Faculty of Engineering and Information Technology
University of Technology Sydney

Visual Saliency Prediction For Stereoscopic Image

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Hao Cheng

July 2018

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

Production Note:
Signature removed prior to publication.

Acknowledgments

Foremost, I would like to express my gratitude to my chief supervisor, Assoc. Prof. Jian Zhang, my co-supervisor Assoc. Prof. Qiang Wu at the University of Technology Sydney, and my supervisor Prof. Ping An at Shanghai University. I am extremely grateful for all the advice and guidance so unselfishly given to me over the last three and half years by these three distinguished academics. This research would not have been possible without their encouragement, continuous support and insight, especially during my leave of absence from the University.

Meanwhile, I am grateful to the doctors and nurses of Royal Prince Alfred Hospital. Without their help and treatment, I could not have recovered from myelitis. I would also like to thank the physiotherapists at MetroRehab Hospital for helping me to stand up again during physical therapy. My sincere appreciation and gratitude go to the people who helped me in my leave of absence for all their freely-given encouragement and care.

The wonderful support and assistance provided by many people during this research are very much appreciated by me and my family. I am very grateful for the help that I received from all of these people, but there are some special individuals who deserve thanks by name.

A very sincere thank you is definitely owed to my loving wife, Yaping Wang. Thank you for taking care of me when I was receiving treatment for six months. A very special thanks to my little son Luojia Cheng for bringing me a lot of happiness. Special thanks are also owed to my mother Hua Shao for all her tremendous assistance and unfailing support.

I thank my fellow labmates in the Big Data Technologies Centre (GBDTC) and the Advanced Analytics Institute: Shangrong Huang and Yucheng Wang for their invaluable advice and insightful discussions throughout my research, and for all the help during my leave of absence due to my illness. I would also like to thank my colleagues: Renhua Song, Jing Ren, Wenbo Wang, Dongyan Guo, Ying Cui, Jinsong Xu, Xiaoshui Huang and many more, for their selfless support over the course of my PhD candidature and for all the fun that we have shared over the last three and half years. Finally, to anyone whom I have not mentioned, please forgive me. I can most definitely assure you that you have occupied a unique and special place in my thoughts. Thank you!

Hao Cheng
January 2018 @ UTS

Contents

List of Figures

List of Tables

List of Publications

Papers Published

- **Hao Cheng**, Jian Zhang, Qiang Wu, and Ping An (2017), Stereoscopic Saliency Detection Based on Stereo Contrast And Stereo Focus. The EURASIP Journal on Image and Video Processing. Volumn 2017, pp.32-61.

- **Hao Cheng**, Jian Zhang, Ping An, Zhi Liu (2015), A Novel Saliency Model for Stereoscopic Images. *in* 'Proceedings of the Internetional Conference on Digital Image Computing: Techniques and Applications (**DICTA2015**)', pp.1-7.

Papers to be Submitted/Under Review

- **Hao Cheng**, Jian Zhang, Qiang Wu, and Ping An (2016) A Computational Model for Stereoscopic 3D Visual Saliency Based on Surrounding Region, Contrast and Interest Point, to be submit the IEEE Transactions on Multimedia.

Abstract

Saliency prediction is considered to be key to attentional processing. Attention improves learning and survival by compelling creatures to focus their limited cognitive resources and perceptive abilities on the most interesting region of the available sensory data. Computational models for saliency prediction are widely used in various fields of computer vision, such as object detection, scene recognition, and robot vision. In recent years, several comprehensive and well-performing models have been developed. However, these models are only suitable for 2D content. With the rapid development of 3D imaging technology, an increasing number of applications are emerging that rely on 3D images and video. In turn, demand for computational saliency models that can handle 3D content is growing. Compared to the significant progress in 2D saliency research, studies that consider depth factor as part of stereoscopic saliency analysis are rather limited. Thus, the role depth factor in stereoscopic saliency analysis is still relatively unexplored.

The aim of this thesis is to fill this gap in the literature by exploring the role of depth factors in three aspects of stereoscopic saliency: how depth factors might be used to leverage stereoscopic saliency detection; how to build a stereoscopic saliency model based on the mechanisms of human stereoscopic vision; and how to implement a stereoscopic saliency model that can adjust to the particular aspect of human stereoscopic vision reflected in specific 3D content. To meet these three aims, this thesis includes three distinct computation models for stereoscopic saliency prediction based on the past and present outcomes of my research. The contributions of the thesis are as

follows:

Chapter 3 presents a preliminary saliency model for stereoscopic images. This model exploits depth information and treats the depth factor of an image as a weight to leverage saliency analysis. First, low-level features from the color and depth maps are extracted. Then, to extract the structural information from the depth map, the surrounding Boolean-based map is computed as a weight to enhance the low-level features. Lastly, a stereoscopic center prior enhancement based on the saliency probability distribution in the depth map is used to determine the final saliency.

The model presented in Chapter 4 predicts stereoscopic visual saliency using stereo contrast and stereo focus. The stereo contrast submodel measures stereo saliency based on color, depth contrast, and the pop-out effect. The stereo focus submodel measures the degree of focus based on monocular vision and comfort zones. Multi-scale fusion is then used to generate a map for each of the submodels, and a Bayesian integration scheme combines both maps into a stereo saliency map.

However, the stereoscopic saliency model presented in Chapter 4 does not explain all the phenomena in stereoscopic content. So, to improve the models robustness, Chapter 5 includes a computational model for stereoscopic 3D visual saliency with three submodels based on the three mechanisms of the human vision system: the pop-out effect, comfort zones, and the background effect. Each mechanism provides useful cues for stereoscopic saliency analysis depending on the nature of the stereoscopic content. Hence, the model in Chapter 5 incorporates a selection strategy to accurately determine which submodel should be used to process an image. The approach is implemented within a purpose-built, multi-feature analysis framework that assesses three features: surrounding region, color and depth contrast, and points of interest.

All three models were verified through experiments with two eye-tracking databases. Each outperforms the state-of-the-art saliency models.

Chapter 1

Introduction

1.1 Background

In computer science, digital image processing is the use of computer algorithms to perform image processing on digital images. As a subcategory or field of digital signal processing, digital image processing has many advantages over analog image processing. It allows a much wider range of algorithms to be applied to the input data and can avoid problems such as the build-up of noise and signal distortion during processing. Since images are defined over two dimensions (perhaps more) digital image processing may be modeled in the form of multidimensional systems (Solomon & Breckon 2011).

With the development of the 3D technology, stereo image processing becomes the new branch in the computer vision. Computer stereo vision is the extraction of 3D information from digital images, such as those obtained by a CCD camera. By comparing information about a scene from two vantage points, 3D information can be extracted by examining the relative positions of objects in the two panels. This is similar to the biological process Stereopsis. Stereoscopic images are often stored as MPO (multi-picture object) files. It is widely used in different kinds of fields (Shapiro 1992).

Saliency detection for the stereoscopic images can significantly improve the performance of the other computer vision algorithm. Recently, it be-

comes a new branch in computer vision. The main focus of this thesis is to develop three different stereoscopic saliency models that are each able to generate a saliency map of a stereoscopic image. Each model is based on a different way of leveraging depth information. One model uses depth information as a weight in saliency analysis. A second stereo-vision model is based on the mechanisms of the human vision system. And a third three-mechanism model incorporates a multi-feature analysis selection strategy. Accordingly, this section briefly introduces saliency detection in computer vision, the development of 2D saliency detection and emerging 3D saliency detection models.

1.1.1 Visual saliency

Humans have a unique ability to understand complex scenes in real time, despite the limited speed of the neurons we have available for such tasks. Intermediate and higher visual processes appear to select a subset of the available sensory information before further processing (Tsotsos, Culhane, Wai, Lai, Davis & Nuflo 1995), most likely to reduce the complexity of a scene for analysis (Niebur & Koch 1998). This selection appears to occur as a spatially circumscribed region of the visual scene, the so-called "visual attention", which allows humans to scan the scene in both a rapid, saliency-driven, bottom-up, task-independent manner, and in a slower, volition-controlled, top-down, task-dependent manner (Tsotsos et al. 1995).

The definition of visual attention is the behavioral and cognitive process of selectively concentrating on a discrete aspect of information, whether subjective or objective, while ignoring other perceived information. It is the mind taking possession, in clear and vivid form, of one of several simultaneous objects or trains of thought. Focalisation and concentration of consciousness are its essences. Visual attention has also been referred to as the allocation of limited processing resources (Anderson 1990).

Visual attention is a major area of investigation within education, neuropsychology, psychology and neuroscience. Areas of active investigation

involve determining the source of the sensory cues and signals that generate attention, the effects of those sensory cues and signals on the tuning properties of sensory neurons, and the relationships between attention and other behavioral and cognitive processes like working memory and vigilance (Eriksen & James 1986).

A key property of visual attention is that attention can be selective given some cues, such as luminance, contrast, shape, and orientation. These cues reflect some mechanisms of the human vision system, and most computational models of visual attention are based on these mechanisms.

1.1.2 Application of visual saliency

Computational models of visual attention simulate the attention mechanism of humans. These models are used in many fields, such as visual neuroscience, computer vision, and multimedia processing (Borji & Itti 2013). Visual attention enables the discovery of an object or region that efficiently represents a scene and, thus, can be harnessed to solved complex vision problems like scene understanding. Visual attention helps us to distinguish between the foreground and background of an image, and highlights regions that attract people's attention. Therefore, visual attention is a very important research topic in computer vision and is widely used for such purposes as object detection (Viola & Jones 2001)(Felzenszwalb, Girshick, McAllester & Ramanan 2010), object recognition (Rutishauser, Walther, Koch & Perona 2004)(Ren, Gao, Chia & Tsang 2014), image retrieval (Li, Jiang, Zha, Wu & Huang 2013)(Smeulders, Worring, Santini, Gupta & Jain 2000), image segmentation (Cheng, Jiang, Sun & Wang 2001), and so on.

1.1.3 Classification of visual saliency

The approach to the prediction of visual attention are usually divided into two categories: bottom-up and top-down (Yarbus, Haigh & Rigss 1967). Bottom-up is a rapid, data-driven, task-independent process and is usually

feed-forward. A prototypical example of a bottom-up attention model is the act of looking at a scene that only has one horizontal bar among several vertical bars; attention is immediately drawn to the horizontal bar (Treisman & Gelade 1980). Although many models use this mechanism, they can only explain a small fraction of the eye movements since the majority of visual attention is top-down, i.e., task-driven (Henderson & Hollingworth 1999), such as recognizing an object or a human face. For example, years of evolution have made people very sensitive to human faces in a scene. The top-down mechanism is a slower, task-driven, task-dependent process. It relies on a humans subjective intentions, experiences, and the target (Frintrop, Rome & Christensen 2010). Top-down attention models consider high-level cognitive features to quantify visual saliency, such as human faces (Judd, Ehinger, Durand & Torralba 2009) and prior knowledge about the target (Frintrop et al. 2010). Among these top-down features, prior knowledge about the target is the most difficult to model. Recently, a number of saliency models have incorporated both top-down and bottom-up feature detection processes in an effort to improve prediction accuracy (Jiang, Ling, Yu & Peng 2013). Wei *et al.* (Wei, Wen & Sun 2013) turned to background priors to guide saliency detection. Goferman *et al.* (Goferman, Zelnik-Manor & Tal 2010) and Judd *et al.* (Judd et al. 2009) integrated high-level information, making their methods potentially suitable for specific tasks.

However, visual attention can also be classified in other ways. Based on the results of prediction, visual attention can also be categorized into two types of models: salient object detection and fixation prediction. Salient object detection models are designed to highlight salient objects while completely suppressing the background regions in the generated saliency maps (Liu, Sun, Zheng, Tang & Shum 2007)(Achanta, Estrada, Wils & Süsstrunk 2008)(Liu, Zou & Le Meur 2013)(Liu, Shi, Shen, Xue, Ngan & Zhang 2012). Fixation prediction models predict the region's humans will fixate on within a saliency map to verify the efficacy of a saliency model and to understand human visual attention. A typical prediction model was proposed by Itti *et*

al. based on the feature contrast of intensity, color, and orientation (Itti, Koch & Niebur 1998) based on the feature contrasts in intensity, color, and orientation. However, the boundary between fixation prediction models and salient object detection models has blurred, as these models now share many concepts associated with established areas of computer vision, such as object segmentation algorithms (Li, Hou, Koch, Rehg & Yuille 2014), which use a fixation model to perform object detection, and fixation prediction models (Zhang & Sclaroff 2013)(Erdem & Erdem 2013), which threshold their saliency maps to detect and segment the salient proto-objects.

1.1.4 Development of 3D saliency detection

Most current saliency detection models are designed for 2D images. However, with the rapid development of 3D technology, an increasing number of applications are emerging for 3D images or videos, such as 3D rendering (Chamaret, Godeffroy, Lopez & Le Meur 2010), 3D visual quality assessment (Huynh-Thu, Barkowsky & Le Callet 2011) and 3D video coding (Shao, Jiang, Yu, Chen & Ho 2012). These 3D applications increase the demand for saliency models that are designed for 3D visual content.

Unlike 2D saliency detection, 3D saliency models must consider the depth dimension. Thus, the key issues in a 3D saliency detection model are how to incorporate depth as a factor and how to combine that depth factor with 2D information.

Some researchers have built models based on the mechanisms of stereoscopic perception in the human vision system, which are depth and color information. Bruce and Tsotsos (Bruce & Tsotsos 2005*a*) extended a 2D model that uses a visual pyramid processing architecture by adding neuronal units to model stereoscopic vision. However, this study does not include a computational model. Thus, the mechanisms of stereoscopic vision still pose several research challenges, such as how to build a neural vision model.

Other researchers have used depth saliency as a feature for saliency measurement. Depth saliency is extracted from a depth map, or disparity map,

to create an additional depth saliency map. The final result is a combination of the 2D saliency map and the depth saliency map. Niu *et al.* (Niu, Geng, Li & Liu 2012) explored stereo saliency by analyzing the characteristics of stereo vision and proposed a depth saliency model for a depth map that would expand the 2D saliency model. However, the proposed model does not fully explore the relationship between the depth model and the 2D saliency model.

There have also been some attempts to build a model directly. These models are designed to fuse the depth feature and 2D features into one saliency measurement (Fan, Liu & Sun 2014). Composite 2D and 3D saliency models to analyze the stereoscopic saliency have also emerged in recent years. However, despite the many and varied stereoscopic saliency detection models that have been proposed, no one model has become the standard for stereoscopic saliency analysis.

1.2 Research issues

This overview of the scholarly progress in stereoscopic saliency analysis reveals some current research limitations. These issues are discussed next.

1.2.1 Depth factor

Unlike 2D saliency detection, stereoscopic saliency detection needs to consider depth as a factor. Direct solutions treat the depth factor as a weight to leverage saliency analysis, with two main benefits. First, most existing 2D saliency models can easily be complemented with a weighted depth factor to extend their applicability to 3D content. Second, the classic 2D saliency models have already been proven to be effective, so good results should be relatively easy to attain. However, depth factors are able to convey different kinds of information in stereoscopic saliency analysis. Therefore, different ways of using depth factor as a weight need to be explored.

1.2.2 Mechanisms of stereoscopic vision

Simply treating depth as a weight in stereoscopic saliency does not necessarily provide a complete analysis of 3D content. Other mechanisms of stereoscopic vision, such as the pop-out effect and comfort zones, need to be considered for a stereo-vision model to effectively detect stereoscopic saliency. These mechanisms reflect some of the characteristics of the depth factor, and they are theoretically more consistent with human visual attention. Therefore, further exploration into how to design a computational model based on the mechanisms of the human visual system is required.

1.2.3 Relationship among these mechanisms

Further, expanding the features used to analyze stereoscopic saliency to include the pop-out effect and comfort zones still does not explain all phenomenon in stereoscopic content (Cheng, Zhang, Wu, An & Liu 2017). Some conditions create conflicts between the pop-out effect and comfort zones, which may negatively impact stereoscopic saliency analysis. Therefore, stereoscopic saliency detection models could be improved by reducing these conflicts. Additionally, a new mechanism needs to be defined for special cases to explain the "background effect". The background effect occurs when salient regions are located in or near a background region - a phenomenon that neither the pop-out effect or the comfort zone can explain.

1.3 Research contributions

The purpose of this thesis is to study stereoscopic saliency detection. Given the characteristics of the depth factor in the human visual system and the gaps in the literature identified above, this research project has three primary objectives.

- Develop a preliminary saliency model for stereoscopic images.

Chapter 3 presents a preliminary saliency model for stereoscopic images. This model exploits depth information and treats the depth factor of an image as a weight to leverage saliency analysis. First, low-level features from the color and depth maps are extracted. Then, to extract the structural information from the depth map, the surrounding Boolean-based map is computed as a weight to enhance the low-level features. Lastly, a stereoscopic center prior enhancement based on the saliency probability distribution in the depth map is used to obtain the final saliency.

- Develop a stereoscopic saliency detection model based on stereo contrast and stereo focus.

  Two characteristics of the stereoscopic vision, the pop-out effect and comfort zones, are explored in Chapter 3. In line with these two characteristics, the model presented in Chapter 4 predicts stereoscopic visual saliency using stereo contrast and stereo focus. The stereo contrast submodel measures stereo saliency based on color, depth contrast and the pop-out effect. In parallel, the stereo focus submodel measures the degree of focus based on monocular vision and comfort zones. The resulting values from these two submodels are clustered. Multi-scale fusion is then used to generate a map for each of the submodels, and a Bayesian integration scheme combines both maps into a stereo saliency map. Experimental results on two eye-tracking databases show that this method outperforms the state-of-the-art saliency models.

- Develop a computational model for stereoscopic 3D visual saliency

  Chapter 5 explores the role of depth information in analyzing stereoscopic saliency and presents a computational model that predicts stereoscopic visual saliency based on three aspects of human vision: the pop-out effect, comfort zones, and the background effect. Most salient stereoscopic regions can be explained by analyzing these three phenomena. Therefore, the model presented in this chapter model comprises

three submodules, each describing one aspect of saliency distribution, and a control function that can be used to independently adjust the three models. The relationship between the three models is not mutually exclusive. One, two, or three phenomena may appear in one image. Therefore, to accurately determine which phenomena the image conforms to, the model incorporates a strategy that selects the appropriate combination of submodels based on the content of the image. The approach is implemented within a purpose-built, multi-feature analysis framework that considers three features surrounding regions, color and depth contrast and points of interest to further enhance prediction.

1.4 Thesis structure

The structure of this thesis follows and is also illustrated in Fig. 1.1.

Chapter 1 provides the background to this thesis, beginning with a brief discussion on the stereoscopic saliency analysis, the research questions and the contributions of this thesis to the literature. The structure of the thesis is also provided.

Chapter 2 includes a literature review of 3D saliency models, including their classifications and the common methods. Useful tools and 2D features in saliency analysis are also discussed along with the benchmark stereoscopic saliency datasets.

Chapter 3 presents a preliminary saliency model for stereoscopic images. It is a direct stereoscopic saliency detection model based on the 2D contrast model and depth weight. However, this model enhances the 2D contrast model in three ways: at a low-level, depth contrast is treated as a weight; at the middle-level, depth structure information is treated as a weight; and at a high-level aspect, the saliency probability distribution of the depth map is treated as a weight.

Chapter 4 presents a stereoscopic saliency detection model based on stereo contrast and stereo focus. Two characteristics of stereoscopic vision are ex-

ploited: stereo contrast and stereo focus. Stereo contrast is based on the contrast in color and depth and the pop-out effect. Stereo focus describes the binocular and monocular focus regions of human vision. The values resulting from the two modules are enhanced individually to make the salient regions more distinct, and then each is converted into a saliency map using multi-scale fusion. The two saliency maps are integrated using Bayesian integration.

Chapter 5 presents a computational model for stereoscopic 3D visual saliency. The model comprises three modules based on the pop-out effect, comfort zones and the background effect, respectively. Given most salient stereoscopic regions can be explained by analyzing these three phenomena, the model incorporates a strategy that selects the appropriate combination of submodels based on the content of the image. The approach is implemented within a purpose-built, multi-feature analysis framework that considers three features surrounding regions, color and depth contrast and points of interest.

Chapter 6 concludes the thesis and outlines the scope of future work.

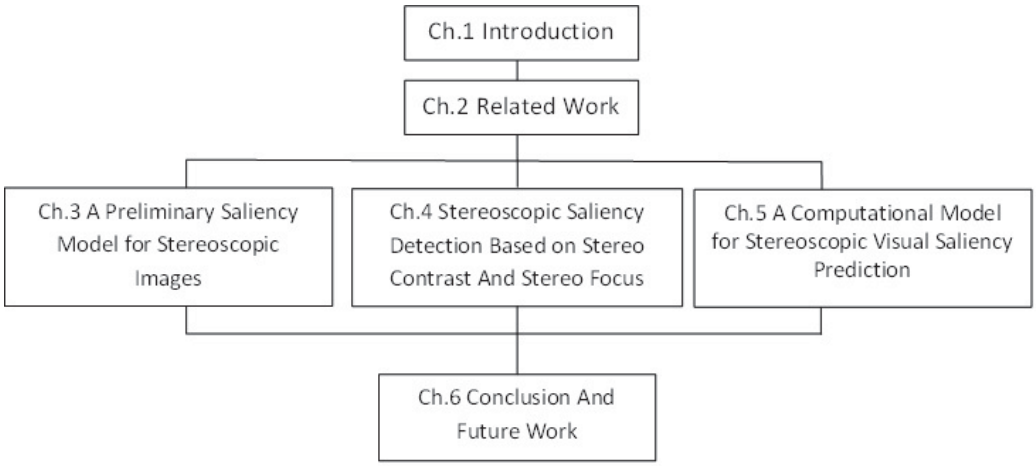

Figure 1.1: Thesis Structure. Ch. 1 introduced the research background. Ch.2 provides the literature review. Ch. 3 presents a preliminary saliency model for stereoscopic images. Ch. 4 propose stereoscopic saliency detection based on stereo contrast and stereo focus. Ch. 5 presents a computational model for stereoscopic 3D visual saliency based on three mechanisms: pop-out effect, comfort zone, and background effect.Ch. 6 provides a final summary of this research and also suggests some future directions.

Chapter 2

Literature Review

This chapter reviews the related work, which includes 3D saliency models, fundamental methods and stereoscopic datasets. The development and classification of 3D saliency detection are introduced in Section 1, and the related fundamental methods for saliency analysis are introduced in Section 2. Section 3 introduces the stereoscopic saliency datasets and their limitations. A summary is provided in the last section.

2.1 The 3D saliency model

3D saliency modeling is a new field of computer vision and is continuing to develop. To understand the concept of stereoscopic saliency analysis, the development of 3D saliency models and their classifications are introduced first, followed by recent research on 3D saliency detection.

2.1.1 Development of 3D saliency detection

In some applications, 2D saliency models can be directly applied to 3D saliency detection (Lang, Hornung, Wang, Poulakos, Smolic & Gross 2010)(Jeong, Ban & Lee 2008) with some degree of effectiveness. However, some researchers have found that the depth cues in 3D content have an effect on human visual attention (Häkkinen, Kawai, Takatalo, Mitsuya & Nyman

2010)(Wang, Le Callet, Tourancheau, Ricordel & Da Silva 2012), and the quality of predictions can be improved by incorporating depth information into saliency detection models. This discovery has lead to an increase in research attention of the role of depth information in 3D saliency analysis.

Based on the results of psychophysical experiments, some studies have begun to exploit how visual attention is influenced by 2D visual features and additional depth cues, both qualitatively and quantitatively.

One of the earliest works was proposed by Jansen *et al.* (Jansen, Onat & König 2009). They studied the influence of disparity on viewing behaviors with 2D and 3D still images. A free-viewing task on 2D and 3D versions of the same set of images was conducted. They found that the additional depth information led to an increased number of fixations, shorter and faster saccades and broader spatial exploration. However, there was no significant difference between the viewing of 2D and 3D stimuli concerning the saliency of several 2D visual features, such as mean luminance, luminance contrast and texture contrast. These results imply that: (a) the influence of 2D low-level visual features are important in 3D visual saliency analysis; and (b) adapting existing 2D visual attention models to the design of 3D models is feasible.

Liu *et al.* (Liu, Cormack & Bovik 2010) investigated the visual features at fixation positions in stereo images with natural content. Rather than comparing the viewing behavior with 2D images compared to 3D images, they paid more attention to the visual features extracted from fixations and random locations when viewing still 3D images. The study illustrates that the values of some 2D visual features, such as luminance contrast and luminance gradient, are generally higher in fixation areas. The results also show that the disparity contrast and disparity gradients are lower at fixation locations than randomly selected locations. However, these findings are inconsistent with the results of Jansen *et al.* (Jansen et al. 2009), which show that observers consistently look more at depth discontinuities (high disparity contrast areas) than at planar surfaces. The contradiction in Liu *et al.*'s findings may lie in

the quality of the ground-truth disparity map, which was generated through a simple correspondence approach rather than from depth-range sensing systems or a sophisticated depth estimation approach. Therefore, Liu *et al.*'s final results might have been affected by a considerable amount of noise in the estimated disparity maps.

Hakkinen *et al.* (Häkkinen et al. 2010) examined the difference in eye-movement patterns when viewing 2D and 3D versions of the same video content. The results show that eye movements are more widely distributed for 3D content. Compared to 3D content, viewers not only look at the main actors but they also look at other targets in typical movie content. Their results indicate that depth information provides viewers with additional information and, thus, forms new salient regions in a scene. These results also suggest that a stereoscopic saliency map could combine both 2D saliency and depth saliency. Moreover, Ramasamy *et al.*s results (Liu et al. 2010) show that viewers' gaze points are more concentrated when viewing 3D versions of some content, for example, scenes containing long deep hallways.

In terms of studies that focus on the places where fixations on depth tend to be located, Wang *et al.* (Wang et al. 2012) investigated "depth-bias" in a free-viewing task of still stereoscopic synthetic stimuli. They found that the objects closest to the viewers always attract the most fixations. And that the number of fixations on each object decreases as the depth order of the object increased, except for the furthest object which receives slightly more fixations than the one or two objects in front of it. They also found that the number of fixations on the objects in the different depth planes is time-dependent. This is consistent with the results of Jansen *et al.* (Jansen et al. 2009). Considering the influence of center bias in 2D visual attention, these results indicate an additional location prior exists according to the depth information in the 3D content viewed. This location prior implies the possibility of integrating depth information using a weight.

