's distortion factors are all calculated in the same steps. Firstly, the video streams are divided into 3D Spatial-Temporal (S-T) sub-regions typically sized by 8 pixel $\times$ 8 lines $\times$ 0.2 second; then feature values are extracted from each of these 3D S-T regions by using statistics (mean, standard deviation, etc.) of the gradients obtained by a 13-coefficient spatial filter, and these feature values are clipped to prevent them from measuring unperceivable distortions; finally these feature values are compared and their differences combined together for quality prediction. Three feature comparison methods used by VQ Model are Euclidean distance, ratio comparison, and log comparison. Also as aforementioned, RR-LHS [165] employed discriminative local harmonic strength with motion consideration to evaluate the distorted video quality. The gradient information of each frame is employed for harmonic and discriminative analysis. Furthermore, Zeng *et al.* [191] [192] extended the RR IQA to VQA, by modeling the video natural temporal statistics. In [191], the temporal motion smoothness of a video sequence is proposed to examine the temporal variations of local phase structures in

the complex wavelet transform domain. In [192], both intra- and inter-frame RR features are calculated based on the statistical modeling of natural videos. Together with a robust video watermarking approach, a quality-aware video system is developed. It has been demonstrated that these two RR VQAs present good measurement of the individual distortion level. However, these metrics are not evaluated over the subjective quality video database, which leads to a deficiency of the evaluation results.

This study deals with the RR quality assessment by extending the previous work RR IQA introduced in Section 4 for compressed video sequences. With inspiration from the RR IQAs, an efficient RR VQA for compressed video sequences is proposed. Firstly, from the spatial perspective, an energy variation descriptor (EVD) is proposed to measure the energy change of each distorted frame. The proposed EVD can also be utilized to simulate the texture masking property of the HVS. For the temporal distortion, the generalized Gaussian distribution (GGD) is employed to model the histogram distribution of the inter frame difference. The city-block distance (CBD) is used to calculate the histogram difference between the original video and the distorted one. Finally, the perceptual quality index is derived by combining the spatial EVD together with temporal CBD. The rest of the chapter is organized as follows. The detailed algorithm will be introduced in Section 5.2. Section 5.3 will demonstrate the performance comparisons. Finally, the conclusion will be given in Section 5.4.

5.2 Proposed Reduced Reference Video Quality Metric

The general framework of the RR VQA system is illustrated in Figure 5.1. In the sender side, the RR features which are sensitive to the HVS perception are firstly extracted from the original video sequence. Then the original video is encoded and transmitted to the receiver side. The corresponding RR features can be embedded into the coded bit-streams or transmitted through an ancillary data channel to the receiver side. After decoding, the processed features can be calculated from the compressed video sequence. By comparing the processed features with the ones of the original video sequence, the visual quality index of the compressed video can be generated.

As aforementioned, in order to develop an efficient RR VQA, several challenges need to be considered. In the sender side, the extracted features need to be sensitive to a variety of video coding distortions, not only from the spatial perspective but also from

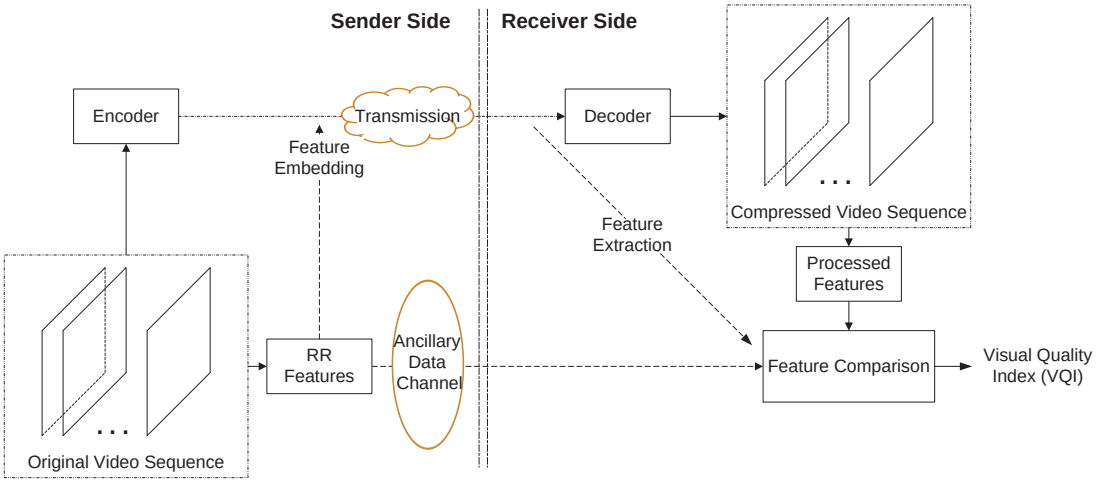


Figure 5.1: *General framework of the RR VQA system.*

the temporal perspective. Also these features have to be relevant to the HVS perception of the video quality. The second important issue is the computational complexity of the RR feature calculation. If the complexity is too high, the receiver cannot easily compute the processed features from the compressed video. Consequently it cannot practically monitor the visual quality of the distorted video. Therefore, the feature computation process should be efficient. Another important aspect is that the RR feature selection should consider not only the prediction accuracy of the quality metric, but also the data rate of the RR features. For a higher data rate, one may include more information of the reference video. Thus a good performance can be obtained, but this on the other hand will introduce a heavy burden to the RR feature transmission. Actually, the FR VQA is one extreme case of RR VQA, with the data rate being the whole reference video. With a smaller data rate, little information of the reference image/video is available, resulting in poor quality prediction accuracy. As such, we can regard the NR VQA as another extreme case of RR VQA, with no information from the reference video. How to balance the data rate and performance is the key point for RR feature selection.

The framework of extracting the RR features in the sender side is illustrated in Figure 5.2. For each original video frame, the RR feature representing the distortions from the spatial perspective is calculated. As the difference frame can depict the temporal relationship between adjacent frames, the temporal features are extracted from each difference frame. After the feature extraction, the compression process is performed to represent the RR features in limited bits, which can be easily transmitted

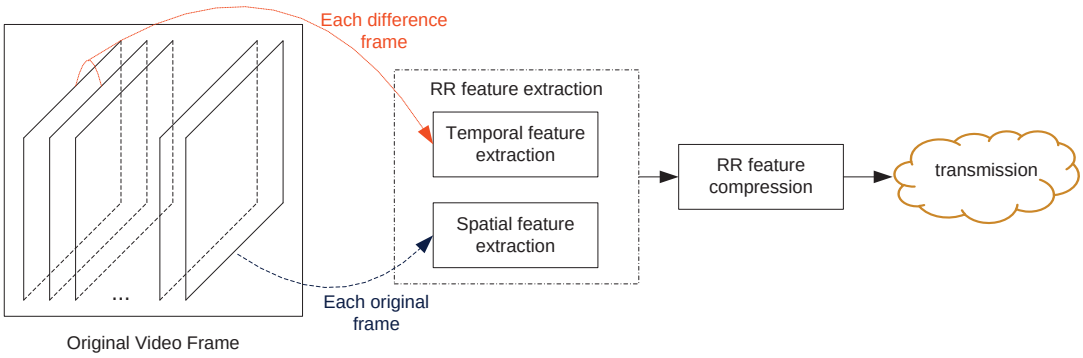


Figure 5.2: *RR feature extraction in the sender side.*

to the receiver side for visual quality analysis. The following sections will introduce the detailed information of feature extraction from both spatial and temporal perspectives.

5.2.1 Reduced Reference Feature Extraction from Spatial Perspective

The distortion of the video sequence encoded by MPEG-2 and H.264 is introduced during the quantization process, which quantizes the DCT coefficients of the spatial blocks into different levels. It can help to efficiently reduce bit-rates for representing the video sequence. However, the quantization process results in the useful information loss. Intuitively, the larger the quantization step, the more is the information loss is, and the worse is the perceptual quality of the encoded video. Therefore, the information loss has certain implicit relationship with the video perceptual quality. In this study, we propose an energy variation descriptor (EVD) to represent the spatial information loss.

For each block-based DCT (take an 8×8 DCT for an example), the DCT sub-bands can be categorized into different frequency bands, namely, high frequency (HF), medium frequency (MF), and low frequency (LF). In JND estimation [119] [122], the authors employed the energies of different subbands to indicate different block types. Based on these different types, the visual texture masking property is described. The frequency categorization of DCT subbands is illustrated in Figure 5.3. Let L, M, and H represent the sums of the absolute DCT coefficient values in the LF, MF and HF groups, respectively. It should be noted that the quantization matrix is not uniformly distributed. The higher the DCT frequency, the larger the quantization parameter is. The reason is that the HVS is more sensitive to the LF components, which should be

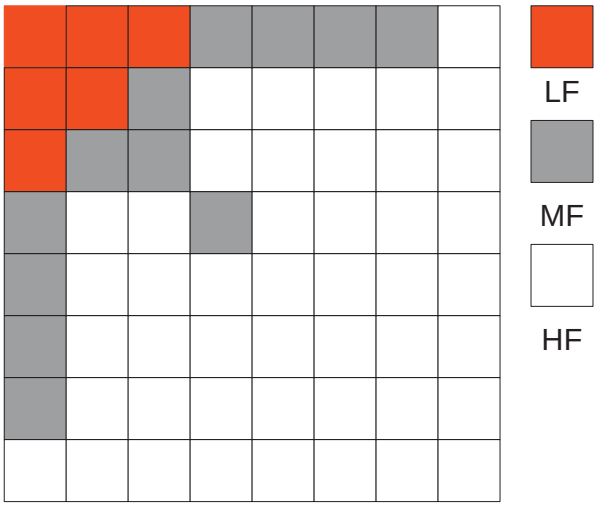


Figure 5.3: *RR feature extraction in the sender side.*

preserved during the quantization process. Therefore, it is not reasonable to record the absolute values of L, M, and H, which cannot effectively depict information loss. In this part, the corresponding frequency ratio EVD is proposed to depict the HVS-related information loss, which is defined as:

$$EVD = \frac{M + H}{L} \tag{5.1}$$

The above definition is for each 8×8 DCT block. You can sum all the L, M, and H values over all the blocks to get the EVD value for a whole image/frame. From the definition, we can see that the EVD depicts the frequency energy proportion of the original video frame. When the distortion is introduced, specifically in the quantization process, the energies of MF and HF components will change more significantly than the LF ones. Thus, the EVD can accurately depict the changes and effectively capture the information losses. Furthermore, the larger the value of EVD, the more energy the MF and HF components possess. It means that the DCT block is more likely to contain texture information. For the plain block, the energy mostly concentrates in the LF components. For the edge block, there will be only a small number of DCT coefficients in the HF group. Consequently, the texture block will present higher EVD. As discussed in the JND models [119] [122], the texture block can tolerate more distortions than the plain and edge block, which is interpreted as the texture masking property of the HVS. Therefore, the proposed EVD can be employed to simulate the

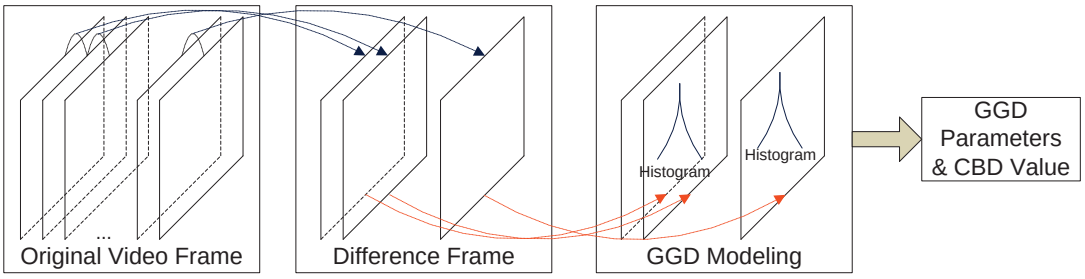


Figure 5.4: *RR feature extraction in the sender side.*

texture masking property for the derivation of the final video quality metric.

5.2.2 Reduced Reference Feature Extraction from Temporal Perspective

The temporal RR feature extraction strategy is illustrated in Figure 5.4. Firstly, the temporal relationship between adjacent frames needs to be depicted. The block-based motion estimation [193] [194] and optical flow [195] are employed to explore the motion information between the corresponding blocks or pixels of adjacent frames. However, although they can provide much more accurate information for describing the motion, the computational complexity is too high for practical implementations, especially in the receiver side. Therefore, we simply employ the difference image for characterizing the temporal relationship between adjacent frames:

$$D(i) = I(i) - I(i-1), i \in 2, 3, \ldots, N \tag{5.2}$$

where $I(i)$ is the $i-th$ original video frame, $D(i)$ is the corresponding difference frame, N is the total frame number of the video sequence. This simple scheme has been proved to be effective for detecting the visual saliency map of the natural video sequences [147]. Since luminance is more important than chrominance for our visual system, only the luminance information is considered to compute the difference frame. In order to illustrate the statistical property of the difference image, several original video sequences, such as PA, PR, RB, and TR are selected from the LIVE video quality database [17] [18] for demonstration, as illustrated in Figure 5.5. In order to provide a better visualization, the difference image has been reconstructed by $128+(Pixel_Value)$. It can be observed that the pixel values of the difference image mostly concentrate around zero, which generates a highly kurtotic distribution (with a sharp peak at zero and a fat-tail distribution). As demonstrated in [167] [168], the histogram distribution

of the wavelet coefficient is highly kurtotic. And this highly kurtotic distribution can be well fitted by the generalized Gaussian density (GGD) function. Furthermore, the coefficient distribution of the RDCT subband, presenting highly kurtotic, can also be modeled by GGD as shown in Section 4.4.1. Therefore, in this study the GGD is employed to model the histogram distribution of the difference image. The probability density function (PDF) of GGD is defined as:

$$p_{\alpha,\beta}(x) = \frac{\alpha}{2\beta\Gamma\left(\frac{1}{\alpha}\right)} exp\Big\{ -\left(\frac{|x|}{\beta}\right)^{\alpha}\Big\} \tag{5.3}$$

where α models the width of the PDF peak (standard deviation), while β is inversely proportional to the decreasing rate of the peak. α and β are also referred to as the scale and shape parameters, respectively. Γ is the Gamma function given by:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} exp(-t) dt \tag{5.4}$$

The GGD model can accurately model the histogram distribution, as demonstrated in Figure 5.5, where the actual histogram distribution and the fitted GGD curve overlap with each other. Furthermore, it can be observed that the GGD model can work effectively with different types of video sequences. For example, the PA video sequence is captured by a static camera, which results in a great proportion of the pixel value around zero, whereas the PR video sequence is captured by a moving camera, hence the pixel value distribution is much flatter. On the other hand, the RB video sequence is rich of dynamic texture information, and the TR video sequence is captured with a camera zooming effect.

By considering the maximum-likelihood estimation and assuming $\beta > 0$, we can obtain the approximated $\hat{\alpha}$ [196] according to:

$$\hat{\alpha} = \left(\frac{\beta}{L}\sum_{i=1}^{L} |x_i|^{\beta}\right)^{\frac{1}{\beta}} \tag{5.5}$$

where x_i is the pixel sample from the corresponding difference image, L denotes the total number of the pixels. From Eq. 5.5, it can be observed that the estimated $\hat{\alpha}$ is related to the energy of the difference image in the β-norm. The difference energy can somewhat reflect the temporal changes between adjacent frames. That is the reason why we introduce GGD to model the histogram distribution of the difference image,

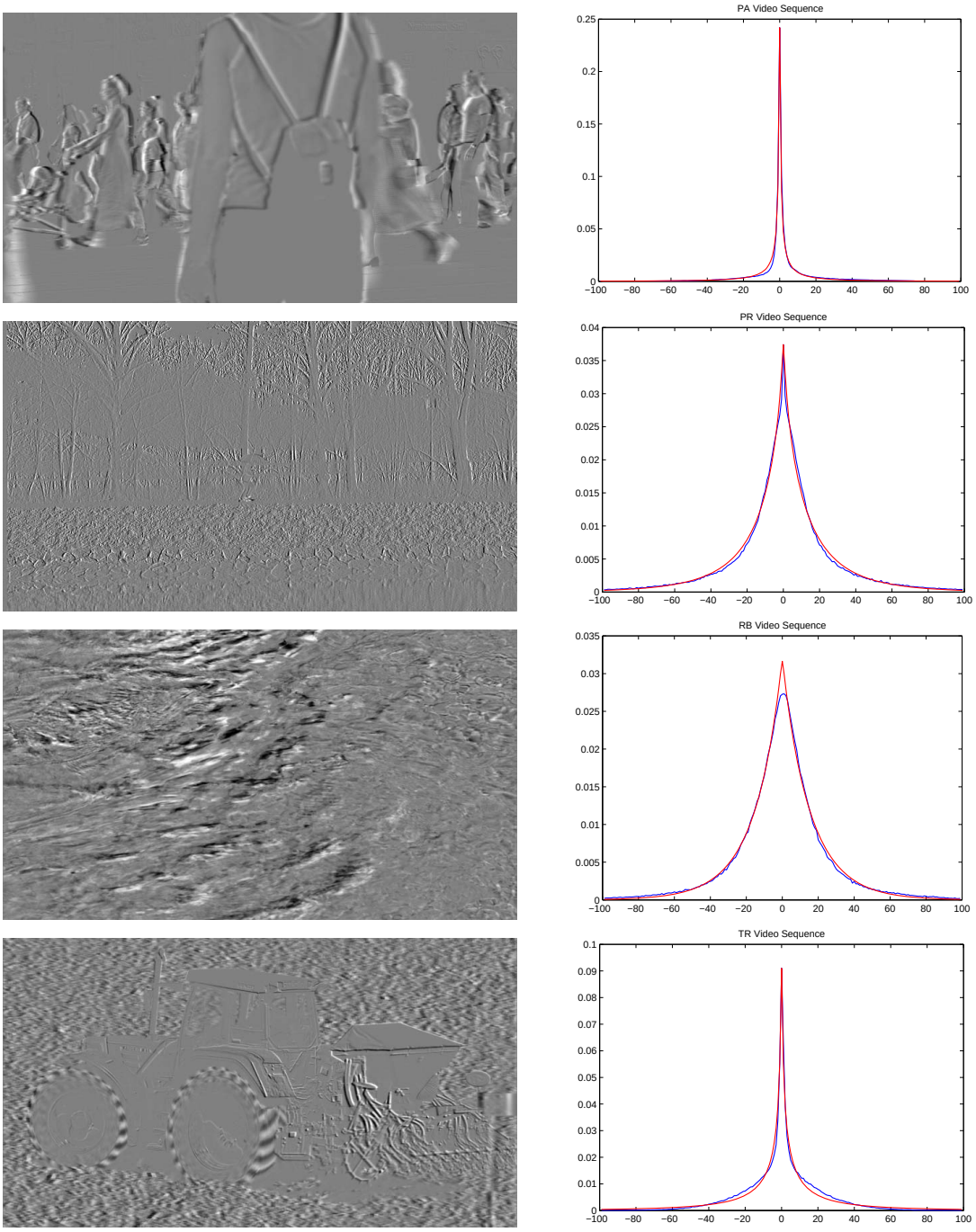


Figure 5.5: *Left: the 11^{th} difference image of the original video sequence, right: its corresponding histogram (blue line), and the fitted GGD curve (red line). From top to bottom: the PA, PR, RB, and TR video sequence from the LIVE video quality database [17] [18]*

not only because of the modeling accuracy but also its ability to indicate the energy of frame difference. As demonstrated in [154] [158], the energy of the frame difference is useful to measure the temporal content for video quality assessment. Furthermore, in order to improve the modeling accuracy, another parameter besides (α, β) is introduced,

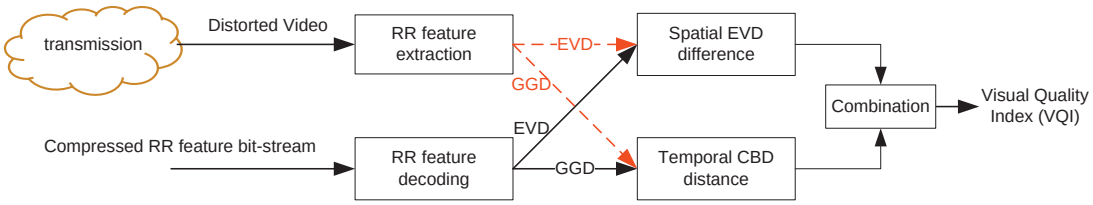


Figure 5.6: *Framework of visual quality analysis in the receiver side.*

which is named as CBD, which has been introduced:

$$d_{CBD}(p, p_{\alpha,\beta}) = \sum_{i=1}^{h_L} |p(i) - p_{\alpha,\beta}(i)| \tag{5.6}$$

where $p(i)$ is the actual histogram of the difference image, $p_{\alpha,\beta}(i)$ is the fitted GGD curve, and h_L is the total number of the histogram bins. Compared with KLD, CBD is symmetrical, which makes it more reasonable for evaluating the histogram distance as discussed in Section 4.4.1.

For each video frame, one parameter EVD is recorded to depict the spatial information loss, and three GGD parameters $\{\alpha, \beta, d_{CBD}(p, p_{\alpha,\beta})\}$ are extracted from each difference image for describing the temporal information. Therefore, there will be 4 parameters per frame in total to be recorded and transmitted to the receiver side for the quality assessment. For the EVD parameter, it is quantized into 8-bit precision for transmission. For the 3 GGD parameters, same as in [167], β and $d_{CBD}(p, p_{\alpha,\beta})$ are quantized into 8-bit precision, and α is represented using 11-bit floating point, with 8 bits for mantissa and 3 bits for exponent. The quantization steps are set uniformly to represent the corresponding parameters in a limited number of bits. Therefore, for each frame, only $8 + 8 + 8 + 8 + 3 = 35$ bits are required to represent the RR features. As the data rate is very small, the features can be easily transmitted through an ancillary data channel. Furthermore, they can also be embedded into the same video signal with a robust watermarking scheme [192].

5.2.3 Visual Quality Analysis in Receiver Side

In the receiver side, as shown in Figure 5.1, we need to evaluate the visual quality of the compressed video sequence based on the RR features of the original video. The framework of the visual quality analysis in the receiver side is illustrated in Figure 5.6.

Firstly, the feature calculation procedure is performed on the distorted sequence to obtain the processed features, which consist of the spatial EVD and temporal GGD. The original RR features are decoded from the transmitted bit-streams. By comparing the original features with the processed ones, the spatial EVD difference and temporal CBD distance are obtained. By combining the two distances together, the visual quality score of each frame is generated. The final video quality index (VQI) of the corresponding video is obtained by temporally pooling the frame-level scores together.

For the spatial EVD, as the compression process will discard more HF and MF components than the LF ones, the degradation of EVD can effectively represent the information loss caused by the compression:

$$EL = |EVD_{ori} - EVD_{pro}| \tag{5.7}$$

where EVD_{ori} is the original feature, and EVD_{pro} is calculated from the compressed video sequence. For the coded video sequences, the compression artifacts are superposed onto the original video sequence, which is regarded as the masker signal. Therefore, the original sequence is utilized to mask the compression artifacts, which are introduced by quantization process. As discussed before, larger EVD value indicates more texture information. Consequently, more distortion can be masked by a larger EVD. Therefore, the extracted EVD can be utilized to simulate the HVS texture masking property. The information loss in Eq. 5.7 is weighted by the original feature EVD_{ori} :

$$EL_v = \frac{EL}{EVD_{ori}} = \frac{|EVD_{ori} - EVD_{pro}|}{EVD_{ori}} \tag{5.8}$$

where EL_v is the final HVS-related features for depicting the spatial information loss.

