

Puneet Sharma

# Towards three-dimensional visual saliency

Thesis for the degree of Philosophiae Doctor

Trondheim, May 2014

Norwegian University of Science and Technology  
Faculty of Information Technology, Mathematics and  
Electrical Engineering  
Department of Computer and Information Science



**NTNU – Trondheim**  
Norwegian University of  
Science and Technology

**NTNU**

Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology, Mathematics and Electrical Engineering  
Department of Computer and Information Science

© Puneet Sharma

ISBN 978-82-326-0214-8 (printed ver.)  
ISBN 978-82-326-0215-5 (electronic ver.)  
ISSN 1503-8181

Doctoral theses at NTNU, 2014:146

Printed by NTNU-trykk

# Abstract

A salient image region is defined as an image part that is clearly different from its surround in terms of a number of attributes. In bottom-up processing, these attributes are defined as: contrast, color difference, brightness, and orientation. By measuring these attributes, visual saliency algorithms aim to predict the regions in an image that would attract our attention under free viewing conditions, i.e., when the observer is viewing an image without a specific task such as searching for an object. To quantify the interesting locations in a scene, the output of the visual saliency algorithms is usually expressed as a two dimensional gray scale map where the brighter regions correspond to the highly salient regions in the original image. In addition to advancing our understanding of human visual system, visual saliency models can be used for a number of computer vision applications. These applications include: image compression, computer graphics, image matching & recognition, design, and human-computer interaction.

In this thesis the main contributions can be outlined as: first, we present a method to inspect the performance of Itti's classic saliency algorithm in separating the salient and non-salient image locations. Based on our results we observed that, although the saliency model can provide a good discrimination for the highly salient and non-salient regions, there is a large overlap between the locations that lie in the middle range of saliency. Second, we propose a new bottom-up visual saliency model for static two-dimensional images. In our model, we calculate saliency by using the transformations associated with the dihedral group  $D_4$ . Our results suggest that the proposed saliency model outperforms many state-of-the-art saliency models. By using the proposed methodology, our algorithm can be extended to calculate saliency in three-dimensional scenes, which we intend to implement in the future. Third, we propose a way to perform statistical analysis of the fixations data from different observers and different images. Based on the analysis, we present a robust metric for judging the performance of the visual saliency algorithms. Our results show that the proposed metric can indeed be used to alleviate the problems pertaining to the evaluation of saliency models. Fourth, we introduce a new approach to compress an image based on the salient locations predicted by the saliency models. Our results show that the compressed images do not exhibit visual artifacts and appear to be very similar to the originals. Fifth, we outline a method to estimate depth from eye fixations in three-dimensional virtual scenes that can be used

for creating so-called gaze maps for three-dimensional scenes. In the future, this can be used as ground truth for judging the performance of saliency algorithms for three-dimensional images.

We believe that our contributions can lead to a better understanding of saliency, address the major issues associated with the evaluation of saliency models, highlight on the contribution of top-down and bottom-up processing based on the analysis of a comprehensive eye tracking dataset, promote use of human vision steered image processing applications, and pave the way for calculating saliency in three-dimensional scenes.

# Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) for partial fulfillment of the requirements for the degree of philosophiae doctor. This project was funded by the Department of Informatics & e-Learning (AITeL), Sør-Trøndelag University College (HiST), Trondheim.

## Acknowledgments

This thesis would not have been possible without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

First, I would like to thank my supervisors Dr. Torbjørn Skramstad and Dr. Ali Alsam for their constant encouragement during these past four years. Torbjørn has provided tremendous support and has given me the freedom to pursue various projects without objection. Ali has given unreserved support during my PhD and generously spent countless hours mentoring me. I would have been lost without him. I am grateful to Dr. Hans Jakob Rivertz and Dr. Anette Wrålsen for collaborating with me on the research articles included in this thesis. Their meticulous comments were an enormous help to me.

I would like to express my gratitude to Per Borgesen and Thorleif Hjeltnes, who gave me the opportunity to commence on this PhD project. I am deeply grateful to my colleagues at AITel who volunteered for eye tracking experiments, and provided their services in creating and maintaining the experiment setup. Finally, I would like to thank my fellow PhD candidates, Rune Havnung Bakken and Knut Arne Strand for their friendship, support and insightful discussions.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Outline of the thesis . . . . .	2
<b>2</b>	<b>Background</b>	<b>5</b>
2.1	Attention . . . . .	5
2.1.1	Psychological and philosophical perspective . . . . .	5
2.1.2	Computational perspective . . . . .	5
2.2	Attention mechanisms . . . . .	6
2.2.1	Selective attention . . . . .	6
2.2.2	Overt and covert attention . . . . .	7
2.2.3	Bottom-up and top-down attention . . . . .	7
2.3	Attention and eye movements . . . . .	9
2.4	State-of-the-art in modeling visual attention . . . . .	11
2.5	Summary . . . . .	19
2.6	Evaluation metrics . . . . .	20
2.7	Suitable candidate for evaluating the saliency algorithms . . . . .	22
<b>3</b>	<b>Research and contributions</b>	<b>25</b>
3.1	Publications . . . . .	25
3.2	Research issues, answers, and future work . . . . .	26
3.2.1	Proposed robust metric for the evaluation of saliency models . . . . .	26
3.2.2	Validating the visual saliency model . . . . .	27
3.2.3	Proposed group based asymmetry algorithm . . . . .	28
3.2.4	Proposed saliency based image compression algorithm . . . . .	29
3.2.5	Depth estimation in three-dimensional scenes . . . . .	30
3.3	Research questions and contributions . . . . .	31
<b>4</b>	<b>Research methodology</b>	<b>33</b>
<b>5</b>	<b>Summaries of research papers</b>	<b>37</b>
5.1	<b>Paper 1:</b> Analysis of eye fixations data . . . . .	37
5.1.1	Synopsis . . . . .	37
5.2	<b>Paper 2:</b> A robust metric for the evaluation of visual saliency models . . . . .	37

5.2.1	Synopsis . . . . .	37
5.3	<b>Paper 3:</b> A robust metric for the evaluation of visual saliency algorithms (extended) . . . . .	38
5.3.1	Synopsis . . . . .	38
5.4	<b>Paper 4:</b> Validating the visual saliency model . . . . .	38
5.4.1	Synopsis . . . . .	38
5.5	<b>Paper 5:</b> Asymmetry as a measure of visual saliency . . . . .	39
5.5.1	Synopsis . . . . .	39
5.6	<b>Paper 6:</b> Calculating saliency using the dihedral group $D_4$ . . . . .	39
5.6.1	Synopsis . . . . .	39
5.7	<b>Paper 7:</b> What the eye did not see—a fusion approach to image coding . . . . .	39
5.7.1	Synopsis . . . . .	39
5.8	<b>Paper 8:</b> What the eye did not see—a fusion approach to image coding (extended) . . . . .	40
5.8.1	Synopsis . . . . .	40
5.9	<b>Paper 9:</b> Evaluation of geometric depth estimation model for virtual environment . . . . .	40
5.9.1	Synopsis . . . . .	40
5.10	<b>Paper 10:</b> Estimating the depth in three-dimensional virtual environment with feedback . . . . .	40
5.10.1	Synopsis . . . . .	40
5.11	<b>Paper 11:</b> Estimating the depth uncertainty in three-dimensional virtual environment. . . . .	41
5.11.1	Synopsis . . . . .	41
<b>6</b>	<b>Discussion</b> . . . . .	<b>43</b>
6.1	Validating the visual saliency model . . . . .	43
6.2	Proposed group based asymmetry algorithm . . . . .	44
6.3	Proposed robust metric for the evaluation of saliency models . . . . .	45
6.4	Proposed saliency based image compression algorithm . . . . .	48
6.5	Depth estimation in three-dimensional scenes . . . . .	49
6.6	Towards three-dimensional visual saliency . . . . .	51
<b>A</b>	<b>Research papers in full text</b> . . . . .	<b>65</b>
A.1	Analysis of eye fixations data . . . . .	65
A.2	A robust metric for the evaluation of visual saliency models . . . . .	74
A.3	A robust metric for the evaluation of visual saliency models (extended) . . . . .	87
A.4	Validating the visual saliency model . . . . .	97
A.5	Asymmetry as a measure of visual saliency . . . . .	107
A.6	Calculating saliency using the dihedral group $D_4$ . . . . .	118
A.7	What the eye did not see—a fusion approach to image coding . . . . .	132
A.8	What the eye did not see—a fusion approach to image coding (extended) . . . . .	143
A.9	Evaluation of geometric depth estimation model for virtual environment. . . . .	157
A.10	Estimating the depth in three-dimensional virtual environment with feedback. . . . .	170



A.11 Estimating the depth uncertainty in three-dimensional virtual environment . . . . .	180
<b>B Statements of co-authorship</b>	<b>189</b>



# List of Figures

2.1	The picture of Dalmatian sniffing at leaves (credited to Richard Gregory). . . . .	6
2.2	The eye movements for an observer under free viewing conditions, and for six different tasks (Yarbus, 1967). In each case, the observers' viewed the image for a period of 3 minutes. . . . .	10
2.3	The general architecture of the saliency model by (Itti, Koch, & Niebur, 1998). . . . .	11
2.4	(a) Contrast detection filter showing inner square region $R_1$ and outer square region $R_2$ . (b) The width ( $w$ ) of $R_1$ remains constant while that of $R_2$ ranges from $w/2$ to $w/8$ . (c) The image is filtered at one of the scales in a raster scan fashion (Achanta, Estrada, Wils, & Süsstrunk, 2008). . . . .	14
4.1	The general research methodology of design science research, from (Vaishnavi & Kuechler, 2004). . . . .	34
6.1	Probability histograms and relative probabilities for the fixated and non-fixated regions for an average observer. X-axis shows the saliency values obtained by using the visual saliency algorithm (Itti, Koch, & Niebur, 1998). . . . .	43
6.2	Comparison of visual saliency algorithms, both algorithms return the region containing the boat at the center as salient, which is also in agreement with the fixations map obtained from the eye fixations data. . . . .	45
6.3	Eigenvector for an average observer. It shows a concentration of fixations in the center region of the image. . . . .	46
6.4	Ranking of visual saliency models using the shuffled AUC metric. . . . .	47
6.5	Ranking of visual saliency models using the robust AUC metric. . . . .	47
6.6	In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. . . . .	48

6.7	Distributions and histograms of depth estimates for two experiments: without compensatory cue, and with compensatory cue. Depth estimates were calculated using the line-intersection method.	50
6.8	Distributions and histograms of depth estimates for two experiments: without compensatory cue, and with compensatory cue. Depth estimates were calculated using the cone-intersection method.	51
6.9	Number of axes with opposite diagonals like this = 4. We can rotate by 120 or 240 degrees around these axes. These operations give 8 elements.	52
6.10	Number of axes with opposite faces like this = 3. We can either rotate by 90, 180 or 270 degrees around these axes. These operations give 9 elements.	52
6.11	Number of axes with opposite edges like this = 6. We can rotate by 180 degrees around these axes. These operations give 6 elements.	52

# List of Tables

2.1	top-down versus bottom-up (adapted from (Suder & Worgotter, 2000)) . . . . .	8
2.2	The various visual attention models and their categories according to the study by (Borji & Itti, 2013). . . . .	18
3.1	Relations between research papers and research questions. . . . .	32
4.1	Research papers and eye tracking datasets. . . . .	35



# Chapter 1

## Introduction

### 1.1 Motivation

Our visual system is selective, i.e., we concentrate on certain aspects of a scene while neglecting other things. This is evident from studies on change blindness (Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1998; O'Regan, Rensink, & Clark, 1999), that show that large changes can be made in a visual scene which can remain unnoticed. The reason our visual system is selective is because our brains do not process all the visual information in a scene. In fact, while the optic nerve receives information at the rate of approximately  $3 \times 10^6$  bits/sec, the brain processes less than  $10^4$  bits/sec of this information (Anderson, Essen, & Olshausen, 2005). In other words, the brain uses a tiny fraction (less than 1 percent) of the collected information to build a representation of the scene, a representation that is good enough to perform a number of complex activities in the environment such as walking, aiming at objects, and detecting objects. Based on this, we ask what mechanisms are responsible for building this representation of the scene?

In the literature, two main attention mechanisms are discussed: bottom-up and top-down (Braun & Sagi, 1990; Desimone & Duncan, 1995; Steinman & Steinman, 1998; Mozer & Sitton, 1998; Suder & Worgotter, 2000; Itti & Koch, 2001; Navalpakkam & Itti, 2006). Bottom-up factors, also mentioned as visual saliency, are fast, involuntary, and driven by the properties of a visual scene that pop-out. These properties include: color, intensity, orientation, and motion (Koch & Ullman, 1985; Itti, Koch, & Niebur, 1998). For example, a yellow ball on a green background or a flashing light bulb would instantly capture our attention. Top-down factors on the other hand, are voluntary, slower than bottom-up, and driven by task. They involve cognitive aspects such as memory, thought, and reasoning. As an example of top-down, we might consider the problem of locating an item such as the room keys on a table. Here we would be trying to browse the scene in search of an object that best fits the mental description of a key and disregarding other properties of the scene.

In the past two decades, modeling visual saliency has generated a lot of interest in the research community. In addition to contributing towards the understanding of human vision, it has also paved the way for a number of computer

vision applications. These applications include: target detection (Itti & Koch, 2000), image and video compression (Itti, 2004; Yu & Lisin, 2009), image segmentation (Achanta, Estrada, Wils, & Ssstrunk, 2008), robot localization (Sia-gian & Itti, 2007; Frintrop, Jensfelt, & Christensen, 2006), image retrieval (Kadir & Brady, 2001), image and video quality assessment (Feng, Liu, Yang, & Wang, 2008; Ma & Zhang, 2008), dynamic lighting (El-Nasr, Vasilakos, Rao, & Zupko, 2009), advertisement (Rosenholtz, Dorai, & Freeman, 2011), artistic image rendering (Judd, Ehinger, Durand, & Torralba, 2009) and human-robot interaction (Breazeal & Scassellati, 1999; Ajallooeian, Borji, Araabi, Ahmadabadi, & Moradi, 2009). Furthermore, saliency algorithms can be used to identify the image locations that are robust to affine transformations (Lowe, 2004). This is useful for applications such as: image matching and recognition (Lowe, 2004).

A number of visual saliency models are based on the feature integration theory (FIT) proposed by (Treisman & Gelade, 1980). The FIT based models such as (Koch & Ullman, 1985; Itti, Koch, & Niebur, 1998; Itti & Koch, 2000, 2001; Frintrop, 2006a; Walther & Koch, 2006; Harel, Koch, & Perona, 2006), suggest that regions in a scene that are different from their surround with respect to properties such as color, brightness, and orientation, are salient, and these are calculated in parallel. For an image scene this is accomplished by calculating these differences and storing the results in so-called feature maps which are then combined in a saliency map. Thus, the resultant saliency map is a two-dimensional gray-scale map where the brighter regions represent higher saliency.

Although there are a number of visual saliency models in the literature, none of them can fully account for the viewing pattern of observers. In fact, it is well known that the observers agree best with the viewing patterns of other observers. This raises several questions, such as, are visual saliency models good classifiers of so-called salient and non salient regions? Given that the visual saliency models calculate image features such as orienting gradients, color difference and brightness, can we then find a mathematical unified metric that groups these expressions in a mathematical description? Given that such a metric exists, how does it perform as compared to other visual saliency models? Is this metric fast as compared to other visual saliency models? Can this metric be extended to calculate visual saliency of a three-dimensional scene? What are the challenges associated with calculating visual saliency of a three-dimensional scene? This thesis is an attempt to answer such issues.

## 1.2 Outline of the thesis

The dissertation is organized as follows:

**Chapter 1** This chapter describes the motivation behind the thesis.

**Chapter 2** This chapter gives an overview of attention and the relevant mechanisms associated with attention. In addition, we examine the state-of-the-art saliency models and the metrics used for judging the performance of the saliency algorithms.

**Chapter 3** This chapter elaborates on the research issues, the answers found from our analysis, and the future research directions associated with the research papers. Based on this, the main research questions and contributions are outlined.



**Chapter 4** This chapter introduces the research methodology employed for the research effort.

**Chapter 5** This chapter gives an overview of the results from the research papers and highlights the main direction for future work.

**Appendix A** This chapter contains the complete research papers.



# Chapter 2

## Background

### 2.1 Attention

#### 2.1.1 Psychological and philosophical perspective

In psychology and prior disciplines, attention has been described in different ways. For instance, Hobbes (1655) suggested (Itti, Rees, & Tsotsos, 2005): *“While the sense organs are occupied with one object, they cannot simultaneously be moved by another so that an image of both arises. There cannot therefore be two images of two objects but one put together from the action of both.”* In contrast to this view, Hamilton (1788-1856) argued that people can attend to more than one object at a time (Johnson & Proctor, 2004). Hamilton’s view was supported by the findings of Jevons (1871), who estimated the number of objects to be four. James (1890) suggested a plain language definition of attention as: *“Everybody knows what attention is. It is taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought.”* In contrast, Groos (1896), believed that: *“To the question, what is attention, there is not only no generally recognized answer, but the different attempts at a solution even diverge in the most disturbing manner.”* For a complete overview of various definitions, one can refer to (Itti, Rees, & Tsotsos, 2005). Thus from this perspective, it is safe to say that attention is not a unitary concept, but, a collection of different mechanisms which enable us to understand and interact with the environment (Styles, 2005). Although, we have a vague notion of what we mean when we attend to something, what we attend to in one situation can vastly differ in another situation (Styles, 2005).

#### 2.1.2 Computational perspective

It is well established that the brain does not process all the visual information in the environment. In this way, it is comparable to an information processing unit with limited capacity, i.e., less than  $10^4$  bits/sec. While viewing an image, each part of the scene can be matched to many different objects or scenes in the memory, and the number of these part-to-object combinations can exceed the processing capacity of the brain (Tsotsos, 2011). This can be illustrated by

the picture of “*Dalmatian sniffing at leaves*” (credited to Richard Gregory). As shown in figure 2.1, the picture is reasonably complex such that each part of the image is either strongly or weakly related to a number of other possible objects, thus, leading to a large number of possible interpretations for the brain to choose from. This raises the question, how does the brain manages to successfully interpret such a vast amount of visual information? According to (Tsotsos, 2011), this can be explained by employing a computational approach to visual attention. In this approach, it is assumed that while looking the brain is not solving a generic viewing problem, but, instead the problem is reshaped through approximations such that it can be solved by using the available processing power for vision. Although, the term computational can be used to describe any computer simulated mathematical model that uses equations to solve the problem, Marr (2010) proposed that a computational model should be described at three levels of analysis defined as: computational, algorithmic, and implementation. At the computational level, the system should explain what problem is solved by it and why does it solve that problem. At the algorithmic level, the system should describe how the problem is solved, and what mathematical or machine learning methods are employed to solve the problem. Finally, at the implementation level, the system should define the physical mechanism used to perform these calculations and its structure.



Figure 2.1: The picture of Dalmatian sniffing at leaves (credited to Richard Gregory).

## 2.2 Attention mechanisms

### 2.2.1 Selective attention

Selective attention is defined as the ability to focus on a specific aspect of a scene while ignoring other factors. It is often compared to the spotlight model of attention (Posner, Snyder, & Davidson, 1980; Eriksen & St. James, 1986), which suggests that the information from a scene is extracted in the form of a spotlight of arbitrary radius, which can shift from one location to another either involuntarily or voluntarily. In addition to this, it is assumed that the information is acquired by the spotlight shifts in a serial manner. Selective attention enables us to engage with our surroundings in an intelligent manner to

perform activities that use visual information. The activities range from reading, walking, aiming to safely driving a car on the road.

### 2.2.2 Overt and covert attention

In overt attention mechanism, the information from a scene is selected by explicit movement of the sensory organs (Geisler & Cormack, 2011; Johnson & Proctor, 2004). For instance, our visual system can dynamically focus on a region of interest by moving the eyes. Here, it is natural to assume that attention is linked to the center of the focus. However, this assumption is not always valid, as covert attention does not involve explicit movement of the sensory organs (Geisler & Cormack, 2011). Covert attention is often compared to observing something out of the corner of the eye without focusing at it. This mechanism was introduced in the study by (von Helmholtz, 1860 / 1962), which suggested that it is possible to attend to different regions of an image on the retina without eye movements. For example, while holding our focus steady on a word in the text such as “this”, we can read the words on different spatial locations like on the lines above and below. The mechanisms associated with the overt and covert attentions normally work together (Frintrop, 2006b). In fact in most cases, prior to eye movement on a target location, the covert attention shifts to this location (Tsotsos, 2011). Studies have shown that covert attention enhances the visual information at a particular location in the scene, which leads to faster discrimination of objects (Carrasco, 2011). While covert attention can be measured by using reaction times in target detection or neurobiological methods such as changes in firing rates of single cells, overt attention is usually measured by employing eye trackers (Frintrop, 2006b).

### 2.2.3 Bottom-up and top-down attention

Visual attention can be classified as: top-down, and bottom-up. Top-down, is voluntary, goal-driven, and slow, i.e., typically in the range between 100 milliseconds to several seconds (Suder & Worgotter, 2000). It is assumed that the top-down attention is closely linked with cognitive aspects such as memory, thought, and reasoning. For example, by employing top-down mechanisms, we can attend to a person sitting next to us in a busy restaurant and neglect other people and visual information in the background. In contrast, bottom-up attention also known as visual saliency is associated with attributes of a scene that draw our attention to a particular location. These attributes include: motion, contrast, orientation, brightness, and color (Koch & Ullman, 1985). Bottom-up mechanisms are involuntary, and faster as compared to top-down (Suder & Worgotter, 2000). For instance, flickering lights, a yellow target among green objects, and a horizontal target among vertical objects are some stimuli that would automatically capture our attention in the environment. Studies (Chun & Wolfe, 2001; Wolfe, Butcher, Lee, & Hyle, 2003) show that in search tasks, such as looking for a target object among distractors both bottom-up and top-down mechanisms work together to guide our attention. While bottom-up attention is based on elementary attributes of a scene, top-down is quite complex and strongly influenced by task demands (Jasso & Triesch, 2008). For example, studies by (Land, Mennie, & Rusted, 1999; Pelz, Hayhoe, & Loeber, 2001)

suggest that for tasks such as picking up and placing objects by hand in the environment, attention is mainly driven by top-down mechanisms.

The differences between the top-down and bottom-up mechanisms are summarized in table 2.1.

Table 2.1: top-down versus bottom-up (adapted from (Suder & Worgotter, 2000))

	<b>top-down</b>	<b>bottom-up</b>
driven by	task or cognition	visual stimuli
controlled by	conscious, voluntary	unconscious, involuntary
time scale	sustained ( 100 ms to several seconds)	transient ( 0 to 300 ms)
responsible for	searching and highlighting	pop-out effects

## 2.3 Attention and eye movements

Visual attention can be studied by analyzing eye movements. This can be explained by the classic example from (Yarbus, 1967), where an image depicting an unexpected visitor arriving in a Victorian living room is shown to an observer under free viewing conditions, and for six different tasks. The tasks given to the observer were:

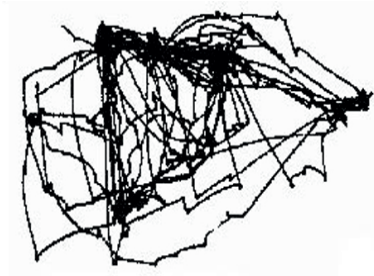
1. Estimate the economic status of the family.
2. Estimate the ages of the people.
3. Infer what the family was doing before the arrival of the visitor.
4. Remember the clothes worn by the people.
5. Remember the locations of people and objects in the room.
6. Estimate how long has the unexpected visitor been away from the family.

As shown in figure 2.2, the eye movements associated with different tasks were different, suggesting that eye movements reflect the observers thought process, i.e., the reason for looking at an image strongly influences the parts or the objects the observer looks at (Tsotsos, 2011). Similar observations were made in the study by (Just & Carpenter, 1976) leading to the formulation of eye-mind hypothesis. The eye-mind hypothesis suggests that where an observer is looking in the scene indicates what he or she is processing and the duration of this reflects how much processing effort is needed.

The eye movements can be broadly classified as fixations and saccades. Fixation is defined as the momentary pause of the eye on a location in the scene, while saccade is the rapid eye movement that usually occurs in between fixations. It is assumed that by using eye fixations the brain acquires most of the visual information and no useful information is taken in during saccades (Henderson, 2003).



(a) Picture



(b) Free viewing



(c) Estimate the economic status of the family.



(d) Estimate the ages of the people.



(e) Infer what the family was doing before the arrival of the visitor.



(f) Remember the clothes worn by the people.



(g) Remember the locations of people and objects in the room.



(h) Estimate how long has the unexpected visitor been away from the family.

Figure 2.2: The eye movements for an observer under free viewing conditions, and for six different tasks (Yarbus, 1967). In each case, the observers' viewed the image for a period of 3 minutes.



## 2.4 State-of-the-art in modeling visual attention

In this section, the computer models for predicting eye fixations in still images are discussed. The models are presented in a chronological order.

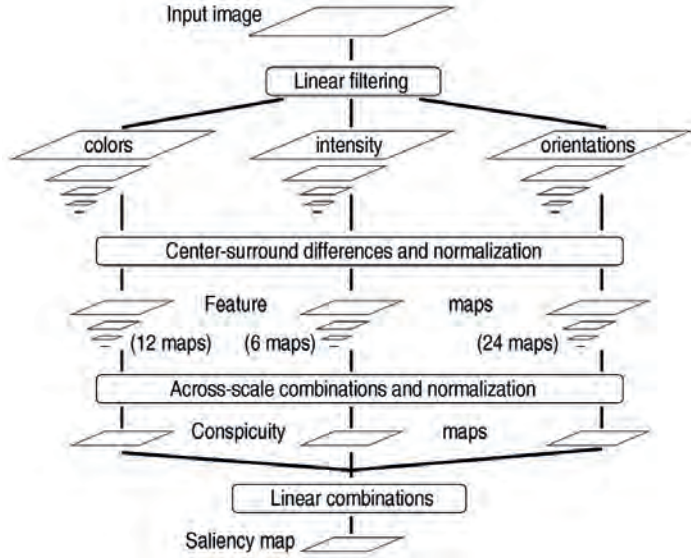


Figure 2.3: The general architecture of the saliency model by (Itti, Koch, & Niebur, 1998).

The classic model of visual saliency proposed by (Itti, Koch, & Niebur, 1998), calculates salient regions by decomposing the input image in three different channels namely, color, intensity, and orientation as shown in figure 2.3. The color channel consists of two maps: red/green and blue/yellow color opponencies, the intensity channel consists of a gray scale representation of the input image, and the orientation channel contains four local orientation maps associated with angles 0, 45, 90, and 135 degrees. For each channel map nine spatial scales are created by repeatedly low pass filtering and sub-sampling the input channel. After that, feature maps are computed by using center-surround operations, which are inspired by vision studies such as (Treisman & Gelade, 1980; Koch & Ullman, 1985). The center-surround operations are defined as the difference between fine and coarse scales. For example, if the center is a pixel at scale  $c \in \{2, 3, 4\}$ , the surround is the corresponding pixel at scale  $s = c + d$ , with  $d \in \{2, 3\}$ , and  $\ominus$  denotes the across scale difference, then the center-surround feature maps for a channel  $I$  are represented as:

$$I(c, s) = |I(c) \ominus I(s)|.$$

These operations generate 42 feature maps: six for intensity, 12 for color opponencies, and 24 for orientation. Finally, the resulting feature maps from different channels are normalized and combined linearly to get the so-called saliency map. The VOCUS model proposed by (Frintrop, 2006b), and the saliency toolbox im-

plemented by (Walther, 2006; Walther, Itti, Riesenhuber, Poggio, & Koch, 2002) are based on this saliency model.

Lee & Yu (2000) proposed a theoretical model based on the assumption that our visual system operates on the principle of information maximization, i.e., we fixate at a location in the image that provides maximum amount of information. They proposed that mutual information among cortical representations of the retinal image, the priors constructed from our long term visual experience, and a dynamic short term internal representation constructed from recent saccades, provides the map for the guidance of eye movements. Based on this approach, a similar model was defined by (Renninger, Coughlan, Verghese, & Malik, 2005).

Rao et al. (2002) introduced a model that uses a top-down search template matching approach to locate the salient regions. In their model, first, a saliency map is obtained from the input image by employing oriented spatiochromatic filters. After that, a template of the desired target object is moved across different regions of the saliency map, and the similarity between a selected region and the target is measured by calculating their euclidean distance. Finally, the  $N$  most similar regions are represented as salient.

Torralba (2003) and Oliva et al. (2003) defined a model that combines three factors: bottom-up saliency, object likelihood, and contextual prior. The local saliency is calculated as:  $S(x) = \frac{1}{p(v_L/v_C)}$ , where  $v_L$  encodes local features in the neighborhood of a location which is represented by the outputs of multi-scale oriented bandpass filters, and  $v_C$  represents the contextual properties of the scene or background which include: global image statistics, color histograms, and wavelet histograms. In the object likelihood factor, the locations corresponding to features different from the target object are suppressed, and the locations with similar features are maintained. The contextual prior stage uses the past search experience from similar images and the strategies that were successful in finding the target.

Bruce & Tsotsos (2005) introduced a saliency model based on the principle of maximizing information that uses Shannon’s self information measure. The saliency is defined by the self information associated with each local image region. The self information is given by:  $I(X) = -\log(p(X))$ , where  $X$  is a  $n$ -dimensional feature vector extracted from an image region, and  $p(X)$  is the probability of observing  $X$  based on its surround. The authors state that there is insufficient data in a single image to provide a reasonable estimate of the probability distribution. To address this issue, they employ independent component analysis (ICA) in order to learn the bases from a large database of natural images. After that, the probability of observing an image region is calculated for each basis coefficient. Finally, for a given image region the likelihood of observing it is represented by the product of corresponding ICA basis probabilities for that region.

