Visual attention modeling and its applications

Fang, Yuming

2012

Fang, Y. (2012). Visual attention modeling and its applications. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/51079>

<https://doi.org/10.32657/10356/51079>

NANYANG
TECHNOLOGICAL
UNIVERSITY

Visual Attention Modeling and its Applications

Yuming FANG

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2012

Acknowledgments

I would like to express my profound gratitude to Dr. Weisi Lin, my mentor and research supervisor. Without his patience and guidance, I would certainly not have been able to complete the work in this thesis. His expertise and professional attitude have greatly influenced my academic career. I also thank Dr. Bu Sung Lee and Dr. Chiew Tong Lau, who all contributed much to the work in this thesis. They provided many insightful suggestions on my research work. Sincere thanks are also expressed to Dr. Zhenzhong Chen for his solid methods, scientific ideas and patience. I appreciate Dr. Chia-Wen Lin for the fruitful collaborations during my visiting to National Tsing-Hua University in Taiwan.

I really enjoyed the research life in the CeMNet lab at the School of Computer Engineering in Nanyang Technological University. I wish to thank my teammates for their valuable suggestions and encourage during my Ph.D. study: Manoranjan Paul, Chenwei Deng, Narwaria Manish, Zhouye Gu, Lu Dong, Huan Yang, Nevrez Imamoglu, Wei Liu, Feng Shao, Jinjian Wu, Anmin Liu, Lin Ma, Huawei Tian and Qianqian Xu. I also express my gratitude towards the lab staff for their help and support. Throughout this research work, a number of people have contributed many suggestions and criticisms during the development of the proposed models. I would like to express my gratitude to them: Dr. Wei Xiong, Chia-Ming Tsai, Chih-Chung Hsu, Lining Zhang, and all whom I forget. I am also thankful to all the people who participated in my subjective experiments for sparing their valuable time to help.

My special thanks go to my families, who have always given me love and unconditional support. I would like to express my appreciation to all my friends for their support and help in my academic and non-academic matters.

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

Abstract

The visual environment for observers is usually complex, and it is impossible for the human visual system (HVS) to process all signal components and figure out their relationships immediately. Selective attention in the HVS allocates most processing resources to the salient regions rather than the entire visual view equally. There are two different types of visual attention mechanism: bottom-up and top-down. Visual attention mechanism will cause the salient regions automatically ‘pop out’ in visual scenes. In this thesis, we explore the visual attention modeling and its applications in visual signal processing.

Firstly, we propose a saliency detection model for images based on human visual sensitivity and amplitude spectrum. The amplitude spectrum is adopted to represent color, intensity, and orientation distributions for image patches. The saliency value of each image patch is calculated by not only the differences between amplitude spectrum of this patch and other patches in the whole image, but also the visual impacts of these differences determined by human visual sensitivity. Due to the integration of the characteristics of the HVS and better feature representation, the proposed saliency detection model can achieve better performance than existing ones.

Secondly, we design two novel saliency detection models in the compressed domain for images and video respectively. Existing saliency detection models are all implemented in uncompressed domain. However, most images/video over Internet are typically stored in the compressed domain. Thus, we propose the saliency detection

models in the compressed domain for images and video based on features extracted from image and video bit-stream respectively. The saliency value of each image block is obtained based on feature contrast calculation and feature map fusion. New feature extraction and fusion methods are designed to calculate saliency maps for images/video frames in the compressed domain.

Besides the bottom-up mechanism, we also investigate the influence of top-down mechanism in visual attention modeling. A visual attention model by combining bottom-up and top-down information is designed for man-made object detection. The statistical knowledge of orientation features for man-made objects is used to obtain the top-down information to help for determining locations of salient objects in natural scenes. This model demonstrates the usability of orientation features in simulating top-down mechanism.

Additionally, we explore the applications of the proposed saliency detection models in image retargeting and visual search. Saliency map from the proposed saliency detection model is adopted to measure the visual importance of image pixels for image resizing. A new multi-operator image retargeting algorithm in the compressed domain is designed based on the proposed saliency detection model in the compressed domain. A novel idea of texture homogeneity is designed to determine the number of removed seams and proves promising in image resizing. We also explore the influence of the proposed top-down visual attention model in the application of visual search. Experimental results show the benefit of the top-down mechanism in visual search for man-made objects.

Contents

Acknowledgments	i
Abstract.....	iii
Contents	v
List of Figures.....	viii
List of Tables	x
List of Abbreviations	xi
Chapter 1 Introduction.....	1
1.1 Background and Motivation	1
1.2 Objective and Scope of this Work	5
1.3 Thesis Contributions	5
1.4 Organization of the Thesis	8
Chapter 2 Literature Survey	10
2.1 Bottom-up Saliency Detection Models for Images.....	11
2.2 Bottom-up Saliency Detection Models for Video	15
2.3 Top-down Saliency Detection Models.....	17
2.4 Image Retargeting	19
2.5 Summary	22
Chapter 3 Bottom-up Saliency Detection Based on Human Visual Sensitivity.....	24
3.1 Introduction.....	25
3.2 The Proposed Model Based on Human Visual Sensitivity	29
3.2.1 Formulation	31
3.2.2 Amplitude Spectrum for Each Image Patch.....	33
3.2.3 Differences between Image Patches.....	35
3.2.4 Patch Size and Scale for Final Saliency Value.....	37
3.3 Experimental Results and Discussions	40
3.4 Application: Image Retargeting	44
3.5 Conclusions.....	50
Chapter 4 Saliency Detection in Compressed Domain for Images.....	51
4.1 Introduction.....	52

4.2	Framework of the Proposed Saliency Detection Model	55
4.2.1	Feature Extraction from JPEG Bit-stream	56
4.2.2	Saliency Calculation	61
4.3	Application: Image Retargeting in Compressed Domain	63
4.4	Overall Experimental Evaluation	68
4.4.1	Saliency Detection Evaluation	68
4.4.2	Image Retargeting Evaluation	75
4.5	Conclusion	80
Chapter 5	Saliency Detection in Compressed Domain for Video	82
5.1	Introduction	83
5.2	The Proposed Video Saliency Detection Model	85
5.2.1	Feature Extraction from Video Bit-stream	86
5.2.2	Saliency Detection in the Compressed Domain	91
5.3	Experimental Results and Analysis	98
5.3.1	Evaluation Methodologies	98
5.3.2	Experiment I	100
5.3.3	Experiment II	106
5.4	Conclusion	109
Chapter 6	Combination of Bottom-up and Top-down Mechanisms	111
6.1	Introduction	112
6.2	Combining Bottom-up and Top-down Mechanisms	113
6.2.1	Proposed Multi-scale Bottom-up Saliency	115
6.2.2	Top-down Saliency	116
6.2.3	Integration	118
6.3	Performance Evaluation: Visual Search	119
6.4	Conclusion	122
Chapter 7	Summary and Future Work	124
7.1	Summary	124
7.1.1	Integration of HVS' Characteristics in Visual Attention Modeling	125
7.1.2	Saliency Detection in Compressed Domain	126
7.1.3	Top-down Mechanism Modeling	128
7.1.4	Applications of Saliency Detection Models	128
7.2	Future Work	130

7.2.1	Improved Visual Attention Modeling	130
7.2.2	Multi-Factor Top-down Mechanism Modeling.....	131
7.2.3	Applications of Saliency Detection in Compressed Domain	132
Appendix.....		134
References		137
Publications		152
	Journal Papers	152
	Conference Papers	153

List of Figures

Figure 1.1 The major content and organization of the thesis.	8
Figure 2.1 The saliency detection model by Itti <i>et al.</i> [24].	12
Figure 2.2 The retargeted image samples from different algorithms	20
Figure 3.1 Original images and saliency maps	26
Figure 3.2 Object loss in saliency map of FT based model.	27
Figure 3.3 The proposed saliency detection model.	30
Figure 3.4 The relationship between visual acuity and eccentricity [85].	32
Figure 3.5 Relationship between viewing distance and retina eccentricity.	37
Figure 3.6 Original images and its different saliency maps with different patch sizes	39
Figure 3.7 The comparison between our proposed model and others.	41
Figure 3.8 Saliency maps from different saliency detection models	43
Figure 3.9 Original images, visual significance maps and retargeted images	46
Figure 3.10 The mean score for each retargeted image from 10 participants for four different algorithms for the first image dataset.	46
Figure 3.11 The mean score for each retargeted image from 10 participants for four different algorithms for the second image dataset.	47
Figure 3.12 Overall mean scores and variance values of 23 retargeted images from four different algorithms for the first image dataset.	47
Figure 3.13 Overall mean scores and variance values of 23 retargeted images from four different algorithms for the second image dataset.	48
Figure 3.14 Retargeted image samples from the first image dataset for four algorithms.	49
Figure 3.15 Retargeted image samples from the second image dataset for four algorithms	49
Figure 4.1 Comparison of different image retargeting algorithms	54
Figure 4.2 Syntax for DCT-based modes of operation in JPEG standard [95].	57
Figure 4.3 DCT coefficients and the zig-zag scanning in one 8×8 block.	58
Figure 4.4 Different types of DCT coefficients in one 8×8 block.	60
Figure 4.5 Comparison results of ROC curves from different saliency detection models.	70
Figure 4.6 Saliency map samples from different saliency detection algorithms.	71

Figure 4.7 Comparison between saliency maps of images with different compression ratios.	73
Figure 4.8 Comparison between saliency maps of images with different sizes.....	73
Figure 4.9 Overall result of retargeted images from different image retargeting algorithms.	74
Figure 4.10 Detailed results of retargeted images from different image retargeting algorithms. .	75
Figure 4.11 Comparison I of different image retargeting algorithms.	76
Figure 4.12 Comparison II of different image retargeting algorithms.....	77
Figure 4.13 Comparison III of different image retargeting algorithms.....	78
Figure 4.14 Comparison IV of different image retargeting algorithms.	79
Figure 5.1 The proposed video saliency detection model.....	85
Figure 5.2 Data organization.....	87
Figure 5.3 Progressive coding.....	88
Figure 5.4 Motion feature extraction.	91
Figure 5.5 The saliency distributions at human saccade locations (narrow blue bars) and random locations (wide green bars) from different algorithms	101
Figure 5.6 The ROC curves for the static saliency map, the motion saliency map and the final saliency maps from different fusion methods.....	102
Figure 5.7 Comparison samples from different saliency maps	104
Figure 5.8 The saliency distributions at human saccade locations (narrow blue bars) and random locations (wide green bars) from different models.....	105
Figure 5.9 The ROC curves for different video saliency detection models in comparison.	106
Figure 5.10 Comparison samples of saliency maps from different models.	107
Figure 5.11 Continuous video frames and their saliency maps. First row: original images; second row: saliency maps.	109
Figure 6.1 The proposed visual attention model by combining bottom-up and top-down mechanisms.	114
Figure 6.2 Comparison between the original phase-based model and the improved multi-scale phase-based model.....	115
Figure 6.3 Difference of orientation features for vehicles and trees.....	116
Figure 6.4 Gabor filters with 12 different orientations.	117
Figure 6.5 Comparison results of two sample images between the integrated model and the bottom-up visual attention model	119
Figure 6.6 Comparison results of two other sample images between the integrated model and the bottom-up visual attention model	120
Figure 7.1 The comparison samples from different parameters for the Gaussian model	135
Figure 7.2 The ROC curves of saliency results from different values for parameter σ	136

List of Tables

Table 4.1: Comparison of AUC from different saliency detection models.	69
Table 5.1: Comparison results from the static saliency map, the motion saliency map and the final saliency maps from different fusion methods.	103
Table 5.2: Comparison results from different video saliency detection models.	108
Table 6.1 Comparison of average saliency area for the bottom-up model and the proposed integrated model.....	122
Table 7.1 The AUC of saliency results from different values for parameter σ	134

List of Abbreviations

ASP	Advanced Simple Profile
AUC	Area Under ROC Curve
CRF	Conditional Random Field
DCT	Discrete Cosine Transform
DoG	Difference of Gaussian
FIT	Feature Integration Theory
FPR	False Positive Rate
FT	Fourier Transform
GBVS	Graph-based Visual Saliency
GMM	Gaussian Mixture Model
GOVP	Group of Video Object Planes
HF	High-frequency
HVS	Human Visual System
ICL	Incremental Coding Length
IEC	International Electrotechnical Commission
IFT	Inverse Fourier Transform
ISO	International Organization for Standardization
ITU	International Telecommunication Union
IWT	Inverse Wavelet Transform
JPEG	Joint Photographic Experts Group
JTC1	Joint Technical Committee 1
J2K	JPEG 2000
KL	Kullback-Leibler
LF	Low-frequency
MCU	Minimum Coded Unit
MF	Medium-frequency
MOS	Mean Opinion Score
MV	Motion Vector
NM	Normalization and Maximum
NP	Normalization and Product
NS	Normalized and Sum
PNSP	Parameterized Normalization, Sum and Product
QFT	Quaternion Fourier Transform
ROC	Receiver Operating Characteristic
ROI	Regions of Interest
SER	Site Entropy Rate
SOI	Start of Image
SR	Spectral Residual
TPR	True Positive Rate
VLC	Variable Length Coding
VO	Video Object
VOL	Video Object Layer

VOP	Video Object Plane
VS	Video Session
WT	Wavelet Transform
WTA	Winner-Take-All

Chapter 1

Introduction

1.1 Background and Motivation

With the rapid increase of various multimedia services, efficient perception-aware image or video processing technology becomes more and more important for delivering high-quality images/video to users. The saliency detection technology, which exploits auto-detection of regions of interest (ROI) for natural scenes, is very useful in practice, since it makes the perceptual-friendly image or video processing possible by modeling the relevant functionalities of the human visual system (HVS). The saliency detection technology has already been used widely in various multimedia processing applications such as coding, retrieval, adaptation, classification and so on. Studies within the context of visual signal compression have tried to use a relatively larger amount of bits available to represent the most representative visual content [1, 3, 21]; this can save much bandwidth and/or storage space without loss of perceived image or video quality. The image retrieval systems require efficient image representation algorithms to obtain the best searching results. The study in [22] uses the saliency detection model for semantics-based image retrieval, while the study in [23] uses the local descriptors of salient regions for image retrieval. As to the mobile devices, images or video have to be retargeted for

small display sizes. In [2, 4], the saliency detection models are utilized to capture the most representative visual content for small displays. In [122], the visual attention model is used to capture the ‘gist’ feature of scenes for classification. Studies in [123, 124] adopt the saliency detection model in visual quality assessment. The saliency detection models have also been widely used in object recognition [125 - 129]. Other applications of saliency detection models include image segmentation [131, 132], video summarization [130], image super-resolution [135], visual tracking [137], robot localization [138, 140], and advertising [139]. A detailed introduction of the applications of saliency detection techniques can be found in the study [141].

Visual attention is an important characteristic in the HVS and research on visual attention was first reported in 1890 [6]. It is a cognitive process of selecting the relevant regions while acquiring the most significant information from the visual scene. Generally, information captured by human eyes is much more than that the central nervous system can process. When observers look at a natural scene, it is impossible for them to recognize all the components and their relationship in the scene immediately [7 - 10]. Generally, selective attention will allocate processing resources to a certain small region of the scene rather than the entire scene equally [11 - 13]. This attended region is considered as ROI in the natural scene. There are two different approaches in visual attention mechanism: bottom-up approach and top-down approach [14 - 17, 142]. Bottom-up approach, which is data-driven and task-independent, is a perception processing for automatic salient region selection for images [143]. It is a fast and involuntary perception process [141]. On the contrary, top-down approach is related to the recognition processing influenced by the prior knowledge such as tasks to be performed, the feature distribution of the target, the context information of the visual scene and so on [18 - 20].

During the past ten years, various computational models of visual attention have been proposed for salient region detection for images/video [2, 3, 20, 24 - 34]. Although some of these models have been widely used in many multimedia processing applications, rooms for further performance improvement are still to be explored for enabling new functionalities. In addition, the applications of saliency detection models can be further investigated.

For bottom-up saliency detection models, most of existing models are implemented by calculating the center-surround differences based on the Feature Integration Theory (FIT) [14]. They always calculate saliency map for images based on the contrast of low-level features including intensity, color, orientation, etc. These low-level features used in existing models are extracted simply to represent the image for saliency calculation. Thus, more reasonable and effective feature representation should be explored in visual attention modeling. In addition, most of the existing models neglected some key characteristics of the HVS, such as the human visual sensitivity due to foveation (i.e. the influence of an image patch decreases with the increase of the spatial distance). The visual impact of the characteristics in the HVS should be investigated in modeling the visual attention mechanism.

Furthermore, existing saliency detection models for images and video are all implemented in the uncompressed domain. However, images are typically stored over Internet in the compressed domains such as Joint Photographic Experts Group (JPEG), and JPEG2000 (J2K), while video are always stored in the compressed domain such as H.264, MPEG2, MPEG4 Visual, etc. These compressed images and video are widely used in various Internet-based multimedia applications, since they reduce the storage space and increase the downloading speed for Internet users. In order to extract features from the compressed images or video, existing saliency detection models have to

decompress these compressed images or video into the spatial domain. The full decompression process is not only time-consuming but computation-consuming as well. Therefore, saliency detection in the compressed domain is much desired for various multimedia processing applications.

As introduced above, saliency detection models are widely used in various multimedia processing applications today. One popular application is image retargeting [72 - 74]. In many image retargeting algorithms, a saliency map is used to measure the visual importance of image pixels for image resizing. In these image retargeting algorithms, the used saliency map will influence the performance of image resizing greatly. A superior saliency detection model used for measuring the visual importance of image pixels is much desired in the application of image retargeting. In addition, designing effective saliency-based image retargeting algorithms is also meaningful for delivering high-quality images to Internet users. Another similar application of saliency detection models is visual quality assessment [155]. The saliency map is adopted to measure the importance of pixels to determine the weighting of image pixels for visual quality assessment [18, 152, 153, 154].

Besides the bottom-up mechanism, the focus of attention is also influenced by the top-down mechanism. The top-down mechanism is a perceptual process related to the prior knowledge including the performed tasks, the feature distributions of images and so on. Thus, the top-down saliency detection models can be designed differently from each other due to the used prior knowledge. Recent studies have used target templates, context information or object appearance features to simulate the top-down mechanism for building saliency detection models [19, 20, 35, 36]. Besides these prior knowledge used in existing studies, other types of prior knowledge should be also investigated to design more reasonable top-down saliency detection models, such as orientation features.

1.2 Objective and Scope of this Work

The objective of this thesis is to build new computational models for both bottom-up and top-down visual attention mechanisms, and explore the applications of the proposed saliency detection models. In particular, the following five aspects are addressed in this thesis.

- How to design new feature extraction methods to better represent images for saliency calculation.
- How to explore and use characteristics of the HVS to build effective perception-based saliency detection models to stimulate the visual attention mechanism.
- How to design efficient saliency detection models in the compressed domain.
- How to use prior knowledge to simulate the top-down mechanism for visual attention modeling.
- How to adopt the proposed saliency detection models in various multimedia processing applications.

1.3 Thesis Contributions

In this thesis, new computational models of visual attention are proposed from the following aspects: the new feature extraction method for images, the exploration of characteristics of the HVS in visual attention modeling, the investigation of visual attention modeling in the compressed domain, the exploration of other types of prior knowledge in top-down visual attention modeling, and the applications of the proposed saliency detection models. The key contributions of this work are outlined as follows:

- Propose a novel bottom-up saliency detection model based on human visual sensitivity.

The major advantages of the proposed bottom-up saliency detection model are as follows: (a) we divide an image into small patches for local information extraction and combine information from different patches in a global perspective; (b) we investigate the visual impact of image patch differences based on human visual sensitivity, which is a key characteristic of the HVS; (c) we utilize the amplitude spectrum of Quaternion Fourier Transform (QFT) to represent color, intensity and orientation distributions for image patches; (d) we exploit the characteristics of the HVS to determine the patch size and perform the multi-scale operation.

- Propose a novel bottom-up saliency detection model in the compressed domain for images.

We demonstrate how to extract low-level features of intensity, color and texture from JPEG bit-stream. These extracted low-level features in the compressed domain prove promising in image representation. In addition, the Hausdorff distance is adopted to calculate feature contrast based on the extracted features in the compressed domain. A new saliency calculation method for Discrete Cosine Transform (DCT) blocks is proposed based on the feature contrast weighted by a Gaussian model of Euclidean distances between image blocks.

- Propose a novel bottom-up saliency detection model in the compressed domain for video.

We demonstrate how to extract low-level features of luminance, color, texture and motion from video bit-stream. A new algorithm of video saliency detection is proposed based on the extracted features in the compressed domain. In addition, we design a new fusion method for the combination of the static saliency map and motion saliency map for the final saliency calculation.

- Propose an integrated visual attention model combining bottom-up and top-down mechanisms for man-made object detection.

Different from existing studies which use target templates, context information or object appearance features as top-down information, we utilize the statistical knowledge of orientation features as the top-down information. We demonstrate the usability of orientation features in building top-down visual attention models.

- Explore the applications of the proposed saliency detection models in image retargeting and visual search.

Firstly, we explore the application of the proposed saliency detection models in image retargeting in the following two aspects. (a) We use the saliency map from the proposed bottom-up saliency detection model to measure the visual importance of image pixels in the framework of Seam Carving algorithm. The study demonstrates the effectiveness of the proposed saliency detection model in image retargeting. (b) A novel adaptive image retargeting algorithm is designed in the compressed domain based on the proposed saliency detection model in the compressed domain. The block-based seam carving and image scaling are used for image resizing in the proposed image retargeting algorithm. Different from existing multi-operator image retargeting algorithms, the novel texture homogeneity is defined to determine the number of removed block-based seams and proves promising in image resizing. Furthermore, we investigate the influence of the top-down mechanism in visual search task. We adopt the proposed visual attention model of bottom-up and top-down mechanisms in visual search for man-made objects and obtain better performance than existing bottom-up saliency detection models.

Figure 1.1 The major content and organization of the thesis.

1.4 Organization of the Thesis

Figure 1.1 illustrates the major content and organization of this thesis. The thesis has been divided into seven chapters as described as follows.

Chapter 1 (this chapter) gives a brief introduction about the thesis, including the background and motivation, objective and scope, thesis contributions, and thesis organization.

Chapter 2 introduces the major related existing work and algorithms in visual attention modeling for bottom-up and top-down mechanisms. As we explore the application of the proposed saliency detection model in image retargeting, the state-of-the-art image retargeting algorithms are also reviewed in this chapter. The advantages and shortcomings of some state-of-the-art algorithms are briefly given in this chapter. More specific literature survey to each proposed technique in this thesis will be further introduced whenever appropriate in Chapters 3 - 6.

Chapter 3 investigates the benefits of using human visual sensitivity and amplitude

spectrum in building the bottom-up saliency detection model. In addition, the application of the proposed saliency detection model in image retargeting is explored based on the framework of Seam Carving.

Chapter 4 describes a novel bottom-up saliency detection model in the compressed domain for JPEG images. Based on the proposed model, a new adaptive image retargeting algorithm is designed in the compressed domain.

Chapter 5 presents an effective bottom-up saliency detection model in the compressed domain for video. This model is an extension work of the one in Chapter 4. A new fusion method for the static saliency map and motion saliency map for the final saliency map is devised in this chapter.

Chapter 6 focuses on developing a visual attention model by combining bottom-up and top-down mechanisms for man-made object detection. The proposed model uses the statistical knowledge of orientation information as the prior knowledge to simulate the top-down mechanism.

Chapter 7 concludes the thesis with a summary of the main research work performed and the possible directions of further studies.

The next chapter provides literature review for the research study in this thesis.

Chapter 2

Literature Survey

When human eyes gaze at a natural scene, the HVS will rapidly focus on the salient region which is also termed as ROI. During this process, the HVS filters out redundant visual information to select the most important visual information to process and thus focus on the salient regions in natural scenes. Existing studies have explored the visual attention mechanism from various aspects such as psychology, biology, computer vision, etc. [10 - 18, 37, 38]. In the 1980s, Treisman *et al.* developed the well-known FIT [14]. According to that theory, the early selective attention mechanism leads some image regions to be salient for their different features (including color, intensity, orientation, motion and so on) from their surrounding regions [14, 39]. Meanwhile, Koch *et al.* proposed a neurophysiological model of visual attention in [40]. It indicates that the selective visual attention includes the following three stages: elementary parallel feature representation across the visual field; the Winner-Take-All (WTA) mechanism singling out the most salient location; the routing selection for next most salient locations [40]. Recently, researchers in the area of computer vision have started to build computational models of visual attention for the emerging interest in the HVS for various multimedia processing applications [3, 4, 24 - 36, 141, 144].

In this chapter, we give a brief overview of the major relevant existing work in saliency

detection models for both bottom-up and top-down mechanisms. In addition, relevant work of image retargeting is also given in this chapter. We also discuss the limitations of some existing methods in order to provide the motivation for the remaining thesis. For a better organization, we divide this chapter into four separate sections: one for bottom-up saliency detection models for images, one for bottom-up saliency detection models for video, one for top-down saliency detection models, and the last one for image retargeting.