For the temporal difference image, the CBD is employed to measure the difference between the reference video and the distorted one:

$$d_{CBD}(p, p_d) = \sum_{i=1}^{h_L}(p(i) - p_d(i)) \tag{5.9}$$

where p depicts the difference image histogram of the original video, and p_d is the distorted one. However, as the original video is unavailable at the receiver side, the

fitted GGD curve is employed to approximate the distance:

$$\acute{d}_{CBD}(p, p_d) = |d_{CBD}(p_{\alpha,\beta}, p_d) - d_{CBD}(p_{\alpha,\beta}, p)| \tag{5.10}$$

where $d_{CBD}(p_{\alpha,\beta}, p)$ is the third parameter introduced in the sender side. In the receiver side, only $d_{CBD}(p_{\alpha,\beta}, p_d)$ needs to be calculated. Their difference will be recorded to represent the statistical feature distance from the temporal perspective. As in [167], the logarithm process is employed to scale the temporal CBD distance as $\log_{10}(1 + \acute{d}_{CBD}(p, p_d)/c)$, where c is utilized to scale the CBD distance to avoid the variation being too small, and it is set as 0.001 for simplicity.

After obtaining the spatial EL_v value and temporal $\log_{10}(1 + \acute{d}_{CBD}(p, p_d)/c)$ value, how to combine them together remains a problem. In [192], the authors employed the averaging process to combine the spatial and temporal values together. However, it is not suitable for our obtained spatial and temporal values, because their magnitudes are quite different. In order to make the spatial EL_v value and temporal $\log_{10}(1 + \acute{d}_{CBD}(p, p_d)/c)$ value contribute equally to the final quality score Q_s for each frame, the simple multiplication process is employed:

$$Q_s = El_v \times \log_{10}(1 + \acute{d}_{CBD}(p, p_d)/c) \tag{5.11}$$

Based on the frame-level quality score Q_s, the VQI for depicting the perceptual quality of the entire compressed video is obtained by temporally pooling the Q_s scores together. In our implementation, the averaging process is employed to generate the final VQI:

$$VQI = \frac{\sum_{i=1}^{N} Q_s(i)}{N} \tag{5.12}$$

where N is the total number of the video frames. According to the definition of VQI, the smaller the VQI, the better visual quality the compressed video sequence is. And the VQI of the original sequence is 0 according to its definition.

5.3 Experimental Results

In this section, different VQAs are compared to demonstrate the effectiveness of the proposed RR VQA for evaluating the video perceptual quality. Firstly, similar to [191] [192], the consistency between the quality index generated by our proposed method

and the distortion level is evaluated. Subsequently, the effectiveness of the proposed RR VQA is evaluated based on the LIVE video quality database [17] [18], compared with the other VQAs. Finally, each component of the proposed algorithm is evaluated separately to demonstrate their corresponding contributions.

5.3.1 Consistency Test of the Proposed RR VQA over Compressed Video Sequences

We first tested the consistency of our proposed RR VQA on the coding artifacts, specifically, the MPEG-2 compression and H.264 compression. The LIVE video quality database contains the coded video sequences and their corresponding DMOS values. The consistency results of our proposed RR VQA on the coded video sequences are illustrated in Figure 5.7. It can be observed that the relationship between the VQI and DMOS values is monotonic for a given source video, specifically the VQI value is monotonically increasing with the DMOS value for a given source video. The larger the VQI, the worse visual quality is the compressed video sequence, which possesses larger DMOS value. For all the original video sequences, the relationship between the VQI and DMOS value is approximately linear for both MPEG-2 and H.264 coded video sequences. For each original video sequence, if a new MPEG-2 or H.264 coded video sequence is introduced, we can utilize the slope information which can be derived from Figure 5.7, and its corresponding VQI value to predict its DMOS value with high accuracy. Consequently, the true perceptual quality of the coded sequence is obtained. In the following section, we will further evaluate the proposed RR VQA metric in the standardized way by measuring the relationship of the obtained VQI values and the provided subjective DMOS values.

The middle column of Figure 5.7 shows the EVD values of the original and distorted videos, respectively. The MPEG-2 and H.264 compression will change the EVD value of each frame. During the compression process, more HF and MF components have been discarded than the LF components, which results in a smaller value of $M + H$ in Eq. 5.1. Therefore, a smaller EVD value of each frame is obtained, compared to the original value. The histogram of the difference image indicating the temporal information is illustrated in the right column of Figure 5.7. Compared to the fitted GGD curve, the histogram distribution has been changed. As MPEG-2 and H.264 introduce more zero

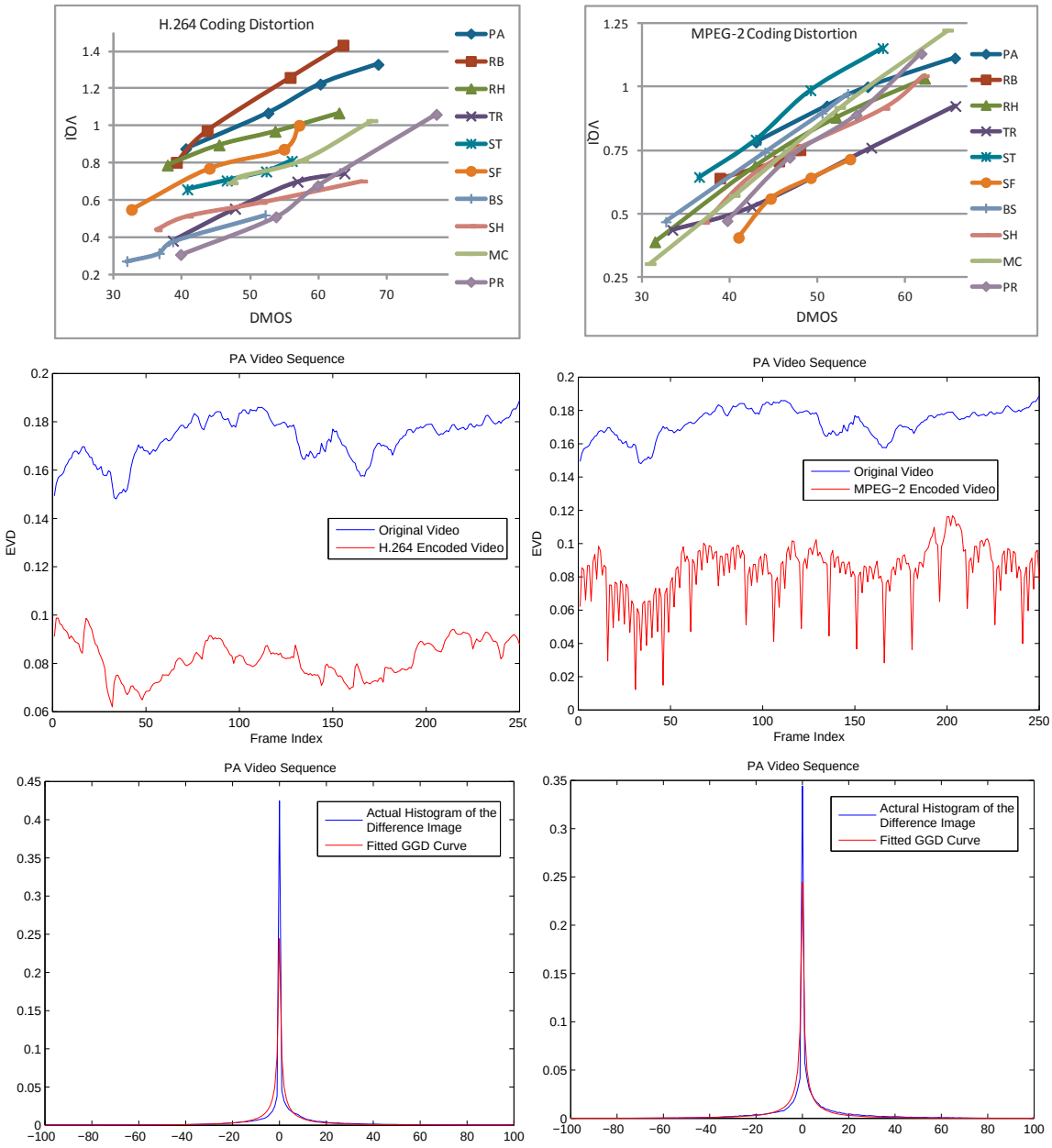


Figure 5.7: *Consistency evaluation of the proposed RR VQA over H.264 (left) and MPEG-2 (right) coded video sequences. Top: the proposed distortion measure VQI versus the DMOS value of each distorted video sequence; middle: spatial EVD value of the PA video sequence (with the largest VQI value); bottom: temporal histogram of the* 11^{th} *difference image of the distorted video PA (with the largest VQI value) and the fitted GGD curve.*

coefficients during the quantization process, a sharper and narrower distribution can be obtained from the distorted video sequence. Moreover, the actual histogram of the difference frame and the fitted GGD curve appear very close, while the EVD curves of the original and coded videos are quite different. The EVD is believed to affect the final VQI value more. As the compression distortion increases, although the LF component

starts to be affected, the HF and MF components are quantized even more severely. Therefore, by computing EVD in Eq. 5.1, its value becomes even smaller. However, the quantization step of each frame is usually the same during the compression process. The temporal CBD changes will not be as significant as the spatial EVD variations. That is the reason why the spatial EVD contributes more to the final quality score than the temporal EVD, as to be demonstrated in the following section.

As shown in Figure 5.7, the MPEG-2 coded sequences with the same perceptual quality (same DMOS value) demonstrate similar VQI values for different original sequences. On the contrary, the VQI values of the H.264 coded sequences with the same perceptual quality (same DMOS value) appear more dispersed. It means that the performance of the proposed RR VQA on MPEG-2 coded video sequences is more robust than that of H.264 coded video sequences. The reason is that the EVD is calculated based on 8×8 DCT. The energy variation of MPEG-2 can be accurately depicted, as the transform and quantization are performed based on 8×8 block. For H.264, different block-size based intra prediction, inter motion estimation, and DCT result in an inaccurate energy variation calculation. The final VQI values of H.264 coded video sequences of the same perceptual quality will be different.

5.3.2 Consistency Test of the Proposed RR VQA over Video Sequences with Simulated Distortions

Furthermore, as in [191] [192], the consistency property of the proposed RR VQA is evaluated over the simulated video sequences with five distortion types at different distortion levels. These five distortions include (1) Gaussian noise contamination, where the mean value is set as 0 and the distortion level is defined as the variance. (2) Gaussian blur distortion, where the filter is fixed as a 7×7 window, and the corresponding distortion level is determined by the standard deviation. (3) Line jittering, where each line in a frame is shifted horizontally by a random number uniformly distributed between $[-S, S]$, and S defines the line jittering level. (4) Frame jittering, where the whole frame is shifted together by a random number uniformly distributed between $[-S, S]$, and S defines the frame jittering level. (5) Frame dropping, which is simulated by discarding every 1 of N frames and repeating the previous frame to fill the empty frame, and $10 - N$ defines the distortion level. As claimed in [191], all these distortion

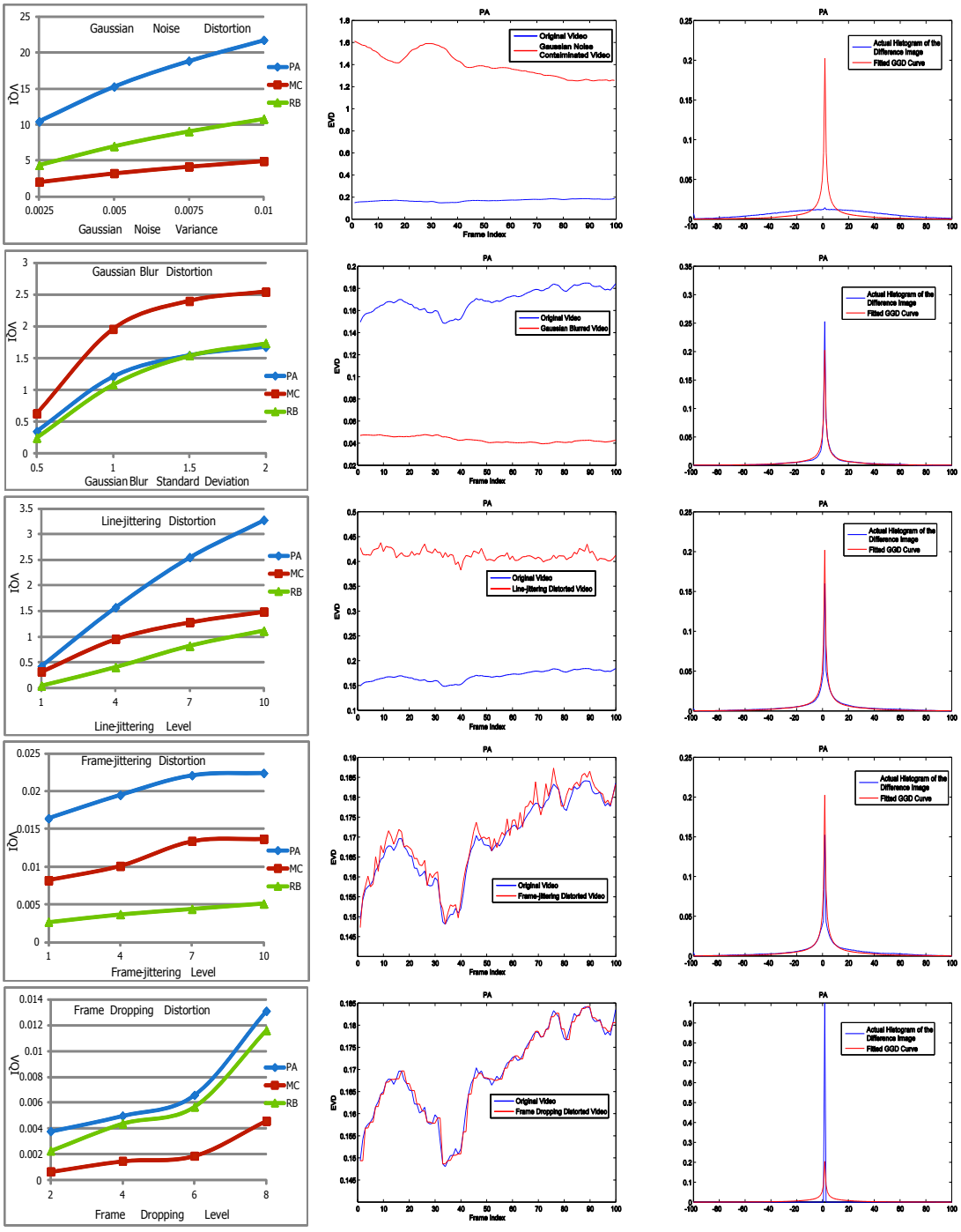


Figure 5.8: *Consistency evaluation of the proposed RR VQA over different distortions of different levels. Left: proposed distortion measure VQI versus the distortion level; middle: spatial EVD value of the PA video sequence (at the largest distortion level); right: temporal histogram of the* 10^{th} *difference image of the distorted video PA (at the largest distortion level) and the fitted GGD curve.*

types are associated with certain real-world scenarios. For example, frame jittering is often caused by irregular camera movement; line jittering often occurs when two fields

of interlaced video signals are not synchronized.

Figure 5.8 illustrates the consistency evaluation results over different distortion types of different levels. As we do not have the DMOS values of the video sequences contaminated by the aforementioned 5 distortions, the corresponding distortion level is utilized to indicate its perceptual quality. For each distortion type, the higher the distortion level, intuitively the worse is the perceptual quality of the processed video. Similar to MPEG-2 and H.264 coded video sequences, the relationship between the distortion level and the VQI is monotonic. Specifically the VQI value is monotonically increasing with the distortion level for a given source video. Therefore, from this aspect, we can conclude that the proposed VQI is sensitive to the levels of different distortions. It demonstrates a consistent relationship with the distortion level of different distortion types. The spatial EVD and the temporal CBD information of the PA video sequence (at the largest distortion level) are illustrated. For Gaussian noise contamination and line-jittering distortion, the EVD value of the distorted video is larger than that of the original video. It means that the HF and MF components increase more than the LF components. For the Gaussian noise contamination, the Gaussian noise dominates the histogram distribution of the difference image. It demonstrates a much flatter distribution, compared to the fitted GGD curve. For the line-jittering and frame-jittering distortion, as the temporal relationship still exists, the histogram distribution of the difference image appears to be similar with the GGD fitted curve. However, the pixel values of the difference image will increase due to the jittering distortion. Therefore, there will not be so many zero values, which results in a smaller peak value as shown in Figure 5.8. For Gaussian blur distortion, more HF and MF components are discarded compared with LF component. The EVD value decreases after the Gaussian blur process. And a sharper and narrower histogram distribution is obtained as more zero pixel values appear due to the filtering process. For the frame dropping distortion, the spatial EVD varies slightly, because of the close temporal relationship between adjacent frames. However, the pixel values of the difference image are all zero, as the previous frame is simply copied to fill the empty frame.

5.3.3 Performance Evaluation of the Proposed RR VQA on Compressed Video Sequences

In order to provide a more convincing result of the proposed RR VQA, we tested the proposed method on the LIVE video quality database [17] [18]. The performance can be evaluated by depicting the relationship of the obtained VQI values and the provided subjective ratings, specifically the DMOS value of each distorted video. As usual, three statistical measurements LCC, SROCC, and RMSE, are employed to evaluate the corresponding performances. According to the definitions, larger values of LCC and SROCC mean that the objective and subjective scores correlate better, that is to say, a better performance of the VQA. And the smaller RMSE values indicate smaller errors between the two scores, therefore a better performance.

We compare the performance of our proposed RR VQA with the representative RR video quality metric VQ Model [80], Yang's metric [155], RR-LHS [165], and J.246 [164], as well as several FR metrics: PSNR, SSIM [68], MSSIM [151], VSNR [72], and VIF [144]. The corresponding results together with the reference type and RR data rates are illustrated in Table 5.1. As PSNR, SSIM, MSSIM, VSNR, VIF, Yang's metric, and J.246 only provide frame-level quality scores, the final quality index of the video sequence is generated by averaging their outputs of each frame. For PSNR, SSIM, MSSIM, VSNR, and VIF, are FR metrics, the whole original frame should be available for quality analysis. Therefore, the RR data rates are regarded as the whole original video sequence. As for the RR VQAs, in order to ensure a fair comparison, the RR data rate is calculated based on video sequences of 25fps. For J.246, the locations and edge pixel values need to be encoded. As shown in [164], 14 extracted edge pixels per frame will result in the data rate as about 10 kbps. For Yang's metric, the only one extracted ratio parameter can be quantized in 8-bit precision. The data rate (about 0.2 kbps) is relatively small. For the RR-LHS, as shown in [165], the bit rate of the RR data is 64 kbps. For VQ Model, the compression method has been researched in [163], which ensures a more than 30× compression ratio compared to the original VQM features. The bit rate of the RR feature is about 150 kbps. For the proposed method, as only 35 bits for each frame are required to encode all the features, $35 \times 25 = 875$ bps are required to represent the features. Compared with the other RR VQAs except Yang's metric, the RR data rate of the proposed metric is much smaller. However, the performance

of the proposed metric is better. Furthermore, if only the spatial EVD is employed for constructing the RR metric, the RR data rate will be the same as Yang's metric. The performance is better, as to be illustrated in the following section.

	LCC	SROCC	RMSE	Reference type	Data rate(25fps)
PSNR	0.4488	0.4157	9.188	FR	-
SSIM	0.5946	0.5969	8.267	FR	-
MSSIM	0.6671	0.6944	7.717	FR	-
VSNR	0.3097	0.3041	9.777	FR	-
VIF	0.6447	0.6350	7.860	FR	-
J.246	0.5036	0.4460	8.883	RR	10 kbps
Yang's metric	0.5654	0.5366	8.484	RR	0.2 kbps
RR-LHS	0.4557	0.4082	9.152	RR	64 kbps
VQM	0.7003	0.6790	7.340	RR	150 kbps
Proposed	0.7567	0.7486	6.722	RR	0.875 kbps

Table 5.1: *Performances of different VQAs over the LIVE video quality database (MPEG-2 and H.264 encoded videos).*

From Table 5.1, it can be observed that the FR PSNR performs poorly, because it is not related to the HVS perception. Also the VSNR performs badly, which can be attributed to two reasons. The first is that VSNR analyzes the HVS perception of the distortion in the wavelet domain. But the MPEG-2 and H.264 compression schemes introduce the distortions during the quantization process in DCT domain. The second one is that VSNR is an image quality metric designed to capture the spatial distortions. For video quality assessment, the temporal information is very important and needs to be accounted for. This is also the reason why SSIM, MSSIM and VIF perform successfully in image quality evaluation, but not so well on the video quality assessment. Yang's metric employs the DCT coefficient ratio to measure the video quality. Although a small RR data rate is required, Yang's metric only depicts the DCT coefficient distortion from the spatial aspect. The temporal information is not considered. For J. 246, only the edge pixels in spatial domain are extracted for quality comparison. For RR-LHS metric, the harmonic and discriminative analysis is employed to depict the blocking and blur artifacts in the spatial domain. And the temporal motion information is employed to finally correct the quality values. From Table 5.1, it can be observed that the performances of these metrics are not good enough, with SROCC values smaller than 0.6. The reason is that the temporal information is not accurately

modeled. For video quality assessment, the temporal distortion is very important and needs to be considered for developing an effective video quality metric. The RR VQ Model [80] is derived by recording several features which depict the spatial information losses, edge information changes, contrast information, and the color impairments. However, the feature extraction process is of high complexity. And the RR data rate after compression is still very large.

As for our proposed method, it outperforms the VQ Model and the other FR quality metrics. It means that the proposed metric can effectively depict the perceptual quality of the compressed videos. Furthermore, the RR data rate is very small compared with the other RR VQAs, which will not introduce heavy burden for transmitting the RR features from the sender to the receiver side. The scatter-plots of different VQAs over the LIVE video quality database are illustrated in Figure 5.9. It can be observed that for our proposed method, the sample points scatter more closely around the fitted line. It means that the values predicted by the proposed method correlate better with the subjective ratings, specifically the DMOS values, demonstrating a better performance.

Moreover, for the proposed RR VQA and VQ Model [80], the triangles representing MPEG-2 coded videos scatter more closely to the fitted line, while several star points indicating H.264 coded videos are under or over estimated. As mentioned before, such scattering may be attribute to the fact that the features are calculated based on fixed block size, specifically EVD from the 8×8 DCT for the proposed RR VQA and the quality-related features from $(8 \times 8) \times 0.2$ second S-T region for the VQ Model. By considering the fixed 8×8 block in the spatial domain, the distortion of MPEG-2 can be accurately depicted, as the transform and quantization are performed based on 8×8 block. However, for H.264, different block-size based intra prediction, inter motion estimation, and DCT result in an inaccurate energy variation calculation. Therefore, the DMOS values correlate worse with the quality values of H.264 coded videos than that of MPEG-2 coded videos. In the future, we will consider the information of the H.264 coded video, specifically the transform and quantization block size. Then the EVD calculation can be extended to different block sizes for accurately capturing the energy variation, which is believed to be able to improve the performance of our proposed RR VQA.