Harel et al. (2006) proposed a bottom-up model that uses graph algorithms for saliency computations. In their model, the first step comprises of calculating feature maps using a procedure similar to (Itti, Koch, & Niebur, 1998). After that, a fully connected graph for the locations of the feature maps is build. A graph comprises of nodes or vertices connected by links or edges. The weights between two nodes are calculated based on their dissimilarity and their closeness. Given two locations  $(i, j)$  and  $(p, q)$  in the feature map  $M$ , the dissimilarity

between their respective nodes  $M(i, j), M(p, q)$  is defined as:

$$d((i, j) || (p, q)) \triangleq \left| \log \frac{M(i, j)}{M(p, q)} \right|.$$

Next, the graphs obtained are treated as Markov chains, and the equilibrium distribution of these chains are adopted as the activation maps. Finally, these activation maps are normalized using another Markovian algorithm to highlight the conspicuity, and admitting their combination to form the saliency map.

Meur et al. (2006) presented a saliency model inspired by various properties of human visual system such as: contrast sensitivity function, visual masking, and perceptual grouping. This model is based on the saliency framework proposed by (Koch & Ullman, 1985), and the saliency map is build by linearly combining the different feature maps. The authors showed that their model outperforms the saliency model proposed by (Itti, Koch, & Niebur, 1998).

Navalpakkam & Itti (2006) introduced a model that combines top-down and bottom-up aspects of attention. The bottom-up component is calculated by using the saliency model by (Itti, Koch, & Niebur, 1998), and the top-down component uses the information about the target and the background objects to maximize the ratio between the saliency values of the targets to that of the background objects. This model was evaluated using a search task, i.e., the observers were instructed to search for a specific object in the scene. Their results showed that a combined top-down and bottom-up model yields faster search than a bottom-up model.

Hou & Zhang (2007) proposed a saliency model based on analyzing the log spectrum of the input image. First, the log spectrum is defined as:  $L(f) = \log(A(f))$ , where  $A(f)$  is the amplitude of the Fourier spectrum of the image. After computing the log spectrum, the spectral residue is calculated as:  $R(f) = L(f) - A(f)$ . Finally, the spectral residue is transformed to spatial domain to get the saliency map. The results from the authors suggested that their model predicts the fixations better than the saliency model by (Itti, Koch, & Niebur, 1998).

Mancas (2007) defined saliency as a measure of two components: contrast, and rarity, i.e., rare features in an image are interesting.. To account for contrast two methods are proposed: global and local. Global contrast is measured using histogram, and local contrast is calculated using center-surround operations similar to that of (Itti, Koch, & Niebur, 1998). The rarity is quantified by employing Shannon’s self-information measure. First, a low level saliency map is calculated by describing each location by the mean and the variance of its neighborhood. After that, rarity is measured based on the features such as size and orientation, where smaller areas and lines corresponding to the orientations get higher saliency values on the saliency map. Finally, high-level methods such as Gestalt laws of grouping are employed to find the salient regions.

Cerf et al. (2007) proposed a model that combined the bottom-up feature channels of color, intensity, and orientation, from (Itti, Koch, & Niebur, 1998), with a face-detection channel, based on the algorithm by (Viola & Jones, 2001). Their results showed that the combined model improves the correspondence between the fixated and the salient image regions.

The SUN model by (Zhang, Tong, Marks, Shan, & Cottrell, 2008), defined saliency as a combination of three components: the first contains self information, which depends only on the visual features at a location. Here, rarer

features are considered more informative. In the second, top-down information such as the knowledge about the attributes of the target is used to obtain a log likelihood. The third component, consists of the probability associated with the knowledge of the location of the target. In their algorithm, the saliency map was calculated using difference of Gaussians and independent component analysis derived features.

Rajashekar et al. (2008) proposed a bottom-up model that calculates salient image regions based on four foveated low-level image features, namely, luminance, contrast, luminance-bandpass, and contrast-bandpass. The input image is divided into uniform regions, and the feature maps associated with the four low level features are calculated. Finally, the four maps are linearly combined using a weighted average to get the saliency map. For evaluation, they used 101 static gray-scale images that contained no high level features such as animals, faces, and other items of high-level semantic interest.

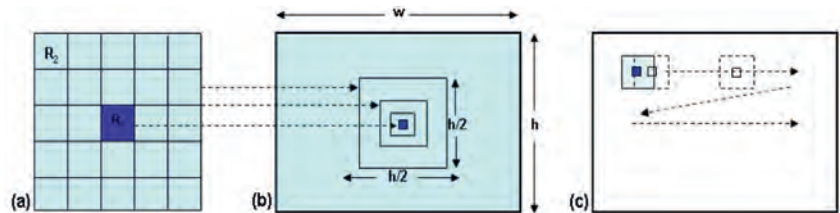


Figure 2.4: (a) Contrast detection filter showing inner square region  $R_1$  and outer square region  $R_2$ . (b) The width ( $w$ ) of  $R_1$  remains constant while that of  $R_2$  ranges from  $w/2$  to  $w/8$ . (c) The image is filtered at one of the scales in a raster scan fashion (Achanta, Estrada, Wils, & Ssstrunk, 2008).

Achanta et al. (2008) presented a model that represents saliency as the local contrast of an image pixel with respect to its neighborhood at different scales. For a given scale, the saliency value at a pixel  $(i, j)$  is calculated as the distance  $D$  between the mean vectors of pixel features of the inner region  $R_1$  and the outer region  $R_2$  as:

$$c_{i,j} = D \left[ \left( \frac{1}{N_1} \sum_{p=1}^{N_1} v_p \right), \left( \frac{1}{N_2} \sum_{q=1}^{N_2} v_q \right) \right],$$

where  $N_1$ , and  $N_2$  are the number of pixels associated with the regions  $R_1$  and  $R_2$  as depicted in figure 2.4. In their model *CIELAB* color space is used to generate feature vectors for color and luminance. The final saliency map is obtained by summing the saliency values across the different scales.

Guo et al. (2008) calculated saliency in a manner similar to the spectral residue approach by (Hou & Zhang, 2007), with the exception that this model excludes the computation of spectral residue in the amplitude spectrum. They state that by excluding the amplitude computation the saliency map is obtained faster. For a given image  $I(x, y)$ , the saliency map is defined as:

$$sM(x, y) = g(x, y) * \|F^{-1}[e^{i \cdot p(x,y)}]\|^2,$$

such that  $f(x, y) = F(I(x, y))$  and  $p(x, y) = P(f(x, y))$ , where  $F$  and  $F^{-1}$  represent Fourier Transform and Inverse Fourier Transform respectively.  $P(f)$

denotes the phase spectrum of the image, and  $g(x, y)$  is a two-dimensional Gaussian filter.

Gao et al. (2009) defined saliency as equivalent to discrimination, i.e., they state that the most salient features are the ones that best separate the target class from all others. In their model saliency is represented by two components: feature selection and saliency detection. The best feature subset is selected by computing the marginal mutual informations as:

$$I(X; Y) = \sum_i P_Y(i) D_{KL}(P_{X|Y}(x|i) || P_X(x)),$$

where  $X$  is a set of features, and  $Y$  is a class label with prior probabilities  $P_Y(i)$ , such that the probability density of  $X_k$  given class  $i$  is  $P_{X_k|Y}(x|i)$ , and  $D_{KL}$  is the Kullback-Leibler divergence (Wilming, Betz, Kietzmann, & Konig, 2011). In the saliency detection, the features that are considered highly non-salient are eliminated by employing Barlows principle of inference (Barlow, 1987).

Judd et al. (2009) used a machine learning approach to train a combined bottom-up, top-down model based on low, mid, and high-level image features. In their model, the low level features are described by models such as (Itti, Koch, & Niebur, 1998; Rosenholtz, 1999; Oliva & Torralba, 2001), the mid level features are represented by a horizon line detector, and the high level features consist of people and face detectors. The authors collected eye fixations of 15 observers from a comprehensive dataset (with 1003 images) which was also used for evaluation. The model proposed by the authors showed better correspondence with the fixations than several other models such as (Itti, Koch, & Niebur, 1998; Rosenholtz, 1999; Oliva & Torralba, 2001; Cerf, Harel, Einhauser, & Koch, 2007).

Seo & Milanfar (2009) introduced a bottom-up model based on self resemblance measure. In their model, image features are obtained by using local regression kernels, which are quite robust to noise and efficient at capturing the underlying structure of the image. After that, matrix cosine similarity is used to compute the resemblance of each location to its surroundings. The saliency for a given location  $i$  is represented as:

$$S_i = \frac{1}{\sum_{j=1}^N \exp(\frac{-1+\rho(F_i, F_j)}{\sigma^2})},$$

where  $\sigma$  is a weight parameter, and  $\rho(F_i, F_j)$  is the matrix cosine similarity between two feature maps  $F_i$ , and  $F_j$ . Here the matrix cosine similarity is defined as Frobenius inner product between two normalized matrices  $F_i$ , and  $F_j$ . The authors showed that their model predicts fixations better than the models by (Bruce & Tsotsos, 2005; Zhang, Tong, Marks, Shan, & Cottrell, 2008).

Bian & Zhang (2009) adopted a spectral approach similar to (Guo, Ma, & Zhang, 2008) for calculating salient image regions. In their model, the input image is resized to a fixed scale, and a windowed Fourier transform of the image is calculated to get a spectral response. The spectral response denoted by  $f(u, v)$  is then normalized as:  $n(u, v) = f(u, v) / \|f(u, v)\|$ . After that,  $n(u, v)$  is transformed to spatial domain by using inverse Fourier transform followed by squaring to promote the salient regions. The resultant saliency map is convolved with a Gaussian filter  $g$  to model the spatial pooling operations of complex

cells as:  $S(x, y) = g(u, v) * \|F^{-1}[n(u, v)]\|$ , where  $F^{-1}$  denotes inverse Fourier transform.

Kienzle et al. (2009) proposed a non-linear machine learning approach for calculating saliency. In their model, the intensities pertaining to local image regions are used as feature vectors. The authors employ support vector machine to train the feature vectors of fixated regions to yield positive values and the feature vectors of randomly selected regions to negative values. The resultant saliency is modeled with four perceptive fields, two most likely image structures and two least likely patterns for driving fixations. For the training and evaluation a dataset of 200 gray scale images was used.

Chikkerur et al. (2010) presented a Bayesian model of attention based on the concept that the task of the visual system is to recognize what is where and this is achieved by localizing sequentially, i.e, one object at a time. Their model extends the template based approach used in the model by Rao et al. (2002), in the following ways: first, both feature and object priors are included, which allows to combine top-down feature-based attention and spatial attention. Second, this model allows a combination of  $N$  feature vectors that share common spatial modulation. Third, in the spatial attention, scale/size information is used in addition to the location information. The authors state that their model combines bottom-up, feature-based, and context-based attention mechanisms, and in so doing it is able to explain part of the basic functional anatomy of attention.

Li et al. (2010) introduced a model that measures saliency as minimum conditional entropy. In their model, the minimum conditional entropy represents the uncertainty of the center-surround local region, when the surrounding area is given and the perceptual distortion is considered. The authors state that the larger the uncertainty the more salient the center is, and vice versa. The minimum conditional entropy is approximated by the lossy coding length of Gaussian data. Finally, the saliency map is segmented by thresholding to detect the salient objects. In their results it was shown that their model outperforms the saliency model by (Itti, Koch, & Niebur, 1998).

Goferman et al. (2010) proposed a context aware saliency model based on four principles of visual attention: first, low level attributes such as contrast, and color. Second, global considerations, which suppress frequently occurring features, while maintaining features that deviate from the norm. Third, visual organization rules, which state that visual forms may possess one or several centers of gravity about which the form is organized. Four, high-level factors, such as human faces. Their results showed that the context aware saliency model performs better than the models by (Walther & Koch, 2006; Hou & Zhang, 2007).

Avraham & Lindenbaum (2010) presented a stochastic model of visual saliency. In their model, first, the input image is segmented into regions which are considered as candidates for attention. An initial probability for each candidate is set using preferences such as small number of expected targets. After that each candidate is represented by a feature vector, and visual similarity between every two candidates is evaluated using Pearson correlation coefficient. Next, a tree based Bayesian network is employed for clustering the candidates. Finally, the saliency map is obtained by selecting the most likely candidates.

Liu et al. (2011) introduced a supervised approach to calculating salient image regions. The salient object detection is formulated as an image segmen-

tation problem, where the objective is to separate the salient object from the image background. To do this in their model, ground truth salient objects are obtained from the regions labeled by the observers as salient. After that, a set of features including multi-scale contrast, center-surround histogram, and color spatial distribution are used to describe a salient object locally, regionally, and globally. Finally, these features are optimally combined through Conditional Random Field (CRF) learning. The CRF was trained and evaluated for a large dataset containing 20,840 labeled images by multiple users.

Kootstra et al. (2011) proposed a model that calculates saliency on the basis of symmetry. In their model, three local symmetry operators namely, isotropic symmetry (Reisfeld, Wolfson, & Yeshurun, 1995), radial symmetry (Reisfeld, Wolfson, & Yeshurun, 1995), and color symmetry (Heidemann, 2004) are defined. These three symmetry features are calculated at five image scales. The resulting saliency map is obtained by normalizing and combining the feature maps. For the evaluation of this model, the authors used a dataset containing 99 images belonging to different categories such as natural symmetries, animals, street scenes, buildings, and natural environments. The authors showed that their symmetry model outperforms the saliency model by (Itti, Koch, & Niebur, 1998) in predicting the eye fixations.

Murray et al. (2011) calculated salient image regions in three steps: first, the input image is processed according to operations consistent with early visual pathway (color-opponent and luminance channels, followed by a multi-scale decomposition). Second, a simulation of the inhibition mechanisms present in cells of the visual cortex is performed, this step effectively normalizes their response to stimulus contrast. Third, the model integrates information at multiple scales by performing an inverse wavelet transform directly on weights computed from the non-linearization of the cortical outputs. Their saliency model showed better correspondence with the fixations than the saliency models by (Bruce & Tsotsos, 2005; Seo & Milanfar, 2009).

Wang et al. (2011) proposed a computational model based on the principle of information maximization. Their model considers three key factors, namely, reference sensory responses, fovea-periphery resolution discrepancy, and visual working memory. In their model, first, three multi-band filter response maps are calculated as a coherent representation for the three factors. After that, the three filter response maps are combined into multi-band residual filter response maps. Finally, the saliency map is obtained by calculating the residual perceptual information at each location. The results from the authors showed that their model performs significantly better than the saliency model by (Itti, Koch, & Niebur, 1998).

Garcia-Diaz et al. (2012) introduced a saliency model based on adaptive whitening of color image and feature maps. First the input image is transformed from  $(r, g, b)$  to  $(z_1, z_2, z_3)$ , a whitened representation. The whitening is done through de-correlation by employing principal component analysis. The feature maps are calculated for  $(z_1, z_2, z_3)$  using a bank of log-Gabor filters for orientations  $(0^\circ, 45^\circ, 90^\circ, 135^\circ)$ , and seven scales are calculated for  $z_1$  and only five for  $z_2$ , and  $z_3$ . Next, for each chromatic component the feature maps are whitened and contrast normalization is performed in several steps in a hierarchical manner. Saliency is computed as the square of the vector norm in the resulting representation. The authors showed that their model outperforms the state-of-the-art models in predicting fixations. These results were confirmed in



an independent study by (Borji, Sihite, & Itti, 2013), which concluded that the saliency model by (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012) is the top performing model for natural images.

Table 2.2: The various visual attention models and their categories according to the study by (Borji & Itti, 2013).

<b>Bayesian models</b>	Torralba (2003), Oliva et al. (2003), and Zhang et al. (2008)
<b>Cognitive models</b>	Itti et al. (1998), Walther (2006), Walther et al. (2002), Frintrop (2006b), Meur et al. (2006), Rajashekar et al. (2008), and Cerf et al. (2007)
<b>Decision theoretic models</b>	Gao & Vasconcelos (2004), Gao et al. (2009), Li et al. (2010), and Wang et al. (2011)
<b>Graphical models</b>	Harel et al. (2006), Achanta et al. (2008), Avraham & Lindenbaum (2010), Chikkerur et al. (2010), and Liu et al. (2011)
<b>Information theoretic models</b>	Bruce & Tsotsos (2005), Mancas (2007), and Seo & Milanfar (2009)
<b>Pattern classification models</b>	Judd et al. (2009), and Kienzle et al. (2009)
<b>Spectral analysis models</b>	Hou & Zhang (2007), Guo et al. (2008), Achanta et al. (2008), and Bian & Zhang (2009)
<b>Other models</b>	Rao et al. (2002), Goferman et al. (2010), and Garcia-Diaz et al. (2012)

In the study by (Borji & Itti, 2013), the authors state that the visual attention models in the literature can be divided into eight classes: bayesian, cognitive, decision theoretic, graphical, information theoretic, pattern classification, spectral analysis, and others. The different classes and the attention models associated with these classes are shown in table 2.2. In bayesian models, prior knowledge about the scene, and sensory information such as target features are employed to calculate salient image regions. For instance, the models such as Torralba (2003), Oliva et al. (2003), and Zhang et al. (2008) fall in this category. Cognitive models are the ones that are strongly based on psychological and neurophysiological findings. This category includes models such as Itti et al. (1998), Walther (2006), Walther et al. (2002), Frintrop (2006b), Meur et al. (2006), Rajashekar et al. (2008), and Cerf et al. (2007). Decision theoretic models are based on the concept of identifying the optimal factors based on which people make decisions. For instance, models such as Gao & Vasconcelos (2004), Gao et al. (2009), Li et al. (2010), and Wang et al. (2011) are classified under this category. A graphical model is a probabilistic model in which graphs are used to represent probabilistic relationships between different variables. For example, models such as Harel et al. (2006), Achanta et al. (2008), Avraham & Lindenbaum (2010), Chikkerur et al. (2010), and Liu et al. (2011) belong to this class. Information theoretic models are based on the concept that localized



saliency computation serves to maximize information sampled from one’s environment. In other words, these models select the most informative parts of the image and discard the rest. This class consists of models such as Bruce & Tsotsos (2005), Mancas (2007), and Seo & Milanfar (2009). In pattern classification models, a machine learning procedure is employed to model visual attention. For the learning, typically eye fixations data or labeled salient regions are used. For instance, models such as Judd et al. (2009), and Kienzle et al. (2009) are classified under this category. Spectral analysis models calculate saliency in the frequency domain. This category consists of models such as Hou & Zhang (2007), Guo et al. (2008), Achanta et al. (2008), and Bian & Zhang (2009). The models that do not conform to the above categories are classified as other models. This class includes models such as Rao et al. (2002), Goferman et al. (2010), and Garcia-Diaz et al. (2012).

## 2.5 Summary

In a comprehensive study by (Borji, Sihite, & Itti, 2013), 35 state-of-the-art visual saliency models were evaluated for 54 challenging synthetic patterns, three natural image datasets, and two video datasets. For the evaluation the authors employed three metrics namely, correlation coefficient, normalized scan-path saliency, and shuffled *AUC*. Their results suggest: first, all existing databases are highly center-biased and there is a need to develop datasets that are less center-biased. Second, the correlation coefficient and normalized scan-path saliency metrics suffer from the influences of the center-bias and the authors discourage their use in future model evaluations. Third, the feature integration theory based models such as the classic saliency model by (Itti, Koch, & Niebur, 1998), the saliency toolbox implemented by (Walther, 2006; Walther, Itti, Riesenhuber, Poggio, & Koch, 2002), the GBVS model proposed by (Harel, Koch, & Perona, 2006), the saliency model proposed by (Bian & Zhang, 2009), the VOCUS model by (Frintrop, 2006a) and the *AWS* model by (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012) work better in locating a target over synthetic images. Four, the best model for static and dynamic images is the *AWS* model proposed by (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012). In addition to this, the *AWS* model performed second best with synthetic patterns. Five, models such as those proposed by (Itti, Koch, & Niebur, 1998; Torralba, 2003; Hou & Zhang, 2007; Bian & Zhang, 2009; Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012) are fast and effective in predicting fixations. In fact their results suggest that among the models implemented by using Matlab, the model introduced by (Hou & Zhang, 2007) is the fastest (0.30 sec.), while the model proposed by (Judd, Ehinger, Durand, & Torralba, 2009) is the slowest (98.58 sec.). Six, there is still a gap between current saliency algorithms and inter-observer performance, where inter-observer performance is defined as the level of agreement between the fixations of an observer viewing an image and the fixations of other observers viewing the same image. The authors suggest that the performance of the saliency models can be improved by the addition of top-down factors such as task and semantic cues (i.e., faces, people, and text).

## 2.6 Evaluation metrics

In the literature, various metrics have been employed to measure the performance of saliency models. In this section, these metrics are briefly discussed.

### Area under the receiver-operating-characteristic curve (*AUC*)

*AUC* (Fawcett, 2004; Borji & Itti, 2013) is commonly employed in vision studies to evaluate the correspondence between fixated regions and salient image regions predicted by visual saliency models. For this, the fixations pertaining to a given image are averaged into a single two dimensional map which is then convolved with a two dimensional Gaussian filter. The resultant fixations map is then thresholded to yield a binary map with two classes—the positive class consisting of fixated regions, and the negative class consisting of non-fixated regions. Next, from the two dimensional saliency map, we obtain the saliency values associated with the positive and negative classes. Using the saliency values, a receiver-operating-characteristic (*ROC*) curve is drawn that plots the true positive rate against the false positive rate. The area under the *ROC* curve gives us a measure of the performance of the classifier. *AUC* gives a scalar value in the interval  $[0,1]$ . If *AUC* is 1 then it indicates that the saliency model is perfect in predicting fixations. An *AUC* of 0.5 implies that the performance of the saliency model is not better than a random classifier or by chance prediction. For a detailed description of *AUC*, see the study by (Fawcett, 2004).

### Chance adjusted salience

Chance adjusted salience (Kienzle, Franz, Schlkopf, & Wichmann, 2009; Wilming, Betz, Kietzmann, & Konig, 2011) is calculated by the difference between the mean saliency values of two sets of image regions, the first set consists of parts that are fixated by an observer and the second consists of non-fixated parts. The non-fixated parts are selected from the fixations of the observer for an unrelated image. If the difference value obtained is greater than zero then it suggests that the saliency model is better than a random classifier. The range of this metric is governed by the interval of saliency values which can be arbitrary.

### Eightieth percentile measure

To calculate eightieth percentile measure the saliency maps are thresholded to top 20 percent of the salient image locations (Torralba, Castelhana, Oliva, & Henderson, 2006; Wilming, Betz, Kietzmann, & Konig, 2011). After that, the percentage of fixations falling inside these locations are calculated. In this way, this measure calculates the true positive rate of a classifier that uses eightieth percentile as threshold for the saliency values (Wilming, Betz, Kietzmann, & Konig, 2011). This evaluation metric gives a scalar value in the range  $[0,100]$ .

### Kullback Leibler divergence ( $D_{KL}$ )

$D_{KL}$  (Itti & Baldi, 2009; Wilming, Betz, Kietzmann, & Konig, 2011) is a measure of logarithmic distance between two probability distributions. For evaluat-

ing saliency models, it is calculated as:

$$D_{KL}(P\|Q) = \sum_i P(i) \ln \left( \frac{P(i)}{Q(i)} \right),$$

where  $P$  is the fixations probability distribution, i.e., the fixations map normalized in the interval  $[0,1]$  and  $Q$  refers to the normalized saliency map. As  $D_{KL}$  is not a symmetric measure, i.e.,  $D_{KL} \neq D_{KL}$ , a symmetric version of  $D_{KL}$  is calculated as:

$$KL = D_{KL}(P\|Q) + D_{KL}(Q\|P).$$

A  $KL$  value of zero indicates that the saliency model is perfect in predicting fixations. The  $KL$  metric does not have a well defined upper bound, thus its interval is  $[0,\infty)$ .

### Normalized scan-path saliency ( $NSS$ )

$NSS$  (Peters, Iyer, Itti, & Koch, 2005; Wilming, Betz, Kietzmann, & Konig, 2011) is calculated by normalizing the saliency maps such that the saliency values have zero mean and unit standard deviation. After that, the mean of the saliency values for the fixated regions is calculated. A  $NSS$  value greater than zero suggests that the saliency model shows better correspondence with the fixations than a random classifier. If  $NSS$  is less than or equal to zero then it implies that the prediction by the saliency model is not better than chance prediction. For a detailed insight on the  $NSS$  metric, see the study by (Peters, Iyer, Itti, & Koch, 2005).

### Pearson correlation coefficient

Pearson correlation coefficient (Hwang, Higgins, & Pomplun, 2009; Wilming, Betz, Kietzmann, & Konig, 2011) is a measure of linear dependence between two variables. It is calculated as:

$$r = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}},$$

where  $X$ , and  $Y$  are the two variables,  $\bar{X}$ , and  $\bar{Y}$  are the sample means, and  $r$  is the correlation coefficient.  $r$  returns a value in the range  $[-1,1]$ . If  $r$  is 1 then it suggests a perfect prediction of the fixated regions by the saliency model, while a value of -1 implies that the predicted regions are the exact opposite of the fixations. A value of 0 suggests that there is no linear relation between the salient image regions and the fixated regions.

### Ratio of medians

To calculate ratio of medians (Parikh, Itti, & Weiland, 2010; Wilming, Betz, Kietzmann, & Konig, 2011), two sets of saliency values are selected, the first set consists of the saliency values of the fixated regions and second pertains to the saliency values of regions chosen from random points on the image. The saliency value for a fixation point is calculated as the maximum of the saliency values within in a circular area of diameter 5.6 degree with the fixation point as

the center. The saliency values for the random points are computed in the same manner as that of the fixation points. Next, for a given image the median of the saliency values for the fixated regions and the median of the saliency values for the randomly selected regions are calculated. The ratio of the two medians is used for the evaluation of saliency model. A higher ratio implies that the prediction of fixations by the saliency model is better than the prediction by chance.

### String editing distance

To calculate the string editing distance (Brandt & Stark, 1997; Privitera & Stark, 2000; Borji & Itti, 2013) for a given image, the fixations and the saliency values are clustered using methods such as k-means. After that, regions of interest (*ROIs*) are defined around these clusters which are labeled by alphabetic characters. Next, the *ROIs* are ordered based on the values assigned by the saliency model or the time sequence in which the *ROIs* were fixated on by the observer. The character strings obtained after ordering the *ROIs* for the saliency model and the fixations are then compared by using a string editing similarity index  $S_s$ , which is defined by the cost associated with performing operations such as deletion, insertion, and substitution on the strings. A  $S_s$  value of zero implies that the saliency model perfectly predicts the fixated regions and their temporal sequence. For a detailed description of string editing distance, see the study by (Privitera & Stark, 2000).

## 2.7 Suitable candidate for evaluating the saliency algorithms

While viewing images, observers tend to look at the center regions more as compared to peripheral regions. As a result of that a majority of fixations fall at the image center. This effect is known as center bias and is well documented in vision studies (Tatler, Baddeley, & Gilchrist, 2005; Tatler, 2007). The two main reasons for this are: first, the tendency of photographers to place the objects at the center of the image. Second, the viewing strategy employed by observers, i.e., to look at center locations more in order to acquire the most information about a scene (Tseng, Carmi, Cameron, Munoz, & Itti, 2009). The presence of center bias in fixations makes it difficult to analyze the correspondence between the fixated regions and the salient image regions. This can be explained by the fact in a study by (Judd, Ehinger, Durand, & Torralba, 2009), it was observed that a dummy classifier consisting of a two-dimensional Gaussian shape drawn at the center of the image outperformed all saliency models. The center bias is implicitly linked with a so-called edge effect discussed by (Zhang, Tong, Marks, Shan, & Cottrell, 2008). Edge effect (Borji, Sihite, & Itti, 2013) is defined as adding a varied image border of zeros to a saliency map as a result of which it can yield different values from evaluation metrics. For example, in the study by (Zhang, Tong, Marks, Shan, & Cottrell, 2008), it was observed that a dummy saliency map consisting of all ones with a four-pixel image border consisting of zeros gave an *AUC* value of 0.62. Meanwhile, an *AUC* of 0.73 was obtained with a dummy saliency map using eight-pixel border. In the presence of center bias and edge effect, a fair comparison of the performance of the saliency algorithms

becomes a challenging task. To alleviate the influence of the center bias and the edge effect, a shuffled *AUC* metric was employed in the study by (Zhang, Tong, Marks, Shan, & Cottrell, 2008).

To calculate the shuffled *AUC* metric for a given image and one observer, the regions fixated by the observer are associated with the positive class, however, the regions corresponding to the negative class are defined differently. The regions for the negative class are selected randomly from the fixated regions of the rest of the images, such that they do not coincide with the regions from the positive class. Finally, recent studies by (Borji, Sihite, & Itti, 2013; Zhang, Tong, Marks, Shan, & Cottrell, 2008), have suggested that the shuffled *AUC* metric is quite robust as compared to other evaluation metrics and the most suitable candidate for judging the performance of saliency models.



# Chapter 3

## Research and contributions

### 3.1 Publications

This thesis is based on eleven papers that are published or under review in national and international peer-reviewed conference proceedings and journals.