2.1 Bottom-up Saliency Detection Models for Images

One of the earliest bottom-up saliency detection models was proposed by Itti *et al.* [24]. The framework of that model is depicted in Figure 2.1. It is implemented based on the behavior and neuronal architecture of the early primate visual system. The model obtains feature maps through calculating multi-scale center-surround differences for color, orientation and intensity [24]. The final saliency map is achieved by a linear combination for these feature maps. Later, Harel *et al.* designed a Graph-based Visual Saliency (GBVS) model [25] by using a better dissimilarity measure for saliency based on Itti's model. Different from the study in [24], the GBVS model uses the graph theory to form saliency maps from low-level features. In [41], Walther *et al.* extended Itti's model [24] for object detection. That study demonstrates the visual attention model for attending proto-objects can be used for serializing visual processing by other models of object recognition. Le Meur *et al.* devised a visual attention model based on the structure of the HVS [42]. The contrast sensitivity functions, perceptual decomposition, visual masking and center-surround interactions are utilized for detecting salient regions [42].

Figure 2.1 The saliency detection model by Itti *et al.* [24].

Besides the above mentioned biologically plausible models, many studies have started to propose saliency detection models based on information theory [28, 45 - 47]. Bruce *et al.* described a visual attention model based on the principle of information maximizing [28]. That saliency detection model yields the saliency map for images based on Shannon's self-information measure. It demonstrates that the model has close ties with circuitry existent in the primate visual cortex [28]. In [45], Hou *et al.* measured the entropy gain of visual features based on the defined Incremental Coding Length (ICL) for saliency detection. The saliency map is calculated by maximizing the entropy gain of visual features. Seo *et al.* calculated likeness of a pixel to its surrounding by local

regression kernels for salient region detection [46]. The final saliency map is computed based on the defined ‘self-resemblance’ measure, which is obtained from non-parametric kernel density estimation for local regression kernels [46]. Wang *et al.* proposed a saliency detection model based on the information maximization principle [47]. In that study, they defined the Site Entropy Rate (SER) to measure the average information transmitted from one node to all the others during the random walk on the graph [47]. The features used in that model are extracted from learned sparse basis.

Some studies have tried to obtain saliency map for images from the transform domain. In [26], Hou *et al.* devised a saliency detection model based on a concept defined as Spectral Residual (SR). The authors claimed that the SR model is obtained by log spectra representation of images [26]. It can be implemented simply by Fourier Transform (FT) and Inverse Fourier Transform (IFT). However, Guo *et al.* later found that Hou’s model was actually caused by phase spectrum and they designed a new phase-based saliency detection model [3, 27]. The model in [3] achieves the final saliency map by IFT on a constant amplitude spectrum and the original phase spectrum from images. In that model, the QFT instead of FT is used for saliency calculation. Compared with FT which processes each feature channel of color images separately, QFT allows color images to be transformed as a whole [84]. Murray *et al.* proposed a non-parametric low-level saliency detection model to extract the salient regions for images [43]. In that model, Inverse Wavelet Transform (IWT) on scale-weighted center-surround features is adopted for scale integration and a Gaussian Mixture Model (GMM) on eye-fixation data is trained to determine the sizes of center-surround inhibition windows [43]. Achanta *et al.* tried to obtain more frequency information to get a better saliency measure [48]. Difference of Gaussian (DoG) is used to extract frequency information in that model [48]. Gopalakrishnan *et al.* designed a saliency detection model according to the color and

orientation distributions in images [32]. In that model, the orientation saliency map is calculated based on the FT amplitude spectrum. The final saliency map is selected as the color saliency map or the orientation saliency map based on the defined criteria [32].

Recently, some studies have tried to use machine learning techniques to learn the saliency map for images. In [30], Liu *et al.* used a machine learning technique to achieve the saliency map for images. That study obtains the salient objects based on features of multi-scale contrast, center-surround histogram and color spatial distribution. Then a Conditional Random Field (CRF) is determined for these features to detect salient objects in images [30]. Judd *et al.* also adopted a machine learning technique to calculate the saliency map for images based on a set of low, middle, and high-level features [50]. The saliency detection model is learned from a large-scale eye tracking database of 15 viewers on 1003 images [50].

In [31, 44], Gao *et al.* calculated the center-surround discriminant for saliency detection. The saliency value for image pixels is obtained by the power of a Gabor-like feature set to discriminate the center-surround visual appearance [31]. That model was used for visual recognition and obtained promising performance [44]. Goferman *et al.* designed a context-aware saliency detection model by including more context information in the final saliency map [49]. The center-surround differences of image patches are used for saliency detection and Euclidean distances between image patches are used to weight these center-surround differences for saliency detection [49]. A saliency detection model by Valenti *et al.* is advanced through calculating the center-surround differences of edges, color and shape for images [33]. A new fusion operator is proposed for combining the feature contrast to infer global information, which is used to segment salient objects out of the scene background [33].

To evaluate the performance of saliency detection models, we can compare the saliency

map obtained from saliency detection algorithms with the ground truth. There are two types of ground truth for saliency detection: one is the ground truth map labeled by subjects; the other is the ground truth map obtained from eye tracking data. Currently, there are two widely used datasets for ground truth labeled by subjects: Microsoft Database including 5000 images and their ground truth maps [30], and EPFL Database including 1000 images and their ground truth maps [48]. As to the ground truth from eye tracking data, there are also two public databases: the one in [28] includes 120 images and their eye tracking data; the other in [50] includes 1000 images and their eye tracking data. Generally, the following two evaluation methods are used for the performance evaluation of saliency detection models: F-measure and Receiver Operating Characteristic (ROC). F-measure is used to evaluate the performance of saliency detection models by calculating the precision, recall and F-measure values from the saliency map and ground truth. ROC uses the ROC curve and ROC area to evaluate the performance of saliency detection models. The detailed introduction of F-measure and ROC will be presented in experimental sections in Chapter 3 and Chapter 4, respectively.

2.2 Bottom-up Saliency Detection Models for Video

Compared with image saliency detection models, video saliency detection models have to calculate the motion saliency map, since motion is an important factor to attract human attention [56 - 58]. Currently, many studies have tried to propose video saliency detection models to extract the salient regions in video frames [3, 45, 59, 60 - 64].

Guo *et al.* proposed a phase-based saliency detection model for video in [3]. That model obtains the saliency map through IFT on constant amplitude and the original phase

spectrum of input video frames. In that study, the phase spectrum of QFT is adopted based on the following features: intensity, color and motion. Itti *et al.* developed a model to detect the low-level surprising events in video [59]. In that study, the surprising events are defined as the important information attracting human beings' attention in video. A Bayesian framework of low-level features is adopted to extract the surprise events in video [59]. Zhai *et al.* built a video saliency detection model by combining the spatial and temporal saliency maps [60]. The color histogram of images is used for calculating the spatial saliency map, while the planar motion between images (estimated by applying RANSAC on point correspondences in the scene) is adopted for computing the temporal saliency map.

The study [45] introduced in the previous section can also be used to extract salient regions in video. As introduced previously, that visual attention model is built based on the rarity of features. The ICL is defined to measure the entropy gain of visual features for saliency calculation. In [61], Le Meur *et al.* extended the image saliency detection model in [42] to the spatio-temporal domain for video saliency detection. The saliency map for video frames is calculated from the combination of features maps based on the achromatic, chromatic and temporal information [61]. A new fusion method is proposed to combine feature maps to obtain the final saliency map for video frames [61]. Mahadevan *et al.* devised a spatio-temporal saliency detection model based on the biological mechanisms of motion-based perceptual grouping and the discriminant formulation of center-surround saliency [62]. That model is an extension work from the discriminant saliency detection in [31]. Marat *et al.* presented a spatio-temporal saliency detection model inspired by the HVS [64]. The input video can be split into spatial and motion information by two pathways (magnocellular and parvocellular) of the HVS simulated in that model. The static and dynamic saliency maps can be calculated based

on the spatial and motion information respectively. A fusion method is employed to combine the static saliency map and motion saliency map into the final saliency map for video frames [64].

Ma *et al.* built a video saliency detection model by integrating the top-down mechanism into the classical bottom-up saliency detection model for video summarization [63]. Semantic cues such as face, speech and so on are used to simulate the top-down mechanism in that model. A fusion method is adopted to combine the bottom-up saliency and top-down saliency maps into the final saliency map for video frames. In addition, the application of the saliency detection model in video summarization is demonstrated in that study [63].

Generally, the database for performance evaluation of video saliency detection models includes the original video and the ground truth from eye tracking data. There are several eye tracking databases in studies [59], [61] and [65]. Most studies use ROC and Kullback-Leibler (KL) distance [114, 115] to evaluate the performance of video saliency detection models. KL distance is used to calculate the similarity between the saliency map and ground truth for the performance evaluation of saliency detection models. The detailed introduction of KL distance can be found in the experiment section in Chapter 5.

2.3 Top-down Saliency Detection Models

Most of the models described in the two previous sections are data-driven, bottom-up visual attention models for salient region detection. Existing studies indicate that the human visual processing is also influenced by the top-down mechanism of prior knowledge such as tasks alongside with the bottom-up mechanism [15, 66]. The influencing factors of the top-down mechanism include the task to be performed, the

prior knowledge about feature distributions of targets, the context information of visual scenes, and so on [19, 20, 36, 67].

During the past ten years, many approaches have been presented to simulate the top-down mechanism [18, 19, 20, 35, 36, 67]. In [18], Lu *et al.* took advantage of human skin color and face as the top-down information for saliency detection in video. The top-down saliency map is combined with the bottom-up saliency map to obtain the final saliency map for visual quality assessment for video [18]. Navalpakkam *et al.* [35] proposed an integrated model for optimal object detection through choosing appropriate weights for various feature channels. Appropriate weights are calculated from the prior statistical knowledge of feature distributions for target and background.

Torralba *et al.* [19, 68] utilized the top-down information from visual context to implement their model. That scene-based top-down mechanism can be understood as follows: if humans search for an object (e.g. a car) in an image, they will focus on the related context part (e.g. the street part) but not an unrelated context part (e.g. a building) in images [19]. Kanan *et al.* [20] used the statistical feature information of object appearance in the top-down visual attention model. A Bayesian framework is adopted to combine the top-down saliency and bottom-up saliency [20] for the final saliency map. The bottom-up saliency is implemented by the self-information of visual features, while the top-down saliency is calculated based on the mutual information between local image features and search target's features.

Li *et al.* measured video saliency by using a machine learning technique [65]. That model presents a Bayesian learning framework by integrating the stimulus-driven and task-related components for visual saliency. The bottom-up saliency is calculated based on the multi-scale wavelet transform (WT), while the top-down saliency is computed from the trained components by a multi-task learning algorithm. Peters *et al.* designed a

visual attention model by combining bottom-up and top-down mechanisms in [69]. The bottom-up saliency is computed from the low-level multi-scale feature maps, while the top-down saliency is calculated from the low-level signature of images. Rao *et al.* implemented a top-down visual attention model by comparing the similarity between input image patches and templates of the target [36].

Similar with the performance evaluation for bottom-up saliency detection models, the ROC and KL distance are used to evaluate the performance of top-down saliency detection models. Usually, most studies of top-down visual attention modeling built their own databases including original images and ground truth for performance evaluation [19, 20, 65].

2.4 Image Retargeting

One popular application of saliency detection models is image retargeting. Due to the heterogeneity of end devices, images have to be resized to fit various display screens with different sizes or aspect ratios. One traditional image resizing method is to scale images by down-sampling. The problem with image scaling is that it will result in worse viewing experience and loss of some detailed information as the salient objects turn to be smaller. Image cropping is an alternative solution which preserves ROI in images by discarding other non-interest regions [145 - 147]. The defect of this technique is that the context information in images will be lost [70, 76, 80, 81]. To overcome the limitations of image scaling and cropping, many advanced image retargeting algorithms [70-79] have been proposed. In these algorithms, content awareness is taken into consideration and the visual significance map is designed for measuring the visual importance of each pixel for image resizing operation. The visual significance maps used in these algorithms

Figure 2.2 The retargeted image samples from different algorithms (image samples are from [80]). The first column: original images; the second to the final column: retargeted images from cropping, image scaling, algorithm [78] and algorithm [73], respectively.

are generally composed of the gradient map, the saliency map and some high-level feature maps such as facial map, motion map and so on [70 - 79]. Some retargeted image samples from different algorithms are provided in Figure 2.2.

Recently, Avidan *et al.* proposed a popular image retargeting algorithm named seam carving [70]. A seam is defined as 8-connected path of low-energy pixels (from top to bottom or left to right) in images. These pixels include only one pixel in each row or column. The seam carving aims to reduce the width (or height) by removing those unimportant seams. A gradient map is used to determine the visual importance of each pixel in images. Later, Rubinstein *et al.* extended the seam carving algorithm to video retargeting by introducing the forward energy method [71]. The removed 1D seam from 2D images is extended into the 2D seam manifolds for 3D spatio-temporal volumes in video [71]. Based on the seam carving technique, Achanta *et al.* proposed a content-aware image retargeting algorithm by using the saliency map instead of the gradient map [72]. In [77], Grundmann *et al.* extended the seam carving technique through the

discontinuous seam carving in both space and time for video retargeting. Instead of using geometrically smooth and continuous seams, that study adopts temporally discontinuous seams based on a designed appearance-based temporal coherence calculation method.

Other advanced image retargeting algorithms have also been proposed. Wolf *et al.* introduced a video retargeting algorithm through introducing a linear system to determine the new pixel position [73]. In that study [73], the visual importance of each image pixel is measured by the visual importance map composed of local saliency map, facial map and motion map. Ren *et al.* proposed an image retargeting algorithm based on global energy optimization, in which the saliency map and facial map are combined to determine the visual importance of each image pixel [74]. In that algorithm [74], a constrained linear programming method is adopted to maximize the retained energy during resizing process. Jin *et al.* presented a content-aware image resizing algorithm through warping a triangular mesh over images by regarding salient line features and curved features as important regions [75]. A standard quadratic programming method is utilized to resize images [75]. Guo *et al.* advanced an image retargeting algorithm through utilizing saliency-based mesh parametrization [76].

In [148], Cho *et al.* described a patch transform method and demonstrated its application in image editing. The method divides the input image into non-overlapping patches and the new image is reconstructed based on the defined constraints [148]. A Markov network is devised to reconstruct the image with specified constraints [148]. Barnes *et al.* proposed the PatchMatch algorithm for image editing by finding approximate nearest-neighbor matches between patches [149]. That algorithm can obtain substantial performance improvement over the existing ones and thus can be used in interactive editing tools [149]. Pritch *et al.* designed an image editing algorithm based on an optimal graph labeling [105]. In that study, the defined data term and smoothness term

are used to calculate the optimal shift-map [105].

Recently, Rubinstein *et al.* conducted a user study and found that applying multi-operators (such as seam carving, cropping and so on) can obtain better results than those from only a single operator in image retargeting [78]. In that study, the authors proposed a multi-operator media retargeting algorithm which combines seam carving, scaling and cropping operators to resize images. The size amount for each operation is determined by the optimal result for maximizing the similarity between the input image and the retargeted image. In [79], Dong *et al.* introduced an image retargeting algorithm by combining seam carving and image scaling. The authors utilized a bidirectional similarity function of image Euclidean distance, a dominant color descriptor similarity and seam energy variation to determine the best number of seam carving operation [79].

The performance evaluation method for image retargeting algorithms is to invite subjects to evaluate the retargeted results from different retargeting algorithms. The performance of the image retargeting algorithms can be obtained from the subjective evaluation. Currently, there is still no effective objective metric for the performance evaluation of image retargeting algorithms. There is one popular database for retargeting performance evaluation in [80]. In some existing retargeting algorithms, the authors also build their own databases to evaluate the performance of their algorithms.

2.5 Summary

Although various existing saliency detection models have been proposed in the past decades, some key characteristics of HVS are still neglected in building saliency detection models. The characteristics or theories of HVS from biology, neuroscience, psychology and other areas should be investigated further for visual attention modeling.

Another aspect of visual attention modeling is to design the saliency detection models in the compressed domain. Today, most images/video over Internet are stored in the compressed domain for saving the storage space and increasing the delivering speed over Internet. Thus, saliency detection in the compressed domain is much desired for various multimedia processing applications (such as image retargeting, object detection and so on) in the compressed domain. Finally, the practical applications of saliency detection models should also be explored. As introduced above, the saliency map can be used to measure the visual importance of image pixels for various multimedia processing applications, such as image retargeting, visual quality assessment, etc. In this thesis, we explore the applications of saliency detection models in image retargeting and visual object search.

From the next chapter, we will present our approaches by starting with a novel bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum.

Chapter 3

Bottom-up Saliency Detection Based on Human Visual Sensitivity

With the wide applications of saliency information in visual signal processing, various saliency detection methods have been proposed. However, some key characteristics of the HVS are still neglected in building these saliency detection models. In this chapter, we propose a new saliency detection model based on human visual sensitivity and amplitude spectrum of QFT. Image patches are first extracted from the input image. Then the amplitude spectrum of QFT is adopted to represent the color, intensity and orientation distributions for these image patches. The saliency value of each image patch is calculated by not only the differences between QFT amplitude spectrum of this patch and other patches in the whole image, but also the visual impacts for these differences determined by human visual sensitivity. Experimental results show that the proposed saliency detection model outperforms the state-of-the-art saliency detection models. In addition, we apply our proposed saliency detection model in the application of image retargeting based on the framework of Seam Carving and achieve better performance over the traditional image retargeting algorithms.

3.1 Introduction

As introduced in the chapter of literature review, various bottom-up saliency detection models have been proposed during the past decades. Among these saliency detection models, the FT based models are more efficient compared with others because of the availability of fast FT algorithms [3, 26, 27]. As being one of the fundamentals of signal processing, FT has been widely used in image processing since the 1960s with many applications such as convolutions, filtering, compression and reconstruction. For visual attention modeling, FT is used to get the saliency map for images [3, 26, 27, 32]. It is commonly accepted that the phase spectrum carries location information, while the amplitude spectrum includes the appearance and orientation information for visual scenes [51 - 53]. Based on this understanding, FT has been used in various studies of human visual perception and understanding [3, 26, 32, 54, 55]. In the studies of [3, 26, 27], the saliency map is obtained based on phase spectrum, whereas in [32], the amplitude spectrum of image patches is applied to obtain the orientation saliency map. We analyze these two types of saliency detection models in detail below.

In phase-based saliency detection models [26, 27], the amplitude spectrum and phase spectrum are first obtained by FT. The saliency map is then calculated through IFT on a user-defined constant amplitude spectrum (always set as 1 in the models [26, 27]) and the original phase spectrum. This will amplify the intensity of regions with less periodicity or less homogeneity and suppress the intensity of regions with more periodicity or more homogeneity in original images [26, 27]. In these approaches, the FT is operated on the whole images, and thus they mainly consider the feature contrast for images from the global perspective. Therefore, these approaches suffer some defects in saliency detection.

Figure 3.1 Original images and saliency maps; (a) Original images; (b) The saliency maps obtained from the phase-based model without resizing the original images before FT; (c) The saliency maps obtained from phase-based model with resizing the original images into smaller ones (64 pixels for the input image width) [26]; (d) The saliency maps obtained from the saliency detection model in [32]; (e) The saliency maps obtained from the saliency detection model in [24]; (f) The saliency maps obtained from our proposed saliency detection model .

One problem with these phase-based saliency detection models is that they cannot detect smooth-texture salient objects in the complex-texture background (as shown in the first row of Figure 3.1 (c)), since the complex-texture regions are less homogeneous in those images. On the other hand, the phase-based saliency detection models have to resize input images into appropriately smaller sizes to allow the *major part* of a salient object to be less homogeneous, or they will just get the contour of the salient object, as shown in the second row of Figure 3.1 (b). Even when the original image is resized into a smaller one, the saliency map from phase-based models will still ignore some information of salient objects, as shown in the second row of Figure 3.1 (c). The disadvantage can be also seen from Figure 3.2. From this figure, we can see that resizing the original image into a small one will cause some small salient objects to be lost, as shown in Figure 3.2 (b). On the contrary, the proposed model can obtain the salient smooth-texture objects with the complex-texture background in images and will not ignore much information of salient objects, as shown in the Figure 3.1 (f) and Figure 3.2.

Figure 3.2 Object loss in saliency map of FT based model: (a) the original image; (b) the saliency map from the model [26]; (c) the saliency map from the proposed saliency detection model.

The study in [32] gets the final saliency map based on color and orientation distributions of images. The color and orientation saliency sub-maps are achieved separately by using two different algorithms. The final saliency map will be selected as either the color saliency sub-map or the orientation saliency sub-map by identifying which sub-map leads to the identification of salient regions [32]. The model uses amplitude spectrum of FT to calculate the orientation distribution for images by computing the global orientation and orientation entropy contrast. One problem is that the orientation distribution in [32] is calculated by histograms of 18 special orientations of image patches, which causes the loss of other orientation information. In addition, the final saliency map chosen as either color saliency sub-map or orientation saliency sub-map means that the final saliency map is determined by only color distribution or orientation distribution. Therefore, the saliency map from the model in [32] will lose much information of salient objects, as shown in Figure 3.1 (d). On the contrary, our proposed model uses the color, intensity and orientation features together to get the final saliency map and it uses all orientation information in saliency calculation. Thus, the saliency map from our proposed model can better preserve the information of salient objects (Figure 3.1 (f)).

Some classic visual attention models such as the model in [24] also use the low-level

features including color, intensity and orientation to calculate saliency map for images. As these models calculate the multi-scale center-surround differences of low-level features from images to get the final saliency map, they mainly focus on the local contrast of low-level features for saliency detection [24]. One problem with the model in [24] is that it might regard the non-salient regions (such as regions in the background) as salient, as shown in Figure 3.1 (e), due to the lack of consideration for global contrast in the image. In the saliency map from the first row of Figure 3.1 (e), some non-salient regions from the background are considered as salient regions. On the contrary, our proposed model considers both local and global contrast for images, and thus it can obtain much better saliency maps (Figure 3.1 (f)).

As can be seen from the analysis above, a key factor of successful saliency detection is the proper treatment of local and global information. Another important and related issue is how to combine (or pool) different features at a location and a feature from different locations. In [24] and [32], linear combinations are used; in [32], the Euclidian distances are used for the weighting of patch differences. However, there is lack of perceptual ground for all these approaches.

In this chapter, we propose a novel saliency detection model based on FIT and human visual sensitivity variations. According to FIT, the salient regions in an image can be distinguished according to the differences of low-level features between this region and its neighbors. In our proposed model, we first divide the input image into small image patches and measure the saliency value for each image patch through calculating the differences of color, intensity and orientation distributions between this image patch and all other patches (all the neighbor patches) in the image. Unlike existing methods which only consider local contrast or global contrast [24, 29], we exploit both local and global contrast by considering the differences between this patch and all the other image patches

in the image. In addition, the contributions of these differences to the saliency value of image patches are different with the consideration of foveation behavior. We use the foveation-tuned human visual sensitivity to determine the weightage for these patch differences.

In sum, our proposed model first divides images into small image patches. Then it uses QFT instead of FT to obtain the amplitude spectrum of each image patch. Compared with FT which processes each feature channel of color images separately, QFT allows color images to be transformed as a whole [84]. The saliency value of each patch is obtained by two factors: the differences of QFT amplitude spectrum between this image patch and other image patches in the whole image, and the weights for these patch differences determined by human visual sensitivity. The novel saliency detection utilizes the characteristics of the HVS and is proven promising, as shown in the following sections.

Furthermore, we also explore the application of the proposed saliency detection model in image retargeting. We use the saliency map from the proposed saliency detection model to measure the visual importance of image pixels in the framework of Seam Carving [71]. Experimental results demonstrate the effectiveness of the proposed model in the application of image retargeting.

3.2 The Proposed Model Based on Human Visual Sensitivity

In this section, we describe the proposed model in detail. As mentioned above, the model first divides each original input image into small image patches for gathering local information. In our work, the patch size is chosen as 8×8 and partially overlapping. Here we select the size of image patches based on the fovea size. Details of defining image

Figure 3.3 The proposed saliency detection model.

patches will be given in Section 3.2.4. The saliency value of each image patch is obtained through calculating the QFT amplitude spectrum differences between a patch and its neighbor patches, and the weights for these differences determined by human visual sensitivity. The proposed model is illustrated as Figure 3.3. We will describe the details

step by step in the following subsections.

3.2.1 Formulation

In the proposed model, the saliency value of each image patch is determined by two factors: one is the patch differences between this image patch and all other image patches in the input image; the other is the weighting for these patch differences determined by human visual sensitivity. If these differences between an image patch and all other image patches are big, then the saliency value for this image patch is large. In addition, we take the influence of the foveation behavior of human visual sensitivity into consideration in the proposed model. Here, we use $\mathcal{D}_{(i,j)}$ to represent the difference between image patch i and image patch j , the saliency value for image patch i can be expressed as follows:

$$S_i = \sum_{j \neq i} \alpha_{ij} \mathcal{D}_{(i,j)} \quad (3.1)$$

where α_{ij} is the weight for the patch difference between image patches i and j , which is determined by human visual sensitivity.