As illustrated in Table 5.1 and Figure 5.9, the effectiveness of our proposed RR VQA

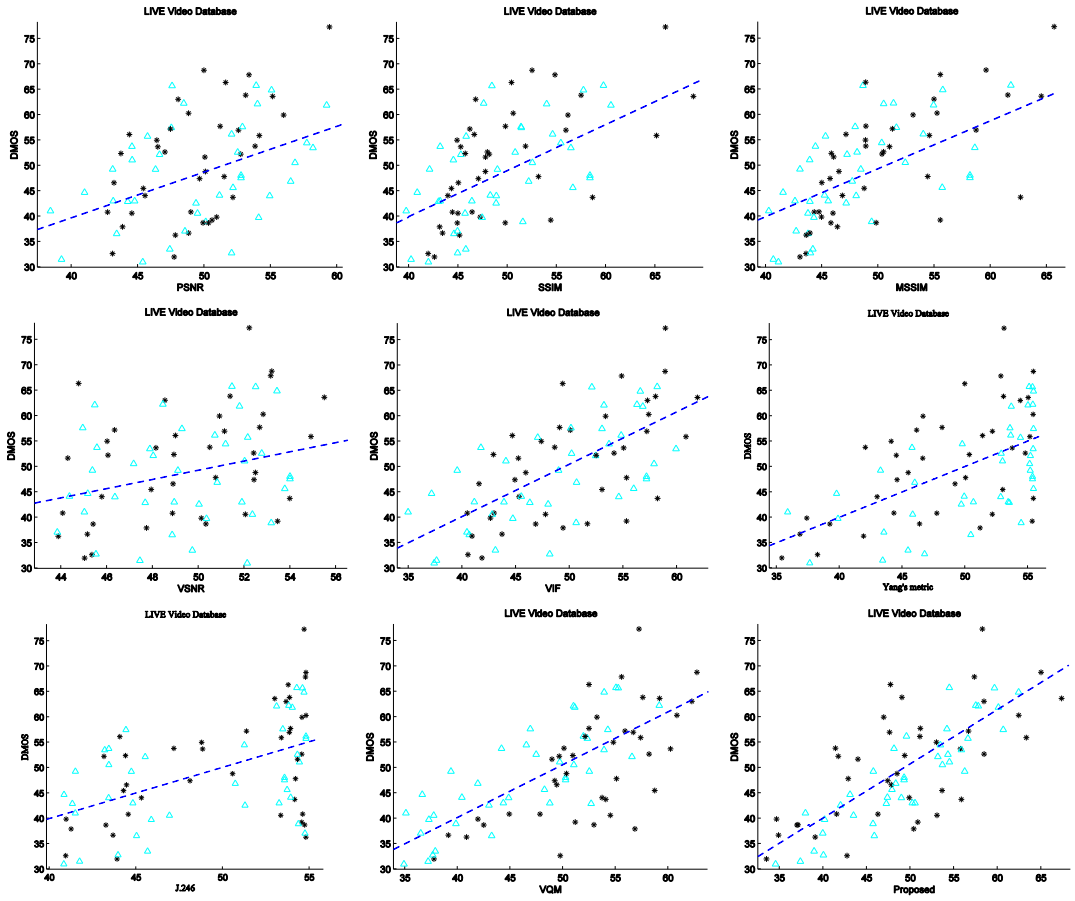


Figure 5.9: *Scatter plots of the DMOS values versus model predictions on the LIVE video quality database. Each sample point represents one test video. (The star indicates H.264 encoded video sequence, while the triangle indicates the MPEG-2 compressed one.) First row from left to right: PSNR, SSIM, and MSSIM; second row from left to right: VSNR, VIF, and Yang's metric; third row from left to right: J.246, VQ Model and the proposed method.*

has been clearly demonstrated compared with the other RR metrics or even FR metrics in terms of both performance and required RR data rate. Therefore, as in [163], we may consider incorporating the proposed RR VQA into video quality monitoring system, where the computational complexity of feature extraction and comparison needs to be evaluated. The spatial EVD of the proposed RR VQA only requires several addition and division processes after DCT, which can be calculated during the DCT process of the video encoding and decoding procedure. For the temporal GGD modeling and CBD calculation, the processing complexity in the sender side is different from that in the receiver side. In the sender side, as shown in Figure 5.4, the difference image is firstly obtained. Then the histogram depicting the pixel value distribution is modeled by the GGD. Finally, the CBD distance as shown in Eq. 5.6 is calculated to indicate

the modeling error. We implement the temporal feature extraction in Matlab. During our implementation, no optimizations are performed. A speed test is performed on our PC with a 3.0GHz Quad CPU and 6.0GB memory. For each difference frame, it only requires 0.7s on average for obtaining the temporal features. In the receiver side, we only construct the histogram of the difference image and compare it with the fitted GGD. The distance shown in Eq. 5.10 is approximated. As the fitting process is not performed, the computation of the temporal information is faster. The speed test is performed on the same PC, which indicates that only 0.14s per difference frame on average is needed for the temporal quality analysis. If further optimization is employed, it is believed that the quality analysis in the receiver side can perform even faster, which can be incorporated into the video quality monitoring system.

5.3.4 Performance Evaluation of the Proposed RR VQA on Video Sequences Containing Transmission Distortions

As the transmission errors over wireless channel and IP network are more realistic for the video quality monitoring, the proposed RR VQA and other representative RR quality metrics were evaluated on the LIVE video sequences containing transmission distortions. These distortions are simulated transmission of H.264 compressed bitstreams through error-prone IP networks and wireless channel. The performance of these RR video quality metrics are illustrated in Table 5.2. It can be observed that the proposed RR VQA can outperform J. 246, Yang's metric, and RR-LHS . However, it performs worse than VQ Model. For Yang's metric, it only employs the ratio between the parent coefficient (second DCT coefficient) and the child coefficient (the third and fourth DCT coefficient) to measure the video perceptual quality. Therefore, the distortion introduced by compression can be depicted, as the quantization process will change the ratio of DCT coefficients. However, the transmission distortion is not related to the ratio of DCT coefficients. Consequently, the perceptual qualities of the video sequences containing transmission distortion cannot be accurately depicted, which results in a bad performance of Yang's metric on these video sequences, as illustrated in Table 5.2. For J.246 and RR-LHS, the performances are not as good as the proposed one, although they required more RR data rates for representing the original video sequence. On the other hand, VQ Model extracts many features for quality analysis, which are

related to the specific distortions, such as blur, edge shifting, chroma spreading, color impairments, and so on. Most of these features can help to depict the distortions introduced during the video transmission. Therefore, the VQ Model performs well on these distorted video sequences. However, considering the data rates of different RR VQAs shown in Table 5.1 the RR data rates of VQM after compression is 150 kbps, which is about 170 times of the proposed VQA (0.875 kbps). It will introduce a heavy burden for the RR features' transmission.

Distortion		Proposed	RR-LHS	J.246	VQ Model	Yang's metric
IP distortion	LCC	0.6000	0.4602	0.3168	0.6553	0.2654
	SROCC	0.5582	0.3766	0.3437	0.6383	0.1462
	RMSE	7.488	8.873	8.873	7.071	9.017
Wireless distortion	LCC	0.5546	0.4684	0.5061	0.7416	0.0842
	SROCC	0.5386	0.4638	0.4051	0.7220	0.1041
	RMSE	8.586	9.117	8.900	6.922	10.282

Table 5.2: *Performances of different VQAs over the LIVE video quality database (IP and wireless distortion).*

For the proposed RR VQA, although it can outperform the other RR metrics except VQ Model employing very low RR data rates, the performance is still not good enough. It can be attributed to two reasons. Firstly, the transmission distortions over wireless channel and IP network are simulated from the H.264 compressed bit-streams. As discussed in Section 5.3.3, the proposed spatial EVD is calculated based on the fixed block size, specifically from the 8×8 DCT. However, for H.264, different block-size based intra prediction, inter motion estimation, and DCT are utilized, which result in an inaccurate energy variation calculation. Therefore, the transmission distortions simulated from H.264 compressed bit-streams cannot be accurately depicted. Secondly, in the RR VQA, the properties of the transmission errors, such as the error patterns, are not considered. If some features related with these errors are further incorporated, the RR VQA can more accurately depict the perceptual qualities of these degraded video sequences. In this study, the focus is on the RR VQA for the compressed video sequences. In future, as the RR data rate of the proposed metric is relatively small, we will consider incorporating more RR features to better handle the transmission errors.

5.3.5 Performance Analysis of Each Component

In this section, we evaluate the corresponding contribution for each component of our proposed metric in Eq. 5.11. To this end, we derive three different metrics to generate the frame-level quality score. The first one is the spatial EVD distance, which means $Q_s = EL$ as in Eq. 5.7. The second one is the weighted spatial EVD distance, which means $Q_s = EL_v$ as in Eq. 5.8. The third one is the temporal CBD distance defined as:

$$Q_s = \log +10(1 + \frac{d_{CBD}(p,p_d)}{c}) \tag{5.13}$$

where c is also set as 0.001. Their corresponding performances are illustrated in Table 5.3.

It can be observed that all of these three components are necessary for the proposed RR VQA. The spatial EVD distance as in Eq. 5.7 performs the worst. The reason is that it only considers the absolute difference of corresponding DCT coefficients, which captures the information loss during the quantization process. Therefore, it does not correlate well with the HVS perception. Furthermore, the distance in Eq. 5.7 is performed in the spatial domain. It does not consider the temporal information, which is critical to the video quality assessment. The HVS related weighting strategy of the EVD distance was tested as formulated in Eq. 5.8. As discussed before, the EVD of the original frame can represent its texture characteristic. The higher the EVD value, the more texture information it may contain. And the more texture information, the more distortion it can mask. Therefore, the EVD value is employed to simulate the texture masking property of the HVS as shown in Eq. 5.8. Compared with Eq. 5.7, the performance is significantly improved. It means that the EVD can accurately model the texture masking property of the HVS. Actually, Yang's metric also employs the ratio of DCT coefficients to measure the video quality in spatial domain. It employs the ratio between the parent coefficient (second DCT coefficient) and the child coefficient (the third and fourth DCT coefficient). However, it does not consider the texture masking effect of the HVS. Therefore, the performance, as illustrated in Table 5.1, is not as good as that of Eq. 5.8.

For the coded video sequences, the artifacts in the processed video are superposed onto the original video sequence, which is regarded as the masker signal. Therefore, as

	Spatial EVD distance as in Eq. 5.7	Weighted spatial EVD distance as in Eq. 5.8	Weighted spatial EVD distance as in Eq. 5.14	Temporal CBD distance as in Eq. 5.13
LCC	0.3986	0.5965	0.5958	0.4135
SROCC	0.3475	0.5992	0.5717	0.3950
RMSE	9.430	8.253	8.258	9.362

Table 5.3: *Performances of different components of the proposed RR VQA over the LIVE video quality database (MPEG-2 and H.264 encoded videos).*

shown in Eq. 5.8, we employed EVD_{ori} to mask the compression artifacts, which are introduced by quantization process. However, as discussed in [187], for the content of video sequence and the compression artifacts, one's presence will affect the visibility of the other. It is believed that the coded video sequence lacking of detailed information can also mask the artifacts. Therefore, we also evaluated the weighted spatial EVD distance, where EVD_{pro} is employed for simulating the HVS texture masking property:

$$\acute{EL}_v = \frac{EL}{EVD_{pro}} = \frac{|EVD_{ori} - EVD_{pro}|}{EVD_{pro}} \tag{5.14}$$

The corresponding performance is shown in Table 5.3. It can be observed that $\acute{EL}_v$ performs better than the spatial EVD distance formulated in Eq. 5.7. It means that EVD_{pro} can also simulate the texture masking property of HVS. However, EVD_{ori} as the masker signal can generate a better performance. Therefore, we only consider employing the original video signal to simulate the HVS texture masking effect in this study. It means that EVD_{ori} is employed to weight the spatial EVD distance as in Eq. 5.8. In future, we will research on how to accurately model the HVS texture masking effect by considering both the original and processed video signal.

The temporal CBD distance is evaluated as expressed in Eq. 5.13. The temporal CBD distance depicts the temporal statistical characteristic. It has been demonstrated to be related to HVS perception, as shown in [191] [192]. The distortions in the video will result in the statistical characteristic changes. By accurately capturing these changes, the corresponding perceptual quality can be described. Comparing the performances in Table 5.3 with those in Table 5.1, it is clear that the spatial distance or the temporal distance alone cannot outperform the integrated one. It means that only the spatial or temporal distortion alone is not sufficient to depict the perceptual quality of the video sequence. An effective RR VQA needs to accurately capture not

only the spatial distortion but also the temporal one. This is the main reason why our RR VQA outperforms the other quality metrics, such as PSNR, SSIM, VSNR, Yang's metric, J.246, and VQ Model.

5.4 Conclusion

In this chapter, an effective RR VQA is proposed by depicting the distortions from both the spatial and temporal perspectives. The EVD captures the information loss of each individual frame, which is also employed to simulate the texture masking property of the HVS. The GGD function and CBD distance are utilized to describe the temporal statistical characteristics. Evaluation results on the subjective quality video database show that the proposed RR VQA outperforms the representative RR metric VQM, and also the FR metrics. Due to its simplicity and efficiency in terms of feature representation, the proposed metric can be considered for incorporation into the video quality monitoring system.

Part III

Retargeted Visual Signal Quality Assessment

Chapter 6

Image Retargeting Perceptual Quality Assessment

6.1 Introduction

The previous chapters discuss the perceptual quality assessment of visual signals corrupted by traditional distortions, such as JPEG image compression, H.264 video compression, and so on. In this chapter, we investigate the newly encountered distortions, which are introduced during the image and video retargeting process. Nowadays, the diversity and versatility of the display devices have imposed new demands on digital image processing. The same image needs to be displayed with different resolutions on various devices. The image retargeting methods [197]- [207] have been proposed to adjust the source images into arbitrary sizes and simultaneously keep the salient content of the source images. These developed methods, such as seam carving [200]- [202], warp [198], and multi-operator [203], try to preserve the salient shape and content information of the source image, and shrink (or expend) the unimportant regions of the image into the given resolution. For most of these methods, a simple visual comparison was conducted for the results (comparing the results of different retargeting methods based on a small set of images) to demonstrate the efficiencies of the retargeting methods. Such a method cannot be used for on-line manipulation. In order to obtain an image with good quality, quality assessment of retargeted images should be performed and used to maximize the perceptual quality during the retargeting process. Therefore, there is a new challenge of objectively evaluating the retargeted image perceptual quality, where the resolution has been changed, the objective shape may be distorted, and some content information may be discarded.

Given that the ultimate receivers of images are human eyes, the human subjective opinion is the most reliable value for indicating the image perceptual quality. The subjective opinions are obtained through the subjective testing, where a large number

of viewers participate in the subjective test and provide their personal opinions of the image quality on some pre-defined scale. After processing these subjective scores across the human subjects, a score is finally generated to indicate the perceptual quality of the image. The subjective testing method is time-consuming and expensive, which makes it impractical for most image applications. However, the subjective rating obtained can be recognized as the ground truth of the image perceptual quality. Therefore, they can be employed to evaluate the performances of the objective quality metrics, which evaluate the image quality automatically [16]- [37]. Moreover, subjective studies can also enable the improvement in the performance of the quality metric towards attaining the ultimate goal of matching human perception. Then the developed quality metric can be utilized to guide the corresponding application. Furthermore, the subjective studies can also benefit the image applications for better perceptual quality experience, specifically improving the perceptual quality of the retargeted image. Therefore, there is a need to build an image retargeting database with subjective testing results, based on which we can evaluate the current developed quality metrics for retargeted images.

Until now, the only publicly available subjective image retargeting database is built by M. Rubinstein *et al.* [34]. The main purpose of building the database concentrates on a comparative study of existing retargeting methods. The authors compared which retargeting method generates the retargeted image with the highest perceptual quality. The subjective test is performed in a pair comparison way, where the participants are shown two retargeted images at a time, side by side, and are asked to simply choose the one they like better. The resulting database comprises the retargeted image and the corresponding number of times that the retargeted image is favored over another one. This is distinct from the traditional subjective testing [16]- [32], where the MOS or DMOS of each visual signal is obtained. Therefore, the perceptual quality metric for retargeted images cannot be evaluated in the standardized way [81], where the statistical measurements are used to match the scores between metric values and MOS/DMOS values. Moreover, as only the number of times that the retargeted image is favored over another image is recorded, the actual perceptual quality of the image is not clearly indicated. For one image with a larger number of favored times, it may possess a low perceptual quality if it is compared with images of even lower perceptual qualities. The image may be favored the most by comparing with other images, whereas its perceptual

quality may still not be accepted. It is also the main reason why the quality metric cannot be evaluated in the standardized way. Furthermore, the total number of possible paired comparisons is too large. It is unaffordable and unrealistic for employing many human subjects and taking long time. Therefore, the authors in [34] sample the space of possible comparisons to reduce the labors for the subjective testing. However, the completeness of the comparison is not ensured, which may affect the robustness of the subjective ratings. The most serious shortcoming of the database is that subjects have difficulties to arrive at an agreement on the perceptual quality of the retargeted image. The Kendall μ-coefficient [208] obtained for all the images is only 0.095. It is a relatively low value suggesting that the subjects in general had difficulty judging.

In this chapter, a subjective study is conducted to assess the perceptual quality of the retargeted image to build a publicly available database. Totally 171 retargeted images (in two different scales) are generated by different retargeting methods from 57 source images. With the source image as the reference, the perceptual quality of each retargeted image has been subjectively rated by at least 30 human viewers on a pre-defined scale. After processing the subjective ratings, the MOS value and the corresponding standard deviation are obtained for each image. Based on the MOS values, the constructed image retargeting database is analyzed from the perspectives of the retargeting scale, the retargeting method, and the source image content. Moreover, some publicly available quality metrics for retargeted images are evaluated on the database in the standardized way. Furthermore, a specifically designed subjective testing process is carried out to provide further information for developing an effective quality metric for retargeted images.

Our constructed database mainly focuses on evaluating perceptual quality of the retargeted images other than pair-wise comparing the retargeting methods [34]. Therefore, based on our database, the objective quality metrics can be evaluated in the standardized way. For the database [34], only the Kendall τ distance [209] is employed to measure the degree of correlation between two rankings. Same as traditional image/video quality assessments where multiple image/video databases [16]- [32] were created, the image retargeting quality assessment also requires multiple image databases. When constructing different databases, different subjects participated in the subjective testing with different rating scales. Meanwhile, the source image content and image

distortions introduced by retargeting are quite different. In these respects, the subjective quality databases can be ensured to be of great diversity, which can be employed to evaluate the effectiveness and robustness of the developed objective quality metric. Therefore, our database and the one in [34] can be further viewed as complementary to each other.

The rest of this chapter is organized as follows. In Section 6.2, we will introduce the subjective testing process for building the image retargeting database. In Section 6.3, the obtained subjective ratings will be processed and analyzed. In Section 6.4, some objective quality metrics are introduced and evaluated on the built database. Finally, Section 6.5 will conclude the work.

6.2 Preparation of Database Building

6.2.1 Source Image

Content-aware retargeting methods generate images with high perceptual quality where some background content can be removed or efficiently compacted, and the clear foreground object will be preserved. However, for some images with geometric structures and faces, the perceptual quality of the retargeted image cannot be ensured. In order to build a reasonable image retargeting database, we need to consider the source images containing the frequently encountered attributes, such as the face and people, clear foreground object, natural scenery (containing smooth or texture region), and geometric structure (evident lines or edges). The detailed information of the attributes can be referred to Appendix A.

In order to build the database, we select 57 source images in which the frequently encountered attributes have been included. The corresponding resolutions of source images are diverse, in order to alleviate the influence of the image resolution on the subjective testing. Figure 6.1 illustrates some samples of the source images for generating the retargeted images. The source images are roughly categorized into four classes according the aforementioned attributes. It should be pointed out that one image may contain more than one attributes. For example, the image 'umdan' contains the attributes of people and geometric structure. The image 'bicycle1' contains the attributes of clear foreground object, people, and natural scenery. And the image 'fishing' contains the attributes of people and natural scenery. The attribute information

Figure 6.1: *Samples of the source images utilized in the subjective testing. The images in the top row mostly contain the attribute of face and people; the images in the second row mostly contain the attribute of clear foreground object; the images in the third row mostly contain the attribute of natural scenery; the images in the bottom row mostly contain the attribute of geometric structure.*

of the source image can be found in Appendix A. As the image retargeting methods are content-aware, the perceptual qualities of retargeted results from different source images will be different. The attributes of the images are critical to the perceptual quality of the final retargeted images. The human subjects are very sensitive to the distortion of the faces and geometric structures, while they can tolerate more distortions on the natural scenery, especially for the texture regions. By including the images with different attributes, the subjective database can reflect how the retargeted images are favored by the human subjects.

6.2.2 Retargeting Methods

In order to efficiently demonstrate the perceptual quality of the retargeted images, the resolution changes are restricted in only one dimension. The retargeting methods change the resolution of the source images in either the width or height dimension. As shown in [197]- [207], most of the retargeting methods generate the retargeted images in two ratios, shrinking the image to 75% and 50%. Therefore, only these two retargeting ratios are employed to generate the retargeted image for constructing our

database. In the constructed database, three retargeted results of each source image are included. They may be in different retargeting scales. The reason why the database is built in this way is that we only care about the perceptual quality of the retargeted image, no matter how it is generated and what the resolution is. For some source images, the retargeted results in 50% scale appear to have very high perceptual quality, which perfectly preserve the salient information of the source image. For some source images, even the retargeted images in the 75% scale are of low perceptual quality. For the subjective testing of different scales separately, how the scale influences the perceptual quality may not be clearly revealed. Therefore, it is more reasonable to mix retargeted images with different scales together to examine its perceptual quality through subjective testing. Ten recently developed retargeting methods are employed to generate the retargeted images, which are detailed below.

- Cropping (CROP): manually choosing a window of the target size from the source image to maximize the salient information.
- Scaling (SCAL): simple scaling the source image into the target size.
- Seam carving (SEAM) [200]- [202]: removing the contiguous chains of pixels that lie in the regions of the smallest gradient magnitude values in the source image. The dynamic programming is employed to find the seams for removing.
- Optimized seam carving and scale (SCSC) [207]: a measurement named as "seam carving distance" is proposed to measure the similarity of retargeted image and the source one. A combination of linear scaling and seam carving is considered to optimize the measurement.
- Non-homogeneous retargeting (WARP) [198]: a warping function is optimized to find the optimal squeezed image by reducing the image width. The gradient magnitude together with the face detection is employed to indicate the saliency region of the source image, which needs to be preserved with high priority during the retargeting process.
- Scale and stretch (SCST) [204]: an objective function is optimized by uniformly scaling the salient regions to preserve the shape information. The saliency map

is detected by combining the gradient magnitude and the saliency map detected by Itti *et al.* [148].

- Shift-map editing (SHIF) [205]: graph cut is used to remove an entire object at a time rather than a seam. The smoothness is depicted by the color differences and the gradient information.

- Multi-operator process (MULT) [203]: seam carving, scaling, and cropping are combined together to generate the retargeted image. And a bi-directional warping measurement determines how to choose these operators.

- Energy-based deformation (ENER) [206]: similar as the SCST method, warping is also used to generate the retargeting image.

- Streaming video (STVI) [199]: the warping method is also used. The saliency map is obtained by combining the visual attention map, the line detection, and important objects.

Referring to these retargeting methods, it can be observed that the cropping, scaling, seam carving, and warping are the basic tools for image retargeting. Many research works are proposed to combine these tools together by optimizing a defined objective measurement. As the foreground objects, including the faces and people, represent the most salient information to the human viewers, the saliency map is incorporated into retargeting. It can be utilized to guide the image retargeting by preserving the shape information in the salient regions.