- Paper 1** Alsam, A., & Sharma, P. (2011). Analysis of eye fixations data. In *Proceedings of the IASTED International Conference, Signal and Image Processing (SIP 2011)*, (pp. 342–349)
- Paper 2** Sharma, P., & Alsam, A. (2014 (accepted)). A robust metric for the evaluation of visual saliency models. In *International Conference on Computer Vision Theory and Applications (VISAPP 2014)*
- Paper 3** Alsam, A., & Sharma, P. (2014). Robust metric for the evaluation of visual saliency algorithms. *Journal of the Optical Society of America A (JOSA A)*, 31(3), 1–9
- Paper 4** Alsam, A., & Sharma, P. (2013). Validating the visual saliency model. In *SCIA 2013, Lecture Notes in Computer Science (LNCS)*, vol. 7944, (pp. 153–161). Springer-Verlag Berlin Heidelberg
- Paper 5** Alsam, A., Sharma, P., & Wrålsen, A. (2013b). Asymmetry as a measure of visual saliency. In *SCIA 2013, Lecture Notes in Computer Science (LNCS)*, vol. 7944, (pp. 591–600). Springer-Verlag Berlin Heidelberg
- Paper 6** Alsam, A., Sharma, P., & Wrålsen, A. (2014). Calculating saliency using the dihedral group d4. *Journal of Imaging Science & Technology*, *accepted*
- Paper 7** Alsam, A., Rivertz, H. J., & Sharma, P. (2012). What the eye did not see – a fusion approach to image coding. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, C. Fowlkes, S. Wang, M.-H. Choi, S. Mantler, J. Schulze, D. Acevedo, K. Mueller, & M. Papka (Eds.) *Advances in Visual Computing*, vol. 7432 of *Lecture Notes in Computer Science*, (pp. 199–208). Springer

- Paper 8** Alsam, A., Rivertz, H. J., & Sharma, P. (2013a). What the eye did not see—a fusion approach to image coding. *International Journal on Artificial Intelligence Tools*, 22(6), 13
- Paper 9** Sharma, P., Nilsen, J. H., Skramstad, T., & Cheikh, F. A. (2010). Evaluation of geometric depth estimation model for virtual environment. In *Norsk informatikkonferanse (NIK-2010)*
- Paper 10** Sharma, P., & Alsam, A. (2012a). Estimating the depth in three-dimensional virtual environment with feedback. In *Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012)*, (pp. 9–17)
- Paper 11** Sharma, P., & Alsam, A. (2012b). Estimating the depth uncertainty in three-dimensional virtual environment. In *Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012)*, (pp. 18–25)

## 3.2 Research issues, answers, and future work

In this section, we discuss the research issues addressed in each paper and the answers found from the analyses associated with the given paper.

### 3.2.1 Proposed robust metric for the evaluation of saliency models

Based on the research work in **Paper 1**, **Paper 2**, and **Paper 3**, we discuss the research issues, and answers.

#### Research issues

1. For an observer viewing a selection of different images obtained from a database, are the fixations random, i.e., there exists no intelligible pattern while viewing different images? On the other hand, if the fixations are not random then the data contains one or more patterns. In that case, we ask are the patterns repeated over different images with different content, or are they indeed image dependent? The visual saliency model suggests that the viewing patterns are image dependent (Itti & Koch, 2001), nevertheless if there is a pattern that is repeated in a mechanical fashion then that would mean that the visual saliency model is not underlying the process of fixations, thus, leading us to ask what mechanisms are responsible for driving the fixations?
2. Given a large number of different images, different observers and a varied number of fixations, how can we perform a meaningful statistical analysis of the data?
3. Given that a certain percentage of the fixations data is common across different images—that is to say, some fixations are not driven by image content, how can we compare the performance of different saliency models such that the effect of content independent fixations is neglected?



## Answers

1. To analyze the fixations data, we superimpose a grid on the image and then create a spatial histogram of locations where the fixations are falling. Using the spatial histogram, we were able to group the fixations from different images and different observers into histograms of the same size, the histograms were represented in the form of vectors. Once the vectors are obtained, the relation between those vectors can be analyzed by using any statistical method.
2. We did not find a clear answer to the question regarding the mechanisms driving the fixations. It was observed that about 23 percent of the data was common across different images. This pattern is repeating spatially to some intensity variation.

When the vectors from the histograms were grouped and we looked at the agreement between different observers on one image, we found a higher agreement than across images with a single observer. The agreement between different observers suggested that part of the viewing mechanism is indeed image dependent. Next, we looked at the images that showed large correspondence between observers that comes from image features. From the results, we observed that the images with clear top-down features such as faces, people, and text ranked higher in correspondence between observers. Images that were more complex, ranked lower in correspondence between viewers. However, some images lay between the two ranks. In addition to this, there were no images where there was a 100 percent agreement between observers. This analysis suggested that there was a stronger agreement on images with so-called top-down features and a weaker agreement on complex images such as landscapes, buildings, and street views.

3. To mitigate the influence of content independent fixations in the performance evaluation of saliency models, we proposed a robust AUC (area under the receiver operating characteristic curve) metric based on the statistical analysis of the fixations data. The proposed metric for a given image is calculated as follows: first, the locations fixated by the observer are associated with the positive class in a manner similar to regular AUC. Next, the locations for the negative class are selected from the fixations associated with high probability in the repeated viewing pattern. In other words, the negative class locations are chosen from the part of the fixations that are most likely image content independent. The results suggest that the proposed metric is a good candidate for ranking the performance of saliency models.

### 3.2.2 Validating the visual saliency model

Based on the research work in **Paper 4** and **Paper 6**, we discuss the research issue, answer, and future work.

### Research issue

1. Assuming that the visual saliency model by (Itti, Koch, & Niebur, 1998) is a good classifier of salient and non-salient regions, can we use linear discrimination methods to separate the parts that are salient from those that are not salient?

### Answer

1. For a given image, we selected parts of the image that received fixations and those that received no fixations. After this, we grouped the two parts into matrices of the same dimensions. On the given matrices A and B, one that pertains to data in image parts that received fixations and the other that encapsulates data from image parts that received no fixations, we used linear discrimination to separate the data of these two matrices. Here the data is the values returned by the visual saliency algorithm by (Itti, Koch, & Niebur, 1998). We found that, we got good discrimination for the parts of the images that were returned by the visual saliency algorithm as highly salient, and the parts that were returned as highly non-salient. However, we found a large overlap in the middle region.

### Future work

1. As a part of our future work, we would like combine the knowledge gained from **Paper 1** with **Paper 4**, and ask if the parts that are in the middle range of saliency are responsible for the seemingly repeated pattern, and if the parts that are highly salient and highly non-salient are responsible for the agreement between different observers?

### 3.2.3 Proposed group based asymmetry algorithm

Based on the research work in **Paper 5** and **Paper 6**, we discuss the research issues, answers, and future work.

#### Research issues

1. Given that the visual saliency model is represented by image features such as gradients, contrasts, and lightness across different scales, can we find a unified metric that groups these expressions in a rigorous description? Given that such a metric exists, what might we deduce from that as to the nature of how vision works? Working under the assumption that the  $D_4$  group transformations are a good representation of saliency, we asked whether we need to use the center-surround operations that constitute the core of the visual saliency model?

#### Answers

1. We found that the transformations pertaining to the dihedral group  $D_4$  are a good unified metric, and they give better results than the visual saliency model by (Itti, Koch, & Niebur, 1998). Hence, the  $D_4$  group transformations were employed to give us an estimation of saliency. Performing the

same validation on the group model as that performed on the saliency model in **Paper 4**, we found a better discrimination between the salient and non-salient regions detected by using the group model as compared to the visual saliency model by (Itti, Koch, & Niebur, 1998). We chose to implement our model without the center-surround operations, however, we represented this group metric in three different scales. The proposed algorithm can be implemented much faster than the visual saliency model.

#### **Future work**

1. As a part of our future work, we would like to look into how to implement the proposed group model faster. To this end, we might employ the representation theory.

### **3.2.4 Proposed saliency based image compression algorithm**

Based on the research work in **Paper 7** and **Paper 8**, we discuss the research issues, answers and future work.

#### **Research issues**

1. Given the knowledge that as the eye browses a scene, it is not fixating everywhere in the image and where the eye fixates is the only part that returns high frequency information, can we then use the information from the fixations data to steer image compression?
2. Given that we get a number of regions that are said to be salient or a number of regions that have received fixations, how can we then use this information to compress an image?

#### **Answers**

1. We propose an algorithm that allows us to compress an image based on the fixations data obtained from an eye tracker or predicted by the visual saliency model.
2. An algorithm that is fast, works in Fourier domain to extract the gradients that have received fixations, disregards the gradients that did not receive any fixations, and then integrates those gradients with the constraint that the resultant is similar to the original. In this way, we maintain the gradients at regions that received fixations, while dampening the gradient information in the regions that did not receive fixations. In so doing, we get a smoothing effect away from the fixated image regions, while maintaining the original sharpness in the regions that received fixations. The resultant image is seamless, does not exhibit visual artifacts and appears to be very similar to the original.

### Future work

1. As a part of future work, we would like to perform a pairwise comparison of the original image and the compressed image, whereby we ask the viewers if they can detect any changes, and to what level we can compress an image before the observer detects changes.

### 3.2.5 Depth estimation in three-dimensional scenes

Based on the research work in **Paper 9**, **Paper 10**, and **Paper 11**, we discuss the research issues, answers and future work.

#### Research issues

1. Can we estimate depth in a three-dimensional virtual scene using an eye tracker?
2. What is the uncertainty surrounding depth estimation? In the presence of noise, slight head movements, and error in the eye tracker, how can we incorporate the uncertainty of the depth estimation?
3. Studies (Duchowski, Shivashankaraiah, Rawls, Gramopadhye, Melloy, & Kanki, 2000; Duchowski, Medlin, Cournia, Murphy, Gramopadhye, Nair, Vorah, & Melloy, 2002; Essig, Pomplin, & Ritter, 2006; Pfeiffer, Latoschik, & Wachsmuth, 2008) have shown that interaction in the virtual environment is difficult as result of the uncertainty recovered in depth estimation. Based on this, we ask can we improve depth estimation, i.e., improve the interaction with the virtual environment if we were to provide a clue to the observer?

#### Answers

1. To answer the question of whether we can accurately estimate depth in a three-dimensional virtual scene using an eye tracker, we replicated an experiment based on the study by (Pfeiffer, Latoschik, & Wachsmuth, 2008). Our results suggest that depth estimation for a three-dimensional virtual scene is possible, given that the virtual scene is designed within the range of the personal space (< 1 meter). However, the resultant depth estimates are not always accurate which can be attributed to noise, slight head movements, and error in the eye tracker.
2. To estimate the the uncertainty in the depth estimation the points recorded by the eye tracker are defined as circles of confusion, instead of points on the plane. Thus from both eyes we get two circles of confusion. Using the two circles as bases and the actual eye locations as vertices, we define two cones. After that, these two cones are extended beyond the plane to a fixed distance (within the range of the personal space). The two extended cones intersect in a region of confusion which gives us a depth estimation with uncertainty measure.
3. By using an audible feedback, we were able to able to clearly improve on the interaction between the user and the object at a certain depth.

## Future work

1. Based on the knowledge gained from **Paper 4**, **Paper 5**, **Paper 6**, **Paper 9**, **Paper 10**, and **Paper 11**, we ask how can we encode visual saliency in three-dimensional scenes such as computer generated scenes or those taken by binocular cameras? We know that using the visual saliency model by (Itti, Koch, & Niebur, 1998), for coding visual saliency in three-dimensional scenes would lead to large computational problems. This is because, the concept of center-surround operations for a three-dimensional scene is not clear. As a part of future work for estimating saliency in three-dimensions, we can employ the symmetry groups for a cube. A cube has 48 different arrangements that can be represented by the transformations of the product of groups  $S_4$  and  $S_2$ . This would provide the link between the work on depth estimation discussed in **Paper 9**, **Paper 10**, and **Paper 11** and the saliency estimation discussed in **Paper 5**, and **Paper 6**. There we envisage using the same operations as employed in the two-dimensional space using the  $D_4$  transformations, but instead perform them in the three-dimensional space using the  $S_4 \times S_2$  transformations. We would be rotating and reflecting a cube in the three-dimensional scene and recording the values and combining them to give us a representation of visual saliency for the three-dimensional scene. In this case, the operations are simple, because we can resize each of the three planes, i.e., X-Y, Y-Z, Z-X, and repeat the  $S_4 \times S_2$  transformations and encode those in a three-dimensional map. This is left as future work, and we hope that the knowledge gained from this research can act as the bridge to go from two-dimensional saliency using  $D_4$  to three-dimensional saliency using  $S_4 \times S_2$ .

## 3.3 Research questions and contributions

Based on the discussion in the previous section, the main objectives of this thesis can be summarized in the form of five main research questions as:

- R1** Is the classic visual saliency algorithm by (Itti, Koch, & Niebur, 1998) a good classifier for salient and non-salient image regions?
- R2** Can salient image regions be calculated in a novel way? How can we calculate saliency for a three-dimensional scene?
- R3** How can we perform a meaningful statistical analysis of the fixations data from different images and observers? Can we use the statistical information obtained from the analysis to create a robust metric for judging the performance of the saliency models?
- R4** How can we use the salient image locations to design an algorithm that compresses an image such that the compressed image is nearly identical to the original?
- R5** How can we estimate depth from the fixations in a three-dimensional virtual scene? How is the depth information useful in the context of three-dimensional visual saliency?

The relations between the research questions and research papers are shown in table 3.1.

Table 3.1: Relations between research papers and research questions.

	R1	R2	R3	R4	R5
Paper 1			•		
Paper 2			•		
Paper 3			•		
Paper 4	•				
Paper 5		•			
Paper 6	•	•			
Paper 7				•	
Paper 8				•	
Paper 9					•
Paper 10					•
Paper 11					•

The major contributions of this research effort are:

- C1** A novel method to inspect the performance of the classic visual saliency algorithm by (Itti, Koch, & Niebur, 1998) in separating the salient and non-salient image regions.
- C2** A visual saliency model that calculates salient image regions in a novel way, i.e., by using the transformations pertaining to the dihedral group  $D_4$ . The proposed model performs better than the saliency model by (Itti, Koch, & Niebur, 1998), and it is among one of the four best models in the literature. In addition, the proposed model can be extended to calculate saliency in three-dimensional virtual scenes.
- C3** A new method for the statistical analysis of the eye fixations data from different images and different observers. Based on the analysis, a new robust metric is proposed that can be used for the evaluation of the visual saliency algorithms.
- C4** A novel algorithm that compresses an image based on the salient locations predicted by the visual saliency algorithm. The compressed images do not exhibit visual artifacts and they appear to be very similar to the originals.
- C5** A new method for estimating the depth in a three-dimensional virtual scene by using the fixations from both eyes. As a part of future work, we intend to use the depth information obtained by showing the observer a virtual scene to create a three-dimensional fixations map, which can be used as the ground truth for the evaluation of three-dimensional saliency algorithms.

The research contributions **C1** to **C5** and the research questions **R1** to **R5** have one-to-one correspondence.

# Chapter 4

## Research methodology

For the research work in this thesis, we employed the methodology of design science research. According to (Iivari, 2007), design science research has been practiced in the disciplines such as Computer Science, Software Engineering and Information Technology for decades without explicitly naming it. Studies by (Iivari, 2007; Hevner & Chatterjee, 2010) suggest that by using the methodology of design science research computer scientists have developed new architectures for computers, new programming languages, new compilers, new algorithms, new data and file structures, new data models, new database management systems, and more.

Design science research as discussed by (Hevner, March, Park, & Ram, 2004), consists of creating novel artifacts, i.e., something new that does not exist in nature, and using it to understand a natural or man-made phenomenon (Vaishnavi & Kuechler, 2004). In this way, it is quite useful for vision studies, where new algorithms or statistical methods are frequently used to analyze different aspects of human vision. In the general methodology of design science research as shown in figure 4.1, the process begins with the *Awareness of Problem* and terminates with *Conclusion*. We discuss the various steps of the design science research methodology and how they were used for this research.

### Awareness of problem

The first step in this process is the awareness of an interesting problem in the given field. This can come from developing an understanding of the relevant field by using sources such as scientific literature or new industrial developments.

The output of this stage is a proposal for a new research project, and in our case this PhD project.

### Suggestion

In this step, to analyze the problem and provide possible solutions, either new methods are created or methods are employed from existing literature in a new way. Based on the employed methods, a tentative design is suggested. In this thesis, the formulation of the research questions associated with all the papers and the proposed methods suggested to investigate them, constituted this step.

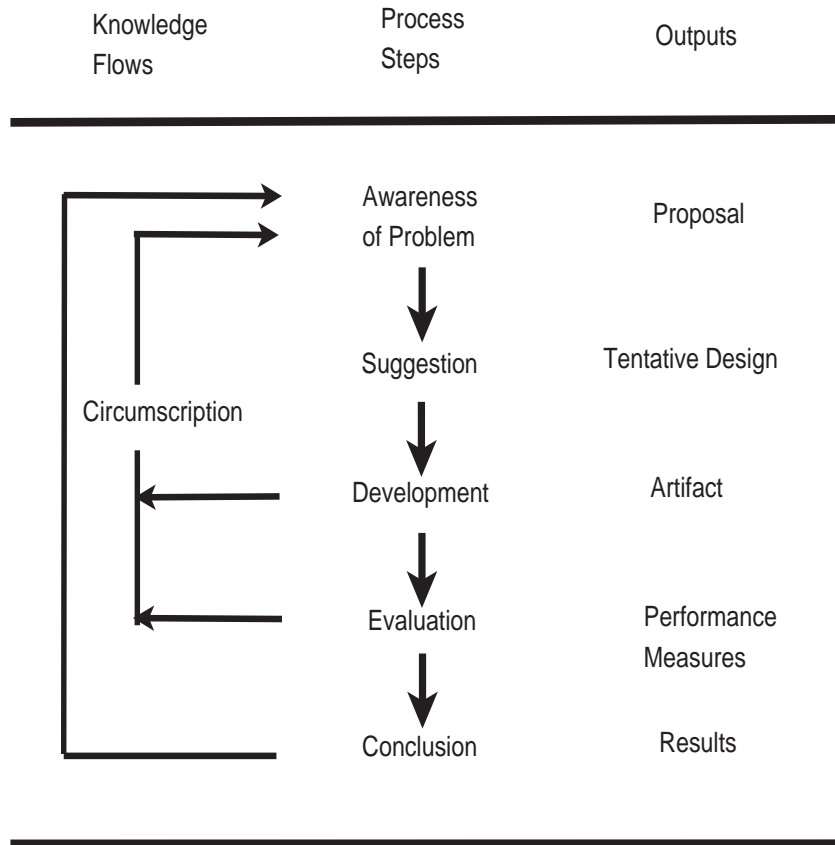


Figure 4.1: The general research methodology of design science research, from (Vaishnavi & Kuechler, 2004).

### Development

During the development step, the tentative design is evolved to completion. This is achieved by using techniques relevant to the construction of an artifact. In our case, the technique used was algorithm development and a number of algorithms were developed using Matlab/C++ to answer the research questions.

### Evaluation

The artifact created from the previous step is expected to behave in a certain way. In this step, the deviation from the expected behavior is measured using quantitative or qualitative methods and the results are analyzed to confirm or contradict the hypothesis. In case the initial hypothesis is too broad, the knowledge gained here is fed back to the first step as depicted by circumscription arrow, such that the hypothesis is modified based on an improved understanding of the problem. In this thesis, the algorithms developed were evaluated using well known methods from the literature, for example, linear



discrimination analysis, singular value decomposition, and receiver operating characteristic curve. As shown in table 4.1, **Paper 1** to **Paper 8** were evaluated using a publicly available dataset by (Judd et al., 2009). **Paper 9** to **Paper 11** were evaluated by recording the data from the eye tracking experiments performed at Sør-Trøndelag University College (HiST). Five observers took part in the experiments.

Research papers	Eye tracking data
<b>Paper 1</b> <b>Paper 2</b> <b>Paper 3</b> <b>Paper 4</b> <b>Paper 5</b> <b>Paper 6</b> <b>Paper 7</b> <b>Paper 8</b>	publicly available dataset by (Judd et al., 2009)
<b>Paper 9</b> <b>Paper 10</b> <b>Paper 11</b>	eye tracking experiments performed at HiST

Table 4.1: Research papers and eye tracking datasets.

## Conclusion

This is the final step of the research effort. Even though the results obtained might still stray from the expected behavior, but they are considered good enough for improving the understanding of the problem. The knowledge gained here is expected to contribute towards future research projects. In our case, this is highlighted by the contributions from the research papers and this thesis.



# Chapter 5

## Summaries of research papers

### 5.1 Paper 1: Analysis of eye fixations data

#### 5.1.1 Synopsis

In this paper, we analyzed eye fixations data obtained from 15 observers and 1003 images. When studying the correlation matrix constructed based on the fixations data of one observer viewing all images, it was observed that 23 percent of the data can be accounted for by one eigenvector. This finding implies a repeated viewing pattern that is independent of image content. The examination of this pattern revealed that it was highly correlated with the center region of the image. Next, we analyzed the correlation matrix based on the fixations data across different observers viewing the same image. We found a higher agreement across different observers than across different images with a single observer. The agreement between different observers suggested that part of the viewing mechanism is indeed image dependent. We looked at the images that showed large correspondence between observers that comes from image features. From the results, we observed that the images with clear top-down features such as faces, people, and text ranked higher in correspondence between observers. Images that were more complex, ranked lower in correspondence between viewers. This analysis suggested that there was a stronger agreement on images with so-called top-down features and a weaker agreement on complex images such as landscapes, buildings, and street views. The main contribution of this paper can be outlined as a new method to perform the statistical analysis of the fixations data. This is strongly linked to the first part of the research contribution **C3**.

### 5.2 Paper 2: A robust metric for the evaluation of visual saliency models

#### 5.2.1 Synopsis

Finding a robust metric for evaluating the visual saliency algorithms has been the subject of research for decades. Motivated by the shuffled AUC metric in this

paper, we propose a robust AUC metric that uses the statistical analysis of the fixations data to better judge the goodness of the different saliency algorithms. To calculate the robust AUC metric, we use the first eigenvector obtained from the statistical analysis to define the area from which non-fixations are selected thus mitigating the effect of the repeated viewing pattern also known as the center bias. Our results show that the proposed metric results in similar performance when compared with the shuffled AUC metric, but given that the proposed metric is derived from the statistics for the data set, we believe that it is more robust. The main contribution of this paper is a new robust metric that can be used for evaluating the performance of the saliency algorithms. This is most relevant to the second part of the research contribution **C3**.

### **5.3 Paper 3: A robust metric for the evaluation of visual saliency algorithms (extended)**

#### **5.3.1 Synopsis**

This paper combines the research work done in **Paper 1** and **Paper 2** and is strongly linked with both the first and second parts of the research contribution **C3**.

### **5.4 Paper 4: Validating the visual saliency model**

#### **5.4.1 Synopsis**

Bottom up attention models suggest that human eye movements can be predicted by means of algorithms that calculate the difference between a region and its surround at different image scales where it is suggested that the more different a region is from its surround the more salient it is and hence the more it will attract fixations. Recent studies have however demonstrated that a dummy classifier which assigns more weight to the center region of the image out performs the best saliency algorithm calling into doubt the validity of the saliency algorithms and their associated bottom up attention models. In this paper, we performed an experiment using linear discrimination analysis to try to separate between the values obtained from the saliency algorithm for regions that have been fixated and others that haven't. Our working hypothesis was that being able to separate the regions would constitute a proof as to the validity of the saliency model. Our results show that the saliency model performs well in predicting non-salient regions and highly salient regions but that it performs no better than a random classifier in the middle range of saliency. The main contribution of this paper is the validation that the classic saliency model is good at discriminating between the parts of the images that are returned as highly salient, and the parts that are returned as highly non-salient. This is strongly linked to the research contribution **C1**.

## 5.5 Paper 5: Asymmetry as a measure of visual saliency

### 5.5.1 Synopsis

A salient feature is a part of the scene that stands out relative to neighboring items. By that we mean that a human observer would experience a salient feature as being more prominent. It is, however, important to quantify saliency in terms of a mathematical quantity that lends itself to measurements. Different metrics have been shown to correlate with human fixations data. These include contrast, brightness and orienting gradients calculated at different image scales. In this paper, we show that these metrics can be grouped under operations pertaining to the dihedral group  $D_4$ , which is the symmetry group of the square image grid. Our results show that salient features can be defined as the image features that are most asymmetric in their surrounds. The main contribution of this paper is that the transformations pertaining to dihedral group  $D_4$  are a good representation of saliency. This is most relevant to the research contribution **C2**.

## 5.6 Paper 6: Calculating saliency using the dihedral group $D_4$

### 5.6.1 Synopsis

This paper combines the research work in **Paper 4** and **Paper 5** and is explicitly linked with the research contributions **C1** and **C2**.

## 5.7 Paper 7: What the eye did not see—a fusion approach to image coding

### 5.7.1 Synopsis

The concentration of the cones and ganglion cells is much higher in the fovea than the rest of the retina. This non-uniform sampling results in a retinal image that is sharp at the fixation point, where a person is looking, and blurred away from it. This difference between the sampling rates at the different spatial locations presents us with the question of whether we can employ this biological characteristic to achieve better image compression. This can be achieved by compressing an image less at the fixation point and more away from it. It is, however, known that the vision system employs more than one fixation to look at a single scene which presents us with the problem of combining images pertaining to the same scene but exhibiting different spatial contrasts. This article presents an algorithm to combine such a series of images by using image fusion in the gradient domain. The advantage of the algorithm is that unlike other algorithms that compress the image in the spatial domain our algorithm results in no visual artifacts. The algorithm is based on two steps, in the first we modify the gradients of an image based on a limited number of fixations and in the second we integrate the modified gradient. Results based on measured and predicted fixations verify our approach. The main contribution of this paper is

a new method to compress an image based on the salient locations predicted by the saliency algorithms or the fixation data obtained from an eye tracker. This is most relevant to the research contribution **C4**.

## **5.8 Paper 8: What the eye did not see—a fusion approach to image coding (extended)**

### **5.8.1 Synopsis**

This paper is an extended version of **Paper 7** with more results and is strongly linked to the research contribution **C4**.

## **5.9 Paper 9: Evaluation of geometric depth estimation model for virtual environment**

### **5.9.1 Synopsis**

Three-dimensional virtual environment is a computer generated experience which gives us a feeling of presence in the environment. Objects displayed in virtual environment unlike the real world have no physical depth. Due to the distance between the eyes, the images formed on the retina are different, this facilitates our perception of depth. In the range of personal space, eyes converge at different angles to look at objects in different depth planes, known as convergence angle. Since we cannot get images of the scene viewed by the two eyes, the convergence angle cannot be calculated by standard photogrammetry principles such as triangulation. However, we can measure the point of focus (fixations) of the eyes on two-dimensional display plane, by using eye tracker. Each eye gets a different view of the virtual scene. Knowing the physical location of both eyes and their corresponding fixations, we can calculate the estimated depth using geometry. In this paper, we replicate an experiment based on the study by (Pfeiffer, Latoschik, & Wachsmuth, 2008). Our results suggest that depth estimation for a three-dimensional virtual scene is possible given that the virtual scene is designed within the range of the personal space. The main contribution of this paper is recreating the experiment setup necessary to estimate depth in a virtual environment. This is most relevant to the research contribution **C5**.

## **5.10 Paper 10: Estimating the depth in three-dimensional virtual environment with feedback**

### **5.10.1 Synopsis**

Visual interaction in three-dimensional virtual space can be achieved by estimating objects depth from the fixations of the left and right eyes. Training a PSOM neural network to estimate depth, from eye fixations, has been shown to result in good level of accuracy. Instead of training a neural network we

postulate that it is possible to improve the accuracy of the fixation data by providing the observer with feedback. In order to test this hypothesis we introduce a closed-loop feedback in the environment. When the user’s visual axes intersect, within a range of the correct depth, a sound is produced. This mechanism trains the users to correct their fixations in a fashion that results in improved depth estimation. Our results show that indeed the accuracy of depth estimation improves in the presence of feedback. The main contribution of this paper is a method to improve the depth estimation in a virtual environment. This is implicitly linked to the research contribution **C5**.

## **5.11 Paper 11: Estimating the depth uncertainty in three-dimensional virtual environment.**

### **5.11.1 Synopsis**

Visual interaction in three-dimensional virtual space can be achieved by estimating objects depth from the fixations of the left and right eyes. The current depth estimation methods, however do not account for the presence of noise in the data. To address this problem we note that any measured fixation point is a member of a statistical distribution defined by the level of noise in the measurement. We thus propose a new numerical method that provides a range of depth values based on the uncertainty in the measured data. The main contribution of this paper is a new method to estimate the depth uncertainty in a virtual environment. This is explicitly linked to the research contribution **C5**.





# Chapter 6

## Discussion

This chapter concludes the dissertation with an overview of the results obtained from the research papers, and the main research direction for future work.

### 6.1 Validating the visual saliency model

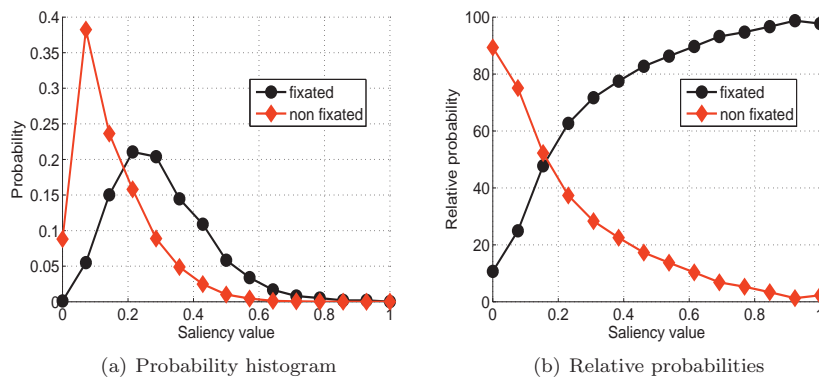


Figure 6.1: Probability histograms and relative probabilities for the fixated and non-fixated regions for an average observer. X-axis shows the saliency values obtained by using the visual saliency algorithm (Itti, Koch, & Niebur, 1998).

In papers 4 and 6, we performed an experiment using linear discrimination analysis to try to separate between the saliency values obtained from the model by (Itti, Koch, & Niebur, 1998) for locations that received fixations and those that received no fixations. The data was based on a subset of the images and corresponding fixations obtained by (Judd, Ehinger, Durand, & Torralba, 2009), where we used 200 landscape images and all the fifteen observers. In the experiment, we defined a fixated area as a square region of dimensions 100 by 100 pixels where the center was located at the fixation point. Non-fixated areas were chosen randomly from parts of the image that had a region of a 100

by 100 pixels without any fixations. By collecting the values returned by the saliency algorithm local to those regions into two matrices we were able to use discrimination analysis to determine whether the data of the two matrices is separable. In figure 6.1(a), we show the probability histograms of the fixated and non-fixated regions for all the observers. Here, the histogram is normalized such that the area under the curve is one. We note that the separation between the two sets is not ideal but rather we find a considerable overlap between the two histograms specifically in the middle range. We further note that there is a clear separation between the two sets for regions of the images that received no fixations indicating that the method is good at predicting non-salient regions of the images. At a value of 0.3 the classification of the two sets is random. To gain better insight into the ability of the algorithm to separate the image regions into fixated and non-fixated, we plotted the relative probabilities of the histograms. For the non-fixated histogram, the relative probabilities were obtained by dividing the area under the non-fixated probability histogram curve of a specific bin  $i$  of the histogram by the area under the fixated histogram curve for the same bin. For the relative probability of the fixated histogram the reciprocal value was calculated. This curve is plotted in figure 6.1(b) where we observe that for low salience values the separation of non-fixated regions is ideal and that the extent of the separation declines to a level that is random. We also note that the separation of the highly salient regions, is nearly ideal. Based on this we can conclude that the saliency algorithm by (Itti, Koch, & Niebur, 1998) is good in predicting non-salient and highly salient regions but its performance drops in the middle range.