It is generally believed that the HVS is highly space-variant because the retina in human eyes has different density of cone photoreceptor cells [82]. On the retina, the fovea owns the highest density of cone photoreceptor cells. Thus, the focused region has to be projected on the fovea to be perceived at the highest resolution. The density of the cone photoreceptor cells becomes lower with larger retinal eccentricity. Therefore, the visual sensitivity decreases with the increased eccentricity from the fixation point, as shown in Figure 3.4 [82, 83, 85].

As to the saliency value of image patch i in (3.1), all patch differences between the image patch i and the other image patches are considered and summed together. We use the human visual sensitivity to determine weights for the patch differences. In this study, the eccentricity from the center of the fixation (the center of the image patch i) is not

Figure 3.4 The relationship between visual acuity and eccentricity [85].

directly used as a weighting factor for calculating the saliency value of the image patch i but a weighting factor for calculating the importance of patch-difference pairs. Here, the weights for patch differences are determined by human visual sensitivity, and this means that the weights of patch differences from its nearer neighbor patches (with smaller eccentricities) are larger compared with these from farther neighbor patches. With larger eccentricity of image patches from the image patch i (which means farther image patches from the image patch i), the visual sensitivity decreases and thus the weighting for patch differences between these image patches and image patch i becomes smaller. Therefore, the contributions of patch differences to the saliency value of image patch i will decrease with larger-eccentricity image patches from image patch i . On the contrary, the contributions of patch differences to the saliency value of the image patch i will increase with smaller-eccentricity image patches from image patch i . This is reasonable, as human eyes are more sensible to the patch differences from nearer image patches compared with those from farther image patches. Our proposed saliency detection model takes both local and global center-surround differences into account, for it uses the patch differences from all other image patches in the image to calculate the saliency value of image patch i . We

will describe how to get α_{ij} and $\mathcal{D}_{(i,j)}$ in detail in the following subsections.

3.2.2 Amplitude Spectrum for Each Image Patch

In the proposed model, we use the color and intensity channels for QFT to get the amplitude spectrum for each image patch, which is used to compute the differences between image patches. As we know, the amplitude spectrum indicates the presence of respective spatial frequencies and their strengths can represent the orientation distributions in images [52]. Thus, the amplitude spectrum of QFT can represent the color, intensity and orientation distributions for image patches when we use the color and intensity channels as the input for QFT. The differences between amplitude spectrum of QFT for image patches can show the differences for color, intensity and orientation distributions between image patches. Here we use opponent color space to represent the color information for image patches [86]. If r , g and b denote the red, green and blue color components respectively, four broadly-tuned color channels are generated as: $R = r - (g + b)/2$ for red, $G = g - (r + b)/2$ for green, $B = b - (r + g)/2$ for blue, and $Y = \frac{r+g}{2} - \frac{|r-g|}{2} - b$ for yellow. Each color channel is then decomposed into red-green and blue-yellow double opponency according to the related property of the human primary visual cortex [86]:

$$\mathcal{C}_{rg} = R - G \quad (3.2)$$

$$\mathcal{C}_{by} = B - Y \quad (3.3)$$

The intensity channel can be computed as $I = (r + g + b)/3$. We use one intensity channel I , and two color channels $\mathcal{C}_{rg}$ and $\mathcal{C}_{by}$, as three features for calculating the amplitude spectrum of QFT. Based on these three features, the quaternion representation for each image patch is as follows:

$$q(n, m) = \mathcal{I}(n, m)\mu_1 + \mathcal{C}_{rg}(n, m)\mu_2 + \mathcal{C}_{by}(n, m)\mu_3 \quad (3.4)$$

where μ_1, μ_2 and μ_3 are unit pure quaternion; $\mu_1^2 = \mu_2^2 = \mu_3^2 = -1$; $\mu_1 \perp \mu_2, \mu_2 \perp \mu_3, \mu_1 \perp \mu_3$ and $\mu_3 = \mu_1\mu_2$.

The symplectic decomposition for the above quaternion image patch can be represented as follows:

$$q(n, m) = f_1(n, m) + f_2(n, m)\mu_2 \quad (3.5)$$

$$f_1(n, m) = \mathcal{I}(n, m)\mu_1 \quad (3.6)$$

$$f_2(n, m) = \mathcal{C}_{rg}(n, m) + \mathcal{C}_{by}(n, m)\mu_1 \quad (3.7)$$

The study in [84] indicates that QFT can be calculated by using two standard complex fast Fourier transform. According to the study in [84], QFT of $q(n, m)$ in formula (3.5) can be calculated as follows:

$$\mathcal{Q}[u, v] = F_1[u, v] + F_2[u, v]\mu_2 \quad (3.8)$$

$$F_i[u, v] = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\mu_1 2\pi \left(\left(\frac{mv}{M} \right) + \left(\frac{nu}{N} \right) \right)} f_i(n, m) \quad (3.9)$$

where $i \in \{1, 2\}$; (n, m) and (u, v) represent locations for image patches in spatial and frequency domains respectively; N and M represent height and width of image patches; $f_i(n, m)$ is obtained from (3.6) and (3.7).

According to (3.4) - (3.9), we can get QFT result $\mathcal{Q}[u, v]$ for each image patch. Now we describe $\mathcal{Q}[u, v]$ in polar form as follows:

$$\mathcal{Q}[u, v] = \mathcal{A}e^{\mu\varphi} \quad (3.10)$$

where $\mathcal{A}$ is the QFT amplitude spectrum of the image patch; φ is the corresponding QFT phase spectrum; μ is a unit pure quaternion.

Actually, after we get the QFT result for image patches, the FT amplitude spectrum $\mathcal{A}$ can be calculated as:

$$\mathcal{A} = |\mathcal{Q}[u, v]| \quad (3.11)$$

From (3.11), we can get the amplitude spectrum of QFT for each image patch, to be used to represent the features of each image patch. In this study, we use the color and intensity channels for QFT, so this amplitude spectrum from QFT includes color information as well as intensity information. As we know, the amplitude spectrum indicates the presence of respective spatial frequencies in images. The value of the amplitude spectrum in a special direction of the center-shifted FT indicates the strength of the orientation in a perpendicular direction [32, 52]. We can get the strength of different orientations through using the amplitude spectrum of the center-shifted FT, as shown in [32]. Thus, the amplitude spectrum used here can represent the orientation distribution of image patches. As we use the color and intensity channels for QFT in this study, the QFT amplitude spectrum can represent color, intensity and orientation distributions for image patches.

3.2.3 Differences between Image Patches

As described previously, the saliency value of each image patch is determined by the weighted differences between this patch and its neighboring patches including all other image patches in the image. If an image patch is significantly different from its neighbors, it has a higher probability to be a salient region. The saliency value for an image patch should be larger with the larger differences between this patch and its neighbors. As the spatial distance (eccentricity) between the patch and its neighboring patches increases, the weight of this difference to the saliency value of the patch decreases. The saliency value for an image patch is calculated according to formula (3.1). Now we discuss how $\mathcal{D}_{(i,j)}$ in formula (3.1) is calculated.

We use Euclidian distance of QFT amplitude spectrum to represent the feature differences between each image patch and its neighbors. To reduce the dynamic range of

amplitude coefficients, we use logarithm operation and add the constant 1 to each original amplitude coefficient value to avoid the undefined case when $\mathcal{A}$ approaches zero. Using this algorithm, we can calculate the feature difference between image patches i and j as:

$$\mathcal{D}_{(i,j)} = \sqrt{\sum_{m_i} (\log(\mathcal{A}_{m_i}^i + 1) - \log(\mathcal{A}_{m_i}^j + 1))^2} \quad (3.12)$$

where m_i indexes all pixels in an image patch.

We use human visual sensitivity to determine the weights of QFT amplitude spectrum differences between image patches. A well accepted model is used to measure the human contrast sensitivity as a function of eccentricity [85]. The contrast sensitivity $C_s(f_s, w)$ is defined as the reciprocal of the contrast threshold $C_t(f_s, w)$ as follows:

$$C_s(f_s, w) = 1/C_t(f_s, w) \quad (3.13)$$

According to the study in [85], the contrast threshold is defined as:

$$C_t(f_s, w) = C_0 \exp(\alpha_s f_s \frac{w+w_2}{w_2}) \quad (3.14)$$

where f_s is the spatial frequency (cycles/degree), w is the retinal eccentricity (degree); C_0 is the minimum contrast threshold; α_s is the spatial frequency decay constant; w_2 is the half-resolution eccentricity. According to the experiments reported in [85], these parameters are set to $C_0 = 1/64$, $\alpha_s = 0.106$, and $w_2 = 2.3$.

The retina eccentricity w is calculated according to its relationship with viewing distance v' as Figure 3.5. Given the position of the fixation point (x_0, y_0) (the center of an image patch), the retinal eccentricity w for the position (x, y) (the center of another image patch) can be computed as follows:

$$w = \tan^{-1}(d/v') \quad (3.15)$$

where d is the Euclidian distance between (x, y) and (x_0, y_0) . The typical ratio of the

Figure 3.5 Relationship between viewing distance and retina eccentricity.

viewing distance to the picture height is in the range of 3 to 6 [87]. Here we use a ratio of 4 to determine the viewing distance.

Thus, we can get the weight α_{ij} as the normalized $C_s(f_s, w)$ based on (3.13) - (3.15). The weighting parameters α_{ij} in formula (3.1) can be calculated as follows:

$$\alpha_{ij} = \frac{1}{c_0 \exp(\alpha_s f_s \frac{w+w_2}{w_2})} \quad (3.16)$$

Based on (3.12) and (3.16), we can get the patch difference $\mathcal{D}_{(i,j)}$ and its weighting α_{ij} for the final saliency calculation in (3.1). From the description above, we can see that the saliency value for the image patch i is represented as all the contributions from the patch differences between the image patch i and all other image patches in the image (as calculated in (3.1)).

3.2.4 Patch Size and Scale for Final Saliency Value

In this model, the final saliency map is influenced by the image patch size. The conventional computational visual attention models [28, 32] choose a fixed patch size empirically. Different from the relevant saliency detection models, here we consider the characteristics of the HVS and the fovea size to determine the patch size. Given an image

patch with the size $p \times p$, the relationship between eccentricity w and viewing distance v' can be computed as follows:

$$w = \tan^{-1}(p/2v') \quad (3.17)$$

Studies have shown that the 1 to 2 degree retinal area in the fovea is with the best visual acuity and in the parafovea surrounding the fovea has lower visual acuity [88]. Here we use w_0 to represent the eccentricity for the best visual acuity, which is set as 1 degree. We set $w = \beta_0 w_0$, where $\beta_0 < 1$ to make sure that with w good visual acuity is maintained. As mentioned previously, the typical viewing distance is 3 to 6 times of the image height. Here we set the view distance as 4 times of the image height, while setting β_0 as 0.2. Setting $\beta_0 = 0.2$ means that the maximum eccentricity for the width of the image patch is 0.2 degree, which can guarantee that the whole image patch is in the area with the best visual acuity. In addition, for better effect, we divide input images into partially overlapping image patches, which is determined by the overlap-eccentricity $\varepsilon \beta_0 w_0$. We choose the parameter $\varepsilon = 0.5$.

As we can see, saliency values for all pixels in an image patch obtained based on $\mathcal{D}_{(i,j)}$ and α_{ij} are the same. Thus, the patch size would influence the final saliency map. With a smaller patch size, the final saliency map will become more distinguishable, as shown in Figure 3.6 where the saliency map with the smallest image patch size (the eccentricity of 0.15 degrees) is more distinguishable than the other two with larger patch sizes. Of course, to obtain more accurate saliency map, we can divide images into smaller image patches with larger overlapping; however, in this situation, the computational complexity will increase. Given an input image with size of $M \times N$ (where M is the width and N is the height): with the patch size of $p \times p$, the computational complexity of the proposed algorithm is $(M * N)^2 / ((1 - \lambda_o)^2 p)^2$ with λ_o overlapping. Therefore, with the smaller

Figure 3.6 Original images and its different saliency maps with different patch sizes: (a) Original images; (b) Saliency maps with the image patch, the eccentricity of whose width is 0.25 degree; (c) Saliency maps with the image patch, the eccentricity of whose width is 0.2 degree; (d) Saliency maps with the image patch, the eccentricity of whose width is 0.15 degree.

patch size or more overlapping, the computational complexity will increase. Thus, we choose the suitable patch size to compute the saliency map based on the consideration of fovea characteristics, saliency detection performance and the computational complexity.

As well as the patch size, the scale will also influence the final saliency map. In the saliency map, the saliency value for image patches from salient regions are much higher than that for these patches belonging to background. For the images with different scales, the saliency values of background are similarly low, while the saliency values of salient regions are high. Thus, using multi-scale operation can strengthen the saliency for these salient regions. We adopt the steerable pyramid algorithm in [89] to obtain the multi-scale images. This algorithm obtains multi-scale images through low-pass filtering and subsampling the input image. For simplicity, here we use the linear combination for the saliency maps from multi-scale images to obtain the final saliency map. Therefore, the saliency value of image patch i is expressed as follows:

$$S_i = \frac{1}{N_s} \sum_k S_i^k \quad (3.18)$$

where N_s is the scale number; S_i^k is the saliency value for image patch i at the k th scale. The image with the lowest scale level should not be too small for a good performance of the final saliency map. Our experiments show that the lowest scale level should not be smaller than one fourth of the original scale. Therefore, we use three different scales to get the final saliency map: the original scale, a half of the original scale and one fourth of the original scale.

3.3 Experimental Results and Discussions

From the discussion in the first section of this chapter, we know that there are some defects for saliency maps from existing saliency detection models. We have also compared the performance of the proposed method with others by providing some image samples in Figures 3.1 and 3.2. As a further experiment, we give a quantitative evaluation of the saliency map obtained from our proposed model and other relevant saliency detection models on a public database [30].

Saliency map can give the salient regions for images, which can provide the locations for salient object candidates. One efficient quantitative evaluation method for saliency detection algorithms is to detect salient objects for natural images. Many studies have used saliency map to detect salient objects for natural images [26, 27, 30, 41]. The quantitative evaluation of this experiment is based on a public database including 5000 images from Microsoft [30]. This image database includes the original images and their corresponding ground-truth indicated with bounding boxes by 9 subjects. We calculate the ground-truth map for images by averaging the 9 users' labeled-data (similar with [30]). Thus, the quantitative evaluation for a saliency detection algorithm is to see how

much the saliency map from the algorithm overlaps with the ground-truth map. Similar with the study [32], we use precision, recall, and F-measure to evaluate the performance of our proposed model. Precision is computed as the ratio of correctly detected saliency region to the detected salient region from the saliency detection algorithm. Recall is calculated as the ratio of correctly detected salient region to the ground-truth salient region. Given the ground-truth map G and the saliency map S for an image, we have:

$$\text{precision} = \sum_x g_x s_x / \sum_x s_x \quad (3.19)$$

$$\text{recall} = \sum_x g_x s_x / \sum_x g_x \quad (3.20)$$

F-measure, a harmonic mean of precision and recall, is a measure that combines precision and recall. It is calculated as follows:

$$F_{\beta_p} = \frac{(1+\beta_p) * \text{precision} * \text{recall}}{\beta_p * \text{precision} + \text{recall}} \quad (3.21)$$

where β_p is a positive parameter to decide the importance of precision over recall in computing F-measure.

Generally, the precision indicates the performance of saliency detection algorithms compared with ground-truth map. To compare the proposed model with others, we

Figure 3.7 The comparison between our proposed model and others.

always see the precision value for different algorithms, for precision value is the ratio of the correctly detected region over the whole detected region. We set $\beta_p = 0.3$ in this experiment as in [32] for fair comparison. The comparison results (precision, recall, and F-measure values) are shown in Figure 3.7.

Here we use the original experiment results of other models including models [26, 32, 34, 41] from [32]. The model in [41] is an improved version based on the model in [24]. The models for the comparison experiment are selected based on the relevance and the acceptability of existing methods in the research community. For example, the models in [26] and [32] are the recent models and adopt FT to calculate the saliency map for images (the proposed model also uses FT) and thus their results are very relevant to compare with our scheme. On the other hand, the models in [34] and [41] are two classical saliency detection models (with high citations) which provide a baseline for comparison. From Figure 3.7, we can see that the overall performance of our proposed model is better than the others under comparison in terms of all three measures.

In Figure 3.8, we give some comparison samples of saliency maps from our proposed model and others. From this figure, we can see that saliency maps from the proposed model are better than those from other existing ones. From the first, third and fourth rows of this figure, the existing saliency detection models detect some background regions as the salient regions, while the proposed model can detect the exact salient regions for images. From fifth and sixth rows of Figure 3.8, the existing models mainly detect the contour of the salient objects, while the proposed model can detect the whole salient objects exactly. Compared with the ground truth, saliency maps from our proposed model are much more accurate than those from others. Overall, the saliency detection results from the proposed model are much better than those from existing models.

Figure 3.8 Saliency maps from different saliency detection models: the first column: Original images; the second to the final columns are saliency maps from Itti's model [41], Hou's model [26], Ma's model [34], Gopalakrishnan's model [32], the proposed model, and the ground-truth from 9 subjects respectively.

3.4 Application: Image Retargeting

Saliency map plays an important role in many image processing applications such as image retargeting [2, 4, 76]. With different screen resolutions in devices such as smart phones and PDAs, the displayed image has to be resized to fit for various devices with different display sizes or aspect ratios. An effective image retargeting algorithm should preserve the visually important content without much distortion to the image context. The saliency map, which can detect the salient regions for images, measures the visual importance for image pixels in image retargeting for better representations [74, 76, 90]. In this work, we show how the good saliency map can improve the image retargeting performance and therefore demonstrate the effectiveness of the proposed saliency detection algorithm in the application of image retargeting.

The performance of image retargeting algorithms greatly depends on the visual significance map, which is used to measure the visual importance for each pixel in images. As introduced in the previous chapter, the visual significance map used in existing image retargeting algorithms includes the gradient map, the saliency map and some high-level feature maps such as the facial map, motion map and so on. In this section, we apply the saliency map from the proposed model in the framework of Seam Carving [71] to demonstrate the effectiveness of the proposed model in the application of image retargeting.

In the experiment, we compare the performance of the proposed image retargeting algorithm and three other existing image retargeting techniques [71, 73, 74] based on Microsoft database [30]. We found that our image retargeting algorithm using the saliency map from the proposed model outperforms others greatly for the images with complex background; while the performance of our proposed image retargeting

algorithm is similar with others for other images with simple background. The reason is that the saliency map from the proposed saliency detection model measures the importance of each pixel in images with complex background more accurately compared with the visual significance map from other algorithms, as shown in the first row of Figure 3.9. For the images with simple background, the visual significance maps used in other image retargeting algorithms can get similar result with the saliency map from our proposed model, as shown in the second row of Figure 3.9.

To better compare the performance of different visual significance maps in image retargeting, we have conducted a user study based on two image datasets: one includes 23 images with complex background, while the other includes 23 images with simple background. All these 46 images are selected from Microsoft database [30]. Three existing image retargeting algorithms [71, 73, 74] are utilized for performance comparison. Ten participants were involved in this experiment. Among these participants, three are female and the rest are male; their average age is around 27; and they have varying degrees of image processing knowledge, being naïve as to the design of the experiment. The experiment was conducted in the typical laboratory environment. The resolution of the screen for this user study is 1920 x 1280, which is sufficient for displaying the images. The size of the original images is 300 x 400, while the size of the retargeted images is 300 x 300. Of course, we can also choose other sizes (such as 50% size of the original images). The original image was displayed in the middle of the display screen as the reference image, while four retargeted images from four different algorithms (our proposed algorithm and other three existing ones) were displayed in random orders surrounding the reference image. Mean opinion score (MOS) (1 - 5) was recorded by participants where 1 means bad viewing experience and 5 means excellent viewing experience. Each participant voted for these 46 images. The mean score for each

Figure 3.9 Original images, visual significance maps and retargeted images from different algorithms: (a) Original images; (b) Gradient maps from algorithm [71]; (c) Retargeted images from algorithm [71]; (d) Saliency maps from Itti's model [24]; (e) Retargeted images from [74] (Itti's model is integrated into this algorithm to measure the importance of each pixel for images); (f) Saliency maps from the proposed model; (g) Retargeted images from our image retargeting algorithm based on the proposed saliency detection model.

Figure 3.10 The mean score for each retargeted image from 10 participants for four different algorithms for the first image dataset.

Figure 3.11 The mean score for each retargeted image from 10 participants for four different algorithms for the second image dataset.

Figure 3.12 Overall mean scores and variance values of 23 retargeted images from four different algorithms for the first image dataset.

retargeted image from ten participants for four different algorithms are shown in Figures 3.10 and 3.11 respectively for these two image datasets.

Figure 3.10 shows that the mean scores of retargeted images from the proposed

algorithm are much higher than the mean scores from other algorithms for the first dataset. Figure 3.12 shows the overall mean scores and the variance values of 23 retargeted images in the first dataset for four different algorithms. These two figures demonstrate that the performance of the proposed algorithm is the best and the performance of algorithms [73] and [74] are better than that from algorithm [71] for the images with complex background in the first image dataset. In Figure 3.11, it might be noted that the results of the four algorithms are close for the images with simple background in the second dataset. Figure 3.13 shows the overall mean scores and variance values of 23 images in the second dataset for four different image retargeting algorithms. In this figure, it can be seen that the overall mean score of the retargeted images from our algorithm is slightly higher than those from the other algorithms, while the variance value of the proposed algorithm is a little lower than those from other algorithms. As the overall mean scores and variance values from four algorithms are close to each other, the overall performances of four algorithms are comparable for the images with simple background.

Figure 3.13 Overall mean scores and variance values of 23 retargeted images from four different algorithms for the second image dataset.

Figure 3.14 Retargeted image samples from the first image dataset for four algorithms: (a) Original images; (b) Retargeted images from [71]; (c) Retargeted images from [74]; (d) Retargeted images from [73]; (e) Retargeted images from our algorithm. The sizes of the original images and the retargeted images are 300×400 and 300×300 respectively.

Figure 3.15 Retargeted image samples from the second image dataset for four algorithms: (a) Original images; (b) Retargeted images from [71]; (c) Retargeted images from [74]; (d) Retargeted images from [73]; (e) Retargeted images from our algorithm. The sizes of the original images and the retargeted images are 300×400 and 300×300 respectively.

Overall, Figures 3.10 - 3.13 show that the image retargeting algorithm based on our proposed saliency detection algorithm outperforms the others. Some retargeted result samples from these two datasets are depicted in Figures 3.14 and 3.15. As can be seen in Figure 3.14, the retargeted images from the other three existing image retargeting algorithms suffer some distortion in salient objects, whereas the retargeted images from our algorithm preserve the salient objects accurately. In Figure 3.15, it can be seen that the retargeted images from all four algorithms are somewhat similar and without much distortion.

3.5 Conclusions

In this chapter, we proposed a novel bottom-up saliency detection model based on both local and global feature contrast, human visual sensitivity and QFT amplitude spectrum. The proposed model first divides the input images into small image patches. It then uses QFT amplitude spectrum to represent the color, intensity and orientation distributions for image patches. The saliency value of each image patch is obtained by calculating the differences between the QFT amplitude spectrum of this patch and all other patches in the image, and the weights for these differences determined by the visual impacts of human visual sensitivity. The proposed saliency detection model also utilizes the characteristics of the HVS for the selection of patch size and multi-scale operations. Compared with existing models, the proposed model has better performance with regard to the ground truth of human-labeled salient objects. In addition, we demonstrate the advances of our proposed saliency detection model in image retargeting.

The next chapter introduces a new saliency detection model in the compressed domain for JPEG images.

Chapter 4

Saliency Detection in Compressed Domain for Images

Today, various saliency detection models have been proposed for image processing applications such as image resizing. However, all these existing saliency detection models are built in the uncompressed domain. Since most images over Internet are typically stored in the compressed domain such as JPEG and J2K, we propose a novel saliency detection model in the compressed domain for images in this chapter. The intensity, color and texture features of images are extracted from DCT coefficients in JPEG bit-stream. The saliency value of each DCT block is obtained based on the Hausdorff distance calculation of features and feature map fusion. The proposed saliency detection model in the compressed domain proves promising, as shown in the experiments. Based on the proposed saliency detection model, we further design an adaptive image retargeting algorithm in the compressed domain to demonstrate the application of the proposed saliency detection model. The proposed image retargeting algorithm utilizes multi-operator operation composed of the block-based seam carving and image scaling to resize images. A novel idea of texture homogeneity is devised to

determine the amount of removed block-based seams. Thanks to the directly derived accurate saliency information from the compressed domain and the novel defined texture homogeneity, the proposed image retargeting algorithm effectively preserves the visually important regions for images, efficiently removes the less crucial regions, and therefore significantly outperforms the relevant state-of-the-art algorithms, as demonstrated with the in-depth analysis in extensive experiments.