With these 10 retargeting methods, if each source image is to be retargeted into two different aspect ratios (75% and 50%), there should be 20 retargeted results for each source image. However, some retargeting methods, such as SCSC [207], MULT [203], ENER [206], and STVI [199], do not provide the source code or executive file. Therefore, we can only include the retargeted results provided by the developers of the corresponding retargeting methods. For some source images, the retargeted results cannot be generated. Including all of 20 retargeted images seems impossible. Secondly, we do not aim to compare the performances between different image retargeting methods as the authors in [34] did. Therefore, we need not include all the retargeted images at

each retargeting ratio into our database. The database we built mainly focus on evaluating the perceptual quality of the retargeted image. It needs only to ensure that the perceptual qualities of the selected images are sampled in an approximately uniform fashion as shown in [16]- [18]. In this respect, 3 retargeted images for each source image are manually selected according to the coarse judgment of the authors. Although 3 out of 20 seems a bit sparse sampling, different retargeted images obtained by different methods are selected, whose perceptual qualities are expected to be distributed uniformly from low to high qualities. The constructed database demonstrates a uniform distribution and good separation of the perceptual quality, as will be illustrated in the following section.

6.2.3 Subjective Testing

ITU-R BT.500-11 [5] has specified several methodologies for the subjective assessment of the quality of television pictures. These methods can be roughly categorized into two types: the double stimulus and single stimulus approaches. The double stimulus approach asks the subjective viewers to rate the quality or change in quality between two videos/images (reference and impaired). For the single stimulus approach, the subjective viewers only rate the quality of just one impaired video/image. As discussed in [210], each subjective test methodology has its own advantages. The double stimulus approach is claimed to be less sensitive to the context, where the subjective ratings are less influenced by the severity and ordering of the impairments within the test session. The single stimulus approach yields more representative quality estimates for quality monitoring. Also the single stimulus approach can ensure a faster and more efficient subjective testing process [211], compared with the double stimulus one.

However, for our subjective testing process of retargeted images, we not only care about the distortions perceived in the retargeted image, but also how much information of the source image has been conveyed. Therefore, in order to provide more convincing results, the source image needs to be presented to the subjective viewers as the reference simultaneously. Otherwise, if we employ the single stimulus approach, the CROP method will always yield the best quality, as no distortions are introduced. Without the source image as the reference, the viewers are not able to detect the discarded information, which may be the most important part of the source image. Therefore,

in this work, the simultaneous double stimulus for continuous evaluation (SDSCE) as specified in Section 1.2 is employed.

Two images are juxtaposed on the screen for the human subject. One is the source image for reference and the other is the retargeted image to be evaluated. The human subjects are aware of which one is the reference image and which one is the retargeted. The subjects are requested to check the difference between the two images and judge the perceptual quality of the retargeted one. After that, they provide their own opinions on the retargeted image quality. The only difference of the subjective testing in this work was the use of the ITU-R absolute category rating (ACR) scale rather than a continuous scale. The ACR scale employs a 5-category discrete quality judgment, as illustrated in Table 1.1. As discussed in [7], the subjective rating scales can be increased to more than 5 categories, such as 9 or 11 categories, which are particularly designed for the assessment of special applications, such as low bit-rate video codecs. Also an additional possibility is to use continuous scale rating, which can provide more precise subjective values. In [211], the experimental data has demonstrated that there are no overall statistical differences between different rating scales, which include (i) 5-category discrete scale, which is the one we employed in our subjective testing process; (ii) 11-category continuous scale; (iii) 5-category continuous scale; (iv) 9-category discrete scale. Moreover, for the subjective testing of retargeted image, the resolutions of the images and the introduced shape distortions are very different. The subjective viewers may have difficulties in judging the perceptual quality of the image and provide a precise subjective value. Therefore, in order to make the scoring process simpler to the subjective viewers and the subjective values more distinguishable, the 5-category discrete scale is employed to obtain the subjective opinions to build the image retargeting subjective quality database.

The user interface for the subjective testing is developed by using **MATLAB**, as shown in Figure 6.2. The two images, including the source and the retargeted one, are loaded into the memory before displaying. In order to avoid strong visual contrast, the remaining regions of the display area are gray (the pixel values are set equal to 128). The quality scales are labeled to help the human subjects to do the quality evaluation. The quality scales are labeled as "Bad", "Poor", "Fair", "Good", and "Excellent" (same as the one shown in Table 1.1), which range from the lowest to the

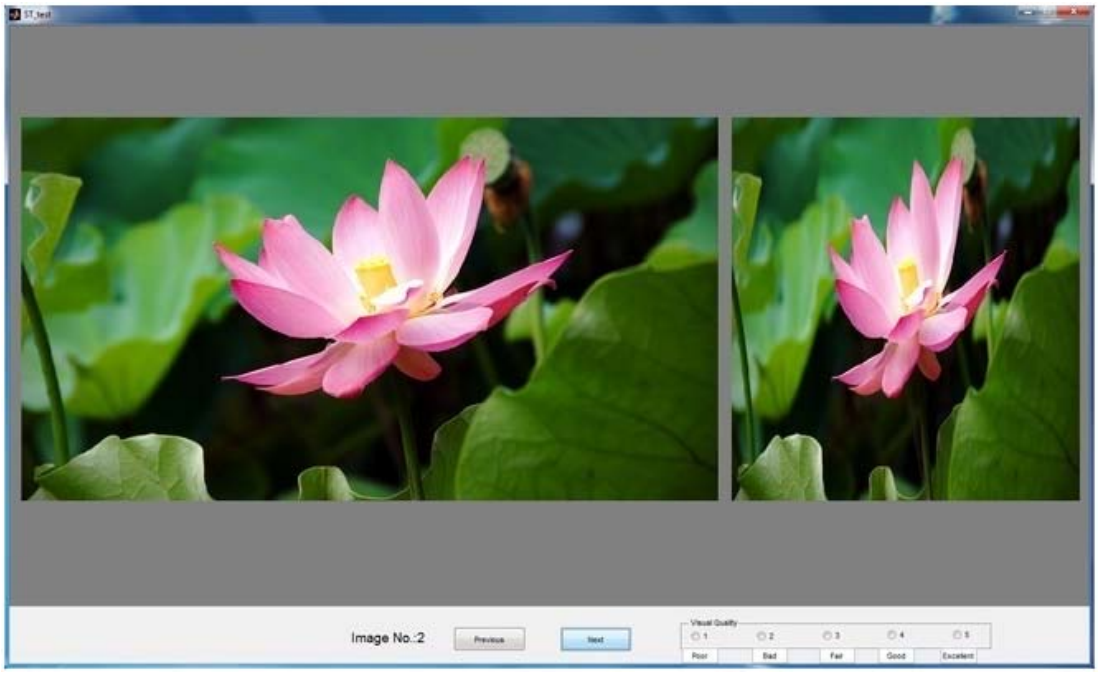


Figure 6.2: *Screenshot of the subjective study interface displaying the images to the human subject.*

highest perceptual quality index. During the subjective testing, the subjective values are recorded in numerical values. As shown in Figure 6.2, the "Bad" corresponds to 1 and the "Excellent" corresponds to 5. Therefore, for the obtained subjective ratings, the larger the value, the better is the image perceptual quality. The human subjects select the appropriate quality index according to their own opinions. After choosing the quality of one image, the subjects can go on evaluating the next image. The subject was allowed to take as much time as needed to evaluate the image quality.

In order to reduce the effect of the viewer fatigue, the 171 retargeted images are divided into 2 sessions. In the first session, the subjective testing is performed in two steps. In the first step, the subjective viewers are asked to provide their personal opinions on the perceptual quality of the retargeted image. After that, in the second step they are further asked to provide their personal opinions on the two distortion levels: (i) the level of shape distortion; (ii) the level of content information loss. The detailed process of the second step will be described in following Section 6.4.2. However, for the second session, the subjective viewers are only asked to take the first step of the subjective testing. Therefore, compared with the first session, the second one will take shorter time for each image. In order to reduce the effect of the viewer fatigue,

the number of the images in the first session should be smaller than that of the second one. But the number of images of the two sessions can be different. Considering this, we simply separate the images into two parts. The first session contains 69 images, while the second one contains 102 images. For each session, it will take the viewer about 10-20 minutes to accomplish the subjective testing. The order of the image pairs (the source image and the retargeted image) is randomly arranged, which is distinct for different viewers. Furthermore, in order to avoid the contextual and memory effects on the subjects' judgment of the quality, the retargeted images which are generated from the same source image will not be presented consecutively. In order to prevent the scaling effect, which is critical to the image retargeting results, the source image and the retargeted image must be displayed in their native resolution. In our experiment, the resolution of the screen for subjective testing is 1920×1280, which is sufficient for displaying the images in their original resolution.

During the subjective test, each viewer is briefed on the objective of this subjective study and told how to do the quality evaluation. Before starting the test, a training session is conducted for all the human subjects. There are in total 7 retargeted images in the training session. They are generated from different source images by different methods in different scales. Also their corresponding perceptual qualities span from "Bad" to "Excellent". The suggested quality scale is explained to each subject. After the training session, each subject should be clear on what they should do and how to provide their opinions on the retargeted image quality.

All the subjects participating in the subjective testing are the students from the Chinese University of Hong Kong in Hong Kong, and Nanyang Technological University in Singapore. They have normal vision (with or without corrective glasses) and have passed the color blindness test. For the first session, 30 subjects provided their personal ratings on the perceptual quality of each image, where 15 viewers are experts in image processing and the others are not. And each image in the second session was rated by 34 subjects, where 18 viewers are experts in image processing and the others are not.

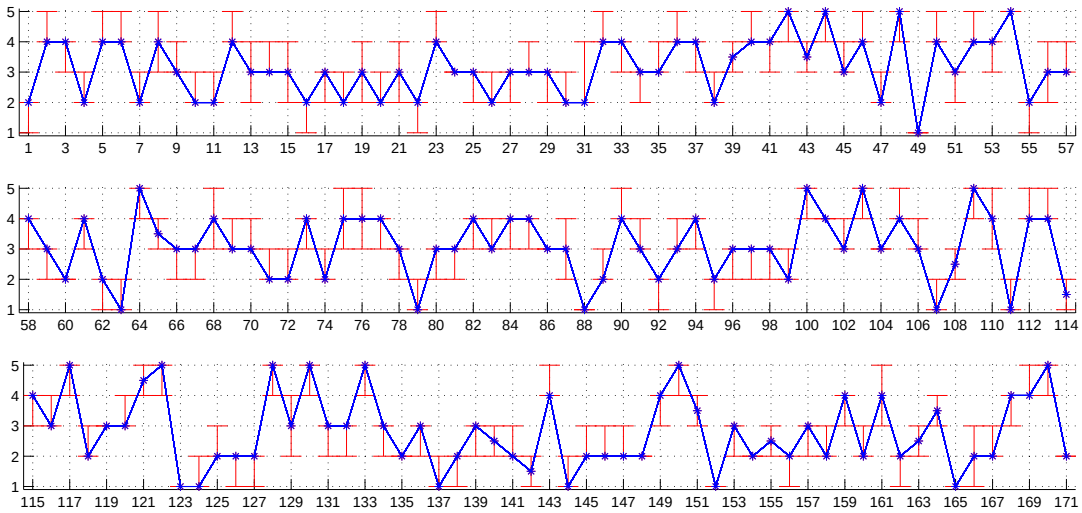


Figure 6.3: *The subjective scores for each image (the horizontal axes corresponds to the image number, and the vertical axes corresponds to the subjective scores of the viewers. The blue asterisk indicates the median value among all the viewers. And the red error bar indicates the corresponding 25^{th} and 75^{th} percentiles of the subjective scores).*

6.3 Data Processing and Analysis for the Database

6.3.1 Processing of Subjective Ratings

Subjective Agreement

Before we process the subjective ratings to build the database, we need to firstly examine the similarity of choices between participants. Each subject has its own opinion to interpret the image quality. However, for a large proportion of the images in the database, most of the participants should have agreements on the perceptual quality. If the subjective results demonstrate diversity among the human subjects, the corresponding image is not suitable for inclusion into the database.

In this work, we employ the quartiles of the subjective scores for each image to analyze the subject agreement, which is illustrated in Figure 6.3. The lower and higher bound of the red error bar denotes the 25^{th} and 75^{th} percentiles of subjective ratings obtained for each image. After sorting the subjective scores, the central 50% of subject ratings lie within the range. The blue asterisk indicates the median value of the subjective scores. The detailed information of the image number and the corresponding retargeted image name can be found in the Appendix B. An outlier coefficient (OC) is

For each subject i, find the P_{ik} and Q_{ik}

if $2 \leq \beta_j \leq 4$ (normally distributed)
  if $S_{ijk} \geq \mu_{jk} + 2\sigma_{jk}$, then $P_{ik} = P_{ik} + 1$;
  if $S_{ijk} \leq \mu_{jk} - 2\sigma_{jk}$, then $Q_{ik} = Q_{ik} + 1$;
else
  if $S_{ijk} \geq \mu_{jk} + \sqrt{20}\sigma_{jk}$, then $P_{ik} = P_{ik} + 1$;
  if $S_{ijk} \leq \mu_{jk} - \sqrt{20}\sigma_{jk}$, then $Q_{ik} = Q_{ik} + 1$;

if $\frac{P_{ik}+Q_{ik}}{N_{jk}} > 0.05$ and $\frac{P_{ik}-Q_{ik}}{P_{ik}+Q_{ik}} < 0.3$, then **REJECT** the subject i.

Figure 6.4*: Detailed algorithm of the subject rejection process.*

introduced to quantify the subjective agreement of the database:

$$OC = \frac{N_{outlier}}{N_{total}} \tag{6.1}$$

where N_{total} denotes the total number of the retargeted images in the database, and $N_{outlier}$ denotes the number of the images, which are regarded as the outlier. If the interval between the higher bound and lower bound error bar in Figure 6.3 is larger than 1, the image is recognized as the outlier image. The reason is that viewers may have different opinions on the image quality, but they should at least have the similar judgement. For one image, different viewers may interpret the same image as "Good" or "Excellent", which are neighboring values. In most cases, the same image will not be scored with greatly differences, such as "Poor" or "Good". Therefore, if the central 50% subjective ratings are constrained within the interval of 1, we believe that the participants have arrived at an agreement of the retargeted image quality. For the constructed database, 15 out of 171 are recognized as the outlier images, which implies $OC = 8.77\%$. Therefore, 91.2% of the images in the database have shown the agreement among participants. It is believed that the images in the database will be rated as the similar quality if subjectively tested by the others. Consequently, these images can be included for building the database and further employed for evaluation of the quality metrics

Screening of the Observers

In the previous section, we have examined the subject agreement on the retargeted image quality. The central 50% subjective ratings of the images have shown high agreement. However, in order to obtain the final MOS and standard deviation value

for each image, the subject rejection process is suggested by [5]. Let S_{ijk} denotes the subjective rating by the subject i to the retargeted image j in session $k = 1, 2$. The S_{ijk} values are firstly converted to Z-scores per session [212]:

$$\begin{aligned} \mu_{ik} &= \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} S_{ijk} \\ \sigma_{ik} &= \sqrt{\frac{1}{N_{ik}-1} \sum_{j=1}^{N_{ik}} \left(S_{ijk} - \mu_{ik}\right)^2} \\ z_{ijk} &= \frac{S_{ijk} - \mu_{ik}}{\sigma_{ik}} \end{aligned} \tag{6.2}$$

where N_{ik} is the number of the test images seen by the subject i in session k. It is noted that Z-scores are obtained per session, which accounts for any differences in subject preferences for the reference images, and the different human subjects between sessions.

After converting the obtained subjective ratings into Z-scores, the subject rejection procedure specified in the ITU-R BT 500.11 [5] is then used to discard the scores from unreliable subjects. The converting process and subject rejection procedure used should be superior to the VQEG studies [82] [65] [213]. The mean value μ_{jk} and the variance value σ_{jk} are firstly computed for each image by accounting for the differences of the subjective viewers. Then whether the scores assigned by a subject are normally distributed is determined by the kurtosis β_j of the computed scores:

$$\begin{aligned} \mu_{jk} &= \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} S_{ijk} \\ \sigma_{jk} &= \sqrt{\frac{1}{N_{jk}-1} \sum_{j=1}^{N_{jk}} \left(S_{ijk} - \mu_{jk}\right)^2} \\ \beta_j &= \frac{m_4}{(m_2)^2} \quad with \quad m_\Delta = \frac{\sum_{j=1}^{N_{ik}} \left(S_{ijk} - \mu_{jk}\right)^\Delta}{N_{jk}} \end{aligned} \tag{6.3}$$

If the kurtosis value β_j falls between 2 and 4, the scores are regarded to be normally distributed. The subject rejection procedure is detailed in Figure 6.4. By performing the procedure, 1 out of 30 subjects and 3 out of 34 subjects are rejected in session 1 and session 2, respectively.

After subject rejection, Z-scores are then linearly rescaled to lie in the range of $[0, 100]$. Assuming that the Z-scores assigned by a subject are distributed as a standard Gaussian [17] [18], 99% of the scores will lie in the range $[-3, +3]$. Re-scaling is

accomplished by linearly mapping the range $[-3, +3]$ to $[0, 100]$ by:

$$\tilde{Z}_{ijk} = \frac{100(z_{ijk}+3)}{6} \tag{6.4}$$

Finally, the MOS value of each retargeted image is computed as the mean of the rescaled Z-scores, together with the standard deviation:

$$MOS_{jk} = \frac{1}{M_k} \sum_{i=1}^{M_k} \tilde{Z}_{ijk} \tag{6.5}$$

$$std_{jk} = \sqrt{\frac{1}{M_k - 1} \sum_{i=1}^{M_k} \left(\tilde{Z}_{ijk} - MOS_{jk}\right)^2}$$

where M_k is the number of remaining subjects of session k after the subject rejection. The MOS value together with the standard deviation is recorded for each retargeted image, which is recognized as the ground truth representing the retargeted image perceptual quality. They can be further analyzed and used for evaluating the performances of the quality metrics. The final subjective scores after conversion, with the standard deviation indicating the error bar, are illustrated in Figure 6.5.

As we mentioned above, the perceptual qualities of the retargeted images in the database should span the entire range of visual quality and exhibit good perceptual quality separation [16]- [18]. The histogram of the MOS values is shown in Figure 6.6. It can be observed that the perceptual qualities of the images range from low to high values. Also it demonstrates that the subjective study samples a range of perceptual quality in an approximately uniform fashion. The image perceptual qualities exhibit a good separation.

6.3.2 Analysis and Discussion of the Subjective Ratings

After the processing of the subjective ratings, the image retargeting database is built, which comprises the retargeted images and their corresponding MOS values. The database is analyzed from three aspects, specifically the scale, the retargeting method, and the source image content.

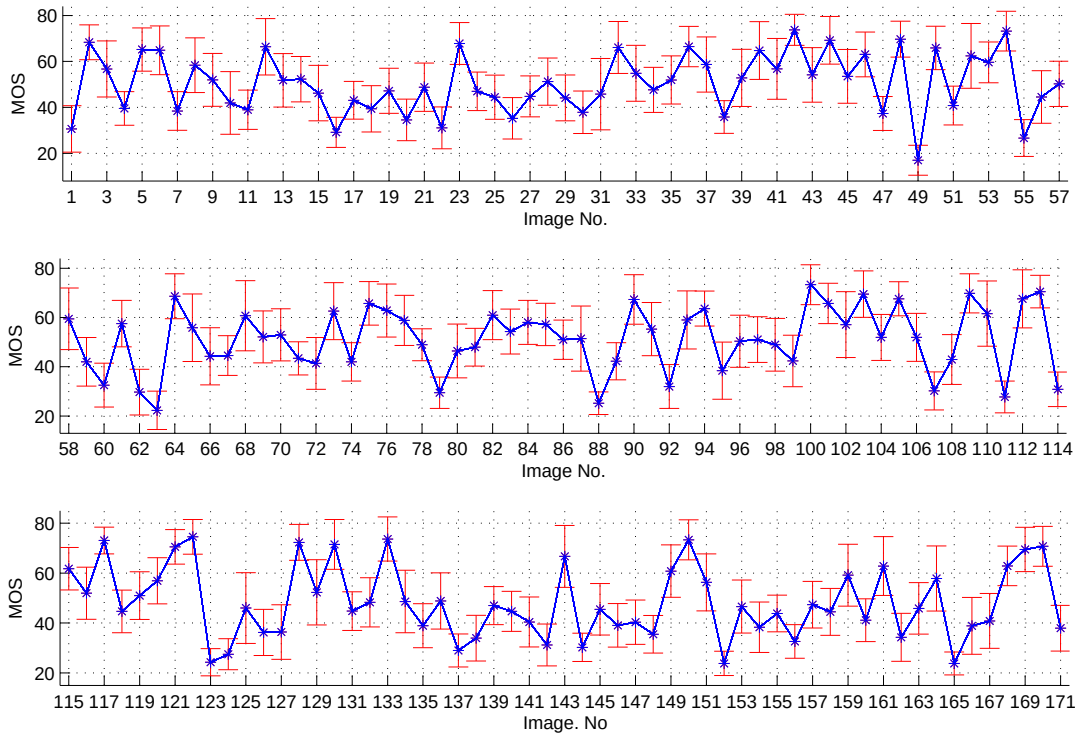


Figure 6.5: *The obtained MOS value of each retargeted image after processing (the horizontal axes corresponds to the image number, and the vertical axes corresponds to the MOS value. The blue asterisk indicates the obtained MOS value. And the red error bar indicates the standard deviation of the subjective scores).*

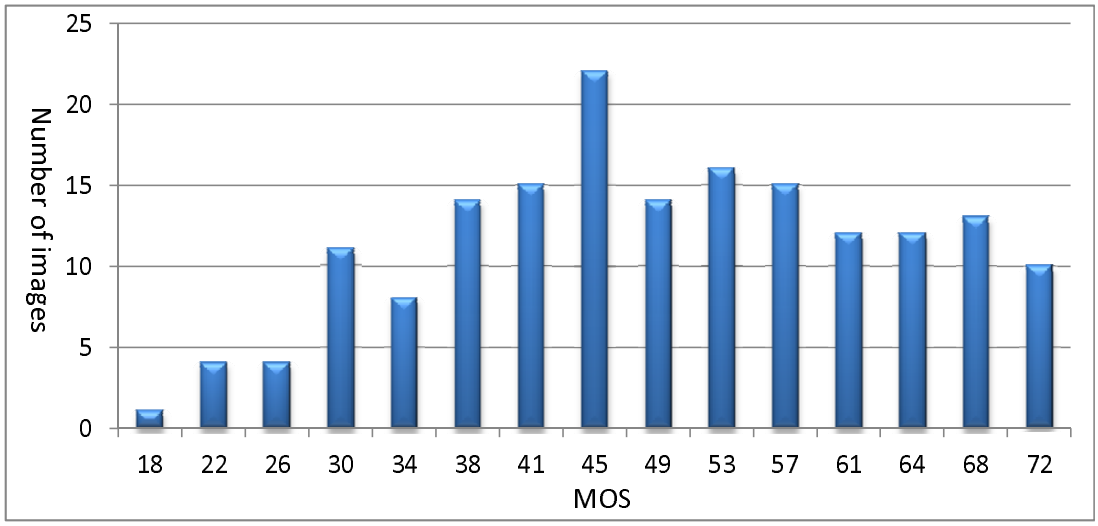


Figure 6.6: *Histogram of the MOS values in 15 equally spaced bins between the minimum and maximum MOS values of the image retargeting database.*

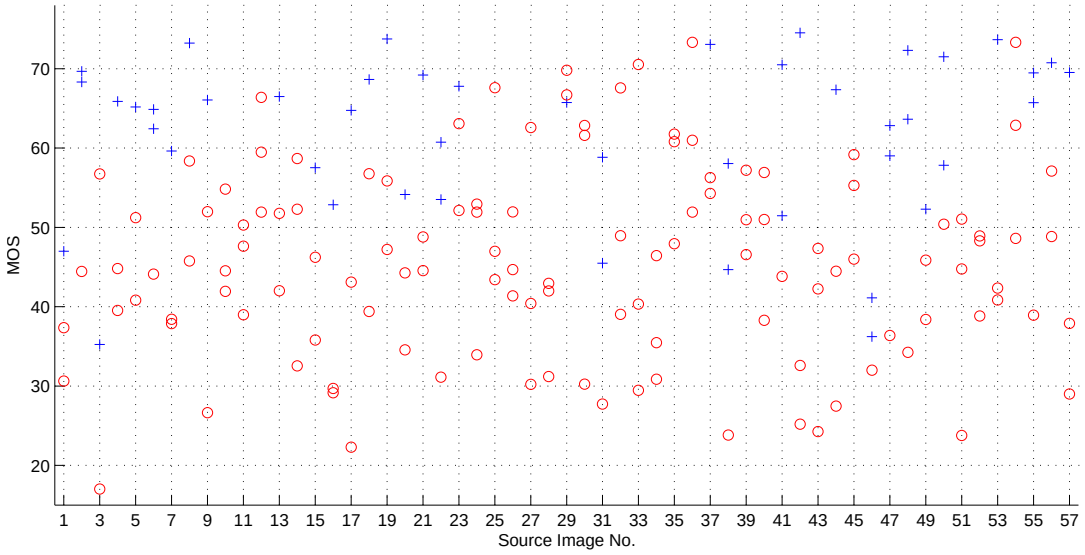


Figure 6.7: *The obtained MOS value versus the source image from the scale perspective. (The blue cross indicates the retargeted image in 75% scale; the red circle indicates the retargeted image in 50% scale).*

Retargeting Scale

The MOS values of the retargeted images in two different scales are illustrated in Figure 6.7. The detailed information of the image number and the corresponding source image can be found in the Appendix C. Generally, it can be observed that the retargeted images in 75% scale (with average MOS value as 61.79) exhibit higher perceptual quality than the retargeted images in 50% scale (with average MOS value as 45.66). There are two exceptions, which were generated from the source images 'kodim04' and 'bicycle1'. For the 'kodim04' containing the human face, the CROP method in 50% scale can preserve the shape information but sacrifice some content information, while the SCSC method in 75% will distort the human face. For 'bicycle1' with clear foreground object, the SEAM and SHIF methods in 50% scale will accurately preserve the shape and the content information, while the SCAL method in 75% scale will introduce some shape distortion. Therefore, the two images in 50% scale present better quality than the images in 75% scale. The reason is that the subjects prefer information loss rather than shape deformation.