## 6.2 Proposed group based asymmetry algorithm

In papers 5 and 6, we set about unifying the mathematical description of saliency in a single metric. Based on the knowledge gained from research in image processing where it has been shown that the dihedral group  $D_4$  can be used to encode edges and contrast which are the main current descriptions of saliency, we chose to devise an algorithm that represents the level of saliency in an image region by virtue of the transformations of  $D_4$ . In our experiment, we used a receiver operating characteristic (ROC) curve to compare the performance of the proposed method with that of (Itti, Koch, & Niebur, 1998). For the analysis, we used fixations data from 200 images and fifteen observers. We found that the proposed group based asymmetry (GBA) algorithm results in an AUC value of 0.81 which is better than that achieved with the visual saliency algorithm by (Itti, Koch, & Niebur, 1998) which gives AUC of 0.77. Based on the results, we conclude that the transformations pertaining to the dihedral group  $D_4$  are a good metric to estimate salient image regions. In figure 6.2, we offer a visual comparison between the two algorithms, we show the fixations map, and the saliency maps obtained from the proposed GBA algorithm and the visual saliency algorithm by (Itti, Koch, & Niebur, 1998) for an example image. We can see that the maps from both the algorithms are quite similar. In fact both of them return the region containing the boat at the center as salient, which is also in agreement with the fixations map. The performance of the proposed GBA algorithm is compared with other state-of-the-art saliency models in the next section.

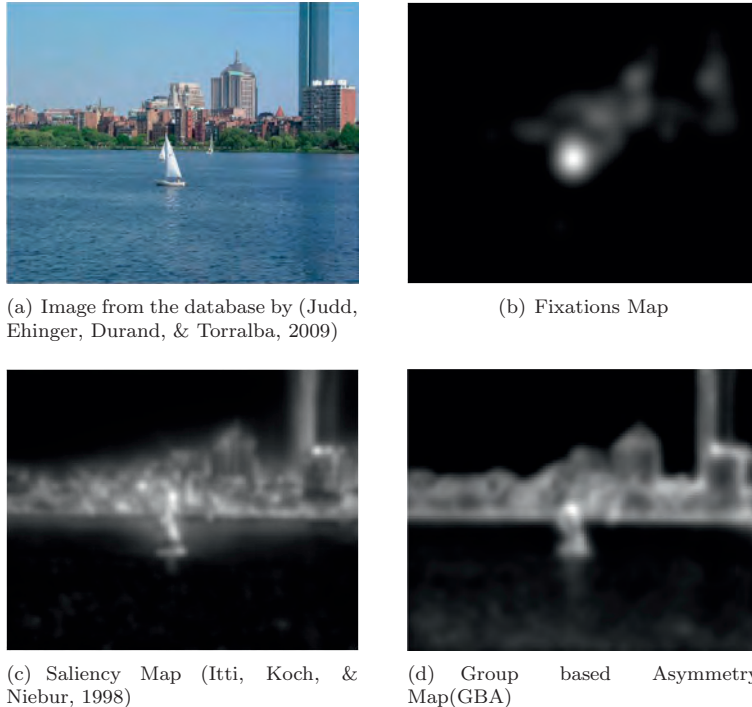
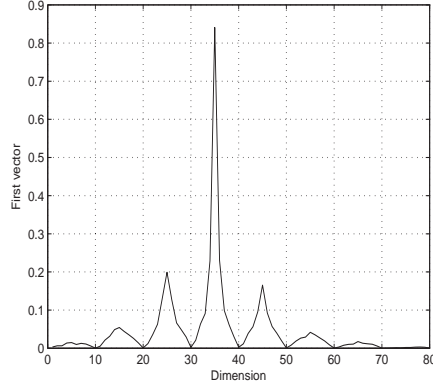


Figure 6.2: Comparison of visual saliency algorithms, both algorithms return the region containing the boat at the center as salient, which is also in agreement with the fixations map obtained from the eye fixations data.

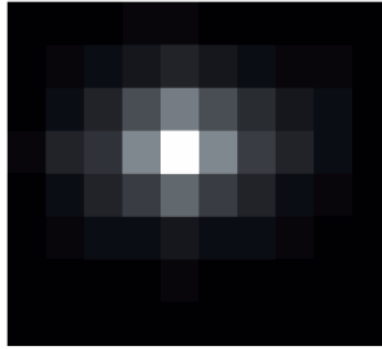
### 6.3 Proposed robust metric for the evaluation of saliency models

In papers 1, 2, and 3, we analyzed the fixations data from 15 observers and 1003 images collected as a part of the study by (Judd, Ehinger, Durand, & Torralba, 2009). The database consisted of portrait and landscape images. For our analysis we chose 463 landscape images of size 768 by 1024 pixels. When studying the eigen-decomposition of the correlation matrix constructed based on the fixations data of one observer viewing all images, it was observed that 23 percent of the data can be accounted for by one eigenvector. This finding implies a repeated viewing pattern that is independent of image content. Figure 6.3 shows the repeated viewing pattern, i.e., the first eigenvector for all observers and images. We note that it depicts a concentration of fixations in the center region of the image. This center bias in the fixations has been observed in other studies (Meur, Callet, Barba, & Thoreau, 2006; Tatler, 2007; Judd, Ehinger, Durand, & Torralba, 2009) and it is likely responsible for the high correlation of fixations data with a dummy Gaussian classifier as noted in the study by Judd et al. (Judd, Ehinger, Durand, & Torralba, 2009).

Guided by recent studies on the creation of a metric that normalizes for the influence on the center region, we studied the work by (Zhang, Tong, Marks,



(a) Eigenvector for an average observer.



(b) Probability histogram for the shared eigenvector.

Figure 6.3: Eigenvector for an average observer. It shows a concentration of fixations in the center region of the image.

Shan, & Cottrell, 2008), in which a shuffled AUC (area under the receiver-operating-characteristic curve) metric was used by the authors to abate the effect of center-bias in fixations. Instead of selecting non-fixated regions from single image as is the case in the shuffled metric by (Zhang, Tong, Marks, Shan, & Cottrell, 2008), we decided to use the repeated viewing pattern obtained from the statistical analysis of the fixations data. We reasoned that for a given image the repeated pattern is the part which is most likely to be fixated upon, thus choosing a non-fixated region from within it for the analysis by the AUC metric would indeed counteract the influence of the repeated fixations pattern. The results obtained by employing the shuffled AUC metric are shown in figure 6.4. We note that, **AIM** by (Bruce & Tsotsos, 2005), **Hou** by (Hou & Zhang, 2007), our proposed group based asymmetry (**GBA**) model, and **AWS** by (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012) are the four best models. In-line with

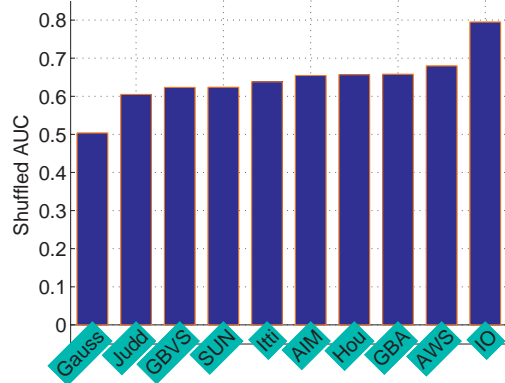


Figure 6.4: Ranking of visual saliency models using the shuffled AUC metric.

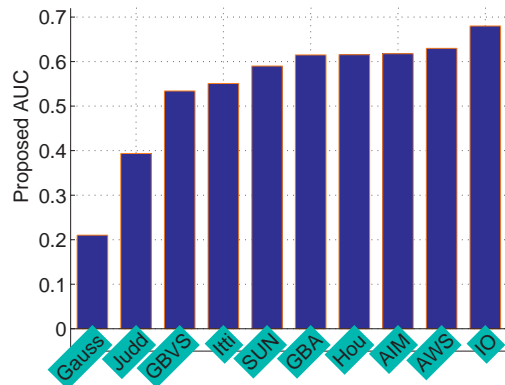


Figure 6.5: Ranking of visual saliency models using the robust AUC metric.

the study by (Borji, Sihite, & Itti, 2013), our results show that the **AWS** model is the best among all. Figure 6.5 shows the ranking of saliency models obtained by using the proposed robust AUC metric. We observe that the ranking is almost the same as the shuffled AUC metric, with the **AWS** model performing the best and the **Gauss** model performing the worst. We note that the robust AUC metric gives a lower value for the **Gauss** model, and the saliency models are closer to the inter-observer (**IO**) model. Based on the results, we conclude that the robust AUC metric a good candidate for the evaluation of saliency algorithms.

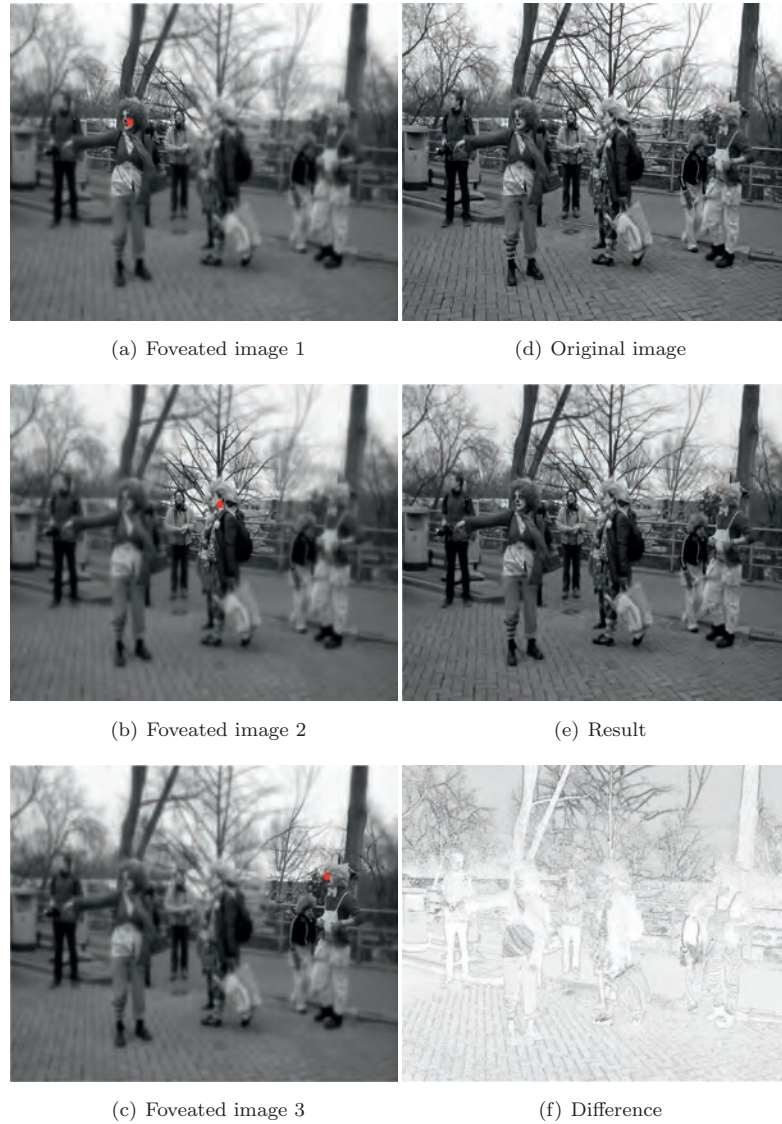


Figure 6.6: In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image.

## 6.4 Proposed saliency based image compression algorithm

In papers 7 and 8, we proposed an algorithm to compress an image based on the eye fixations from an eye tracker or the salient image locations predicted by the

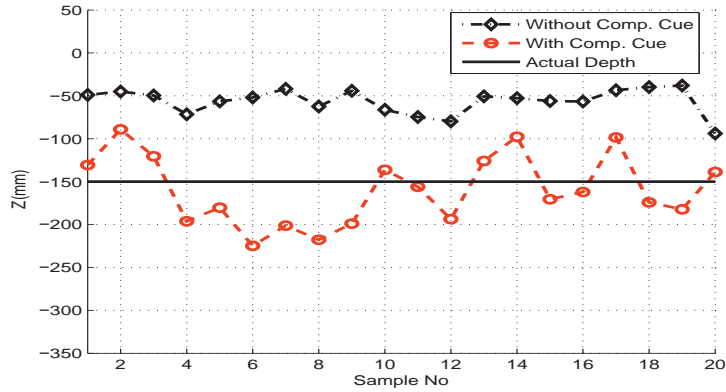
saliency models. This is achieved by using image fusion in the gradient domain. The algorithm is based on two steps, in the first we modify the gradients of an image based on a limited number of fixations and in the second we integrate the modified gradient. The use of human vision steered compression is seen by researchers as the most promising path toward further improvements. In this regard, the proposed algorithm can be used as part of an image compression pipeline with very promising results. From our initial tests, we have noticed that the algorithm results in reduced storage requirements without the added artifacts associated with frequency based compressions in the wavelets domain. The results for an example images and the associated fixations are shown in figure 6.6. In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. In agreement with the predicted results for the application of the contrast function by (Wang & Bovik, 2001), we notice that the regions around the fixation points are sharper than the rest. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient.

## 6.5 Depth estimation in three-dimensional scenes

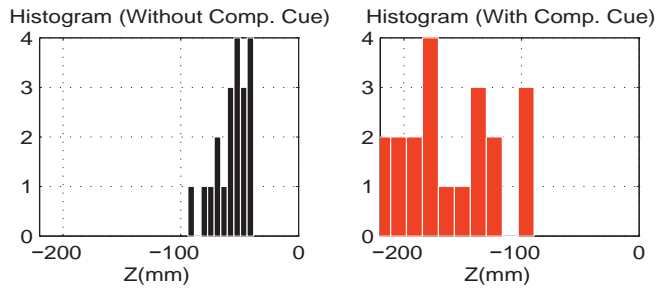
In papers 9, 10, and 11, we presented two main contributions.

The first is the hypothesis that the introduction of a closed loop feedback in the form of a compensatory cue improves the estimation of perceived depth in virtual environments. To test our hypothesis we designed a simple three-dimensional virtual environment which included a checkerboard background and spherical objects appearing at different depth values. The depth range used in the experiment varied from 50 to 300 mm behind the screen. This range corresponds to the users personal space which is believed to be the range in which convergence is a significant cue. Furthermore, we included an audible cue into the design of the environment. The audible cue was provoked when the fixation-data obtained from the eye tracker resulted in a depth estimate that was within a predefined error value. Here the calculations were based on a line-intersection method. To examine the local variations in the data we sub-sampled the distribution into twenty regions. For each sub-sample we calculated the average values of the depth obtained by employing the line-intersection method. Figure 6.7(a) shows the variation over time of the local average values for a depth of 150 mm. We note that the introduction of the compensatory cue is indeed improving the estimated depth over time. Further, the comparison of the histograms, figure 6.7b, for the two experiments reflects that the introduction of the compensatory cue results in a higher frequency of depth estimates that are in the vicinity of the actual depth.

The second contribution is the introduction of a new method that allows designers of virtual environments to estimate the uncertainty in the measured depth value. The proposed method is based on the principle of intersection of convex sets where two sets are defined. The first set is defined by the statistical distribution of the left eye fixations together with the center of the eye. A



(a) Distributions of depth estimates for the sub-sampled data of two experiments over twenty samples of the total time. In the experiment with compensatory cue we see a clear convergence towards the actual depth of the object, that is 150 mm behind the screen.

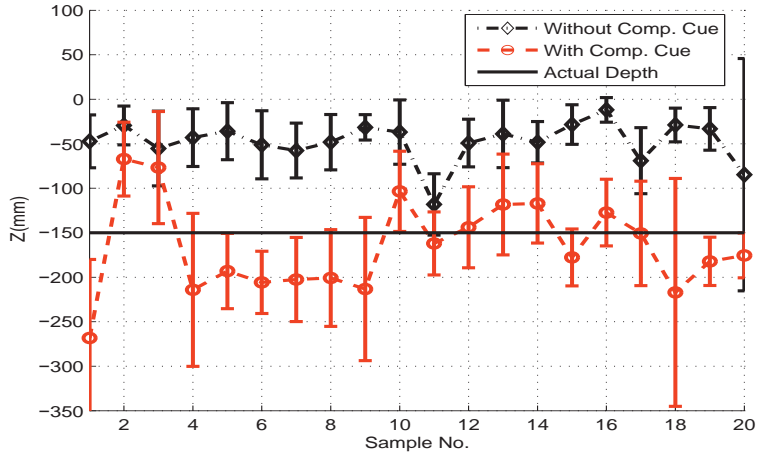


(b) Histograms of the sub-sampled data for two experiments.

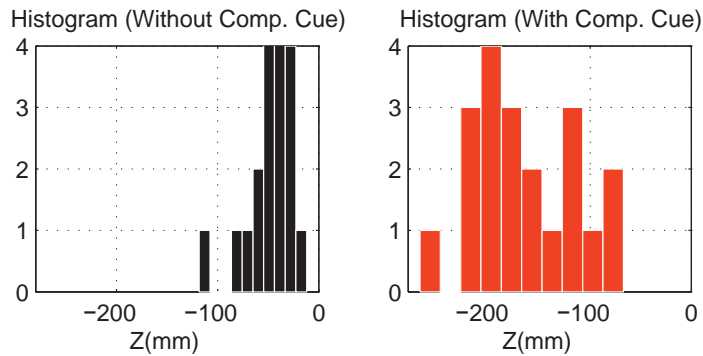
Figure 6.7: Distributions and histograms of depth estimates for two experiments: without compensatory cue, and with compensatory cue. Depth estimates were calculated using the line-intersection method.

corresponding set is defined for the right eye. In an ideal situation i.e., when no noise is present in the data these two sets are reduced to the visual lines and the method is identical to the line-intersection method. When noise is present, however, the sets represent conical volumes and their intersection is the feasible solution space where any point is equally likely to be the actual depth. Based on that we represented the uncertainty in the estimate by means of three standard deviations from the average value. Figure 6.8 shows the results obtained based on a depth value of 150 mm behind the screen. We note that the result obtained with the compensatory cue represents a clear improvement over that achieved without. We also note that while the average values of the cone intersection region are a fair representation of the actual depth, the uncertainty depicted by the error-bars offers a more comprehensive view into the estimation. We observe that the real depth is almost always within the uncertainty range.





(a) Distributions of depth estimates for the sub-sampled data of two experiments over twenty samples of the total time. In the experiment with compensatory cue we see a clear convergence towards the actual depth of the object, that is 150 mm behind the screen. Furthermore we notice that the actual depth is almost always within the uncertainty range.



(b) Histograms of the sub-sampled data for two experiments.

Figure 6.8: Distributions and histograms of depth estimates for two experiments: without compensatory cue, and with compensatory cue. Depth estimates were calculated using the cone-intersection method.

## 6.6 Towards three-dimensional visual saliency

In order to calculate saliency in three-dimensional virtual scenes, we can use the symmetry groups for a cube. A cube has 48 symmetries that can be represented by the transformations of products of the groups  $S_4$  and  $S_2$ .  $S_2$  is the symmetric group of degree 2 and has two elements: the identity, and the permutation interchanging the two points (Dummit & Foote, 2004).  $S_4$  is a symmetric group of degree 4, i.e., all permutations on a set of size four (Dummit & Foote, 2004). This group has 24 elements that are obtained by rotations about opposite faces,

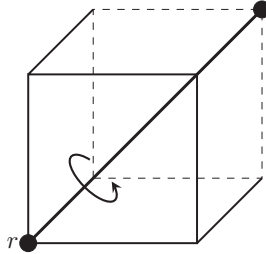


Figure 6.9: Number of axes with opposite diagonals like this = 4. We can rotate by 120 or 240 degrees around these axes. These operations give 8 elements.

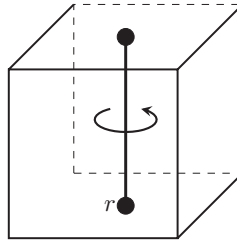


Figure 6.10: Number of axes with opposite faces like this = 3. We can either rotate by 90, 180 or 270 degrees around these axes. These operations give 9 elements.

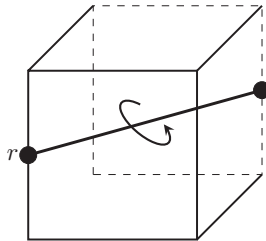


Figure 6.11: Number of axes with opposite edges like this = 6. We can rotate by 180 degrees around these axes. These operations give 6 elements.

opposite diagonals and opposite edges of the cube. For instance, figures 6.9 to 6.11 show the different rotational symmetries of the cube. We note that from the rotations along opposite diagonals, faces, and edges we get 8, 9, and 6 elements respectively. These elements along with the identity form the 24 elements of the  $S_4$  group.

Saliency in a three-dimensional virtual scene can be calculated by employing the same procedure as discussed in papers 5 and 6, but instead of computing in two-dimensional space using the  $D_4$  group, we can calculate in three-dimensional space using the  $S_4 \times S_2$  transformations. For example, after dividing the three-dimensional scene into uniform size cubes, we can rotate and reflect a cube and record the values associated with the transformations. The recorded values can

be collected in a matrix and rescaled along each of the three planes, i.e., X-Y, Y-Z, Z-X to get a three-dimensional feature map. The resulting feature maps corresponding to the 48 elements can be combined to get a representation of saliency for the three-dimensional scene. Similar to the implementation discussed in papers 5 and 6, different cube sizes can be used to capture both the local and global salient details. This is left as future work and we hope that this will help future researchers to venture towards three-dimensional saliency.



# Bibliography

- Achanta, R., Estrada, F., Wils, P., & Süsstrunk, S. (2008). Salient region detection and segmentation. In *Proceedings of the 6th international conference on Computer vision systems*, (pp. 66–75). Springer-Verlag.
- Ajallooieian, M., Borji, A., Araabi, B., Ahmadabadi, M., & Moradi, H. (2009). An application to interactive robotic marionette playing based on saliency maps. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, (pp. 841–847).
- Alsam, A., Rivertz, H. J., & Sharma, P. (2012). What the eye did not see – a fusion approach to image coding. In G. Bebis, R. Boyle, B. Parvin, D. Koricin, C. Fowlkes, S. Wang, M.-H. Choi, S. Mantler, J. Schulze, D. Acevedo, K. Mueller, & M. Papka (Eds.) *Advances in Visual Computing*, vol. 7432 of *Lecture Notes in Computer Science*, (pp. 199–208). Springer.
- Alsam, A., Rivertz, H. J., & Sharma, P. (2013a). What the eye did not see—a fusion approach to image coding. *International Journal on Artificial Intelligence Tools*, 22(6), 13.
- Alsam, A., & Sharma, P. (2011). Analysis of eye fixations data. In *Proceedings of the IASTED International Conference, Signal and Image Processing (SIP 2011)*, (pp. 342–349).
- Alsam, A., & Sharma, P. (2013). Validating the visual saliency model. In *SCIA 2013, Lecture Notes in Computer Science (LNCS)*, vol. 7944, (pp. 153–161). Springer-Verlag Berlin Heidelberg.
- Alsam, A., & Sharma, P. (2014). Robust metric for the evaluation of visual saliency algorithms. *Journal of the Optical Society of America A (JOSA A)*, 31(3), 1–9.
- Alsam, A., Sharma, P., & Wrålsen, A. (2013b). Asymmetry as a measure of visual saliency. In *SCIA 2013, Lecture Notes in Computer Science (LNCS)*, vol. 7944, (pp. 591–600). Springer-Verlag Berlin Heidelberg.
- Alsam, A., Sharma, P., & Wrålsen, A. (2014). Calculating saliency using the dihedral group d4. *Journal of Imaging Science & Technology*, accepted.

- Anderson, C. H., Essen, D. C. V., & Olshausen, B. A. (2005). *Neurobiology of Attention*, chap. Directed Visual Attention and the Dynamic Control of Information Flow, (pp. 11–17). Elsevier.
- Avraham, T., & Lindenbaum, M. (2010). Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(4), 693–708.
- Barlow, H. (1987). Cerebral cortex as model builder. In L. Vaina (Ed.) *Matters of Intelligence*, vol. 188 of *Synthese Library*, (pp. 395–406). Springer Netherlands.
- Bian, P., & Zhang, L. (2009). Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In M. Kppen, N. Kasabov, & G. Coghill (Eds.) *Advances in Neuro-Information Processing*, vol. 5506 of *Lecture Notes in Computer Science*, (pp. 251–258). Springer Berlin Heidelberg.
- 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.
- Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69.
- Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, *9*, 27–38.
- Braun, J., & Sagi, D. (1990). Vision outside the focus of attention. *Perception and Psychophysics*, *48*(1), 45–58.
- Breazeal, C., & Scassellati, B. (1999). A context-dependent attention system for a social robot. In *IJCAI*, (pp. 1146–1153).
- Bruce, N. D. B., & Tsotsos, J. K. (2005). Saliency based on information maximization. In *NIPS'05*, (pp. 155–162).
- Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*, 14841525.
- Cerf, M., Harel, J., Einhauser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems (NIPS)*, vol. 20, (pp. 241–248).
- Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A bayesian inference theory of attention. *Vision Research*, *50*(22), 2233 – 2247.
- Chun, M. M., & Wolfe, J. M. (2001). Visual attention. *Blackwell Handbook of Sensation and Perception*, (pp. 272–310).
- Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Reviews in the Neurosciences*, *18*, 193–222.
- Duchowski, A., Medlin, E., Cournia, N., Murphy, H., Gramopadhye, A., Nair, S., Vorah, J., & Melloy, B. (2002). 3d eye movement analysis. *Behavior Research Methods, Instruments, and Computers (BRMIC)*, *34*(4), 573–591.

- Duchowski, A. T., Shivashankaraiah, V., Rawls, T., Gramopadhye, A. K., Melloy, B. J., & Kanki, B. (2000). Binocular eye tracking in virtual reality for inspection training. In *Eye Tracking Research & Applications Symposium*, (pp. 89–96). ACM.
- Dummit, D. S., & Foote, R. M. (2004). *Abstract Algebra*. John Wiley & Sons.
- El-Nasr, M., Vasilakos, A., Rao, C., & Zupko, J. (2009). Dynamic intelligent lighting for directing visual attention in interactive 3-d scenes. *Computational Intelligence and AI in Games, IEEE Transactions on*, 1, 145–153.
- Eriksen, C., & St. James, J. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, 40(4), 225–240.
- Essig, K., Pomplin, M., & Ritter, H. (2006). A neural network for 3d gaze recording with binocular eye trackers. *The International Journal of Parallel, Emergent and Distributed Systems*, 21(2), 79–95.
- Fawcett, T. (2004). Roc graphs with instance-varying costs. *Pattern Recognition Letters*, 27(8), 882–891.
- Feng, X., Liu, T., Yang, D., & Wang, Y. (2008). Saliency based objective quality assessment of decoded video affected by packet losses. In *15th IEEE International Conference on Image Processing (ICIP 2008)*, (pp. 2560–2563).
- Frintrop, S. (2006a). *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, vol. 3899 of *Lecture Notes in Artificial Intelligence (LNAI)*. Springer.
- Frintrop, S. (2006b). *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*. Ph.D. thesis, University of Bonn.
- Frintrop, S., Jensfelt, P., & Christensen, H. I. (2006). Attentional landmark selection for visual slam. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*.
- Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6), 989–1005.
- Gao, D., & Vasconcelos, N. (2004). Discriminant saliency for visual recognition from cluttered scenes. In *In Proc. NIPS*, (pp. 481–488).
- Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1), 51 – 64.
- Geisler, W. S., & Cormack, L. (2011). *Oxford Handbook of Eye Movements*, chap. Models of overt attention, (pp. 439–454). Oxford University Press.
- Goferman, S., Zelnik-Manor, L., & Tal, A. (2010). Context-aware saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, (pp. 2376–2383).

- Guo, C., Ma, Q., & Zhang, L. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, (pp. 1–8).
- Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Proceedings of Neural Information Processing Systems (NIPS)*, (pp. 545–552). MIT Press.
- Heidemann, G. (2004). Focus-of-attention from local color symmetries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(7), 817–830.
- Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.
- Hevner, A., & Chatterjee, S. (2010). *Design Research in Information Systems: Theory and Practice*, vol. 22. Springer.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Q.*, *28*(1), 75–105.
- Hou, X., & Zhang, L. (2007). Computer vision and pattern recognition, 2007. cvpr '07. iee conference on. In *Saliency Detection: A Spectral Residual Approach*, (pp. 1–8).
- Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, *9*(5).
- Iivari, J. (2007). A paradigmatic analysis of information systems as a design science. *Scandinavian Journal of Information Systems*, *19*(2), 39–64.
- Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, *13*(10), 1304–1318.
- Itti, L., & Baldi, P. F. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. Top cited article 2008-2010 award from Vision Research.
- Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506.
- Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*, 194–203.
- 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.
- Itti, L., Rees, G., & Tsotsos, J. K. (Eds.) (2005). *Neurobiology of Attention*. San Diego, CA: Elsevier.