4.1 Introduction

As introduced in previous chapters, various saliency detection models [24 - 34] have been proposed to extract the salient regions for image processing applications. Image retargeting is one of the popular applications of saliency detection models [70 - 79]. In these existing image retargeting algorithms, the saliency map is used to measure the visual importance of image pixels for image resizing operation.

Existing saliency detection algorithms [24 - 34] and image retargeting algorithms [70 - 79] mentioned above are all implemented in the uncompressed domain. However, most images over Internet are typically stored in the compressed domain of JPEG [91]. The compressed JPEG images are widely used in various Internet-based multimedia processing applications, since they reduce the storage space and increase the downloading speed greatly for Internet users. In order to extract various low-level features (such as intensity, color, orientation, etc.) from compressed JPEG images, existing saliency detection models and image retargeting algorithms have to decompress these JPEG images from the compressed domain into the spatial domain. The full decompression from these saliency detection or image retargeting algorithms is not only computation-consuming but time-consuming as well. Compared with these existing

saliency detection and image retargeting algorithms which obtain the results in the uncompressed domain, it is crucial to design effective saliency detection and image retargeting algorithms in the compressed domain.

In this chapter, we propose a new saliency detection model based on DCT coefficients in the compressed domain. We only need partial decompression of images (not complete decompression) to obtain the DCT coefficients for saliency detection. Based on the proposed saliency detection model, we further design an adaptive image retargeting algorithm in the compressed domain to demonstrate the application of the proposed saliency detection model in image retargeting. As DCT is used in JPEG compression at 8×8 block level, the DCT coefficients are used to extract intensity, color and texture features for each 8×8 block for saliency detection. Previous studies have shown that many features can be extracted directly from DCT coefficients for image processing applications [92 - 94]. In [92], the authors extracted the color and texture features from DCT coefficients for object localization. In [93], the authors adopted several DCT coefficients in each block for object indexing. In [94], the color and texture features are extracted from DCT coefficients to measure image similarity by a statistical graph matching method. Those studies have demonstrated the effective feature extraction in the compressed domain for image processing applications. Here, we extract the intensity, color and texture features directly from the DCT coefficients for saliency detection for JPEG images. Although the Minimum Coded Unit (MCU) blocks can be as large as 16×16 (for 4:2:0 component subsampling format) [91, 95], we perform our saliency detection and image retargeting at the 8×8 block level for each DCT block. After obtaining the intensity, color and texture features from the DCT coefficients, the saliency map of images are calculated from weighted feature differences between DCT blocks.

Based on the proposed saliency detection model, we design an adaptive image

retargeting algorithm in the compressed domain. As stated in the study [78], the multi-operator image resizing algorithm can obtain better performance than the single-operator image retargeting methods. Here, we devise an adaptive multi-operator image retargeting algorithm based on seam carving and image scaling operations. The saliency map in the compressed domain is used to determine the visual importance of each 8×8 image block in JPEG images. The multi-operators including block-based seam carving and image scaling are utilized for image resizing. Existing multi-operator image retargeting algorithms [78, 79] use patch-based image similarity algorithms to determine the number of times to apply each image resizing operator (seam carving or image scaling); these algorithms aim to find the similar patches in original and retargeted images globally. The

Figure 4.1 Comparison of different image retargeting algorithms: (a) the original image; (b) the gradient map (used in [71] and [73]); (c) the saliency map from Itti's model [24] (used in [74]); (d) the saliency map from the model in [48] (used in [72]); (e) the saliency map from our proposed model; (f) - (j) the retargeted images from [71], [73], [74], [72] and our proposed algorithm respectively. The width of the retargeted images is 75 percent of that from the original image.

problem with these algorithms is that the deformed patches in retargeted images may be mismatched with wrong patches in the original images [80]. In the proposed multi-operator image retargeting algorithm, we define the texture homogeneity based on spatial distribution and connectedness of the energy map to determine the number of the times to apply each image resizing operator. It does not have the said problem in existing multi-operator image algorithms [78, 79]. Thanks to the directly derived saliency map from the compressed domain and the defined texture homogeneity, the proposed image retargeting algorithm effectively preserves the objects of attention and removes the less crucial regions, as shown in Figure 4.1. From Figure 4.1, we can see that our saliency map identifies the salient regions more accurately than the gradient map and the saliency maps from [24] and [48]. Therefore, the retargeted image from our proposed algorithm is better than these from other algorithms, as shown in Figure 4.1 (f) - (j). From this figure, the retargeted results from other algorithms suffer much distortion (Figure 4.1 (f) - (i)), while the retargeted result from our proposed algorithm can preserve the salient object without distortion (Figure 4.1 (j)). More details and comparisons will be provided in the following sections.

4.2 Framework of the Proposed Saliency Detection Model

In this section, we introduce the proposed saliency detection model in the compressed domain in detail. We first describe how to extract the intensity, color and texture features for images from DCT coefficients in JPEG bit-stream. Then the saliency calculation method is given based on these extracted low-level features. Following this section, the adaptive multi-operator image retargeting algorithm based on block-based seam carving

and image scaling are provided in the next section.

4.2.1 Feature Extraction from JPEG Bit-stream

In this model, we extract the saliency information for images from JPEG bit-stream. The saliency information is derived directly from DCT coefficients rather than from the fully decoded JPEG images. Therefore, we should extract the DCT coefficients from JPEG bit-stream. In the following subsections, we firstly introduce how to obtain DCT coefficients from JPEG bit-stream. Then we give the description of the feature extraction based on DCT coefficients.

a) Extracting DCT Coefficients from JPEG Bit-Stream

The *Baseline method* of JPEG, which is implemented based on DCT, is the most widely used compression method [91]. Thus, we mainly focus on the *Baseline method* of JPEG in this study. In the *Baseline method* of JPEG, entropy decoding is used to decode JPEG bit-stream to obtain quantized DCT coefficients. As Huffman coding is utilized to encode quantized DCT coefficients in the *Baseline method* of JPEG [91, 95], we decode JPEG bit-stream into quantized DCT coefficients according to the two sets of Huffman tables (one AC table and DC table per set). Then the de-quantization operation is applied on these quantized DCT coefficients to obtain the DCT coefficients.

The syntax for DCT-based modes of operation in JPEG standard is shown in Figure 4.2. In JPEG standard, markers are used to identify various structural parts of compressed data formats. We obtain the marker ‘Start of Image (SOI)’ in JPEG bit-stream to identify the start of a compressed image. The frame header presented at the start of a frame (JPEG image) specifies the source image characteristics, the components in the frame, and the sampling factors for each component, and specifies the destinations from which the quantized tables to be used with each component are retrieved. The parameter of Tq in

the frame header specifies the quantization table destination from which the quantization table to use for de-quantization of DCT coefficients.

Following frame header, the scan header specifies which components and which DCT quantized coefficients are contained in the scan. The parameters T_{dj} and T_{aj} in the scan header specify the DC and AC entropy coding table destinations respectively. The data following the scan header includes *ECS* and *RST* data. Each *ECS* is comprised of a sequence of entropy-coded MCUs. The *RST* is a conditional marker placed between two *ECS*s only if restart label is enabled. Detailed information of JPEG bit-stream can be found in [95]. The JPEG bit-stream can be decoded into quantized DCT coefficients based on DC and AC entropy coding tables (T_{dj} and T_{aj}) from the scan header. According to the quantization table from T_q , quantized DCT coefficients are further decoded through de-quantization operation to finally get DCT coefficients.

Figure 4.2 Syntax for DCT-based modes of operation in JPEG standard [95].

Figure 4.3 DCT coefficients and the zig-zag scanning in one 8×8 block.

b) Feature Extraction Based on DCT Coefficients

In this study, three features including intensity, color and texture are extracted based on DCT coefficients to build the saliency detection model. The DCT coefficients in one 8×8 block are shown as Figure 4.3. DCT coefficients in one block are composed of a DC coefficient and 63 AC coefficients. In each block, the DC coefficient is a measure of the average energy over all the 8×8 pixels, while the remaining 63 AC coefficients represent the detailed frequency properties of this block. The JPEG compression standard takes advantage of the fact that most energy is included in the first several low-frequency coefficients, which are in the left-upper corner of the block in Figure 4.3. The high-frequency coefficients from the right-bottom of the block are almost close to zero and thus they are neglected during quantization of DCT coefficients. The AC coefficients are ordered by zig-zag scanning from low-frequency to high-frequency in Figure 4.3.

The YCrCb color space is used to encode color images in JPEG standard. The Y

channel represents the Luminance information while the Cr and Cb channels include the Chrominance information for JPEG images. As discussed above, the DC coefficients represent the average energy of each block including 8×8 pixels. Here we firstly transfer the DC coefficients from YCrCb color space to the RGB color space to extract the intensity and color features for JPEG images. We calculate the color and intensity features by the following steps: let r , g and b denote the red, green and blue color components from DC coefficients, and four broadly-tuned color channels are generated as (same as the previous chapter): $R = r - (g + b)$ for new red component, $G = g - (r + b)/2$ for new green component, $B = b - (r + g)/2$ for new blue component and $Y = \frac{r+g}{2} - \frac{|r-b|}{2} - b$ for new yellow component. The intensity feature can be calculated as: $I = (r + g + b)/3$. Each color channel is then decomposed into red/green and blue/yellow double opponency according to the related properties of the human primary visual cortex [86]: $C_{rg} = R - G$ and $C_{by} = B - Y$.

Based on the above description, I , C_{rg} and C_{by} are the three extracted intensity and color features from DCT coefficients for an 8×8 block in JPEG images. It is noted that a 16×16 MCU consists of four 8×8 Luminance blocks and two 8×8 Chrominance blocks (one for Cb component and the other for Cr component). Thus, four Luminance blocks share the same Chrominance blocks in a typical 4:2:0 component subsampling JPEG encoding system.

The AC coefficients include the detailed frequency information for each image block and previous studies have shown that the AC coefficients can be used to represent the texture information for images blocks [92, 94, 96 - 98]. Here, we use the AC coefficients in YCrCb color space to extract the texture feature for each 8×8 block. In YCrCb color space, Cr and Cb components represent the color information and their AC coefficients

provide little information for texture. In addition, a 16×16 MCU consists more Luminance blocks compared with Chrominance blocks in a typical 4:2:0 scheme. Therefore, we use the AC coefficients from Y component only to extract the texture feature T for image blocks. As to the AC coefficients in one 8×8 DCT block, the low-frequency components capture most of the detailed information, while the high-frequency components include less information. Following the studies in [150, 151], we classify AC coefficients into three parts: Low-frequency (LF), Medium-frequency (MF), and High-frequency (HF) parts, as shown in Figure 4.4. The coefficients in each part are summed as one value to obtain three corresponding elements (t_{LF} , t_{MF} and t_{HF}) to represent the texture feature for each DCT block. Thus, the texture feature T for each DCT block can be expressed as follows.

$$T = \{t_{LF}, t_{MF}, t_{HF}\} \quad (4.1)$$

where t_{LF} , t_{MF} and t_{HF} is the sum of all the coefficients in LF, MF and HF parts respectively in Figure 4.4.

Figure 4.4 Different types of DCT coefficients in one 8×8 block.

4.2.2 Saliency Calculation

From the above subsection, the intensity, color and texture features (I , C_{rg} , C_{by} and T) can be extracted from the DCT coefficients of each 8×8 block. In this study, we use these four features (one intensity feature, two color features and one texture feature) to obtain four feature maps respectively. Then we use the coherent normalization based fusion method to combine these four feature maps to get the saliency map for JPEG images. The details of the saliency detection model are shown in the following.

a) Feature Differences between DCT Blocks

As to the intensity and color features (I , C_{rg} , C_{by}), the feature differences between blocks i and j can be computed as:

$$D_{ij}^{\iota} = \mathcal{C}_i^{\iota} - \mathcal{C}_j^{\iota} \quad (4.2)$$

where ι ($\iota \in \{1, 2, 3\}$) represents the intensity and color features respectively (one intensity feature and two color features); $\mathcal{C}^{\iota} \in \{I, C_{rg}, C_{by}\}$.

We use the vector T including three elements from (4.1) to represent the texture feature for each DCT block in JPEG images. The Hausdorff distance [99] is used to calculate the difference between two vectors of texture feature from two different blocks. Hausdorff distance is widely used to calculate the dissimilarity between two point sets through examining the fraction of points in one set that lie near points in the other set (and perhaps vice versa). The texture difference D_{ij}^4 between two blocks i and j can be computed as follows:

$$D_{ij}^4 = \max(h_m(T_i, T_j), h_m(T_j, T_i)) \quad (4.3)$$

where 4 means the texture feature is the fourth feature (the first three features include one intensity and two color features described above); T_i and T_j represent the vectors of

texture feature for blocks i and j , respectively. $h_m(T_i, T_j)$ in (4.3) can be calculated as:

$$h_m(T_i, T_j) = \max_{t_i \in T_i} \min_{t_j \in T_j} \|t_i - t_j\| \quad (4.4)$$

where $\| \cdot \|$ is the $L2$ norm.

b) Feature Maps in the Compressed Domain

In this study, the saliency value of each DCT block in each feature map is determined by two factors: one is the block differences between this DCT block and all other DCT blocks in the input image; the other is the weighting for these block differences. If these differences between this DCT block and all other DCT blocks are larger, then the saliency value of this DCT block is larger. In addition, a Gaussian model of Euclidean distances between DCT blocks are used to determine the weighting for these DCT block differences. We use the Gaussian model of Euclidean distances for its generality. In this saliency detection model, we use S_i^t to represent the saliency value calculated from the ut h feature for the DCT block i . The feature map for the ut h feature can be obtained as follows.

$$S_i^t = \sum_{j \neq i} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{d_{ij}^2}{2\sigma^2}} D_{ij}^t \quad (4.5)$$

where σ is a parameter for the Gaussian model, which determines the weighting of local and global contrast for saliency detection; D_{ij} is the Euclidean distance between DCT blocks i and j ; D_{ij}^t is calculated as (4.2) and (4.3). Based on the experiment conducted in Appendix Section, we set $\sigma = 20$.

From (4.5), the saliency value of block i considers all the block differences between this block and the other blocks in the image. The saliency value of the block i is larger with greater block differences from all other blocks in the image. Here we use a Gaussian model of Euclidean distances between blocks to weight the block differences. Based on

(4.5), the weights of the block differences from nearer neighbor blocks are larger compared with those from farther neighbor blocks. Therefore, the contributions of the block differences to the saliency value of block i will decrease with larger-distance blocks from block i . On the contrary, the contributions of the block differences to the saliency value of block i will increase with smaller-distance blocks from block i . According to (4.5), we can obtain four feature maps (one intensity feature map, two color feature maps and one texture feature map) based on the intensity, color and texture features. The final saliency map is a combination of these four feature maps. We will describe how to get the final saliency map for images in the next subsection.

c) Final Saliency Map in the Compressed Domain

After obtaining the four feature maps $S^l (l \in \{1, 2, 3, 4\})$, one intensity, two color and one texture feature maps), the saliency map for JPEG images can be obtained by fusing these four feature maps. In this study, we use the coherent normalization based fusion method to combine these four feature maps into the final saliency map S as follows:

$$S = \sum \gamma_\theta \aleph_{(\theta)} + \beta_\theta \prod \aleph_{(\theta)} \quad (4.6)$$

where $\aleph$ is the normalization operation; $\theta \in \{S^l\}$; γ_θ and β_θ are parameters determining weights for components in (4.6). In this work, we set $\gamma_\theta = \beta_\theta = 1/5$. The parameter choice is based on the assumption that weighting of different feature maps for the final saliency map is the same. The second term in (4.6) represents these regions which all the four feature maps S^l detect as salient regions.

4.3 Application: Image Retargeting in Compressed Domain

As described above, we build a saliency detection model in the compressed domain by

extracting intensity, color and texture features from DCT coefficients. To demonstrate its advantages and applications, we design a novel image retargeting algorithm in the compressed domain. The proposed saliency detection model is used to obtain the saliency map to measure the visual importance of each DCT block for image retargeting. Multi-operators including the block-based seam carving and image scaling are utilized to resize JPEG images in the proposed image retargeting algorithm. The number of block-based seam carving operation is determined by the defined texture homogeneity.

The image resizing operation steps are given as follows: (1) determine the number of block-based seam carving operation based on the defined texture homogeneity; (2) use the block-based seam carving operation to resize the original image; (3) use the image scaling operation to resize the retargeted image from the block-based seam carving to obtain the final retargeted image.

a) Block-based Seam Carving Operation

It is noted that, since our final saliency map is at 8×8 block level, each seam indicates connected blocks instead of connected pixels in the original image. We use DCT blocks with the size of 8×8 to calculate the saliency map, thus the final saliency map is only $1/64$ times of the original image and each value in the final saliency map represent the saliency value for one 8×8 DCT block. We design a block-based seam carving method based on the forward energy [71] to determine the optimal block seams. Based on the saliency map S in (4.6), the block-based seam carving uses the following dynamic programming technique to determine the optimal block-based seams.

$$M(i, j) = S(i, j) + \min \begin{cases} M(i-1, j-1) + C_L(i, j) \\ M(i-1, j) + C_U(i, j) \\ M(i-1, j+1) + C_R(i, j) \end{cases} \quad (4.7)$$

where $M(i, j)$ determines the position (i, j) of the saliency map for the optimal block-

based seams; $C_L(i, j)$, $C_U(i, j)$ and $C_R(i, j)$ are costs due to generation of new neighbor blocks separated by the removed seam previously. These costs are calculated as:

$$\begin{cases} C_U(i, j) = ||S(i, j + 1) - S(i, j - 1)|| \\ C_L(i, j) = C_U(i, j) + ||S(i - 1, j) - S(i, j - 1)|| \\ C_R(i, j) = C_U(i, j) + ||S(i - 1, j) - S(i, j + 1)|| \end{cases} \quad (4.8)$$

b) Adaptive Image Retargeting

The optimal block-based seams can be determined by (4.7) and (4.8). As introduced previously, the proposed image retargeting algorithm first utilizes the block-based seam carving operation to resize the image. Then the image scaling operation is used to obtain the final retargeted images. In this study, we propose to use texture homogeneity to decide the number of removed block-based seams. The number of removed block-based seams in dimension κ (horizontal or vertical) can be calculated as follows.

$$n_\kappa^s = \lambda_\kappa(n_\kappa - n_\kappa^r)/8 \quad (4.9)$$

where n_κ^s is the number of removed block seams in dimension κ ; λ_κ represents the texture homogeneity of the image in dimension κ , which is used to determine the number of removed block seams; n_κ is the length of the original image in dimension κ (width or height); n_κ^r is the length of the retargeted image in dimension κ (width or height). The value of n_κ^r is decided by the size of the display screen based on initial communication between the server and client in real applications. As the proposed algorithm is based on DCT blocks and the size of DCT blocks is 8×8 , we use the number 8 to calculate the number of removed block-based seams in (4.9).

Texture homogeneity is widely measured in various image processing applications [30, 100-102]. In this study, we define a measurement for texture homogeneity λ_κ to determine the number of removed blocks in dimension κ . The texture homogeneity defined here is dependent on the energy spatial-distribution and the energy connectedness

(here the saliency map from the proposed saliency detection model is used as the energy map). If the image energy is more centralized and connected, there may be only one or several small salient objects in the image with simple background. In this case, we will use more seam carving operation to remove block-based seams. On the contrary, with more disconnected and decentralized energy distribution, the image may include one or several big salient objects, or the context of the image is complex. In this case, we should use more image scaling operation to resize the image for preserving these salient objects or the context information. In this study, the texture homogeneity of an image in dimension κ (horizontal or vertical) can be computed as follows.

$$\lambda_{\kappa} = (1 - \tau_{\kappa})\varsigma_{\kappa} \quad (4.10)$$

where τ_{κ} represents the spatial variance of energy pixels in dimension κ ; ς_{κ} represents the connectedness of energy pixels in dimension κ . In this study, Otsu's threshold algorithm [103] is used to binarize the energy map (the saliency map) into energy pixels (the value is 1) and non-energy pixels (the value is 0).

To simplify the description, we just demonstrate how to calculate the horizontal variance of energy pixels here. The calculation process of the vertical variance of energy pixels is similar. The horizontal variance of energy pixels τ_1 in the image can be calculated as follows.

$$\tau_1 = \frac{1}{\mathbb{N}} \sum_{(i,j)} |i - H_v|^2 * \mathbb{E}_{(i,j)} \quad (4.11)$$

$$H_v = \frac{1}{\mathbb{N}} \sum_{(i,j)} i * \mathbb{E}_{(i,j)} \quad (4.12)$$

where $\mathbb{E}_{(i,j)}$ is the energy value for the position (i, j) ; $\mathbb{N}$ represents all the energy pixels in the image, $\mathbb{N} = \sum_{(i,j)} \mathbb{E}_{(i,j)}$.

From (4.12), we can see that H_v is the expected value of spatial locations for energy pixels in the image. Thus, we can obtain the horizontal variance of energy pixels for the

image based on (4.11) and (4.12). We set $\tau_1 = 1$ when all energy pixels are centralized into one square in the image. In this case, the image texture is the most homogeneous. On the contrary, we set $\tau_1 = 0$ when all energy pixels are distributed uniformly over the image. In this case, the image texture is considered the most inhomogeneous. We normalize τ_1 from (4.11) based on these two cases and then use the normalized τ_1 to calculate λ_1 in (4.10).

The connectedness of energy pixels in the image is measured by the number of energy pixels in the neighborhood of all energy pixels in the image. For each dimension (horizontal or vertical) of the image, there are at most 6 neighbor pixels for each energy pixel. The other two neighbor pixels are from the other dimension and thus not considered. We use ς_κ^i to represent the connectedness of the energy pixel i for dimension κ . It can be computed as follows [100].

$$\varsigma_\kappa^i = \frac{1}{6} \sum_{z \in M_i} f(z) \quad (4.13)$$

where M_i includes all 6 neighbor pixels around i ; $f(z)$ is the function to denote whether the neighbor pixel z is an energy pixel or not.

The connectedness of image energy in dimension κ is obtained as the sum of the connectedness of all energy pixels in the image as follows.

$$\varsigma_\kappa = \frac{1}{N} \sum_i \varsigma_\kappa^i \quad (4.14)$$

where N is the number of energy pixels in the image, as that in (4.11).

We can obtain the connectedness of image energy according to (4.13) and (4.14). We set $\varsigma_\kappa = 1$ when energy pixels in the image are centralized as a connected square. In this case, the image texture has the greatest connectedness with this amount of energy pixels. On the contrary, we set $\varsigma_\kappa = 0$ when the energy pixels in the image are distributed uniformly over the image. Based on these two cases, ς_κ from (4.13) is normalized

between 0 and 1 as the relative connectedness to be used in (4.10).

Therefore, the amount of removed block-based seams for images can be obtained according to (4.9) - (4.14). After we utilize the block-based seam carving to remove the optimal block-based seams, the image scaling is used to scale the retargeted image from the block-based seam carving to obtain the final retargeted image.

4.4 Overall Experimental Evaluation

In this section, we evaluate the overall performance of the proposed algorithms from two aspects: one is the performance evaluation of the proposed saliency detection algorithm; the other is the performance evaluation of the proposed image retargeting algorithm in resizing images. The following two subsections give the performance comparisons between the proposed algorithms and existing ones from these two aspects respectively.

4.4.1 Saliency Detection Evaluation

As described in the previous chapter, saliency detection models are widely used to detect salient objects [3, 30, 41], for it can provide the positions of salient objects in images. An efficient qualitative evaluation method for the performance of saliency detection algorithms is to extract the salient objects in images. Currently, salient object detection is widely utilized in the performance evaluation of saliency detection algorithms [3, 24, 26, 30]. In this work, we use the public database provided by Achanta *et al.* [48] to evaluate the proposed saliency detection algorithm. This database includes 1000 original images from Microsoft database [30] and the ground-truth maps. The ground truth maps in this database [48] include the accurately human-labeled salient objects, rather than the human-labeled boxes including the salient objects in [30]. Thus,

the ground-truth in the database [48] is more accurate than that from the database [30], as shown in the final row of Figure 4.6. We compare the proposed saliency detection algorithm with other six existing ones: IT [24], FT [48], SR [26], CA [49], MR [3] and MZ [34]. Three (FT, SR and MR) of these algorithms are implemented in the frequency domain while the other three (IT, CA and MZ) are built in the spatial domain. In the public database [48], the sizes of the human-labeled ground truth are the same as those of the original images. Therefore, the saliency maps from all saliency detection algorithms are resized to the sizes of the original images for fair comparison.