Wismeijer *et al.* (Wismeijer, Erkelens, van Ee & Wexler 2010) studied whether saccades were aligned with either individual depth cues or with a

combination of depth cues, by presenting stimuli in which monocular perspective cues and binocular disparity cues conflicted. They found that there is a weighted linear combination of cues when the conflicts are small and a cue dominance when the conflicts are large. The results also show that vergence is only dominated by binocular disparity and the interocular distance recorded by the binocular eye-tracking experiment. Hence, 3D content should compensate for the local disparity value.

The combined results of these comprehensive research efforts into studying viewing behavior with 3D content indicate that viewing behavior has strong relationships to saliency detection, and that there are some relationships between 3D content and additional depth cues that affect viewing behaviours. They also imply that 3D saliency detection relies on 2D features and additional depth cues.

2.1.2 Classification of 3D saliency detection

Compared to the plethora of 2D visual saliency models, only a few 3D visual saliency models have been proposed. Further, the few existing 3D visual saliency models are all based on 2D saliency features and depth cues. These models can be divided into two categories based on their use of the depth information.

One category relies on depth cues for visual attention. These models are called "depth models". They exploit depth cues for saliency analysis in an attempt to use depth information to enhance 2D features. Depth models can be divided into two subclasses: depth weight models and depth saliency models, as shown in Fig.2.1.

Depth weight models do not contain any depth feature extraction processes based on a depth map. Rather, they treat depth information as a weight factor for 2D visual saliency features. The saliency of each location in the scene, such as pixels, targets or the depth plane, is directly related to its depth. Maki *et al.* (Maki, Nordlund & Eklundh 1996) proposed a saliency model based on assigning the target closest to the observer with the highest

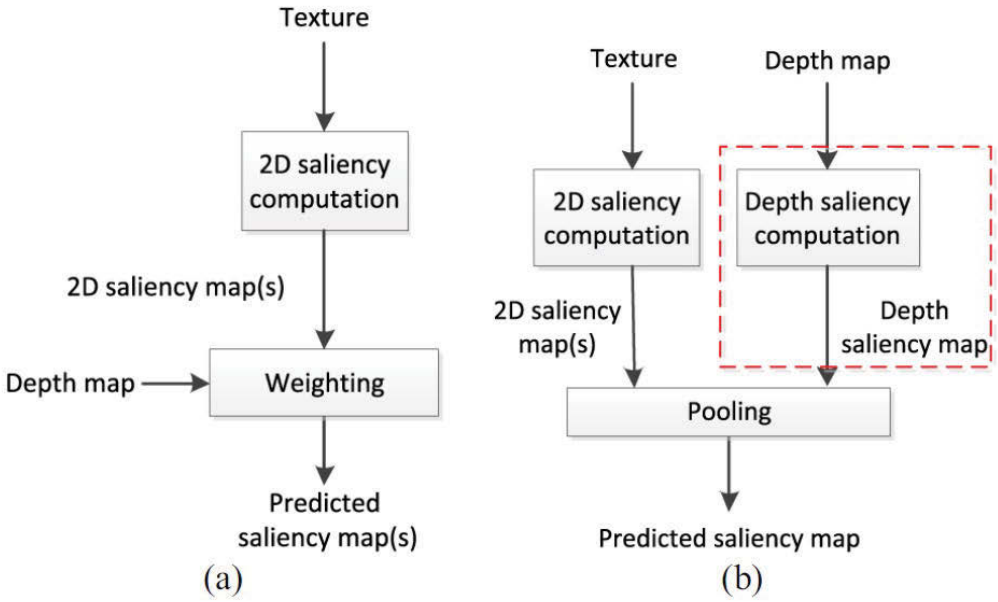

Figure 2.1: (a) a typical depth weight model and (b) a typical depth saliency model.

priority. Zhang *et al.* (Zhang, Jiang, Yu, Chen & Dai 2010) proposed a saliency model based on irregular space conversion and an assumption that the pixels closer to the observer and at the front of the screen are considered to be more salient. This model only considers one mechanism of the human vision system: the pop-out effect. Chamaret *et al.* (Chamaret et al. 2010) weights each pixel in a 2D saliency map according to its depth value. It is worth noting thatThese three models are not evaluated by the quantitative measure in the eye-tracking data.

Depth saliency models treat depth saliency as additional information. Depth features are first extracted from a depth map to compute a depth saliency map. Then, the resulting stereoscopic saliency map is combined with a 2D visual saliency model and the depth saliency model using a saliency map pooling strategy. Niu *et al.* (Niu et al. 2012) proposed using the characteristics of stereoscopic vision to enhance existing saliency models; however, their method does not fully exploit the relationship between existing saliency

models and a depth model. Ramenahalli *et al.* (Ramenahalli & Niebur 2013) extended the Itti model by treating depth information as an additional channel. This model uses color, intensity, orientation and depth channels to generate a 3D saliency map. However, the characteristics of stereoscopic vision, such as comfort zones and the pop-out effect (Häkkinen et al. 2010), are not considered in this model.

The second category of 3D visual saliency models considers the mechanisms of the human vision system when building the model by fusing depth information with other 2D features. These models are called "stereo-vision models". Fan *et al.* (Fan et al. 2014) proposed the use of region-level depth, colour contrast and spatial information to measure saliency. Peng *et al.* (Peng, Li, Xiong, Hu & Ji 2014) proposed an RGBD model based on both depth and appearance cues derived from color and depth contrast features. A study by Khaustova *et al.* (Khaustova, Fournier, Wyckens & Le Meur 2013) shows that the characteristics of natural disparity improve prediction accuracy in 3D visual saliency. It is, therefore, reasonable to consider that a 3D saliency models performance can be improved by incorporating human vision characteristics into the design (Bruce & Tsotsos 2005*a*).

2.1.3 Recent research about 3D saliency detection

All the above 3D visual saliency models have proven to be effective in experiments under some conditions. However, to the best of my knowledge, none of these models are able to explain all the phenomena of human visual attention. Depth models simply rely on depth factors to enhance 2D features, neglecting the mechanisms of associated with depth information. Depth saliency models are effective at individually extracting 2D features and depth information for saliency detection. However, these models ignore the relationships between depth and the other features. Stereoscopic models do consider the relationships between depth and other features, but designing a suitable and reasonable model for stereoscopic visual attention is a difficult task.

Hence, to explain more phenomena in human visual attention, some scholars have designed mixed saliency models that combine multiple features from different models to increase prediction performance. Iana *et al.* (Iatsun, Larabi & Fernandez-Maloigne 2015) proposed a new stereoscopic saliency model by considering two separate spatial saliency models. One model relies on the characteristics of points of interest; the other relies on depth-color saliency. Jiang *et al.*(Jiang, Shao, Jiang, Yu, Peng & Yu 2015) designed a saliency model by fusing three models: a 2D saliency model, a depth saliency model and a visual comfort saliency model. The multiple features extracted from different models reflect different cues for saliency detection. Combining these different features improved prediction performance in saliency analysis because multiple features simultaneously rely on multiple mechanisms of the human vision system (Pylyshyn & Storm 1988)(Awh & Pashler 2000), which effectively limits noise.

2.2 The fundamental method of saliency analysis

The analysis of saliency in stereoscopic images discussed in this thesis is built on several foundational techniques. These methods include simple linear iterative clustering (SLIC) segmentation, multi-scale integration, Bayesian integration and center bias. These methods supply a very good basis for stereoscopic saliency analysis. The following subsections explain each method in more detail.

2.2.1 SLIC and multi-scale integration

Segmentation is a useful tool for both 2D and 3D saliency detection. The main idea of this method is to divide an image into many segments based on an edge or texture. We can see that in different size of the superpixel, SLIC can segment the image into small part based on the edge and texture.

Clustering these segments according to the different cues can create an initial saliency area (i.e., a saliency map). SLIC is based on K-means clustering (Achanta, Shaji, Smith, Lucchi, Fua & Susstrunk 2012). The method is very simple to use and easy to understand. The algorithms do not have many parameters, and generating superpixels is straightforward and efficient. Simply setting the number of superpixels initializes the center of each segment, and then the algorithms automatically set the weighted distance measure to balance color and spatial weight. The technique is as sensitive to image boundaries as previous methods, if not more sensitive.

SLIC has two parameters in the algorithms is k and S. The first parameter is the number of superpixels, which can initialize the center of each segment. The second parameter is a little difficult to explain but it can be simply understood as a weighted distance measure to balance color and spatial weight.

For color images in LAB color space, the first step starts at an initialization where k initialize cluster centers c_i and search area is on a regular grid spaced S pixels apart. Then the centers are moved to another location relying on the lowest gradient position in a 33 neighborhood. This can avoid move a center to an edge and reduce the chance of seeding a superpixel with a noisy pixel. In the assignment step, every pixel i is associated with the nearest cluster center whose search area overlaps the location. This is the reason that SLIC speeds up because limiting the search area can significantly reduce the number of distance calculations. When each pixel has been associated to the nearest cluster center, an update step can adjust the cluster centers location. Experiments show that 10 iterations suffice is enough for most images. Finally, a post-processing step connects disjoint pixels to nearby superpixels. The result is shown in Fig.2.2

The models presented in this thesis use SLIC as the method of processing the color and depth maps. However, in practice, SLIC was always combined with multi-scale integration to facilitate stereoscopic saliency analysis. The input images (color/depth image) were segmented several times by control-

Figure 2.2: Examples of SLIC.

ling the numbers (scale) of the superpixels in the SLIC segmentation. For each scale, the images were segmented into a set of non-overlapping superpixels. Each superpixel was described by the mean color/depth feature and the mean coordinate of the pixels. Then, according to the different saliency cues, different models were built to calculate the saliency value of each superpixel.

A multi-scale integration of all the scales was conducted to produce a pixel-level saliency image following the method in (Lu, Li, Zhang, Ruan & Yang 2016) of fusing the segmentation feature values in the different scales. This method considers the multi-scale value and its textural information using the textural features of the pixel and its corresponding superpixel as the weight value to average the multi-scale value. The textural feature of pixel p and its corresponding superpixel t_s are used as the weight value to average the multi-scale value as follows:

$$S(p) = \frac{\sum_{s=1}^{M} z_{pt_s} \cdot S(t_s)}{\sum_{s=1}^{M} z_{pt_s}} \tag{2.1}$$

$$z_{pt_s} = \frac{1}{\left\| P_p - x_{t_s} \right\|_2} \tag{2.2}$$

where M is the numbers of all scales. $S(t_s)$ is the value of superpixel t_s in scale s. P_p is a 6-dimensional feature of pixel p that includes the colour/depth and position information as H, S, V, x, y, z and x_{t_s} is a 6-dimensional feature of the superpixel t_s including the pixel p in scale s. The pixel-level saliency map s is computed after the multi-scale integration.

2.2.2 Bayesian integration

When two saliency maps have been built by two different cues or methods, they need to be integrated. However, as discussed in (Gopalakrishnan, Hu & Rajan 2009), the quality of good individual saliency maps may deteriorate when they are combined with other maps using a weighting technique. A Bayesian model is a good choice for integrating two saliency maps to avoid

this problem, as Bayesian models tend to be very robust to various types of images.

In terms of probability, the stereoscopic saliency of each pixel is defined as being equal to the probability of the point being viewed. A Bayesian formulation is used to measure the saliency by posterior probability (Chamaret et al. 2010)(Sun, Lu & Li 2012), which can be expressed as

$$p(g|h(p)) = \frac{p(g)p(h(g)|g)}{p(g)p(h(p)|g) + p(b)p(h(p)|b)} \tag{2.3}$$

$$p(b) = 1 - p(g) \tag{2.4}$$

where $h(p)$ is a feature vector of pixel p. g is the gaze area and b is the background. $p(g)$ and $p(b)$ represent the prior distribution of the gaze area and background respectively. $p(h(p)|g)$ and $p(h(p)|b)$ are the observation likelihoods (Huynh-Thu et al. 2011), as shown below:

$$p(h(p)|g) = \prod_{i \in H,S,V} \frac{N_{g(h(p))}}{N_g} \tag{2.5}$$

$$p(h(p)|b) = \prod_{i \in H,S,V} \frac{N_{b(h(p))}}{N_b} \tag{2.6}$$

where N_g is the number of pixels in the gaze area and $N_{g(h(p))}$ is the number of pixels whose colour features fall into the gaze area g, which contains feature $h(p)$. The colour feature distribution of the background is likewise denoted as N_b and $N_{b(h(p))}$.

The Bayesian integration method is as follows. Take one saliency map S_i $(i = 1, 2)$, as the prior probability and use the other map S_j $(i \neq j, j = 1, 2)$ to compute the likelihood. Then, based on the Bayesian formulation, compute the posterior probability as the final saliency.

In practice, however, the threshold of the map S_i is set to its mean saliency value. The gaze area and background region can be described by g_i and b_i.

The likelihood can then be computed by comparing S_j and S_i in terms of the gaze area and background bins at pixel p:

$$p(S_j(p)|g_i) = \frac{N_{g_i(S_j(p))}}{N_{g_i}} \tag{2.7}$$

$$p(S_j(p)|b_i) = \frac{N_{b_i(S_j(p))}}{N_{b_i}} \tag{2.8}$$

The posterior probability is computed with S_i as the prior by

$$p(g_i|S_j(p)) = \frac{S_i(p) \cdot p(S_j(p)|g_i)}{S_i(p) \cdot p(S_j(p)|g_i) + (1 - S_i(p)) \cdot p(S_j(p)|b_i)} \tag{2.9}$$

Similarly, the posterior saliency with S_j as the prior can also be computed. Two posterior probabilities are used to produce one integrated saliency map.

$$S(S_1(p), S_2(p)) = p(g_2|S_1(p)) + p(g_1|S_2(p)) \tag{2.10}$$

2.2.3 Center bias

The center bias describes the saliency probability distribution in the X-axis and the Y-axis. Since many datasets have a property that locates the salient object or region in the center of the image (Borji, Cheng, Jiang & Li 2015), the center bias G can be used to process the saliency map in both the X and Y directions. In general, the centre bias can be modeled using Gaussian standard deviations, as follows:

$$G(p) = \exp[-\frac{(x-u)^2}{2\sigma_x^2} - \frac{(y-v)^2}{2\sigma_y^2}] \tag{2.11}$$

where (x, y) is the coordinate of the pixel p, and u and v denote the centre of the image. Thus, $\sigma_x = 0.25 \times H$ and $\sigma_y = 0.25 \times W$, where H and W represent the height and width of the image, respectively.

2.3 Experimental datasets and measurements

One obstacle to the development of 3D saliency is the lack of enough stereoscopic image material. Additionally, the quality of the stereoscopic images captured by various 3D devices differs significantly. For example, a Panasonic AG-3DA1 3D camera can supply high-quality left/right images image/video for saliency analysis in experiments. But the stereoscopic image generated by the Kinect-1 is a 640x480 image with holes that may cause noise. When both are used in a saliency analysis experiment, the low-quality stereoscopic images may introduce noise into the results, which makes designing a stereoscopic saliency model difficult. In this thesis, all the depth map are supplied by datasets, which is generated by the different depth-capture sensors. Some of them are captured by the stereoscopic camera including the Panasonic AG-3DA1 3D camera and Kinect-1. The others are generated by two-view image with the peoples adjustment.

Two public datasets were selected to evaluate the performance of the saliency models presented in this thesis. The first was published in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). This dataset is an eye-tracking database with 600 stereoscopic 3D images of outdoor and indoor scenes containing different numbers of objects of different sizes. Additionally, different degrees of interaction or activity are depicted in the scene. The images were generated by the Kinect-1 at 640x480 pixels, so they are likely to introduce noise and contain a number of holes that need to be smoothed. Stereoscopic image pairs were produced by pre-processing, calibration and post-processing. Eye-tracking data from 80 participants (ranging from 20 to 33 years old) were captured by an eye-tracker in both 2D and 3D free-viewing experiments. In order to produce a continuous fixation map of an image, we convolve a Gaussian filter across all corresponding viewers fixation locations. Human fixation maps were constructed from the fixations of viewers to globally represent the spatial distribution of human fixations. A Gaussian kernel was used to obtain the continuous fixation density maps as the ground-truth maps. Both the 2D and 3D fixation maps are supplied with the dataset to

facilitate comparisons. In this thesis, the 3D fixation maps were used as the stereoscopic 3D ground-truth map.

The second eye-tracking database was published in a study by (Wang et al. 2013). It includes 18 stereoscopic images of various types (e.g., indoor scenes, outdoor scenes and scenes containing various numbers of objects). Ten images were taken from the Middlebury database 2005/2006, with their accompanying accurate depth images, which is adjusted by people. The other eight images were recorded by the authors with a Panasonic AG-3DA1 3D camera as 1080P stereoscopic images. To avoid 3D fatigue (Hoffman, Girshick, Akeley & Banks 2008) resulting from a conflict in the depth field, when, for example, one object is seen by the right eye but is missed by the left eye, the degree of vergence in human vision within the stereoscopic 3D viewing environment is considered in this eye-tracking experiment. The stereoscopic image pair is produced by pre-processing, calibration and post-processing. The eye-tracking data are captured in both 2D and 3D free-viewing experiments by the eye-tracker from 80 participants (ranging in age from 20 to 33 years old). Human fixation maps are constructed from the fixation of viewers to globally represent the spatial distribution of human fixation. Then a Gaussian kernel is used to obtain the continuous fixation density maps as the ground-truth maps

To quantitatively evaluate the performance of our proposed saliency model, similar quantitative measure methods to those used in (Wang et al. 2013) were followed. The performance of each model was measured by comparing the computed saliency map with the ground-truth map supplied in the database. Because any stereoscopic image pair comprises two images (left and right), the saliency map of the left image was used for comparison, similar to the study in (Wang et al. 2013). The area under the receiver operating characteristics curve (AUC) and correlation coefficient (CC) were used to evaluate quantitative performance for each model. Using AUC, human fixations are considered to be the positive set, and some points from the image are sampled to form the negative set. The saliency map S was then treated as a

binary classifier to separate the positive samples from the negatives. A ROC curve was generated for each image by thresholding over the saliency map and plotting the true positive rate vs the false positive rate. The resulting ROC curves were then averaged over all images, and the area underneath the final ROC curve was calculated as the AUC. The CC measures the strength of a linear relationship between the predicted saliency map and the ground-truth saliency map. When the CC is close to +1/1, there is almost a perfectly linear relationship between the two variables.

2.4 Summary

This chapter introduced the development of 3D saliency detection and some fundamental saliency tools used within this thesis. The stereoscopic datasets used in the experiments were also described. Section 2.1 provided the background on 3D saliency detection and related research through a review of the historical development of 3D saliency detection and current classifications of 3D saliency detection models. Section 2.2 explained the fundamental methods in saliency analysis used in this thesis. The SLIC segmentation method and multi-scale integration were introduced, followed by Bayesian integration, which analyses saliency from the perspective of Bayesian theory, and center bias as another a useful tool in saliency analysis. Section 2.3 presented two eye-tracking datasets, which have been used repeatedly for measuring and evaluating the performance of visual saliency detection, and are also used in the experiments for this thesis.

Chapter 3

A Preliminary Saliency Model for Stereoscopic Images

3.1 Introduction

Visual saliency is a fundamental problem in neuroscience, psychology, and vision perception, and refers to the measurement of low-level stimuli that attract human attention in visual processing (Itti et al. 1998). It measures the distinctiveness of a region from its neighboring regions and a more salient region conveys more information (Thomas & Thomas 2006). The computation of visual saliency is widely used in various applications of image processing, such as image segmentation (Ko & Nam 2006), content-aware image/video re-targeting (Luo, Yuan, Xue & Tian 2011), video quality assessment (Wang & Li 2011) and adaptive image compression based on region-of-interest (ROI) detection (Guo & Zhang 2010).

In general, visual saliency relies on four kinds of features for saliency detection (Goferman, Zelnik-Manor & Tal 2012): local features, global features, visual forms and high-level factors. The local features include low-level factors such as intensity, color and contrast. These always assume that human attention is sensitive to high-contrast regions. If a region is distinctive in intensity, color, texture or motion, it is considered a high salient region.

Global features detect the overall structure or arrangement of the features which form the salient region. A prototypical example of the attention model is looking at a scene with only one horizontal bar among several vertical bars, where attention is immediately drawn to the horizontal bar (Treisman & Gelade 1980). Visual forms or shapes assume that visual saliency can possess one or several centers of gravity about which the form is organized. A region with a more compact form should be assigned as more salient. High-level factors include semantic objects, such as human faces or objects which have more saliency because of our empirical knowledge.

Based on these features, various saliency detection algorithms are proposed. For example, local intensity, orientation and color features have been utilized to obtain visual saliency, based on center-surround enhancement (Itti et al. 1998). Global features have been exploited by contrast-based methods (Cheng, Zhang, Mitra, Huang & Hu 2011) and by spectral analysis methods (Hou & Zhang 2007). The shape feature has been studied by a Boolean map method (Zhang & Sclaroff 2013) which uses a set of binary images to compute the shape of the saliency region. High-level factors have been used to improve the accuracy of the saliency, such as object detection (Jiang et al. 2013), face detection (Cerf, Harel, Einhäuser & Koch 2008) and center bias (Borji et al. 2015).

The above models are designed for 2D saliency detection. However, with the rapid development of 3D technology, the number of applications for a 3D image or video is increasing, such as 3D visual quality assessment (Huynh-Thu et al. 2011), 3D video coding (Shao et al. 2012), 3D rendering (Chamaret et al. 2010), and more. These 3D applications require more and more saliency models for 3D visual content.

Compared to the significant progress in 2D saliency research, the work leveraging depth information for saliency analysis is rather limited (Peng et al. 2014). Niu *et al.* (Niu et al. 2012) analyzed the characteristics of stereo vision and proposed a depth saliency model for a depth map that would leverage the 2D saliency model for stereo saliency analysis. However,

the proposed model does not fully exploit the relationship between the depth model and the 2D saliency model. Fan *et al.* (Fan et al. 2014) utilized the region-level depth, color and spatial information to analyze saliency for stereoscopic images. Fang *et al.* (Fang, Wang, Narwaria, Le Callet & Lin 2013) proposed a stereoscopic saliency model based on the low-level features extracted by DCT (Discrete Cosine Transform) coefficients of image patches and feature contrast. However, they did not consider the characteristics of human stereo vision such as 3D fatigue or the pop-out effect.

According to the above analysis, the key issue for a 3D saliency detection model is how to adopt the depth factor and how to combine the depth factor with 2D information. In this chapter, a preliminary saliency detection model for stereoscopic images is proposed. This model utilizes depth information to leverage stereo saliency analysis from three aspects. In the low-level aspect, the local-global features are used to analyze saliency by considering the color and depth contrast in local and global ranges. In the mid-level aspect, the surrounding map based on the Boolean map is obtained as a weight value to enhance the local-global features. Lastly, by analyzing the saliency probability distribution in depth information, a stereo center prior enhancement is used to form the final saliency.

The rest of the chapter is organized as follows: Section 3.2 proposes a preliminary saliency detection model for stereoscopic images; Section 3.3 describes the quantitative comparison of the proposed model with the state-of-the-art algorithms; Section 3.4 gives the research outcome and discussion.

3.2 Proposed stereo saliency detection

The framework of the proposed stereo saliency detection method is shown in Fig.3.1. Firstly, the local and global features are extracted from the left image and depth map. Then, surrounded enhancement based on boolean map is used to increase the accuracy of the saliency. Lastly, the stereo center prior enhancement is utilized by considering the saliency probability distribution

in the depth map and color map to obtain the final saliency map. Each step in detail is described in the following subsection.

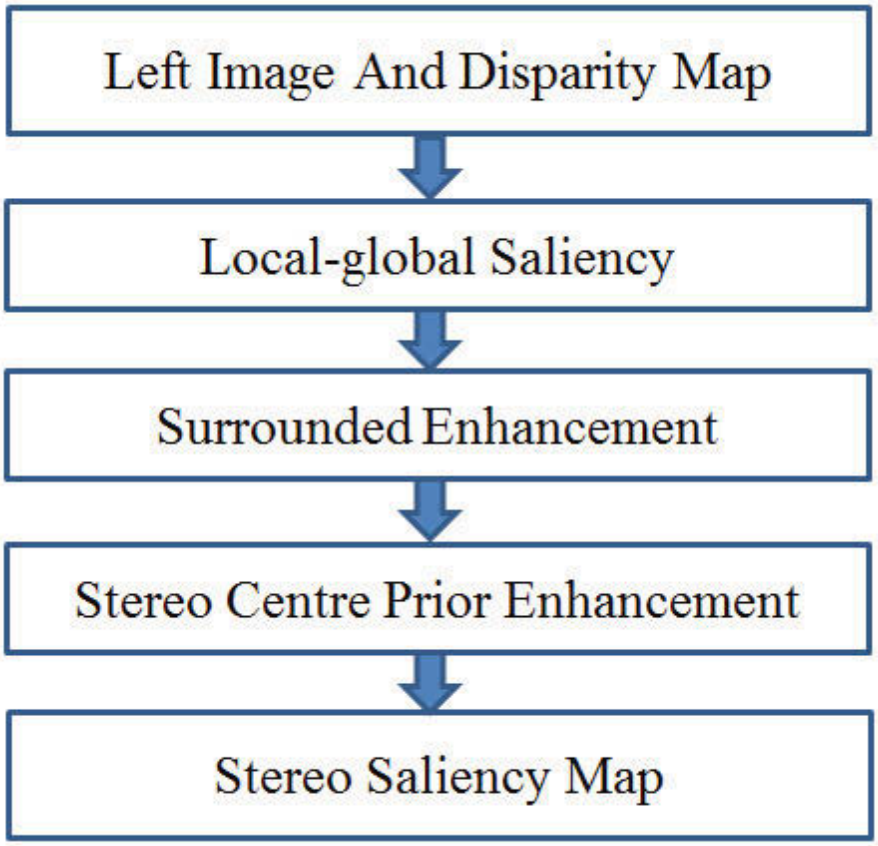

Figure 3.1: The framework of the proposed stereo saliency detection method

3.2.1 Local-global saliency

Based on observations of biological vision, where the vision system is sensitive to contrast in visual signals, a local-global saliency, defined by color and depth contrast, is proposed. The colors of the salient region are usually distinctive and show contrast with the other regions. However, if the color contrast is not distinct enough, the depth cue can leverage the saliency analysis under

the assumption that an absolute region is usually more salient for human visual attention. Therefore, a region is defined as a salient region if its color and depth are more distinct than the other regions, which include local and global ranges.