Furthermore, it can be observed that the MOS values of the retargeted images in 75% scale are mostly larger than 50, except 'kodim01', 'kodim04', 'buddha', 'face', and

'kodim15. Referring to the attribute information of the source images, these images only contain either 'face and people' or 'geometric structure' attributes. It is known that human eyes are very sensitive to these attributes, which will greatly influence the perceptual quality of the retargeted image. For 'buddha' and 'face' images, other retargeting methods can generate higher quality images. Therefore, retargeting methods should be carefully selected for these images, which should not distort the shape information. For the retargeted images in 50% scale, the MOS values vary greatly.

Some source images, such as 'bicycle1' and 'eagle', generate retargeted images with very good quality. Also some source images, such as 'volleyball', generate retargeted images with very poor quality. Therefore the source image content will influence the perceptual quality of the retargeted images. Moreover, the retargeted images from the same source image also possess perceptual qualities with great differences, such as 'blueman'. It means that the retargeting method will also affect the image perceptual qualities. In the following sub-sections, the perceptual qualities of the retargeted images in 50% scale are analyzed from the two aspects: retargeting method, and source image content.

Retargeting Methods

As we discussed in the previous subsection, most of the algorithms produce the retargeted images in 75% scale with acceptable perceptual quality. In order to analyze the influence of the retargeting method, only the retargeted images in 50% scale are considered. The MOS values of the images by different retargeting methods are illustrated in Figure 6.8. As we mentioned before, the basic tools for retargeting are CROP, SCAL, WARP and SEAM. We firstly analyze these basic tools and then discuss the performances of the other methods.

The images generated by SEAM method [200]- [202] (denoted by the blue cross in Figure 6.8) are always of the worst perceptual quality. The reason is that the SEAM method tries to remove the seams in the regions with low gradient magnitudes. For some images, such as 'kodim04' and 'kodim15', some regions of the salient object appear to be very smooth, which will be discarded during the retargeting process. Therefore, some annoying shape distortion will be introduced. And as revealed by [34], the human subjects prefer sacrificing some image information rather than having deformation.

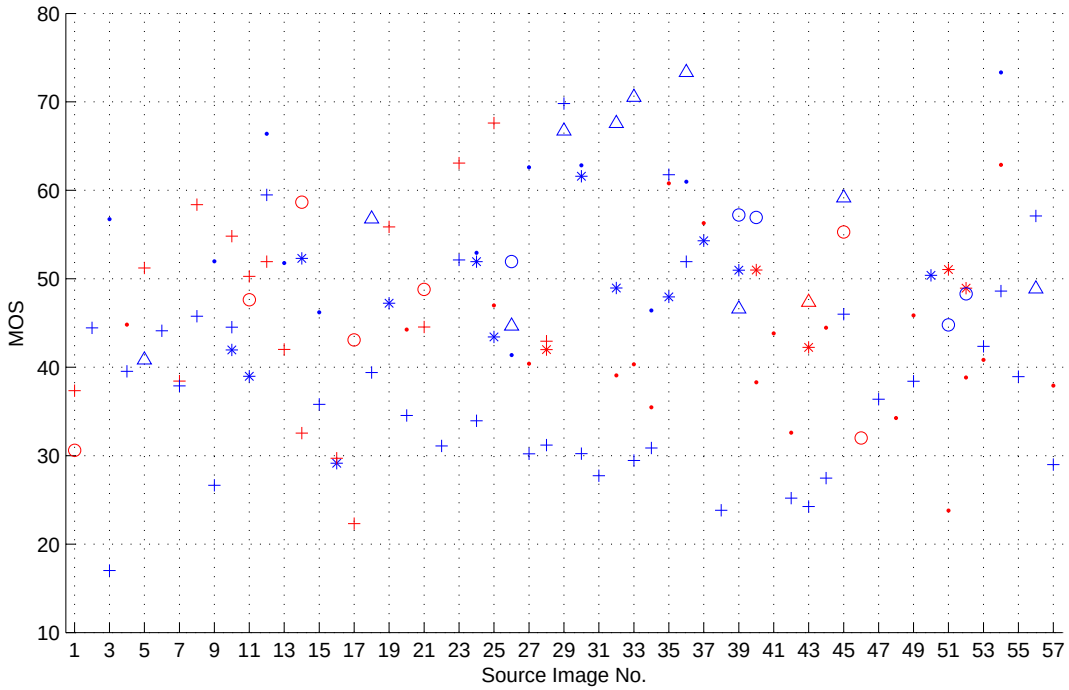


Figure 6.8: *The obtained MOS value versus the source image from the retargeting method perspective (in 50% scale). The blue dot is the CROP method; the blue star is the SCAL method, the blue cross is the SEAM method [200]- [202]; the blue triangle is the SHIF method [205]; the blue circle denotes the MULT algorithm [203]; the red dot denotes the WARP algorithm [198]; the red star denotes the ENER algorithm [206]; the red cross denotes the SCST [204]; the red triangle denotes the STVI method [199]; the red circle denotes the SCSC method [207].*

The SEAM method does not consider any approaches to preserve the object shape. Therefore, it exhibits the worst perceptual quality, especially for images containing salient objects.

The CROP method (denoted by the blue dot) can only retarget some images with good perceptual quality. As it only keeps part information of the source image, its performance depends on the source image content. For some images with a small region containing the salient content, the CROP method can retarget a good quality image, such as 'surfer'. For some images, such as 'perissa_ santorini', where all regions contain meaningful information, the CROP method retargets images with bad quality. In [34], the CROP method is suggested as the most reliable and simplest method to retarget images.

The WARP algorithm [198] (denoted by the red dot) tries to squeeze the source image to a target size by optimizing a warping function. The shape of the object cannot be preserved. Therefore, the retargeted images are of bad perceptual quality. In most

cases, it only outperforms SEAM method, while is inferior to other methods. The SCAL method (denoted by the blue star) retargets images with medium perceptual quality. It will introduce some shape deformation into the retargeted image, but not as severe as the SEAM and WARP method. Therefore, the SCAL method always outperforms SEAM and WARP, but worse than the other methods under study.

The other methods try to combine these basic tools together to produce an optimal retargeted image. Some methods, such as SCST [204] and SHIF [205], have considered using the saliency map to guide the retargeting. The shape information of the objects in the salient regions is preserved to avoid introducing unpleasant deformation. Therefore, these methods can obtain better performances. As shown in Figure 6.8, in most cases the SHIF algorithm (denoted by the blue triangle) and SCST (denoted by the red cross) can retarget the test images with better perceptual quality, compared with the other methods.

Source Image Contents

As mentioned above, the source images can be categorized by the containing attributes, which are 'face and people', 'clear foreground object', 'natural scenery', and 'geometric structure'. The 'clear foreground object' attribute is defined as the salient object occupying an image region smaller than 50% of the source image. If the salient object is preserved, the perceptual quality of the retargeted image (in 50% and 75% ratios) will not be very bad, as the crop margin (how much can be cropped without losing the object/regions of interest) is larger than 50%. The 'natural scenery' attribute is for an image with a large proportion of it containing the texture or smooth information. These images contain information with symmetric similar patterns. Therefore, cropping or scaling some part of the image will not introduce significant degradations in perceptual quality. The crop margin of these images is large. Therefore, the retargeted images in 50% and 75% ratios are of good perceptual quality. These two attributes are regarded as non-salient. The 'geometric structure' attribute denotes that there are evident edges or lines in the source image, and 'face and people' attribute means that the faces or persons occupy most regions of one source image. The subjective viewers will be very sensitive to the edges, shapes, and faces. The distortion introduced by the retargeting method will severely affect the judgment of the subjective viewer. For some images

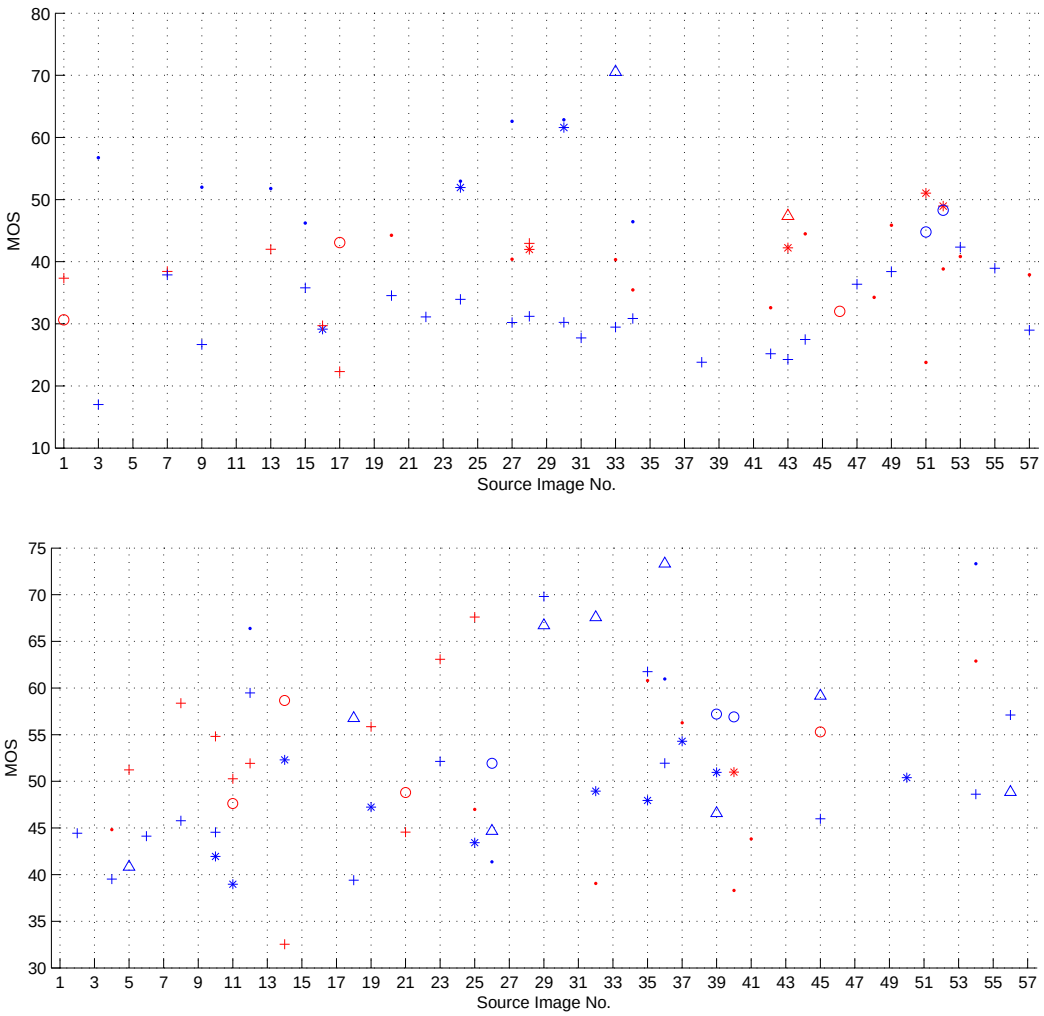


Figure 6.9: *The obtained MOS value versus the source image. Top: source images with salient attributes; bottom: source images with non-salient attributes*

containing 'face and people' attribute, such as the image 'face', 'kodim15', 'kodim04', and 'buddha', the entire image is a human face, which is of great importance. If we crop some part of the image, some important content is discarded, which will result in very bad perceptual quality. In this respect, the crop margin of these images is very small (nearly 0). Therefore, if we retarget these images with unsuitable methods, the perceptual quality will not be good. These two attributes are regarded as the salient attributes.

Each source image may contain more than one attribute. However, one attribute dominates each source image, while other attributes are not so significant. The detailed

attribute information of each source image is illustrated in the Appendix A. The attributes are sorted according to their significances. According to the attribute saliency, the source images are divided into two classes. Note that we only utilized the most significant attribute to classify the source images. After the separation, we obtained 30 images with salient attributes and the other 27 images with non-salient attributes. The MOS values versus the source images of different attributes are illustrated in Figure 6.9. In this subsection, as we only care about the influence of the image content on the perceptual quality, the retargeting methods are not considered. The retargeted image with the worst perceptual quality is utilized for comparison. They are all in the 50% scale, which ensures a fair comparison. We calculated the mean MOS values of the retargeted images in the two classes. The mean MOS value of the images with non-salient attributes is 45.55, which is higher than the average MOS value of the database. The images with non-salient attributes contain some texture and smooth information, such as 'kodim13' and 'fishing', which ensures a large crop margin value. Therefore, the shape deformation will not be easily detected. And the region discarded during the retargeting process mostly contains information with symmetric or similar patterns, or unimportant background information. Therefore, the perceptual quality will not be significantly influenced. However, the mean MOS value of the images with salient attributes is 31.1292, which is lower than that of non-salient attribute image. As most regions of the source image contain salient or meaningful information, the crop margin of such image is very small (nearly 0). And the contents and shapes of the objects, faces, or humans are critical for judging the perceptual quality. Retargeting these images into 50% scale will significantly distort the shapes or discard important content information. Therefore, the perceptual quality will be very unpleasant.

For the source images with salient attributes, Figure 6.9 shows that the CROP method always retargets image with the highest MOS values, such as 'kodim04' and 'sanfrancisco'. Although only a few source images employ the CROP method to retarget image, it can be deduced that the CROP will retarget the other images with highest quality as claimed by [34]. The reason is that the shape deformation is much more annoying to the human viewers, compared with the information discarded. For the salient attributes, such as 'face and human' and 'geometric structures', the shape distortion can be easily detected and rated badly by the subjective viewers. Therefore,

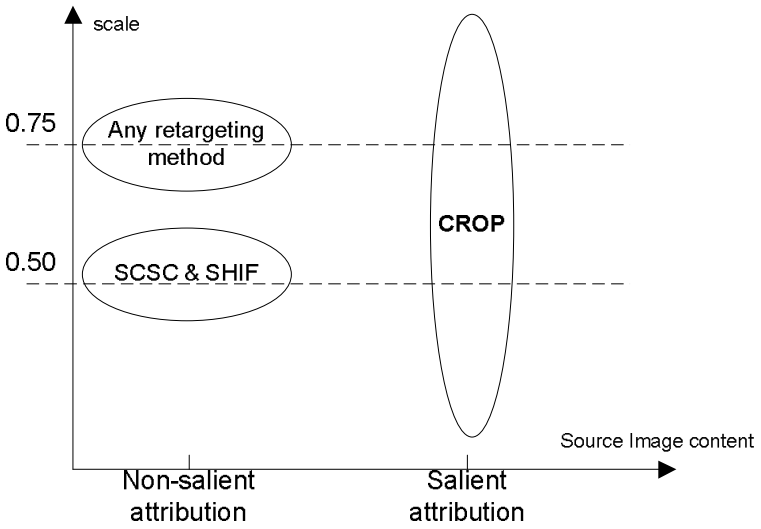


Figure 6.10: *Recommended retargeting methods by considering the retargeting scale and source image content.*

for the images containing salient attributes, the CROP process is recommended, not only because of its simplicity but also for its best performance. This is also applied to retarget image into 75% scale. For the images with salient attributes, such as 'kodim01', 'kodim04', 'buddha', 'face', and 'kodim15', the other methods other than CROP can introduce deformation to the object shape. That is the reason why the retargeted images in 75% are of low MOS values. For the images with non-salient attributes, most of the retargeted images are of good qualities, with MOS values larger than 45. However, there are several exceptions, such as 'perissa_ santorini', 'butterfly', and 'fishing'. It can be observed from Figure 6.9 that SCAL and SEAM also generate images with bad quality. The other methods, such as SCST, will preserve much more information, while the introduced shape deformation can be hardly detected by the human viewers.

Considering the above analysis, different retargeting methods are recommended for different images, as shown in Figure 6.10. For images with salient attributes, the CROP methods are suggested for its effectiveness and low complexity. For images with non-salient attribute, if retargeting them into 75% scale, all the retargeting method can generate acceptable results, because the shape distortion can hardly perceived and the loss of the image content is negligible. To retarget the images with non-salient attributes into 50% scale, we recommend the SCSC and SHIF method. They have considered the saliency map, which can help to preserve the object in the image.

6.4 Objective Quality Metric for Retargeted Images

6.4.1 Quality Metric Performances on the Constructed Image Retargeting Database

Image retargeting quality metric has been recently researched [214]- [220], in order to not only evaluate the retargeted image quality automatically and reliably in lieu of the subjective testing, but also help to improve the performance of the retargeting methods. One problem is that several quality metrics are licensed or patented, such as the bidirectional warping in [203], and the quality metric in [219], which are not made publicly. In this section, we only tested the metrics which are publicly available and suggested in [34], specifically the earth mover's distance (EMD) [214] [215], the bidirectional similarity (BDS) [216] [217], edge histogram (EH) [220], and SIFT-flow [218]. The information about the metrics is detailed in the following.

- EMD is based on the minimal cost that must be paid to transform one distribution into the other. The signature $\{S_j = (m_j, w_j)\}$, which represents a set of feature clusters, is viewed as the histogram distribution. The point m_j is the central value in bin j of the histogram, and w_j is to indicate the corresponding proportion. The definition of cluster is open. The color, position, and texture information can be employed to obtain the feature clusters. Only the size of the clusters in the feature space needs to be limited. Let $P = \{(p_1, w_{p_1}), \cdots, (p_m, w_{p_m})\}$ be the first signature with m clusters; $Q = \{(q_1, w_{q_1}), \cdots, (q_n, w_{q_n})\}$ is the second signature with n clusters. And $D = [d_{ij}]$ is the ground distance matrix, where d_{ij} is the ground distance between clusters p_i and q_j. d_{ij} can be any distance and will be chosen according to the problem at hand. The purpose is to find a flow $F = [f_{ij}]$, with f_{ij} as the flow between p_i and q_j, that minimizes the overall cost:

$$WORK(P, Q, F) = \sum_i^m \sum_j^n d_{ij} f_{ij} \tag{6.6}$$

  After obtaining the optimal flow F, EMD is defined as the work normalized by the total flow:

$$EMD(P, Q) = \frac{\sum_i^m \sum_j^n d_{ij} f_{ij}}{\sum_i^m \sum_j^n f_{ij}} \tag{6.7}$$

- Two signals S (original image) and T (retargeted image) are considered to be 'visually similar' if as many as possible patches of S (at multiple scales) are

contained in T, and vice versa. The dissimilarity can be formulated as:

$$(S,T) = \underbrace{\frac{1}{N_S}\sum_{P\subset S} min_{Q\subset T} D(P,Q)}_{d_{complete}(S,T)} + \underbrace{\frac{1}{N_T}\sum_{Q\subset T} min_{P\subset S} D(Q,P)}_{d_{cohere}(S,T)} \tag{6.8}$$

P and Q denote patches in S and T, respectively. And let N_S and N_T denote the number of patches in S and T. For each patch $Q \subset T$ we search for the most similar patch $P \subset S$, and measure their distance $D(P,Q)$, and vice-versa. The patches are taken around every pixel at multiple scales, resulting in significant patch overlap. $D(P,Q)$ can be any distance measurements between two patches, such as sum squared distances (SSD) or SSIM [68]. The two terms have important commentary roles. The first term, $d_{complete}(S,T)$ measures the deviation of the target T from 'completeness' w.r.t. S. Namely, it measures if all patches of S have been preserved in T. The second term $d_{cohere}(S,T)$ measures if there are any 'newborn' patches in T which have not originated from S. Therefore, the $d_{complete}(S,T)$ tries to represent the input image well (be complete), and the $d_{cohere}(S,T)$ makes sure the retargeted image is visually pleasing (coherent). The dissimilarity measurement is minimized in order to generate a retargeted image [216] [217].

- EH captures the spatial distribution of edges in the image. In order to depict the local edge distribution, the image is divided into 4×4 sub-images, each of which is examined by 5 different orientations: vertical, horizontal, two diagonals, and isotropic (non-directional). For each sub-image, a normalized 5-bin histogram is obtained by classifying apparent edges to these five categories. The feature is defined to be the combination of these histograms, which results in $4 \times 4 \times 5 = 80$ length description. Only the intensity component is employed for edge detection. And the L_1-norm distance is employed to measure the feature distance between two images, which is defined as $EH(S,T) = \| \ EHF(S) - EHF(T) \ \|_1$, where EHF is the edge histogram feature.

- SIFT-flow descriptors characterize view-invariant and brightness-independent image structures. Matching SIFT descriptors allows establishing meaningful correspondences across image with significantly different image content. Furthermore,

the pixel displacement (indicating by the SIFT correspondence matching) should be spatial coherent, which means that close-by pixels should have similar displacement. The cost function is defined as:

$$E(w) = \sum_p \parallel s_1(p) - s_2(p+w) \parallel_1 + \frac{1}{\sigma^2} \sum_p (\mu^2(p) + \nu^2(p)) + \tag{6.9}$$

$$\sum_{(p,q)\in\epsilon} \big(\min(\alpha|\mu(p) - \mu(q)|, d) + \min\big(\alpha|\nu(p) - \nu(q)|, d\big)\big)$$

where $w(p) = \big(\mu(p), \nu(p)\big)$ is the displacement vector at pixel location $p = (x, y)$, $s_i(p)$ is the SIFT descriptor extracted at location p in image i and ϵ is the spatial neighborhood of a pixel. SIFT flow employs the SIFT for feature matching. And the local smoothness is preserved by the vector difference constraint.