Different from the evaluation method in the previous chapter (Precision, Recall and F-Measure), here we use ROC to evaluate the performance of the proposed saliency detection model. ROC is widely used to evaluate the performance of the saliency map [3, 49]. The saliency map obtained by a computational saliency detection model can be divided into salient points and non-salient points through a defined threshold. The ground truth map marked by subjects includes target points and background points. The percentage of target points falling into salient points from a computational saliency detection model is True Positive Rate (TPR), while the percentage of background points falling into salient points is False Positive Rate (FPR). The ROC curve for a specified saliency detection algorithm can be achieved as the curve of TPR vs. FPR through choosing different thresholds. Therefore, the ROC curve can avoid the problem of choosing one threshold for calculating the results in F-Measure evaluation method. The overall quantitative performance for this specified saliency detection algorithm can be

Table 4.1: Comparison of AUC from different saliency detection models.

Models	IT in [24]	SR in [26]	MR in [3]	MZ in [34]	FT in [48]	CA in [49]	Model in Chapter 3	Our Model
AUC	0.8028	0.8025	0.8630	0.7951	0.8229	0.8834	0.8643	0.9414

Figure 4.5 Comparison results of ROC curves from different saliency detection models.

given by the area under the ROC curve (AUC). Generally, the larger the AUC is, the better the predication performance of the saliency detection algorithm is for the salient object detection.

The ground truth map from the database [48] includes target points (the accurately labeled salient objects in images) and background points (the accurately labeled non-salient areas), which can be shown in the last row of Figure 4.6. After obtaining saliency maps from different saliency detection models, we choose different thresholds (100 thresholds varied uniformly) from 0 to 255 for these saliency maps to get the ROC curves for different saliency detection models. The ROC curves for different saliency detection models are shown in Figure 4.5. Table 4.1 shows the AUC for all these saliency detection algorithms (we also show the experimental result of our proposed model in Chapter 3). From Figure 4.5 and Table 4.1, our proposed algorithm obviously outperforms other existing ones. According to Table 4.1, the AUC from MR in [3], CA in

Figure 4.6 Saliency map samples from different saliency detection algorithms. The first row: the original images. The second to ninth rows: the saliency maps from IT in [24], SR in [26], MR in [3], MZ in [34], FT in [48], CA in [49], our proposed model and the ground truth, respectively.

[49] and our proposed model in Chapter 3 are larger than those from IT in [24], SR in

[26], and MZ in [34]. This is coincident with the experimental results in the studies [3], [49] and Chapter 3.

We also give some comparison results visually in Figure 4.6. From the second and fifth rows in the figure, we can see that the saliency maps from [24] and [34] mainly detect the contour of the salient objects. The reason for this is that the saliency detection models in [24] and [34] mainly consider the local contrast for calculating the saliency map for the image. Furthermore, the saliency detection models from [24] and [34] also detect some non-salient regions as salient, as shown in saliency maps from the second and fifth rows in Figure 4.6. The saliency maps in the third and fourth rows are obtained from the phase-based saliency detection models [3, 26]. In these models, the FT is used for images to obtain the amplitude and phase firstly, then the saliency map is calculated as the reconstructed image through IFT on user-defined constant amplitude (always set as 1) and the original phase. These models obtain the saliency maps for images mainly by global contrast. Thus, the saliency maps from the third and fourth rows mainly include the high-frequency regions (such as edges) as salient. To improve the saliency detection model [26], the authors in [48] built the saliency detection model by retaining more frequency content for images. However, this model still loses some visually important information in the saliency map (saliency maps from the first, fifth, sixth and seventh columns in the sixth row of Figure 4.6) or detects some background regions as salient (saliency maps from the second and fourth columns in the sixth row of Figure 4.6). Although the saliency maps in the seventh row seems better than those from models [3, 24, 26, 48], they suffer the defect that the contour of salient objects is much more salient than other parts of salient objects in images, as shown in the seventh row of Figure 4.6. On the contrary, our proposed saliency detection can obtain more accurate salient regions for images, as shown from the saliency maps in the eighth row of Figure 4.6.

Figure 4.7 Comparison between saliency maps of images with different compression ratios. The first row: original images with compression ratio about 1.34 bpp. The second to the fourth rows: the saliency maps from the images with compression ratios about 1.34 bpp, 1.15 bpp and 0.86 bpp respectively.

Figure 4.8 Comparison between saliency maps of images with different sizes. The first row: original images. The second to fourth rows: the saliency maps of images with the original size, 80% and 60% of the original size respectively.

It is commonly accepted that the texture features of DCT blocks will change with different compression ratios. However, the change will not significantly affect saliency map in the proposed algorithm. We have conducted experiments to investigate the influence of the compression ratio. Experimental results demonstrate that the image compression has little influence on the saliency map. Figure 4.7 gives the saliency maps from some sample images with different compression ratios. From this figure, we can see the saliency maps from the second to fourth rows for each image are very similar. On the other hand, one recent study has shown that video coding impairments would not disturb the visual attention regions from observers [104]. Therefore, the normal image/video compression will not change the saliency results from the proposed algorithm and the observers. In addition, we have also conducted experiments for the images with different sizes. Experimental results have shown that the saliency map will not change greatly with the changed image sizes. Figure 4.8 gives the saliency maps from some sample images with different sizes. From this figure, we can see that there is little difference between the saliency maps from images with different sizes. In this study, the saliency value for each

Figure 4.9 Overall results of retargeted images from different image retargeting algorithms.

Figure 4.10 Detailed results of retargeted images from different image retargeting algorithms.

DCT block is calculated through computing the weighted differences between this DCT block and all other DCT blocks in the image. In JPEG standard, the size of each DCT block is 8×8 and the DCT blocks are less in images with smaller sizes. Therefore, the computational cost will decrease with smaller image sizes.

4.4.2 Image Retargeting Evaluation

As we know, the performance of image retargeting algorithms greatly depends on two factors: the used visual significance map and the image resizing operation. In this study, the saliency map in the compressed domain is used as the visual significance map to measure the visual importance for each pixel in images. The performance of the proposed saliency detection in the compressed domain is demonstrated in the above experiment. In this experiment, we compare the performance of the proposed image resizing operation with that from other existing ones.

Figure 4.11 Comparison I of different image retargeting algorithms. The first column: the original images. The second to sixth columns: the retargeted images from [71], [73], [105], [74] and our proposed algorithm respectively. The width of the retargeted images is 75% of the width of the original images.

To demonstrate the performance of our proposed image retargeting algorithm in the compressed domain, we use an image dataset including 25 images to conduct a user study. The compression ratio is about 1.34 bpp (bits per pixel) for these images. Three state-of-the-art image retargeting algorithms [71, 73, 105] are adopted for the performance evaluation. 11 participants involved in this experiment. Among these participants, four are female and the rest are male; their average age is around 27; and they have varying degrees of image processing knowledge, being naïve as to the design of the experiment. The experiment was conducted in the typical laboratory environment. The resolution of the screen for this user study is 1920 x 1280, which is sufficient for displaying the images. The size of the original images is 300 x 400, while the size of the retargeted images is 300 x 300. The original images are used as reference images in this

Figure 4.12 Comparison II of different image retargeting algorithms. The first column: the original images. The second to sixth columns: the retargeted images from [71], [73], [105], [74] and our proposed algorithm respectively. The height of the retargeted images is 75% of the height of the original images.

user study. All retargeted images from four different algorithms are displayed in random orders on the screen. MOS (1- 5) was recorded by participants where 1 means bad viewing experience and 5 means excellent viewing experience. Each participant votes for this image dataset. The statistical results for the retargeted images are shown in Figures 4.9 and 4.10.

Figure 4.9 shows the overall mean scores and variances from four different image retargeted algorithms based on the 25 images. From this figure, we can see that the mean score from our proposed algorithm is higher than those from other three ones. This means that the overall viewing experience from the retargeted images from our proposed algorithm is better than those from others. Meanwhile, the score variance from our proposed algorithm is lower than those from others. This means that the retargeted result

Figure 4.13 Comparison III of different image retargeting algorithms. The first column: the original images. The second to sixth columns: the retargeted images from [71], [73], [105], [74] and our proposed algorithm respectively. The width and height of the retargeted images are 75% of the width and height of the original images respectively.

from our proposed algorithm is more stable than those from others. From this figure, we can see that the overall viewing experience from the algorithm [105] is better than other two algorithms [71, 73]. However, the stability from the algorithm [105] is not as good as the other two [71, 73]. Figure 4.10 presents the number of the retargeted images under each score. From this figure, we can see that most of the retargeted images from our proposed algorithm provide better viewing experience for users.

Here, we give some visual comparison results in Figures 4.11 - 4.14. In these figures, we add the retargeted results of the saliency-based image retargeting algorithm [74] for

comparison. From Figures 4.11 - 4.14, we can see that the retargeted images from the algorithm [71] suffer serious distortion. The viewing experience of the retargeted images from algorithms [73] and [74] is better than that from algorithm [71]. However, there is still some distortion in these retargeted images from [73] and [74], as shown in the first and third rows in Figure 4.11, and the second row in Figure 4.13. The retargeted coin from [73] and [74] in the second row of Figure 4.13 obviously suffer serious distortion. In addition, compared with the original images, the salient objects in the retargeted

Figure 4.14 Comparison IV of different image retargeting algorithms. The first row: the original images; the second to the last rows: the retargeted images with the 80%, 60% and 50% width of the original images respectively. For the retargeted images in the second to the last rows, the first to fifth columns are the retargeted images from [71], [73], [105], [74] and our proposed algorithm respectively.

images from [73] and [74] become much smaller, as shown in the first row in Figure 4.11 and the third row in Figure 4.13. From the fourth column in Figures 4.11 - 4.13, we can see that the retargeted images from [105] obviously lose some important visual information. In addition, there is some serious distortion in some retargeted images from [105], as shown in the third row in Figures 4.12 and 4.13. On the contrary, the retargeted images from our proposed image retargeting algorithm are obviously better than those from other algorithms. In Figure 4.14, the retargeted images with different sizes from one sample image are given for different image retargeting algorithms. From this figure, we can see that there is much distortion with the retargeted images from algorithms [71], [73] and [74]. From the third column in Figure 4.14, the retargeted image with 80 percent width from [105] can preserve salient objects well in the image. However, the retargeted images with smaller width sizes lose much important visual information, as shown in the second and third retargeted images from [105] in the third column of Figure 4.14.

4.5 Conclusion

Saliency detection is widely used to extract ROI for images in various image processing applications. Existing saliency detection algorithms are implemented in the uncompressed domain. However, images over Internet are typically stored in the compressed format such as JPEG, J2K and so on. In this work, we propose a novel saliency detection model in the compressed domain. Based on the proposed saliency detection in the compressed domain, we further design a novel adaptive image retargeting algorithm in the compressed domain to demonstrate the advantages and applications of the proposed saliency detection in the compressed domain.

Firstly, we extract the intensity, color and texture features from DCT coefficients in

JPEG bit-stream to calculate DCT block differences based on Hausdorff distance. Combining the Gaussian model of Euclidean distances between DCT blocks, we utilize DCT block differences to calculate the saliency map for JPEG images. Experimental results show that the proposed saliency detection model in the compressed domain outperforms other existing ones.

Furthermore, based on the proposed saliency detection model, we devise a novel adaptive image retargeting algorithm in the compressed domain. The saliency map from our proposed saliency detection model is used as the visual significance map to measure the visual importance of image pixels for our image retargeting algorithm. The multi-operator operation including block-based seam carving and image scaling is utilized for image resizing. Different from existing studies which use the block-based image similarity to determine the number of seam carving operation, a novel idea of texture homogeneity is defined to determine the number of removed block-based seams in this study. Experimental results show that the performance of the proposed image retargeting algorithm is better than that of existing ones.

In the next chapter, we extend the proposed model in this chapter to video saliency detection in the compressed domain.

Chapter 5

Saliency Detection in Compressed Domain for Video

Both of the proposed saliency detection models in the previous two chapters are designed for images. Besides images, more and more video files are uploaded to various websites by Internet users. As introduced in the chapter of literature review, the motion factor has to be considered in video saliency detection compared with the saliency detection for images. Recently, several saliency detection models have been proposed for video in the uncompressed domain. However, video over Internet are always stored in the compressed domain such as MPEG2, H.264, MPEG4 Visual, etc. In this work, we propose a video saliency detection model based on the feature contrast in the compressed domain. Four types of features including luminance, color, texture and motion are extracted from DCT coefficients and motion vectors in video bit-stream. The static saliency map of the non-predicted frames (I frames) is calculated based on the luminance, color and texture features, while the motion saliency map of the predicted frames (P and B frames) is computed by the motion feature. A new fusion method is then designed to combine the static saliency map and motion saliency map to get the final

saliency map for video frames. Due to the directly derived features in the compressed domain, the proposed model can predict the salient regions efficiently for video frames. Experimental results on a public database show superior performance of the proposed video saliency detection model in the compressed domain.

5.1 Introduction

Currently, many saliency detection models have been proposed for video in the uncompressed domain [3, 45, 59, 60 - 64]. As introduced previously, compressed video are widely used in various Internet-based multimedia applications, since they can reduce the storage space and increase the delivering speed for Internet users greatly. Existing saliency detection models for video have to decompress compressed video into the spatial domain for feature extraction. The full decompression process for video is not only time-consuming but computation-consuming as well. Therefore, the video saliency detection algorithm in the compressed domain is much desired for various Internet-based multimedia applications.

In the previous chapter, we propose a saliency detection model in the compressed domain for images. The saliency map is calculated based on features extracted from DCT coefficients in JPEG bit-stream. However, this model is designed for JPEG images and does not include motion feature extraction, which has to be considered in video saliency detection models. In addition, the encoding algorithms for video and images differ greatly. Therefore, the model proposed in the previous chapter cannot be used for video saliency detection. To the best of our knowledge, there is still no saliency detection model in the compressed domain for video. As an extension to our work in the previous chapter, in this chapter, we propose a saliency detection model in the compressed domain

for video.

In the study of this chapter, the MPEG4 Visual (Advanced Simple Profile, ASP) coded video are used for saliency detection experiments. Video of other types such as MPEG2 can be processed in the similar way. As we know, the YCrCb color space is used in MPEG4 ASP video [106], where Y represents the luminance component, and the other two (Cr and Cb) are used to represent the Chroma components. According to MPEG4 ASP standard, video frames are divided into 16×16 macroblocks for luminance component and 8×8 macroblocks for Chroma components in 4:2:0 Chroma subsampling [106]. Each 4:2:0 coded unit block consists of six 8×8 blocks: four 8×8 Luminance blocks (Luminance channel) and two 8×8 Chrominance blocks (one is Cr channel block while the other is Cb channel block). The data in video bit-stream is always processed by 8×8 blocks and thus we perform the saliency detection at 8×8 block level here.

In this model, the features of luminance, color and texture are extracted from DCT coefficients of the non-predicted frames (I frames) in video bit-stream, and these features are adopted for static saliency map calculation for these non-predicted frames. Meanwhile, the motion feature is extracted from motion vectors of the predicted frames (P and B frames) in video bit-stream, and this motion feature is adopted for motion saliency map calculation for these predicted frames. A new fusion method is proposed to combine the static saliency map and motion saliency map to get the final saliency map for video frames. Because of the directly derived features from the compressed domain and the new fusion method for the static saliency map and motion saliency map, the proposed saliency detection model in the compressed domain obtains promising results, as shown in the experimental section.

Figure 5.1 The proposed video saliency detection model.

5.2 The Proposed Video Saliency Detection Model

The proposed framework is depicted in Figure 5.1. Firstly, three features including luminance, color and texture are extracted from video bit-stream for the non-predicted

frames (I frames), and the motion feature is extracted from motion vectors in video bit-stream for the predicted frames (P and B frames). Then, the static saliency map is calculated based on the features of luminance, color and texture for the non-predicted frames, while the motion saliency map is calculated based on the motion feature for the predicted frames. Finally, the static saliency map and the motion saliency map are combined to get the final saliency map for video frames. Here, we use the MPEG4 ASP video to extract the features in the compressed domain. The features of other video such as MPEG2 video can be extracted in the similar way. In the following subsections, we firstly demonstrate how to extract features from video bit-stream. Then we present how to obtain the saliency map for video frames from these extracted features in the compressed domain.

5.2.1 Feature Extraction from Video Bit-stream

The proposed model uses DCT coefficients of the non-predicted frames (I frames) to get the luminance, color and texture features, while motion vectors of the predicted frames (P and B frames) are extracted to get the motion feature for the motion saliency detection. Here, we do not use DCT coefficients of the predicted frames, since these DCT coefficients represent the inter-predicted block residue information. This residue information cannot be used to obtain the static saliency map by the method in our model. Meanwhile, there are no motion vectors for the non-predicted frames in video bit-stream. Therefore, in this chapter, the DCT coefficients of the non-predicted frames are used to calculate the static saliency map for these non-predicted frames, while the motion vectors of the predicted frames are used to compute the motion saliency map for the predicted frames. In the following sections, we will firstly introduce how to obtain the DCT coefficients and motion vectors from the video bit-stream. Then we will present how to

Figure 5.2 Data organization. The first row: the data organization in one video packet. The second row: The data organization in one macroblock within the video packet.

calculate the features of luminance, color, texture and motion from these DCT coefficients and motion vectors.

a) DCT Coefficient and Motion Vector Extraction from Video Bit-stream

To extract the DCT coefficients from video bit-stream, we firstly give a brief description of MPEG4 video coding [106 - 107]. Generally, a MPEG4 visual scene consists of several Video Objects (VOs). The VO can be a rectangular frame, or it can be an arbitrarily shaped object corresponding to an object or background of the scene. A VO can consist of several Video Object Layers (VOLs). A VOL can represent different parts of a VO or different layers of a scalable bit-stream. The instances of a VOL at a given time are called Video Object Plane (VOP). The motion, texture and shape coding tools are adopted by the video encoder to encode VOP using I, P and B modes. For process convenience, consecutive VOPs are grouped into Group of Video Object Planes (GOVP). The highest syntactic structure of Video Session (VS) consists of several VOs. The details can be found in [106]. The VSs, VOs, VOLs, GOVPs and VOPs can be accessed by corresponding start code values in video bit-stream.

A natural video object is composed of a sequence (at different time points) of 2D representations, which are refereed as VOPs [106]. The VOPs are coded using macroblocks by exploiting both temporal redundancies and spatial redundancies. Usually, a VOP consists of one or several video packets (slices) and each video packet is

composed of an integer number of consecutive macroblocks. The data organization in one video packet is shown in the first row of Figure 5.2 [106, 108]. For the data in one video packet, the *Resync* is the periodic resynchronization marker. The *MB No.* indicates the macroblock number of the first macroblock in this video packet. *QP* is the quantization parameter. *HEC* and *Header* are the extended head markers. The last is the macroblock data in this video packet. The data organization in one macroblock is shown in the second row of Figure 5.2. The *COD* is the marker to indicate whether a certain macroblock is coded or not. The *MCBPC* is a data field to indicate the mode of the macroblock and which of the two color blocks is coded. The *DQUANT* is an optional field related to quantization values. The *MV Data* and *DCT Data* are the coded motion vector data and coded DCT coefficients respectively. It is noted that the coded motion vector data is the motion vector differences (predictively coded with respect to the neighboring motion vectors) after the Variable Length Coding (VLC). The coded DCT coefficients are the 64 DCT coefficients encoded by zig-zag scanning and run-length-encoded, and the VLC.

Figure 5.3 Progressive coding: DCT coefficients and zig-zag scanning in one 8×8 block.

The differential motion vectors can be extracted from the coded motion vector data based on VLC tables. Then it is added to a motion vector predictor component to form the real motion vector for the predicted frames [106]. In the similar way, the VLC tables of DCT coefficients are used to decode the coded DCT coefficients. Then the fixed length decoding is used to obtain the real DCT coefficients for video frames [106].

b) Feature Calculation Based on DCT Coefficients

In MPEG4 ASP video, DCT coefficients in one 8×8 block are composed of one DC coefficient and 63 AC coefficients, as shown in Figure 5.3. In each block, the DC coefficient is a measure of the average energy for the 8×8 block, while the other 63 AC coefficients represent the detailed frequency properties of this block. As mentioned above, the YCrCb color space is used in MPEG4 video bit-stream. In the YCrCb color space, Y channel represents the Luminance component, while Cr and Cb represent two Chroma components. Here, the DC coefficients in DCT blocks from the Y, Cr, and Cb channels are used to represent one luminance feature and two color features for 8×8 blocks as follows:

$$L = DC_Y \quad (5.1)$$

$$C_1 = DC_{C_r} \quad (5.2)$$

$$C_2 = DC_{C_b} \quad (5.3)$$

where L , C_1 and C_2 represent one luminance and two color features in each 8×8 DCT block respectively; DC_Y , DC_{C_r} and DC_{C_b} are the DC coefficients from Y, Cr, Cb components in each DCT block respectively. It is noted that four 8×8 Luminance blocks share two 8×8 Chrominance blocks in 4:2:0 Chrominance format.

As mentioned above, the AC coefficients include the detailed frequency information and existing studies have shown that AC coefficients can represent the texture

information for image blocks [94, 96]. In the YCrCb color space, the Cr and Cb components mainly include color information and little texture information is included in these two channels. Thus, only AC coefficients in Y component are used to represent the texture information for images. In one DCT block, most of the energy is included in the first several low-frequency coefficients, which are in the left-upper corner of the block in Figure 5.3. The AC coefficients in the right-bottom of the DCT block are equal or close to zero and they are neglected during quantization in the coding process. In the progressive coding, the AC coefficients in one DCT block are ordered by zig-zag scanning, as shown in Figure 5.3. As the high-frequency AC coefficients include little energy for each DCT block, we use the first several AC coefficients to represent the texture feature of DCT blocks. The existing study in [94] has shown that the first 9 AC coefficients can represent most energy in each DCT block. Therefore, here we use the first 9 AC coefficients in each DCT block to represent the texture feature as follows.

$$T = \{AC_{01}, AC_{10}, AC_{20}, AC_{11} \dots, AC_{30}\} \quad (5.4)$$

c) Motion Feature Calculation based on Motion Vectors

In this study, the extracted motion vectors from video bit-stream are used to calculate the motion feature for the predicted frames. In MPEG4 ASP video, there are two types of predicted frames: P frames use motion compensated prediction from a past reference frame, while B frames are bi-directionally predictive-coded by using motion compensated prediction from a past and/or a future reference frame. As there is just one prediction direction (predicted from a past reference frame) for P frames, the original motion vector MV are used to represent the motion feature for P frames. As B frames might include two types of motion compensated prediction (the backward and forward prediction), we calculate the motion vectors for B frames as follows: assume the motion compensated prediction values from the past reference and the future reference frames

Figure 5.4 Motion feature extraction.

are MV_p and MV_f respectively; the motion feature of B frames is calculated as follows.

$$V = MV_p + (-1) * MV_f \quad (5.5)$$

The process of the motion feature extraction is demonstrated as Figure 5.4. The motion feature of each DCT block in B frames is obtained from (5.5), while the original motion vector is used to represent the motion feature for each DCT block in P frames. Now, we can get the luminance, color and texture features for the non-predicted frames as L , C_1 , C_2 and T . The motion feature of the predicted frames can be obtained as V . We will describe how to calculate the saliency map for video frames based on these features.

5.2.2 Saliency Detection in the Compressed Domain

Based on the above description, the luminance, color and texture features (L , C_1 , C_2 and T) for the non-predicted frames can be extracted from DCT coefficients in video bit-stream. The motion feature (V) of the predicted frames can be obtained from motion vectors in video bit-stream. In this subsection, we first describe how to calculate the static saliency map based on these extracted features of luminance, color and texture. Then the motion saliency calculation based on the extracted motion feature is given. Lastly, a new fusion method of static and motion saliency maps is introduced.

a) Static and Motion Saliency Map Calculation

Existing studies have shown that observers will be attracted by the regions with

different features from its surrounding regions when looking at a natural scene [14, 38]. The features which can be used to discriminate the image regions include intensity, color, motion and so on. Based on FIT [14], the center-surround differences of 8×8 DCT blocks are used to detect the salient regions for images. The features of luminance, color, texture and motion extracted from DCT coefficients and motion vectors are used to calculate the DCT block differences for saliency detection in this study.

It is commonly accepted that the HVS is highly space-variant since there are different densities of cone photoreceptor cells in the retina in human eyes [82]. On the retina, the fovea has highest density of cone photoreceptor cells and thus the focus area is perceived at the highest resolution. The visual acuity decreases with the increasing eccentricity from the fixation areas [82]. This means that the HVS will be more sensitive to the center-surround differences from blocks with nearer distance compared with those from farther blocks. As similar with the previous chapter, we use a Gaussian model to simulate this mechanism. It is used to weight the center-surround differences between image blocks for saliency detection. The feature map of video frames is calculated as follows.

$$S_i^l = \sum_{j \neq i} \alpha_{ij} D_{ij}^l \quad (5.6)$$

$$\alpha_{ij} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d_{ij}^2}{2\sigma^2}} \quad (5.7)$$

where S_i^l indicates saliency value of the i th DCT block in the feature map with feature l ; $l \in \{L, C_1, C_2, T, V\}$; σ is a parameter of the Gaussian distribution, which determines the weighting of local and global contrast for saliency detection (we set $\sigma = 20$ based on the experiment in Appendix Section); d_{ij} is the Euclidean distance between DCT blocks i and j ; D_{ij}^l is the feature differences between DCT blocks i and j with feature l .