Specifically, to capture the structural information of the stereoscopic image, a simple linear iterative clustering (SLIC) algorithm (Achanta et al. 2012) is used for the segmentation. SLIC can segment an input image (left image) into N non-overlapping patches (superpixels). Let $DC(i,j)$ be the Euclidean distance between the vectored superpixels i and j in Lab color space, normalized to the range $[0,1]$. $L(i,j)$ is the position distance between the superpixel i and j. $DP(i,j)$ is the depth distance between superpixel i and j. Both of them are normalized to the range $[0,1]$. Based on the observation above, a distinctive measure $D(i,j)$ between a pair of superpixels i and j based on color, spatial and depth information can be defined as:

$$D(i,j) = (\frac{DC(i,j)}{1 + c * L(i,j)}) * \omega_j * DP(i,j) \tag{3.1}$$

where ω_j is the pixel number of superpixel j and c is a control value about spatial information ($c = 3$ in our implementation). As mentioned above, the saliency of a superpixel z can be defined by its distinctive measure as:

$$SC_R(z) = \sum_{i \neq z, i \in R} D(z,i) \tag{3.2}$$

where R is the range and $SC_R(z)$ is the saliency value of superpixel z in the range R. For different ranges, the local and global saliency are computed, as shown in Fig.3.2. All the distance is normalized to $[0,1]$. Based on experience, we set the $R = 0.3$.

In practice, to measure a superpixel distinctiveness, its distinctiveness with respect to all other superpixels. the proposed model simply considers the K most similar superpixels. As the most similar superpixels are extremely different from the current superpixel, clearly all image superpixels are extremely different from it. Therefore, the proposed model searches for

the K most similar superpixels $\{t_1, t_2, ...t_k\}$, based on $DC(i, j)$. The local-global saliency of superpixel z is expressed as:

$$SC(z) = \sum_R SC_R(z) \tag{3.3}$$

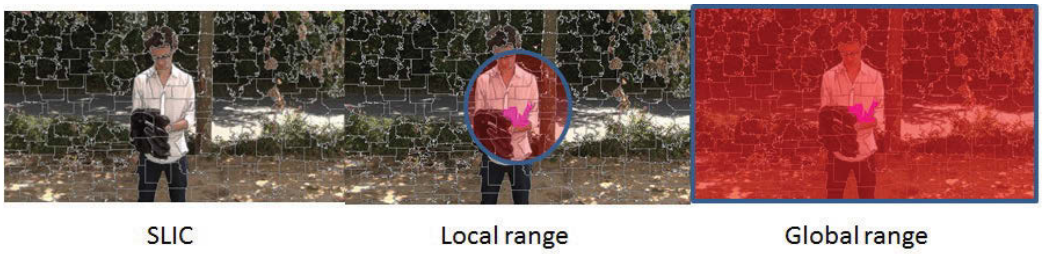

Figure 3.2: Global and local range.

3.2.2 Surrounded enhancement

In addition, to improve the performance of the local-global contrast feature, the topological structural information is used to leverage saliency detection. For example, Markov chain graphs (Harel, Koch & Perona 2006), a graph model measured (Wei et al. 2013) by its shortest edges and a Boolean map approach (Zhang & Sclaroff 2013) can obtain the structural information of the image for saliency analysis. In this chapter, a Boolean map for the depth channel is proposed, because it is easy to implement and produces favorable results from the experiments.

The Boolean map concept comes from Boolean map theory (Huang & Pashler 2007) of visual attention, where a viewer's momentary conscious awareness of a scene can be represented by a Boolean map. We assume the Boolean map is computed by a feature channel (here, we use the depth channel). The influence of a Boolean map B on visual attention can be expressed by an attention map $A(B)$, in which the salient regions are highlighted on

B.

$$BD = \int_{min}^{max} A(B)w(B|I)dB \tag{3.4}$$

where I is a feature channel and $w(B|I)$ is a weight value to describe the salient probability of the attention map B. BD represents a weighted map obtained by a depth map. The Boolean map is used to process the depth map because the depth map can supply stereo information to reflect the positional relationships between the different objects. The weighted map obtained by the depth map leverages the stereo saliency analysis.

In practice, at first, the thresholds by clustering the depth information is computed. Then, the depth map is processed by the thresholds to a set of Boolean maps $B = \{B_1, B_2, ...B_n\}$. Each Boolean map B_i computes an attention map A_i. Lastly, a linear combination of all attention maps forms the depth weight map.

Thresholds

The depth weight map relies on a set of Boolean maps by thresholding the depth map.

$$B_i = Threshold(DM, \theta_i) \tag{3.5}$$

where DM represents the depth map and θ_i is the threshold. If a superpixel depth value is bigger than θ_i, the function $Threshold(DM, \theta_i))$ is 1, otherwise it is 0.

The threshold θ_i divides a depth map into two regions: background region (0 region) and surrounded region (1 region), with the latter having a greater chance of being a salient region than the former (Huang & Pashler 2007). Therefore, a reasonable threshold is important for salient analysis. A clustering approach is used to determine the threshold, which can separate the background and surrounded region efficiently.

For each superpixel, the mean depth value of the pixels in one superpixel is d. The K-means algorithm is used to cluster N superpixels into K clusters,

via the depth value. To enlarge the difference between neighboring clusters, we adjust the d_i in cluster k by considering its own value and the other superpixels in cluster k as follows:

$$Su(i) = \delta \sum_{j=1, j\neq i}^{N_c} r_{ij} d_j + (1-\delta) d_i \tag{3.6}$$

where $\{1, 2, ..., N_c\}$ represents the N_c segment labels in cluster k and $Su(i)$ is the adjusted depth value of superpixel i and δ is a weight value, based on the spatial information of a pair of superpixels:

$$r_{ij} = \frac{L(i,j)}{\sum_{j=1, j\neq i}^{N_c} L(i,j)} \tag{3.7}$$

where $L(i, j)$ is the Euclidean distance between superpixels i and j.

Through the above process, all superpixel depth values are modified. Then, the center of each cluster is computed. The threshold is the median value of neighboring-cluster centers. For all Boolean maps, an opening operation with kernel ψ is then applied to each Boolean map for noise removal, such as an isolated point or making the boundary of the object more clearly visible (Huang & Pashler 2007).

Boolean map

By thresholding the depth map, we can obtain a Boolean map B, while an attention map $A(B)$ is computed based on the Gestalt principle for figure-ground segregation, in which the surrounded regions are more likely to be perceived as figures (Palmer 1999). Surrounded regions in a Boolean map are defined as having the property of a connected region (either of value 1 or 0) that has a closed outer contour (Huang & Pashler 2007). Under this definition, the regions connected to the image borders are not surrounded. To compute the attention map, 1 is the union of the surrounded regions and 0 is the background regions. A flood fill algorithm is used to implement this operation by masking out all the pixels connected to the image borders.

Before the linear combination step, the weight value $w(B|I)$ should be computed, which describes the salient probability of an attention map. Based on observation, the attention map with small concentrated active areas will receive more emphasis. Therefore, the weight value is inversely proportional to the surrounded region as follows:

$$w(B|I) = \exp \frac{1}{\omega_s} \tag{3.8}$$

where ω_s is the pixel number of surrounded region. An exponential function is used in Eq. (3.8) to emphasize the significance of surrounded region. All w are normalize to $[0, 1]$.

The weighted map from the depth map is expressed:

$$BD = \sum_{i=1, i\in\theta} A(B_i) w(B_i|I) \tag{3.9}$$

3.2.3 Stereo center prior enhancement

In the final stage, we fuse the saliency map, combined with stereo center prior enhancement. Stereo center prior enhancement relies on two aspects: depth bias and center bias.

Depth bias describes the saliency probability distribution on the Z-axis and the center bias describes the saliency probability distribution on the X-axis and Y-axis. Before the model of depth bias is built, we should know the characteristics of the depth information. When viewers spend a long time watching stereoscopic images or video, they may experience fatigue. The reason is that the focus range of the stereoscopic image or video may not be suitable for human focus. A good stereoscopic image needs to avoid and minimize 3D viewer fatigue. To avoid and minimize 3D fatigue for viewers, a comfort zone is proposed (Niu et al. 2012), which is a zone around the zero disparity plane. When photographers capture a stereoscopic image or video, they always place more important objects or regions in a comfort zone rather than in other zones. Based on this characteristic of depth information, the

important object or regions have a high probability of being located in a comfort zone. This is one characteristic of depth information. In addition, when viewers watch a stereo image or video, some objects may look like they are popping out of the screen because these objects have negative disparity. This phenomenon is named as the pop-out effect. Studies show that an object which has a pop-out effect has a high probability of catching the viewer's attention (Häkkinen et al. 2010). Based on these two characteristics of depth information, a Gaussian model to describe the depth bias Z based on the combination of the pop-out effect and comfort zone, which can be expressed as follows:

$$Z(p) = \begin{cases} \exp(\frac{d_p^2}{-2\sigma_2^2}) & d_p \geq 0 \\ \alpha \cdot \exp(\frac{d_p^2}{-2\sigma_2^2}) + (1-\alpha) & d_p < 0 \end{cases} \tag{3.10}$$

where d_p denotes the disparity value of pixel p. σ_2 represents the range of the negative and positive disparity and α is the weight to control the weight of the negative disparity. For the negative disparity, if we directly use a comfort value to measure saliency, it may conflict with the pop-out effect. For example, if the pixel has a big negative disparity and is far away from the comfort zone, based on the pop-out effect, its saliency value becomes big; however, based on the comfort zone, its saliency value is small. In this case, it is hard to determine its saliency. To avoid the conflict of negative disparity with the pop-out effect, we set a weight value α to balance the comfort value of negative disparity. Similar to (Niu et al. 2012), we set $\alpha = 0.5$.

The center bias describes the saliency probability distribution in the X-axis and Y-axis. As many datasets have a property that locates the salient object or region in the center of the image (Borji et al. 2015), we use the center bias G to process the saliency map in the X-axis and Y-axis direction which can be modeled by Gaussian standard deviations, in general.

$$G(p) = \exp[-\frac{(x_z - u_x)^2}{2\sigma_x^2} - \frac{(y_z - u_y)^2}{2\sigma_y^2}] \tag{3.11}$$

where u_x and u_y denote the center of the image, we set $\sigma_x = 0.25 \times H$ and

$\sigma_y = 0.25 \times W$, where H and W represent the width and height, respectively, of the image.

The saliency map S is expressed as:

$$S = SC * BD * (G + Z) \tag{3.12}$$

The proposed model is based on the superpixels of SLIC. Since the content of each superpixel may have more than one object or texture, a single scale segmentation scheme is not suitable for objects of different sizes. We conduct multi-scale segmentation, based on controlling the numbers of superpixels in the SLIC. To combine the multi-scale saliency maps, we adopt the multi-scale integration proposed in (Li, Lu, Zhang, Ruan & Yang 2013). The complete saliency detection algorithm can be summarized as:

Table 3.1: The Pseudo-code

Algorithm: A preliminary saliency detection for stereoscopic images
Input: Left image and disparity map
Output: Stereoscopic saliency Map
1. SLIC segmentation, superpixel number $s = \{600, 800, 1000, 1200\}$
2. For each scale s
3. For each superpixel i
4. Local-global saliency: $SC(i)$, in Eq.3.3
5. Surrounded map from depth map: BD, in Eq.3.9
6. For each pixel p
7. Stereo center prior enhancement: $Z(p)$, $G(p)$, in Eq.3.10,3.11
8. After multi-scale fusion, the final stereo saliency map are computed: S in Eq.3.12

3.3 Experiments

In this section, we evaluate the performance of our proposed model on two eye-tracking datasets (Wang et al. 2013)(Lang, Nguyen, Katti, Yadati,

Kankanhalli & Yan 2012) including color and depth information. In Part A, we present the quantitative metrics of the evaluation for the proposed method. In Part B, we provide the performance evaluation by comparing the proposed methods and the state-of-the-art methods. All depth map are supplied by datasets, which is generated by the different depth-capture sensors.

3.3.1 Experimental setup

Our stereo saliency framework is based on the segmentation of SLIC. In the experiment, we set the superpixel count as $\{600, 800, 1000, 1200\}$. All the distance are normalized to [0, 1]. Based on the experiment, we set R=0.3. In K-means clustering of the surrounded enhancement, we set $K = 32$. Similar to (Zhang & Sclaroff 2013), we set the kernel ψ to 7 pixels. In stereo center prior enhancement, we set α as 0.3, similar to (Fang et al. 2013).

One eye-tracking database is published in the study to evaluate the performance of our method (Wang et al. 2013). It has 18 stereoscopic images of various types (e.g. outdoor scene, indoor scene, and scenes containing different numbers of objects). To avoid and minimize 3D fatigue from the conflict in different depth fields (for example, one object is seen by the right eye but missed by the left eye because it is blocked or obscured), in the eye-tracking experiment, the degree of vergence in human vision was considered within a stereoscopic 3D viewing environment. The disparity of the used stereoscopic images is computed and meets the requirements of the comfort viewing zone (Wang et al. 2013). The conflict in different depth fields is not detected by the observers in the eye tracking experiments. The gaze points are recorded by the eye-tracker and processed by a Gaussian kernel to form the fixation density maps, which are used as the ground-truth maps. The other eye-tracking database is proposed in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). It includes 600 stereoscopic images of various scenes. The depth maps are captured by Kinect and do post-process to obtain smooth depth maps. These images are diverse, with different numbers and sizes of

objects and degrees of interaction or activity depicted in the scene. The eye tracker is used to record the human fixation of 80 participants and processed by a Gaussian kernel to form the fixation density maps, which are used as the ground-truth maps.

For a quantitative evaluation of the performance of the proposed model, we conduct similar quantitative measure methods to the study in (Wang et al. 2013). The performance of the proposed model is evaluated by comparing the saliency map with the ground truth supplied by the database. As there are two images (the left and right images) for any stereoscopic image pair, we only use the saliency map of the left image for comparison, similar to the study in (Wang et al. 2013). The Area under the Receiver Operating Characteristics Curve (AUC) and Correlation Coefficient (CC) are used to measure the quantitative performance of the proposed saliency detection model. Of these measures, CC is calculated directly from the fixation density map and the predicted saliency map, while AUC is computed from the actual fixation density map and the predicted saliency map. We adopt these two measures to quantitatively compare the eye-tracking ground truth and predicted saliency map.

3.3.2 Experimental results and comparisons

In addition to comparisons with other state-of-the-art models, we also evaluate each component of the proposed model on the two eye-tracking databases. **Performance of each component**: we show the results of five components: local saliency (LS), global saliency (GS), local-global saliency (LGS), surrounded enhancement (SE) and stereo center prior enhancement (SCP). The local-global saliency extracts the contrast feature from the local and global range to form a low-level contrast map. Then, we use the weighted map from the Boolean map to increase the performance from the mid-level surrounded region. Lastly, stereo center prior enhancement is used to enhance the results from the high-level factors. The performance of each component is shown in Table I and Table II.

Table 3.2: Comparison between each component in database (Wang et al. 2013)

Component combination	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
LS	0.588	0.198
GS	0.648	0.257
LGS	0.674	0.293
SE	0.703	0.312
SCP	**0.838**	**0.597**

Table 3.3: Comparison between each component in database (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012)

Component combination	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
LS	0.588	0.198
GS	0.648	0.257
LGS	0.662	0.287
SE	0.719	0.308
SCP	**0.859**	**0.416**

In Table 3.1 and 3.2, we can see that the performance of "LGS" is improved, compared with "LS" and "GS" in AUC and CC. This is because the content of the images are complex. We should consider the saliency from local and global features. "SE" has some improvement in AUC and CC. This is because the surrounded region can leverage the saliency. "SCP" enhances the performance significantly because the high-level features, such as probability saliency distribution, is important for saliency analysis.

Comparisons with State-of-the-Art Methods: We compare the proposed stereo saliency detection framework with the best methods in (Wang et al. 2013) and (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). We firstly compare the performance of existing 2D saliency models: IT (Itti et al. 1998), AIM (Bruce & Tsotsos 2005*b*), SR (Hou & Zhang 2007) and GBVS (Harel et al. 2006). Then, we fuse these models with the depth saliency models proposed by (Chamaret et al. 2010) and (Wang et al. 2013). Lastly, we compare the proposed model with the 3D saliency models proposed by (Fang et al. 2013). Please note that the AUC and CC values of these two models are from the original paper.

It can be concluded from Table 3.4 that performance is not improved significantly using depth information as a weighted value (2D $\times$ Depth (chamaret)) in AUC and CC. We can see that directly using depth information as a weighted value for the stereo saliency analysis does achieve a good result. This is because the method does not consider the characteristics of depth information. In contrast, the performance of the 2D + Depth Contrast methods does increase compared with 2D $\times$ Depth (chamaret) in IT, AIM and SR, because it considers the characteristics of depth information. However, in relation to GBVS, neither 2D $\times$ Depth (chamaret) nor 2D + Depth Contrast improve the performance in AUC and CC. It is shown that two depth models are not suitable for the 2D model GBVS. When we design the stereoscopic saliency detection model like 2D+Depth, we should consider the relationship between the 2D model and the depth saliency map. DSM is the best of the three models in (Wang et al. 2013). The performance of our proposed

Table 3.4: Comparison between the proposed framework with others. DSM represents the depth saliency map in (Wang et al. 2013)

Model		AUC($\rightarrow 1$)	CC($\rightarrow 1$)
2D Model	IT	0.538	0.137
	AIM	0.638	0.326
	SR	0.63	0.291
	GBVS	0.809	0.54
2D $\times$ depth (chamaret)	IT $\times$ depth	0.54	0.137
	AIM $\times$ depth	0.636	0.299
	SR $\times$ depth	0.634	0.292
	GBVS $\times$ depth	0.771	0.515
2D + Depth Contrast	IT + Depth Contrast	0.596	0.211
	AIM + Depth Contrast	0.644	0.343
	SR + Depth Contrast	0.662	0.307
	GBVS + Depth Contrast	0.799	0.53
DSM (Wang et al. 2013)	Model 1	0.656	0.356
	Model 2	0.675	0.424
	Model 3	0.67	0.41
Stereo Model (Fang et al. 2013)		0.703	0.55
Our Model		**0.838**	**0.597**

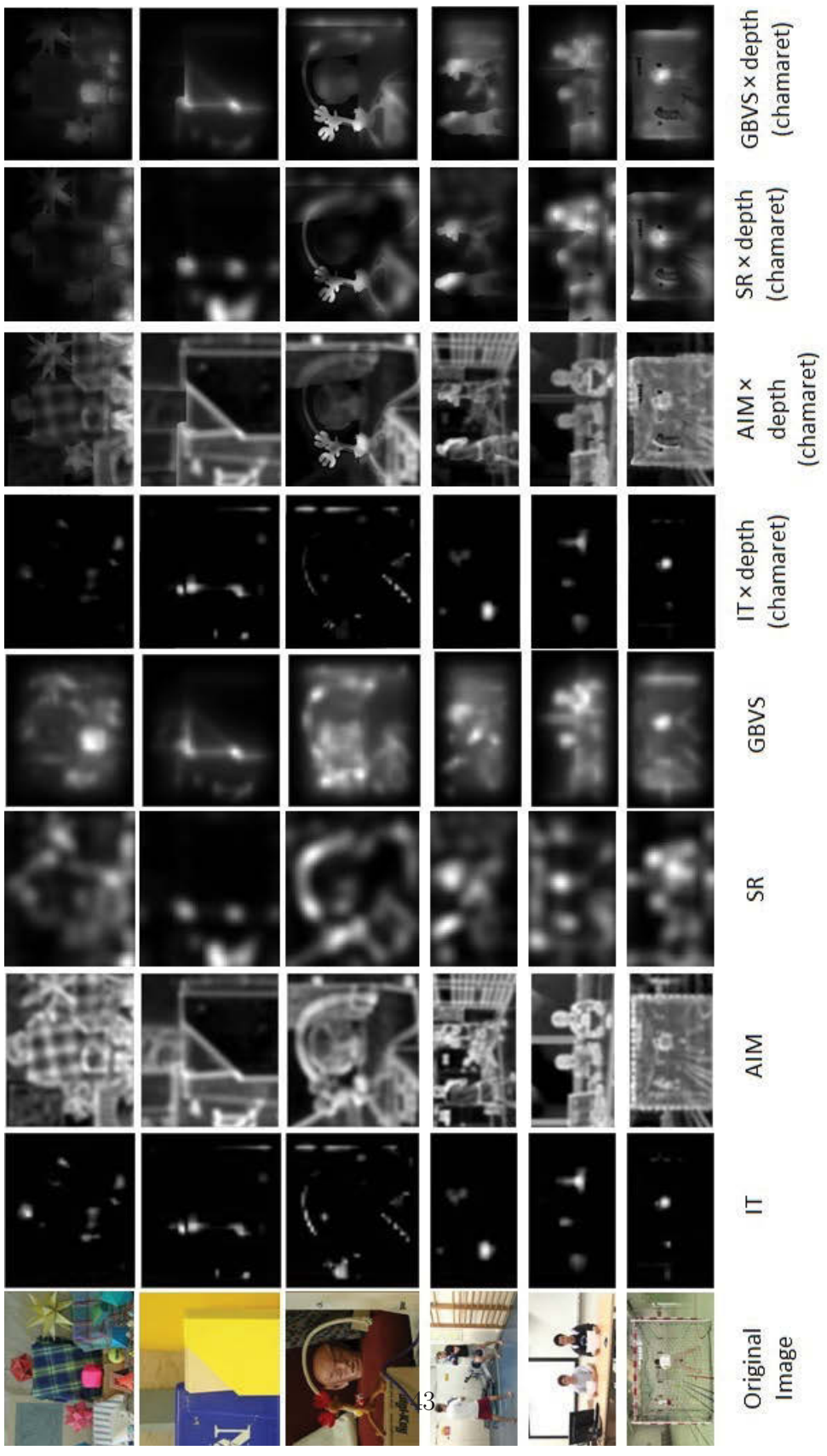

Figure 3.3: Visual comparison of various saliency detection models.

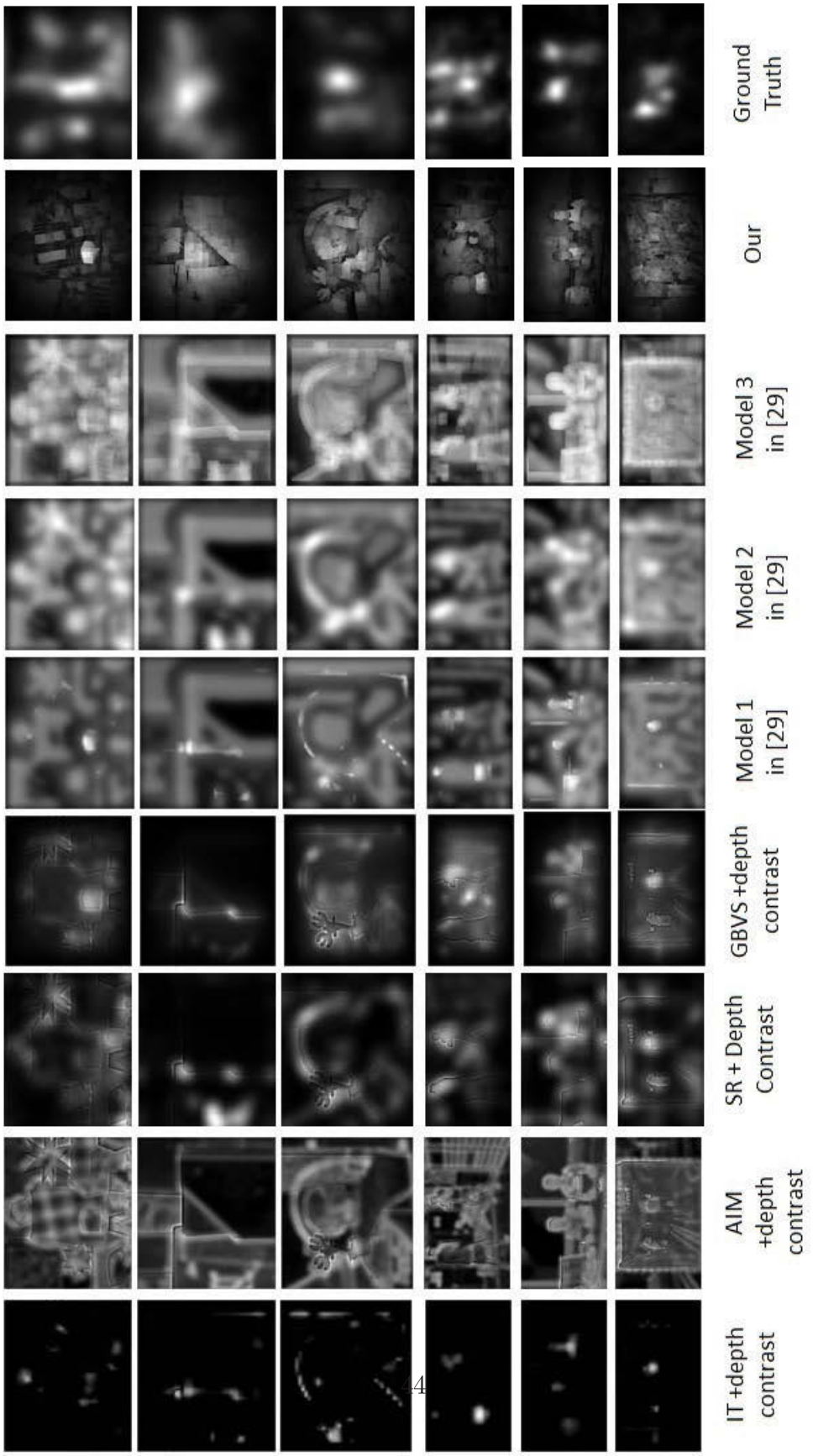

Figure 3.4: Visual comparison of various saliency detection models.