- Jasso, H., & Triesch, J. (2008). Learning to attend—from bottom-up to top-down. In L. Paletta, & E. Rome (Eds.) *Top-Down Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, chap. Learning to Attend – From Bottom-Up to Top-Down, (pp. 106–122). Berlin, Heidelberg: Springer-Verlag.
- Johnson, A., & Proctor, R. W. (2004). *Attention: Theory and Practice*. SAGE Publications.
- Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *International Conference on Computer Vision (ICCV)*, (pp. 2106–2113). IEEE.
- Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, (pp. 441–480).
- Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, *45*(2), 83–105.
- Kienzle, W., Franz, M. O., Schlkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, *9*(5), 1–15.
- Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, *4*, 219–227.  
URL <http://view.ncbi.nlm.nih.gov/pubmed/3836989>
- Kootstra, G., de Boer, B., & Schomaker, L. R. B. (2011). Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive Computation*, *3*(1), 223–240.
- Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*, 1311–1328.
- Lee, T. S., & Yu, S. (2000). An information-theoretic framework for understanding saccadic behaviors. In K.-R. M. S.A. Solla, T.K. Leen (Ed.) *Advance in Neural Information Processing Systems*, vol. 12. MIT Press.
- Li, Y., Zhou, Y., Yan, J., Niu, Z., & Yang, J. (2010). Visual saliency based on conditional entropy. In H. Zha, R.-i. Taniguchi, & S. Maybank (Eds.) *Computer Vision ACCV 2009*, vol. 5994 of *Lecture Notes in Computer Science*, (pp. 246–257). Springer Berlin Heidelberg.
- Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H.-Y. (2011). Learning to detect a salient object. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(2), 353–367.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.
- Ma, Q., & Zhang, L. (2008). Saliency-based image quality assessment criterion. In D.-S. Huang, I. Wunsch, Donald C., D. Levine, & K.-H. Jo (Eds.) *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, vol. 5226 of *Lecture Notes in Computer Science*, (pp. 1124–1133). Springer Berlin Heidelberg.

- Mancas, M. (2007). *Computational Attention: Towards attentive computers*. Ph.D. thesis, Facult Polytechnique de Mons FPMs.
- Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.
- Meur, O. L., Callet, P. L., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5), 802–817.
- Mozer, M. C., & Sittton, M. (1998). *Computational modeling of spatial attention*, chap. 9, (pp. 341–393). Psychology Press.
- Murray, N., Vanrell, M., Otazu, X., & Parraga, C. A. (2011). Saliency estimation using a non-parametric low-level vision model. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11*, (pp. 433–440). Washington, DC, USA: IEEE Computer Society.
- Navalpakkam, V., & Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2, 2049 – 2056.
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 145–175.
- Oliva, A., Torralba, A., Castelhana, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 1, (pp. 253–256).
- O’Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of /‘mudsplashes/’. *Nature*, 398, 34.
- Parikh, N., Itti, L., & Weiland, J. (2010). Saliency-based image processing for retinal prostheses. *Journal of Neural Engineering*, 7(1), 016006.
- Pelz, J., Hayhoe, M., & Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, 139(3), 266–277.
- Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(8), 2397–2416.
- Pfeiffer, T., Latoschik, M. E., & Wachsmuth, I. (2008). Evaluation of binocular eye trackers and algorithms for 3d gaze interaction in virtual reality environments. *Journal of Virtual Reality and Broadcasting*, 5(16). urn:nbn:de:0009-6-16605, ISSN 1860-2037.
- Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology*, 109(2), 160–174.

- Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9), 970–982.
- Rajashekar, U., van der Linde, I., Bovik, A. C., & Cormack, L. K. (2008). Gaffe: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing*, 17(4), 564–573.
- Rao, R. P., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, 42(11), 1447–1463.  
URL <http://www.sciencedirect.com/science/article/pii/S0042698902000408>
- Reisfeld, D., Wolfson, H., & Yeshurun, Y. (1995). Context free attentional operators: The generalized symmetry transform. *International Journal of Computer Vision*, 14, 119–130.  
URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.4228>
- Renninger, L., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. *Advances in Neural Information Processing Systems*, 17, 1121–1128.
- Rensink, R. A., O’Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8(5), 368–373.
- Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39, 31573163.
- Rosenholtz, R., Dorai, A., & Freeman, R. (2011). Do predictions of visual perception aid design? *ACM Trans. Appl. Percept.*, 8(2), 12:1–12:20.
- Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12), 1–27.
- Sharma, P., & Alsam, A. (2012a). Estimating the depth in three-dimensional virtual environment with feedback. In *Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012)*, (pp. 9–17).
- Sharma, P., & Alsam, A. (2012b). Estimating the depth uncertainty in three-dimensional virtual environment. In *Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012)*, (pp. 18–25).
- Sharma, P., & Alsam, A. (2014 (accepted)). A robust metric for the evaluation of visual saliency models. In *International Conference on Computer Vision Theory and Applications (VISAPP 2014)*.
- Sharma, P., Nilsen, J. H., Skramstad, T., & Cheikh, F. A. (2010). Evaluation of geometric depth estimation model for virtual environment. In *Norsk informatikkonferanse (NIK-2010)*.
- Siagian, C., & Itti, L. (2007). Biologically-inspired robotics vision monte-carlo localization in the outdoor environment. In *IEEE International Conference on Intelligent Robots and Systems (IROS’07)*.

- Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5*, 644–649.
- Steinman, S. B., & Steinman, B. A. (1998). Vision and attention. i: Current models of visual attention. *Optometry and Vision Science*, *75*(2), 146–155.
- Styles, E. A. (2005). *The psychology of attention*. Taylor & Francis.
- Suder, K., & Worgotter, F. (2000). The control of low-level information flow in the visual system. *Reviews in the Neurosciences*, *11*, 127–146.
- Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*, 1–17.  
URL <http://www.journalofvision.org/content/7/14/4.abstract>
- Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*(5), 643–659.
- Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America. A, Optics, image science, and vision*, *20*(7), 1407–1418.
- Torralba, A., Castelhana, M. S., Oliva, A., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, *113*, 1–23.
- Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.
- Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7), 1–16.
- Tsotsos, J. K. (2011). *A Computational Perspective on Visual Attention*. MIT Press.
- Vaishnavi, V., & Kuechler, W. (2004). Design research in information systems.  
URL <http://desrist.org/design-research-in-information-systems>
- 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, (pp. I-511–I-518 vol.1).
- von Helmholtz, H. (1860 / 1962). *Physiological optics*. Dover, New York.
- Walther, D. (2006). *Interactions of Visual Attention and Object Recognition: Computational Modeling, Algorithms, and Psychophysics*. Ph.D. thesis, California Institute of Technology ,Pasadena, California.
- Walther, D., Itti, L., Riesenhuber, M., Poggio, T., & Koch, C. (2002). Attentional selection for object recognition - a gentle way. *Lecture Notes in Computer Science*, *2525*, 472–479.

- Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*, 1395–1407.
- Wang, W., Chen, C., Wang, Y., Jiang, T., Fang, F., & Yao, Y. (2011). Simulating human saccadic scanpaths on natural images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, (pp. 441–448).
- Wang, Z., & Bovik, A. C. (2001). Embedded foveation image coding. *IEEE Transactions on Image Processing*, *10*(10), 1397–1410.
- Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and limits of models of fixation selection. *PLoS ONE*, *6*(9), 1–19.
- Wolfe, J. M., Butcher, S. J., Lee, C., & Hyle, M. (2003). Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 483502.
- Yarbus, A. L. (1967). *Eye Movements and Vision*. Plenum Press.
- Yu, S. X., & Lisin, D. A. (2009). Image compression based on visual saliency at individual scales. In *ISVC '09 Proceedings of the 5th International Symposium on Advances in Visual Computing Part I*, (pp. 157–166).
- Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 1–20.



# Appendix **A**

## Research papers in full text

### **A.1 Analysis of eye fixations data**

**Authors:** Ali Alsam, and Puneet Sharma.

**Full title:** Analysis of eye fixations data.

**Published in:** Proceedings of the IASTED International Conference on Signal and Image Processing (SIP 2011), ACTA Press.

# ANALYSIS OF EYE FIXATIONS DATA

Ali Alsam and Puneet Sharma  
Department of Informatics & e-Learning(AITeL)  
Sør-Trøndelag University College(HiST)  
Trondheim, Norway  
email: er.puneetsharma@gmail.com

## ABSTRACT

In this paper, we analyzed eye fixations data obtained from 15 observers and 1003 images. When studying the correlation matrix constructed based on the fixations data of one observer viewing all images, it was observed that 23 percent of the data can be accounted for by one eigenvector. This finding implies a repeated viewing pattern that is independent of image content. The examination of this pattern revealed that it was highly correlated with the center region of the image. The presence of a repeated viewing pattern raised the following question: Is our visual attention driven by salient image content? In order to answer this question, we analyzed the data across different observers viewing the same image. Our analysis showed that there was good agreement among observers for images containing people, faces, and text while poor agreement was observed for complex images such as landscapes, buildings, and street views. Our findings suggest that strong agreement between observers was due to top-down features of the image, i.e., context driven rather than bottom up features associated with low level image attributes.

## KEY WORDS

Eye Fixations, Saliency

## 1 Introduction

A salient image region is defined as an image part that is clearly different from its surround [1]. This difference is measured in terms of a number of attributes, namely, contrast, brightness and orientation [2, 3, 4]. By measuring these attributes, visual saliency algorithms aim to predict the regions in an image that would attract our attention under free viewing conditions [2], i.e., when the observer is viewing an image without a specific task such as searching for an object. Finally, the output of the visual saliency algorithms is a so called saliency map which is a two dimensional gray scale map where the brighter regions represent higher saliency.

To evaluate the performance of visual saliency algorithms, the two dimensional saliency maps are compared with the image regions that attract observers' attention [5, 6, 7, 8, 9]. This is done by displaying to the observers a set of images and using an eye tracker to record their eye fixations. Further, it is thought that a higher num-

ber of fixations correspond to salient image regions. The recorded fixations are thus compared with the associated visual saliency maps in a pair wise manner [10, 11, 12]. Unfortunately, most studies have shown that while the saliency algorithms do predict a certain percentage of fixations they are far from being able to fully account for observers' visual attention [13]. In fact, in a recent comprehensive eye tracking study by Judd et al. [14], it was shown that a dummy classifier defined by a Gaussian blob at the center of the image was better at predicting the eye fixations than any of the visual saliency models [1, 15, 6]. In other words, assuming that the eye fixations fall at the center of the image results in better prediction than an analysis of the image content. This finding is surprising and raises the question of whether our attention is indeed guided by salient image features.

As part of the Judd et al. [14] study, a large database of 1003 images and the associated fixations data from 15 independent observers was made public to researchers in the field. This database represents the largest set of available data to date. Thus to address the issue of whether eye fixations are guided by salient image features we performed a statistical analysis of these fixations data. In this analysis, we examined the fixations data independent of the associated images. Given that a Gaussian blob was shown to predict fixations better than saliency algorithms [14], we started our examination by the assumption that fixations data are independent of image features. In order to examine the truthfulness of this assumption, we represented the fixations data from each image in terms of a  $k$  dimensional probability histogram. Representing fixations using probability histograms allowed us to compare fixations from different observers and images. Given that each histogram can be represented as a vector in a  $k$  dimensional space, we grouped all the 1003 vectors from each observer into a data matrix. Knowing that the images used in the Judd et al. [14] study were carefully chosen to be as different from each other as possible, we would expect an analysis of the associated fixations to reveal that the fixations in turn are different. This difference was analyzed by performing an eigen-decomposition of the  $k \times k$  correlation matrix obtained from the probability histograms. It is well known that the eigen-analysis of a data matrix whose vectors are linearly independent, results in a set of eigenvalues which are comparable in magnitude. On the other hand, analyz-



ing a matrix whose columns are linearly dependent results in a limited number of non-zero eigenvalues. Such matrices are known as rank deficient where the number of non-zero eigenvalues corresponds to the actual rank,  $r$ , of the matrix where  $r$  is smaller or much smaller than the dimension of the matrix.

By analyzing the correlation matrices obtained from the fixations data of the 15 observers, we found that the matrices are strongly rank deficient. Specifically, we found that the first dimension of the matrix accounts for approximately 23 percent of the data and over 93 percent is captured in the first thirty five dimensions ( $k = 80$  in our experiment). The fact that 23 percent of the data is accounted for in the first dimension can explain the high correlation between the fixations data and a Gaussian blob observed in the Judd et al. [14] study. In fact, by examining the first eigenvector of the correlation matrices we found that it represented a center like filter.

The finding that 23 percent of the data is captured by the first dimension indicates that eye fixations are image feature independent. To examine the level of independence we carried out a different experiment. In this case, we examined the correlation matrix whose vectors are the probability histograms of the 15 fixations data obtained from different observers viewing a single image. We worked under the assumption that if the first eigenvector of the correlation matrix accounts for the same data percentage as that obtained from the fixations data of one observer and different images then the fixations data can be said to be image feature independent. We found, however, that the correlation between the 15 observers viewing a single image is generally higher than that of a single observer viewing different images. This finding is again in keeping with the results obtained in the Judd et al. [14] study where it was found that the agreement among observers is higher than the Gaussian blob filter. Furthermore, this finding indicates that a certain percentage of fixations cannot be said to be image feature independent. However, by performing a visual inspection, we found that the correlation between observers is highest for images containing faces, people and text and lowest for images containing street views and landscapes. Indeed for the latter category the order of correlation was similar to that obtained based on the analysis of fixations data from a single observer viewing different images. This leads us to conclude that, based on the current data, visual attention is guided by top-down rather than bottom up image features. Further, the correspondence observed between different observers viewing complex images seems to be better explained by a common viewing mechanism rather than salient image features. Finally, the main contribution of this paper is the idea of grouping fixations data from different images which allows us to analyze the fixations independent of the corresponding images.

## 2 Method

In this section, we discuss the procedure used for the statistical analysis of the data. To allow us to compare fixations data of different durations and counts, we represented the data from each image in terms of an 8 by 10 probability histogram. For instance, figure 1 represents a typical histogram obtained by overlapping the 8 by 10 grid over the fixations data. This operation generated an 80 dimensional vector  $V_i$  for each image in the dataset. Second, the vectors  $V_1, V_2, \dots, V_n$  for all the images were normalized by their sum and grouped in a matrix  $A$  along the rows as,

$$A = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ \dots \\ V_n \end{bmatrix}$$

, where the size of  $A$  is  $n$  by 80, and  $n$  is the number of images in the database. Third, we calculated the correlation matrix  $C$  of  $A$  as,

$$C = A^T A, \quad (1)$$

where the size of  $C$  is 80 by 80. Finally, we employed the singular value decomposition(SVD) for the analysis of the correlation matrix. For a matrix  $C$ , the SVD is defined as follows:

$$C = USV^T, \quad (2)$$

where  $U$  and  $V$  are orthogonal matrices of size  $m$  by  $m$  and  $n$  by  $n$  respectively, and  $S$  is a diagonal matrix of the same size as  $C$  [16]. The diagonal elements of  $S$  are arranged in decreasing order. Since  $C$  is a symmetric matrix the singular value decomposition of  $C$  is identical to its eigen decomposition [17].

## 3 Results

### 3.1 Step 1: SVD Analysis of One Observer and Different Images

The database [14] consisted of portrait and landscape images. For our analysis we chose 463 landscape images of size 768 by 1024 pixels. We started the analysis by grouping the probability histograms based on one observer and the fixations obtained from all the images into a data matrix. As a second step, the corresponding correlation matrix was constructed and its SVD was computed using the standard matlab svd algorithm. The eigenvalues were normalized by dividing them by the sum of all the eigen values. As depicted in figure 2, the eigenvalues of the correlation matrix show that the first dimension accounts for 25 percent of the data for observer no. 1. Similar trends were observed for all the other observers, as an example, see figures 3 to 5, where we note that the first dimension represents 17 percent of the data for observer no. 2, 20 percent of the

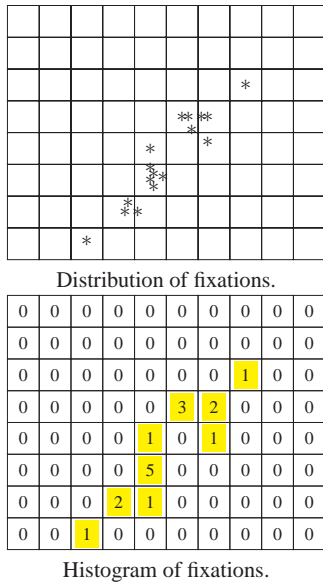


Figure 1. Histogram obtained from the fixations data.

data for observer no. 3, and 31 percent of the data for observer no. 4. To analyze the overall distribution of eigenvalues across all the observers we grouped the fixations' histograms from all the observers and images into a single data matrix and calculated its svd. Here the idea was to estimate the information captured in the first dimension for an average observer. The distribution of eigenvalues for an average observer is shown in figure 6. We observe that 23 percent of the data is accounted for by the first dimension. Considering that the images chosen in the dataset were different from each other. The fact that 23 percent of the data is represented by the first vector implies the presence of a repeated viewing pattern shared between all the observers.

Figure 7 shows the first eigenvector for the average observer. We note that it depicts a concentration of fixations in the center region of the image. This center bias in the fixations has been observed in other studies [18, 13, 14] as well. It can be responsible for the high correlation of fixations data with a dummy Gaussian classifier as observed in the study by Judd et al. [14].

### 3.2 Step 2: SVD Analysis of Different Observers and One Image

To examine the common fixations among different observers, we grouped the fixations' histograms corresponding to a single image and the 15 observers into a data matrix and computed the singular value decomposition based on the correlation matrix. The SVD of the correlation ma-

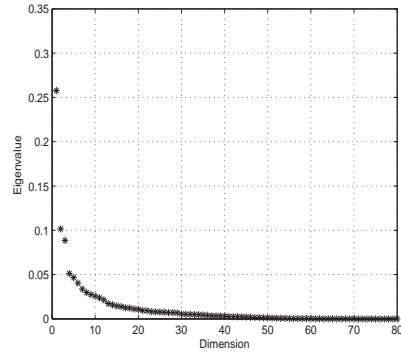


Figure 2. Distribution of eigenvalues for Observer No 1 and different images. First dimension represents 25 percent of the data.

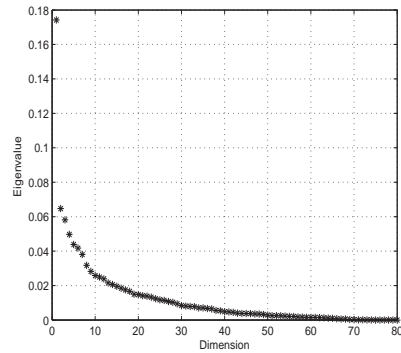


Figure 3. Distribution of eigenvalues for Observer No 2 and different images. First dimension represents 17 percent of the data.

trices, of the different images, show that depending on the image the agreement between observers' varies from 25 to 92 percent. Figure 8 shows the histogram distribution of the eigenvalues of all the images.

Our assumption is that if the fixations data are driven by image content then the first eigenvector of the correlation matrix should account for a higher percentage of the data than the fixations data of one observer and different images. As shown in figure 9, the distribution of eigenvalues for an average image shows that the first vector captures 50 percent of the data. This is clearly more than the percentage of data captured by one observer and different images discussed in the previous section. Based on this analysis we might assume that the content driven mechanisms between different observers play a significant role in the observed pattern of viewing.

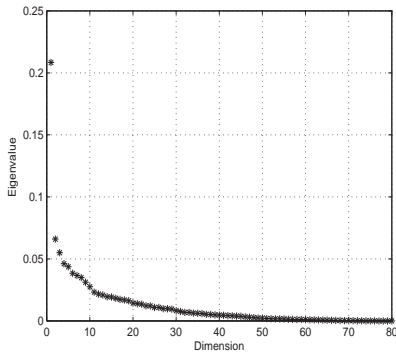


Figure 4. Distribution of eigenvalues for Observer No 3 and different images. First dimension represents 20 percent of the data.

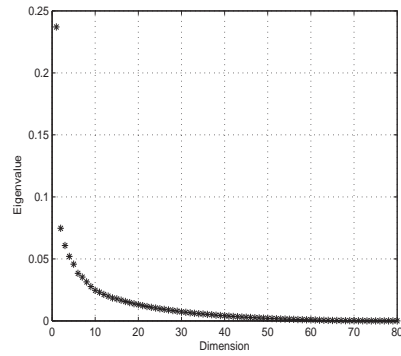


Figure 6. Distribution of eigenvalues for an average observer. 23 percent of the data is captured by the first dimension.

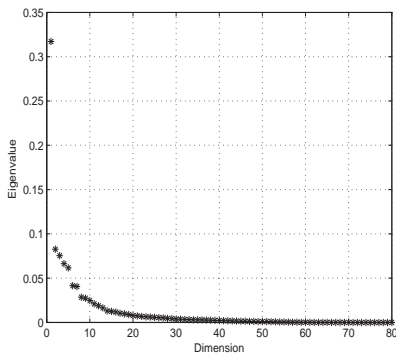
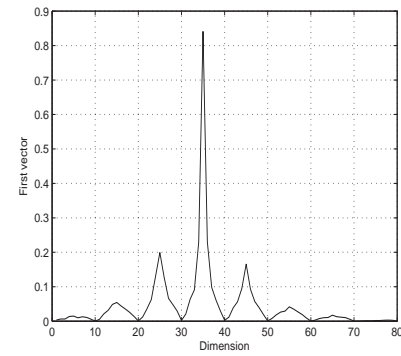
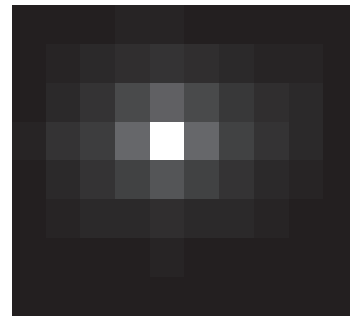


Figure 5. Distribution of eigenvalues for Observer No 4 and different images. First dimension represents 31 percent of the data.



(a) Eigenvector for an average observer.



(b) Probability histogram for the shared eigenvector.

Based on the fact that for an average image the first vector captures 50 percent of the data, we classified the fixations data into two categories. Category I consisted of the fixations data for which the first vector accounted for more than 70 percent of the data. In other words it represented the fixations where the observers were in good agreement. Category II consisted of the fixations data for which first vector accounted for less than 30 percent of the data, i.e. in the order obtained based on one observer viewing different images. In other words it represented the fixations where observers were in poor agreement. Figure 10 shows the images and the probability histograms obtained from their first vectors for the fixations where observers were in good agreement. We note that observers show good agreement for images containing people, faces, and text which has also been observed in other eye tracking studies [7, 19, 14]. Fig-

Figure 7. Eigenvector for an average observer. It shows a concentration of fixations in the center region of the image.

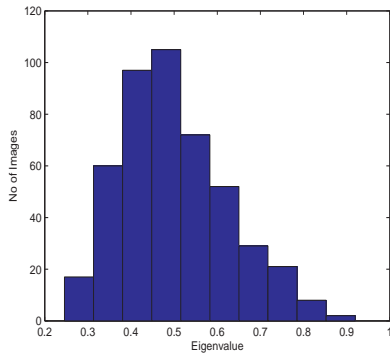


Figure 8. Histogram of first eigenvalues for all the images where mean, minimum, and maximum values of the distribution are 0.50, 0.25, and 0.92 respectively. It represents that the degree of agreement between observers' varies from image to image.

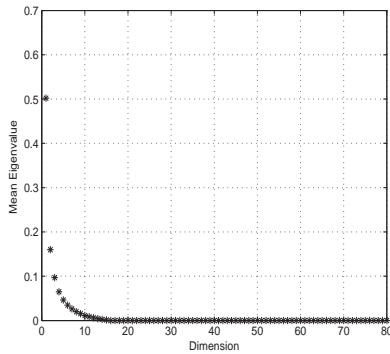


Figure 9. Distribution of eigenvalues for an average image. 50 percent of the data is captured by the first dimension. 90 percent of the data is captured by the first 5 dimensions.

Figure 11 shows the images and the probability histograms of first vectors for the fixations where observers strongly disagree. Observers show poor agreement for complex images which contain a lot of objects such as landscapes, buildings, and street views. Further, the dispersion of the first vector is more for this category of images.

## 4 Conclusions

In this paper, we analyzed the comprehensive eye fixations database by Judd et al. [14]. To allow us to compare fixations from different images and observers, the data were represented in the form of probability histograms. We observed that for an average observer, approximately 23 per-

cent of the fixation data can be represented by one eigenvector. Since the images in the database were carefully chosen to be different from each other, this indicated the presence of a repeated viewing pattern. Analysis of this pattern revealed that it represents fixations at the center of the image. This finding along with the fact that a center Gaussian blob performs better than the visual saliency models for predicting eye fixations raised the question whether visual attention is driven by the image content or by a mechanism that is unique to the observer. In order to address this issue, we analyzed the data for the cases where individual images were viewed by different observers. Results show that depending on the image the correlation between different observers varied from 25 to 92 percent. It was observed that observers were in good agreement for images containing faces, people, and text while there was poor agreement on complex images such as landscapes, buildings, and street views. Thus, for this dataset, we conclude that observers' attention is driven by two different mechanisms the first is a general pattern of viewing an arbitrary image while the second is top-down attention driven by the existence of faces, people and text.

## References

- [1] L. Itti, C. Koch, and E. Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1998, 1254–1259.
- [2] L. Itti and C. Koch, Computational modelling of visual attention, *Nature Reviews Neuroscience*, 2001, 194–203.
- [3] D. Walther and C. Koch, Modeling attention to salient proto-objects, *Neural Networks*, 19, 2006, 1395–1407.
- [4] G. Underwood, L. Humphreys, and E. Cross, Congruency, Saliency and Gist in the inspection of objects in natural scenes *Eye movements: A window on mind and brain* ( Elsevier, 2007) 563–579.
- [5] D. Walther, *Interactions of Visual Attention and Object Recognition: Computational Modeling, Algorithms, and Psychophysics*, PhD thesis ( California Institute of Technology ,Pasadena, California, 2006).
- [6] J. Harel, C. Koch, and P. Perona, Graph-based visual saliency. , *Proceedings of Neural Information Processing Systems (NIPS)*, 2006.
- [7] M. Cerf, J. Harel, W. Einhauser, and C. Koch, Predicting human gaze using low-level saliency combined with face detection, *Advances in Neural Information Processing Systems (NIPS)*, 20, 2007, 241–248.
- [8] J. M. Henderson, J. R. Brockmole, M. S. Castelhano, and M. Mack, Saliency Does Not Account for

- Eye Movements during Visual Search in Real-World Scenes, *Eye movements: A window on mind and brain*, chapter Visual (Elsevier, 2007) 537–562.
- [9] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, Gaffe: A gaze-attentive fixation finding engine, *IEEE Transactions on Image Processing*, 17(4), 2008, 564–573.
  - [10] D. Parkhurst, K. Law, and E. Niebur, Modeling the role of saliency in the allocation of overt visual attention, *Vision Research*, 42, 2002, 107–123.
  - [11] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, Top-down control of visual attention in object detection, *Proceedings International Conference on Image Processing (ICIP)*, 1, 2003, 253–256.
  - [12] J. M. Henderson, Human gaze control during real-world scene perception, *Trends in Cognitive Sciences*, 7(11), 2003, 498–504.
  - [13] B. W. Tatler, The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions, *Journal of Vision*, 7, 2007, 1–17.
  - [14] T. Judd, K. Ehinger, F. Durand, and A. Torralba, Learning to predict where humans look, *International Conference on Computer Vision (ICCV)*, 2009.
  - [15] R. Rosenholtz, A simple saliency model predicts a number of motion popout phenomena, *Vision Research*, 39, 1999, 3157–3163.
  - [16] G. Golub and W. Kahan, Calculating the singular values and pseudo-inverse of a matrix, *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, 2(2), 1965, 205–224.
  - [17] D. Kalman, A singularly valuable decomposition: The svd of a matrix, *College Math Journal*, 27, 1996, 2–23.
  - [18] O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau, A coherent computational approach to model bottom-up visual attention, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5), 2006, 802–817.
  - [19] M. Cerf, E. P. Frady, and C. Koch, Faces and text attract gaze independent of the task: Experimental data and computer model, *Journal of Vision*, 9(12), 2009, 1–15.

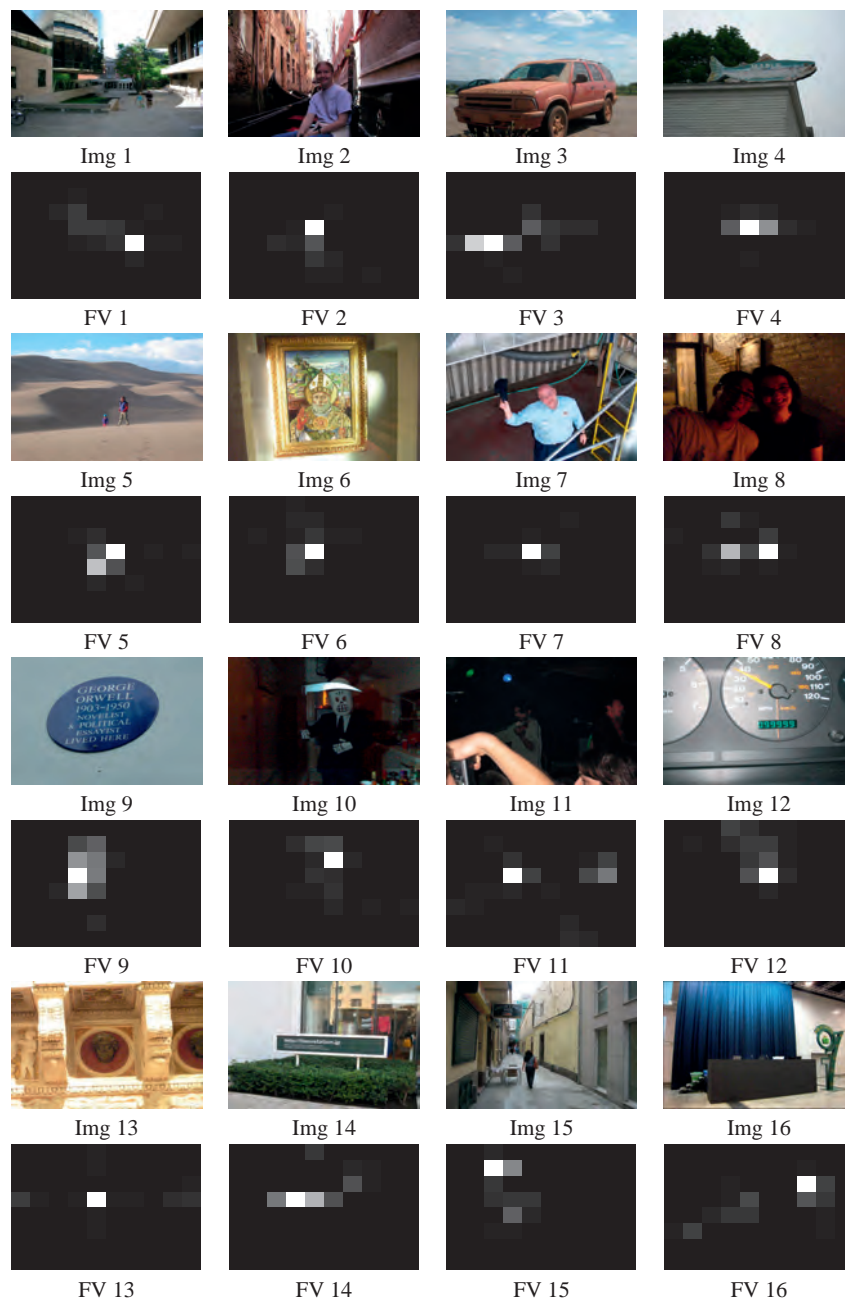


Figure 10. Category I: Images together with their first vectors. Observers show good agreement on people, faces, and text.