The algorithms are provided by the respective authors, which were tested on our built image retargeting quality database following the traditional evaluation process as introduced in Section 1.3.4. As usual, LCC, SROCC, RMSE, and OR statistical measurements are employed to indicate the corresponding performance. According to the definitions, larger values of LCC and SROCC mean that the objective and subjective scores correlate better, that is to say, a better performance of the metric. And the smaller RMSE and OR values indicate smaller errors between the two scores, therefore a better performance.

	EH	EMD	BSD	SIFT-flow	Fusion(EH,EMD and SIFT-flow	Fusion(EH,EMD BSD, SIFT-flow
LCC	0.3422	0.2760	0.2896	0.3141	0.4361	0.5217
SROCC	0.3288	0.2904	0.2887	0.2899	0.4203	0.4514
RMSE	12.686	12.977	12.922	12.817	12.149	11.484
OR	0.2047	0.1696	0.2164	0.1462	0.1462	0.1287

Table 6.1: *Performances of different metrics on the image retargeting database.*

The performances of different metrics are illustrated in Table 6.1. It can be observed that all of the metrics perform poorly on our database. For the EMD, the composed histogram only represents the feature distribution of the image, which cannot accurately depict the object shape and the content information of the image. Therefore, the shape distortions and content information loss, introduced during the retargeting process,

are not effectively described. BDS tries to capture how much information one image conveys of the other image in a bidirectional way. However, although it is claimed that the spatial geometric relationship is considered by a multiple scale approach, the order-relationship can still not be preserved, such as the local-order of each pixel or patch. Therefore, the dissimilarity metric of BDS does not accurately depict the object shape distortion either. SIFT-flow employs the SIFT descriptor to detect the correspondence between two images. It is claimed that the order-relationship of the pixels or patches is captured. However, the content information loss during the retargeting process is not considered. EH employs the edge histograms to describe the image, which are organized in order for comparison. EH can somehow represent the object shape information in the image. Same as the SIFT-flow, the content information loss is not accounted. These are the reasons why the metrics cannot perform effectively on our image retargeting database.

6.4.2 Subjective Analysis of the Shape Distortion and Content Information Loss

As shown in the previous sections, accounting for the object shape or content information loss alone cannot effectively evaluate the retargeted image quality. In order to investigate how the object shape and content information loss influence the perceptual quality, a subjective testing was designed.

During the first session of our subjective test, after the human subjects provided their personal opinions on the retargeted image quality, they were also asked to provide their personal opinions on the two distortion levels: (i) the level of shape distortion; (ii) the level of content information loss. The shape distortion describes the distortion, such as face deforming, object squeezing, object boundary discontinuity, and so on. The content information loss gives that part information of the object or content that is missing in the retargeted image, compared with the source image. Both of the two distortion levels are recorded in 5-scale, same as introduced in Section 6.2.3. After the subjective testing, not only the visual quality of the retargeted image is evaluated, but also the distortion levels of the two factors (shape distortion and content information loss) that may affect the visual quality are recorded.

Same as in Section 6.3, the level scores of the shape distortion and content information loss are processed independently by following the Z-score conversion, the subject

rejection, and Z-score inverse conversion. After these procedures, the level values are re-scaled in the range $[0, 100]$, same as the MOS values. LCC and SROCC between the level scores and the MOS values are utilized to evaluate their correlation, which is shown in Table 6.2. It can be observed that the level of shape distortion correlates much more closely with MOS values than the content information loss. It means that the viewers are more sensitive to the shape distortions introduced in the retargeted images. In most cases, the human subjects tend to sacrifice the information loss rather than the shape distortion for recognizing a good quality image. For the information loss, although it correlates badly with the MOS values, it still affects the visual quality of the retargeted image.

	LCC	SROCC
MOS *vs.* Shape Distortion	0.8243	0.8371
MOS *vs.* Content Information Loss	0.3264	0.4680
MOS *vs.* Fusion of Shape Distortion and Content Information Loss	0.9218	0.9267

Table 6.2*: Relationship between MOS values and the levels of shape distortion and information loss.*

From Table 6.2, it can be observed that the shape distortion correlates closely with the final perceptual quality of the retargeted image. However, the three metrics, EH, EMD, and SIFT-flow describing the shape distortion do not prove to be efficiency, as shown in Table 6.1. The reason may be attributed to that none of them are able to accurately capture the shape distortion. Therefore, a fusion strategy is tested by combining the three metrics together through average process. The performance is also illustrated in Table 6.1. Compared with the three metrics, the fusion one performs better. It means that the current descriptor for capturing the shape distortion is not accurate enough. Furthermore, a fusion strategy by summing the shape distortion and content information loss together was tested. As shown in Table 6.2, the fusion result correlates more closely with the MOS value. The observation provides us some hints for designing the quality metric from the perspective of shape distortion and content information loss. The descriptors of shape distortion and content information loss should be combined together for evaluating retargeted image quality. For the current available metrics, EH, EMD and SIFT-flow tries to capture the object shape of the image. BSD tries to depict the content information loss in a bidirectional way. If

they are combined together, these two distortions are considered to build a quality metric, the performance of which is illustrated in Table 6.1. It can be observed that a better performance is obtained, which means that considering the shape distortion and content information loss together can help to improve the performances.

6.4.3 Discussion

As demonstrated in previous subsections, the performances of the objective quality metrics for retargeted images are still not good enough. The statistical correlations between the subjective MOS values and the metric outputs are not close. Even fusing EH, EMD, BSD, and SIFT-flow together, the LCC and SROCC values are smaller than 0.6, which indicates a bad performance of the objective metric. In this sub-section, we will discuss and try to figure out how to design an effective objective quality metric for evaluating the perceptual quality of the retargeted image. The source image content, retargeting scale, the shape distortion and content information loss measurement, and the HVS properties are the candidate factors, which are believed to benefit the objective metric performance.

- Shape distortion description. As illustrated Table 6.2, the shape distortion is closely related to the perceptual quality of the retargeted image. Therefore, the recently developed metrics, such as EH, EMD, and SIFT-flow, try to capture the object shape of the image and measure the corresponding differences between the source and retargeted image. However, the performances are not good enough, where the LCC and SROCC values are only about 0.35 as shown in Table 6.1. Even combining these metrics together, we can obtain a better performance; but the result is still unsatisfactory. Therefore, in order to accurately depict the perceptual quality of the retargeted image, the shape distortions that introduced by retargeting process need to be captured more precisely. Recently, A. D'Angelo [221] [222] proposed a full-reference quality metric to evaluate the geometrical distortions of the images. The approaches are based on that the HVS is sensitive to the image structures, such as edges and bars, which are identified by employing the Gabor filter. By considering this descriptor for evaluating the geometrical distortion, the shape distortion introduced during the retargeting process is believed to be more accurately described. Therefore, it can help to

improve the performance of the objective quality metric.

- Fusion of the shape distortion and content information loss. As illustrated in Table 6.2, the content information loss alone is not closely related to the final perceptual quality of the retargeted image. But combining the shape distortion and content information loss together can improve the performance, which has also been illustrated in Table 6.1. The combinations of the four objective quality metrics can beat the other metrics. Therefore, if we develop accurate metrics to capture the shape distortion and content information loss, how to fuse them together needs to be further considered. The fusion strategy of the two factors should consider their corresponding contributions to the final retargeted image quality.

- Source image quality and retargeting scale. The source images that we employed to build our database are of different resolutions and different qualities, which may affect the subjective viewers' judgment of the retargeted image perceptual quality. Moreover, the retargeting scale will also affect the retargeted image quality. Given one source image, the larger the retargeting ratio, the better is the perceptual quality of the retargeted image. Therefore, the final perceptual quality index of the retargeted image needs to account for the quality of the source image as well as the retargeting scale.

- Image content. As discussed in previous sections, the image content correlates closely to the crop margin of the source image (how much can be cropped without losing the object/regions of interest). If the source image contains the 'clear foreground object' or 'natural scenery' attribute, the crop margin will be very large. Therefore, retargeting the source image into 75% and 50% ratios will not significantly affect the perceptual quality. Otherwise, if the source image contains the 'face and people' or 'geometric structure' attribute, the crop margin will be very small. Then any retargeting methods will severely degrade the perceptual quality. In this respect, the image content and the crop margin of each source image need to be included to depict the perceptual quality of the retargeted image.

- HVS saliency. Additionally, the HVS demonstrates different conspicuities over

different regions of the image. The shape distortions and content information loss in the salient regions are more sensitively perceived by the viewers than those in the non-salient regions. That is also the reason why several retargeting methods consider the saliency or visual attention map during the retargeting process, such as WARP [198], SCST [204], and STVI [199]. The viewers' assessment on the quality of the retargeted image is prejudiced during the subjective testing process. Therefore, the effect of the HVS saliency needs to be considered to model the subjective viewer's behavior, which will lead to a more effective quality metric for retargeted images. The simplest way of incorporating the HVS saliency is to weight the corresponding shape distortion and content information loss by the saliency map detected from the source image, which has been demonstrated to be effective in evaluating the perceptual quality of the traditional distorted image.

6.5 Conclusion

An image retargeting database is built through the subjective study in this chapter. Based on the subjective ratings of the human viewers, the database is analyzed from the perspectives of retargeting scale, retargeting method, and source image content. Also the publicly available quality metrics for the retargeted images are evaluated on the constructed database. By combining the metrics together, which independently depict shape distortion and content information loss, the performance can be improved.

Chapter 7

Conclusions

This thesis mainly discusses perceptual quality assessment and processing for visual signals. A successful perceptual quality metric can release human beings from laborious works, such as visual quality monitoring in communication, visual system performance evaluation, vision-related tests in manufacturing environment, etc. Also the perceptual quality metric can employed to optimize the performances of many image/video processing applications, such as visual signal compression, communication, watermarking, and so on. In this chapter, we will conclude our work in Section 7.1. And the future work will be discussed in Section 7.2.

7.1 Conclusion

- **Visual Horizontal Effect**

  In Section 2.1, visual horizontal effect (HE) are researched to address the HVS sensitivities to stimuli of different orientations over contents of different orientations. The visual HE is simply modeled by a polynomial function based on the obtained psycho-visual data. The visual HE is further employed to rectify the structural distortion map. Experimental results demonstrate that the visual HE modeling the HVS orientation sensitivity can improve the PQA performance.

- **Adaptive Block-Based Super-Resolution Directed Down-Sampling for Image Compression**

  In Section 2.2, a novel perceptual image coding scheme via adaptive block-based super-resolution directed down-sampling is proposed. For each MB of a given image, whether down-sampling or not depends on the contents of the visual signal itself, which will be determined by the rate distortion optimization (RDO)

process [104]. And the joint method of down-sampling and super-resolution is proposed to minimize the reconstruction errors between the original and the restored MB inferred by the super-resolution method from the down-sampled block. At the decoder side, the super-resolution method performed in DCT domain is employed to recover the full-resolution MB for its simplicity. Experimental results demonstrated that images with much better subjective and objective quality can be constructed with a small number of computations introduced.

- **Adaptive Block-Based Just Noticeable Difference**

  In Section 3.1, extension from 8×8 DCT-based JND to 16×16 DCT-based JND is performed by conducting a psychophysical experiment to parameterize the CSF for the 16×16 DCT. For still images or the intra video frames, a new spatial selection strategy based on the spatial content similarity (SCS) is utilized to yield the JND map. For the inter video frames, a temporal selection strategy based on the motion characteristic similarity (MCS) is employed to determine the transform size for generating the JND map. With the developed ABT-based JND, a simple PQA is derived in Section 3.2. By evaluating on the publicly available databases, the metric is believed to be reliable for evaluating perceptual qualities of the visual signals. Furthermore, as the proposed PQA is very simple, it can be easily integrated into the video coding strategy for perceptual-based video coding in Section 3.4. And experimental results demonstrate that the proposed method can generate higher quality video sequences in terms of both objective and subjective measurements.

- **Motion Trajectory Based Visual Saliency Map for Quality Assessment**

  In Section 3.3, we propose to incorporate the motion trajectory for efficiently detecting the visual saliency of video sequences. A quaternion representation (QR) for each frame is constructed, which comprises the spatial image content, the motion trajectories, and the temporal residuals. Based on the QR, the quaternion Fourier transform (QFT) is employed to construct the visual saliency. Finally, the visual saliency is incorporated with several video quality metrics for evaluating its efficiency. Experimental results demonstrate that the proposed visual saliency map can improve the performances of the video quality metrics.

- **Reduced Reference Image Quality Assessment**

  In Chapter 4, we proposed an efficient RR IQA, which can evaluate the perceptual quality of the image based on a limited number of bits. The statistical dependencies between the DCT subbands after performing DCT still exist. Applying the reorganization strategy, the intra RDCT subband statistical characteristic, specifically the identical natural of the coefficient distribution within the RDCT subband, is exploited by the GGD modeling. The inter RDCT subband dependency is captured by the mutual information (MI) between the DCT coefficient pair in corresponding RDCT subbands, such as parent-child pair coefficient, brother-child pair coefficient, and cousin-child pair coefficient. And a frequency ratio descriptor (FRD) computed in the RDCT domain is employed to measure the energy distribution among different frequency components. It can be further utilized to simulate the HVS texture masking property. By considering the intra RDCT subband GGD modeling, inter RDCT subband MI values, and the image FRD value, an effective RR IQA is developed. Experimental results demonstrate that the proposed RR IQA outperforms the representative RR IQAs, and even the FR IQAs, such as PSNR, and SSIM.

- **Reduced Reference Video Quality Assessment**

  In Chapter 5, the study deals with the RR quality assessment for compressed video sequences by extending the previous work RR IQA introduced in Chapter 4. Firstly, from the spatial perspective, an energy variation descriptor (EVD) is proposed to measure the energy change of each distorted frame. The proposed EVD can also be utilized to simulate the texture masking property of the HVS. For the temporal distortion, the generalized Gaussian distribution (GGD) is employed to model the histogram distribution of the inter frame difference. The city-block distance (CBD) is used to calculate the histogram difference between the original video and the distorted one. Finally, the perceptual quality index is derived by combining the spatial EVD together with temporal CBD. Also these EVD and GGD features are efficiently encoded and represented with a small number of bits. And experimental results demonstrate that the proposed RR VQA outperforms the representative RR metrics VQM and also other quality metrics. Due to its

simplicity and efficiency in terms of feature representation, the proposed metric can be considered to be incorporated into the video quality monitoring system.

- **Image Retargeting Perceptual Quality Assessment**

  In Chapter 6, a subjective study is conducted to assess the perceptual quality of the retargeted image to build a publicly available database. Totally, 171 retargeted images (in two different scales) are generated by different retargeting methods from 57 source images. With the source image as the reference, the perceptual quality of each retargeted image has been subjectively rated by at least 30 human viewers on a pre-defined scale. After processing the subjective ratings, the MOS value and the corresponding standard deviation are obtained for each image. Based on the MOS values, the built image retargeting database is analyzed from the perspectives of the retargeting scale, the retargeting method, and the source image content. Moreover, some publicly available quality metrics for retargeted images are evaluated on the built database in the standardized way. Furthermore, a specifically designed subjective testing process is carried out to provide further information for developing an effective quality metric for retargeted images.

7.2 Future Work

- **Quality monitoring system**

  As discussed in Chapter 4 and Chapter 5, RR PQAs are designed for on-line monitoring the perceptual quality of the visual signals. In order to develop an effective and efficient quality monitoring system, the performances of RR PQAs, compression and transmission of RR features, robustness of the RR feature to the distortions introduced during the transmission need to be considered. This will be part of our future work, which can provide better quality of experience for users.

- **NR metrics**

  As discussed in Chapter 1, in many real-world applications, we cannot access the reference visual signal for the quality evaluation, such as image/video denoising, restoration, etc., where only the distorted visual signal is available for analysis. Therefore, the NR PQAs are thus needed to evaluate and control the perceptual

quality of the processed image. In the future, NR PQAs will be researched to better handle the practical applications, such as image/video denoising, super-resolution, and so on.

- **Quality metrics for retargeted images and the retargeting algorithms**

  As discussed in Chapter 6, how to evaluate the perceptual quality of the retargeted image is still under investigation. In the future, we will consider, HVS saliency property, shape distortion description, source image content, etc., to develop an accurate quality metric. With the effective quality metric, retargeting algorithms can be developed, which can preserve the salient and semantic information of the image content, while no distortions are introduced.

- **High definition, 3D and mobile**

  As electronic and communication techniques progress with a surprisingly rapid speed, consumers are no longer satisfied with traditional standard-definition video services. New types of multimedia services are being developed and delivered to satisfy all kinds of needs of the end users. High definition and 3D videos can provide us with more realistic and immersive viewing experiences. They will become dominant in the consumer market in the foreseeable future. Meanwhile, mobile has become an indispensable part of everyone's life. As mobile services become much cheaper and faster, more and more videos will be watched on mobile devices. Therefore, there are great demands for accurate subjective and objective quality assessment methods for evaluating visual quality of high definition, 3D, and mobile videos. To this end, we need to investigate how the display resolution, the viewing distance, the environment brightness, etc., affect the perceptual visual quality. Further investigation will provide deeper understanding and more valuable information about how to optimize these fast developing multimedia services.

- **Perception-based applications**

  As introduced in Chapter 1, perceptual visual quality assessment can be used to compare system performances, monitor multimedia service quality, and develop perception-based image/video applications. As the accuracy of perception models

improves, more and more perception-based applications have been developed. For example, we have done some work on perceptual video coding by implementing the ABT-based JND into the H.264 video codec. With deeper understanding of the processing mechanism of the human visual system, more accurate perception models will be developed. How to incorporate these perception models seamlessly into many imaging and computer vision applications should be a promising research direction.

Appendix A

Attributes of the Source Image

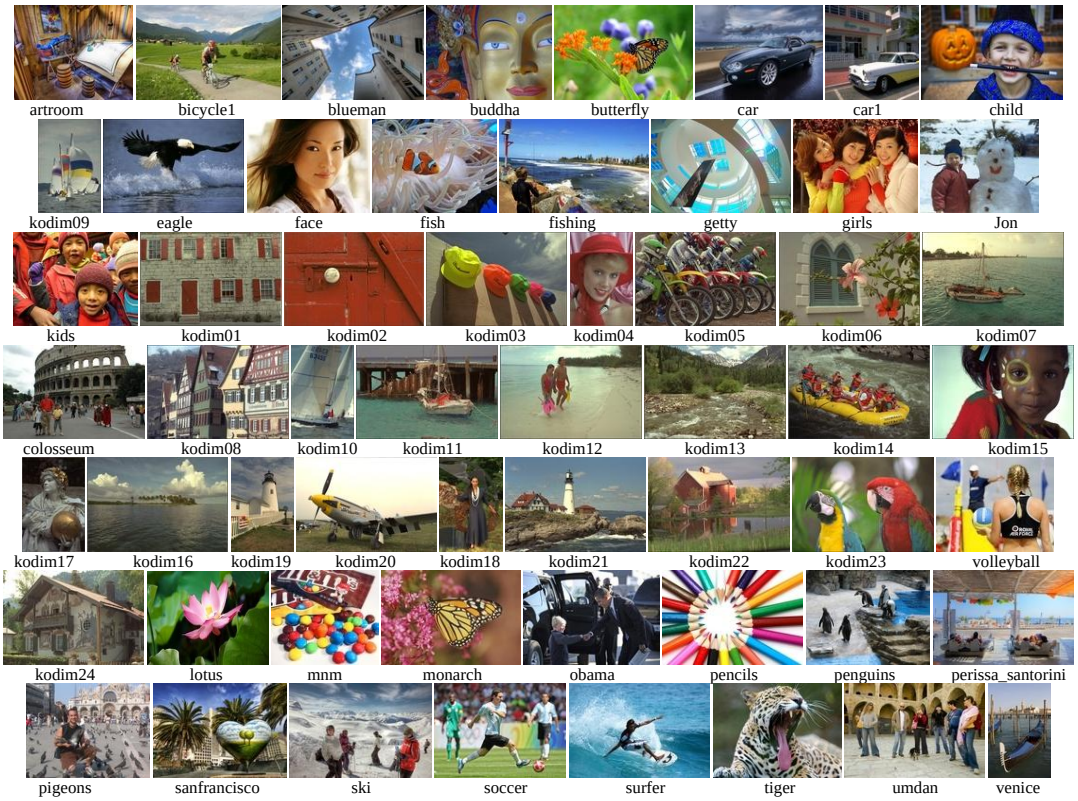


Figure A.1: *Source images for building the image retargeting database.*

The source images for building the image retargeting database are illustrated in Figure A.1. We have considered four attributes, specifically 'face and people', 'clear foreground object', 'natural scenery', and 'geometric structure'. We define the clear foreground object attribute as that the salient object should occupy the image region smaller than 50% of the source image. The 'natural scenery attribute means that a large proportion of the image contains the texture or smooth information. And the 'geometric structure

attribute denotes that there are evident edges or lines in the source image. The detailed attribute information of each source image is illustrated in Table A.1. Firstly, it can be observed that one image may contain more than one attributes. Secondly, the dominant attribute of each source image is illustrated. We sort the attributes of each source image according to the attribute significance.

Table A.1: *The attribute information of the source image. (1 indicates the attribute of 'face and people'; 2 indicates the attribute of 'clear foreground object'; 3 indicates the attribute of 'natural scenery'; 4 indicates the attribute of the 'geometric structure'. The attributes are sorted according to their corresponding significances. The left attribute denotes the most significant, while the right one is the least significant.)*

Source Image Name	Attributes	Source Image Name	Attributes
'kodim01.png'	4	'kodim22.png'	4;3
'kodim03.png'	2	'kodim23.png'	2
'kodim04.png'	1	'kodim24.png'	4
'kodim05.png'	3;1	'monarch.png'	2
'kodim06.png'	2;3	'ArtRoom.png'	4
'kodim07.png'	2;3	'Lotus.png'	2
'kodim08.png'	4	'Perissa_ Santorini.png'	3
'kodim09.png'	2;3	'Sanfrancisco.png'	4;3
'kodim10.png'	4	'Umdan.png'	1;4
'kodim11.png'	3;2	'bicycle1.png'	2;1;3
'kodim12.png'	2;1;3	'blueman.png'	4
'kodim13.png'	3	'buddha.png'	1
'kodim14.png'	1;2;3	'butterfly.png'	2
'kodim16.png'	3	'car1.png'	4
'kodim17.png'	1;4	'car.png'	4;2
'kodim18.png'	1;3	'child.png'	2;1
'kodim19.png'	4	'colosseum.png'	2;1;4
'kodim20.png'	2;4	'eagle.png'	2
'kodim21.png'	3;4	'face.png'	1
'obama.png'	1	'fish.png'	2
'pencils.png'	4	'fishing.png'	3;1

Table A.1 – continued from previous page

'penguins.png'	3	'getty.png'	3;4
'pigeons.png'	1;2;4	'girls.png'	1
'ski.png'	1;3	'jon.png'	1
'soccer.png'	1	'kids.png'	1
'surfer.png'	2;1	'kodim02.png'	3;4
'tiger.png'	4	'kodim15.png'	1
'venice.png'	2;3	'mnm.png'	4
'volleyball.png'	1		

Appendix B

Retargeted Image Name and the Corresponding Number

There are in total 171 retargeted images in the built database. The retargeted image name and the corresponding number are shown in Table B.1, which corresponds to the image no. listed in Figure 6.3 and Figure 6.4 of the thesis. It can be observed that each source image generates 3 retargeted images by different methods and in different scales.