Table 3.5: Comparison between different 3D saliency detection models. "+" means the combination by simple summation by study (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). "×" means the combination by point-wise multiplication (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). DSM represents the depth saliency map in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012).

Component combination	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
IT+DSM	0.849	0.375
IT×DSM	0.854	0.398
GBVS+DSM	0.851	0.39
GBVS×DSM	0.855	0.413
AIM+DSM	0.85	0.342
AIM×DSM	0.85	0.391
FT+DSM	0.797	0.315
FT×DSM	0.745	0.268
SR+DSM	0.846	0.385
SR×DSM	0.808	0.325
Our Model	**0.859**	**0.416**

framework shows a significant increase over all the methods. The stereo model (Fang et al. 2013) extracts four features including color, luminance, texture and depth based on DCT coefficients to measure the saliency for image patches. Our model is based on superpixel and measures the saliency based on the color contrast. Then we use topological structural information computed from depth Booleaning Map and stereo center prior to enhance the results. Our model is sensitive to the boundary of the object. The depth information supplies the topological structural information and saliency probability distribution on Z-axix. Hence, our model has better performance then stereo saliency models (Fang et al. 2013) in CC and AUC. Fig. 3.33.4 shows some examples of all models. It is shown that IT, IT $\times$ Depth (chamaret) and IT + Depth Contrast mainly detect the contour of the saliency area in the images. The models related to AIM and SR detect some background areas as the saliency area in the images. In contrast, our stereo saliency detection framework estimates the saliency region accurately with regard to the ground truth map from the eye-tracking data.

Although the eye-tracking database (Wang et al. 2013) greatly assists in stereo saliency analysis, its samples (only 18 stereoscopic images) are insufficient to demonstrate the performance statistically. Hence, we use another published eye-tracking database (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012) with 600 diverse stereoscopic images, including outdoor and indoor scenes, to evaluate performance. As we could not find the code of the depth saliency map (DSM) in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012), we can only compare our results with the best methods listed in their original paper. The experimental results are shown in Table 3.5. Note that the AUC and CC values of other existing models are taken from the original paper. From this table, we see that the performance of our proposed model is the best of the 11 stereo saliency detection models.

3.4 Conclusion and discussion

In this Chapter, we proposed a preliminary saliency detection method for stereoscopic images, which is based on local-global contrast features, followed by surrounded region enhancement and stereo center prior enhancement. Experimental results show that our proposed saliency detection method achieves the best performance on two eye-tracking databases compared to existing methods. The proposed model is based on contrast and surrounded cues.

In the present study, our model still suffers from some limitations. The main one is that this model does not deeply explore the characteristics of pop-out effect and comfort zone. We use one model to describe the possible saliency distribution based on the pop-out effect and comfort zone, which will decrease the performance of the pop-out effect. Meanwhile, we only treated the depth information as a weight.

Chapter 4

Stereoscopic Visual Saliency Prediction Based on Stereo Contrast and Stereo Focus

The stereoscopic saliency detection model in the previous chapter only treats the depth information as a weight. It ignores the importance of the characteristics of the human visual system. In this chapter, we will deeply explore the characteristics of the human visual system: pop-out effect and comfort zone, which supply two important cues for stereoscopic saliency analysis. Based on them, we propose the stereoscopic visual saliency prediction.

4.1 Introduction

The models of visual attention are usually divided into two categories: bottom-up and top-down (Yarbus et al. 1967). The bottom-up approach is a rapid data-driven task-independent process and is usually feed-forward. A prototypical example of a bottom-up attention model is the act of looking at a scene which has only one horizontal bar among several vertical bars, in which attention is immediately drawn to the horizontal bar (Treisman & Gelade 1980). Top-down model considers high-level cognitive features to

quantify the visual saliency, such as human faces (Judd et al. 2009) and prior knowledge about the target (Frintrop et al. 2010). Of these top-down features, prior knowledge about the target is difficult to model. Recently, a number of saliency models have incorporated both top-down and bottom-up feature detection in an effort to improve prediction accuracy (Jiang et al. 2013). Wei *et al.* (Wei et al. 2013) turned to background priors to guide the generic object level saliency detection. Goferman *et al.* (Goferman et al. 2010) and Judd *et al.* (Judd et al. 2009) integrate high-level information, making their methods potentially suitable for specific tasks.

These models are mainly designed for 2D images. With the rapid development of 3D technology, many devices for stereoscopic capture have appeared. For example, the Panasonic 3D camera captures the stereoscopic images and video for 3D movies. The Kinect-1 device by Microsoft for the XBox captures both the color map and the depth map at the same time, which can generate the stereoscopic images (the depth map of the Kinect-1 may have holes that need to be smoothed (Camplani & Salgado 2012), which may cause noise). These devices make up a number of applications for 3D images or videos, such as 3D rendering (Chamaret et al. 2010), 3D visual quality assessment (Huynh-Thu et al. 2011), 3D video detection (Kim, Lee & Bovik 2014), and more. These 3D applications increase the need for saliency modeling for 3D visual content.

Stereo saliency models can be classified into two categories according to the way they use the depth factor: stereo-vision models and depth-saliency models.

Stereo-vision models take into account the mechanisms of stereoscopic perception in the human visual system (HVS). This type of model considers the characteristics of depth factors and color information. Bruce and Tsotsos extended the 2D model, which uses a visual pyramid processing architecture (Bruce & Tsotsos 2005*a*), by adding neuronal units to model the stereo vision; however, they did not propose a computational model in that study. Based on our knowledge, designing the stereo-vision model is a hard work and we

only find two models in (Wang et al. 2013), because the mechanisms of stereo vision still pose several research challenges, such as how to build then apply the model for the stereoscopic vision mechanism.

Depth-saliency models take depth saliency as a feature of saliency measurement, and methods of formulating and using depth saliency fall into two further categories. One category relies on a depth-saliency map (DSM)(Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). The depth saliency is extracted from the depth map or disparity map (usually based on depth contrast or the depth pop-out effect) to create an additional depth-saliency map. The final result combines the 2D saliency maps (from 2D saliency models usually using color contrast, intensity, or image texture) and the depth-saliency maps (DSM). The other category builds the model directly. In other words, it builds the stereoscopic visual saliency prediction model by taking the mechanisms of stereoscopic perception in the HVS into account. It designs the model by fusing the depth and 2D features into the saliency measurement, based on the mechanisms of the HVS (Fan et al. 2014).

Kim *et al.* (Kim et al. 2014) designed a stereoscopic visual attention algorithm for 3D video based on multiple perceptual stimuli, which assumes that pixels closer to observers and at the front of the screen are more salient. Niu *et al.* (Niu et al. 2012) explored stereo saliency by analyzing the characteristics of stereo vision and proposed a depth saliency model for a depth map that would expand the 2D saliency model for stereo saliency analysis. However, the proposed model does not fully explore the relationship between the depth model and the 2D saliency model. Fan *et al.* (Fan et al. 2014) proposed a stereo saliency model based on region-level depth, color and spatial information. Wang *et al.* (Wang et al. 2013) proposed a computational model that took the depth factors as an additional visual dimension and provided a public database with a ground truth of eye-tracking data. Fang *et al.* (Fang et al. 2013) proposed a visual attention model for stereoscopic images based on the feature contrast of low-level features. However, they did not consider the characteristics of human stereo vision, such as the pop-out

effect or 3D fatigue.

According to the above analysis, the key issue for a 3D visual saliency prediction model is how to adopt the depth factor and how to combine the depth factor with 2D information based on the mechanisms of HVS. In Chapter 3, a preliminary saliency model for stereoscopic images was proposed. However, this model did not deeply explore the HVS characteristics of pop-out effect and comfort zone and only treated the depth information as a weight. In this chapter, we deeply analyze two characteristics of the stereoscopic vision: pop-out effect and comfort zone. Based on two characteristics, we design two stereo-vision models for visual saliency prediction: one based on stereo contrast and the other based on stereo focus. We enhance these two models by clustering and then integrate them into the final stereoscopic saliency map.

The main contributions of this chapter are:

1. We propose a stereo contrast model for detecting stereo saliency. This model detects saliency based on color and depth contrast and the pop-out effect.
2. We propose a stereo focus model for detecting stereo saliency. This model detects the degree of focus via monocular focus and the comfort zone.
3. We propose an enhancement to increase the performance of the stereo contrast and stereo focus models.

The rest of the chapter is organized as follows: In Section 4.2, we introduce the two mechanisms of stereo human vision for stereo saliency analysis. Section 4.3 proposes a new stereo visual saliency prediction method based on the stereo contrast and stereo focus models. Section 4.4 describes a quantitative comparison of the proposed model and state-of-the-art algorithms. Section 4.5 provides the research outcomes and discussion.

4.2 Methodology

When watching a stereoscopic image, people experience different effects, such as the pop-out effect and deep-in effect (Beato 2011). When we watch the

stereoscopic image/video, the pop-out effect occurs when an object looks like it is going to pop out of the screen and the deep-in effect occurs when an object looks like it is behind the screen. To obtain these two effects, we can control the parallax of objects, such as the negative or positive parallax as in Fig.4.1. This finding is based on the research about human stereo vision (Zhang, An, Zhang & Shen 2010). These effects make viewers feel immersed in the image, which is the most attractive aspect of stereoscopic images. Moreover, studies show that an object that has the pop-out effect often catches a viewers attention (Häkkinen et al. 2010). This phenomenon provides a useful depth cue for stereo saliency analysis, since the object having a pop-out effect is usually more salient than objects that have a deep-in effect. We assume that the object having pop-out effect has more salient than other objects. In addition, we use color/depth contrast for the stereo saliency analysis. Hence, we propose a stereo contrast model to simulate the pop-out effect by combining the color/depth contrast and pop-out value.

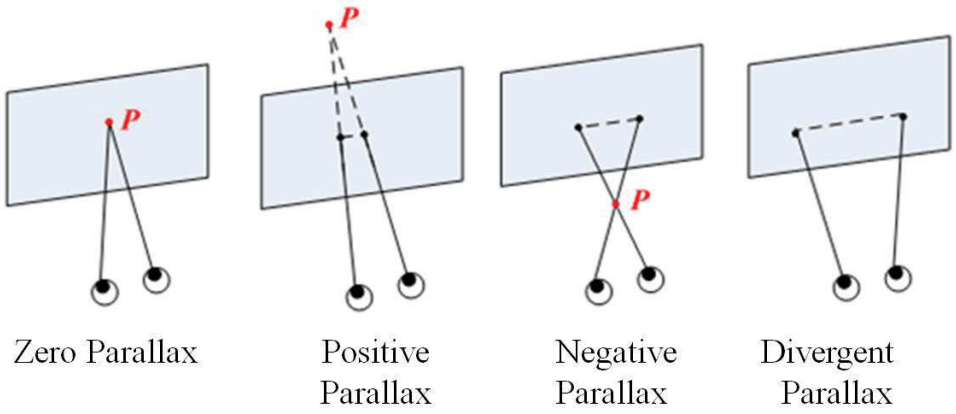

Figure 4.1: Stereo perception based on the different parallax

Another property of stereo vision is the viewing comfort zone based on binocular information. Viewers may experience fatigue when they spend a long time watching stereoscopic images or video. The reason may be accommodation-vergence conflict or too much divergence (Yano, Ide, Mit-

suhashi & Thwaites 2002). A good stereoscopic image needs to minimize 3D viewer fatigue. A common cause of 3D fatigue is the vergence-accommodation conflict (Mendiburu 2009). This conflict increases as the perceived depth of an object becomes further away from the screen, as shown in Fig. 4.2 (Niu et al. 2012). The zone close to the screen plane is called the comfort zone. Photographers usually make sure the more important objects are in the comfort zone when they capture a stereoscopic image or video. This is another depth cue for saliency analysis: the object in the comfort zone tends to be more salient than other zones. Studies show that the object near the zero disparity plane is more salient than those which are away from the zero disparity plane, which can be described by the linear formulation (Niu et al. 2012). When a person watches one salient object, this object should be in the focus region (Jiang et al. 2013). According to the above phenomenon, in the perspective of the comfort zone, this object should meet two conditions: one is that it is located in or near the comfort zone; the second is that it is in the focus region. Therefore, we use monocular focus and comfort zone to analyse stereo saliency. The monocular focus assumes that the salient object is usually located in the focus region. The comfort zone is treated as a weight to adjust the importance of the object located in focus region. The proposed stereo focus model is based on the comfort zone and monocular focus.

In order to describe the two mechanisms of the human visual system: pop-out effect and comfort zone, we have chosen to develop our proposed model on a combination of the stereo contrast and stereo focus models of the stereo-vision model. The stereo saliency of an object can be determined by the values calculated from the stereo contrast and stereo focus models. However, in some cases, the values obtained by these two models can be substantially different. For example, if an object has negative parallax and is far from the comfort zone, or if the object has zero parallax, the two values are quite different. To obtain the benefits from two models and detect the saliency for different stereoscopic content, our stereo visual saliency prediction model considers both the stereo contrast model and the stereo focus model.

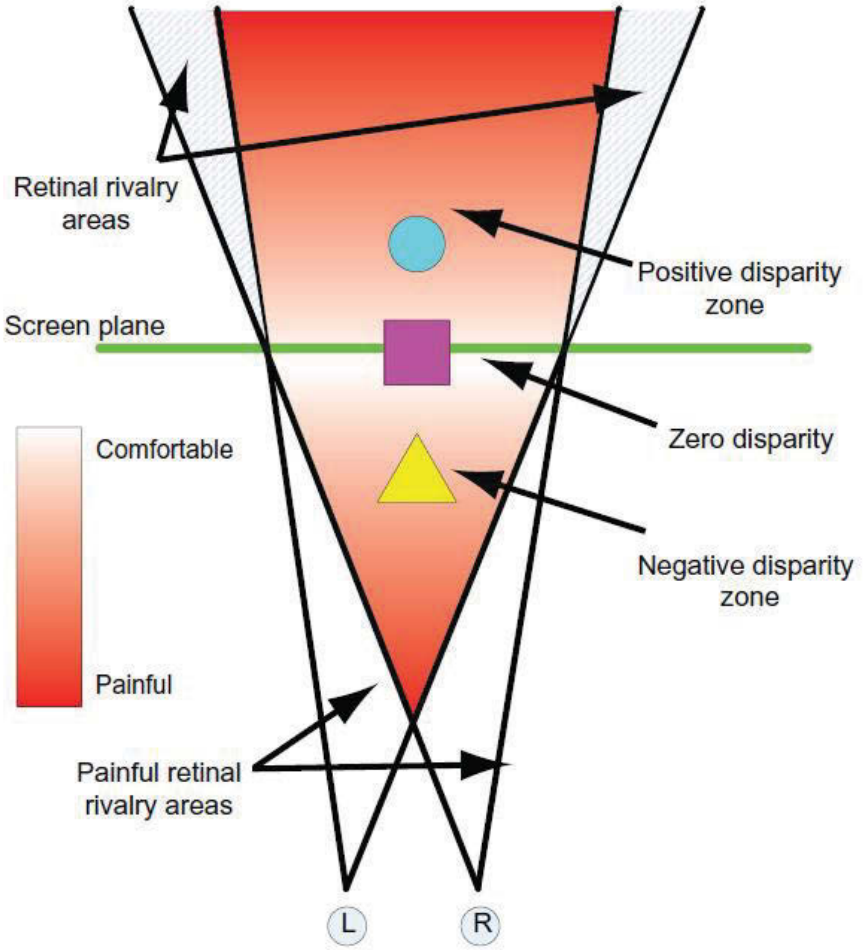

Figure 4.2: Stereo comfort zone based on human stereo vision

4.3 Proposed stereoscopic visual saliency prediction model

The proposed stereoscopic visual saliency prediction framework is shown in Fig.4.3. To capture the structural information of the stereoscopic image, we first adopt a simple linear iterative clustering (SLIC) algorithm (Achanta et al. 2012) for the segmentation. The SLIC algorithm can segment an input image (left image) into multiple uniform and compact superpixels. By controlling the number of superpixels in the SLIC algorithm, the image is segmented into multi-scale images. Then we calculate the saliency values individually by applying the stereo contrast and stereo focus models for each superpixel based on the left image and disparity map. An enhancement is based on clustering and increases the performance of two models according to the experiments. Multi-scale fusion is then used to form the pixel-level stereo contrast and stereo focus maps. Last, the two maps are integrated by Bayesian integration to form the final stereo saliency map.

4.3.1 Pre-processing

In this chapter, we convert the stereoscopic images from the RGB color space to the Hue-saturation-value (HSV) color space. Compared to the RGB color space, the HSV colour space is more consistent with the characteristics of human vision attention, and using it leads to a saliency value with higher accuracy (Mendiburu 2009).

As mentioned previously, we conduct multi-scale visual saliency prediction. Based on the number of superpixels, the input image (left image) is segmented into a set of non-overlapping superpixels in the scale s using the SLIC algorithm. s represents the scale of the segmentation. We chose the SLIC algorithm as the segmentation method because it is a fast and highly efficient segmentation algorithm that is sensitive to the boundary of the object (Lee & Song 2010). Each superpixel t is described by the mean color feature $\{H, S, V\}$, coordinates of the superpixels $\{x, y\}$, and the mean dis-

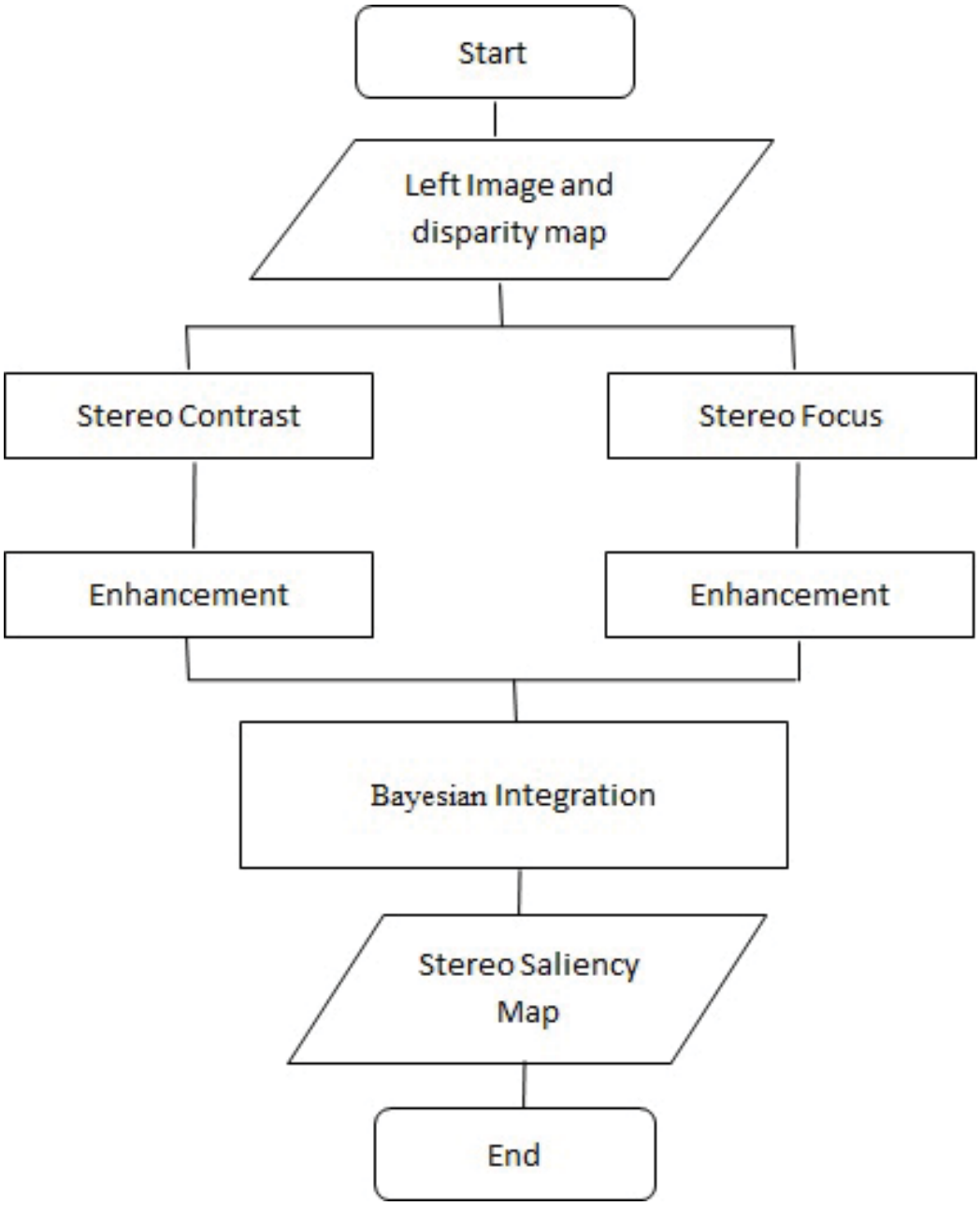

Figure 4.3: The framework of the proposed stereo saliency model

parity value d, $x_t = \{H, S, V, x, y, d\}_t$. The entire image can be represented as $X = [x_1, x_2, \ldots, x_N]_s$.

4.3.2 Stereo contrast model

We propose the stereo contrast model based on the color/depth contrast and the pop-out effect to calculate the saliency value (using disparity map to analyze the pop-out effect). According to the human vision system, the human attention is sensitive to contrast region that includes color contrast and depth contrast (Häkkinen et al. 2010). The colors of the salient region are distinctive and show contrast with the other regions. The depth discontinuity region may attract the viewers attention when view positions or angles are changed. Therefore, the distinctive region may attract the viewer's attention in color/depth information. According to (Einhäuser & König 2003)(Cheng et al. 2011), humans pay more attention to those image regions that contrast strongly with their surroundings. Based on our observation, the distance between neighboring regions and the area of the region plays an important role in human visual attention. To simulate the above mechanism, we define the contrast value to measure the contrast of stereoscopic information.

Let $DC(i, j)$ be the Euclidean distance between the vectorized superpixels i and j in HSV color space and $DD(i, j)$ be the Euclidean distance between superpixel i and j in disparity. DC and DD are normalized to the range $[0, 1]$. We define the contrast measure $C(i, j)$ between superpixel i and j as:

$$C(i, j) = (1 - a) * DC(i, j) + a * DD(i, j) \tag{4.1}$$

where a is a control weight to balance the color and disparity contrast. Although several approaches (Wang et al. 2013)(Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012)(Fang, Wang, Narwaria, Le Callet & Lin 2014) combining depth saliency maps with 2D visual features have been proposed, any specific and standardized approaches still lack the combination of saliency maps from depth with 2D visual features. Reference (Wang et al. 2013)(Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012) treats depth with the same

importance as color. Reference (Fang et al. 2014) uses the adaptive weight for color and depth. In our experiments, we adopt a straightforward approach to merge color and depth contrast, treating the depth contrast with the same importance as the color contrast. We set $a = 0.5$ empirically.

Let $L(i, j)$ be the Euclidean distance between the position of superpixel i and j normalized to the range $[0, 1]$. According to the analysis above, we define the stereo contrast measure $S(i, j)$ between a pair of superpixels i and j based on color, disparity and spatial information:

$$S(i,j) = (\frac{C(i,j)}{1 + c * L(i,j)}) * \omega_j \tag{4.2}$$

where ω_j is the number of pixels in superpixel j and c is a control value for spatial information ($c = 3$ in our implementation). As mentioned above, the saliency of a superpixel z can be defined by its stereo contrast measure as:

$$SC_R(z) = \sum_{i \neq z, i \in R} S(z,i) \tag{4.3}$$

where R is the range and $SC_R(z)$ is the saliency value of superpixel z in the range (All the distance is normalized to $[0, 1]$. We set $R = 0.3$ empirically). Fig.4.4 shows the global and local range. Then, we compute the global and local saliency maps.

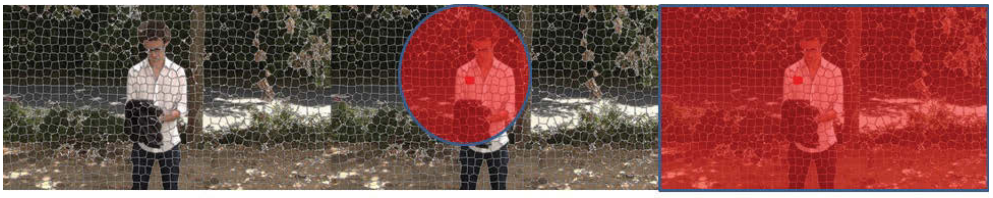

Figure 4.4: Global and local range.