Figure 11. Category II: Images together with their first vectors. Observers show poor agreement for complex images such as landscapes, buildings, and street views.

## **A.2 A robust metric for the evaluation of visual saliency models**

**Authors:** Puneet Sharma and Ali Alsam.

**Full title:** A robust metric for the evaluation of visual saliency models.

**Accepted for publication in:** 9th International Conference on Vision Theory and Applications (VISAPP 2014).



# A robust metric for the evaluation of visual saliency models

Puneet Sharma and Ali Alsam

Department of Informatics & e-Learning (AITeL),  
Sør-Trøndelag University College (HiST),  
Trondheim, Norway  
`er.puneetsharma@gmail.com`

**Abstract.** Finding a robust metric for evaluating the visual saliency algorithms has been the subject of research for decades. Motivated by the shuffled *AUC* metric in this paper, we propose a robust *AUC* metric that uses the statistical analysis of the fixations data to better judge the goodness of the different saliency algorithms. To calculate the robust *AUC* metric, we use the first eigenvector obtained from the statistical analysis to define the area from which non-fixations are selected thus mitigating the effect of the center bias. Our results show that the proposed metric results in similar performance when compared with the shuffled *AUC* metric, but given that the proposed metric is derived from the statistics for the data set, we believe that it is more robust.

**Keywords:** Saliency evaluation, saliency models, fixations

## 1 Introduction

A salient image region is defined as an image part that is clearly different from its surround [1]. This difference is measured in terms of a number of attributes, namely, contrast, brightness, and orientation. By measuring these attributes, visual saliency algorithms aim to predict the regions in an image that would attract our attention under free viewing conditions [2, 1, 3, 4], i.e., when the observer is viewing an image without a specific task such as searching for an object. The output of the visual saliency algorithms is a so-called saliency map, which is a two dimensional gray scale map where the brighter regions represent higher saliency.

In the past two decades, modeling visual saliency has generated a lot of interest in the research community. In addition to contributing towards the under-

standing of human vision, it has also paved the way for a number of multimedia applications. These applications include: image and video compression [5, 6], image segmentation [7], image retrieval [8], image and video quality assessment [9, 10], and artistic image rendering [11].

To evaluate the performance of visual saliency algorithms, the two dimensional saliency maps are compared with the image regions that attract observers' attention [12–14]. This is done by displaying to the observers a set of images and using an eye tracker to record their eye fixations. Further, it is thought that a higher number of fixations correspond to salient image regions. The recorded fixations are thus compared with the associated visual saliency maps in a pair wise manner [4, 15, 16]. Unfortunately, studies [17, 11, 18] have shown that while viewing images observers tend to fixate on the center of the image more than the peripheral regions. This effect is known as center bias and is well documented in vision studies [19, 17]. The two main reasons behind this are: first, the tendency of photographers to place the objects at the center of the image. Second, the viewing strategy employed by observers, i.e., to look at center locations more in order to acquire the most information about a scene [20]. The presence of center bias in fixations makes it difficult to evaluate the correspondence between the fixated regions and the salient image regions. This can be explained by the fact that in a comprehensive eye tracking study by Judd et al. [11], it was shown that a dummy classifier defined by a Gaussian blob at the center of the image was better at predicting the eye fixations than any of the visual saliency models [1, 21, 12].

Guided by recent studies on the creation of a metric that normalizes for the influence on the center region, we studied the work by Zhang et al. [22], in which a so called shuffled *AUC* (area under the receiver-operating-characteristic curve) metric was used by the authors to mitigate the effect of center-bias in fixations.

Instead of selecting non-fixated regions from single image as is the case in the shuffled metric by Zhang et al. [22], we decided to use the repeated viewing pattern obtained from the statistical analysis of the fixations data done in the study by [18]. In their study, the repeated pattern represents the fixations that are image feature independent and is calculated as follows: first, the fixations are represented as probability histograms and each histogram is defined as a  $k$  dimensional vector. Second, the vectors for all observers and images are grouped together into a data matrix. Finally, an eigen-decomposition is performed on the  $k$  by  $k$  correlation matrix obtained from the data matrix and the first eigenvector,

i.e., the eigenvector with the highest eigenvalue is used to represent the repeated viewing pattern. We reasoned that for a given image the repeated pattern is the part which is most likely to be fixated upon, thus choosing a non-fixated region from within it for the analysis by the *AUC* metric would indeed counteract the influence of the repeated fixations pattern.

The steps in our implementation were as follows—first, the probability histogram of the repeated viewing pattern defined by the first eigenvector, is calculated and represented as a two dimensional map, second, the locations for the negative class, i.e., non-fixated are chosen from the locations where the intensity is high.

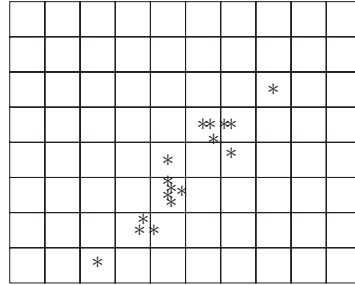
Thus the contribution of this paper is the introduction of a robust *AUC* metric that is based on the statistical analysis of fixations data and isolating that part that is common among images and observers. This metric unlike the shuffled version is not based on choosing non-fixated regions from parts that have been fixated upon in other images but rather on the whole data set making it more robust.

Finally, when we compared the performance of the new robust metric with the shuffled *AUC*, we found that the two metrics returned a similar order of best and worst algorithms with some variations that we elaborate on in the results section.

## 2 Method

In this section, we review the procedure used for the statistical analysis of the data. For the analysis 463 landscape images of size 768 by 1024 pixels are used. To allow us to compare fixations data of different durations and counts, we represented the data from each image in terms of an 8 by 10 probability histogram. For instance, figure 1 represents a typical histogram obtained by overlapping the 8 by 10 grid over the fixations data.

This operation generated an 80 dimensional vector  $V_i$  for each image in the dataset. Second, the vectors  $V_1, V_2, \dots, V_n$  for all the images were normalized by



Distribution of fixations.

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	3	2	0	0	0
0	0	0	0	1	0	1	0	0	0
0	0	0	0	5	0	0	0	0	0
0	0	0	2	1	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0

Histogram of fixations.

Fig. 1: Histogram obtained from the fixations data.

their sum and grouped in a matrix  $A$  along the rows as,

$$A = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ \dots \\ V_n \end{bmatrix}$$

where the size of  $A$  is  $n$  by 80, and  $n$  is the number of images in the database. Third, we calculated the correlation matrix  $C$  of  $A$  as,

$$C = A^T A, \quad (1)$$

where the size of  $C$  is 80 by 80. Finally, we employed the singular value decomposition (SVD) for the analysis of the correlation matrix. For a matrix  $C$ , the

SVD is defined as follows:

$$C = USV^T, \quad (2)$$

where  $U$  and  $V$  are orthonormal matrices of size  $m$  by  $m$  and  $n$  by  $n$  respectively, and  $S$  is a diagonal matrix of the same size as  $C$  [23]. The diagonal elements of  $S$  are arranged in decreasing order. Since  $C$  is a symmetric matrix the singular value decomposition of  $C$  is identical to its eigen decomposition [24].

## 2.1 Evaluation metrics

To evaluate the saliency models, an area under the receiver-operating-characteristic curve (*AUC*) metric is normally employed.

In order to calculate the *AUC* [25, 26], the fixations pertaining to a given image are averaged into a single two dimensional map which is then convolved with a two-dimensional Gaussian filter. The resultant fixations map is then thresholded to yield a binary map with two classes—the positive class consisting of fixated locations, and the negative class consisting of non-fixated locations. Next, from the two dimensional saliency map, we obtain the saliency values associated with the positive and the negative classes. Using the saliency values, a receiver-operating-characteristic (*ROC*) curve is drawn that plots the true positive rate against the false positive rate. For a detailed description of *ROC*, see the study by [25]. The area under the *ROC* curve gives us a measure of the performance of the classifier. *AUC* gives a scalar value in the interval  $[0,1]$ . If *AUC* is 1 then it indicates that the saliency model is perfect in predicting fixations. An *AUC* of 0.5 implies that the performance of the saliency model is not better than a random classifier or by chance prediction.

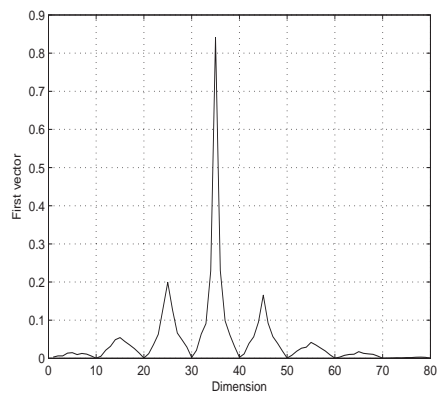
As stated previously, the shuffled *AUC* metric was used by Zhang et al. [22] to mitigate the effect of center-bias in fixations. To calculate the shuffled *AUC* metric for a given image and one observer, the locations fixated by the observer are associated with the positive class in a manner similar to the regular *AUC*, however, the locations for the negative class are selected randomly from the fixated locations of other unrelated images, such that they do not coincide with the locations from the positive class.

It is well known that the choice of locations for the negative class can influence the robustness of an *AUC* metric. To improve the robustness, we propose a modified *AUC* metric in which the negative class locations are chosen from the regions associated with high probability as described by the repeated viewing pattern (defined by the first eigenvector of the correlation matrix  $C$ ). In other

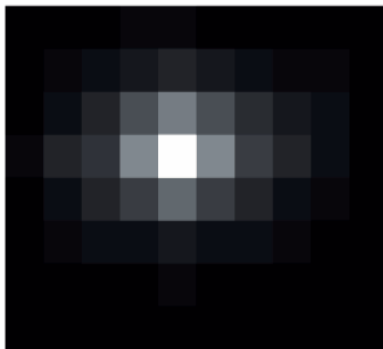
words, the locations for the negative class are selected from within the image regions that are likely to attract fixations.

### 3 Results

#### 3.1 Repeated viewing pattern



(a) First eigenvector for all images and observers.



(b) Probability histogram for the shared eigenvector.

Fig. 2: The first eigenvector for all images and observers. It shows a concentration of fixations in the center region of the image.

Figure 2 shows the repeated viewing pattern, i.e., the first eigenvector for all observers and images obtained in the study by [18]. We note that it depicts a concentration of fixations in the center region of the image. This center bias in the fixations has been observed in other studies [27, 17, 11] and it is likely responsible for the high correlation of fixations data with a dummy Gaussian classifier as noted in the study by Judd et al. [11].

### 3.2 Evaluating the visual saliency models

For evaluation, we chose eight state of the art saliency models, namely, **AIM** by Bruce & Tsotsos [28], **AWS** by Garcia-Diaz et al. [29], **SUN** by Zhang et al. [22], **Hou** by Hou & Zhang [30], **GBA** by Alsam et al. [31], **GBVS** by Harel et al. [12], **Itti** by Itti et al. [1], **Judd** by Judd et al. [11]. Figure 3 illustrates a given image and the associated saliency maps obtained from the different saliency models. In line with the study by Borji et al. [32], we used two models to provide a baseline for the evaluation. **Gauss** is defined as a two-dimensional Gaussian blob at the center of the image. This model corresponds well with the fixations falling at the image center. **IO** model is based on the fact that an observer’s fixations can be predicted best by the fixations of other observers viewing the same image. In this model the map for an observer is calculated as follows: first, the fixations corresponding to a given image from all the observers except the one under consideration are averaged into a single two-dimensional map. Having done that the fixations are spread by smoothing the map using a Gaussian filter. The **IO** model gives us an upper bound on level of correspondence that is expected between the saliency models and the fixations.

Figure 4 shows the ranking of the visual saliency models obtained by using the ordinary *AUC* metric. We observe that all saliency models used in this paper perform above chance. We also observe that **SUN**, **GBA**, **AWS**, **Hou**, **AIM**, and **Itti** perform worse than the **Gauss** model, with **GBVS**, and **Judd** being the two best models. This finding can be explained by the fact that the center regions are weighted more in both the **GBVS**, and **Judd** models.

The results obtained by employing the shuffled *AUC* metric are shown in figure 5. We note that as compared to the ordinary *AUC*, this metric changes the ranking of the saliency models significantly. As an example, the **Gauss** classifier changes from being one of the best to being clearly the worst. Further, the **GBVS**, and **Judd** models drop significantly in the rankings. In fact in this case, **AIM**, **Hou**, **GBA**, and **AWS** models are the four best models. In-line

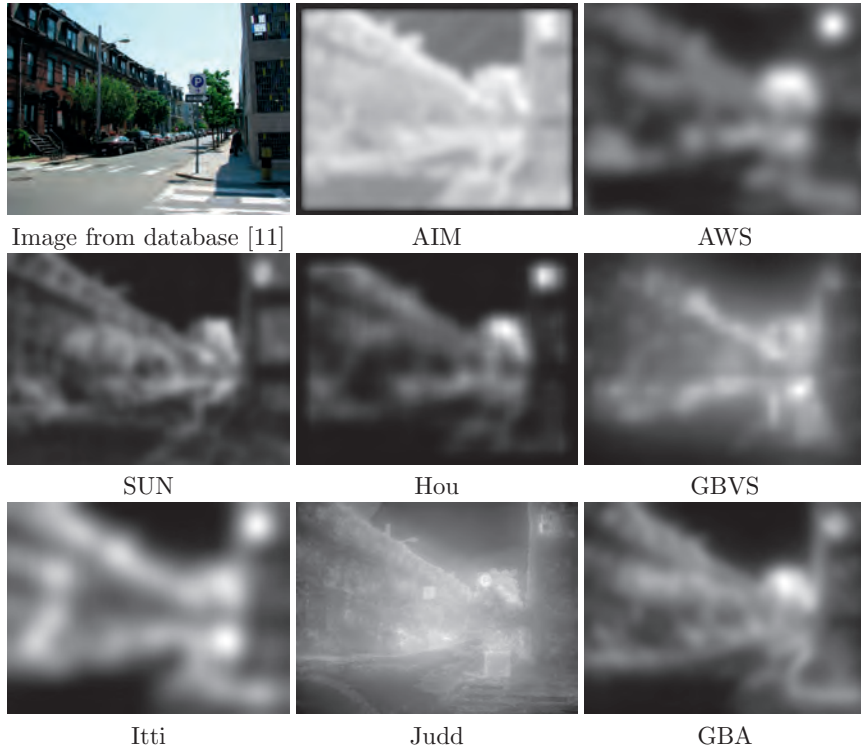


Fig. 3: Figure shows a given image and the associated saliency maps obtained from the different saliency models used in this paper.

with the study by Borji et al. [32], our results show that the **AWS** model is the best among all. The results imply that the shuffled *AUC* metric is robust to the influence of the fixations associated with the center-bias compared to the ordinary *AUC* metric.

Finally, in figure 6 we show the ranking of saliency models obtained by using the proposed robust *AUC* metric. We observe that the ranking is almost the same as the shuffled *AUC* metric, with the **AWS** model performing the best and the **Gauss** model performing the worst. We note that the robust *AUC* metric gives a lower value for the **Gauss** model, and the saliency models are closer to the **IO** model, thus, making the robust *AUC* metric a good candidate for the evaluation of saliency algorithms.



A robust metric for the evaluation of visual saliency models

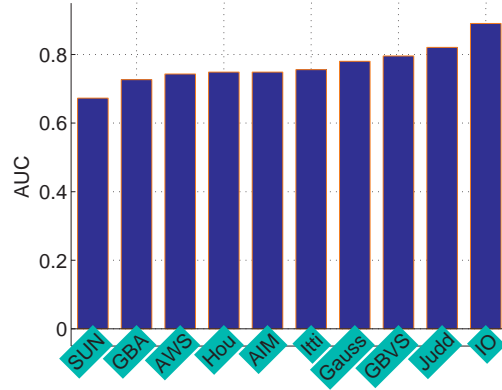


Fig. 4: Ranking of visual saliency models using the ordinary AUC metric.

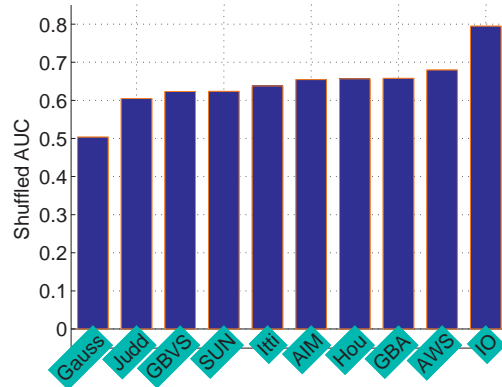


Fig. 5: Ranking of visual saliency models using the shuffled AUC metric.

## 4 Conclusions

Inspired by the shuffled *AUC* metric in this paper, we present a robust *AUC* metric that uses the statistical analysis of the fixations data to evaluate the performance of the different saliency algorithms. In order to calculate the robust *AUC* metric, we use the first eigenvector obtained from the statistical analysis to define the area from which non-fixations are selected thus abating the influence of fixations associated with the center bias. Our results show that the proposed metric results in similar performance when compared with the shuffled *AUC* but

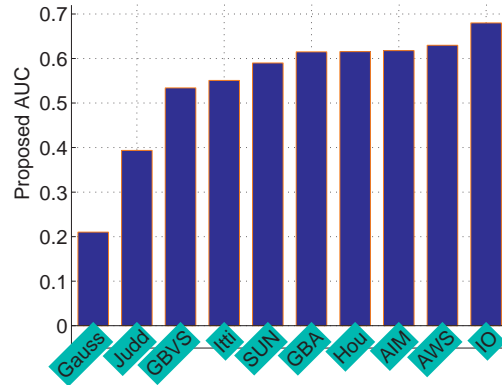


Fig. 6: Ranking of visual saliency models using the robust AUC metric.

given that the proposed metric is derived from the statistics for the data set, we believe that it is more robust.

## References

1. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (1998) 1254–1259
2. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology* **4** (1985) 219–227
3. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* **40** (2000) 1489–1506
4. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. *Vision Research* **42** (2002) 107–123
5. Itti, L.: Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing* **13** (2004) 1304–1318
6. Yu, S.X., Lisin, D.A.: Image compression based on visual saliency at individual scales. In: *ISVC '09 Proceedings of the 5th International Symposium on Advances in Visual Computing Part I*. (2009) 157–166
7. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Salient region detection and segmentation. In: *Proceedings of the 6th international conference on Computer vision systems*, Springer-Verlag (2008) 66–75
8. Kadir, T., Brady, M.: Saliency, scale and image description. *International Journal of Computer Vision* **45** (2001) 83–105

9. Feng, X., Liu, T., Yang, D., Wang, Y.: Saliency based objective quality assessment of decoded video affected by packet losses. In: 15th IEEE International Conference on Image Processing (ICIP 2008). (2008) 2560–2563
10. Ma, Q., Zhang, L.: Saliency-based image quality assessment criterion. In Huang, D.S., Wunsch, Donald C., I., Levine, D., Jo, K.H., eds.: *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*. Volume 5226 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg (2008) 1124–1133
11. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: *International Conference on Computer Vision (ICCV)*. (2009)
12. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: *Proceedings of Neural Information Processing Systems (NIPS)*. (2006)
13. Cerf, M., Harel, J., Einhauser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: *Advances in Neural Information Processing Systems (NIPS)*. Volume 20. (2007) 241–248
14. Henderson, J.M., Brockmole, J.R., Castelano, M.S., Mack, M.: Visual Saliency Does Not Account for Eye Movements during Visual Search in Real-World Scenes. In: *Eye movements: A window on mind and brain*. Elsevier (2007) 537–562
15. Oliva, A., Torralba, A., Castelano, M.S., Henderson, J.M.: Top-down control of visual attention in object detection. In: *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. Volume 1. (2003) 253–256
16. Henderson, J.M.: Human gaze control during real-world scene perception. *Trends in Cognitive Sciences* **7** (2003) 498–504
17. Tatler, B.W.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision* **7** (2007) 1–17
18. Alsam, A., Sharma, P.: Analysis of eye fixations data. In: *Proceedings of the IASTED International Conference, Signal and Image Processing (SIP 2011)*. (2011) 342–349
19. Tatler, B.W., Baddeley, R.J., Gilchrist, I.D.: Visual correlates of fixation selection: effects of scale and time. *Vision Research* **45** (2005) 643659
20. Tseng, P.H., Carmi, R., Cameron, I.G.M., Munoz, D.P., Itti, L.: Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision* **9** (2009) 1–16
21. Rosenholtz, R.: A simple saliency model predicts a number of motion popout phenomena. *Vision Research* **39** (1999) 31573163
22. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision* **8** (2008) 1–20
23. Golub, G., Kahan, W.: Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics:Series B, NumericalAnalysis* **2** (1965) 205–224

24. Kalman, D.: A singularly valuable decomposition: The svd of a matrix. *College Math Journal* **27** (1996) 2–23
25. Fawcett, T.: Roc graphs: Notes and practical considerations for researchers. *Pattern Recognition Letters* **27** (2004) 882–891
26. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35** (2013) 185–207
27. Meur, O.L., Callet, P.L., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28** (2006) 802–817
28. Bruce, N.D.B., Tsotsos, J.K.: Saliency based on information maximization. In: *NIPS'05*. (2005) 155–162
29. Garcia-Diaz, A., Fdez-Vidal, X.R., Pardo, X.M., Dosal, R.: Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing* **30** (2012) 51 – 64
30. Hou, X., Zhang, L.: Computer vision and pattern recognition, 2007. *cvpr '07. ieee conference on*. In: *Saliency Detection: A Spectral Residual Approach*. (2007) 1–8
31. Alsam, A., Sharma, P., Wrålsen, A.: Asymmetry as a measure of visual saliency. In: *SCIA 2013, Lecture Notes in Computer Science (LNCS)*. Volume 7944., Springer-Verlag Berlin Heidelberg (2013) 591–600
32. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing* **22** (2013) 55–69

### **A.3 A robust metric for the evaluation of visual saliency models (extended)**

**Authors:** Ali Alsam and Puneet Sharma.

**Full title:** A robust metric for the evaluation of visual saliency models.

**Published in:** Vol. 31, No. 3, Journal of the Optical Society of America A (JOSA A) 2014.

# Robust metric for the evaluation of visual saliency algorithms

Ali Alsam and Puneet Sharma\*

Department of Informatics & e-Learning (AITeL), Sor-Trøndelag University College (HiST), Trondheim, Norway

\*Corresponding author: er.puneetsharma@gmail.com

Received May 28, 2013; revised October 13, 2013; accepted November 22, 2013;  
posted December 23, 2013 (Doc. ID 191237); published February 13, 2014

In this paper, we analyzed eye fixation data obtained from 15 observers and 1003 images. When studying the eigen-decomposition of the correlation matrix constructed based on the fixation data of one observer viewing all images, it was observed that 23% of the data can be accounted for by one eigenvector. This finding implies a repeated viewing pattern that is independent of image content. Examination of this pattern revealed that it was highly correlated with the center region of the image. The presence of a repeated viewing pattern raised the following question: can we use the statistical information contained in the first eigenvector to filter out the fixations that were part of the pattern from those that are image feature dependent? To answer this question we designed a robust AUC metric that uses statistical analysis to better judge the goodness of the different saliency algorithms. © 2014 Optical Society of America

OCIS codes: (070.0070) Fourier optics and signal processing; (100.2960) Image analysis; (100.5010) Pattern recognition; (330.4060) Vision modeling.  
<http://dx.doi.org/10.1364/JOSAA.31.000532>

## 1. INTRODUCTION

A salient image region is defined as an image part that is clearly different from its surround [1,2]. This difference is measured in terms of a number of attributes, namely, contrast, brightness, and orientation [3–5]. By measuring these attributes, visual saliency algorithms aim to predict the regions in an image that would attract our attention under free viewing conditions [4], i.e., when the observer is viewing an image without a specific task such as searching for an object. Finally, the output of the visual saliency algorithms is a so-called saliency map, which is a two-dimensional gray scale map in which the brighter regions represent higher saliency.

To evaluate the performance of visual saliency algorithms, the two-dimensional saliency maps are compared with the image regions that attract observers' attention [6–10]. This is done by displaying to the observers a set of images and using an eye tracker to record their eye fixations. Further, it is thought that a higher number of fixations correspond to salient image regions. The recorded fixations are thus compared with the associated visual saliency maps in a pair wise manner [11–13]. Unfortunately, most studies have shown that while the saliency algorithms do predict a certain percentage of fixations, they are far from being able to fully account for observers' visual attention [14,15]. In fact, in a recent comprehensive eye tracking study by Judd *et al.* [16], it was shown that a dummy classifier defined by a Gaussian blob at the center of the image was better at predicting the eye fixations than any of the visual saliency models [7,1,17]. In other words, assuming that the eye fixations fall at the center of the image results in better prediction than an analysis of the image content. This finding is surprising and raises the question of whether our attention is indeed guided by salient image features.

As part of the Judd *et al.* [16] study, a large database of 1003 images and the associated fixation data from 15 independent observers were made public to researchers in the field. This database represents the largest set of available data to date. Thus to address the issue of whether eye fixations are guided by salient image features we performed a statistical analysis of these fixation data. In this analysis, we examined the fixation data independent of the associated images. Given that a Gaussian blob was shown to predict fixations better than saliency algorithms [16], we started our examination by the assumption that fixation data are independent of image features. In order to examine the truthfulness of this assumption, we represented the fixation data from each image in terms of a  $k$ -dimensional probability histogram. Representing fixations using probability histograms allowed us to compare fixations from different observers and images. Given that each histogram can be represented as a vector in a  $k$ -dimensional space, we grouped all the 1003 vectors from each observer into a data matrix. Knowing that the images used in the Judd *et al.* [16] study were carefully chosen to be as different from each other as possible, we would expect an analysis of the associated fixations to reveal that the fixations in turn are different. This difference was analyzed by performing an eigen-decomposition of the  $k \times k$  correlation matrix obtained from the probability histograms. It is well known that eigen-analysis of a data matrix whose vectors are linearly independent results in a set of eigenvalues that are comparable in magnitude. On the other hand, analyzing a matrix whose columns are linearly dependent results in a limited number of nonzero eigenvalues. Such matrices are known as rank deficient where the number of nonzero eigenvalues corresponds to the actual rank,  $r$ , of the matrix where  $r$  is smaller or much smaller than the dimension of the matrix.

By analyzing the correlation matrices obtained from the fixation data of the 15 observers, we found that the matrices are strongly rank deficient. Specifically, we found that the first dimension of the matrix accounts for approximately 23% of the data and more than 93% is captured in the first 35 dimensions ( $k = 80$  in our experiment). The fact that 23% of the data is accounted for in the first dimension can explain the high correlation between the fixation data and a Gaussian blob observed in the Judd *et al.* [16] study. In fact, by examining the first eigenvector of the correlation matrices we found that it represented a center like filter.

The finding that 23% of the data is captured by the first dimension indicates that eye fixations are image feature independent. To examine the level of independence we carried out a different experiment. In this case, we examined the correlation matrix whose vectors are the probability histograms of the 15 fixation data obtained from different observers viewing a single image. We worked under the assumption that if the first eigenvector of the correlation matrix accounts for the same data percentage as that obtained from the fixation data of one observer and different images, then the fixation data can be said to be image feature independent. We found, however, that the correlation between the 15 observers viewing a single image is generally higher than that of a single observer viewing different images. This finding is again in keeping with the results obtained in the Judd *et al.* [16] study, where it was found that the agreement among observers is higher than the Gaussian blob filter. Furthermore, this finding indicates that a certain percentage of fixations cannot be said to be image feature independent. However, by performing a visual inspection, we found that the correlation between observers is highest for images containing faces, people, and text and lowest for images containing street views and landscapes. Indeed for some images in the latter category the order of correlation was similar to that obtained based on the analysis of fixation data from a single observer viewing different images. This might lead us to conclude that visual attention is guided by top-down rather than bottom-up image features; however, separating the implicit spatial bias represented by top-down features from stimulus-dependent and scene-dependent fixations would require a detailed forensic analysis. Given that these different mechanisms, and perhaps others, are at play and cannot be disentangled from one another nor from the scene composition, it is challenging to make assertions about their relative contributions.

The first contribution of this paper is the idea of grouping fixation data from different images, which allows us to analyze the fixations independent of the corresponding images.

Visual inspection of a number of images cannot be said to qualify as a reliable metric of the performance of visual saliency algorithms and is thus short of enabling us to claim that salient image regions have no influence on our visual attention.

Guided by recent studies on the creation of a metric that normalizes for the influence on the center region, we studied the work by Tatler *et al.* [15], in which a so-called shuffled area under the receiver-operating-characteristic curve (AUC) metric was used by the authors to mitigate the effect of center bias in fixations.

To calculate the shuffled AUC metric for a given image and one observer, the locations fixated on by the observer are

associated with the positive class in a manner similar to the regular AUC; however, the locations for the negative class are selected randomly from the fixated locations of other unrelated images, such that they do not coincide with the locations from the positive class.