As depicted in Figure 5.1, the features of luminance, color and texture are used to calculate the static saliency map for non-predicted frames. The luminance and color

features only include one DC coefficient value from the luminance and color channels, thus the feature differences of luminance and color between DCT blocks are represented as the DC coefficient differences from the luminance and color channels. Since the texture feature is represented as a vector including 9 AC coefficients, the Euclidean distance between vectors is used to compute the texture difference between DCT blocks (the reason we use Euclidean distance instead of Hausdoff distance here is that the Euclidean distance calculation is much faster than the Hausdoff distance calculation). The static saliency map S_s for the non-predicted frames is calculated as the linear combination of four feature maps from luminance, color and texture features (L, C_1, C_2, T) as follows:

$$S_s = \sum \beta_{\theta} \aleph_{\theta} \quad (5.8)$$

where $\aleph$ is the normalization operation; $\theta \in \{S^t\}$; β_{θ} is the parameter determining the weighting for each feature map. In this work, we regard the weighting of the four features (one luminance feature, two color features and one texture feature) the same and thus set $\beta_{\theta} = 1/4$.

The motion feature V is used to calculate motion feature differences between DCT blocks for computing the motion saliency map S_m based on (5.6) - (5.7) for the predicted frames. The final saliency map for video frames is calculated by combining the static saliency map S_s and the motion saliency map S_m . We will describe how to compute the final saliency map in the following section.

b) Final Saliency Map Calculation

Based on the above description, we can obtain the static saliency map for the non-predicted video frames (I frames) and the motion saliency map for the predicted video frames (P and B frames). The motion saliency map of the non-predicted frames cannot be

calculated, since there are no motion vectors for non-predicted frames in video bit-stream. However, there may be motion in non-predicted frames (in general, there may be motion in all frames except the first I frame in video). Meanwhile, the static saliency map of the predicted frames cannot be computed, since DCT coefficients of these predicted frames represent the DCT block residue information and cannot be used to calculate the static saliency map. Here, we use the static/motion saliency map of the previous non-predicted/predicted frames to replace that of the current predicted/non-predicted ones based on the implicit memory theory [109 - 110]. Existing studies have shown that the focal attention and eye movements are guided by the recently attended locations and the implicit memory traces of context cueing [109 - 111]. These studies demonstrate that the previous attended targets will trigger the attention traces utilized in the following several fixations. It means that human eyes will continue to focus on the similar locations in future frames with these from the previous frames without much context change. Generally, the content in the consecutive video frames will not change greatly and thus the saliency maps (static or motion) of the consecutive video frames are very similar in video. Therefore, we can use the static or motion saliency map of the previous video frames to represent that of the current ones.

As there is no motion saliency map for the non-predicted frames, the motion saliency map of the previous predicted frame is adopted to represent that of the current non-predicted frame. Thus, the final saliency map for the non-predicted frames (I frames) is calculated as follows.

$$S = f(S_s, S_{m_p}) \quad (5.9)$$

where S is the final saliency map of the current non-predicted frame; S_s is the static saliency map of the current non-predicted frame; S_{m_p} is the motion saliency map of the

previous predicted frame; $f(S_1, S_2)$ is the fusion function to get the final saliency map from the saliency maps of S_1 and S_2 .

Similarly, the static saliency map of the previous non-predicted frame is used to represent that of the current predicted frame, and thus the final saliency map of the predicted frames (P and B frames) is computed as follows.

$$S = f(S_{sp}, S_m) \quad (5.10)$$

where S is the final saliency map of the current predicted frame; S_{sp} is the static saliency map of the previous non-predicted frame; S_m is the motion saliency map of the current predicted frame; $f(S_1, S_2)$ is the fusion function to get the final saliency map from saliency maps of S_1 and S_2 .

According to (5.9) and (5.10), we can calculate the final saliency map for video frames based on the static saliency map and motion saliency map. We will describe the fusion method f (in (5.9) and (5.10)) for the static and motion saliency maps to get the final saliency map in the following section.

c) Saliency Map Fusion

Currently, there are many fusion methods for combining the static saliency map and motion saliency map into the final saliency map [112]. In this study, we have tried several common fusion methods in [112] for the combination of the static saliency map and motion saliency map as follows.

Normalized and Sum (NS): the most simple fusion method which normalizes the static saliency map and motion saliency map to the same dynamic range and then sums these two maps to obtain the final saliency map as follows [112].

$$S = \sum_i \mathfrak{N}(S_\theta) \quad (5.11)$$

where S is the final saliency map; $\mathfrak{N}$ is the normalization operator; S_θ is the static or

motion saliency map; $S_q \in \{S_s, S_m\}$.

Normalization and Maximum (NM): the fusion method which normalizes the static and motion saliency maps to the same dynamic range, and then uses the maximum value in the static and motion saliency maps as the final saliency value at each location.

$$S = \max_q \mathfrak{K}(S_q) \quad (5.12)$$

where $\max$ is the maximum operator.

Normalization and Product (NP): the fusion method which normalizes the static saliency map and motion saliency map to the same dynamic range and then products the static saliency map and motion saliency map to obtain the final saliency map.

$$S = \prod_q \mathfrak{K}(S_q) \quad (5.13)$$

These three methods are the common fusion methods in the research area. However, there is no spatial competition between the static saliency map and motion saliency map with these above fusion methods. In these fusion methods, the static saliency map and motion saliency map are considered with the same weighting whatever the differences between these two maps. To address the drawbacks with these fusion methods, here we propose a new fusion method of **Parameterized Normalization, Sum and Product (PNSP)** based on the characteristics of the static saliency map and motion saliency map. The final saliency map from the proposed fusion method PNSP for video frames is calculated as follows.

$$S = \gamma_1 S_s + \gamma_2 S_m + \gamma_3 S_s S_m \quad (5.14)$$

where S is the final saliency map for video frames; S_s is the static saliency map; S_m is the motion saliency map; γ_1 , γ_2 and γ_3 are parameters of the weighting for each component.

The weighting parameters in (5.14) are determined by the characteristics of the static saliency map and motion saliency map. A good saliency map should include the small

compact salient regions rather than the spread salient points. If the salient regions in the motion saliency map are small and compact, the motion contrast of this frame can be considered strong. In this case, we can use a large weighting parameter γ_2 to weight the motion saliency map. This means that the motion saliency map contribute much to the final saliency map. Similarly, a large weighting parameter γ_1 should be used to weight the static saliency map if the feature contrast is strong in the static saliency map. As we can see, the weighting parameter γ_3 is used to measure the importance of these regions which both static saliency map and motion saliency map detect as salient. Here we set it as: $\gamma_3 = (\gamma_1 + \gamma_2)/2$.

The spatial variance is adopted to measure the weighting of the static saliency map or motion saliency map in calculation of the final saliency map. Given a static saliency map or motion saliency map S_q ($S_q \in \{S_s, S_m\}$), Otsu's threshold method [103] is used to binarize it and the spatial variance can be calculated as follows.

$$v_q = \frac{\sum_{(i,j)} \sqrt{(i-S_{i,e})^2 + (j-S_{j,e})^2} * S_q(i,j)}{\sum_{(i,j)} S_q(i,j)} \quad (5.15)$$

$$S_{i,e} = \frac{\sum_{(i,j)} i * S_q(i,j)}{\sum_{(i,j)} S_q(i,j)} \quad (5.16)$$

$$S_{j,e} = \frac{\sum_{(i,j)} j * S_q(i,j)}{\sum_{(i,j)} S_q(i,j)} \quad (5.17)$$

where v_q is the spatial variance of saliency map S_q ; $S_q(i,j)$ is the saliency value at the location (i, j) in saliency map S_q ; $S_{i,e}$ and $S_{j,e}$ represent the spatial expectation values of the salient regions in horizontal and vertical directions, respectively.

The spatial variance of the static saliency map or motion saliency map can be calculated from (5.15) - (5.17). The spatial variance values of the saliency maps containing small and compact salient regions are smaller than those of the saliency maps containing spread salient points. Therefore, we set $\gamma_q = 1/v_q$ (γ_q is the weighting

parameter determining the weighting of the static saliency map or motion saliency map ($\gamma_q \in \{\gamma_1, \gamma_2\}$). Thus, the final saliency map for video frames can be calculated based on (5.14) - (5.17). We will present the performance of this fusion method compared with existing ones (5.11) - (5.13) in the experiment section.

5.3 Experimental Results and Analysis

In this section, we evaluate the performance of the proposed model based on a public video database [59]. This database includes 50 video-clips totaling over 25 minutes and their saccade data from 8 observers. It includes various types of video such as outdoor video in daytime and nighttime, sports video, news television broadcast, video games, etc. The human saccade data was recorded by a 240-Hz infrared-video-based eye tracker. In this study, we have performed two experiments to evaluate the performance of the proposed model: the first one is conducted to demonstrate why we have to combine the static saliency map and motion saliency map to obtain the final saliency map for video frames; while the other one is performed to compare the performance of the proposed video saliency detection model with other existing ones. All the experiments are conducted in the compressed domain. We extracted DCT coefficients and motion vectors for saliency detection for these video.

5.3.1 Evaluation Methodologies

We use the similar measure methods as the studies [59] and [113] to evaluate the performance of the proposed model. The performance of video saliency detection models is measured by comparing the response values at saccade locations and random locations in the saliency map. Here we calculate the salient value at a human saccade location as the maximum value over a circular aperture around the saccade location in the saliency

map. A high salient value at the human saccade location means the saliency detection model can predict the salient locations exactly. Similarly, the salient value at a randomly chosen location is calculated as the maximum value over a circular aperture around the randomly chosen location.

Generally, an efficient video saliency detection model would have high response values at saccadic locations and have no response at most randomly chosen locations. Here, we firstly calculate the saliency distributions at human saccade locations and random locations with 10 bins for saliency values over the saliency map. Figure 5.5 and 5.8 show the saliency distributions at human saccade locations and random locations in the saliency map from different algorithms. The x-axis represents the saliency value bins from 0 to 1 with 0.1 intervals, while the y-axis represents the number of human saccade locations or random locations with different saliency value bins. To evaluate the performance of saliency detection models, KL distance [114, 115] is used to measure the similarity between these two distributions as follows [113].

$$KL(H', R') = \frac{1}{2} (\sum_{n'} h_{n'} \log \frac{h_{n'}}{r_{n'}} + \sum_{n'} r_{n'} \log \frac{r_{n'}}{h_{n'}}) \quad (5.18)$$

where H' and R' are saliency distributions at human saccade locations and random locations with probability density functions $h_{n'}$ and $r_{n'}$, respectively; n' is the saliency value bins ($n' \in \{1, 2, 3, \dots, 10\}$).

As mentioned above, a good video saliency detection model can get high salient values at human saccade locations and low salient values (even zero) at most randomly chosen locations in the saliency map. In this case, the saliency distributions at human saccade locations and random locations are different greatly and thus the KL distance between these saliency distributions is large. Therefore, the video saliency detection model with a higher KL distance can more easily discriminate human saccadic locations from the

random locations in video frames, and this means the better performance in saliency detection for video [59].

In addition, we use ROC curve [114, 115] to evaluate the performance of the proposed model. As introduced previously, ROC curve is a graphical plot of TPR *VS.* FPR for a binary classifier system with the varied discrimination thresholds [114], as shown in Figures 5.6 and 5.9. Here we first normalize saliency values into the range $[0, 1]$ and then use thresholds from 0 to 1 with the interval 0.1 to get ROC curves for saliency detection models. The saliency distribution at human saccade locations is used as the model to test, while the saliency distribution at random locations is used as the model to discriminate. For each threshold, the TPR is calculated as the percentage of the number of human saccade locations with salient values larger than this threshold over the total number of human saccade locations; while the FPR is computed as the percentage of the number of random locations with salient values larger than this threshold over the total number of random locations. As described previously, the overall quantitative evaluation with the ROC curve is the AUC. The larger the AUC is, the better the saliency predication for the saliency detection model is for video frames.

In the next subsections, we will describe two experiments, all based on the two evaluation methods with KL distance and ROC curve.

5.3.2 Experiment I

In this section, we compare the performance of the static saliency map, the motion saliency map and the combined saliency maps to demonstrate the importance of the combination of these two saliency maps for the final saliency map. In the next experiment, we compare the performance of the proposed method with other existing ones to demonstrate the advantage of the proposed algorithm.

Figure 5.5 The saliency distributions at human saccade locations (narrow blue bars) and random locations (wide green bars) from different algorithms (NS: Normalization and Maximum; NM: Normalization and Maximum; NP: Normalization and Product; PNSP: Parameterized Normalization, Sum and Product). The x-axis and y-axis represent the predicted saliency values from different models, and the number of locations with the corresponding salient values, respectively.

Figure 5.6 The ROC curves for the static saliency map, the motion saliency map and the final saliency maps from different fusion methods.

Figure 5.5 shows the saliency distributions at human saccade locations and random locations for the static saliency map, the motion saliency map, and the final saliency maps from different fusion methods. From the first graph (Motion) in Figure 5.5, the amount of lower saliency values (0 to 0.2) at the random locations is large. This means that the saliency values at most random locations are low. Although the saliency values at most human saccade locations are higher than these at random locations, the amount of the low saliency values (0 to 0.5) at human saccade locations is still large. From the second graph (Static) in Figure 5.5, the saliency values at most human saccade locations and random locations in the static saliency map are larger than these corresponding ones from the motion saliency map. In addition, compared with the motion saliency map, the saliency distribution at human saccade locations is more different from that at random saccade locations in the static saliency map, as shown in Figure 5.6 and Table 5.1. From Figure 5.6 and Table 5.1, we can see that the KL distance and AUC values from the static

Table 5.1: Comparison results from the static saliency map, the motion saliency map and the final saliency maps from different fusion methods.

Models	Motion	Static	NS	NM	NP	PNSP
KL Distance	0.846	1.529	1.596	1.761	1.174	1.828
AUC	0.817	0.89	0.918	0.922	0.837	0.93

saliency map are larger than those from the motion saliency map. Thus, the performance of the static saliency map is better than that of the motion saliency map.

Figures 5.5 and 5.6, and Table 5.1 also present the experimental results of the final saliency maps from different fusion methods for the static saliency map and motion saliency map. Based on the statistic results in Table 5.1, the fusion methods of NS, NM and PNSP can obviously obtain better performance than the static saliency map and motion saliency map, while the NP fusion method cannot get good result. The disadvantage of NP fusion method can be demonstrated by experimental results in Figure 5.5. From this figure, we can see that the saliency distributions at human saccade locations and random saccade locations are more similar, compared with other fusion methods (NS, NM and PNSP). From Table 1, although the KL distance and AUC values from the fusion methods of NS, NM and the proposed one (PNSP) are all larger than those from the static and motion saliency map, the KL distance and AUC values from our proposed fusion method (PNSP) are larger than those from NS and NM fusion methods. This demonstrates that our proposed fusion algorithm PNSP can obtain better performance than other ones (NS, NP and NM).

In Table 5.1, we can see that the KL distance and RUC values of the final saliency map from the proposed fusion method are much larger than the static saliency map and motion saliency map. The good performance of the proposed method demonstrates the importance of the fusion method in (5.14). Additionally, we present some comparison

samples for the static saliency map, the motion saliency map and the final saliency map from the proposed fusion method in Figure 5.7. In Figure 5.7, the human saccade locations in the original images and saliency maps are marked with a circle. The saliency values at human saccade locations are calculated as the maximum value within this circle in the saliency map [59, 113]. From Figure 5.7, we can see that the saliency value at the human saccade location is the largest value in each final saliency map from the proposed fusion method. On the contrary, in the motion saliency map and static saliency map, the

Figure 5.7 Comparison samples from different saliency maps: the first column: the original images with human saccade locations marked with a 64-radius circle; the second column: the motion saliency map; the third column: the static saliency map; the final column: the final saliency map from the proposed fusion method.

Figure 5.8 The saliency distributions at human saccade locations (narrow blue bars) and random locations (wide green bars) from different models. The x-axis and y-axis represent the predicted saliency values from different models, and the number of locations with the corresponding salient values, respectively.

saliency value at human saccade locations may not be the largest value or there may be the largest values at other locations except the human saccade locations.

From Figure 5.7, we can further find that the final saliency map reflects much of the static saliency map. According to FIT, the static saliency map is influenced by feature contrast from various low-level features such as intensity, color, orientation and so on. In most cases, the motion actions are from objects in video. Thus, the motion salient regions are mostly included in the object regions and these moving objects can be also detected

Figure 5.9 The ROC curves for different video saliency detection models in comparison.

by the static saliency map, as shown in Figure 5.7. From Figure 5.7, we can see that most salient regions in the motion saliency map can be also detected by the static saliency map at the same time. Therefore, the fusion of the motion saliency map and static saliency map can be considered as enhancement of the salient regions with motion contrast in the static saliency map. This is the reason why the final saliency results in Figures 5.7 and 5.10 much reflect the static saliency map.

5.3.3 Experiment II

In this study, three existing saliency detection models are used to evaluate the performance of our proposed model: CIOFM [24], Surprise [59] and MRS [3]. Please note that we do not choose many existing image saliency detection models (except CIOFM [24]) to do the comparison, since here we mainly evaluate the performance of the proposed model in saliency detection for video. The models of Surprise [59] and MRS [3] are designed for video saliency detection.

Based on the public video database [59], the experiment results from different models

Figure 5.10 Comparison samples of saliency maps from different models. The first column: the original images with fixation data (the center of the circle is the fixation point). The second to the last column: the saliency maps from CIOFM [24], Surprise [59], MRS [3] and the proposed model.

Table 5.2: Comparison results from different video saliency detection models.

Models	CIOFM[24]	MRS[3]	Surprise[59]	Ours
KL Distance	0.402	0.529	0.593	1.828
AUC	0.720	0.771	0.782	0.93

are shown in Figures 5.8 - 5.10 and Table 5.2. Figure 5.8 shows the saliency distributions at human saccade locations and random locations from different saliency detection models. From this figure, we can see that the difference between saliency distributions at human saccade locations and random locations from our algorithm is larger than those from other algorithms. Table 5.2 shows the KL distance and AUC values for these distributions from different saliency detection models. From Table 5.2, we can see that the KL distance and AUC values of MRS [3] and Surprise [59] are larger than those of CIOFM [24]. This demonstrates that the performance of MRS [3] and Surprise [59] is better than the one [24]. From this table, the KL distance and AUC values of the proposed model is the largest and thus the performance of the proposed model is the best among these compared algorithms.

Figure 5.10 presents some comparison samples of saliency maps from different saliency detection models. In the first column of Figure 5.10, the human saccade locations are labeled by a circle. As mentioned in previous sections, the saliency value of the saccade location is calculated by the maximum value within this circle in the saliency map. From this figure, we can see that the saliency values at human saccade locations from the model CIOFM [24] are very low, some of them even approaching to zero (the first column in Figure 5.10). From the second column of Figure 5.10, we can see that the saliency values at human saccade locations are not the largest ones in the saliency map from the model Surprise [59]. The saliency values at many other locations are larger than those at human saccade locations from the results of the model Surprise [59]. Similarly,

the saliency values at human saccade locations from the model [3] are low, as shown from the fourth column in Figure 5.10. On the contrary, the saliency values at fixation points from the proposed model are almost the largest values in the saliency map. This demonstrates that the saliency map from the proposed model can predict more accurate human saccade locations compared with other existing ones.

Figure 5.11 shows some video frames where the saliency is continuously and consistently detected. From this figure, we can see that the proposed model can get the saliency maps for continuous video frames accurately.

5.4 Conclusion

In this chapter, we propose a video saliency detection model in the compressed domain based on four types of features: luminance, color, texture and motion. These features are extracted from DCT coefficients and motion vectors in video bit-stream. The static saliency map is calculated from the features of luminance, color and texture, while the

Figure 5.11 Continuous video frames and their saliency maps. First row: original images; second row: saliency maps.

motion saliency map is computed from the motion feature. A new fusion method has been designed to combine the static saliency map and motion saliency map to get the final saliency map for video frames. Experimental results based on a public database show that the proposed video saliency detection model outperforms existing ones.

It is noted that existing video saliency detection models are implemented in the uncompressed (pixel) domain (currently, there is no video saliency detection model in the compressed domain). Compared with the video saliency detection in the uncompressed domain, the proposed video saliency detection model in the compressed domain can be used more conveniently in various Internet-based multimedia applications such as video retargeting, video quality assessment, etc. Therefore, the proposed saliency detection model in the compressed domain is significant in this research area. As the next step of the work, we will explore various multimedia applications of the proposed video saliency detection model in the compressed domain.

In the next chapter, the application of visual attention models in visual object search is explored.

Chapter 6

Combination of Bottom-up and Top-down Mechanisms

Existing studies have demonstrated that the top-down mechanism plays an important role during the formation of visual attention besides the bottom-up mechanism, especially with the cases of object detection and location. In existing studies, the top-down mechanism is stimulated by various visual information such as object templates, context information, object appearance features and so on [19, 20, 35, 36, 67]. In this chapter, we put forward a new visual attention model by combining bottom-up and top-down features for salient object detection. Here, we improve an existing phase-based bottom-up saliency detection model to formulate a new multi-scale phase-based saliency detection model. Based on the differences between statistical knowledge of orientation features for salient objects and scene background, we implement the top-down mechanism through extracting orientation features for images. The algorithm of extracting top-down information in this study is different from other existing top-down visual attention models [19, 20, 35, 36, 67]. In this study, we demonstrate the usability of orientation features in building top-down visual attention models. We further explore the

application of the proposed visual attention model in visual search for salient objects in natural scenes. Experimental results show that the proposed model by combining bottom-up and top-down mechanisms can find salient man-made objects in images more efficiently compared with the bottom-up only visual attention model.

6.1 Introduction

Selective attention in the HVS allows humans to focus on the most important regions when observing visual scenes. Many computational visual attention models have been proposed to simulate the bottom-up and top-down mechanisms, as introduced in the chapter of literature review. For the bottom-up mechanism, we know that the salient regions always ‘pop out’ automatically for its different features from their surrounding regions. On the contrary, observers may focus on different salient regions with different situations when performing complex tasks. Without any task (top-down mechanism), observers will focus on these salient regions with different low-level features (such as color, intensity and so on) from their surrounding regions when looking at a natural scene. With a task in mind, observers will focus on these regions related to the task; an example is that they are asked to find cars in forest. In this situation, the orientation feature will contribute much more to find the salient regions compared with other features (such as intensity, color and so on), because a car, being a man-made object, has more orientation alignment than the surrounding regions (like trees).

In this chapter, we put forward a new visual attention model by combining bottom-up and top-down mechanisms for salient object detection. As mentioned above, to implement the top-down visual attention models, existing studies adopt target templates [36], context information [19], or other visual information to simulate the top-down

mechanism. Here, we implement the top-down mechanism through extracting orientation features of images based on the prior knowledge that the orientation features of salient man-made objects are different greatly from those of the scene background. The algorithm of extracting top-down information in this model is different from other existing models. Compared with [36], we use the statistical knowledge of orientation features instead of using target templates. Different from [19] and [20], we use the orientation features to get the top-down information rather than object appearance or visual context information. Furthermore, compared with other existing methods using machine learning algorithms [19, 20], the proposed algorithm is easier to be implemented without resorting to any machine learning technique. Experiments show that the integrated model provides more efficient visual search for salient man-made objects, compared with bottom-up only visual attention models.

6.2 Combining Bottom-up and Top-down Mechanisms

In this section, we introduce the proposed visual attention model by combining bottom-up and top-down mechanisms. For the bottom-up saliency detection models, the phase-based methods in [26, 27] have been widely used for saliency detection for images due to their simple implementation and efficiency. Here, we improve the phase-based method in [27] to achieve the bottom-up saliency map for images, while the top-down information is obtained by extracting the specific orientation features of salient objects.

Please note that other bottom-up saliency detection models including our proposed bottom-up saliency detection models in previous chapters can be also used here. The study in this chapter aims to demonstrate the influence of the top-down mechanism in

visual attention modeling, and the phase-based visual attention model [26] is adopted here for its simplicity and efficiency (the code runs fast for the used FFT in the phase-based saliency detection models). The framework of the proposed model is depicted as Figure 6.1.

Figure 6.1 The proposed visual attention model by combining bottom-up and top-down mechanisms.

6.2.1 Proposed Multi-scale Bottom-up Saliency

In this work, we improve the phase-based saliency detection model [27] toward a multi-scale phase-based visual attention model for bottom-up mechanism. As the original phase-based visual attention model firstly resizes the images into a smaller scale, it causes some information loss in images. Figure 6.2 demonstrates this defect of the saliency map in [27]. The resizing procedure will induce small objects to be lost in the final saliency map (Figure 6.2 (b)).