When we compute the stereo contrast saliency value of the current superpixel, we do not compute all superpixels in the range. We only choose the K most similar superpixels in the range and use them to compute the stereo contrast saliency of the current superpixel. This is based on the experiments

and (Cheng, Zhang, An & Liu 2015), as using the k most similar superpixels to compute the stereo contrast can prevent the stereo contrast saliency value of an abnormal superpixel becoming too great. Therefore, in practice, to measure a superpixels stereo contrast,we simply consider the K most similar superpixels. If the most similar superpixels are extremely different from the current superpixel, clearly all image superpixels are extremely different from it. In other words, to measure a superpixels stereo contrast, there is no need to incorporate its stereo contrast value in all other superpixels in the range. We simply consider the K most similar superpixels. If the most similar superpixels are extremely different from the current superpixel, clearly all image superpixels are extremely different from it. Therefore, we search for the K most similar superpixels $k = \{1, 2, ..., K\}, k \in R$, where R is all superpixels in the range. Based on the observation of the experiments, we set K as 15 empirically. The local-global stereo contrast saliency of superpixel z is expressed as:

$$SC'(z) = \sum_{k=1, k \in R}^{K} S(z, k) \tag{4.4}$$

According to the pop-out effect in Section II, the region that has the pop-out effect may attract people's attention. Therefore the pop-out effect describes the importance of the superpixel in stereoscopic saliency analysis. We treat the pop-out effect as a weight to enhance the stereo contrast saliency. Based on the reference (Niu et al. 2012) and our experiments, the superpixel of the pop-out effect can be represented by an exponential function of the disparity. We use d to represent the disparity, and d_z is the mean disparity for superpixel z which is quantized to $[-1, +1]$. Let o be the pop-out value for superpixel z. If $d_z < 0$, it means that the superpixel has a pop-out effect. The saliency of this superpixel should increase. If $d_z > 0$, it means the superpixel has a deep-in effect and saliency should decrease. The

pop-out value can be expressed as follows:

$$o_z = 2^{-d_z} \tag{4.5}$$

We use the local-global stereo contrast and the pop-out value to simulate the pop-out effect. Fig.4.5 is an example of stereo contrast map. The stereo contrast $SC(z)$ relies on the color/depth contrast, distance contrast, superpixel area and pop-out value, which can be expressed as follows:

$$SC(z) = SC'(z) * o_z \tag{4.6}$$

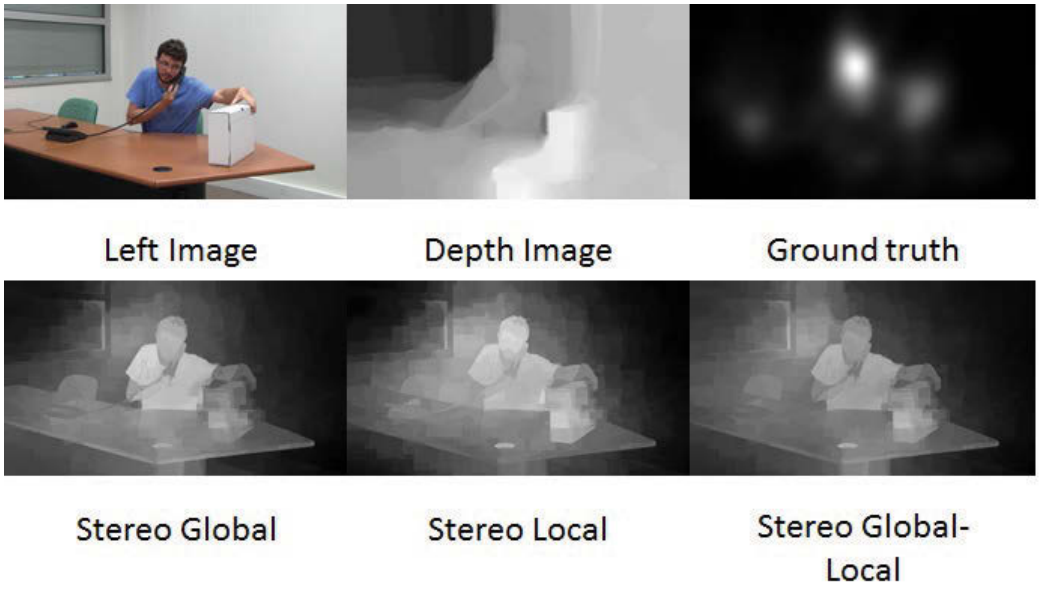

Figure 4.5: A example of stereo contrast map.

4.3.3 Stereo focus model

We propose a stereo focus model based on monocular focus and the comfort zone. According to the comfort zone as mentioned in Chapter 4-2, human visual attention can take the initiative to focus on the salient region by using two monocular focus. Mnocular focus can be detected by the focal blur (Elder & Zucker 1998) and we add the comfort zone to improve its accuracy.

For monocular focus, sharp edges of an object may be spatially blurred when projected on the image plane. The degree of blur model (Jiang et al. 2013) can measure the focus/defocus for the edges of the image by computing the Differential-of-Gaussian (DOG) operation in a different scale for the edge pixels. The monocular focus of the edge pixel p is $F_{2D}(p)$. This value is sensitive to the edge pixels and easy to implement. However, it is a 2D focus measure and only useful for the edge pixels of the image. For stereoscopic analysis, we expand this model to measure the edge of stereoscopic focus by combining the monocular focus and the comfort zone. Then we expand the stereoscopic focus model from edge to region.

According to our experiments, we use a comfort value to measure the comfort zone. The comfort value is a weight to indicate the objects importance by measuring the comfort zone. When multiple objects have zero or small disparity in the stereoscopic images and are located in the comfort zone, our observation is that their comfort values are similar. When they are far away from the zero disparity plane, their comfort values decrease sharply. Based on this observation, the comfort value complies with a Gaussian distribution. $v(p)$ denotes the comfort value of pixel p. This can be expressed as:

$$v(p) = \begin{cases} \exp(\frac{d_p^2}{-2\sigma_1^2}) & d_p \geq 0 \\ \alpha \cdot \exp(\frac{d_p^2}{-2\sigma_1^2}) + (1-\alpha) & d_p < 0 \end{cases} \tag{4.7}$$

where d_p represents the disparity of pixel p. σ_1 is the range of positive and negative disparity. α controls the weight of negative disparity. For negative disparity, we cannot directly follow the comfort zone model (Niu et al. 2012) to design our comfort value. The reason for this is that there is the conflict between the pop-out effect and comfort zone. If we directly use the comfort zone model (Niu et al. 2012) to measure saliency, in some cases, stereo contrast model and stereo focus model may give quite different results for an object with negative disparity, which will reduce the performance our proposed model. For example, if the pixel has a large negative disparity

and is far from the comfort value, its pop-out value becomes big, and its comfort value is small. After the fusion of two models, the results may be not reliable. To reduce the errors caused by such conflicts, we increase the importance of the negative disparity in the comfort zone by using to balance the comfort value of the negative disparity. There are two benefits in this modification. Firstly, this modification increases the importance of the pop-out effect for the object with the negative disparity. Secondly, it still keeps a high importance for the object in the comfort zone in stereoscopic saliency analysis. According to our experiments, our modification for the comfort zone works in most cases and improves the performance of the proposed model.

We set the comfort value as a weight, because the comfort value describes the importance of the stereo saliency analysis. We define the stereo focus value of the edge pixels p by combining the monocular focus value F_{2D} with the comfort value. It is expressed as:

$$F_{3D}(p) = F_{2D}(p) \cdot v(p) \tag{4.8}$$

It would be ideal to analyze the saliency for each object as a whole. However, it is difficult to segment an object accurately. Therefore, we compute the stereo saliency at the superpixel level instead. For each stereo focus value of the edge pixels, we filter by using a Gaussian kernel with σ, equal to 1 degree of visual angle. This processing can effectively reduce the noise, such as isolated point. The stereo focus value of superpixel t relies on the stereo focus degree of all its pixels. Further, our observation is that a region with a sharper boundary usually stands out as being more salient. We set the boundary sharpness as a weight value, which can be represented by the stereo focus value of the boundary pixels. The stereo focus value $SF(t)$ of superpixel t is formulated as:

$$SF(t) = \frac{1}{m} \sum_{p \in B_t} F_{3D}(p) \cdot \frac{1}{n} \sum_{q \in t} (F_{3D}(q)) \tag{4.9}$$

B_t represents all the edge pixels in superpixel t, m is the number of edge

pixels and n is the number of all the pixels in superpixel t. The first term on the right-hand side of Eq. 4.9 is the average value of the stereo focus value for all the edge pixels. The second term is the average value of the stereo focus value for all the pixels in superpixel t. The stereo focus model simulates the stereoscopic focus vision by combining the monocular focus and the comfort value. Fig.4.6 shows the examples of the stereo focus maps.

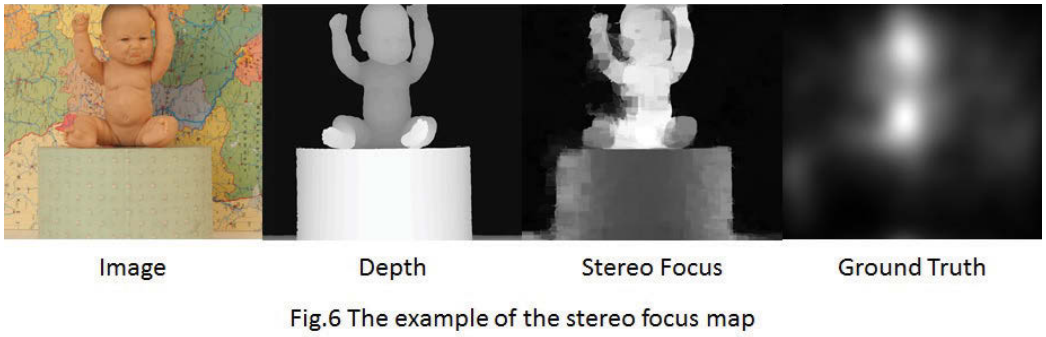

Figure 4.6: The examples of the stereo focus maps.

4.3.4 Enhancement

The stereo contrast model and stereo focus model are superpixel-level. To make the salient region more distinctive and separated easily, we propose an enhancement based on clustering for the two models. In practice, we use the k-means algorithm to cluster N superpixels to K clusters via the value of superpixel t. For simplicity, we use SV to represent SC and SF ($SV = SC = SF$). To enlarge the difference between neighboring clusters, each value of superpixel t belonging to cluster $k(k = 1, 2, 3 \ldots, K)$ is modified by considering its own value and the other superpixels in cluster k:

$$Sm(t) = \delta \sum_{i=1, k_i \neq t}^{N_c} r_{tk_i} SV_{k_i} + (1 - \delta) SV_t \tag{4.10}$$

where $\{k_1, k_2, \ldots, k_{N_c}\}$ denotes the N_c superpixels in cluster k and t is one superpixel in cluster k. δ is the weight parameter. $Sm(t)$ is the value of

superpixel t belonging to cluster k. r_{tk_i} is a weight value that relies on the value of superpixels t and k_i. The first term on the right-hand side of the equation is the weighted average of all the superpixels without superpixel t in cluster k, and the other is the weighted value of superpixel t. The weighted value is more sensitive to the spatial information of superpixel pairs:

$$r_{tk_i} = \frac{\exp \frac{SD(k_i,t)}{-\sigma_2^2}}{\sum_{i=1,k_i \neq t}^{N_c} \exp \frac{SD(k_i,t)}{-\sigma_2^2}} \tag{4.11}$$

$SD(k_i, t)$ is the spatial distance between the superpixels k_i and t. σ_2 is a weight to control the range of the spatial information. After re-calculating the value of each superpixel, the values of the important superpixels in cluster k are enhanced. Fig. 4.7 gives an example in which two maps computed by the stereo contrast and stereo focus models are processed by the enhancement.

Figure 4.7: An example of the proposed visual saliency prediction. (a) is the original left image and depth map. (b) shows the maps computed by the stereo contrast and stereo focus models. (c) shows the maps after clustering. (d) Final saliency map and ground truth.

Since the content of each superpixel may have more than one object or texture, a single scale segmentation scheme is not suitable for objects of different sizes. We conduct multi-scale segmentation based on controlling the number of superpixels in the SLIC algorithm. At each superpixel scale size layer, both the stereo contrast and stereo focus models are individually

applied to calculate their respective saliency values. A multi-scale pixel-level fusion is introduced to fuse the results for each model. Through this fusion, the saliency value for each pixel is calculated based on multi-scale saliency and its texture information.

To deal with the values in the different scales, we adopt (Li, Lu, Zhang, Ruan & Yang 2013) fusing the multi-scale layered value. This method considers the multi-scale value and its textural information, which uses the textural feature of the pixel and its corresponding superpixel as the weight value to average the multi-scale value. For each pixel, the saliency value relies on the saliency value of each scale and its corresponding weight. The weight considers the textural information that relies on the difference between the current pixel value and superpixel value.

4.3.5 Bayesian integration scheme

At this stage, two saliency maps have been built based on the stereo contrast and stereo focus model. The next step is to integrate them; however, as has been discussed (Gopalakrishnan et al. 2009), good individual saliency maps may become worse maps when they are combined by using weights. Therefore, we adopt a Bayesian model to integrate the two saliency maps (Lu et al. 2016). For the Bayesian model, each pixel's saliency can be estimated by the posterior probability. The Bayesian integration approach is suitable for dealing with two saliency maps. When we compute one saliency map, it treats the other saliency map as the prior while the current saliency map computes the likelihood. The specific steps are as follows: when we compute the saliency map $S_2^{'}$ based on the Bayesian formula, using one saliency map S_1 computes the prior probability and using the other saliency S_2 computes the likelihood. After that, we use the saliency maps in the formula in the opposite way. In other words, S_2 then computes the prior and S_1 computes the likelihood. In this way, the saliency map $S_1^{'}$ is computed. Finally, $S_1^{'}$ and $S_2^{'}$ are combined to obtain the final saliency map. Using this approach, it is possible to avoid reintroducing the noise in different saliency features, thereby

obtaining a more accurate posterior probability. This model is very robust with regards to various types of images. After Bayesian integration, we use center bias to conduct post-processing to obtain the final stereo saliency map, because many datasets place the salient object or region in the center of the image (Borji et al. 2015). Fig. 4.7 is an example of the saliency map after Bayesian integration and center bias.

The complete visual saliency prediction algorithm can be summarized as:

Table 4.1: The Pseudo-code

Algorithm: Stereo visual saliency prediction based on stereo contrast and stereo focus
Input: Left image and disparity map
Output: Saliency Map
1. Multi-scale segmentation, superpixel number $\{600, 800, 1000, 1200\}$
2. For each scale $X = [x_1, x_2, \ldots, x_N]_s$
3. For each superpixel $t = [H, S, V, x, y, z]_t$
4. Stereo contrast: $SC(t)$, in Eq.4.6
5. Enhancement for stereo contrast: $Sm\_c(t)$, in Eq.4.10
6. Stereo focus: $SF(t)$, in Eq.4.9
7. Enhancement for stereo contrast: $Sm\_f(t)$, in Eq.4.10
8. After multi-scale fusion, two pixel-level saliency maps are computed: S_i $(i = 1, 2)$
9. Bayesian integration scheme: $S(S_1, S_2)$

4.4 Results and discussion

In this section, we evaluate the performance of our proposed model on two eye-tracking datasets (Wang et al. 2013)(Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). One supplies high-quality stereoscopic images and the other supplies low-quality stereoscopic images generated by Kinect-1. First, we present the quantitative metrics of evaluation for the proposed

method in Section 4.4.1. To demonstrate the effect of the different component combinations of our algorithm, a performance comparison is given in Section 4.4.2. Last, we give a performance evaluation by comparing the proposed methods to state-of-the-art methods in Section 4.4.3.

4.4.1 Experimental setup

Our stereo saliency framework is based on the superpixel. In the experiment, we set the segmentation scale of superpixels in the SLIC algorithm. The number of superpixels were set as $\{600, 800, 1000, 1200\}$. The SLIC algorithm automatically adjusts the shape of each superpixel based on the segmentation scale and texture information of the image, which is sensitive to the boundary of the object. The main parameters of our proposed method are the number of clusters K and δ in Eq.4.10. In the experiment, we varied $K(K = 6, 8, 10, 12)$ and δ $(\delta = 0.4, 0.5, 0.6, 0.7)$, and observed that the saliency results were insensitive to either parameter. We set the number of clusters $K = 10$ and $\delta = 0.5$. The parameters of σ_1 and σ_2 are given in Eq.4.7 4.11 ,we differed these values to $[0.01, 3]$ and observed the saliency results. Then we set $\sigma_1^2 = 0.8$ and $\sigma_2^2 = 0.6$. In Eq. 4.7, α is set to $\alpha = 0.5$, same to (Cheng et al. 2015).

All depth map are supplied by datasets, which is generated by the different depth-capture sensors. We used one of the databases from (Wang et al. 2013). This database is consistent with the characteristics of the HVS, and includes 18 high-quality stereoscopic images of various types (e.g. indoor scenes, outdoor scenes, and scenes containing various numbers of objects). Some images in the database were collected from the Middlebury 2005/2006 dataset (Scharstein & Pal 2007), which has high accuracy depth maps, while others were produced from videos recorded using a Panasonic AG-3DA1 3D camera, which supplies high-quality left/right images. To avoid 3D fatigue resulting from conflict in the depth field (for example, one object is seen by the left eye but missed by the right eye), the degree of vergence in human vision was considered within the stereoscopic 3D viewing environment in this

eye-tracking experiment. The disparity of the stereoscopic images used is within the comfortable viewing zone. The conflict in different depth fields will not be detected by observers during the eye tracking experiments. The gaze points are recorded by the eye-tracker and processed by a Gaussian kernel to generate the fixation density maps, which are used as the ground-truth maps.

The other eye-tracking database was published in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). This database supplies low-quality stereoscopic images compared with (Wang et al. 2013) and has 600 stereoscopic images that include outdoor and indoor scenes. These stereoscopic images generated by Kinect-1 are diverse in term of the numbers and sizes of objects and the degree of interaction or activity depicted. The stereoscopic images only have a resolution of 640x480 and may have some noise because the depth map by the Kinect-1 has some holes and need to be smoothed. The stereoscopic image pair is produced by pre-processing, calibration and post-processing. The eye-tracking data are captured in both 2D and 3D free-viewing experiments by the eye-tracker from 80 participants (ranging in age from 20 to 33 years old). Human fixation maps are constructed from the fixations of viewers to globally represent the spatial distribution of human fixations. Then a Gaussian kernel is used to obtain the continuous fixation density maps as the ground-truth maps. This dataset supplies 2D and 3D fixation maps. To facilitate a comparison, we used 3D fixation maps as the stereoscopic 3D ground-truth maps.

To quantitatively evaluate the performance of the proposed model, we applied similar quantitative measuring methods to (Wang et al. 2013). The performance of the proposed model was measured by comparing the saliency map with the ground-truth map supplied by the database. Because there are two images (left and right) for any stereoscopic image pair, we used the saliency map of the left image for comparison (Wang et al. 2013). The area under the receiver operating characteristics curve (AUC) and the correlation coefficient (CC) were used to evaluate the quantitative performance of the

proposed stereo visual saliency prediction model. Of these measures, the AUC is the area under the receiver operating characteristics (ROC) curve (DAVID 1966). Using this score, human fixations were considered to be the positive set, and some points from the image were sampled to form the negative set. The saliency map S was then treated as a binary classifier to separate the positive samples from the negatives. By thresholding over the saliency map and plotting the true positive rate vs. the false positive rate, an ROC curve was generated for each image. Then, the ROC curves were averaged over all images and the area underneath the final ROC curve was calculated as the AUC (Bruce & Tsotsos 2006). Perfect prediction corresponds to a score of 1 while a score of 0.5 indicates a level of chance. To compute the AUC, each eye fixation density map and saliency map were normalized to [0,1]. In practice, we set different thresholds from [0.01, 1]. The CC measures the strength of a linear relationship between the predicted saliency map and the ground truth saliency map. When CC is close to $+1/1$ there is almost a perfectly linear relationship between the two variables.

4.4.2 Performance comparison with different combinations of components.

Four main components were compared: stereo contrast, stereo focus, and enhancement and integration via the Bayesian scheme. The performance of different combinations of components is shown in Table 4.1 and Table 4.2. SCM is the saliency map based on stereo contrast followed by multi-scale fusion. SFM is the saliency map based on stereo focus followed by multi-scale fusion. SCE is the saliency map based on stereo contrast followed by enhancement. SFE is the saliency map based on stereo focus, followed by the enhancement. OurWE is the proposed stereo saliency map without enhancement. Our model is the proposed stereo saliency map.

Table 4.1 indicates that SFM performs better than SCM on the database (Wang et al. 2013) in AUC and CC. Table 4.2 shows that SFM performs better than SCM on the database (Lang, Nguyen, Katti, Yadati, Kankanhalli

Table 4.2: Comparison between different component orders in database (Wang et al. 2013)

Different combinations of components	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
SCM	0.588	0.198
SFM	0.648	0.257
SCE	0.598	0.213
SFE	0.65	0.258
OurWE	0.864	0.557
Our model	**0.881**	**0.656**

Table 4.3: Comparison between different component orders in database (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012)

Different combinations of components	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
SCM	0.619	0.148
SFM	0.533	0.115
SCE	0.628	0.154
SFE	0.541	0.116
OurWE	0.849	0.37
Our model	**0.861**	**0.419**

& Yan 2012) with AUC and CC. The two models performed differently on each database, so using either one to form the saliency map would not result in good performance. Tables I and II show that the enhancement slightly improves the performance of the two models with AUC and CC. However, if we remove the enhancement from our proposed model, the performance of our model will be affected. In order to verify the improvement of the enhancement, we conduct a significance test for Our model and OurWe. For the dataset in (Wang et al. 2013), we use a paired-samples t-test to compare the average performance of our model with the average performance of the OurWE model. For AUC, the improvement of the enhancement is not significant ($t(18) = 1.61, P(T <= t) = 0.126, P < 0.05$). For CC, the improvement of the enhancement is significant ($t(18) = 3.09, P(T <= t) = 0.0067, P < 0.05$). For the dataset in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012), we use an ANOVA to compare the average performance of our model with the average performance of the OurWE model. The improvement of the enhancement is significant in AUC ($F = 14.89, P-value = 0.00012, P < 0.05$) and CC ($F = 114.948, P-value = 1.13E-25, P < 0.05$). According to the results of the significant test, we can see there are three positive results and one negative result. We believe that the enhancement can increase the performance of our proposed model slightly.

From Tables 4.1 and 4.2, we can see that the contribution of stereo focus varies. In Table 4.1, the stereo focus has a more important contribution than the stereo contrast because the objects of the stereoscopic image from the database in (Wang et al. 2013) lie in different focus regions and the stereo focus works more effectively. In Table 4.2, we can see that the contribution of the stereo focus is less than the stereo contrast because the content of the database in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012) is more sensitive to the color/depth contrast. Thus, to deal with these different types of stereoscopic images, we designed our model based on both stereo focus and stereo contrast. Fig.4.8 shows examples of the proposed visual saliency prediction. We notice that the small cap is not detected as a salient

region in the stereo focus model. The stereo focus is related to the monocular focus and comfort value. In this case, the zero disparity plane is at the big cap according to our comfort value. The monocular focus model detects the big cap as the focus region and the small cap is out of the focus region. Therefore, the salient region is the big cap region and the small cap is not the salient region in the monocular focus model. Even if we increase the weight of the comfort value (because the small cap is near the zero disparity plane and it pops out), it is not detected as the salient region according to the proposed stereo focus model. In stereo contrast model, the small cap is detected as the salient region because of the pop-out effect. Although the conflict between the stereo focus and stereo contrast still exists, our proposed model obtains the acceptable result that has the benefits from the stereo focus and stereo contrast models. This case shows that the stereo focus model may not work in the object with the negative disparity. For improving the performance of the proposed model, it is necessary to take the stereo contrast model into consideration.

4.4.3 Comparison of our proposed method with other methods.

First, we compared the proposed model with other state-of-the-art methods (Wang et al. 2013). We compare it with 2D saliency methods, the mixed models and stereoscopic 3D saliency models. The 2D saliency methods includes IT (Itti et al. 1998), AIM (Bruce & Tsotsos 2005*b*), SR (Hou & Zhang 2007) and GBVS (Harel et al. 2006) (denoted as 2D model in Table 4.4). The mixed model means combining these 2D models with the depth saliency models proposed by (Huynh-Thu et al. 2011) (denoted as 2D $\times$ depth (chamaret)) and (Wang et al. 2013) (which have two models denoted as 2D + depth contrast and 2D +DSM). We used a Bayesian integration (Lu et al. 2016) to process the 2D model and depth contrast saliency. For a fair comparison, we added center bias to process the results of the Bayesian integration. 2D +DSM has considered the center-surrounded mechanisms. We then compared our

Figure 4.8: An example of the proposed visual saliency prediction

proposed model with the stereoscopic 3D saliency model proposed by (Iatsun et al. 2015) (denoted as Stereo Model). We should note that the stereo model in (Iatsun et al. 2015) has already taken the center bias into consideration. From Table III we can see that the performance is not improved significantly using the depth information as a weighted value (2D $\times$ depth (chamaret)) in AUC and CC. Directly using depth information as a weighted value for the stereo saliency analysis does not achieve a good result because the method does not consider the actual characteristics of the depth information. By contrast, the performance of the 2D + DSM and 2D + depth contrast methods are better than the 2D $\times$ depth (chamaret), precisely because both consider the characteristics of the depth information. Bayesian integration and center bias do increase the performance compared with 2D + Depth Contrast methods. The performance of our proposed framework shows a significant increase over all the methods. The 2D + DSM method demonstrated the best performance against four comparative stereo saliency models. As shown in Fig 4.9, IT + DSM mainly detects the contour of the saliency area in the images. In AIM + DSM and SR + DSM, some background areas are detected as the saliency area in the images. By contrast, our stereo visual saliency prediction model estimates the saliency region accurately with regard to the ground truth map from the eye-tracking data.

Second, we used the published eye-tracking datasets in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012) with 600 3D images, including outdoor and indoor scenes, to evaluate performance. We used the 3D fixation maps as the ground-truth maps. Because we could not find the code of depth saliency map (DSM) in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012), we could only compare our results with the best methods listed in their original paper. The comparative model is DSM, and the 2D saliency modes are IT(Itti et al. 1998), AIM (Bruce & Tsotsos 2005*b*), FT(Achanta, Hemami, Estrada & Susstrunk 2009), GBVS(Harel et al. 2006), ICL(Hou & Zhang 2009), LSK(Seo & Milanfar 2009), LRR(Lang, Liu, Yu & Yan 2012). To compare the results of these models, we quantitatively evaluated their

Table 4.4: Comparison between the proposed framework with others. DSM represents the depth saliency map in (Wang et al. 2013)

	Model	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
2D Model	IT	0.538	0.137
	AIM	0.638	0.326
	SR	0.63	0.291
	GBVS	0.809	0.54
2D×Depth(Chamaret)	IT × depth	0.54	0.137
	AIM × depth	0.636	0.299
	SR × depth	0.634	0.292
	GBVS × depth	0.771	0.515
2D + Depth Contrast	IT + Depth Contrast	0.596	0.211
	AIM + Depth Contrast	0.644	0.343
	SR + Depth Contrast	0.662	0.307
	GBVS + Depth Contrast	0.799	0.53
Bayesian Integration	IT $\oplus$ Depth Contrast	0.668	0.254
	AIM $\oplus$ Depth Contrast	0.713	0.336
	SR $\oplus$ Depth Contrast	0.714	0.369
	GBVS $\oplus$ Depth Contrast	0.787	0.511
Center Bias	CB(IT $\oplus$ Depth Contrast)	0.798	0.547
	CB(AIM $\oplus$ Depth Contrast)	0.830	0.61
	CB(SR $\oplus$ Depth Contrast)	0.844	0.629
	CB(GBVS $\oplus$ Depth Contrast	0.856	0.632
2D + DSM	Model 1	0.656	0.356
	Model 2	0.675	0.424
	Model 3	0.67	0.41
Stereo Model	CB(CNSP)	0.79	0.48
	CB(CNMC)	0.78	0.63
	CB(GNLNS)	0.77	0.65
Our Model		**0.881**	**0.656**

Table 4.5: Comparison between different 3D viusal saliency prediction models. "+" means the combination by simple summation by study (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). "×" means the combination by point-wise multiplication (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). DSM represents the depth saliency map in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012).