Table B.1: *171 retargeted images and their corresponding image no.*

Retargeted image name	Image No.	Retargeted image name	Image No.
'kodim01_ scsc_ 0.50.png'	1	'getty_ seam_ 0.75.bmp'	87
'kodim03_ scst_ 0.75.bmp'	2	'girls_ seam_ 0.50.bmp'	88
'kodim04_ crop_ 0.50.bmp'	3	'jon_ ener_ 0.50.png'	89
'kodim05_ seam_ 0.50.bmp'	4	'kids_ crop_ 0.75.png'	90
'kodim06_ scsc_ 0.75.png'	5	'kodim02_ scsc_ 0.50.png'	91
'kodim07_ scst_ 0.75.bmp'	6	'kodim15_ scsc_ 0.50.png'	92
'kodim08_ scst_ 0.50.bmp'	7	'mnm_ scst_ 0.75.png'	93
'kodim09_ scst_ 0.50.bmp'	8	'obama_ ener_ 0.75.png'	94
'kodim10_ crop_ 0.50.bmp'	9	'pencils_ seam_ 0.50.bmp'	95
'kodim11_ scal_ 0.50.bmp'	10	'penguins_ scal_ 0.50.bmp'	96
'kodim12_ scal_ 0.50.bmp'	11	'pigeons_ ener_ 0.50.png'	97
'kodim13_ crop_ 0.50.bmp'	12	'ski_ ener_ 0.50.png'	98
'kodim14_ crop_ 0.50.bmp'	13	'soccer_ seam_ 0.50.bmp'	99
'kodim16_ scal_ 0.50.bmp'	14	'surfer_ crop_ 0.50.bmp'	100

Table B.1 – continued from previous page

'kodim17_ crop_ 0.50.bmp'	15	'tiger_ ener_ 0.75.png'	101
'kodim18_ scal_ 0.50.bmp'	16	'venice_ seam_ 0.50.bmp'	102
'kodim19_ scsc_ 0.50.png'	17	'volleyball_ mult_ 0.75.png'	103
'kodim20_ seam_ 0.50.bmp'	18	'ArtRoom_ scal_ 0.50.bmp'	104
'kodim21_ scal_ 0.50.bmp'	19	'Lotus_ scst_ 0.50.png'	105
'kodim22_ seam_ 0.50.bmp'	20	'Perissa_ Santorini_ mult_ 0.50.png'	106
'kodim23_ scsc_ 0.50.png'	21	'Sanfrancisco_ seam_ 0.50.bmp'	107
'kodim24_ seam_ 0.50.bmp'	22	'Umdan_ scst_ 0.50.png'	108
'monarch_ crop_ 0.75.bmp'	23	'bicycle1_ seam_ 0.50.bmp'	109
'kodim01_ scsc_ 0.75.png'	24	'blueman_ scal_ 0.50.bmp'	110
'kodim03_ seam_ 0.50.bmp'	25	'buddha_ seam_ 0.50.bmp'	111
'kodim04_ scsc_ 0.75.png'	26	'butterfly_ shif_ 0.50.png'	112
'kodim05_ warp_ 0.50.bmp'	27	'car1_ shif_ 0.50.bmp'	113
'kodim06_ scst_ 0.50.bmp'	28	'car_ seam_ 0.50.bmp'	114
'kodim07_ seam_ 0.50.bmp'	29	'child_ seam_ 0.50.bmp'	115
'kodim08_ seam_ 0.50.bmp'	30	'colosseum_ seam_ 0.50.bmp'	116
'kodim09_ seam_ 0.50.bmp'	31	'eagle_ seam_ 0.75.bmp'	117
'kodim10_ scsc_ 0.75.png'	32	'face_ scst_ 0.75.png'	118
'kodim11_ scst_ 0.50.bmp'	33	'fish_ scal_ 0.50.png'	119
'kodim12_ scsc_ 0.50.png'	34	'fishing_ mult_ 0.50.png'	120
'kodim13_ scst_ 0.50.bmp'	35	'getty_ stvi_ 0.75.png'	121
'kodim14_ scal_ 0.75.bmp'	36	'girls_ stvi_ 0.75.png'	122
'kodim16_ scsc_ 0.50.png'	37	'jon_ seam_ 0.50.png'	123
'kodim17_ seam_ 0.50.bmp'	38	'kids_ seam_ 0.50.bmp'	124
'kodim18_ scsc_ 0.75.png'	39	'kodim02_ seam_ 0.50.bmp'	125
'kodim19_ scsc_ 0.75.png'	40	'kodim15_ seam_ 0.75.bmp'	126
'kodim20_ shif_ 0.50.bmp'	41	'mnm_ seam_ 0.50.bmp'	127
'kodim21_ scsc_ 0.75.png'	42	'obama_ mult_ 0.75.png'	128

Table B.1 – continued from previous page

'kodim22_ seam_ 0.75.bmp'	43	'pencils_ stvi_ 0.75.png'	129
'kodim23_ scsc_ 0.75.png'	44	'penguins_ scst_ 0.75.png'	130
'kodim24_ seam_ 0.75.bmp'	45	'pigeons_ mult_ 0.50.png'	131
'monarch_ scst_ 0.50.bmp'	46	'ski_ mult_ 0.50.png'	132
'kodim01_ scst_ 0.50.bmp'	47	'soccer_ seam_ 0.75.bmp'	133
'kodim03_ seam_ 0.75.bmp'	48	'surfer_ seam_ 0.50.bmp'	134
'kodim04_ seam_ 0.50.bmp'	49	'tiger_ seam_ 0.50.bmp'	135
'kodim05_ warp_ 0.75.bmp'	50	'venice_ shif_ 0.50.bmp'	136
'kodim06_ shif_ 0.50.bmp'	51	'volleyball_ seam_ 0.50.bmp'	137
'kodim07_ seam_ 0.75.bmp'	52	'ArtRoom_ seam_ 0.50.bmp'	138
'kodim08_ seam_ 0.75.bmp'	53	'Lotus_ warp_ 0.50.png'	139
'kodim09_ seam_ 0.75.bmp'	54	'Perissa_ Santorini_ shif_ 0.50.png'	140
'kodim10_ seam_ 0.50.bmp'	55	'Sanfrancisco_ warp_ 0.50.bmp'	141
'kodim11_ seam_ 0.50.bmp'	56	'Umdan_ seam_ 0.50.png'	142
'kodim12_ scst_ 0.50.bmp'	57	'bicycle1_ shif_ 0.50.bmp'	143
'kodim13_ seam_ 0.50.bmp'	58	'blueman_ seam_ 0.50.bmp'	144
'kodim14_ scst_ 0.50.bmp'	59	'buddha_ seam_ 0.75.png'	145
'kodim16_ scst_ 0.50.bmp'	60	'butterfly_ warp_ 0.50.png'	146
'kodim17_ warp_ 0.75.bmp'	61	'car1_ warp_ 0.50.bmp'	147
'kodim18_ scst_ 0.50.bmp'	62	'car_ warp_ 0.50.bmp'	148
'kodim19_ scst_ 0.50.bmp'	63	'child_ warp_ 0.50.bmp'	149
'kodim20_ warp_ 0.75.bmp'	64	'colosseum_ shif_ 0.50.bmp'	150
'kodim21_ scst_ 0.50.bmp'	65	'eagle_ warp_ 0.50.bmp'	151
'kodim22_ warp_ 0.50.bmp'	66	'face_ seam_ 0.50.bmp'	152
'kodim23_ scst_ 0.50.bmp'	67	'fish_ shif_ 0.50.png'	153
'kodim24_ shif_ 0.75.bmp'	68	'fishing_ warp_ 0.50.png'	154
'monarch_ seam_ 0.50.bmp'	69	'getty_ warp_ 0.50.bmp'	155
'ArtRoom_ crop_ 0.50.bmp'	70	'girls_ warp_ 0.50.bmp'	156

Table B.1 – continued from previous page

'Lotus_ scal_ 0.50.png'	71	'jon_ stvi_ 0.50.png'	157
'Perissa_ Santorini_ crop_ 0.50.png'	72	'kids_ warp_ 0.50.bmp'	158
'Sanfrancisco_ crop_ 0.50.bmp'	73	'kodim02_ shif_ 0.50.bmp'	159
'Umdan_ ener_ 0.50.png'	74	'kodim15_ shif_ 0.75.bmp'	160
'bicycle1_ scal_ 0.75.bmp'	75	'mnm_ stvi_ 0.75.png'	161
'blueman_ crop_ 0.50.bmp'	76	'obama_ warp_ 0.50.bmp'	162
'buddha_ mult_ 0.75.png'	77	'pencils_ warp_ 0.50.bmp'	163
'butterfly_ scal_ 0.50.png'	78	'penguins_ seam_ 0.75.bmp'	164
'car1_ seam_ 0.50.bmp'	79	'pigeons_ warp_ 0.50.png'	165
'car_ crop_ 0.50.bmp'	80	'ski_ warp_ 0.50.png'	166
'child_ scal_ 0.50.bmp'	81	'soccer_ warp_ 0.50.bmp'	167
'colosseum_ crop_ 0.50.bmp'	82	'surfer_ warp_ 0.50.bmp'	168
'eagle_ scal_ 0.50.bmp'	83	'tiger_ stvi_ 0.75.png'	169
'face_ ener_ 0.75.png'	84	'venice_ warp_ 0.75.png'	170
'fish_ mult_ 0.50.png'	85	'volleyball_ warp_ 0.50.bmp'	171
'fishing_ ener_ 0.50.png'	86		

Appendix C

Source Image Name and the Corresponding Number

The source image and the corresponding image no. are shown in Table C.1, which corresponds to the image no. listed in the Figure 6.7, Figure 6.8, and Figure 6.9 of the thesis.

Table C.1: *57 source images and their corresponding image no.*

Source Image Name	Image No.	Source Image Name	Image No.
'kodim01.png'	1	'blueman.png'	30
'kodim03.png'	2	'buddha.png'	31
'kodim04.png'	3	'butterfly.png'	32
'kodim05.png'	4	'car1.png'	33
'kodim06.png'	5	'car.png'	34
'kodim07.png'	6	'child.png'	35
'kodim08.png'	7	'colosseum.png'	36
'kodim09.png'	8	'eagle.png'	37
'kodim10.png'	9	'face.png'	38
'kodim11.png'	10	'fish.png'	39
'kodim12.png'	11	'fishing.png'	40
'kodim13.png'	12	'getty.png'	41
'kodim14.png'	13	'girls.png'	42
'kodim16.png'	14	'jon.png'	43
'kodim17.png'	15	'kids.png'	44
'kodim18.png'	16	'kodim02.png'	45
'kodim19.png'	17	'kodim15.png'	46

Table C.1 – continued from previous page

'kodim20.png'	18	'mnm.png'	47
'kodim21.png'	19	'obama.png'	48
'kodim22.png'	20	'pencils.png'	49
'kodim23.png'	21	'penguins.png'	50
'kodim24.png'	22	'pigeons.png'	51
'monarch.png'	23	'ski.png'	52
'ArtRoom.png'	24	'soccer.png'	53
'Lotus.png'	25	'surfer.png'	54
'Perissa_ Santorini.png'	26	'tiger.png'	55
'Sanfrancisco.png'	27	'venice.png'	56
'Umdan.png'	28	'volleyball.png'	57
'bicycle1.png'	29		

Bibliography

[1] Quality of Experience (QoE), [Online] Available: http://en.wikipedia.org/wiki/Quality_of_experience.

[2] ITU-T J.144, "Objective perceptual video quaity measurement techniques for digital cable television in the presence of a full reference", 2004.

[3] ITU-T J.341, "Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference", http://www.itu.int/rec/T-REC-J.341-201101-I/en.

[4] ITU-T J.149, "Method for specifying accuracy and cross-calibration of video quality metrics (VQM)", 2004.

[5] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures", ITU, Geneva, Switzerland, 2002.

[6] ITU-R Recommendation BT.710-4, "Subjective assessment methods for image quality in high-definition television", ITU, Geneva, Switzerland, 1998.

[7] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications", ITU, Geneva, Switzerland, 2008.

[8] ITU-R Recommendation BT.814-1, "Specifications and alignment procedures for setting of brightness and contrast of displays", ITU, 1994.

[9] ITU-R Recommendation BT.1129-2, "Subjective assessment of standard definition digital television (sdtv) systems", ITU, 1998.

[10] ITU-R Recommendation BT.1361, "Worldwide unified colorimetry and related characteristics of future television and imaging systems", ITU, 1998.

[11] ITU-R Recommendation BT.815-1, "Specification of a signal for measurement of the contrast ratio of displays", ITU, 1994.

[12] ITU-R Recommendation BT.1082-1, "Studies toward the unification of picture assessment methodology", ITU, 1990.

[13] VQEG RRNR-TV Group Test Plan, Version 2.0, 2007. [Online] Available: ftp://vqeg.its.bldrdoc.gov/Documents/Projects/rrnr-tv/RRNR-tv_draft_2.0_changes_accepted.doc.

[14] VQEG Test Plan for Evaluation of Video Quality Models for Use with High Defini-tion TV Content, Draft Version 3.0, 2009. [Online] Available: ftp://vqeg.its.bldrdoc.gov/Documents/Projects/hdtv/VQEG_HDTV_testplan_v3.doc.

[15] VQEG Hybrid Perceptual/Bitstream Group Test Plan, 2009. [Online] Available: ftp://vqeg.its.bldrdoc.gov/Documents/Projects/hybrid/VQEG_hybrid_testplan_v1_3_changes_highlighted.doc.

[16] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "Wireless video quality assessment: a study of subjective scores and objective algorithms", IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 4, pp. 587-598, Apr. 2010.

[17] K. Soundararajan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video", IEEE Transaction on Image Processing, vol. 19, no. 6, pp. 1427-1441, Jun. 2010. [Online] Available: http://live.ece.utexas.edu/research/quality/live_video.html.

[18] K. Soundararajan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "A subjective study to evaluate video quality assessment algorithms", in SPIE Proceedings of Human Vision and Electronic Imaging, Jan. 2010.

[19] F. Boulos, W. Chen, B. Parrein, and P. Le Callet, "Region-of-interest intra prediction for H.264/AVC error resilience", in Proceedings of International Conference on Image Processing, 2009. [Online] Available: http://www.irccyn.ec-nantes.fr/spip.php?article551.

[20] L. Goldmann, F. D. Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video", in Proceedings of Electronic Imaging, 2D Image Processing and Applications, 2010. [Online] Available: http://mmspg.epfl.ch/3dvqa.

[21] J. S. Lee, F. D. Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: application to scalable video coding", IEEE Transactions on Multimedia, vol. 13, no. 5, pp. 882-893, Oct. 2011. [Online] Available: http://mmspg.epfl.ch/svd.

[22] J. S. Lee, F. D. Simone, N. Ramzan, Z. Zhao, E. Kurutepe, T. Sikora, J. Ostermann, and T. Ebrahimi, "Subjective evaluation of scalable video coding for content distribution", in Proceedings of ACM Multimedia, pp. 65-72, Oct. 2011.

[23] S. Pechard, R. Pepion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm". [Online] Available: http://www.irccyn.ec-nantes.fr/spip.php?article541.

[24] F. Zhang, S. Li, L. Ma, and King N. Ngan, "IVP subjective video quality database". [Online] Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.shtml.

[25] H. R. Sheikh, A. C. Bovik, L. Cormack, and Z. Wang, "LIVE image quality assessment database". [Online] Available: http://live.ece.utexas.edu/research/quality/subjective.htm.

[26] P. Le Callet, and F. Autrusseau, "Subjective quality assessment IRCCyN/IVC database". [Online] Available: http://www2.irccyn.ec-nantes.fr/ivcdb/.

[27] Y. Horita, K. Shibata, and Y. Kawayoke, "MICT image quality evaluation database". [Online] Available: http://mict.eng.u-toyama.ac.jp/mictdb.html.

[28] E. C. Larson, and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy", Journal of Electronic Imaging, vol. 19, no. 1, Mar. 2010. [Online] Available: http://vision.okstate.edu/?loc=csiq.

[29] H. Liu, N. Klomp, and I. Heynderickx, "TUD image quality database: perceived ringing". [Online] Available: http://mmi.tudelft.nl/iqlab/ringing.html.

[30] U. Engelke, T. M. Kusuma, and H. J. Zepernick, "Wireless imaging quality (WIQ) database". [Online] Available: http://www.bth.se/tek/rcg.nsf/pages/wiq-db.

[31] L. Goldmann, F. D. Simone, and T. Ebrahimi, "Impact of acquisition distortions on the quality of stereoscopic images", in Proceedings of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2010. [Online] Available: http://mmspg.epfl.ch/3diqa.

[32] F. Autrusseau, and M. Babel, "Subjective quality assessment of LAR coded art images". [Online] Available: http://www.irccyn.ec-nantes.fr/~autrusse/Databases/.

[33] L. Ma, C. Deng, W. Lin, and King N. Ngan, "Image retargeting subjective quality database". [Online] Available: http://ivp.ee.cuhk.edu.hk/projects/demo/retargeting/index.html.

[34] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting", in Proceedings of Siggraph Asia 2010. [Online] Available: http://people.csail.mit.edu/mrub/retargetme/.

[35] H. Liu, and I. Heynderickx, "TUD image quality database: eye-tracking release 1". [Online] Available: http://mmi.tudelft.nl/iqlab/eye_tracking_1.html.

[36] U. Engelke, A. J. Maeder, and H. J. Zepernick, "Visual attention for image quality database". [Online] Available: http://www.bth.se/tek/rcg.nsf/pages/vaiq-db.

[37] D. M. Chandler, and S. S. Hemami, "A57 dataset". [Online] Available: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html.

[38] W. Lin, and C. C. J. Kuo, "Perceptual visual quality metrics: a survey", Journal of Visual Communication and Image Representation, vol. 22, no. 4, pp. 297-312, 2011.

[39] S. Winkler, "Perceptual video quality metrics: a review",in Digital Video Image Quality and Perceptual Coding, H. R.Wu and K. R. Rao, ed., Boca Raton, FL: CRC Press, ch. 5, 2005.

[40] W. Lin, and M. Narwaria, "Perceptual image quality assessment: Recent progress and trends", in Proceedings of SPIE, vol. 7744, 2010.

[41] B. Girod, "what's wrong with mean-squared error?", in Digital Images and Human Vision, A. B. Watson ed., Cambridge, MIT Press, pp. 207-220, 1993.

[42] Z. Wang, and A. C. Bovik, "MSE: love it or leave it? a new look at signal fidelity measures", IEEE Signal Processing Magazine, vol. 26, no. 1, pp. 98-117, Jan. 2009.

[43] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2002.

[44] S. Winkler, and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," IEEE Transactions on Broadcasting, vol. 54, no. 3, pp. 660-668, Sept. 2008.

[45] S. Winkler, "Metric evaluation", Digital video quality: vision models and metrics, Wiley, New York, 2005.

[46] W. Lin, "Computational models for just-noticeable difference", Digital video image quality and perceptual coding, H. R. Wu, and K. R. Rao eds., CRC Press, 2005.

[47] J. L. Mannos, and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images", IEEE Transactions on Information Theory, vol. 20, no. 4, pp. 525-536, 1974.

[48] F. Lukas, and Z. Budrikis, "Picture quality prediction based on a visual model", IEEE Transactions on Communications, vol. 30, no. 7, pp. 679-1692, 1982.

[49] T. N. Pappas, and R. J. Safranek, "Perceptual criteria for image quality evaluation", in Handbook of Image and Video Processing, Academic Press, Orlando, FL, 2000.

[50] P. C. Teo, and D. J. Heeger, "Perceptual image distortion", in Proceedings of IEEE International Conference on Image Processing, vol. 2, pp. 982-986, 1994.

[51] P. C. Teo, and D. J. Heeger, "Perceptual image distortion", Human Vision, Visual Processing, and Digital Display V, vol. 2179, pp. 127-141, 1994.

[52] Y. K. Lai, and C. C. J. Kuo, "A Haar wavelet approach to compressed image quality measurement", Journal of Visual Communication and Image Representation, vol. 11, no. 1, pp. 17-40, 2000.

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

[54] S. Winkler, "Issues in vision modeling for perceptual video quality assessment", Signal Processing, vol. 78, no. 2, pp. 231-252, 1999.

[55] M. Masry, S.S. Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications", IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 2, pp. 260-273, 2006.

[56] Z. H. Yu, H. Wu, S. Winkler, and T. Chen, "Vision-model-based impairment metric to evaluate blocking artifacts in digital video", Proceedings of the IEEE, vol. 90, no. 1, pp. 154-169, 2002.

[57] C. J. V. Lambrecht, "Color moving pictures quality metric", in Proceedings of International Conference on Image Processing, vol. 1, pp. 885-888, 1996.

[58] A. B. Watson, J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision", Journal of Electronic Imaging, vol. 10, no. 1, pp. 20-29, 2001.

[59] J. A. J. Ahumada, and H. A. Peterson, "A visual detection model for DCT coefficient quantization", in Proceedings of 9th AIAA Computing in Aerospace Conference, pp. 314-318, 1993.

[60] S. Winkler, "Vision", Digital video quality: vision models and metrics, Wiley, New York, 2005.

[61] G. E. Legge, "Contrast masking in human-vision", Journal of the Optical Society of America, vol. 70, no. 12, pp. 1458-1471, 1980.

[62] A. B. Watson, "Model of visual contrast gain control and pattern masking", Journal of the Optical Society of America A - Optics Image Science and Vision, vol. 14, no. 9, pp. 2379-2391, 1997.

[63] J. A. J. Ahumada, B. L. Beard, and R. Eriksson, "Spatio-temporal discrimination model predicts temporal masking functions", in Proceedings of SPIE, 1998.

[64] ITU-R Recommendation BT.1683, "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference", ITU, Geneva, Switzerland, 2004.

[65] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment II", 2009. [online] Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII/downloads/VQEGII_Final_Report.pdf.

[66] F. Zhang, L. Ma, S. Li, and K. N. Ngan, "Practical image quality metric applied to image coding", IEEE Transactions on Multimedia, vol. 13, no. 4, pp. 615-624, Aug. 2001.

[67] F. Zhang, W. Liu, W. Lin, and K. N. Ngan, "Spread spectrum image watermarking based on perceptual quality metric", IEEE Transactions on Image Processing, vol. 20, no. 11, pp. 3207-3218, Nov. 2011.

[68] Z. Wang, A. C. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[69] Z. Wang, and A. C. Bovik, "A universal image quality index", IEEE Signal Processing Letters, Vol. 9, no. 3, pp. 81-84, Mar. 2002.

[70] H. R. Wu, and M. Yuen, "A generalized block-edge impairment metric (GBIM) for video coding", IEEE Signal Processing Letters, vol. 4, no. 11, pp. 317-320, Nov. 1997.

[71] A. Eskicioglu, A. Gusev, and A. Shnayderman,"An SVD-based gray-scale image quality measure for local and global assessment", IEEE Transactions on Image Processing, vol. 15, no. 2, pp. 422-429, 2006.

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

[73] M. Miyahara, K. Kotani, and V. R. Algazi, "Objective picture quality scale (PQS) for image coding", IEEE Transactions on Communications, vol. 46, no. 9, pp. 1215-1226, 1998.

[74] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using nature scene statistics: JPEG 2000", IEEE Transactions on Image Processing, vol. 14, no. 11, pp. 1918-1927, Nov. 2005.

[75] L. Liang, S. Wang, J. Chen, S. Ma, D. Zhao, and W. Gao, "No-reference perceptual image quality metric using gradient profiles for JPEG 2000", Signal Processing: Image Communication, vol. 25, no. 7, pp. 502-516, Aug. 2010.

[76] T. Brandao, and M. P. Queluz, "No-reference image quality assessment based on DCT domain statistics", Signal Processing, vol. 88, no. 4, pp. 822-833, Apr. 2008.

[77] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images", in Proceedings of IEEE International Conference on Image Processing, Sept. 2002.

[78] R. Ferzli, and L. J. Karam, "A no-Reference objective image sharpness metric based on the notion of just noticeable blur (JNB)", IEEE Transactions on Image Processing, vol. 18, no. 4, Apr. 2009.