To overcome the problem with the existing phase-based approaches, here we propose a multi-scale phase-based visual attention model. Given an input image $I(i, j)$, we have the bottom-up saliency map $\mathcal{M}_{bu}$ as an accumulation of that for each scale k , that is:

$$\mathcal{M}_{bu} = \sum_{k=1}^{N_l} \mathcal{M}_k / N_l \quad (6.1)$$

where N_l is the number of the scales used (the first scale is the original size, and the smallest scale is 32×32); $\mathcal{M}_k$ is the saliency map for the image at scale k , which is calculated as [27].

In essence, this multi-scale scheme is capable of detecting salient objects in a scene with different sizes. Figure 6.2 (c) illustrates that this proposed scheme can remedy the abovementioned drawback of the existing phase-based approaches.

Figure 6.2 Comparison between the original phase-based model and the improved multi-scale phase-based model: (a) the original image; (b) the saliency map from single-scale phase-based model [27]; (c) the saliency map from the proposed multi-scale phase-based model.

6.2.2 Top-down Saliency

Man-made objects usually have stronger and denser orientation features than natural scenes, as shown in Figure 6.3, and many existing studies use the orientation information for image processing applications such as scene classification (e.g., [116]). Figure 6.3 depicts the orientation feature differences between vehicles and trees in natural scenes. From this figure, the regions with red remarks denote the orientation strength of the images. Here we extract those orientation strength values larger than two times of the mean of the orientation strength values in the whole image. As shown from this figure, the orientation strength and dense values of vehicles are much higher than those from the trees. Therefore, to distinguish such salient objects from the background of natural scenes, we can use the statistical knowledge of orientation features to stimulate the top-down mechanism in visual attention modeling.

Here, a set of Gabor filters are adopted to extract the orientation information for image pixels. The orientation features for the pixel at location (x, y) can be calculated as:

Figure 6.3 Difference of orientation features for vehicles and trees: images in (a) and (b) show the orientation strength for two sample vehicles; while images in (c) and (d) show the orientation strength for two sample trees.

Figure 6.4 Gabor filters with 12 different orientations.

$$G_{\lambda_t, \theta_t, \varphi_t, \sigma_t, \gamma_t, b_t}(x, y) = \exp\left(\frac{x'^2 + \gamma_t^2 y'^2}{2\sigma_t^2}\right) \cos\left(2\pi \frac{x'}{\lambda_t} + \varphi_t\right) \quad (6.2)$$

where $\lambda_t, \theta_t, \varphi_t, \gamma_t$, and b_t are wavelength, orientation, phase offset, aspect ratio and bandwidth for the Gabor filter, respectively; here, $\frac{\sigma_t}{\lambda_t} = \frac{1}{\pi} \sqrt{\frac{\ln 2}{2} \frac{2^{b_t} + 1}{2^{b_t} - 1}}$; $x' = x \cos \theta + y \sin \theta$ and $y' = -x \sin \theta + y \cos \theta$.

Let $\theta_t = \{0, \frac{\pi}{12}, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{5\pi}{12}, \frac{\pi}{2}, \frac{7\pi}{12}, \frac{2\pi}{3}, \frac{3\pi}{4}, \frac{5\pi}{6}, \frac{11\pi}{12}\}$ such that we can extract the orientation information for 12 different orientations for the image. Figure 6.4 shows these Gabor functions with the 12 different orientations. We have done some experiment for parameter choice. The experimental results show that the following parameter choice can get a good result: $\lambda_t = 8$; $\gamma_t = 0.5$; $b_t = 0.8$; $\varphi_t = \pi/2$; the size of used filters here is 15.

Based on this set of Gabor filters, we can get a vector of orientation features at location (x, y) as $\mathcal{O} = \{\mathcal{O}_i(x, y)\}$. As there are 12 orientations, the size of the vector $\mathcal{O}$ is 12. Thus, for each image pixel, there is a vector feature containing 12 orientation values. Among the orientation values, we choose the strongest value to represent the orientation feature of that image pixel as:

$$\mathcal{L}(x, y) = \max_i(|\mathcal{O}_i(x, y)|) \quad (6.3)$$

where $\max$ operator as being used in (6.3) is to select the largest absolute orientation

value among the 12 orientation values.

Since salient objects have stronger and denser orientation values than the scene background, we obtain the top-down saliency information in images as follows:

$$\mathcal{M}_{tp} = \begin{cases} \mathcal{L}(x, y), & \text{if } \mathcal{L}(x, y) > M_t * \mathcal{L}' \text{ and } \mathcal{T}(x, y) > \mathcal{T}' \\ 0, & \text{otherwise} \end{cases} \quad (6.4)$$

where $\mathcal{L}'$ is the mean value of $\mathcal{L}(x, y)$ and M_t is a threshold parameter for the orientation strength; $\mathcal{T}(x, y)$ indicates the orientation density in an region centering at location (x, y) , while $\mathcal{T}'$ is a threshold parameter for the orientation density. The orientation density $\mathcal{T}(x, y)$ in the region centering at location (x, y) is computed as the amount of strong orientation values larger than the value of $M_t * \mathcal{L}'$. The thresholds of M_t and $\mathcal{T}'$ are adopted to distinguish the salient objects (e.g. a vehicle) from the scene background (e.g. trees). The orientation strength and density values of the vehicles and trees have been shown in Figure 6.3.

6.2.3 Integration

After the bottom-up saliency and top-down saliency are calculated, the next step is to combine them to get the final saliency map for images. Firstly, normalization operation is applied to $\mathcal{M}_{tp}$ and $\mathcal{M}_{bu}$, and results in $\mathcal{M}_{tp}^{\mathcal{N}}$ and $\mathcal{M}_{bu}^{\mathcal{N}}$ respectively. And then, the bottom-up saliency and top-down saliency are compounded to acquire the final saliency map as below [117]:

$$S = \eta \mathcal{M}_{tp}^{\mathcal{N}} + (1 - \eta) \mathcal{M}_{bu}^{\mathcal{N}} + \tau \mathcal{M}_{tp}^{\mathcal{N}} \mathcal{M}_{bu}^{\mathcal{N}} \quad (6.5)$$

where η and τ are the weights for bottom-up and top-down contributions. Here, we consider the contributions of bottom-up and top-down mechanisms are the same to the final saliency map. η and τ are selected as 0.5 and 1 respectively.

Figure 6.5 Comparison results of two sample images between the integrated model and the bottom-up visual attention model: (a): the original Image 1; (b): the original Image 2; (c) the saliency map of Image 1 from the bottom-up model; (d) the saliency map of Image 2 from the bottom-up model; (e) the saliency map of Image 1 from the proposed model; (f) the saliency map of Image 2 from the proposed model.

6.3 Performance Evaluation: Visual Search

We have demonstrated some examples for the effectiveness of the proposed bottom-up and top-down modules as Figures 6.2 and 6.3. In this section, we investigate into the overall performance (as well as the proper benchmarking) of the proposed visual attention model in the application of visual search for salient man-made objects in natural scenes. As our aim is to demonstrate the influence of the top-down mechanism in visual

Figure 6.6 Comparison results of two other sample images between the integrated model and the bottom-up visual attention model: (a): the original Image 3; (b): the original Image 4; (c) the saliency map of Image 3 from the bottom-up model; (d) the saliency map of Image 4 from the bottom-up model; (e) the orientation map of Image 3; (f) the orientation map of Image 4; (g) the saliency map of Image 3 from the proposed model; (h) the saliency map of Image 4 from the proposed model.

attention modeling, we mainly compare the performance of the proposed model by

combining bottom-up and top-down mechanisms with the bottom-up only saliency detection model in visual search.

In this experiment, we explore the application of the proposed visual attention model in visual search for man-made objects based on the Search_2 dataset [118], a collection of scenes with vehicles in forested regions (some image samples can be found in Figures 6.5 and 6.6). We use the images with the resolution 760×510 which contain a tank that can be detected by human eyes in this experiment. The bottom-up visual attention model used here for comparison is the improved version of the phase-based visual attention model [27]. The comparison results of four samples are shown in Figures 6.5 and 6.6.

From Figure 6.5 (c) and (d), we can see that the saliency map achieved from the bottom-up visual attention model is poor for target search. Besides the vehicles, the saliency maps also include other objects such as trees. This similar problem exists when using other bottom-up models as well (like the models [26] and [24], and our proposed models in previous chapters). The reason is that the bottom-up visual attention model obtains the final saliency map by computing the center-surround difference, so trees and even vacant ground may become salient regions in images with complex background. From Figure 6.5 (c) and (d), the saliency maps from the bottom-up model cannot tell the exact locations of vehicles, since too many saliency regions are detected. However, when combining the bottom-up and top-down mechanisms, the saliency map achieved from the proposed integrated model is largely improved in Figure 6.5 (e) and (f). There are still some other salient regions except for the vehicles. However, compared with the saliency maps of the bottom-up model, the search results of saliency map of the proposed integrated model are much better. More experimental results of other two images are shown in Figure 6.6. We add the orientation maps of images in Figure 6.6 (e) and (f). From this figure, we can see that the vehicles in images cannot be found just by

orientation maps. The bottom-up saliency map should be combined with orientation maps to search the salient man-made vehicles in images.

Table 6.1 shows the comparison results from the bottom-up model and the proposed model. We use a window of 30×40 (similar size as a vehicle in images) to scan the saliency maps to compute the salient regions. If there are salient points in one window, then the region of this window is regarded as a salient region. From Table 6.1, we can see that the salient region of the saliency map is reduced greatly when using the proposed model that integrates both bottom-up and top-down modules, without missing vehicles. The reason is that for those salient regions (with trees, etc.) in the bottom-up model, the orientation strength and density is very small. These salient regions with small orientation strength and low orientation density are eliminated by combining the bottom-up and top-down mechanisms in the proposed visual attention model.

Table 6.1 Comparison of average saliency area for the bottom-up model and the proposed integrated model.

	The Bottom-up Model [27]	The Proposed Model
Saliency Area	59.8%	4.7%

6.4 Conclusion

In this chapter, we propose a new visual attention model by combining bottom-up and top-down mechanisms for man-made object detection and search. We improve the phase-based saliency detection model into a multi-scale version for bottom-up mechanism. Different from existing relevant approaches which use target templates, object appearance or context information as top-down information, the proposed model uses the statistical knowledge of orientation features as the top-down information. The proposed

model demonstrates that the statistical knowledge of orientation features can be used as top-down information in visual attention modeling. In the experiments, we explore the application of the proposed visual attention model in visual search for salient man-made objects. The experimental results demonstrate that the proposed visual attention model by combining bottom-up and top-down mechanisms can find salient man-made objects in images more efficiently compared with the bottom-up only visual attention model.

In the next chapter, we summarize the thesis and provide the possible future work based on the current research work.

Chapter 7

Summary and Future Work

7.1 Summary

To extract the salient regions in images/video more reasonably and accurately than existing visual attention models, this thesis has presented several novel saliency detection models to simulate the bottom-up and top-down mechanisms. In this thesis, we design the saliency detection models from the following aspects: the integration of the characteristics of the HVS (human visual sensitivity change due to foveation), the representation of the feature distribution from amplitude spectrum of QFT, the feature extraction in the compressed domain for images and video, the fusion methods for different feature maps, and the representation of top-down mechanism by the statistical knowledge of orientation features. Furthermore, we explore the applications of the proposed saliency detection models in image retargeting and visual search for man-made objects.

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

7.1.1 Integration of HVS' Characteristics in Visual Attention Modeling

Unlike most existing saliency detection models in which the characteristics of the HVS (i.e., human visual sensitivity variation with different eccentricities) have been ignored, a novel bottom-up saliency detection model has been proposed based on human visual sensitivity and amplitude spectrum in Chapter 3. Based on FIT, we build the computational model of visual attention through calculating center-surround differences between image patches. The human visual sensitivity is utilized to weight these differences for saliency detection. In addition, the amplitude spectrum of QFT is used to represent the feature distributions for image patches. In sum, the saliency value of each image patch is calculated by the two factors: the differences of QFT amplitude spectrum between this image patch and other image patches in the whole image, and the weights for these patch differences determined by human visual sensitivity.

In Subsection 3.2.1, we demonstrate how the proposed model is formulated based on human visual sensitivity and QFT amplitude spectrum of image patches. Subsection 3.2.2 presents how to use the amplitude spectrum of QFT to represent the intensity, color and orientation distributions for image patches. Then the feature contrast calculation and the use of human visual sensitivity are introduced. Here we use human visual sensitivity as weights for the feature differences between image patches for the following reason: for the saliency value of each image patch, the contribution of feature differences from the further image patches is smaller than that from the nearer image patches. This contribution can be modeled by human visual sensitivity, which decreases with the larger eccentricities.

In addition, we discuss the choice of patch size and scale in the final saliency map

generation based on the characteristics of the HVS. The size of image patches is selected within the best visual acuity according to the experimental results in psychology, as described in Subsection 3.2.4. Besides, the multi-scale operation is adopted to enhance the final saliency map. Experimental results demonstrate that the proposed saliency detection model outperforms the relevant existing ones.

7.1.2 Saliency Detection in Compressed Domain

With the fast increase of various user devices and Internet-based multimedia applications, there are more and more image/video files uploaded over Internet by users. Most of these image/video files are stored in the compressed domain such as JPEG, H.264 and MPEG4 Visual. However, existing saliency detection models are implemented in the uncompressed domain and they have to decompress the compressed image/video files into the spatial domain for feature extraction. This full decompression process is not only time-consuming but computation-consuming as well. Thus, in Chapters 4 and 5, we build the saliency detection models in the compressed domain for images and video respectively. The saliency detection models in the compressed domain obtain the saliency maps for images/video frames without decompressing bit-stream data completely. The proposed saliency detection models in the compressed domain can be used widely in various multimedia applications such as retargeting, coding, visual quality assessment and so on.

In Chapter 4, a novel saliency detection model in the compressed domain is proposed for images. The intensity, color and texture features are firstly extracted from DCT coefficients in JPEG bit-stream. Subsection 4.2.1 introduces the feature extraction process from DCT coefficients in JPEG bit-stream in detail. In Subsection 4.2.2, we give

saliency calculation process for images based on features extracted in the compressed domain. Similar to most existing saliency detection models, the proposed model calculate image saliency based on FIT. However, the features extracted from the compressed domain are novel and more effective, as demonstrated in Chapter 4. In addition, this model adopts a Gaussian model of Euclidean distances to weight the feature differences between DCT blocks. The saliency calculation method considers both local and global contrast and proves promising, as shown in the experimental part in Chapter 4. Section 4.3 gives the performance evaluation of the proposed saliency detection in the compressed domain. Compared with existing saliency detection algorithms, the proposed saliency detection model can achieve better performance on a public image database.

In Chapter 5, we extend the proposed saliency detection model in Chapter 4 to the saliency detection model in the compressed domain for video. Different from images, the compression standard for video is different greatly and thus, the feature extraction process for compressed video is different greatly from that for compressed images. Furthermore, the motion feature has to be considered in video saliency detection. In Subsection 5.2.1, the feature extraction method is given based on video bit-stream. Besides the luminance, color, and texture features, we introduce the process of motion feature extraction. Subsection 5.2.2 introduces the saliency calculation method based on extracted features in the compressed domain. The proposed saliency detection model calculates the saliency map for video frames based on feature contrast from luminance, color, texture and motion. Furthermore, a new fusion method for feature maps is designed to obtain the final saliency map for video frames. Experimental results over a public eye tracking database demonstrate the superior performance of the proposed video saliency detection model in the compressed domain.

7.1.3 Top-down Mechanism Modeling

In Chapters 3 - 5, we propose saliency detection models for image/video in the spatial/compressed domain to simulate the bottom-up mechanism. As we know, the top-down mechanism will also influence the salient region detection besides bottom-up mechanism in visual attention modeling. To explore the visual attention modeling of top-down mechanism, we propose a visual attention model by combining the bottom-up and top-down mechanisms for man-made object detection in Chapter 6.

A multi-scale phase-based saliency detection model is designed for the bottom-up mechanism for its simple implementation and efficiency in saliency detection. The saliency map from the bottom-up mechanism is calculated as in Subsection 6.2.1. For the top-down mechanism, we use the statistical knowledge of orientation features for visual attention modeling. In natural scenes, it is known that the orientation strength and density of man-made objects are different greatly from these in the scene background. We use this statistical knowledge of orientation features as the top-down information for salient object detection. The used top-down information in this model is different from existing relevant studies. In this part of the thesis, we demonstrate the effectiveness of the orientation features used as the top-down mechanism. Experimental results show that the proposed visual attention model enables much more efficient object search over bottom-up only visual attention models.

7.1.4 Applications of Saliency Detection Models

As introduced previously, saliency detection models are widely used in various multimedia processing applications. In this thesis, we also explore the applications of the proposed saliency detection models in image retargeting and visual search. We use the

proposed bottom-up saliency detection models to devise new image retargeting algorithms, as demonstrated in Chapters 3 and 4. In addition, we adopt the proposed visual attention model by combining bottom-up and top-down mechanisms in visual search for man-made objects in Chapter 6.

In Chapter 3, we use the saliency detection model in the framework of Seam Carving to demonstrate the application of the proposed saliency detection model in image retargeting. The saliency map from the proposed saliency detection model is used to measure the visual importance of image pixels for image resizing operation. The experimental results show that the performance of the proposed image retargeting algorithm is better than that from existing ones because of the better measurement of visual importance for image pixels from the proposed saliency detection model.

In Chapter 4, we devise a multi-operator image retargeting algorithm in the compressed domain based on the proposed saliency detection model in the compressed domain for images. The multi-operator operation including the image scaling and block-based seam carving is used to resizing images in the proposed retargeting algorithm. To overcome the problems with existing multi-operator image retargeting algorithms, a novel idea of texture homogeneity is defined to determine the number of removed block-based seams. Subsection 4.2.3 gives details of the proposed image resizing operation. Experimental results demonstrate the proposed multi-operator image retargeting algorithm outperforms the state-of-the-art algorithms.

Besides the image retargeting, we also explore the application of the proposed saliency detection model in visual search for man-made objects. In Chapter 6, the proposed visual attention model of bottom-up and top-down mechanisms is used to search for man-made objects (vehicles). From experimental results in Chapter 6, the search efficiency increases greatly by integrating the top-down information in bottom-up saliency detection models.

7.2 Future Work

This thesis work has proposed several computational models of visual attention for saliency detection for images/video in spatial/compressed domain. Based on the current work and tracking of the latest development in visual attention modeling, we believe there are several interesting avenues for further research as the possible extension of the thesis work.

7.2.1 Improved Visual Attention Modeling

In this thesis, we integrate the characteristics of the HVS (human visual sensitivity change due to foveation) in building the bottom-up saliency detection model for images. The weighting factor of human visual sensitivity proves promising and thus the proposed model can achieve better performance than other existing ones. The proposed model proves the usability of perceptual theories from psychology in visual attention modeling. Therefore, the cognitive theories from other disciplines such as psychology or neuroscience can be further explored to build the computational model of visual attention more accurately. To understand visual attention mechanism, researchers from different disciplines such as psychology, neuroscience, computer vision, and so on have benefited much from each other [144]. Psychologists can use the theories from neuroscience to conduct their experiments. Neuroscientists can adopt the experimental data from psychology to explain their theories. These findings in cognitive science can help research scientists in computer vision to build more reasonable computational models to simulate the mechanisms in the HVS. The proposed model based on human visual sensitivity demonstrates the effectiveness of the characteristics of the HVS in building computational models of visual attention, and thus, it inspires us to explore other better

cognitive theories to design the computational systems for visual attention.

7.2.2 Multi-Factor Top-down Mechanism Modeling

Another interesting direction is how to simulate the top-down mechanism to design computational models of visual attention. Many studies have explored the visual search in images by modeling the bottom-up visual attention mechanism [119 - 121]. However, the top-down information is rarely considered in these studies. As we know, the top-down mechanism includes many other factors such as previous experiences, emotion status, motivations, and so on. As introduced in the literature review, existing studies use target templates, context information, object appearance features or other visual information to simulate the top-down cues. In Chapter 6, we use the statistical knowledge of orientation features as the top-down information for saliency detection. There are various types of visual information which can be used to simulate the top-down mechanism for saliency detection. Then, which type of visual information is the most efficient and effective one in certain situation? This is still an open question in visual attention modeling and related to applications, and it is meaningful to further investigate how to model the top-down mechanism more efficiently and effectively. Currently, we are investigating the influence of the fixation duration in visual attention modeling from the implicit memory theory [109 - 110]. Existing studies have shown that the focus regions of attention and eye movements are guided by the recently attended regions and the implicit memory traces of context cueing [109 - 111]. These studies demonstrate that the previous attended regions will trigger attention traces utilized in the following several fixations. In addition, existing studies have demonstrated the influence of context information in saliency detection [19, 68]. Therefore, we can formulate the new top-down visual attention model for video based on the fixation duration influence, context information, and specific

features of targets as follows:

$$S = f(S_p, C_c, V_c)$$

where S represents the saliency map of the current video frame; S_p is the saliency map from the previous video frame, which is used to obtain the fixation duration influence; C_c is the context cue in the current video frame; V_c represents the feature vector of the current video frame; f is the integration function to combine these three components (fixation duration influence, context cue and current feature vector). The feature vector includes the specific features of targets (e.g. the used statistical knowledge of orientation features in Chapter 6). Of course, general features for salient region detection can be also included to integrate the bottom-up mechanism in this model.

7.2.3 Applications of Saliency Detection in Compressed Domain

In this thesis, we introduce the applications of the proposed saliency detection models in image retargeting and visual search. Actually, saliency detection models have been widely used in various multimedia processing applications, such as video coding [1, 3], scene classification [122], image/video quality assessment [123, 124], object recognition [125-129], video summarization [63, 130], image segmentation [131, 132], and so on [133-140]. In Chapters 4 and 5, we propose two saliency detection models in the compressed domain for images and video respectively. Compared with saliency detection models in the uncompressed domain, the saliency detection model in the compressed domain can obtain saliency map for images/video frames without decompressing image/video files completely. Chapter 4 has investigated the application of the proposed model in image retargeting in the compressed domain. The image retargeting algorithm

in the compressed domain can be used to transfer different image/video bit-stream to users with the devices of different sizes or aspect ratios. This can save much network bandwidth, compared with the traditional network transmission method which transfers the whole image/video bit-stream to users. Besides the image retargeting, we can further explore other applications of the proposed saliency detection models in the compressed domain. Transcoding is one of the potential applications of the proposed saliency detection models in the compressed domain. With the proposed saliency detection models in the compressed domain, the saliency map can be obtained in the compressed domain without full decompression of compressed images and video. Transcoding can use the saliency map in the compressed domain to encode data: use more bits to encode the ROI and fewer bits to encode the background of images/video frames. Compared with traditional transcoding methods, this new transcoding scheme can finish these operations in the compressed domain without decompressing the compressed data completely. We can also explore other applications such as visual quality assessment in the compressed domain.

Appendix

In this section, we present an experiment to demonstrate how we choose the parameter σ for the Gaussian model used in Formula (4.5) of Chapter 4 (the parameter choice of the Gaussian model in Formula (5.7) of Chapter 5 is similar). For each image patch, the saliency value is calculated by weighted center-surround differences between this image patch and others in the whole input image. The weighting for center-surround differences is determined by a Gaussian model of the spatial distance between image patches. Therefore, the parameter σ for the Gaussian model determines the weighting for local contrast and global contrast for saliency map. With a smaller σ , the weighting for local contrast is larger and thus the local contrast will contribute more to the final saliency map. In this case, the saliency map mainly detects contours of salient objects. On the contrary, the global contrast would contribute more to the saliency map with a larger σ . Figure 7.1 shows some comparison samples from saliency detection models with different values for parameter σ . From this figure, we can see that the saliency maps with $\sigma = 1$ for the Gaussian model mainly detect contours of salient objects in images, since local contrast contributes much more to saliency map compared with global contrast. From Figure 7.1, we can see that the saliency map is more accurate with larger values of σ than that with lower values of σ compared with ground truth. However, the performance will decrease with larger weighting of global contrast than a certain

Table 7.1 The AUC of saliency results from different values for parameter σ .

Parameter σ	= 1	= 5	= 10	= 20	= 40	= 80
AUC	0.8397	0.9060	0.9317	0.9414	0.9389	0.9357

Figure 7.1 The comparison samples from different parameters for the Gaussian model in Formula (4.5). First row: original images; Second to Seven row: saliency results from Gaussian models with parameter $\sigma = 1$, $\sigma = 5$, $\sigma = 10$, $\sigma = 20$, $\sigma = 40$, and $\sigma = 80$, respectively; Last row: ground truth.