Component combination	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
IT+DSM	0.849	0.375
IT×DSM	0.854	0.398
GBVS+DSM	0.851	0.39
GBVS×DSM	0.855	0.413
AIM+DSM	0.85	0.342
AIM×DSM	0.85	0.391
FT+DSM	0.797	0.315
FT×DSM	0.745	0.268
ICL+DSM	0.846	0.385
ICL×DSM	0.808	0.325
LSK+DSM	0.845	0.379
LSK×DSM	0.824	0.351
LRR+DSM	0.856	0.385
LRR×DSM	0.846	0.395
Our model	**0.861**	**0.419**

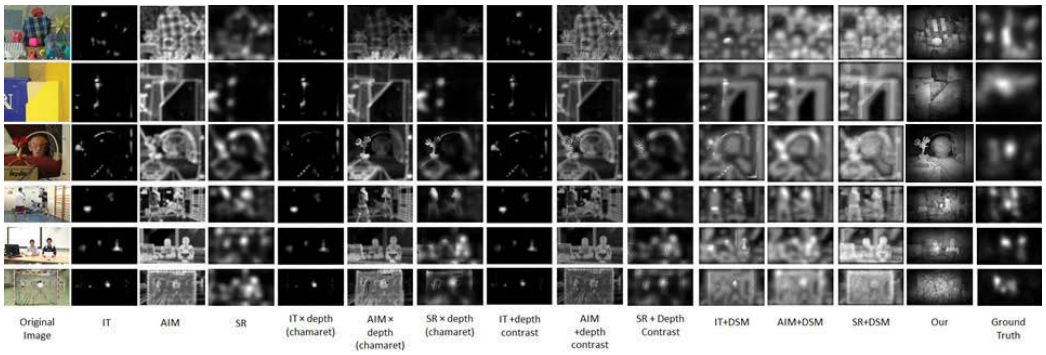

Figure 4.9: Stereo comfort zone based on human stereo vision. DSM represents the depth saliency map in (Wang et al. 2013)

.

performance on the database of the proposed method, using AUC and CC (Ouerhani, Von Wartburg, Hugli & Muri 2003). The experimental results are shown in Table IV. Note that the AUC and CC values of other existing models were taken from the original paper (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). From this table, we see that the performance of our proposed model is the best out of 15 stereo visual saliency prediction models. Here, we notice that our proposed model does slightly better than the GBVS×DSM. The reason is that sometimes the pop-out effect and comfort zone will fail because the salient region may be located at the backgrounds or near the background. Therefore, although the results of our proposed model are better than other existing models, it is not much better than GBVS×DSM.

4.5 Conclusion and discussion

The model in Chapter 3, we only use the combination of two characteristics of stereoscopic vision as an enhancement. It does not fully explore the

stereoscopic characteristics.

In this chapter, we explore two characteristics of stereoscopic vision and propose stereo visual saliency prediction based on stereo contrast and stereo focus. Stereo contrast is a product of color and depth contrast and the pop-out effect, which describes the contrast in objects. Stereo focus is based on the focus mechanism of human stereo vision, which describes the region of human focus. After adopting the stereo contrast and stereo focus models for the stereoscopic saliency analysis, the multi-scale fusion is used to form the respective maps of two models. Lastly, both saliency maps were integrated using Bayesian integration. Experimental results show that our proposed model can process stereoscopic images from different stereoscopic capture devices to achieve the best performance on two eye-tracking databases compared to existing methods.

In the present study, even if the performance of the proposed model is good, our model still suffers from some limitations. The main one is that in some cases, the pop-out effect and comfort zone may fail in stereoscopic saliency analysis. For example, if the salient region is located near the background, the performance of our model will decrease. The reason for this is that this case is not suitable for our assumption that the salient region should be located in the comfort zone or have the pop-out effect. In the next chapter, we will explore more mechanisms of HVS for saliency analysis. We try to find out how to deal with the conflict between pop-out effect and comfort zone, and how to improve the accuracy of the salient region if the pop-out effect and comfort zone are not working very well. Additionally, we will try more features to improve our proposed model in different color spaces.

Chapter 5

A Computational Model for Stereoscopic Visual Saliency Prediction

The characteristics of the human visual system can be used in stereoscopic saliency analysis, which is proved by the previous chapter. However, the proposed stereoscopic saliency model in Chapter 4 still suffers from some limitations. For example, in some cases, it may fail in stereoscopic saliency analysis. The reason is that the pop-out effect and comfort zone can not explain all phenomenon of stereoscopic saliency analysis. In this chapter, we propose a computation model for stereoscopic visual saliency prediction, which can overcome these limitations effectively.

5.1 Introduction

Research into 3D saliency models is still at an early stage compared to the significant progress made with 2D saliency models, and is generally considered to be an explored direction in HVS. In some cases, 3D saliency can be effectively detected by simply applying a 2D saliency model (Lang et al. 2010). However, 2D models ignore depth information, which plays an important

role in saliency analysis (Desingh, Krishna, Rajan & Jawahar 2013)(Peng et al. 2014). Some researchers have proposed using depth contrast to analyze stereoscopic saliency (Song, Liu, Du, Sun, Le Meur & Ren 2017)(Wang et al. 2013). However, these models do not consider the characteristics of human stereo vision; they merely treat depth information as a fourth channel. Stereopsis refers to the perception of depth through retinal disparity, which occurs when two slightly different images are projected onto the retina of each eye in binocular vision (Poggio & Poggio 1984). Depth information, also known as disparity or a depth map, often enables people to distinguish objects from a background with similar visual attributes. Several characteristics of HVS were discovered by analyzing depth information, including the pop-out effect and the comfort zone (Beato 2011)(Lambooij, Fortuin, Heynderickx & IJsselsteijn 2009). Recently, these characteristics have been used to improve the accuracy of stereoscopic saliency analysis and prediction (Niu et al. 2012)(Lin, Lin, Zhao, Xiao & Tillo 2015). Additionally, in some cases, the salient object may be located in or near the background region, which can not be explained by the pop-out effect and comfort zone (Cheng et al. 2017) (in this chapter, we call it the background effect).

Our approach capitalizes on these recent developments. First, we analyze the pop-out effect, the comfort zone, and the background effect (hereafter referred to as PE-CZ-BE) to deeply exploit the depth information in images. Then, we use a control function to govern each mechanism respectively. Finally, we propose a strategy for selecting the most suitable models based on the content of the image.

The main contribution of this chapter follow:

1. We propose three models to describe the PE-CZ-BE mechanisms with the control functions to enhance these models.
2. We propose a strategy for choosing the most suitable models based on the content of the image.

The rest of the chapter is organized as follows: Section II introduces related work. Section III proposes a new stereoscopic visual saliency predic-

tion method. Section IV presents a quantitative comparison of the proposed model and the state-of-the-art algorithms. Section V concludes this chapter.

5.2 Related work

Saliency models for stereoscopic images can be divided into two categories based on their use of the depth information (Cheng et al. 2017). One category relies on a depth saliency map (DSM-based model). It extends 2D saliency models into 3D models using the depth saliency map (DSM). Ramenahalli *et al.* (Ramenahalli & Niebur 2013) extended Itti's model to treat depth information as an additional channel. This model uses color, intensity, orientation, and depth channels to generate a 3D saliency map. The characteristics of stereoscopic vision, such as comfort zone and the pop-out effect (Häkkinen et al. 2010), are not considered in this model. However, Niu *et al.* (Niu et al. 2012) did incorporate these features into their saliency model, although their method might not produce optimal results if the salient region is located in the background. In our earlier work (Cheng et al. 2015), we proposed a preliminary saliency model for stereoscopic images that considers depth bias, based on the center bias, the pop-out effect, and the comfort zone. However, this model only treats depth information as a weight and does not consider the relationship between the pop-out effect and the comfort zone. Therefore, developing a suitable mechanism for stereoscopic saliency analysis is essential.

The second category considers the relationship between the depth feature and the other 2D features. This type of stereoscopic saliency model is usually based on multiple features (multi-feature models), such as depth, contrast, shape, and spatial information. Fan *et al.* (Fan et al. 2014) proposed using region-level depth, color, and spatial information to measure saliency. Peng *et al.* (Peng et al. 2014) proposed an RGB-D model based on both depth and appearance cues derived from color and depth contrast features. While multi-feature analysis has been shown to improve the performance of stereo-

scopic saliency prediction models (Bruce & Tsotsos 2005*a*), it ignores the mechanisms of HVS.

Neither of these categories fully explains all the phenomena of human visual attention. DSM-based models efficiently extract the individual 2D features and depth information for saliency detection but ignore the relationships between depth and the other features. Multi-feature models consider depth and other features but choosing and combining suitable features is a difficult task. As a result, some researchers have designed stereoscopic saliency models that fuse more than one model to increase prediction performance. Iana *et al.* (Iatsun et al. 2015) proposed a stereoscopic saliency model by considering two spatial saliency models one based on points of interest, the other based on depth-color saliency. Jiang *et al.*(Jiang et al. 2015) saliency model fuses three models: a 2D saliency model, a depth saliency model, and a visual comfort saliency model. The features extracted by the different models reflect different cues for saliency detection, and performance is improved by fusing the different features (Awh & Pashler 2000). However, none of these models consider background effect.

The work of these researchers does show that HVS mechanisms can improve the accuracy of stereoscopic saliency predictions. The key points are how to build the model based on these mechanisms and how to use them for stereoscopic saliency analysis. Some of the above models incorporate the pop-out effect and the comfort zone (Niu et al. 2012)(Cheng et al. 2015) but do not consider the conflicts between these features (Cheng et al. 2017). Comfort zone models assume that objects in the comfort zone are more salient than objects in other zones (Niu et al. 2012)(Jiang et al. 2015), and the pop-out effect assumes that objects with a pop-out effect are usually more salient (Zhang, Jiang, Yu, Chen & Dai 2010)(Kobyshev, Riemenschneider, Bódis-Szomorú & Van Gool 2016). In some situations, combining both these concepts of salience can create conflicts. If an object with a pop-out effect is far from the comfort zone, it will have low salience even though its salience should be high. Our comfort zone model considers such conflicts. Addition-

ally, pop-out effect and comfort zone do not fully explain all the phenomena in human visual attention (Cheng et al. 2017). For example, if a salient object is located in or near a background region, relying solely on pop-out effect and comfort zone may fail to distinguish the object as it conforms to neither of these features. The reason is that the furthest object receives a few more fixations than the one or two objects in front of it (Wang et al. 2012)(Wang et al. 2013). In this chapter, we define this phenomenon as the background effect. The background effect can explain salience in a way that the pop-out effect and comfort zone cannot. Thus, our stereoscopic saliency model also takes the background effect into account.

To address these two shortcomings in existing models, we based our stereoscopic saliency model on the pop-out effect, comfort zone, and background effect. Our model actually comprises three models, each describing one aspect of saliency distribution, The comfort zone model considers potential conflicts between the pop-out effect and comfort zone. The control function is used to adjust the three models independently. The relationship between the three models is not mutually exclusive. One, two, or three models may appear in one image. To accurately determine which mechanism the image conforms to, we have devised a selection strategy that chooses the appropriate combination of models based on the content of the image. Our approach is implemented within a framework based on multi-feature analysis. The framework considers surrounding regions, contrast, and points of interest to further enhance prediction. A series of experiments on two recent eye-tracking datasets show that our proposed method outperforms several state-of-the-art saliency models.

5.3 The proposed computational model for stereoscopic visual saliency

This chapter presents a computational model for stereoscopic saliency analysis. This model considers three phenomena of human vision, the pop-out

effect, the comfort zone, and the background effect, to supply possible distributions of the salient region. Our approach combines a multi-feature stereoscopic visual saliency model as input with the proposed model to generate a stereoscopic saliency map that is then used to predict salient objects.

5.3.1 PE-CZ-BE mechanisms

The PE-CZ-BE mechanisms provide useful cues for stereoscopic saliency analysis.

Objects with the pop-out effect look like they are going to pop out of the screen and the deep-in effect occurs when an object looks like it is behind the screen (Zhang, An, Zhang & Shen 2010). When watching a stereoscopic image, viewers feel immersed in the scene because of these effects. Current studies show that objects with the pop-out effect catch viewers attention more than objects with the deep-in effect (Sheng, Liu & Zhang 2016), which provides a useful cue for stereoscopic saliency analysis. We refer to this phenomenon as PE.

Further, viewers may experience fatigue or feel uncomfortable when watching stereoscopic images or video for long periods of time. This has been attributed to accommodation-vergence conflict or too much divergence (Chang, Hsueh, Tung, Jhou & Lin 2016), and comfort zones are thought to minimize 3D viewer fatigue (Jang, Park, Lee, Han, Donghyun & Song 2017). The zone close to the zero-disparity plane (where the disparity of the pixels in the zero-disparity plane is zero) is called the comfort zone. Objects located in the comfort zone can be observed for long periods without causing fatigue. By contrast, objects that are far away from the comfort zone may cause viewer fatigue or discomfort. This phenomenon also provides a useful cue for stereo saliency analysis: objects close to the zero-disparity plane tend to be more salient than the objects far away from the zero-disparity plane (Niu et al. 2012). This phenomenon is referred to as CZ.

The pop-out effect and the comfort zone are able to explain most of the phenomena in human visual attention. However, in some cases, the furthest

object may receive a few more attention than the one or two objects in front of it (Wang et al. 2012)(Wang et al. 2013). For example, if the salient region is located near or in the background, the pop-out effect and comfort zone do not produce good results according to our experiments. This phenomenon is referred to as the background effect (BE), and it supplies another useful cue: objects with a background effect may be more salient than the objects in front.

The relationships between these three phenomena are not mutually exclusive; an image may include one or more of these phenomena, as shown in Fig.5.1. Therefore, to accurately determine the most suitable phenomena for identifying the salient object(s) in the image belongs to, a selection strategy based on the content of the image is needed.

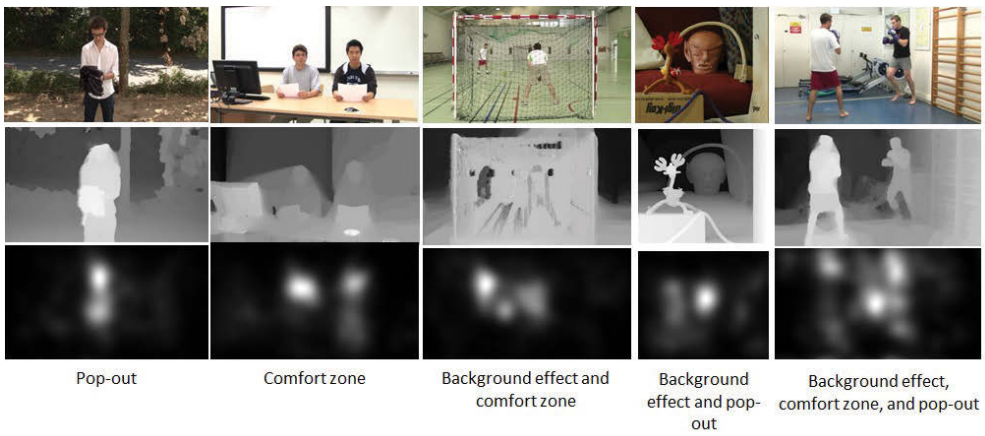

Figure 5.1: Examples of the different combinations of the mechanisms. The first row depicts the left version of several stereoscopic images. The second row shows the corresponding depth maps. The last row shows the ground truths.

5.3.2 The three modules based on the PE-CZ-BE mechanisms

As mentioned above, a region with a pop-out effect (PE) is likely to attract attention. Therefore, a pop-out effect suggests an important object in stereoscopic saliency analysis. We treat the pop-out effect as a weight to enhance the objects saliency. Based on the work in (Niu et al. 2012) and our experiments, the pixels that make up the pop-out effect can be represented using an exponential function of their disparity. We use d_p to represent the disparity of pixel p, which is normalized to $[-1, +1]$. Let PE' be the pop-out value for pixel p. If $d < 0$, the pixel has a pop-out effect. The saliency of this pixel should be increased and, if $d > 0$, it means the pixel has a deep-in effect and saliency should be lowered. The pop-out value can be expressed as

$$PE'(p) = \exp(-d_p) \tag{5.1}$$

The comfort value CZ' is used to measure the comfort zone (CZ). The comfort value is a weight indicating the objects importance by measuring the comfort zone. Through observation, we find that when multiple objects in a stereoscopic image have zero or a small disparity and are located in the comfort zone, they have similar comfort values. As their distance to the zero-disparity plane increases, comfort values decrease sharply. Based on this observation, the comfort value complies with a Gaussian distribution, which can be expressed as

$$CZ'(p) = \begin{cases} \exp(\frac{d_p^2}{-2\sigma^2}) & d_p \geq 0 \\ \alpha \cdot \exp(\frac{d_p^2}{-2\sigma^2}) + (1-\alpha) & d_p < 0 \end{cases} \tag{5.2}$$

where σ is the range of positive and negative disparity. We use α to control the weight of the negative disparity.

We should notice that for reducing the conflict between pop-out effect and comfort zone, we can not directly use the comfort zone model (Niu et al. 2012) to design our comfort value. This is because, in some cases, the pop-out value and the comfort value may produce very different results for an object with

negative disparity, which reduces the performance of the model. For example, a pixel with a large negative disparity that is far from the comfort value returns a high pop-out value and a low comfort value. Hence, the results may be not reliable when the two models are combined. To reduce such errors, we use α to balance the negative disparity and increase the importance of the comfort value. There are two benefits to this modification. First, it increases the importance of the pop-out effect for objects with negative disparity. Second, it maintains high importance for objects in the comfort zone. According to our experiments, this modification works, in most cases, and improves the performance of the proposed model.

According to the description of BE and our experiments, we believe the background effect as a replica of the pop-out effect. Denoted as BE', the background value can be represented by an exponential function of disparity:

$$BE'(p) = \exp(d_p) \tag{5.3}$$

5.3.3 Control function

The pop-out, comfort, and background values describe the possible distribution of the stereoscopic saliency regions based on the PE-CZ-BE mechanisms. However, directly using these values to influence the input saliency map may produce poor results as they also might enhance non-salient regions which would decrease prediction accuracy. For example, using the background value to enhance an image with a salient region located near a background region would also enhance the non-salient regions of the background, decreasing the accuracy of the final prediction. Therefore, non-salient regions are not considered when applying these three values. Instead, a control function G overcomes this problem. The three benefits of this control function are: it does not enhance the non-saliency region (does nothing to the non-saliency region); it can assist the three mechanisms in enhancing the saliency region based on their importance; and it removes ambiguity.

For simplicity, we use M to represent the value $M = PE' = CZ' = BE'$.

The input saliency map is denoted as S_{input}. A threshold value ts is used to classify the salient and non-salient regions. In practice, ts is the mean value of S_{input}. If the current saliency value of pixel $S_{input}(p)$ is less than ts, $G = M^{-1}$, which means nothing happens. If $S_{input}(p) >= ts$, enhancements are conducted based on importance. The control function can be expressed as

$$G(p) = \begin{cases} M^{-1} & S_{input}(p) < ts \\ 1 + \beta * W(p) * H(p) & \text{Other} \end{cases} \tag{5.4}$$

where β is a weight to control the magnitude of the enhancement; in practice, $\beta = 0.5$. W is a window function used to reduce noise. It describes the relationship between the nearby pixel and the current pixel. The saliency map according to the threshold ts is processed through binarization and the number of positive pixels (1) in the range r are counted. W is the percentage of the number of positive values and all pixels in the range r. In practice, all distance is normalized to $[0, 1]$ and we set $r = 0.1$ empirically. H describes the importance of the current pixel, which can be expressed as $H(p) = \exp(S_{input}(p))$.

After adding the control function, the pop-out value (PE), comfort value (CZ), and background value (BE) are changed to

$$\begin{aligned} PE &= PE'(p) * G(p) \\ CZ &= CZ'(p) * G(p) \\ BE &= BE'(p) * G(p) \end{aligned} \tag{5.5}$$

5.3.4 The selection strategy

As shown in Fig.5.1, each of the three mechanisms are not mutually exclusive. Depending on the image, a combination of each of the three values may be required to produce accurate predictions. Therefore, a selection strategy to determine the best combination is required. The proposed stereoscopic saliency model is illustrated in Fig.5.2.

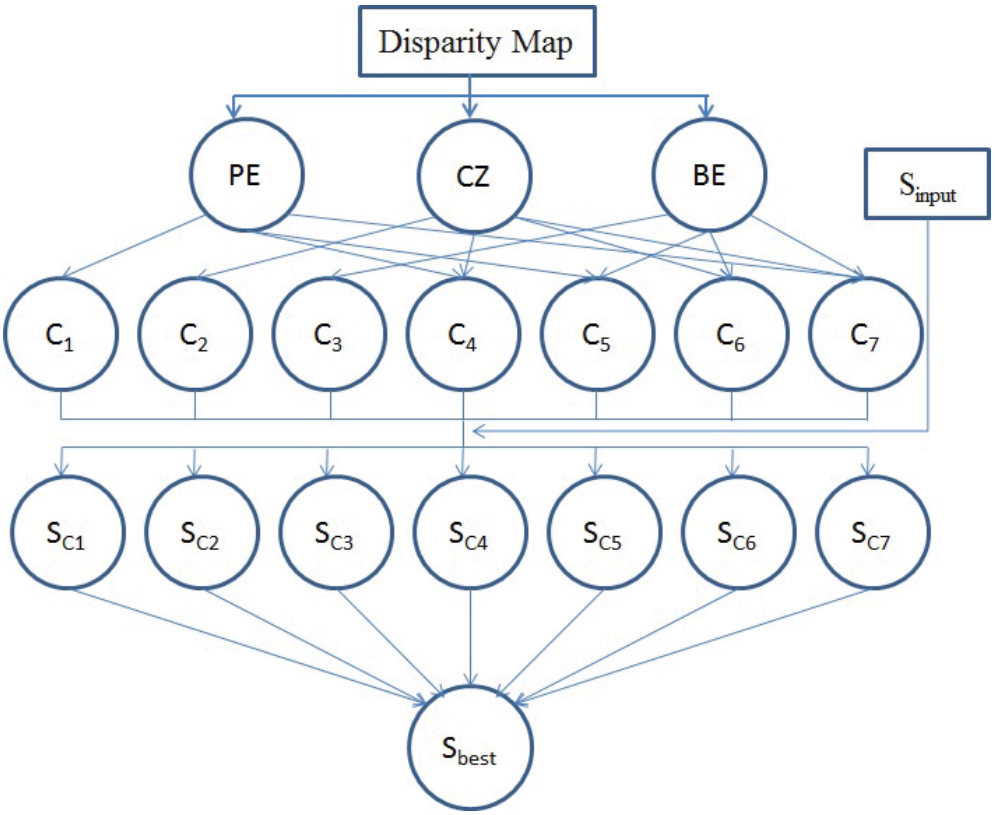

Figure 5.2: Flow chart of the stereoscopic saliency model

First, the three values PE, CZ, BE are computed by the disparity map. We define C as the combination of the three values, which can be expressed as:

$$C = \gamma_1 * PE + \gamma_2 * CZ + \gamma_3 * BE \tag{5.6}$$

where $\gamma_1, \gamma_2, \gamma_3$ are Boolean values. 0 means the corresponding value does not work in the current image. 1 means the corresponding value works. There are eight total combinations. However, the combination $\gamma_1 = \gamma_2 = \gamma_3 = 0$ (C_0) is not considered because this would mean that no mechanism works, as in the cases such as monochrome images, which are not common. In practice, only the remaining seven combinations $(C_1, C_2, ..., C_7)$ are considered.

Second, the combinations of C provide for all possible distributions of the saliency values based on the PE-CZ-BE mechanisms. However, a saliency map S_{input} is needed to compute all possible stereoscopic saliency maps. S_{Ci} represents the corresponding stereoscopic saliency map for each combination of C_i, $i = \{1, 2, ..., 7\}$. This is calculated from the combination C_i and the saliency map S_{input}, which can be expressed as

$$S_{Ci} = C_i * S_{input} \tag{5.7}$$

Lastly, the best saliency map is selected from among all the stereoscopic saliency maps $(S_{C1}, S_{C2}, ..., S_{C7})$. As mentioned, the pop-out, comfort, and background values are only applied to possibly salient regions and do nothing in non-salient regions. Similarly, these combinations are also only applied to possibly salient regions. The greater the similarity between the combination and the corresponding stereoscopic saliency map, the more effective the combination is. Therefore, the most similar map is selected as the final stereoscopic saliency map S_{best} based on a measure of the similarity between C_i and S_{Ci}, $i = \{1, 2, ..., 7\}$. In practice, we use the expectation $E_i(C_i, S_{Ci})$ to measure the similarity between the combination C_i and the stereoscopic saliency map S_{Ci}. The saliency map S_{best} with the least expectation is selected as the final stereoscopic saliency map.

5.3.5 Framework based on a multi-feature saliency model

As outlined in Fig.5.2, we need to construct a saliency map as the input for the stereoscopic saliency model. We propose a framework based on four features that reflect three different aspects of saliency: the surrounding region, contrast, and the point of interest. The main steps are shown in Fig.5.3. The surrounding region is computed according to an attention map based on Boolean map theory, which considers the structure of the salient region (Huang & Pashler 2007). Contrast comprises color and depth contrast maps, which are derived by assuming the salient region is unique (Cheng et al. 2015). The point of interest (IP) is reflective of the distribution of the gaze point (GP), which is processed as the saliency maps ground truth (Nauge, Larabi & Fernandez-Maloigne 2012).