Instead of selecting nonfixated regions from a single image as is the case in the shuffled metric by Tatler *et al.* [15], we decided to use the values returned by the first eigenvector of the correlation matrix as a filter from which we can choose nonfixated regions. We reasoned that the common pattern found in the first part of the experiment is the part that is most likely to be fixated upon; thus choosing a nonfixated region from within it for the analysis by the AUC metric would indeed counteract the influence of the repeated fixations pattern. Here, it is important to state that by choosing the first eigenvector we are not claiming that there is no commonality captured by the second or indeed higher frequency components. That said, principle component analysis returns orthogonal vectors, and combining two or more features to represent commonality requires an in-depth study with strong reasoning. We thus disregard any influence of the second component on the common viewing pattern based on the evidence that the second eigenvalue captures 7% of the data compared to the 23% captured by the first.

The steps in our implementation were as follows: first, the probability histogram of the repeated viewing pattern defined by the first eigenvector is calculated and represented as a two-dimensional map. Second, the locations for the negative class, i.e., nonfixated, are chosen from the locations where the intensity is high.

The second contribution of this paper is the introduction of a robust AUC metric that is based on statistical analysis of fixation data and isolating that part that is common among images and observers. This metric, unlike the shuffled version, is not based on choosing nonfixated regions from parts that have been fixated upon in other images, but rather on the whole dataset, making it more robust.

Finally, when we compared the performance of the new robust metric with the shuffled AUC, we found that the two metrics returned a similar order of best and worst algorithms with some variations that we elaborate on in the results section.

## 2. METHOD

In this section, we discuss the procedure used for statistical analysis of the data. To allow us to compare fixation data of different durations and counts, we represented the data from each image in terms of an 8-by-10 probability histogram. For instance, Fig. 1 represents a typical histogram obtained by overlapping the 8-by-10 grid over the fixation data. This operation generated an 80-dimensional vector  $V_i$  for each image in the dataset. Second, the vectors  $V_1, V_2, \dots, V_n$  for all the images were normalized by their sum and grouped in a matrix  $A$  along the rows as

$$A = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ \dots \\ V_n \end{bmatrix},$$

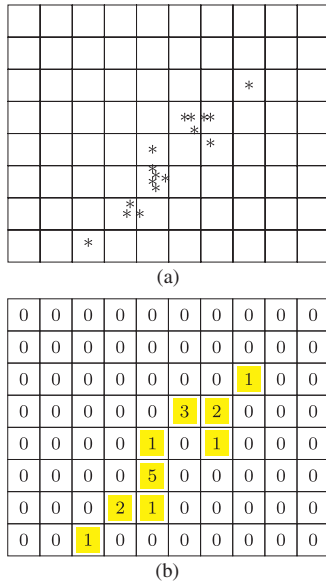


Fig. 1. Histogram obtained from the fixation data. (a) Distribution of fixations and (b) histogram of fixations.

where the size of  $A$  is  $n$  by 80, and  $n$  is the number of images in the database. Third, we calculated the correlation matrix  $C$  of  $A$  as

$$C = A^T A, \tag{1}$$

where the size of  $C$  is 80 by 80. Finally, we employed the singular value decomposition (SVD) for analysis of the correlation matrix. For a matrix  $C$ , the SVD is defined as follows:

$$C = USV^T, \tag{2}$$

where  $U$  and  $V$  are orthonormal matrices of size  $m$  by  $m$  and  $n$  by  $n$ , respectively, and  $S$  is a diagonal matrix of the same size as  $C$  [18]. The diagonal elements of  $S$  are arranged in decreasing order. Since  $C$  is a symmetric matrix, the SVD of  $C$  is identical to its eigen-decomposition [19].

**A. Evaluation Metrics**

To evaluate the saliency models, an AUC metric is normally employed.

In order to calculate the AUC [20,21], the fixations pertaining to a given image are averaged into a single two-dimensional map, which is then convolved with a two-dimensional Gaussian filter. The resultant fixations map is then thresholded to yield a binary map with two classes—the positive class consisting of fixated locations, and the negative class consisting of nonfixated locations. Next, from the two-dimensional saliency map, we obtain the saliency values associated with the positive and the negative classes. Using the saliency values, a receiver-operating-characteristic (ROC) curve is drawn that plots the true positive rate against the false positive rate. For a detailed description of ROC, see the study by [21]. The area under the ROC curve gives us a measure of the performance of the classifier. AUC gives a

scalar value in the interval [0,1]. If AUC is 1, then it indicates that the saliency model is perfect in predicting fixations. An AUC of 0.5 implies that the performance of the saliency model is not better than a random classifier or by chance prediction.

To improve the robustness of the shuffled AUC metric, we propose a modified AUC metric in which the negative class locations are chosen from the regions associated with high probability as described by the repeated viewing pattern (defined by the first eigenvector of the correlation matrix  $C$ ). In other words, the locations for the negative class are selected from within the image regions that are likely to attract fixations.

**3. RESULTS**

**A. Step 1: SVD Analysis of One Observer and Different Images**

The database [16] consisted of portrait and landscape images. For our analysis we chose 463 landscape images of size 768 by 1024 pixels. We did not use the whole database due to the difficulty in combining the histograms of the landscape images with portraits. Further, the portrait images were of different sizes. We thus limited the analysis to the landscape images to avoid any influence of rotating and resizing the rest of the data. We started the analysis by grouping the probability histograms based on one observer and the fixations obtained from all the images into a data matrix. As a second step, the corresponding correlation matrix was constructed, and its SVD was computed using the standard MATLAB SVD algorithm. The eigenvalues were normalized by dividing them by their sum. As depicted in Fig. 2, the eigenvalues of the correlation matrix show that the first dimension accounts for 25% of the data for observer No. 1. Similar trends were observed for all the other observers; as an example, see Figs. 3–5, where we note that the first dimension represents 17% of the data for observer No. 2, 20% of the data for observer No. 3, and 31% of the data for observer No. 4. To analyze the overall distribution of eigenvalues across all the observers we grouped the fixations’ histograms from all the observers and images into a single data matrix and calculated its SVD. Here the idea was to estimate the information captured in the first dimension for an average observer. The distribution of eigenvalues

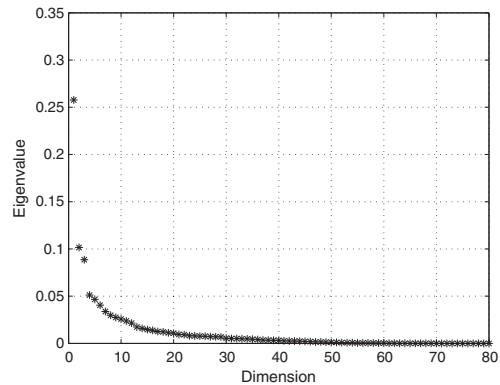


Fig. 2. Distribution of eigenvalues for observer No. 1 and different images. First dimension represents 25% of the data.



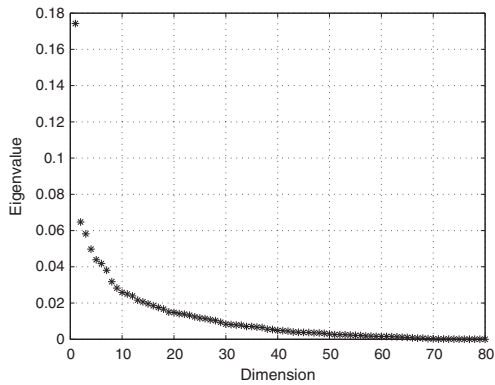


Fig. 3. Distribution of eigenvalues for observer No. 2 and different images. First dimension represents 17% of the data.

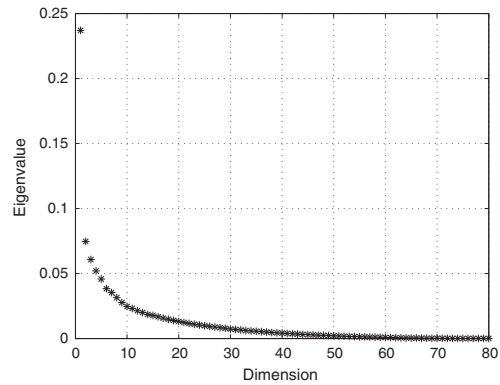


Fig. 6. Distribution of eigenvalues for an average observer. 23% of the data is captured by the first dimension.

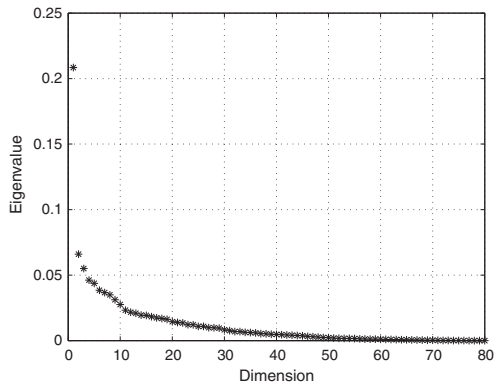
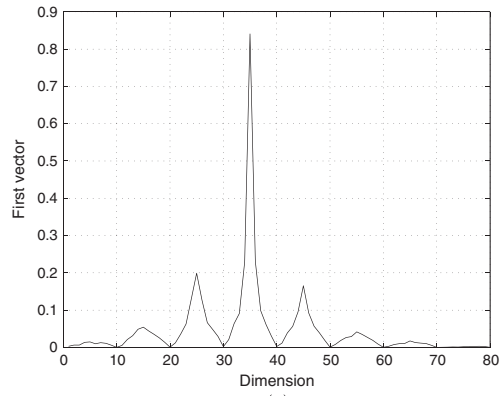


Fig. 4. Distribution of eigenvalues for observer No. 3 and different images. First dimension represents 20% of the data.



(a)

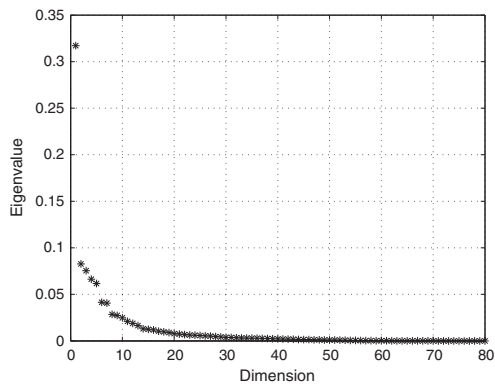


Fig. 5. Distribution of eigenvalues for observer No. 4 and different images. First dimension represents 31% of the data.



(b)

Fig. 7. Eigenvector for an average observer. It shows a concentration of fixations in the center region of the image. (a) Eigenvector for an average observer. (b) Probability histogram for the shared eigenvector.

for an average observer is shown in Fig. 6. We observe that 23% of the data is accounted for by the first dimension, considering that the images chosen in the dataset were different from each other. The fact that 23% of the data is represented

by the first vector implies the presence of a repeated viewing pattern shared between all the observers.

Figure 7 shows the first eigenvector for the average observer. We note that it depicts a concentration of fixations in the center region of the image. This center bias in the fixations has been observed in other studies [16,22,14], and it is

likely responsible for the high correlation of fixation data with a dummy Gaussian classifier as observed in the study by Judd *et al.* [16].

### B. Step 2: SVD Analysis of Different Observers and One Image

To examine the common fixations among different observers, we grouped the fixations' histograms corresponding to a single image and the 15 observers into a data matrix. The SVDs of the correlation matrices, of the different images, show that depending on the image the agreement between observers varies from 25% to 92%. Figure 8 shows the histogram distribution of the first eigenvalues of all the images, where the average value is approximately 50%. Figure 9 shows the average eigenvalues based on all observers viewing a single image. We note that the average agreement between different observers viewing a single image, 50%, is double that obtained when a single observer views different images, 23%. We can thus state that the difference between the commonality observed

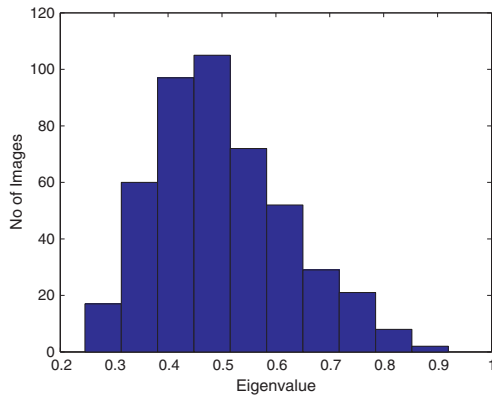


Fig. 8. Histogram of first eigenvalues for all the images where the mean, minimum, and maximum values of the distribution are 0.50, 0.25, and 0.92 respectively. It represents that the degree of agreement between observers varies from image to image.

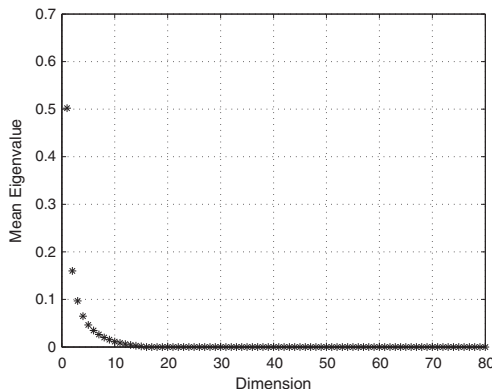


Fig. 9. Distribution of eigenvalues for an average image. 50% of the data is captured by the first dimension. 90% of the data is captured by the first five dimensions.

in the viewing pattern when one observer views different images and that observed in the case of all observers viewing a single image is due to the image-dependent features, including, i.e., implicit spatial bias represented by top-down features, and stimulus-dependent and scene-dependent fixations.

Based on the finding that for an average image the first vector captures 50% of the data, we classified the fixations into two categories. Category I consisted of the fixation data for which the first vector accounted for more than 70% of the data. In other words it represented the fixations where the observers were in good agreement. Category II consisted of the fixation data for which the first vector accounted for less than 30% of the data, i.e., in the order obtained based on one observer viewing different images. Figure 10 shows the images and the probability histograms obtained for the image where observers were in good agreement. We note that observers show good agreement for images containing people, faces, and text, something that has also been observed in other eye tracking studies [24,6,16]. Figure 11 shows the images and the probability histograms for images where observers disagree. Observers show poor agreement for complex images that contain a lot of objects such as landscapes, buildings, and street views. Further, the dispersion of the first vector is more for this category of images.

### C. Evaluating the Visual Saliency Models

For evaluation, we chose 11 state-of-the-art saliency models, namely, AIM by Bruce and Tsotsos [25], AWS by Garcia-Diaz *et al.* [26], Erdem by Erdem and Erdem [27], Hou by Hou and Zhang [28], Spectral by Schauerte and Stiefelwagen [29], SUN by Zhang *et al.* [30], GBA by Alsam *et al.* [31], GBVS by Harel *et al.* [7], Itti by Itti *et al.* [1], Judd by Judd *et al.* [16], and LG by Borji and Itti [32]. In line with the study by Borji *et al.* [33], we used two models to provide a baseline for the evaluation. Gauss is defined as a two-dimensional Gaussian blob at the center of the image. This model corresponds well with the fixations falling at the image center. The IO model is based on the fact that an observer's fixations can be predicted best by the fixations of other observers viewing the same image. In this model the map for an observer is calculated as follows: first, the fixations corresponding to a given image from all the observers except the one under consideration are averaged into a single two-dimensional map. Having done that the fixations are spread by smoothing the map using a Gaussian filter. The size of the Gaussian filter as well as the level of smoothing are two factors known to influence the performance of the models. To avoid introducing too many variables, we used the fixation maps from the Judd *et al.* [16] study without varying the level of smoothing. The IO model gives us an upper bound on the level of correspondence that is expected between the saliency models and the fixations.

Figure 12 shows the ranking of the visual saliency models obtained by using the ordinary *AUC* metric. We observed that all saliency models used in this paper perform above chance. We also observed that SUN, Spectral, GBA, LG, AWS, Hou, AIM, and Itti perform worse than the Gauss model, with GBVS, Erdem, and Judd being the three best models. This finding can be explained by the fact that the center regions are weighted more in the GBVS, Erdem, and Judd models.

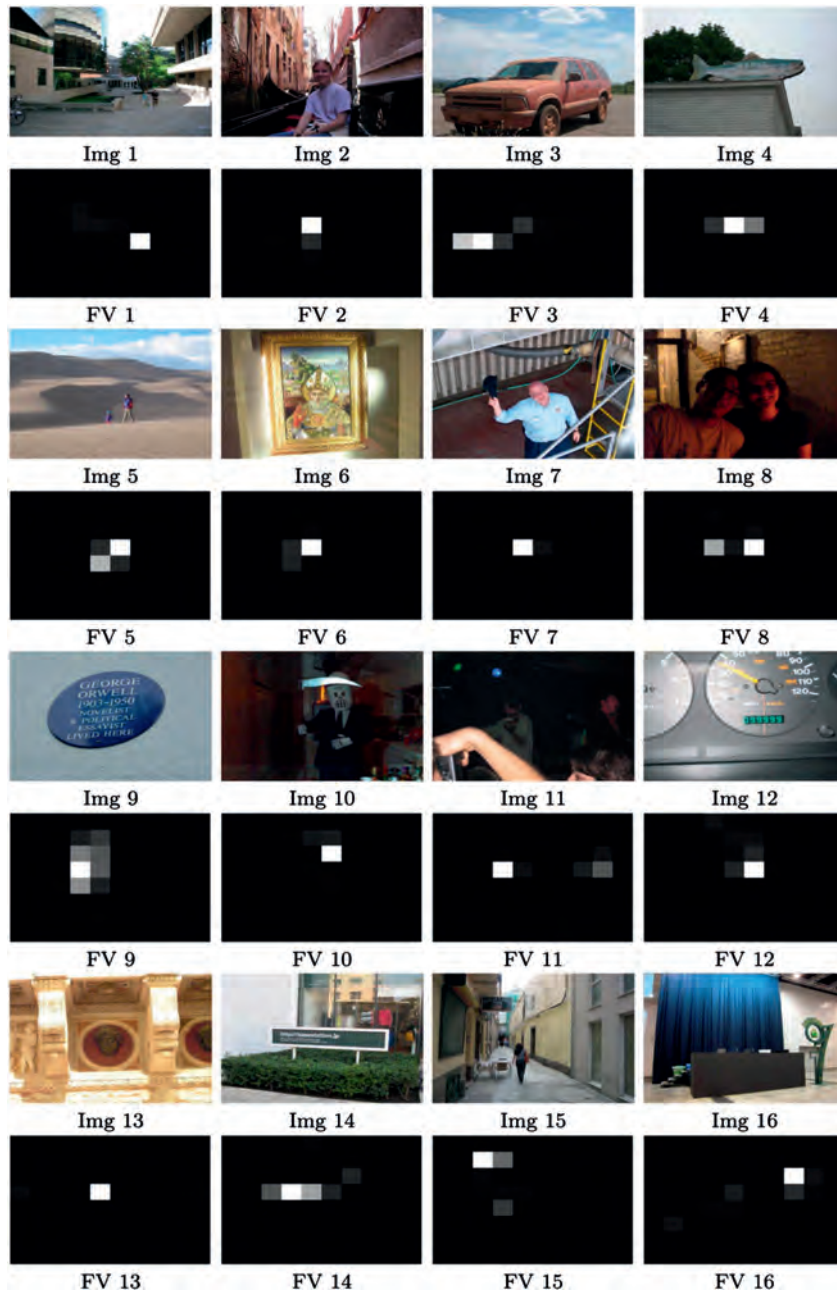


Fig. 10. Category I: images together with their first vectors. Observers show good agreement on people, faces, and text. Img 1–16 used from the database by Judd et al. [16], and the LabelMe dataset by Russell et al. [23].

The results obtained by employing the shuffled *AUC* metric are shown in Fig. 13. We note that as compared to the ordinary *AUC*, this metric changes the ranking of the saliency models significantly. As an example, the Gauss classifier changes from being one of the best to being clearly the worst. Further, the GBVS, Erdem, and Judd models drop significantly in the

rankings. In fact in this case, the LG and AWS models are the two best models. In line with the study by Borji et al. [33], our results show that the AWS model is the best among all. The results imply that the shuffled *AUC* metric is robust to the influence of the fixations associated with the center bias compared to the ordinary *AUC* metric.

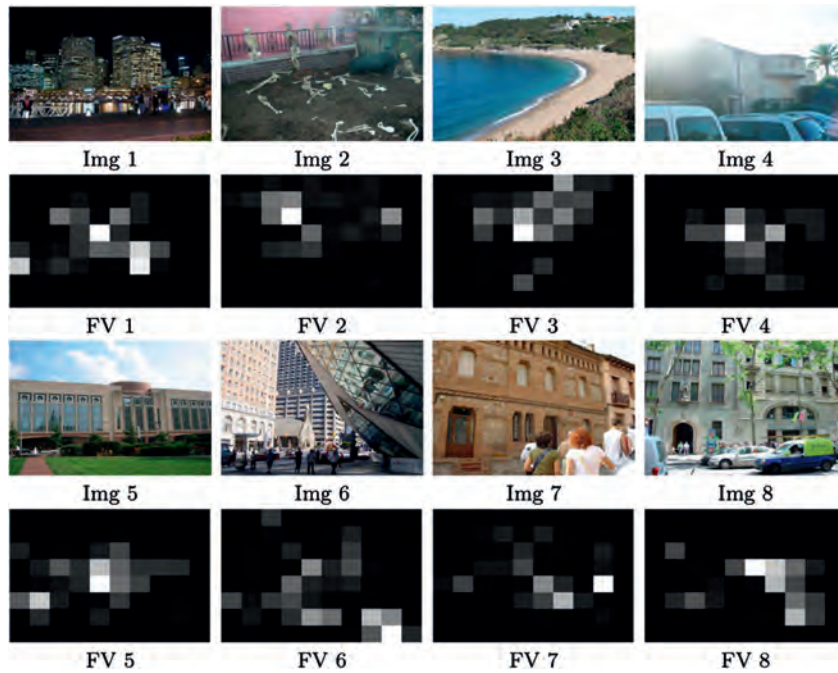


Fig. 11. Category II: images together with their first vectors. Observers show poor agreement for complex images such as landscapes, buildings, and street views. Img 1–8 used from the database by Judd *et al.* [16], and the LabelMe dataset by Russell *et al.* [23].

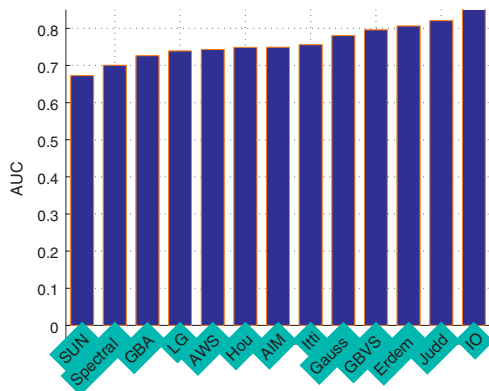


Fig. 12. Ranking of visual saliency models using the ordinary AUC metric.

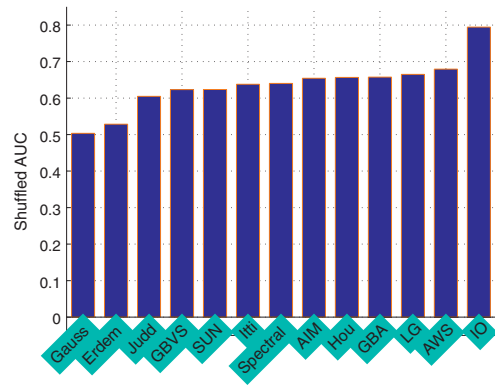


Fig. 13. Ranking of visual saliency models using the shuffled AUC metric.

Next, in Fig. 14 we show the ranking of saliency models obtained by using the proposed robust *AUC* metric. We observe that the ranking is almost the same as the shuffled *AUC* metric, with the AWS model performing the best and the Gauss model performing the worst. We note that the robust *AUC* metric gives a lower value for the Gauss model, and the saliency models are closer to the *IO* model, thus making the robust *AUC* metric a good candidate for the evaluation of saliency algorithms.

Finally, in Figs. 13 and 14, we can see that there is a gap between the performance of visual saliency models and that of human performance represented by the *IO* model; further, this difference is less when we use the proposed *AUC* metric. Although some recent models such as AWS, and LG, have reduced this disparity, adding top-down features in the saliency models is seen as the next step toward bridging this gap. To this end, researchers [16,24,33] have suggested features such as faces, people, text, and objects of interest [16], such as cars, human body parts, and animals.

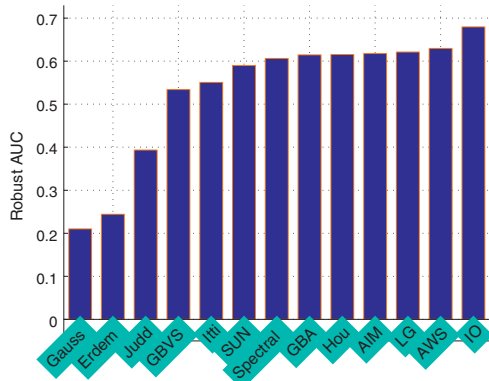


Fig. 14. Ranking of visual saliency models using the proposed AUC metric.

#### 4. CONCLUSIONS

In this paper, we analyzed the comprehensive eye fixations database by Judd *et al.* [16]. To allow us to compare fixations from different images and observers, the data were represented in the form of probability histograms. We observed that for an average observer, approximately 23% of the fixation data can be represented by one eigenvector. Since the images in the database were carefully chosen to be different from each other, this indicated the presence of a repeated viewing pattern. Analysis of this pattern revealed that it represents fixations at the center of the image. This finding along with the fact that a center Gaussian blob performs better than the visual saliency models for predicting eye fixations raised the question of whether visual attention is driven by the image content or by a mechanism that is unique to the observer. In order to address this issue, we analyzed the data for the cases in which individual images were viewed by different observers. Results show that depending on the image the correlation between different observers varied from 25% to 92%. It was observed that observers were in good agreement for images containing faces, people, and text, while there was poor agreement on complex images such as landscapes, buildings, and street views. Thus, for this dataset, we conclude that observers' attention is driven by two different mechanisms: the first is a general pattern of viewing an arbitrary image, while the second is driven by a mixture of top-down and bottom-up attention.

Knowing that the repeated pattern described by the first eigenvector represents fixations at the center region of the image we decided to investigate its usefulness in designing a variant of the *AUC* metric. Inspired by the shuffled *AUC* metric, we used the first eigenvector obtained from the statistical analysis to define the area from which nonfixations are selected, thus mitigating the effect of the center region.

Our results show that the proposed metric results in similar performance when compared with the shuffled *AUC* but given that the proposed metric is derived from

the statistics for the dataset, we believe that it is more robust.

#### REFERENCES

1. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259 (1998).
2. C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiol.* **4**, 219–227 (1985).
3. L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.* **40**, 1489–1506 (2000).
4. L. Itti and C. Koch, "Computational modelling of visual attention," *Nat. Rev. Neurosci.* **2**, 194–203 (2001).
5. D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks* **19**, 1395–1407 (2006).
6. M. Cerf, J. Harel, W. Einhauser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in Neural Information Processing Systems (NIPS)* (2007), Vol. **20**, pp. 241–248.
7. J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems (NIPS)* (2006).
8. J. M. Henderson, J. R. Brockmole, M. S. Castelano, and M. Mack, "Visual saliency does not account for eye movements during visual search in real-world scenes," in *Eye Movements: A Window on Mind and Brain* (Elsevier, 2007), pp. 537–562.
9. U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "Gaffe: a gaze-attentive fixation finding engine," *IEEE Trans. Image Process.* **17**, 564–573 (2008).
10. D. Walther, "Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics," Ph.D. thesis (California Institute of Technology, 2006).
11. J. M. Henderson, "Human gaze control during real-world scene perception," *Trends Cogn. Sci.* **7**, 498–504 (2003).
12. A. Oliva, A. Torralba, M. S. Castelano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proceedings of International Conference on Image Processing (IEEE, 2003)*, pp. 253–256.
13. D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vis. Res.* **42**, 107–123 (2002).
14. B. W. Tatler, "The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.* **7**(14):4 (2007).
15. B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vis. Res.* **45**, 643–659 (2005).
16. T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *International Conference on Computer Vision (ICCV)* (IEEE, 2009).
17. R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," *Vis. Res.* **39**, 3157–3163 (1999).
18. G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *J. Soc. Ind. Appl. Math. Ser. B* **2**, 205–224 (1965).
19. D. Kalman, "A singularly valuable decomposition: the SVD of a matrix," *College Math J.* **27**, 2–23 (1996).
20. A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 185–207 (2013).
21. T. Fawcett, "ROC graphs with instance-varying costs," *Pattern Recogn. Lett.* **27**, 882–891 (2006).
22. O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 802–817 (2006).
23. B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *Int. J. Comput. Vis.* **77**, 157–173 (2008).
24. M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: experimental data and computer model," *J. Vis.* **9**(12):10 (2009).



25. N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems (NIPS)* (2005), pp. 155–162.
26. A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosi, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image Vis. Comput.* **30**, 51–64 (2012).
27. E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.* **13**(4):11 (2013).
28. X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2007), pp. 1–8.
29. B. Schauerte and R. Stiefelwagen, "Predicting human gaze using quaternion dct image signature saliency and face detection," in *Proceedings of the IEEE Workshop on the Applications of Computer Vision (WACV)*, Breckenridge, Colorado, January 9–11, 2012 (IEEE, 2012).
30. L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: a Bayesian framework for saliency using natural statistics," *J. Vis.* **8**(7):32 (2008).
31. A. Alsam, P. Sharma, and A. Wrlsen, "Asymmetry as a measure of visual saliency," in *18th Scandinavian Conference on Image Analysis (SCIA)*, Espoo, Finland (2013).
32. A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island (2012).
33. A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study," *IEEE Trans. Image Process.* **22**, 55–69 (2013).