Figure 7.2 The ROC curves of saliency results from different values for parameter σ .

threshold, as shown in Figure 7.2 and Table 7.1. Therefore, we should consider the weighting of local and global contrast for the parameter choice for the Gaussian model. Based on this experiment, we choose the parameter σ as 20 for the Gaussian models used in Formula (4.5) and (5.7).

References

- [1] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304-1318, 2004.
- [2] L. Q. Chen, X. Xie, X. Fan, W. Y. Ma, H. J. Zhang, and H. Q. Zhou, "A visual attention model for adapting images on small displays," *Multimedia Systems*, vol. 9, no. 4, pp. 353-364, 2003.
- [3] C. Guo and L. Zhang, "A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185-198, 2010.
- [4] H. Liu, S. Jiang, Q. Huang, C. Xu, and W. Gao, "Region-based visual attention analysis with its application in image browsing on small displays," *ACM International Conference on Multimedia*, 2007.
- [5] E. El-Kwae, H. Xu, and M. R. Kabuka, "Content-based retrieval in picture archiving and communication systems," *Journal of Digital Imaging*, vol. 13, no. 2, pp. 70-81, 2000.
- [6] W. James, *The Principles of Psychology*. Cambridge, MA: Harvard University Press, 1890.
- [7] U. Neisser, *Cognitive Psychology*. New York: Appleton, 1967.
- [8] W. Schneider, and R. M. Shiffrin, "Controlled and automatic human information processing: I. Detection, search, and attention," *Psychological Review*, vol. 84, no. 1, pp. 1-66, 1977.
- [9] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry, V. Balasubramanian, and P.

- Sterling, "How much the eye tells the brain," *Current Biology*, vol. 16, no. 14, pp. 1428-1434, 2006.
- [10] J. K. Tsotsos. "Analyzing vision at the complexity level," *Behavioral and Brain Sciences*, vol. 13, no. 3, pp. 423-469, 1990
- [11] H. Pashler, *The psychology of attention*, Cambridge, MA: MIT Press, 1997.
- [12] H. Pashler, (Ed.) *Attention*, Hove, England: Psychology Press, 1988.
- [13] Anderson, and R. John, *Cognitive psychology and its implications*, Worth Publishers, 2004.
- [14] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97-136. 1980.
- [15] J. M. Wolfe, S. J. Butcher and M. Hyle, "Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, no. 2, pp. 483-502, 2003.
- [16] L. Itti, "*Models of bottom-up and top-down visual attention*," Ph.D. dissertation, Dept. Computat. Neur. Syst., California Inst. Technol., Pasadena, Jan. 2000.
- [17] J. Braun, and D. Sagi, "Vision outside the focus of attention," *Perception and Psychophysics*, vol. 48, no. 1, pp. 45-48, 1990.
- [18] 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, 2005.
- [19] A. Torralba, A. Oliva, M. S. Castelhana and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," *Psychological Review*, vol. 113, no. 4, pp. 766-786, 2006.
- [20] C. Kanan, M. Tong, L. Zhang and G. Cottrell, "SUN: Top-down saliency using natural statistics," *Visual Cognition*, vol. 17, no. 6, pp. 979-1003, 2009.

- [21] C. W. Tang, "Spatiotemporal visual considerations for video coding," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 231-238, 2007.
- [22] W. Wang, Y. Song and A. Zhang, "Semantics-based image retrieval by region saliency," *International Conference on Image and Video Retrieval*, LNCS, Vol. 2383, pp. 29-37, 2002.
- [23] J. S. Hare, "Saliency for Image Description and Retrieval," *PhD thesis*, University of Southampton, 2005.
- [24] 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, 1998.
- [25] J. Harel, C. Koch and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems*, 2006.
- [26] X. Hou and L. Zhang, "Saliency Detection: A spectral residual approach," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.
- [27] C. Guo, Q. Ma and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.
- [28] N. D. Bruce and J. K. Tsotsos, "Saliency based on information maximization," *Advances in Neural Information Processing Systems*, vo. 18, pp. 155-162, 2006.
- [29] L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194-203, 2001.
- [30] T. Liu, J. Sun, N. Zheng, X. Tang and H. Y. Shum, "Learning to detect a salient object," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.
- [31] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," *IEEE International Conference on Computer Vision*, 2007.
- [32] V. Gopalakrishnan, Y. Hu and D. Rajan, "Salient region detection by modeling

- distributions of color and orientation,” *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 892-905, 2009.
- [33] R. Valenti, N. Sebe, and T. Gevers, "Image saliency by isocentric curvedness and color", *IEEE International Conference on Computer Vision*, 2009.
 - [34] Y. F. Ma and H. J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," *ACM International Conference on Multimedia*, 2003.
 - [35] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimal object detection," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.
 - [36] R. Rao, G. Zelinsky, M. Hayhoe and D. Ballard, "Eye movements in iconic visual search," *Vision Research*, vol. 42, no. 11, pp. 1447-1463, 2002.
 - [37] J. Wolfe, K. R. Cave and S. L. Franzel, "Guided Search: An alternative to the feature integration model for visual search," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 3, pp. 419-433, 1989.
 - [38] J. Wolfe, "Guided Search 2.0: A revised model of guided search," *Psychonomic Bulletin and Review*, vol. 1, no. 2, pp. 202-238, 1994.
 - [39] A. Treisman, and S. Gormican, "Feature analysis in early vision: evidence from search asymmetries," *Psychology Review*, vol. 95, pp. 15-48, 1988.
 - [40] C. Koch, and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219-227, 1985.
 - [41] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395-1407, 2006
 - [42] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model the bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802-817, 2006.
 - [43] N. Murray, M. Vanrell, X. Otazu, and C. Alejandro Parraga, "Saliency estimation using a non-parametric low-level vision model," *IEEE International Conference on*

Computer Vision and Pattern Recognition, 2011.

- [44] D. Gao, S. Han and N. Vasconcelos, “Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, 2009.
- [45] X. Hou and L. Zhang, “Dynamic visual attention: searching for coding length increments,” *Advances in Neural Information Processing Systems*, 2008.
- [46] H. J. Seo and P. Milanfar, “static and space-time visual saliency detection by self-resemblance,” *Journal of Vision*, vol.9, no. 12, pp. 1-27, 2009.
- [47] W. Wang, Y. Wang, Q. Huang, and W. Gao, “Measuring visual saliency by site entropy rate,” *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010
- [48] R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, “Frequency-tuned salient region detection,” *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.
- [49] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.
- [50] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” *IEEE International Conference on Computer Vision*, 2009.
- [51] A. V. Oppenheim and J. S. Lim, “The importance of phase in signals,” *Proceedings of the IEEE*, vol. 69, pp. 529-541, 1981.
- [52] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.
- [53] L. N. Piotrowski and F. W. Campbell, “A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase,” *Perception*, vol. 11, pp. 337-346, 1982.
- [54] A. Oliva and A. Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelop,” *International Journal of Computer Vision*, vol. 42, pp. 145-

175, 2001.

- [55] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of Optical Society of America*, vol. 4, no. 12, pp. 2379-2394, 1987.
- [56] S. L. Franconeri, and D. J. Simons, "Moving and looming stimuli capture attention," *Perception & Psychophysics*, vol. 65, pp. 999-1010, 2003.
- [57] R. A. Abrams, and S. E. Christ, "Motion onset captures attention," *Psychological Science*, vol. 14, pp. 427– 432, 2003.
- [58] S. Yantis, and J. Jonides. "Abrupt visual onsets and selective attention: Evidence from visual search," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, pp. 601– 621, 1984.
- [59] L. Itti, and P. Baldi, "Bayesian surprise attracts human attention," *Advances in Neural Information Processing Systems*, 2006.
- [60] Y. Zhai, and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," *ACM international conference on multimedia*, 2006.
- [61] O. Le Meur, P. Le Callet and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483-2498, 2007.
- [62] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, 2010.
- [63] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A Generic framework of user attention model and its application in video summarization," *IEEE transactions on multimedia*, vol. 7, no. 5, 2005.
- [64] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guerin-Dugue, "Modeling spatio-temporal saliency to predict gaze direction for short videos," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 231-243, 2009.
- [65] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual

- saliency estimation in video,” *International Journal of Computer Vision*, vol. 90, no. 2, pp. 150-165, 2010.
- [66] W. Einhäuser, U. Rutishauser, and C. Koch, “Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli,” *Journal of Vision*, vol. 8, no. 2, pp. 1-19, 2008.
 - [67] M. Neider and G. Zelinsky, “Scene context guides eye movements during visual search,” *Vision Research*, vol. 46, pp. 614-621, 2006.
 - [68] A. Oliva, A. Torralba, M.S. Castelhana, and J. M. Henderson, “Top-down control of visual attention in object detection,” *IEEE International Conference on Image Processing*, 2003.
 - [69] R. J. Peters and L. Itti, “Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention,” *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.
 - [70] S. Avidan, and A. Shamir, “Seam carving for content-aware image resizing,” *ACM Transactions on Graphics*, vol. 26, no. 3, 2007.
 - [71] M. Rubinstein, A. Shamir, and S. Avidan, “Improved seam carving for video retargeting,” *ACM Transactions on Graphics*, 2008.
 - [72] R. Achanta, and S. Susstrunk, “Saliency detection for content-aware image resizing,” *IEEE International Conference on Image Processing*, 2009.
 - [73] L. Wolf, M. Guttman, and D. Cohen-OR, “Non-homogeneous content-driven video retargeting,” *IEEE International Conference on Computer Vision*, 2007.
 - [74] T. Ren, Y. Liu, and G. Wu, “Image retargeting based on global energy optimization,” *IEEE International Conference on Multimedia and Expo*, 2009.
 - [75] Y. Jin, L. Liu, and Q. Wu, “Nonhomogeneous scaling optimization for realtime image resizing,” *The Visual Computer*, vol. 26, no. 6-8, pp. 769-778, 2010.
 - [76] Y. Guo, F. Liu, J. Shi, Z. Zhou, and M. Gleicher, “Image retargeting using Mesh Parametrization,” *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 856-867,

2009.

- [77] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Discontinuous seam-carving for video retargeting," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.
- [78] M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retargeting," *ACM Transaction on Graphics*, vol. 28, no. 3, pp. 1-11, 2009.
- [79] W. Dong, N. Zhou, J. C. Paul, and X. Zhang, "Optimized image resizing using seam carving and scaling," *ACM Transaction on Graphics*, vol. 28, no. 5, pp. 1-10, 2009.
- [80] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," *ACM SIGGRAPH Asia*, 2010.
- [81] D. Vaquero, M. Turk, K. Pulli, M. Tico, and N. Gelfand, "A survey of image retargeting techniques," *SPIE Applications of Digital Image Processing XXXIII*, vol. 7798, 2010.
- [82] B. A. Wandell, *Foundations of Vision*, Sinauer Associates, 1995.
- [83] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated Just-Noticeable-Distortion model," *IEEE Transactions Circuits and Systems for Video Technology*, vol. 20, no. 6, pp. 806-819, 2010.
- [84] T. Ell and S. Sangwin, "Hypercomplex fourier transforms of color images," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 22-35, 2007.
- [85] W. S. Geisler and J. S. Perry, "A real-time foveated multi-solution system for low-bandwidth video communication," *In Proceedings of SPIE*, vol. 3299, pp. 294-305, Jul. 1998.
- [86] S. Engel, X. Zhang and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, no. 6, pp. 68-71, 1997.
- [87] ITU, Methodology for the subjective assessment of the quality of television pictures, Geneva, Switzerland, ITU-R BT.500-11, 2002.

- [88] M. A. Just and P. A. Carpenter, *The psychology of reading and language comprehension*, Newton, MA: Allyn & Bacon, 1987.
- [89] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson, "Overcomplete Steerable Pyramid Filters and Rotation Invariance," *IEEE International Conference on Computer Vision and Pattern Recognition*, 1994.
- [90] Y.-S. Wang, C.-L. Tai, O. Sorkin, and T.-Y. Lee, "Optimized scale-and-stretch for image resizing," *ACM Transactions on Graphics*, 2008.
- [91] G. K. Wallace, "The JPEG still picture compression standard," *Communications of the ACM*, vol. 34, pp. 30-44, 1991.
- [92] Y. Zhong, and A. K. Jain, "Object localization using color, texture and shape," *Pattern Recognition*, vol. 33, pp. 671-684, 2000.
- [93] J. Wei, "Color object indexing and retrieval in digital libraries," *IEEE Transactions on Image Processing*, vol. 11, no. 8, pp. 912-922, 2002.
- [94] C. Theoharatos, V. K. Pothos, N. A. Laskaris, G. Economou, and S. Fotopoulos, "Multivariate image similarity in the compressed domain using statistical graph matching," *Pattern Recognition*, vol. 39, pp. 1892-1904, 2006.
- [95] JTC1/SC29/WG1: 10918-1: Information technology - Digital compression and coding of continuous-tone still images - requirements and guidelines. International standard, ISO/IEC, 1994.
- [96] Y. L. Huang, and R. F. Chang, "Texture features for DCT-coded image retrieval and classification," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, pp. 3013-3016, 1999.
- [97] Y. Li, X. Chen, X. Fu and S. Belkasim, "Multi-Level Discrete Cosine Transform for content-Based image retrieval by Support Vector Machines," *IEEE International Conference on Image Processing*, 2007.
- [98] H. Li, G. Liu, and Z. Zhang, "A new texture generation method based on pseudo-DCT coefficients," *IEEE Transactions on Image Processing*, vol. 15, pp. 1300-1312,

2006.

- [99] R. T. Rockafellar, and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, 2005.
- [100] L. Macaire, N. Vandenbroucke, and J. G. Postaire, "Color image segmentation by analysis of subset connectedness and color homogeneity properties," *Computer Vision and Image Understanding*, vol. 102, no. 1, pp. 105-116, 2006.
- [101] K. Hirakawa and T. W. Parks, "Adaptive homogeneity-directed demosaicing algorithm," *IEEE Transaction on Image Processing*, vol. 14, no. 3, pp. 360-369, 2005.
- [102] H.-D. Cheng and Y. Sun, "A hierarchical approach to color image segmentation using homogeneity," *IEEE Transaction on Image Processing*, vol. 9, no. 12, pp. 2071-2082, 2000.
- [103] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Transactions on System, Man and Cybernetics*, vol. 9, no. 1, pp. 146-165, 1979.
- [104] O. Le Meur, A. Ninassi, P. Le Callet and D. Barba, "Do video coding impairments disturb the visual attention deployment?" *Signal Processing: Image Communication*, vol. 25, pp. 597-609, 2010.
- [105] Y. Pritch, E. Kav-Venaki, and S. Peleg, "Shift-map image editing," *IEEE International Conference on Computer Vision*, 2009.
- [106] ISO/IEC JTC1. *Information technology - coding of audio-visual objects Part2: visual* ISO/IEC 14492-2, (MPEG-4 Visual), Version 1: Apr. 1999, Version 2: Feb. 2000, Version 3: May 2004.
- [107] T. Ebrahimi and C. Home, "MPEG-4 natural video coding-An overview," *Signal Processing: Image Communication*, vol. 15, pp. 4, pp. 365-385, 2000.
- [108] R. Talluri, "Error-resilient video coding in the ISO MPEG-4 standard," *IEEE Communications Magazine*, vol. 36, no. 6, pp. 112-119, 1998.
- [109] M. M. Chun, and K. Nakayama, "On the functional role of implicit visual memory for the adaptive deployment of attention across scenes," *Visual Cognition*, vol. 7, no.

- 1, pp. 65-81, 2000.
- [110] M. F. Land, and S. Furneaux, "The knowledge base of the oculomotor system," *Philosophical Transactions of the Royal Society of London: Series B. Biological Sciences*, vol. 352, no. 1358, pp. 1231-1239, 1997.
 - [111] R. A. Rensink, "Seeing, sensing, and scrutinizing," *Vision Research*, vol. 40, pp. 1469-1487, 2000.
 - [112] C. Chamaret, J. C. Chevet, and O. Le Meur, "Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies," *International Conference on Image Processing*. 2010.
 - [113] R. J. Peters, and L. Itti, "Applying computational tools to predict gaze direction in interactive visual environments," *ACM Transactions on Applied Perception*, vol. 5, 2008.
 - [114] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.
 - [115] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79-86, 1951.
 - [116] A. Torralba and A. Oliva, "Statistics of natural image categories," *Network: Computation in Neural Systems*, vol. 14, pp. 391-412, 2003.
 - [117] O. Le Meur, P. Le Callet, D. Barba and D. Thoreau, "A spatiotemporal model of the selective human visual attention," *IEEE International Conference on Image Processing*, 2005.
 - [118] A. Toet, P. Bijl, F. L. Kooi and J. M. Valetton, "A high-resolution image dataset for testing search and detection models," *Technical Report TNO-NM-98-A020*, TNO Human Factors research Institute, Soesterberg, The Netherlands, 1998.
 - [119] R. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, pp. 2397-2416, 2005.
 - [120] J. Duncan, and G. W. Humphreys, "Visual search and stimulus similarity," *Psychological Review*, vol. 96, pp. 433-458, 1989

- [121] J. M. Wolfe, "Visual search," In H. Pashler (Ed.), *Attention*, Hove, England: Psychology Press, 1998.
- [122] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300-312, 2007.
- [123] Q. Ma, L. Zhang, and B. Wang, "New strategy for image and video quality assessment," *Journal of Electronic Imaging*, vol. 19, 2010.
- [124] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric," *IEEE International Conference on Image Processing*, 2007.
- [125] A. Salah, E. Alpaydin, and L. Akrun, "A selective attention based method for visual pattern recognition with application to handwritten digit recognition and face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 420-425, 2002.
- [126] S. Frintrop, "VOCUS: A visual attention system for object detection and goal-directed search," *Lecture Notes in Computer Science*, Vol. 3899, 2006.
- [127] S. Han and N. Vasconcelos, "Biologically plausible saliency mechanisms improve feed-forward object recognition," *Vision Research*, vol. 50, no. 22, pp. 2295-2307, 2010.
- [128] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," *Advances in Neural Information Processing Systems*, 2004.
- [129] L. Elazary and L. Itti, "A bayesian model for efficient visual search and recognition," *Vision Research*, vol. 50, no. 14, pp. 1338-1352, 2010.
- [130] S. Marat, M. Guironnet, and D. Pellerin, "Video summarization using a visual attention model," *European Signal Processing Conference*, 2007.
- [131] A. K. Mishra and Y. Aloimonos, "Active segmentation," *International Journal of Humanoid Robotics*, vol. 6, pp. 361-386, 2009.

- [132] A. Maki, P. Nordlund, and J. O. Eklundh, "Attentional scene segmentation: integrating depth and motion," *Computer Vision Image Understanding*, vol. 78, no. 3, pp. 351-373, 2000.
- [133] S. Frintrop and P. Jensfelt, "Attentional landmarks and active gaze control for visual SLAM," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1054-1065, 2008.
- [134] D. Decarlo and A. Santella, "Stylization and abstraction of photographs," *ACM Transactions on Graphics*, vol. 21, no. 3, 2002.
- [135] N. Sadaka and L. J. Karam, "Efficient perceptual attentive super-resolution," *IEEE International Conference on Image Processing*, 2009.
- [136] G. Boccignone, A. Chianese, V. Moscato, and A. Picariello, "Foveated shot detection for video segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 3, pp. 365-377, 2005.
- [137] H. Li and K.N. Ngan, "Saliency model-based face segmentation and tracking in head-and-shoulder video sequences," *Journal of Visual Communication and Image Representation*, vol. 19, pp. 320-333, 2008.
- [138] C. Breazeal, "A context-dependent attention system for a social robot," *International Joint Conference on Artificial Intelligence*, 1999.
- [139] R. Rosenholtz, A. Dorai, and R. Freeman, "Do Predictions of Visual Perception Aid Design?" *ACM Transactions on Applied Perception*, vol. 8, no. 2, 2011.
- [140] N. Ouerhani, A. Bur, and H. Hugli, "Visual Attention-based Robot Self-localization," *European Conference on Mobile Robots*, 2005.
- [141] A. Borji, and L. Itti, "State-of-the-art in Visual Attention Modeling," *Transaction on Pattern Analysis and Machine Intelligence*, in Press.
- [142] R. Desimone, and J. Duncan, "Neural mechanisms of selective visual attention," *Annual Reviews of Neuroscience*, vol. 18, pp. 193-222, 1995.
- [143] H. E. Egeth and S. Yantis, "Visual attention: control, representation, and time course," *Annual Review of Psychology*, vol. 48, 1997.

- [144] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: a survey," *ACM Transactions on Applied Perception*, vol.7, no. 1. 2010.
- [145] M. Zhang, L. Zhang, Y. Sun, L. Feng, and W. Ma, "Auto cropping for digital photographs," *IEEE International Conference on Multimedia and Expo*, 2005.
- [146] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," *ACM International Conference on Multimedia*, 2009.
- [147] I. S. Amrutha, S. S. Shylaja, S. Natarajan, and K. N. Balasubramanya Murthy, "A smart automatic thumbnail cropping based on attention driven regions of interest extraction," *International Conference on Interaction Sciences*, 2009.
- [148] T. S. Cho, M. Butman, S. Avidan, and W. T. Freeman, "The patch transform and its applications to image editing," *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [149] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, 2009.
- [150] H. Y. Tong and A. N. Venetsanopoulos, "A perceptual model for JPEG applications based on block classification, texture masking, and luminance masking," *IEEE International Conference on Image Processing*, 1998.
- [151] 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, 2006.
- [152] X. Feng, T. Liu, D. Yang and Y. Wang, "Saliency based objective quality assessment of decoded video affected by packet losses", *IEEE International Conference on Image Processing*, 2008.
- [153] J. You, A. Perkis, M. Hannuksela, and M. Gabbouj, "Visual quality assessment based on human attention analysis", *ACM International Conference on Multimedia*,

2009.

- [154] J. You, J. Korhonen, A. Perkis, and T. Ebrahimi, “Balancing attended and global stimuli in perceived video quality assessment,” *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp.1269-1285, 2011.
- [155] J. You, U. Reiter, M.M. Hannuksela, M. Gabbouj, and A. Perkis, “Perceptual-based objective quality metrics for audio-visual services – A survey,” *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 482–501, 2010.

Publications

Journal Papers

- **Yuming Fang**, Weisi Lin, Bu-sung Lee, Chiew Tong Lau, Zhenzhong Chen, and Chia-Wen Lin, “Bottom-up Saliency Detection Model Based on Human Visual Sensitivity and Amplitude Spectrum,” *IEEE Transactions on Multimedia*, 14(1):187-198, 2012.
(Chapter 3 in this thesis is based upon this paper)
- **Yuming Fang**, Zhenzhong Chen, Weisi Lin, and Chia-Wen Lin, “Saliency Detection in the Compressed Domain for Adaptive Image Retargeting,” *IEEE Transactions on Image Processing*, 21(9): 3888 – 3901, 2012.
(Chapter 4 in this thesis is based upon this paper)
- **Yuming Fang**, Weisi Lin, Zhenzhong Chen, Chia-Ming Tsai, and Chia-Wen Lin, “A Video Saliency Detection Model in the Compressed Domain,” *IEEE Transactions on Circuits and Systems for Video Technology*, Under Review.
(Chapter 5 in this thesis is based upon this paper)
- Nevrez İmamoğlu, Weisi Lin, and **Yuming Fang**, “A Saliency Detection Model Using Low-Level Features Based on Wavelet Transform,” *IEEE Transactions on Multimedia*, in Press.

Conference Papers

- **Yuming Fang**, Weisi Lin, Zhenzhong Chen, Chia-Ming Tsai, and Chia-Wen Lin, “A Video Saliency Detection Model in the Compressed Domain,” *ACM International Conference on Multimedia (ACM MM)*, 2012.
- **Yuming Fang**, Zhenzhong Chen, Weisi Lin, Chia-Wen Lin, “Saliency-based Image Retargeting in the Compressed Domain,” *ACM International Conference on Multimedia (ACM MM)*, 2011.
- **Yuming Fang**, Weisi Lin, Chiew Tong Lau, and Bu-sung Lee, “A Visual Attention Model Combining Top-down and Bottom-up Mechanisms for Salient Object Detection,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.
(Chapter 6 in this thesis is based upon this paper)
- **Yuming Fang**, Weisi Lin, Bu-sung Lee, and Chiew Tong Lau, “Bottom-up Saliency Detection Model Based on Amplitude Spectrum,” *International Conference on MultiMedia Modeling (MMM)*, 2011.
- **Yuming Fang**, Wei Xiong, Weisi Lin and Zhenzhong Chen, “Unsupervised Malaria Parasite Detection Based on Phase Spectrum,” *International Conference of the IEEE Engineering in Medicine and Bilolgy Society (EMBC)*, 2011.
- **Yuming Fang**, Zhenzhong Chen, Weisi Lin, Chia-Wen Lin and Chia-Ming Tsai, “Image Retargeting Based on the Sensitivity-Tuned Visual Significance Map,” *Visual Communications and Image Processing (VCIP)*, 2011.