A fusion strategy based on the importance of the different features is used to generate the saliency map from the three feature maps above:

$$S_{input} = Fusion(S_{BM}, S_{Col}, S_{Dep}, S_{IP}) \tag{5.8}$$

where S_{BM} represents the saliency map of the surrounding region, S_{Col} and S_{Dep} are the color and depth contrast saliency maps, and S_{IP} is the saliency map of the point of interest. These four features are combined through the fusion function to generate the stereoscopic saliency map S_{input}.

The attention map based on Boolean Map theory

The attention map reflects the surrounding region based on a topological analysis of a Boolean map. These maps are sensitive to the boundaries of objects, but they may not outline an entire object. Fig.5.4 shows an example of some saliency maps based on this Boolean map approach. Figs.5.4(a) and (b) show the left version of the stereoscopic image. In Fig.5.4(d), the human body is not recognized as a whole region. The reason for this is that the colors of the body are different and the surrounding foreground of the Boolean map does not cover the whole body. To overcome this issue, we use

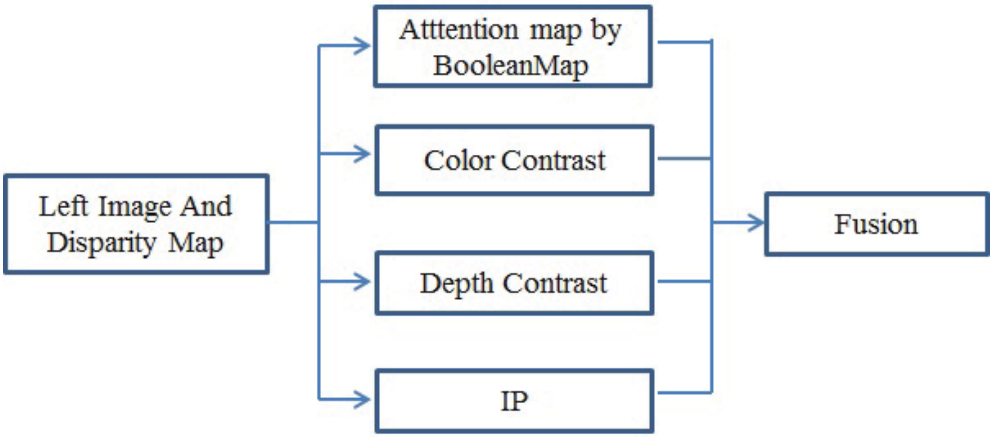

Figure 5.3: The main steps of the framework based on the multi-feature analysis.

superpixels to replace the corresponding pixels. The superpixel-level saliency map of the color/depth image in Fig.5.4(e) demonstrates the improvement over Fig.5.4(d). Fig.5.4(f) is based on the superpixel-level saliency map and is close to the ground truth shown in Fig.5.4(c).

The input image I in the Lab color space is represented by four channels $[L, a, b, depth]$. The range of each channel is translated and scaled to $[0, 255]$. The superpixel value in the channel is the mean value of the pixels in this superpixel, and the channel is processed according to the thresholds in the set of Boolean maps $B = \{B_1, B_2, ..., B_n\}$. The surrounding region is calculated for each Boolean map B_i as the attention map A_i (Huang & Pashler 2007). Then, a linear combination of all the attention maps forms the final attention map.

θ is a set of thresholds $\{\theta_1, \theta_2, ...\theta_n\}$. We define the superpixel-level Boolean map as B_i for threshold θ_i. $A(B_i)$ is the attention map computed by B_i. A linear combination of all the attention maps forms the saliency map S_{BM}. The stereoscopic saliency model, based on the improved Boolean

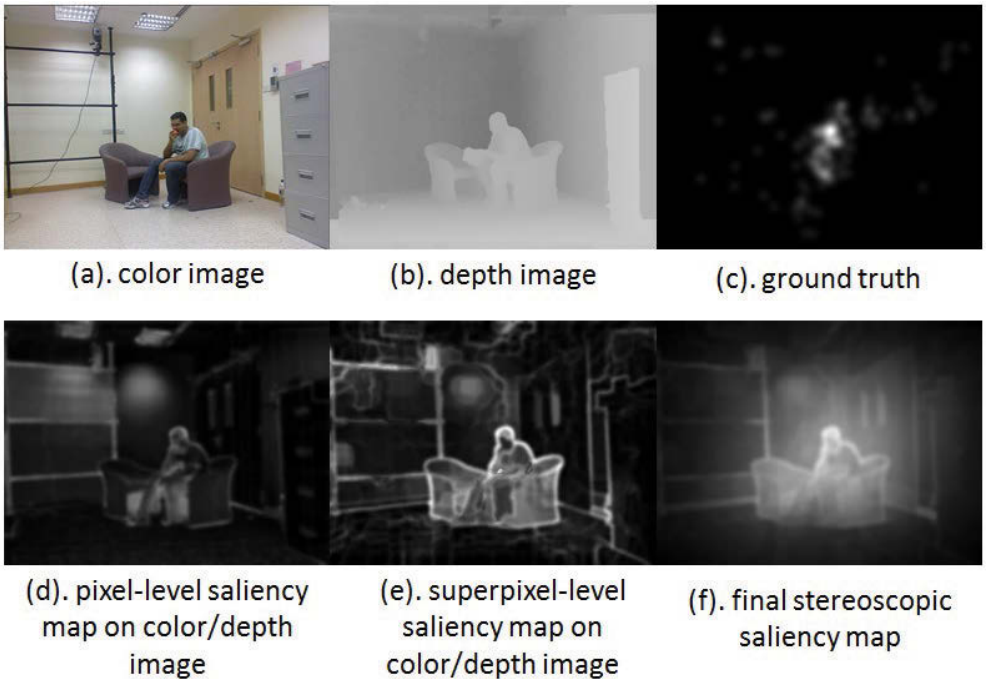

Figure 5.4: Attention maps at pixel-level and at superpixel-level

approach, can be expressed as

$$S_{BM} = \sum_{i=1, i\in\theta_i} A(B_i) w(B_i|I) \tag{5.9}$$

where w is a weight value to enhance small regions and relies on B_i and I. In practice, the weight value is the inversely proportional to the number of pixels in the surrounding region (Cheng et al. 2015).

The contrast map of color and depth information

Human vision is sensitive to contrast in images, and saliency can be defined as contrast or uniqueness (Cheng et al. 2011). Nearer regions tend to attract attention, as do regions with depth discontinuity when viewing positions or angles are changed (Aziz & Mertsching 2010).

Our observations indicate that the contribution both color and depth contrast make to saliency relies on the differences in depth and the distance

between neighboring regions. Therefore, our contrast model measures both features (Cheng et al. 2015):

$$Contrast(i,j) = (\frac{D(i,j)}{1+k*L(i,j)}) * \omega_j \tag{5.10}$$

where ω_j is the number of pixels in superpixel j, and k is a control value for the spatial information ($k = 3$ in our implementation). $D(i,j)$ is the Euclidean distance between the vectored superpixels i and j in the color/depth space, normalized to the range $[0,1]$. $L(i,j)$ is the Euclidean distance between the position of superpixel i and j, normalized to the range $[0,1]$. As mentioned above, the saliency of a superpixel t can be defined by a measure of its distinctiveness as

$$S_{Col/Dep}(t) = \sum_{i \neq t, i \in R} Contrast(t,i) \tag{5.11}$$

where R is the range and $S_{Col/Dep}(t)$ is the color/depth saliency value of superpixel t in the range R. We can compute the local and global saliency for different ranges, as shown in Fig.5.5.

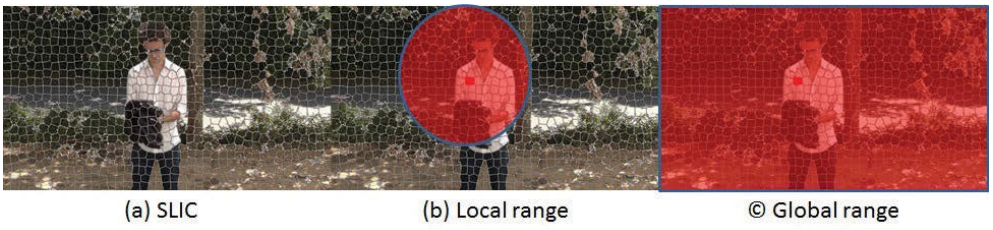

Figure 5.5: Global and local range.

By using Eq.(5.11) to process the color and depth maps, the color contrast S_{Col} and depth contrast S_{Dep} saliency features can be identified, as shown in Fig.5.6.

The IP feature

The IP can be used for saliency analysis because it can reflect the distribution of the GP (Nauge et al. 2012), which can be used to form the ground truth

Figure 5.6: Global and local saliency maps.

of the saliency map. We propose an IP feature extraction approach.

Figure 5.7: A saliency map based on IP, and a ground-truth saliency map based on the GP

We use scale-invariant feature transform (SIFT) to obtain the IPs. SIFT is widely used in scale-invariant feature detection and is suitable for different quality images. The IP feature describes the probability of the GP of the color image that is not affected by noise caused by low-quality depth maps.

For each pixel p, the surrounding IPs are sorted according to distance, from near to far. We use each IP as the center of a Gaussian filter with σ, which is equal to 1 degree of visual angle. A set of Gaussian filter maps $G = \{G_1, G_2, ..., G_n\}$ are computed, with n as the number of IPs. Pixel

saliency $S_{IP}(p)$ can also be represented by these Gaussian filter maps.

$$S_{IP}(p) = \sum G_n(p) \tag{5.12}$$

If an object has complex surfaces or a background has complex textures, IPs may fall more into either object regions or background regions. We observe that if there are many IPs around the current pixel, pixel saliency may be over-strengthened. To avoid this issue, we use η to restrict the number Gaussian filter maps that are computed for the current pixels saliency to a suitable number. If $n > \eta$, we use η to compute the Gaussian filter maps, otherwise we use n. Fig.5.7 provides an example of a pixel-level saliency map based on IPs, demonstrating that using IPs enables the saliency map to be acquired directly.

Pixel-level features

The IP feature map is directly computed from the input images as a pixel-level feature map. The other features are based on multi-scale segmentation and are superpixel-level feature maps. By controlling the number of the superpixels in the SLIC segmentation (the scale), the input color map and depth map become multi-scale maps. However, before the four features are fused, we need to conduct multi-scale fusion to obtain pixel-level feature maps. We adopted the method in (Lu et al. 2016) to fuse the segmentation feature values in different scales. This method considers the multi-scale value and its textural information based on the textural features of the pixel. Its corresponding superpixel is used as a weight value to average the multi-scale value. This multi-scale fusion process produces pixel-level feature maps of the attention map based on the Boolean map, the color contrast map, and the depth contrast map.

Fusion strategy for the four features

The four pixel-level feature maps can then be fused to generate the saliency map. Stereoscopic images may contain noisy feature; therefore, using a

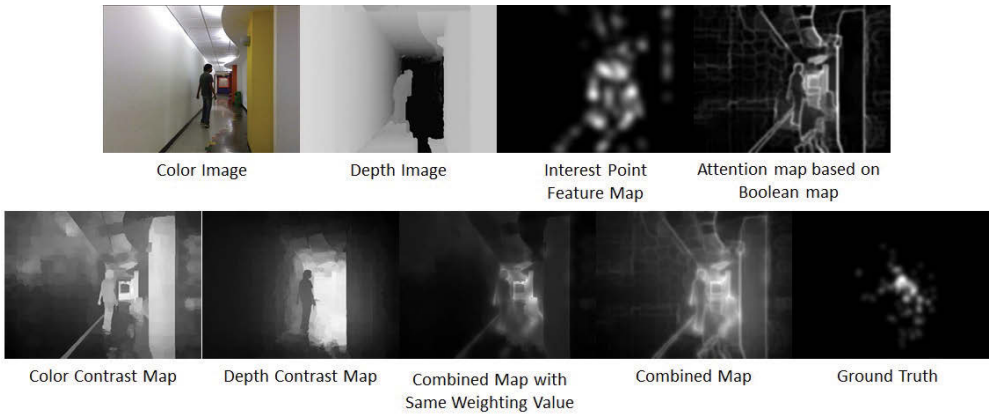

Figure 5.8: Samples of four features and combined saliency maps.

simple linear combination of constant values is not a good fusion strategy. Rather, each feature is given a weight based on the importance of the feature maps. Moreover, the common salient regions across each of the four features are enhanced by the fusion process. For simplicity, we use F_i, $i = \{1, 2, 3, 4\}$ to represent one feature in $\{S_{BM}, S_{Col}, S_{Dep}, S_{IP}\}$. The fusion strategy is computed as

$$Fusion(F_1, F_2, F_3, F_4) = (1 - v) * (\sum_i u_i \cdot F_i) + v * \cdot \prod F_i \tag{5.13}$$

The first term $(1 - v) * (\sum_i u_i \cdot F_i)$ is a linear combination of the four feature maps weighted according to the importance of the feature maps. The second term $v * \cdot \prod F_i$ is the common saliency region, which can be extracted from the four feature maps. The first and second terms are normalized to $[0, 1]$. v is a weight to balance the first term and the second term ($v = 0.5$ in our implementation). u is a weight to measure the importance of one feature and can be computed by

$$u_i = \sum_{j \neq i} |F_i - \overline{F_{ij}}| \tag{5.14}$$

where i, j are any two features in $\{S_{BM}, S_{Col}, S_{Dep}, S_{IP}\}$; $\overline{F_{ij}}$ is the mean of

two feature maps i, j. Fig.5.8 provides an example of the four feature maps and the resulting fused saliency map.

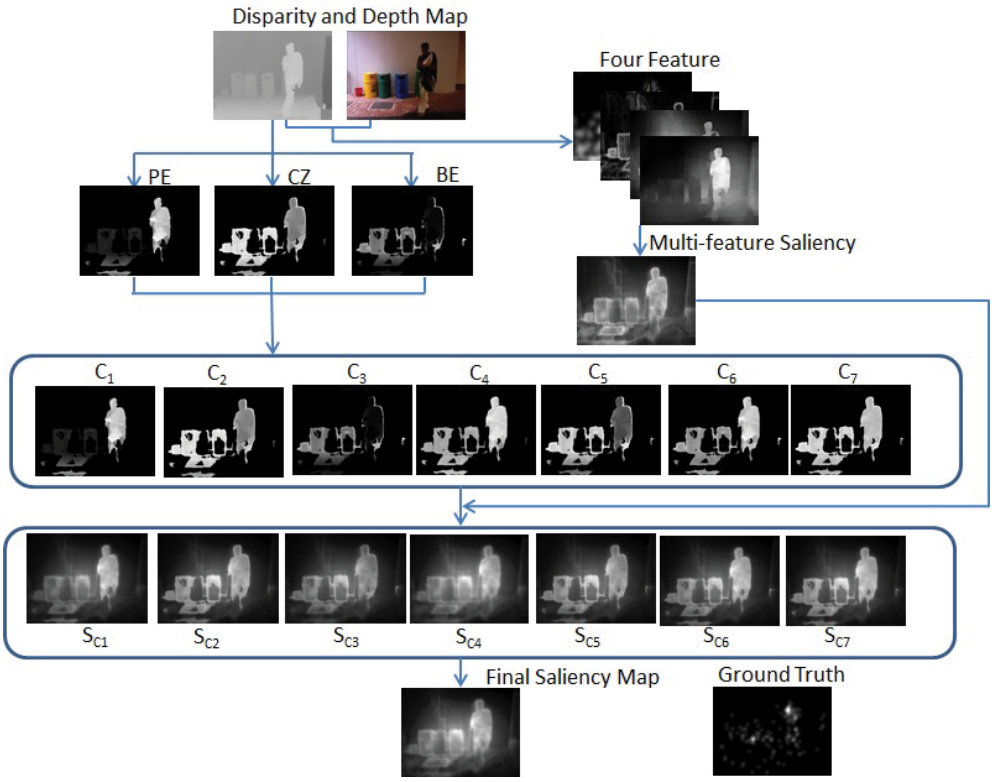

Figure 5.9: Samples of the proposed saliency model.

The final output of this framework (the fused saliency map based on surrounding regions, contrast, and points of interest) is then used as input for the proposed saliency model, which contains the pop-out effect, comfort zone, and background effect modules, to produce the final stereoscopic saliency map. Fig.5.9 illustrates an example of the proposed saliency model process. The disparity map computes three maps using the PE, CZ, and BE modules. The multi-feature saliency map S_{input} as defined in Eq.5.8 is computed by the left image and the disparity maps of the four features $(S_{BM}, S_{Col}, S_{Dep}, S_{IP})$. Then, all combinations of the maps $(C_1, C_2, ..., C_7)$ are computed and the possible stereoscopic saliency maps $(S_{C1}, S_{C2}, ..., S_{C7})$. By comparing the expectation $E_i(C_i, S_{Ci}), i = \{1, 2, ..., 7\}$ between C_i and

S_{Ci}, we choose the S_{best} with the least expectation as our final stereoscopic saliency map. Additionally, by substituting the saliency models S_{input}, our proposed stereoscopic saliency model can be combined with the other 2D or stereoscopic saliency models for stereoscopic saliency analysis.

5.4 Experiments

We evaluated the performance of our proposed saliency model through a series of experiments with two eye-tracking datasets (Wang et al. 2013)(Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). The two datasets provide eye fixation information as ground truths. The evaluation methodology and experimental setup are explained first, followed by the experimental results for each of the components. All depth map are supplied by datasets, which is generated by the different depth-capture sensors. This section concludes with a performance comparison of the proposed method against several state-of-the-art methods.

5.4.1 Experimental setup

All images were first resized to a width of 600 pixels. The IP map was computed directly from the images, while the SLIC algorithm was used to segment the color and depth images for the attention map, the color contrast map, and the depth contrast map. We set the number of segments to a $scale = \{600, 800, 1000, 1200\}$. In Eq.(5.2), we set α to 0.5, as per (Cheng et al. 2015). All distances were normalized to $[0, 1]$ and the range R was empirically set to 0.3, and the global range was set to 1. In the IP feature, the suitable number η was empirically set to 20.

We choose two public datasets to evaluate whether our proposed saliency model is suitable for different stereoscopic capture devices. One is published in the study by (Wang et al. 2013) and includes 18 stereoscopic images of various types (e.g. indoor scenes, outdoor scenes, and scenes containing various numbers of objects). Ten images are chosen from the Middlebury

dataset 2005/2006 (Hirschmüller & Scharstein 2007). The other eight images have been recorded by the authors with a Panasonic AG-3DA1 3D camera. To avoid 3D fatigue (Hoffman et al. 2008) resulting from conflict in the depth field (for example, one object is seen by the right eye but is missed by the left eye), the degree of vergence in human vision within the stereoscopic 3D viewing environment is considered in this eye-tracking experiment. The disparity of the stereoscopic images is suitable for the human comfort viewing zone and has better accuracy than the dataset (Wang et al. 2013). The conflict in different depth fields will not be detected by viewers during the eye-tracking experiments. The eye-tracker is used to record the gaze points, which are processed by a Gaussian kernel to generate the fixation density maps as the ground-truth maps.

The second eye-tracking dataset is published in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). This dataset is an eye-tracking dataset and has 600 3D images that include outdoor and indoor scenes containing different numbers and sizes of objects, and different degrees of interaction or activity are depicted in the scene. The stereoscopic images are generated by Kinect-1. The stereoscopic image pairs are produced by pre-processing, calibration and post-processing. The eye tracking data from 80 participants (age range 20 to 33 years old) are captured by the eye-tracker in both the 2D and 3D free-viewing experiments. Human fixation maps are constructed from the fixations of viewers to globally represent the spatial distribution of human fixations. A Gaussian kernel is used to obtain the continuous fixation density maps as the ground-truth maps. This dataset supplies 2D and 3D fixation maps; to facilitate comparison, we use a 3D fixation map as the stereoscopic 3D ground-truth map.

To quantitatively evaluate the performance of our proposed saliency model, we applied similar measuring methods to those in (Wang et al. 2013). The performance of the proposed model was measured by comparing the computed saliency map with the ground-truth map supplied by the dataset. Because there are two images (left and right) for any stereoscopic image pair,

we used the saliency map of the left image for comparison, similar to the study in (Wang et al. 2013). The area under the receiver operating characteristics curve (AUC) and the correlation coefficient (CC) were used to evaluate the quantitative performance of the proposed stereo saliency detection model. The AUC was computed from the comparison between the actual fixation density map and the predicted saliency map, while CC was calculated directly from the comparison between the fixation density map and the predicted saliency map. We adopted these three measures to quantitatively compare the eye-tracking ground truth and predicted saliency map.

5.4.2 Performance of the features and components

We began by evaluating the performance of the multi-feature saliency model in the dataset (Wang et al. 2013), including the attention map (S_{BM}), the contrast map of color information S_{Col}, the contrast map of the depth information S_{Dep}, the IP features S_{IP}, and the multi-feature saliency map $Fusion$. Then, we compared the three mechanism modules: the pop-out effect (PE), comfort zone (CZ), and background effect (BE), which were directly combined with the multi-feature saliency model. All results are shown in Tables I and II.

As these tables show, directly combining the three mechanism modules into the saliency map Fusion does not produce good results because PE, CZ, and BE cannot be used independently to make salience predictions about stereoscopic images. However, when selecting the best combination of the three mechanisms, our proposed model performed better. Fig.5.10 provides comparative examples.

5.4.3 Comparison with the state-of-the-art methods

We compared our model with three methods that represent state-of-the-arts in each category of stereoscopic saliency modeling: 2D saliency models, DSM-based models, and multi-feature models saliency models. We evaluated the

Table 5.1: Comparison between four features and two components in the dataset (Wang et al. 2013)

Component combination	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
S_{BM}	0.694	0.235
S_{Col}	0.721	0.347
S_{Dep}	0.726	0.389
S_{IP}	0.633	0.219
$Fusion$	0.757	0.455
PE	0.739	0.399
CZ	0.721	0.346
BE	0.694	0.29
Our Model	**0.872**	**0.68**

Table 5.2: Comparison between each component in the dataset (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012)

Component combination	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
S_{BM}	0.72	0.251
S_{Col}	0.778	0.305
S_{Dep}	0.73	0.266
S_{IP}	0.726	0.294
$Fusion$	0.833	0.358
PE	0.804	0.319
CZ	0.798	0.306
BE	0.772	0.282
Our Model	**0.871**	**0.430**

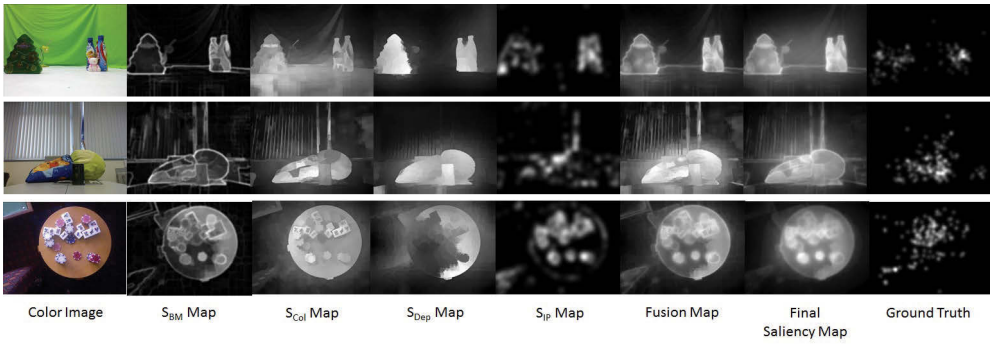

Figure 5.10: Examples of the components of the proposed model

proposed model in the dataset (Wang et al. 2013). The 2D saliency models are IT (Itti et al. 1998), AIM (Bruce & Tsotsos 2005*b*), SR (Hou & Zhang 2007), and GBVS (Harel et al. 2006) (denoted as 2D models in Table III). The DSM-based models combine these 2D models with the depth saliency models proposed in (Huynh-Thu et al. 2011) (denoted as 2D × Depth (Chamaret)) and in (Wang et al. 2013) (denoted as 2D + Depth Contrast and 2D +DSM). The multi-feature saliency models are SDSI (Fang et al. 2013), GNLNS (Iatsun et al. 2015), and SCB (Jiang et al. 2015). The AUC and CC values for these three models were taken from the original paper. From Table III, it can be seen that directly using depth information as a weighted value (such as 2D × Depth (Chamaret)) may not obtain significant improvements in AUC and CC because the combined method may not consider the characteristics of the depth information. In fact, the performance may even be worse. For example, neither 2D × Depth (Chamaret) nor 2D + Depth Contrast improved the performance of GBVS in terms of AUC and CC. Thus, simply combining a 2D saliency model with a depth saliency model may not achieve good performance. Our proposed saliency detection model demonstrated the best performance of all the models, as shown by one example image in Fig. 11. IT, IT × Depth (Chamaret), and IT + Depth Contrast mainly detected the contours of the salient area. AIM and SR included some background areas in the salient area. Whereas, our stereoscopic saliency detection model accurately estimated the saliency region with respect to the

ground-truth map in the eye-tracking data.

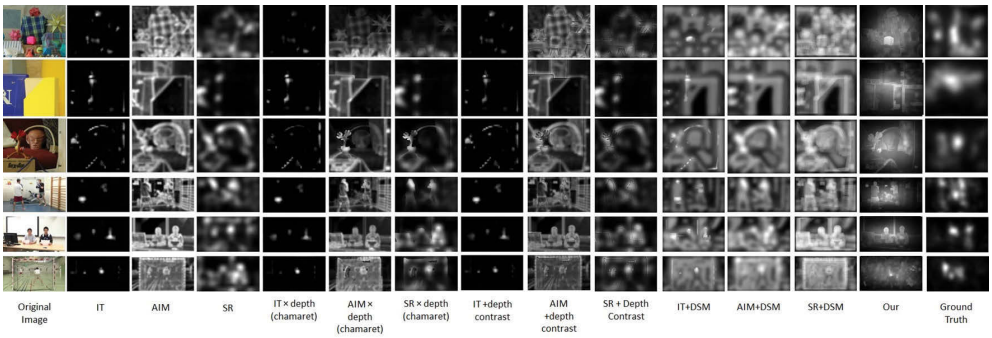

Fig.11 stereo comfort zone based on human stereo vision. DSM represents the depth saliency map in (Wang et al. 2013).