[79] Z. Wang, and A. C. Bovik, "Reduced- and no-reference image quality assessment: the natural scene statistic model approach", IEEE Signal Processing Magazine, vol. 28, pp. 29-40, Nov. 2011.

[80] M.H. Pinson, and S. Wolf, "A new standardized method for objectively measuring video quality", IEEE Transactions on Broadcasting, vol. 50, no. 3, pp. 312-322, 2004.

[81] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms", IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 3440-3451, 2006.

[82] VQEG,"Final report from the video quality experts group on the validation of objective models of video quality assessment", 2000. [Online] Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI/.

[83] K. Seshadrinathan, and A. C. Bovik, "Unifying analysis of full reference image quality assessment", in Proceedings of International Conference on Image Processing, 2008.

[84] X. Hou, and L. Zhang, "Saliency detection: a spectral residual approach", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[85] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment", in the Handbook of Video Databases: Design and Applications, CRC Press, pp. 1041-1078, Sep. 2003.

[86] E. A. Essock, J. K. D. Ford, B. C. Hansen, and M. J. Sinai, "Oblique stimuli are seen best (not worst!) in naturalistic broad-band stimuli: a horizontal effect", Vision Research, vol. 43, no. 12, pp. 1329-1335, Jun. 2003.

[87] B. C. Hansen, and E. A. Essock, "A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes", Journal of Vision, vol. 4, no. 12, pp. 1044-1060, 2004.

[88] B. C. Hansen, and E. A. Essock, "Influence of scale and orientation on the visual perception of natural scenes", Visual Cognition, vol. 12, no. 6, pp. 1199-1234, 2005.

[89] B. C. Hansen, and E. A. Essock, "Anisotropic local contrast normalization: the role of stimulus orientation and spatial frequency bandwidths in the oblique and horizontal effect perceptual anisotropies", Vision Research, vol. 46, no. 26, pp. 4398-4415, 2006.

[90] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles", Vision Research, vol. 20, no. 10, pp. 847-856, 1980.

[91] Z. Wang, and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics", in Proceedings of SPIE, Human Vision and Electronic Imaging, Jan. 2004.

[92] Z. Wang, and X. Shang, "Spatial pooling strategies for perceptual image quality assessment", in Proceedings of International Conference on Image Processing, 2006.

[93] S. Li, and K. N. Ngan, "Influence of the smooth region on the structural similarity index", in Proceedings of Pacific-Rim Conference on Multimedia, 2009.

[94] P. Longere, X. Zhang, P. B. Delahunt, and D. H. Brainard, "Perceptual assessment of demosaicing algorithm performance", Proceedings of the IEEE, vol. 90, no. 1, pp. 123-132, Jan. 2002.

[95] W. Lin, and L. Dong, "Adaptive downsampling to improve image compression at low bit rates", IEEE transactions on Image Processing, vol. 15, no. 9, pp. 2513-2521, Sep. 2006.

[96] X. Zhang, X. Wu, and F. Wu, "Image coding on quincunx lattice with adaptive lifting and interpolation", in Proceedings of Data Compression Conference, Mar. 2007.

[97] X. Zhang, and X. Wu, "Can lower resolution be better?" in Proceedings of Data Compression Conference, Aug. 2008.

[98] B. Zeng, and A. N. Venetsanopoulos, "A JPEG-based interpolative image coding scheme", in Proceedings of International Conference Acoustics, Speech, and Signal Processing, 1993.

[99] A. M. Bruckstein, M. Elad, and R. Kimmel, "Down-scaling for better transform compression", IEEE Transactions on Image Processing, vol. 12, no. 9, pp. 1132-1144, Sep. 2003.

[100] Y. Tsaig, M. Elad, G. Golub, and P. Milanfar, "Optimal framework for low bit-rate block coders", in Proceedings of International Conference on Image Processing, 2003.

[101] X. Wu, X. Zhang, and X. Wang, "Low bit-rate image compression via adaptive downsampling and constrained least squares upconversion", IEEE Transactions on Image Processing, vol. 18, no. 3, pp. 552-561, Mar. 2009.

[102] D. Taubman, and M. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice. Springer, 2001.

[103] T. Wiegand, G. J. Suillivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard", IEEE Transactions on Circuits and System for Video Technology, vol. 13, no. 7, pp. 560-576. Jul. 2003.

[104] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards", IEEE Transactions on Circuits and System for Video Technology, vol. 13, no. 7, pp. 688-703. Jul. 2003.

[105] Y. Zhang, J. Zhang, R. Xiong, D. Zhang, and S. Ma, "Low bit-rate image coding via interpolation oriented adaptive down-sampling", in Proceedings of SPIE Visual Communications and Image Processing, vol. 7744, 2010.

[106] Y. Zhang, D. Zhao, J. Zhang, R. Xiong, and W. Gao, "Interpolation-dependent image downsampling", IEEE Transactions on Image Processing, vol. 20, no. 11, pp. 3291-3296, May. 2011.

[107] I. Shin, and H. W. Park, "Adaptive up-sampling method using DCT for spatial scalability of scalable video coding", IEEE Transactions on Circuits and System for Video Technology, vol. 19, no. 2, pp. 206-214, Feb. 2009.

[108] Z. Wu, H. Yu, and C. W. Chen, "A new hybrid DCT-Wiener-based interpolation scheme for video intra frame up-sampling", IEEE Signal Processing Letters, vol. 17, no. 10, pp. 827-830, Oct. 2010.

[109] Kodak Lossless True Color Image Suite. [Online] Available: http://r0k.us/graphics/kodak/

[110] Weber's Law of Just Noticeable Differences. [Online] Available: http://www.usd.edu/psyc301/WebersLaw.htm

[111] A. J. Ahumada, and H. A. Peterson, "Luminance-model-based DCT quantization for color image compression", in Proceedings of SPIE, Human Vision, Visual Processing, and Digital Display III, vol. 1666 pp. 365-374, 1992.

[112] I. Hontsch, and L. J. Karam, "Adaptive image coding with perceptual distortion control", IEEE Transactions on Image Processing. vol. 11, no. 3, pp. 213-222, Mar. 2002.

[113] 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 Transactions on Circuits and Systems for Video Technology, vol. 15, no. 6, pp. 742-750, Jun. 2005.

[114] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Just noticeable distortion model and its applications in video coding", Signal Processing: Image Communication, vol. 20, no. 7, pp. 662-680, Aug. 2005.

[115] W. Lin, L. Dong, and P. Xue, "Visual distortion gauge based on discrimination of noticeable contrast changes", IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no.7, pp. 900-909, Jul. 2005.

[116] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation", IEEE Transactions on Image Processing, vol. 14, no. 11, pp. 1928-1942, Nov. 2005.

[117] S. J. P. Westen, R. L. Lagendijk, and J. Biemond, "A quality measure for compressed image sequences based on an eye movement compensated spatio-temporal model", in Proceedings of International Conference on Image Processing, 1997.

[118] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, "Perceptual watermarks for digital images and video", Proceedings of the IEEE, vol. 87, no. 7, pp. 1108-1126, Jul. 1999.

[119] X. Zhang, W. Lin, and P. Xue, "Just-noticeable difference estimation with pixels in images", Journal of Visual Communication and Image Representation, vol. 19, no. 1, pp. 30-41, Jan. 2008.

[120] C. Chou, and C. Chen, "A perceptual optimized 3-D subband codec for video communication over wireless channels", IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, no. 2, pp. 143-156, Apr. 1996.

[121] Y. Chin, and T. Berger, "A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancements", IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, no.3, pp. 438-450, Apr. 1999.

[122] X. Zhang, W. Lin, and P. Xue, "Improved estimation for just-noticeable visual distortion", Signal Processing, vol. 85, no. 4, pp. 795-808, Apr. 2005.

[123] Z. Wei, and K. N. Ngan, "A temporal just-noticeable distortion profile for video in DCT domain", in Proceedings of International Conference Image Processing, 2008.

[124] Z. Wei, and K. N. Ngan, "Spatial-temporal just noticeable distortion profile for grey scale image/video in DCT domain", IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 3, pp. 337-346, Mar. 2009.

[125] D. H. Kelly, "Motion and vision II. stabilized spatio-temporal threshold surface", Journal of the Optical Society of America, vol. 69, pp. 1340-1349, 1979.

[126] S. Daly, "Engineering observations from sptaiovelocity and spatiotemporal visual models", in Proceedings SPEI, vol. 3299, pp. 180-191, 1998.

[127] Y. Jia, W. Lin, and A. A. Kassim, "Estimating just-noticeable distortion for video", IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 7, pp. 820-829, Jul. 2006.

[128] J. Dong, J. Lou, C. Zhang, and L. Yu, "A new approach to compatible adaptive block-size transforms", in Proceedings of Visual Communication and Image Processing, 2005.

[129] H. Qi, W. Gao, S. Ma, and D. Zhao, "adaptive block-size transform based on extended integer 8×8/4×4 transforms for H.264/AVC", in Proceedings of International Conference on Image Processing, 2006.

[130] J. Dong, K. N. Ngan, C. Fong, and W. K. Cham, "2D Order-16 integer transforms for HD video coding", IEEE Transaction on Circuit System and Video Technology, vol. 19, no. 10, pp. 1463-1474, Oct. 2009.

[131] C. Zhang, L, Yu, J. Lou, W. K. Cham, and J. Dong, "The technique of prescaled integer transform: concept, design and applications". IEEE Transaction on Circuit System and Video Technology, vol. 18, no. 1, pp. 84-97, Jan. 2008.

[132] S. Gordon, "ABT for Film Grain Reproduction in High Definition Sequences", Doc. JVT-H029, Geneva, Switzerland, May. 2003.

[133] Y. Huh, K. Panusopone, and K. R. Rao, "Variable block size coding of images with hybrid quantization", IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, no. 6, pp. 679-685, Dec. 1996.

[134] K. N. Ngan, K. S. Leong, and H. Singh, "Adaptive cosine transform coding of image in perceptual domain", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 11, pp. 1743-1750, Nov. 1989.

[135] N. Nill, " A visual model weighted cosine transform for image compression and quality assessment", IEEE Transactions on Communications, vol. 33, no. 6, pp. 551-557, Jun. 1985.

[136] B. Li, M. R. Peterson, and R. D. Freeman, "Oblique effect: a neural basis in the visual cortex", Journal of Neurophysiology, pp. 204-217, 2003.

[137] N. Jayant, J. Johnsto, and R. Sagranek, "Signal compression based on models of human perception", Proceedings of the IEEE, vol. 81, no. 10, pp. 1385-1422, Oct. 1993.

[138] R. J. Safranek, and J. D. Johnston, "A perceptually tuned subband image coder with image dependent quantization and post-quantization data compression", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1989.

[139] L. Ma, and K. N. Ngan, "Adaptive block-size transform based just-noticeable difference profile for images", in Proceedings of Pacific-Rim Conference on Multimedia, 2009.

[140] G. Robson, "Spatial and temporal contrast sensitivity functions of the visual system", Journal of Optical Society of America, vol. 56, pp. 1141-1142, 1966.

[141] S. Gordon, "ABT for film grain reproduction in high definition sequences", Doc. JVT-H029, Geneva, Switzerland, May 2003.

[142] T. Wedi, Y. Kashiwagi, and T. Takahashi, "H.264/AVC for next generation optical disc: a proposal on FREext profile", Doc. JVT-K025. Munich, Germany, Mar. 2004.

[143] L. Ma, K. N. Ngan, F. Zhang, and S. Li, "Adaptive block-size transform based just-noticeable difference model for images/videos", Signal Processing: Image Communication, vol. 26, no. 3, pp. 162-174, Mar. 2011. [Online] Available: http://www.ee.cuhk.edu.hk/~lma/welcome_files/SPIC_2011_Experiments.html

[144] D. M. Chandler, and S. S. Hemami, "VSNR: a wavelet-based visual signal-to-noise ratio for natural images", IEEE Transactions on Image Processing, vol. 16, no. 9, pp. 2284-2298, Sep. 2007.

[145] K. Seshadrinathan, and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos", IEEE Transactions on Image Processing, vol. 19, no. 2, pp. 335-350, Feb. 2010.

[146] A. C. Bovik, "The essential guide to video processing", Second Edition. Elsevier, 2009.

[147] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[148] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[149] A. K. Moorthy, and A. C. Bovik, "Efficient video quality assessment along temporal trajectories", IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 11, pp. 1653-1658, Nov. 2010.

[150] T. Ell, and S. J. Sangwine, "Hypercomplex Fourier transforms for color images", IEEE Transactions on Image Processing, vol. 16, no. 1, pp. 22-35, Jan. 2007.

[151] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structure similarity for image quality assessment", in Proceedings of IEEE Asilomar Conference on Signals, Systems and Computers, 2003.

[152] S. Wolf, and M. H. Pinson, "Low bandwidth reduced reference video quality monitoring system", in Proceedings of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2005.

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

[154] P. Le Callet, C. V. Gaudin, and D. Barba, "Continuous quality assessment of MPEG2 video with reduced reference", in Proceedings of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2005.

[155] S. Yang, "Reduced reference MPEG-2 picture quality measure based on ratio of DCT coefficients", Electronics Letters, vol. 47, no. 6, pp. 382-383, Mar. 2011.

[156] T. Oelbaum, and K. Diepold, "Building a reduced reference video quality metric with very low overhead using multivariate data analysis", Journal of Systemics, Cybernetics and Informatics, vol. 6, no. 5, pp. 81-86, 2008.

[157] M. Tagliasacchi, G. Valenzise, M. Naccari, and S. Tubaro, "A reduced-reference structural similarity approximation for videos corrupted by channel errors", Multimedia Tools and Applications, vol. 48, no. 3, pp. 471-492, Jul. 2010.

[158] P. Le Callet, C. V. Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment", IEEE Transactions on Neural Network, vol. 17, no. 5, pp. 1316-1327, May. 2006.

[159] M. Carnec, P. Le Callet, D. Barba, "An image quality assessment method based on perception of structural information", in Proceedings of IEEE International Conference on Image Processing, vol. 3, pp. 185-188, Sep. 2003.

[160] M. Carnec, P. Le Callet, and D. Barba, "Visual features for image quality assessment with reduced reference", in Proceedings of IEEE International Conference on Image Processing, 2005.

[161] D. Tao, X. Li, W. Lu, and X. Gao, "Reduced-reference IQA in contourlet domain", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 39, no. 6, pp. 1623-1627, Dec. 2009.

[162] U. Engelke, M. Kusuma, H. J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging", Signal Processing: Image Communication, vol. 24, no. 7, pp. 525-547, Aug. 2009.

[163] M. Makar, Y.-C. Lin, A. F. de Araujo, and B. Girod, "Compression of VQM features for low bit-rate video quality monitoring", in Proceedings of IEEE workshop on Multimedia Signal Processing, 2011.

[164] ITU-T Recommendation J.246, "Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference", Aug. 2008. [Online] Available: http://www.itu.int/rec/T-REC-J.246/en.

[165] I. P. Gunawan, and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration", IEEE Transactions on Circuits and System for Video Technology, vol. 18, no. 1, pp. 71-83, Jan. 2010.

[166] C. T. E. R. Hewage, and M. G. Martini, "Reduced-reference quality assessment for 3D video compression and transmission", IEEE Transactions on Consumer Electronics, vol. 57, no. 3, pp. 1185-1193, Aug. 2011.

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

[168] Z. Wang, and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model", in Proceedings of SPIE, Human Vision and Electronic Imaging, Jan. 2005.

[169] Q. Li, and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation", IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 2, pp. 201-211, Apr. 2009.

[170] R. Soundararajan, and A. C. Bovik, "RRED indices: reduced reference entropic differencing framework for image quality assessment", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2011.

[171] J. A. Redi, P. Gastaldo, I. Heynderickx, and R. Zunino, "Color distribution information for reduced-reference assessment of perceived image quality", IEEE Transactions on Circuits and System for Video Technology, vol. 20, no. 12, pp. 1757-1769, Dec. 2010.

[172] A. Rehman, and Z. Wang, "Reduced-Reference SSIM estimation", in Proceedings of IEEE International Conference on Image Processing, 2010.

[173] A. Rehman, and Z. Wang, "Reduced-Reference image quality assessment by structural similarity estimation", IEEE Transactions on Image Processing, vol. 21, no. 8, pp. 3378-3389, Aug. 2012.

[174] W. Xue and X. Mou, "Reduced reference image quality assessment based on weibull statistics", in Proceedings of International Workshop on Quality of Multimedia Experience, 2010.

[175] M. Zhang, W. Xue, and X. Mou, "Reduced reference image quality assessment based on statistics of edge", in Proceedings of SPIE 7876, 787611, 2011.

[176] T. M. Cover, and J. A. Thomas, "Element of information theory", New York: Wiley, 1991.

[177] D. Austin, "Image compression: seeing what's not there", Feature Column, American Mathematical Society, [Online] Available: http://www.ams.org/samplings/feature-column/fcarc-image-compression.

[178] Z. Xiong, K. Ramchandran, M. T. Orchard, and Y. Q. Zhang, "A comparative study of DCT- and wavelet-Based image coding", IEEE Transactions on Circuits and System for Video Technology, vol. 9, no. 5, pp. 692-695, Aug. 1999.

[179] D. Zhao, W. Gao, and Y. K. Chan, "Morphological representation of DCT Coefficients for Image Compression", IEEE Transactions on Circuits and System for Video Technology, vol. 12, no. 9, pp. 819-823, Sep. 2002.

[180] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms", IEEE Transactions on Information Theory, vol. 38, no. 2, pp. 587-607, Mar. 1992.

[181] E. P. Simoncelli, and W. T. Freeman, "The steerable pyramid: a flexible architecture for multi-scale derivative computation", in Proceedings of IEEE International Conference on Image Processing, 1995.

[182] S. Lyu, and E. P. Simoncelli,"Nonlinear image representation using divisive normalization", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[183] S. Lyu, "Divisive normalization: justification and effectiveness as efficient coding transform", in Proceedings of Annual Conference on Neural Information Processing Systems, 2010.

[184] R. W. Buccigrossi, and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain", IEEE Transactions on Image Processing, vol. 8, no. 12, pp. 1688-1701, Dec. 1999.

[185] J. Liu, and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients", IEEE Transactions on Image Processing, vol. 10, no. 11, pp. 1647-1658, Nov. 2001.

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

[187] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," IEEE Transactions on Multimedia, vol. 13, no. 5, pp. 935-949, Oct. 2011.

[188] A. P. Bradley, "A wavelet visible difference predictor", IEEE Transactions on Image Processing, vol. 8, no. 5, pp. 717-730, May. 1999.

[189] S. Daly, "The visible difference predictor: an algorithm for the assessment of image fidelity", in Digital Images and Human Vision, A. B. Watson, Ed. Cambridge, MA: MIT Press, pp. 179-206, 1993.

[190] Genetic algorithm toolbox. [Online] Available: http://www.shef.ac.uk/acse/research/ecrg/gat.html.

[191] K. Zeng, and Z. Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2010.

[192] K. Zeng, and Z. Wang, "Quality-aware video based on robust embedding of intra- and inter-frame reduced-reference features", in Proceedings of IEEE International Conference on Image Processing, 2010.

[193] J. R. Jain, and A. K. Jain, "Displacement measurement and its application in interframe image coding", IEEE Transactions on Communications, vol. 29, no. 12, pp. 1799-1808, Dec. 1981.

[194] Y. C. Lin, and S. C. Tai, "Fast full-search block-matching algorithm for motion-compensated video compression", IEEE Transactions on Communications, vol. 45, no. 5, pp. 527-531, May 1997.

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

[196] M. N. Do, and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance", IEEE Transactions on Image Processing, vol. 11, no. 2, pp. 146-158, Feb. 2002.

[197] A. Shamir, and O. Sorkine, "Visual media retargeting", ACM SIGGRAPH Asia Courses, 2009.

[198] L. Wolf, M. Guttmann, D. Cohen-Or, "Non-homogeneous content-driven video-retargeting", in Proceedings of International Conference on Computer Vision, 2007.

[199] P. Krahenbuhl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming Video", in Proceedings of SIGGRAPH Asia, 2009.

[200] S. Avidan, and A. Shamir, "Seam carving for content-aware image resizing", in Proceedings of SIGGRAPH, 2007.

[201] M. Rubinstein, A. Shamir, and A. Avidan, "Improved seam carving for video retargeting", in Proceedings of SIGGRAPH, 2008.

[202] A. Shamir, and S. Avidan, "Seam-carving for media retargeting", Communications of the ACM, vol. 52, no. 1, pp. 77-85, Jan. 2009.

[203] M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retargeting", in Proceedings of SIGGRAPH, 2009.

[204] Y. Wang, C. Tai, O. Sorkine, and T. Lee, "Optimized scale-and-stretch for image resizing", in Proceedings of SIGGRAPH Asia, 2008.

[205] Y. Pritch, E. Kav-Venaki, and S. Peleg, "Shift-map image editing", in Proceedings of International Conference on Computer Vision, 2009.

[206] Z. Karni, D. Freedman, and C. Gotsman, "Energy-based image deformation", in Proceedings of Symposium on Geometry Processing, 2009.

[207] W. Dong, N. Zhou, J. C. Paul, and X. Zhang, "Optimized image resizing using seam carving and scaling", in Proceedings of SIGGRAPH, 2009.

[208] M. G. Kendall, and B. Babington Smith, "On the method of paired comparisons", Biometrika, vol. 31, pp.324-345, 1940.

[209] M. G. Kendall, "A new measure of rank correlation", Biometrika, vol. 30, pp.81-93, 1938.

[210] M. H. Pinson, and S. Wolf, "Comparing subjective video quality testing methodologies", in Proceedings of SPIE, vol. 5150, no. 3, pp. 573-582, 2003.

[211] Q. Huynh-Thu, N. N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video", IEEE Transactions on Broadcasting, vol. 57, no. 1, pp. 1-14, Mar. 2011.

[212] A. M. van Dijk, J. B. Martens, and A. B. Watson, "Quality assessment of coded images using numerical category scaling", in Proceedings of SPIE Advanced Image and Video Communications and Storage Technologies, 1995.

[213] VQEG, "Final report from the video quality experts from group on the validation of objective models of multimedia quality assessment Phase 1". [Online] Available: ftp://vqeg.its.bldrdoc.gov/Documents/Projects/multimedia/MM_Final_Report/.

[214] O. Pele, and M. Werman, "Fast and robust earth mover's distances", in Proceedings of International Conference on Computer Vision, 2009.

[215] Y. Rubner, C. Tomasi, and l. J. Guibas, "The earth mover's distance as a metric for image retrieval". International Journal of Computer Vision, vol. 40, no. 2, pp. 99-121, Nov. 2000.

[216] D. Simakov Yaron Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity", in Proceedings IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[217] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: a randomized correspondence algorithm for structural image editing", in Proceedings of SIGGRAPH, 2009.

[218] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "SIFT flow: dense correspondence across different scenes", in Proceedings of European Conference on Computer Vision, 2008.

[219] Y. Liu, X. Luo, Y. Xuan, W. Chen, and X. Fu, "Image retargeting quality assessment", in Proceedings of EUROGRAPHICS, 2011.

[220] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors", IEEE Transaction on Circuits and System for Video Technology, vol. 11, no. 6, pp. 703-715, Jun. 2001.

[221] A. D'Angelo, G. Menegaz, and M. Barni, "Perceptual quality evaluation of geometric distortions in images", in Proceedings of SPIE, Human Vision and Electronic Imaging, vol. 6492, 2007.

[222] A. D'Angelo, Z. Zhao, and M. Barni, "A full-reference quality metric for geometrically distorted images", IEEE Transactions on Image Processing, vol. 19, no. 4, pp. 867-881, Apr. 2010.