The second experiment with the dataset in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012) relied on 3D fixation maps as the ground truths. We used the DSM from (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012) as the comparative depth saliency model. This model uses global-context depth priors to improve the performance of 2D saliency models. We were unable to locate the code for the DSM in (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012) and, therefore, are only able to compare our results with the best methods listed in the original paper. The 2D saliency models selected for comparison were IT(Itti et al. 1998), AIM (Bruce & Tsotsos 2005*b*), FT(Achanta et al. 2009), GBVS(Harel et al. 2006), ICL(Hou & Zhang 2009), LSK(Seo & Milanfar 2009), LRR(Lang, Liu, Yu & Yan 2012). The stereoscopic saliency map was achieved by simply using the summation "+" or point-wise multiplication "×" as the fusion of the depth and 2D models. Performance was quantitatively evaluated using AUC and CC, and the experimental results are shown in Table IV. Note that the AUC and CC values of existing models were taken from the original paper (Lang, Nguyen, Katti, Yadati, Kankanhalli & Yan 2012). From this table, we see that our proposed models performance exceeded the other 15 stereo saliency detection models. This is because our proposed model chooses the most suitable mechanism depending on the content, and uses it to increase

Table 5.3: Comparison between the proposed framework and others. DSM represents the depth saliency map in (Wang et al. 2013)

Model		AUC($\rightarrow 1$)	CC($\rightarrow 1$)
2D Model	IT	0.538	0.137
	AIM	0.638	0.326
	SR	0.63	0.291
	GBVS	0.809	0.54
2D $\times$ Depth (Chamaret)	IT $\times$ depth	0.54	0.137
	AIM $\times$ depth	0.636	0.299
	SR $\times$ depth	0.634	0.292
	GBVS $\times$ depth	0.771	0.515
2D + Depth Contrast	IT + Depth Contrast	0.596	0.211
	AIM + Depth Contrast	0.644	0.343
	SR + Depth Contrast	0.662	0.307
	GBVS + Depth Contrast	0.799	0.53
2D + DSM	Model 1	0.656	0.356
	Model 2	0.675	0.424
	Model 3	0.67	0.41
Multi-feature Models	SDSI(Fang et al. 2013)	0.703	0.55
	GNLNS(Iatsun et al. 2015)	0.77	0.65
	SCB(Jiang et al. 2015)	0.818	0.491
Our Model		**0.872**	**0.68**

the accuracy of the saliency prediction.

Comparing with the models in Chapter 3 and Chapter 4, the CC value is improved significantly in this model. The reason is that Chapter 5 model can cover all kinds of scene type, which the model in Chapter 3 and 4 can not explain very well. No only the Chapter 5 model considers the background effect, but also it considers the different combination of three mechanisms, including the pop-out effect, comfort zone, and background effect. Comparing with the model in Chapter 4, the AUC value is not improved significantly in datasets (Wang et al. 2013). The reason is the stereo contrast model and stereo focus model can explain the most of the stereo image in this datasets. Even if the Chapter 5 model considers the background effect, the performance is still not improved significantly. In summary, both experiments demonstrate that our proposed model successfully processes the stereoscopic images captured by various devices with good performance compared to other models.

5.5 Conclusions

In this chapter, we thoroughly exploit the depth information for stereoscopic saliency analysis and present a computational model for stereoscopic visual saliency prediction. Our model is based on three mechanisms: the pop-out effect, the comfort zone, and the background effect. Three modules within the model describe each of these three mechanisms, and a control function is used to adjust the weight of each. The best resulting stereoscopic saliency map is chosen through a selection strategy. This approach overcomes the issues of the proposed models in Chapter 3 and Chapter 4. Meanwhile, it is implemented within a framework based on multi-feature analysis that considers the surrounding regions, contrast, and points of interest. The experimental results show that our model performs better than other saliency models in terms of AUC and CC.

Table 5.4: Comparison between 3D saliency detection models.

Component combination	AUC($\rightarrow 1$)	CC($\rightarrow 1$)
IT+DSM	0.849	0.375
IT×DSM	0.854	0.398
GBVS+DSM	0.851	0.39
GBVS×DSM	0.855	0.413
AIM+DSM	0.85	0.342
AIM×DSM	0.85	0.391
FT+DSM	0.797	0.315
FT×DSM	0.745	0.268
ICL+DSM	0.846	0.385
ICL×DSM	0.808	0.325
LSK+DSM	0.845	0.379
LSK×DSM	0.824	0.351
LRR+DSM	0.856	0.385
LRR×DSM	0.846	0.395
Our Model	**0.871**	**0.43**

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In conclusion, this thesis presented three computational models for stereoscopic saliency detection.

Chapter 3 demonstrated a preliminary saliency detection model for stereoscopic images. This model uses depth information to leverage stereo saliency analysis in three respects. At a low-level, local-global features are used to analyse saliency by considering the colour and depth contrast in local and global ranges. At the mid-level, the surrounding map, based on a Boolean map, is obtained as a weight value to enhance the local-global contrast features. Lastly, by analysing the saliency probability distribution in the depth information, a stereocenter prior enhancement is used to form the final saliency.

Chapter 4 presented a stereo visual saliency prediction based on stereo contrast and stereo focus. Stereo contrast, as a product of colour and depth contrast and the pop-out effect, describes the contrast in objects. Stereo focus is based on the focus mechanism of human stereoscopic vision and describes the region of human focus. Two saliency maps, one generated from the stereo contrast information and the other generated from stereo focus information, were integrated using Bayesian integration. Experimental

results on two eye-tracking databases show that the model is able to process stereoscopic images from different stereoscopic capture devices to achieve better performance than existing methods.

Chapter 5 presented a computational model for stereoscopic visual saliency prediction based on three mechanisms in the human vision system: the pop-out effect, comfort zones and the background effect. Three modules each describe one of these three mechanisms, and a control function is used to adjust the weight of each. The best resulting stereoscopic saliency map is chosen through a selection strategy. Further, the model is implemented within a framework based on multi-feature analysis that considers the surrounding regions, contrast and points of interest. Experimental results again show that this model performs better than other saliency models in terms of AUC and CC.

6.2 Future work

Stereoscopic saliency detection is still young, and there is much to be studied. Future research endeavours will focus on several potential directions, both theory-driven and application-driven. These tasks include:

(i). **New saliency detection models**: Three stereoscopic saliency detection modules have been implemented based on three different mechanisms of human vision. However, these models still do not fully explain the saliency in natural scenes. Further study of new and more advanced stereoscopic saliency detection models is required.

(ii). **Applications in the subjective evaluation for stereoscopic images**: Exploring whether the three stereoscopic saliency detection models presented can be used as tools to solve an important problem the subjective evaluation of stereoscopic images is another promising avenue of future research. We can use the stereoscopic saliency model as a measure to determine whether the image meets the characteristics of cognitive perception.

(iii). **Applications in multi-view display systems**: Automatic parallax adjustment is an important problem in multi-view display systems. The proposed saliency detection models might be used as a cue for automatic parallel adjustment, which forms a further avenue of future study.

Bibliography

Achanta, R., Estrada, F., Wils, P. & Süsstrunk, S. (2008), Salient region detection and segmentation, *in* 'Computer Vision Systems', Springer, pp. 66–75.

Achanta, R., Hemami, S., Estrada, F. & Susstrunk, S. (2009), Frequency-tuned salient region detection, *in* 'Computer Vision and Pattern Recognition, IEEE Conference on', IEEE, pp. 1597–1604.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. & Susstrunk, S. (2012), 'Slic superpixels compared to state-of-the-art superpixel methods', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(11), 2274–2282.

Anderson, J. R. (1990), *Cognitive psychology and its implications* ., WH Freeman/Times Books/Henry Holt & Co.

Awh, E. & Pashler, H. (2000), 'Evidence for split attentional foci.', *Journal of Experimental Psychology: Human Perception and Performance* **26**(2), 834.

Aziz, M. Z. & Mertsching, B. (2010), Fast depth saliency from stereo for region-based artificial visual attention, *in* 'Advanced Concepts for Intelligent Vision Systems', Vol. 6474, Lecture Notes in Computer Science, pp. 367–378.

Beato, A. (2011), 'Understanding comfortable stereography', *Techical Report* .

Borji, A., Cheng, M.-M., Jiang, H. & Li, J. (2015), 'Salient object detection: A benchmark', *IEEE Transactions on Image Processing* **24**(12), 5706–5722.

Borji, A. & Itti, L. (2013), 'State-of-the-art in visual attention modeling', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(1), 185–207.

Bruce, N. D. & Tsotsos, J. K. (2005*a*), An attentional framework for stereo vision, *in* 'Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on', IEEE, pp. 88–95.

Bruce, N. & Tsotsos, J. (2005*b*), Saliency based on information maximization, *in* 'Advances in neural information processing systems', pp. 155–162.

Bruce, N. & Tsotsos, J. (2006), 'Saliency based on information maximization', *Advances in neural information processing systems* **18**, 155.

Camplani, M. & Salgado, L. (2012), Efficient spatio-temporal hole filling strategy for kinect depth maps, *in* 'IS&T/SPIE Electronic Imaging', International Society for Optics and Photonics, pp. 82900E–82900E.

Cerf, M., Harel, J., Einhäuser, W. & Koch, C. (2008), Predicting human gaze using low-level saliency combined with face detection, *in* 'Advances in Neural Information Processing Systems', pp. 241–248.

Chamaret, C., Godeffroy, S., Lopez, P. & Le Meur, O. (2010), Adaptive 3d rendering based on region-of-interest, *in* 'IS&T/SPIE Electronic Imaging', International Society for Optics and Photonics, pp. 75240V–75240V.

Chang, Y.-S., Hsueh, Y.-H., Tung, K.-C., Jhou, F.-Y. & Lin, D. P.-C. (2016), 'Characteristics of visual fatigue under the effect of 3d animation', *Technology and Health Care* **24**(s1), S231–S235.

Cheng, H.-D., Jiang, X., Sun, Y. & Wang, J. (2001), 'Color image segmentation: advances and prospects', *Pattern recognition* **34**(12), 2259–2281.

Cheng, H., Zhang, J., An, P. & Liu, Z. (2015), A novel saliency model for stereoscopic images, *in* 'Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on', IEEE, pp. 1–7.

Cheng, H., Zhang, J., Wu, Q., An, P. & Liu, Z. (2017), 'Stereoscopic visual saliency prediction based on stereo contrast and stereo focus', *EURASIP Journal on Image and Video Processing* **2017**(1), 61.

Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X. & Hu, S.-M. (2011), Global contrast based salient region detection, *in* 'Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on', IEEE, pp. 409–416.

DAVID, M. G. (1966), 'Signal detection theory and psychophysics.'.

Desingh, K., Krishna, K. M., Rajan, D. & Jawahar, C. (2013), Depth really matters: Improving visual salient region detection with depth., *in* 'BMVC'.

Einhäuser, W. & König, P. (2003), 'Does luminance-contrast contribute to a saliency map for overt visual attention?', *European Journal of Neuroscience* **17**(5), 1089–1097.

Elder, J. H. & Zucker, S. W. (1998), 'Local scale control for edge detection and blur estimation', *IEEE Transactions on pattern analysis and machine intelligence* **20**(7), 699–716.

Erdem, E. & Erdem, A. (2013), 'Visual saliency estimation by nonlinearly integrating features using region covariances', *Journal of Vision* **13**(4), 11.

Eriksen, C. W. & James, J. D. S. (1986), 'Visual attention within and around the field of focal attention: A zoom lens model', *Perception & psychophysics* **40**(4), 225–240.

Fan, X., Liu, Z. & Sun, G. (2014), Salient region detection for stereoscopic images, *in* 'Digital Signal Processing (DSP), 2014 19th International Conference on', IEEE, pp. 454–458.

Fang, Y., Wang, J., Narwaria, M., Le Callet, P. & Lin, W. (2013), Saliency detection for stereoscopic images, *in* 'Visual Communications and Image Processing (VCIP), 2013', IEEE, pp. 1–6.

Fang, Y., Wang, J., Narwaria, M., Le Callet, P. & Lin, W. (2014), 'Saliency detection for stereoscopic images', *IEEE Transactions on Image Processing* **23**(6), 2625–2636.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. & Ramanan, D. (2010), 'Object detection with discriminatively trained part-based models', *IEEE transactions on pattern analysis and machine intelligence* **32**(9), 1627–1645.

Frintrop, S., Rome, E. & Christensen, H. I. (2010), 'Computational visual attention systems and their cognitive foundations: A survey', *ACM Transactions on Applied Perception (TAP)* **7**(1), 6.

Goferman, S., Zelnik-Manor, L. & Tal, A. (2010), Context-aware saliency detection, *in* 'Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on', IEEE, pp. 2376–2383.

Goferman, S., Zelnik-Manor, L. & Tal, A. (2012), 'Context-aware saliency detection', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **34**(10), 1915–1926.

Gopalakrishnan, V., Hu, Y. & Rajan, D. (2009), 'Salient region detection by modeling distributions of color and orientation', *IEEE Transactions on Multimedia* **11**(5), 892–905.

Guo, C. & Zhang, L. (2010), 'A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression', *Image Processing, IEEE Transactions on* **19**(1), 185–198.

Häkkinen, J., Kawai, T., Takatalo, J., Mitsuya, R. & Nyman, G. (2010), What do people look at when they watch stereoscopic movies?, *in* 'IS&T/SPIE Electronic Imaging', International Society for Optics and Photonics, pp. 75240E–75240E.

Harel, J., Koch, C. & Perona, P. (2006), Graph-based visual saliency, *in* 'Advances in Neural Information Processing Systems', pp. 545–552.

Henderson, J. M. & Hollingworth, A. (1999), 'High-level scene perception', *Annual Review of Psychology,* **50**(1), 243–271.

Hirschmüller, H. & Scharstein, D. (2007), Evaluation of cost functions for stereo matching, *in* 'IEEE 2007 Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 1–8.

Hoffman, D. M., Girshick, A. R., Akeley, K. & Banks, M. S. (2008), 'Vergence–accommodation conflicts hinder visual performance and cause visual fatigue', *Journal of vision* **8**(3), 33.

Hou, X. & Zhang, L. (2007), Saliency detection: A spectral residual approach, *in* 'Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on', IEEE, pp. 1–8.

Hou, X. & Zhang, L. (2009), Dynamic visual attention: Searching for coding length increments, *in* 'Advances in Neural Information Processing Systems', pp. 681–688.

Huang, L. & Pashler, H. (2007), 'A Boolean map theory of visual attention.', *Psychological Review* **114**(3), 599.

Huynh-Thu, Q., Barkowsky, M. & Le Callet, P. (2011), 'The importance of visual attention in improving the 3d-tv viewing experience: Overview

and new perspectives', *IEEE Transactions on Broadcasting* **57**(2), 421–431.

Iatsun, I., Larabi, M.-C. & Fernandez-Maloigne, C. (2015), 'A visual attention model for stereoscopic 3d images using monocular cues', *Signal Processing: Image Communication* **38**, 70–83.

Itti, L., Koch, C. & Niebur, E. (1998), 'A model of saliency-based visual attention for rapid scene analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(11), 1254–1259.

Jang, H., Park, S., Lee, D., Han, S., Donghyun, K. & Song, H. (2017), 'Terminal for increasing visual comfort sensation of 3d object and control method thereof'. US Patent 9,674,501.

Jansen, L., Onat, S. & König, P. (2009), 'Influence of disparity on fixation and saccades in free viewing of natural scenes', *Journal of Vision* **9**(1), 29–29.

Jeong, S., Ban, S.-W. & Lee, M. (2008), 'Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment', *Neural Networks* **21**(10), 1420–1430.

Jiang, P., Ling, H., Yu, J. & Peng, J. (2013), Salient region detection by ufo: Uniqueness, focusness and objectness, *in* 'Computer Vision (ICCV), 2013 IEEE International Conference on', IEEE, pp. 1976–1983.

Jiang, Q., Shao, F., Jiang, G., Yu, M., Peng, Z. & Yu, C. (2015), 'A depth perception and visual comfort guided computational model for stereoscopic 3D visual saliency', *Signal Processing: Image Communication* pp. 38: 57–69.

Judd, T., Ehinger, K., Durand, F. & Torralba, A. (2009), Learning to predict where humans look, *in* 'IEEE 12th International Conference on Computer Vision', pp. 2106–2113.

Khaustova, D., Fournier, J., Wyckens, E. & Le Meur, O. (2013), How visual attention is modified by disparities and textures changes?, *in* 'Proceedings of SPIE 8651 Human Vision and Electronic Imaging XVIII', International Society for Optics and Photonics, pp. 865115–865115.

Kim, H., Lee, S. & Bovik, A. C. (2014), 'Saliency prediction on stereoscopic videos', *IEEE Transactions on Image Processing* **23**(4), 1476–1490.

Ko, B. C. & Nam, J.-Y. (2006), 'Object-of-interest image segmentation based on human attention and semantic region clustering', *JOSA A* **23**(10), 2462–2470.

Kobyshev, N., Riemenschneider, H., Bódis-Szomorú, A. & Van Gool, L. (2016), 3d saliency for finding landmark buildings, *in* '3D Vision (3DV), 2016 Fourth International Conference on', IEEE, pp. 267–275.

Lambooij, M., Fortuin, M., Heynderickx, I. & IJsselsteijn, W. (2009), 'Visual discomfort and visual fatigue of stereoscopic displays: A review', *Journal of Imaging Science and Technology* **53**(3), 30201–1.

Lang, C., Liu, G., Yu, J. & Yan, S. (2012), 'Saliency detection by multitask sparsity pursuit', *IEEE Transactions on Image Processing* **21**(3), 1327–1338.

Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M. & Yan, S. (2012), 'Depth matters: Influence of depth cues on visual saliency', pp. 101–115.

Lang, M., Hornung, A., Wang, O., Poulakos, S., Smolic, A. & Gross, M. (2010), 'Nonlinear disparity mapping for stereoscopic 3D', *ACM Transactions on Graphics (TOG)* **29**(4), 75.

Lee, Y.-J. & Song, J.-B. (2010), 'Autonomous salient feature detection through salient cues in an hsv color space for visual indoor simultaneous localization and mapping', *Advanced Robotics* **24**(11), 1595–1613.

Li, L., Jiang, S., Zha, Z.-J., Wu, Z. & Huang, Q. (2013), 'Partial-duplicate image retrieval via saliency-guided visual matching', *IEEE MultiMedia* **20**(3), 13–23.

Li, X., Lu, H., Zhang, L., Ruan, X. & Yang, M.-H. (2013), Saliency detection via dense and sparse reconstruction, *in* 'Computer Vision (ICCV), 2013 IEEE International Conference on', IEEE, pp. 2976–2983.

Li, Y., Hou, X., Koch, C., Rehg, J. M. & Yuille, A. L. (2014), The secrets of salient object segmentation, *in* 'Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on', IEEE, pp. 280–287.

Lin, H., Lin, C., Zhao, Y., Xiao, J. & Tillo, T. (2015), Depth-based stereoscopic projection approach for 3d saliency detection, *in* 'Pacific Rim Conference on Multimedia', Springer, pp. 664–673.

Liu, T., Sun, J., Zheng, N.-N., Tang, X. & Shum, H.-Y. (2007), Learning to detect a salient object, *in* 'Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on', pp. 1–8.

Liu, Y., Cormack, L. K. & Bovik, A. C. (2010), Natural scene statistics at stereo fixations, *in* 'Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications', ACM, pp. 161–164.

Liu, Z., Shi, R., Shen, L., Xue, Y., Ngan, K. N. & Zhang, Z. (2012), 'Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut', *IEEE Transactions on Multimedia* **14**(4), 1275–1289.

Liu, Z., Zou, W. & Le Meur, O. (2013), 'Saliency tree: A novel saliency detection framework', *IEEE Transactions on Image Processing* pp. 1937–1952.

Lu, H., Li, X., Zhang, L., Ruan, X. & Yang, M.-H. (2016), 'Dense and sparse reconstruction error based saliency descriptor', *IEEE Transactions on Image Processing* **25**(4), 1592–1603.

Luo, Y., Yuan, J., Xue, P. & Tian, Q. (2011), 'Saliency density maximization for efficient visual objects discovery', *Circuits and Systems for Video Technology, IEEE Transactions on* **21**(12), 1822–1834.

Maki, A., Nordlund, P. & Eklundh, J.-O. (1996), A computational model of depth-based attention, *in* 'Pattern Recognition, 1996., Proceedings of the 13th International Conference on', Vol. 4, IEEE, pp. 734–739.

Mendiburu, B. (2009), '3d movie making: Stereoscopic digital cinema from script to screen'.

Nauge, M., Larabi, M.-C. & Fernandez-Maloigne, C. (2012), A statistical study of the correlation between interest points and gaze points, *in* 'Proceedings of SPIE 8291 Human Vision and Electronic Imaging XVII', International Society for Optics and Photonics, pp. 829111–829111.

Niebur, E. & Koch, C. (1998), 'Computational architectures for attention', *The attentive brain* pp. 163–186.

Niu, Y., Geng, Y., Li, X. & Liu, F. (2012), Leveraging stereopsis for saliency analysis, *in* 'Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on', IEEE, pp. 454–461.

Ouerhani, N., Von Wartburg, R., Hugli, H. & Muri, R. (2003), 'Empirical validation of the saliency-based model of visual attention', *Electronic Letters on Computer Vision and Image Analysis* **3**(1), 13–24.

Palmer, S. E. (1999), *Vision science: Photons to phenomenology*, Vol. 1, MIT Press Cambridge, MA.

Peng, H., Li, B., Xiong, W., Hu, W. & Ji, R. (2014), RGBD salient object detection: A benchmark and algorithms, *in* '2014 European Conference on Computer Vision (ECCV)', Springer, pp. 92–109.

Poggio, G. F. & Poggio, T. (1984), 'The analysis of stereopsis', *Annual review of neuroscience* **7**(1), 379–412.

Pylyshyn, Z. W. & Storm, R. W. (1988), 'Tracking multiple independent targets: Evidence for a parallel tracking mechanism', *Spatial Vision* **3**(3), 179–197.

Ramenahalli, S. & Niebur, E. (2013), Computing 3D saliency from a 2d image, *in* 'IEEE 47th Annual Conference on Information Sciences and Systems (CISS),', IEEE, pp. 1–5.

Ren, Z., Gao, S., Chia, L.-T. & Tsang, I. W.-H. (2014), 'Region-based saliency detection and its application in object recognition', *IEEE Transactions on Circuits and Systems for Video Technology* **24**(5), 769–779.

Rutishauser, U., Walther, D., Koch, C. & Perona, P. (2004), Is bottom-up attention useful for object recognition?, *in* 'Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on', Vol. 2, IEEE, pp. II–37.

Scharstein, D. & Pal, C. (2007), Learning conditional random fields for stereo, *in* 'Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on', IEEE, pp. 1–8.

Seo, H. J. & Milanfar, P. (2009), 'Static and space-time visual saliency detection by self-resemblance', *Journal of Vision* **9**(12), 15.

Shao, F., Jiang, G., Yu, M., Chen, K. & Ho, Y.-S. (2012), 'Asymmetric coding of multi-view video plus depth based 3-d video for view rendering', *IEEE Transactions on Multimedia* **14**(1), 157–167.

Shapiro, L. (1992), *Computer vision and image processing*, Academic Press.

Sheng, H., Liu, X. & Zhang, S. (2016), Saliency analysis based on depth contrast increased, *in* 'Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on', IEEE, pp. 1347–1351.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A. & Jain, R. (2000), 'Content-based image retrieval at the end of the early years', *IEEE Transactions on pattern analysis and machine intelligence* **22**(12), 1349–1380.

Solomon, C. & Breckon, T. (2011), *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*, John Wiley & Sons.

Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O. & Ren, T. (2017), 'Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning', *IEEE Transactions on Image Processing* **26**(9), 4204–4216.

Sun, J., Lu, H. & Li, S. (2012), Saliency detection based on integration of boundary and soft-segmentation, *in* 'Image Processing (ICIP), 2012 19th IEEE International Conference on', IEEE, pp. 1085–1088.

Thomas, J. A. & Thomas, J. A. (2006), *Elements of information theory*, Wiley New York.

Treisman, A. M. & Gelade, G. (1980), 'A feature-integration theory of attention', *Cognitive Psychology* **12**(1), 97–136.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N. & Nuflo, F. (1995), 'Modeling visual attention via selective tuning', *Artificial intelligence* **78**(1), 507–545.

Viola, P. & Jones, M. (2001), Rapid object detection using a boosted cascade of simple features, *in* 'Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on', Vol. 1, IEEE, pp. I–511.

Wang, J., DaSilva, M. P., LeCallet, P. & Ricordel, V. (2013), 'Computational model of stereoscopic 3d visual saliency', *IEEE Transactions on Image Processing* **22**(6), 2151–2165.

Wang, J., Le Callet, P., Tourancheau, S., Ricordel, V. & Da Silva, M. P. (2012), 'Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli', *Journal of Eye Movement Research* **5**(5), 1–11.

Wang, Z. & Li, Q. (2011), 'Information content weighting for perceptual image quality assessment', *Image Processing, IEEE Transactions on* **20**(5), 1185–1198.

Wei, Y., Wen, F. & Sun, J. (2013), 'Geodesic saliency using background priors'. US Patent App. 14/890,884.

Wismeijer, D., Erkelens, C., van Ee, R. & Wexler, M. (2010), 'Depth cue combination in spontaneous eye movements', *Journal of vision* **10**(6), 25–25.

Yano, S., Ide, S., Mitsuhashi, T. & Thwaites, H. (2002), 'A study of visual fatigue and visual comfort for 3d hdtv/hdtv images', *Displays* **23**(4), 191–201.

Yarbus, A. L., Haigh, B. & Rigss, L. A. (1967), 'Eye movements and vision', **2**(5–10).

Zhang, J. & Sclaroff, S. (2013), Saliency detection: A boolean map approach, *in* 'Computer Vision (ICCV), 2013 IEEE International Conference on', IEEE, pp. 153–160.

Zhang, Y., Jiang, G., Yu, M., Chen, K. & Dai, Q. (2010), 'Stereoscopic visual attention-based regional bit allocation optimization for multiview video coding', *EURASIP Journal on Advances in Signal Processing* **2010**, 60.

Zhang, Z.-y., An, P., Zhang, Z.-j. & Shen, L.-q. (2010), '2d/3d video processing and stereo display technology'.