## **A.4 Validating the visual saliency model**

**Authors:** Ali Alsam and Puneet Sharma.

**Full title:** Validating the visual saliency model.

**Published in:** SCIA 2013, Lecture Notes in Computer Science (LNCS), Springer-Verlag Berlin Heidelberg.

# Validating the Visual Saliency Model

Ali Alsam and Puneet Sharma

Department of Informatics & e-Learning (AITeL),  
Sør-Trøndelag University College (HiST),  
Trondheim, Norway  
er.puneetsharma@gmail.com

**Abstract.** Bottom up attention models suggest that human eye movements can be predicted by means of algorithms that calculate the difference between a region and its surround at different image scales where it is suggested that the more different a region is from its surround the more salient it is and hence the more it will attract fixations. Recent studies have however demonstrated that a dummy classifier which assigns more weight to the center region of the image out performs the best saliency algorithm calling into doubt the validity of the saliency algorithms and their associated bottom up attention models. In this paper, we performed an experiment using linear discrimination analysis to try to separate between the values obtained from the saliency algorithm for regions that have been fixated and others that haven't. Our working hypothesis was that being able to separate the regions would constitute a proof as to the validity of the saliency model. Our results show that the saliency model performs well in predicting non-salient regions and highly salient regions but that it performs no better than a random classifier in the middle range of saliency.

**Keywords:** Saliency, fixations.

## 1 Introduction

A salient image region is defined as an image part that is clearly different from its surround [1]. This difference is measured in terms of a number of attributes, namely, contrast, brightness and orientation [2–6]. By measuring these attributes, visual saliency algorithms aim to predict the regions in an image that would attract our attention under free viewing conditions [4], i.e., when the observer is viewing an image without a specific task such as searching for an object. Finally, the output of the visual saliency algorithms is a so called saliency map which is a two dimensional gray scale map where the brighter regions represent higher saliency.

To evaluate the performance of visual saliency algorithms, the two dimensional saliency maps are compared with the image regions that attract observers' attention [7–14]. This is done by displaying to the observers a set of images and using an eye tracker to record their eye fixations. Further, it is thought that a higher number of fixations correspond to salient image regions. The recorded



fixations are thus compared with the associated visual saliency maps in a pair wise manner [8, 15–17]. Unfortunately, most studies have shown that while the saliency algorithms do predict a certain percentage of fixations they are far from being able to fully account for observers’ visual attention [18, 19]. In fact, in a recent comprehensive eye tracking study by Judd et al. [20], it was shown that a dummy classifier defined by a Gaussian blob at the center of the image was better at predicting the eye fixations than any of the visual saliency models [1, 21, 22]. In other words, assuming that the eye fixations fall at the center of the image results in better prediction than an analysis of the image content. This finding is surprising and raises the question of whether our attention is indeed guided by salient image features.

In this paper we set about validating the saliency algorithm by means of an experiment in which we divided 200 images into regions which have received fixations and others that didn’t. By collecting the values returned by the saliency algorithm local to those regions into two matrices we were able to use discrimination analysis to determine whether the data of the two matrices is separable. Our working hypothesis was that being able to separate the data using linear discrimination analysis would constitute a proof that the saliency algorithm is indeed in good correspondence with the eye fixations while failing to separate the data would constitute a proof that the saliency algorithm is a poor predictor of eye fixations.

In our experiment we found that the saliency algorithm predicts eye fixations almost perfectly in regions that don’t attract any fixations and also in regions that attract many fixations. It is, however, a poor estimator of fixations in regions with middle saliency where the algorithm performs as a random classifier.

## 2 Brief Description of the Saliency Algorithm

Input to the algorithm is provided in the form of static color images. Three early features: color, intensity, and orientation are calculated from the input image. From these features several spatial scales are created using dyadic Gaussian pyramids [1]. Salient features are detected by using center-surround differences which are grounded in vision studies. The center-surround differences are calculated between the fine and the coarse scales followed by normalization. For details see [1]. Finally the resultant feature maps are combined linearly to form a so-called saliency map.

## 3 Experiments and Results

### 3.1 Data Set

The images and the associated fixations data used in the analysis were obtained from the comprehensive study by Judd et al. [20]. The data-set [20] includes 1003 images which were shown to 15 different observers with normal vision under free viewing conditions, i.e., the observers viewed the images without a specific task such as searching for an object.

**Validating the Saliency Theory.** In this experiment we set about validating the claim that the eye fixates on regions in the image that are salient or different with respect to their surround. To achieve an objective validation we chose to divide each image into two different sets of regions, in the first we have image regions which have attracted observers fixations and in the second set we have image regions that didn't attract fixations. The data was based on a subset of the images and corresponding fixations obtained by Judd et al. [20] where we used 200 landscape images and all the fifteen observers. The images were 1024 by 768 pixels in dimension and a fixated area was defined as a square region of dimensions 100 by 100 pixels where the center was located at the fixation point. Non-fixated areas were chosen randomly from parts of the image that had a region of a 100 by 100 pixels without any fixations. As an example, the fixated and the non-fixated regions for an image and the corresponding feature maps obtained by the saliency algorithm [1] are shown in figure 1.

By dividing the image into square regions that are classed as either fixated or not fixated we were able to assign a value to each square part that corresponds to the average of the intensity of the corresponding pixels in the saliency map obtained by Itti et al. [1]. In so doing we obtained two matrices,  $F$  and  $N_f$  where the elements in the vectors of  $F$  were the values of the averages of the feature maps based on the square regions centered at the fixation points while the vectors of  $N_f$  were the average values for non-fixated areas. Further we chose the number of non-fixated areas to be equal to that of the fixated regions, thus, the size of  $F$  was  $n \times k$  where  $n$  was the number of fixations in all the 200 images and  $k$  was the number of feature maps was defined by the algorithm to be three maps pertaining to intensity, color and orientation.

Our main objective with the creation of the matrices  $F$  and  $N_f$  was to determine whether we can separate between the data of the two matrices using discrimination analysis or not. The basic idea was that being able to separate the data would constitute a proof that the fixations are indeed driven by low level features such as contrast and lightness as is the claim by researchers supporting the bottom up attention model. We further believe that the level of separation achieved between the fixated and non-fixated regions would offer us a clear view as to the goodness of the saliency algorithm in predicting the fixations. Thus if the prediction is random we can conclude, based on the available data set, that the idea that salient regions attract attention is false while a perfect separation would indicate that salient image regions dictate our visual attention.

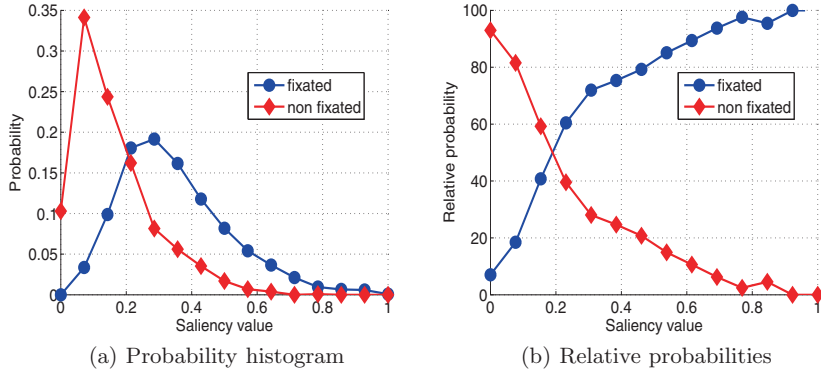
We chose a simple discrimination method which involves calculating the difference vector between the averages of  $F$  and  $N_f$  and then projecting the vectors of  $F$  and  $N_f$  onto the difference vector to judge whether the data is separated along that vector or in other words whether  $F$  and  $N_f$  are significantly different. Mathematically, the operation are:

$$w = \mu_F - \mu_{N_f}, \quad (1)$$

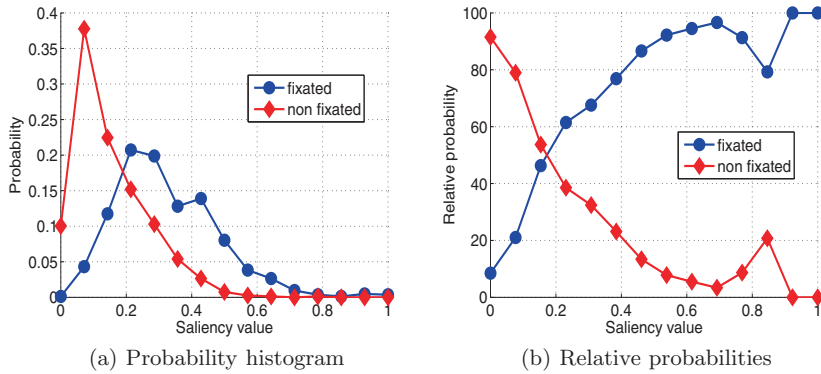
where the size  $w$  corresponds to the number of feature maps (3 for the saliency algorithm), and  $\mu_F$  and  $\mu_{N_f}$  are the means along the columns of  $F$  and  $N_f$ .



**Fig. 1.** Figure shows the fixated and the non-fixated regions for an image and the corresponding feature maps obtained by the saliency algorithm [1]. The fixated regions are marked as blue and the non-fixated regions are marked as red



**Fig. 2.** Probability histograms and relative probabilities for the fixated and non-fixated regions for observer no. 1. X-axis shows the saliency values obtained by using the visual saliency algorithm [1].

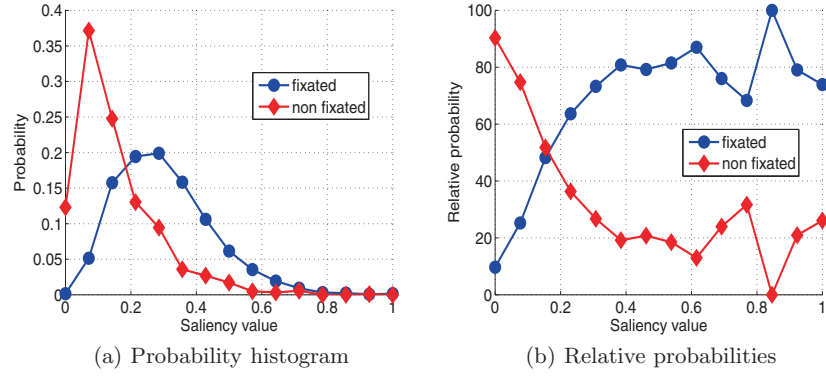


**Fig. 3.** Probability histograms and relative probabilities for the fixated and non-fixated regions for observer no. 2. X-axis shows the saliency values obtained by using the visual saliency algorithm [1].

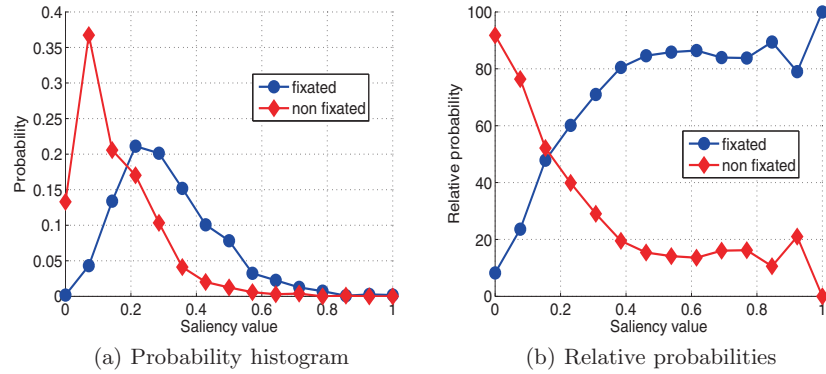
$$p_F = wF \quad (2)$$

$$p_{N_f} = wN_f, \quad (3)$$

where the number of elements of the vectors  $p_F$  and  $p_{N_f}$  are 1 by  $k$ . The distribution of  $p_F$  and  $p_{N_f}$  provides a mathematical description of whether the fixated and non-fixated regions are indeed different as predicted by the saliency algorithm.

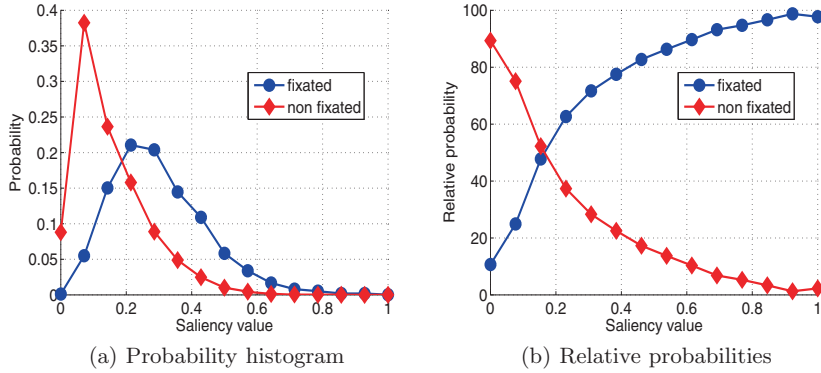


**Fig. 4.** Probability histograms and relative probabilities for the fixated and non-fixated regions for observer no 3. X-axis shows the saliency values obtained by using the visual saliency algorithm [1].



**Fig. 5.** Probability histograms and relative probabilities for the fixated and non-fixated regions for observer no 4. X-axis shows the saliency values obtained by using the visual saliency algorithm [1].

In figure 2, we plotted the probability histograms of  $p_F$  and  $p_{N_f}$ . Here, the histogram was normalized such that the area under the curve is one. We note that the separation between the two sets is not ideal but rather we find a considerable overlap between the two histograms specifically in the middle range. We further note that there is a clear separation between the two sets for regions of the images that received no fixations indicating that the method is good at predicting non-salient regions of the images. At a value of 0.3 the classification of the two sets is random.



**Fig. 6.** Probability histograms and relative probabilities for the fixated and non-fixated regions for an average observer. X-axis shows the saliency values obtained by using the visual saliency algorithm [1].

To gain better insight into the ability of the algorithm to separate the image regions into fixated and non-fixated, we plotted the relative probabilities of the histograms. For the non-fixated histogram, the relative probabilities were obtained by dividing the area under the non-fixated probability histogram curve of a specific bin  $i$  by the area under the fixated histogram curve for the same bin. For the relative probability of the fixated histogram the reciprocal value was calculated. Based on the fixation data of observer number one, this curve is plotted in figure 2 where we observe that for low saliency values the separation of non-fixated regions is ideal and that the goodness of the separation declines to a level that is random. We also note that the separation of the highly salient regions, is nearly ideal. Based on this we can conclude that the algorithm is good in predicting non-salient and highly salient regions but its performance drops in the middle range. Assuming that the algorithm is a good representation of the way in which the human vision system functions we can state that flat regions which are almost never fixated while middle range contrast attracts fixations though not in every part and regions with very high saliency almost always attract fixations. This interpretation is of course dependent on the total number of fixations and the spatial distribution of the salient regions.

To generalize the analysis for the other observers, we performed the same calculations for all the observers and found similar trends in all cases. The results for observers two, three, and four shown in figures 3, and 5 respectively; and similar results were obtained for the fifteen individual observers. The results for the average observer based on all fifteen observers are shown in figure 6.

## 4 Discussion

In this paper, we performed a study to validate the claim that human eye fixations correspond to salient image features. We divided the image into regions

which attracted fixations and others that were deemed by the observers as non-salient. By grouping the associated values for the feature maps obtained from the saliency algorithm by Itti et al. [1] into two matrices one pertaining to the fixated regions and another to the non-fixated areas we were able to use linear discrimination to separate the regions optimally. Our working hypothesis was that being able to distinguish between the local values of the feature maps at fixated and non-fixated regions would indicate that the algorithm is indeed useful in predicting eye fixations. Our findings indicate that saliency algorithm by Itti et al. [1] is nearly ideal at predicting non-salient and highly salient regions with a considerable confusion in the mid saliency region.

## References

1. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 1254–1259 (1998)
2. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology* 4, 219–227 (1985)
3. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* 40, 1489–1506 (2000)
4. Itti, L., Koch, C.: Computational modelling of visual attention. *Nature Reviews Neuroscience* 2, 194–203 (2001)
5. Walther, D., Koch, C.: Modeling attention to salient proto-objects. *Neural Networks* 19, 1395–1407 (2006)
6. Underwood, G., Humphreys, L., Cross, E.: Congruency, Saliency and Gist in the inspection of objects in natural scenes. In: *Eye Movements: A Window on Mind and Brain*, pp. 563–579. Elsevier (2007)
7. Walther, D.: Interactions of Visual Attention and Object Recognition: Computational Modeling, Algorithms, and Psychophysics. PhD thesis, California Institute of Technology, Pasadena, California (2006)
8. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: *Proceedings of Neural Information Processing Systems (NIPS)* (2006)
9. Cerf, M., Harel, J., Einhauser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: *Advances in Neural Information Processing Systems (NIPS)*, vol. 20, pp. 241–248 (2007)
10. Henderson, J.M., Brockmole, J.R., Castelano, M.S., Mack, M.: Visual Saliency Does Not Account for Eye Movements during Visual Search in Real-World Scenes. In: *Eye Movements: A Window on Mind and Brain*, pp. 537–562. Elsevier (2007)
11. Rajashekar, U., van der Linde, I., Bovik, A.C., Cormack, L.K.: Gaffe: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing* 17, 564–573 (2008)
12. Meur, O.L., Callet, P.L., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 802–817 (2006)
13. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 185–207 (2013)
14. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing* 22, 55–69 (2013)

15. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. *Vision Research* 42, 107–123 (2002)
16. Oliva, A., Torralba, A., Castelano, M.S., Henderson, J.M.: Top-down control of visual attention in object detection. In: *Proceedings of the 2003 International Conference on Image Processing, ICIP 2003*, vol. 1, pp. 253–256 (2003)
17. Henderson, J.M.: Human gaze control during real-world scene perception. *Trends in Cognitive Sciences* 7, 498–504 (2003)
18. Tatler, B.W., Baddeley, R.J., Gilchrist, I.D.: Visual correlates of fixation selection: effects of scale and time. *Vision Research* 45, 643–659 (2005)
19. Tatler, B.W.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision* 7, 1–17 (2007)
20. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: *International Conference on Computer Vision (ICCV)* (2009)
21. Rosenholtz, R.: A simple saliency model predicts a number of motion popout phenomena. *Vision Research* 39, 3157–3163 (1999)
22. Cerf, M., Frady, E.P., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision* 9, 1–15 (2009)



## **A.5 Asymmetry as a measure of visual saliency**

**Authors:** Ali Alsam, Puneet Sharma, and Anette Wråsen.

**Full title:** Asymmetry as a measure of visual saliency.

**Published in:** SCIA 2013, Lecture Notes in Computer Science (LNCS), Springer-Verlag Berlin Heidelberg.

# Asymmetry as a Measure of Visual Saliency

Ali Alsam, Puneet Sharma, and Anette Wrålsen

Department of Informatics & e-Learning (AITeL),  
Sør-Trøndelag University College (HiST),  
Trondheim, Norway  
er.puneetsharma@gmail.com

**Abstract.** A salient feature is a part of the scene that stands out relative to neighboring items. By that we mean that a human observer would experience a salient feature as being more prominent. It is, however, important to quantify saliency in terms of a mathematical quantity that lends itself to measurements. Different metrics have been shown to correlate with human fixations data. These include contrast, brightness and orienting gradients calculated at different image scales.

In this paper, we show that these metrics can be grouped under transformations pertaining to the dihedral group  $D_4$ , which is the symmetry group of the square image grid. Our results show that salient features can be defined as the image features that are most asymmetric in their surrounds.

**Keywords:** Saliency, dihedral group  $D_4$ , asymmetry.

## 1 Introduction

We are frequently surprised by the difference between what we observe in our visual world and the observations of others around us. Commonly, we think of these differences as a product of our varying personalities or interests, i.e., we notice what we think of or like. The fact that we observe different visual realities can, however, be explained in a different manner—we are selective because our brains are limited. In other words, we are selective because our brains do not process all the visual information that surrounds us. In this view, which is supported by psychophysical experiments [1–4], visual selection, or attention, is an information reduction method.

Mathematically, information reduction methods start with a process of identifying the most important aspects of the data, i.e., the parts of the data that cannot be disregarded. As an example both factor analysis and principal component analysis are based on the idea that multi-dimensional data can be represented with a set of limited bases that account for them with limited information loss [5, 6]. Based on this mathematical analogy we might wonder how the reduction of visual information is achieved.

In the literature, two main methods have been proposed: Top-down, also known as attention, and bottom-up or pre-attention visual information reduction

[1, 7–13]. As an example of top-down we might consider the problem of locating an item such as the red book on the bookshelf. Here our visual system would be trying to quickly browse the scene, disregarding any other color. As such, top-down visual reduction is task-driven and voluntary, where we would be looking for an aspect in the scene that matches a mental representation. Bottom-up methods on the other hand are involuntary, faster than top-down and not task-driven. Instead they are driven by the identification of a new, unknown, visual experience. The question that arises in bottom-up approaches is: How do we reduce the visual data of an arbitrary scene?

Most of the bottom-up, pre-attention models share the same basic elements. The basic assumption is that the different regions of the visual information field differ in their visual content. Based on that, an area of the scene that is clearly different from its surround, salient, is thought to represent an anchor point for data reduction. In other words, the visual reduction task is similar to statistical methods such as principal component analysis, where the most salient features of the scene represent the set of bases around which the rest of the scene is arranged. To measure the difference between a center and its surround, a number of stimulus characteristics have been proposed. These include color difference, contrast and orientation. For a given scene, these differences are measured and the results stored in so-called feature maps which are then combined in a so-called saliency map.

While salient feature detection algorithms are progressively more efficient at predicting where a person might look under free viewing conditions, the actual definition of a salient feature and thereby the mechanism of selecting such regions is still debatable. Generally, a salient feature is defined as a region in the scene that is different from its surround. The nature of this difference is, however, loosely defined. As previously mentioned, the difference is measured in terms of a number of metrics pertaining to contrast and gradients or orientation at different spatial scales commonly implemented by means of image pyramid decomposition.

The question addressed in this paper is mathematical, namely, we ask if the differences used in estimating the level of saliency at a given scene location can be grouped in a unified mathematical definition. By examining the metrics used to construct the feature maps, we observe that all can be accounted for by transformations described by the dihedral group  $D_4$ . This is the symmetry group of the square image grid and includes two types of symmetries, i.e., rotation and reflection. The transformations defined by  $D_4$  have exhibited immense power in image processing operations including image compression, denoising, and indexing [14–18].

To test the usefulness of the dihedral group in describing salient image features, we constructed a saliency map based on seven elements of  $D_4$ , namely, rotation by 90, 180 and 270 degrees and reflection about the horizontal, vertical and two diagonal axes. These transformations were performed on the blocks obtained by decomposing the image into square regions. The results at the higher and lower scales of image were calculated and stored in separate feature maps (details in the theory section). Finally, the feature maps were grouped into a

saliency map in linear manner, i.e., without the use of center surround operations. Having done that, we evaluated the correspondence between the proposed saliency map and human fixations data. Our results show that a saliency map derived based on the transformations of the dihedral group  $D_4$  matches well with human fixation data, and has very high correspondence with the existing saliency map.

Based on these results and the knowledge that the  $D_4$  transformations represent a mathematical measure of symmetry, we conclude with the hypothesis that a salient image feature is a part of the scene that is highly asymmetric compared to its surround and the more asymmetric a feature is the more salient it is. This hypothesis is strengthened by the knowledge that the transformations of  $D_4$  are extremely fast. This latter aspect of the operations is in agreement with the knowledge that bottom-up operations are fast, in the order of 25 to 50 ms [10].

The rest of this article is organized as follows: In Section 2, we discuss the theory behind the dihedral group  $D_4$  and the implementation of the proposed algorithm, in detail. In section, we examine the results obtained from the evaluation of saliency algorithms.

## 2 Theory

### 2.1 Mathematical Background

Mathematically, the symmetries of geometric objects can be defined by group theory, and in particular the symmetries of the square are encoded in the dihedral group  $D_4$ . In this section we briefly define and describe this group and then show how it can be applied to detect asymmetry in an image.

**The Group  $D_4$ .** A *group* is a set  $G$  together with a binary operation  $*$  on its elements. This operation  $*$  must behave in a very specific way:

- i)  $G$  must be *closed* under  $*$ , that is, for every pair of elements  $g_1, g_2$  in  $G$  we must have that  $g_1 * g_2$  is again an element in  $G$ .
- ii) The operation  $*$  must be *associative*, that is, for all elements  $g_1, g_2, g_3$  in  $G$  we must have that

$$g_1 * (g_2 * g_3) = (g_1 * g_2) * g_3.$$

- iii) There is an element  $e$  in  $G$ , called the *identity element*, such that for all  $g \in G$  we have that

$$e * g = g = g * e.$$

- iv) For every element  $g$  in  $G$  there is an element  $g^{-1}$  in  $G$ , called *the inverse of  $g$* , such that

$$g * g^{-1} = e = g^{-1} * g.$$

Groups appear in many places in mathematics. For instance, the integers form a group with the operation  $+$ , and the real numbers become a group under multiplication. We see that a group has just enough structure that every equation of the form  $g_1 * x = g_2$ , where  $g_1$  and  $g_2$  are elements of  $G$ , has a unique solution  $x = g_2 * g_1^{-1}$  in  $G$ . For a good introduction to group theory, see [19].

In this paper we are interested in  $D_4$ , the symmetry group of the square. This group has eight elements, four rotational symmetries and four reflection symmetries. The rotations are  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$  and  $270^\circ$ , and the reflections are defined along the four axes shown in Figure 1. We refer to these elements as  $\sigma_0, \sigma_1, \dots, \sigma_7$ . Note that the identity element is rotation by  $0^\circ$ , and that for each element there is another element that has the opposite effect on the square, as required in the definition of a group. The group operation is composition of two such transformations. As an example of one of the group elements, consider Figure 2, where we demonstrate rotation by  $90^\circ$  counterclockwise on a square with labeled corners.

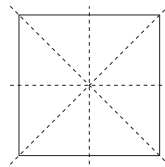


Fig. 1. The four axes of reflection symmetries of the square

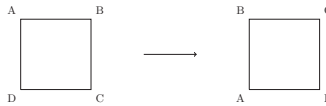


Fig. 2. Rotation of the square by  $90^\circ$  counterclockwise

**The Average Asymmetry Matrix.** The elements of  $D_4$  can be viewed as transformations that act on a square. Such an action on a set which respects the group operation is called a *group action* on the set. We will not define this formally here, just note that this means that we can define the action of  $D_4$  on the entries of a real square matrix in a natural way by letting the group elements rotate or reflect the entries according to the corresponding group elements. We will denote such an action by  $\sigma_i M$ , where  $\sigma_i$  is the element of  $D_4$  acting on a square matrix  $M$ .

Let  $M$  be an  $n \times n$ -matrix and  $\sigma_i$  some element of  $D_4$ . We define the *asymmetry of  $M$  by  $\sigma_i$* , denoted by  $A_i(M)$ , to be the matrix

$$A_i(M) = |M - \sigma_i M|. \tag{1}$$

We note that if  $M$  has a strong degree of the symmetry preserved by  $\sigma_i$ , the entries of this matrix will be close to zero.

Now we are ready to define the *average asymmetry* of  $M$ , denoted by  $A(M)$ . Let  $M$  be an  $n \times n$  matrix. Then we define the average asymmetry matrix  $A$  of  $M$ , denoted  $A(M)$ , as the matrix

$$A(M) = \frac{1}{8} \sum_{i=0}^7 A_i(M). \quad (2)$$

The more symmetries a matrix has, the smaller the entries of  $A(M)$  will be, and in this way we can say that  $A(M)$  provides a way to measure the degree of asymmetry of  $M$ .

## 2.2 Proposed Group Based Asymmetry Algorithm

In this section, we outline the implementation of the proposed group based asymmetry algorithm. From the color image, we calculate three channels, i.e., luminance channel, red-green and blue-yellow color opponency channels as described by Walther and Koch [20]. In order to calculate a feature map, we decompose the channel into square blocks. In the general case when the image dimensions are not perfectly divisible by the selected block size we pad the image borders with neighboring pixels. For example, in our experiments we used a block size of 20 by 20 pixels for an image of size 1024 by 768 pixels, thus after padding the image size becomes 1040 by 780 pixels. For each block, we calculate the absolute difference between the block itself and the result of the D4 group element acting on the block. We take the mean of the absolute difference for each block, which is taken as a measure of asymmetry for the block and has a scalar value in the range [0,1]. The asymmetry values for all the blocks are then collected in an image matrix and scaled up to the size of original image using bilinear-interpolation. In the resultant feature map the saliency of a location is represented by its scalar value, where a greater value represents a higher saliency. From the the D4 group elements i.e., rotations by 90, 180 and 270 degrees, and reflections along the four axes of a square, we get seven feature maps. In order to capture both the local and the global salient details in a channel, we use three scales: the original, 1/2 and 1/4. This gives three scales which combined with the seven D4 group elements give 21 feature maps, i.e., from the three channels we get a total of 63 feature maps which are combined linearly to get a single saliency map.

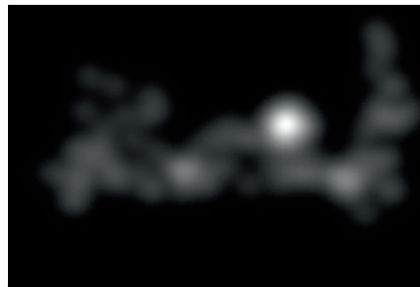
## 2.3 Analysis Using ROC

**Approach.** In this section, we discuss the approach taken for evaluating the performance of the visual saliency models. In keeping with published methods [21–23], we average all the fixations from different observers pertaining to a given image into a single two dimensional map, which is then convolved with a

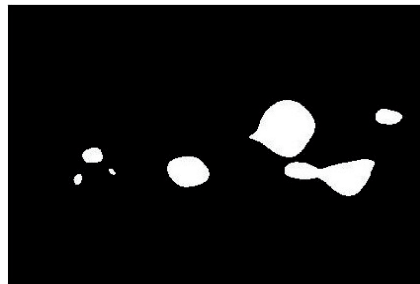
two dimensional Gaussian filter. In the resultant fixations map, the intensity at a given location represents the density of fixations [24], where the more fixations a region receives the more salient its said to be. For example, figure 3(b) shows the fixations map for an image. Similar to the previous experiment, we calculated the fixations maps from the fixations data of 200 images and 15 observers.



(a) Image from database [26].



(b) Fixations map.



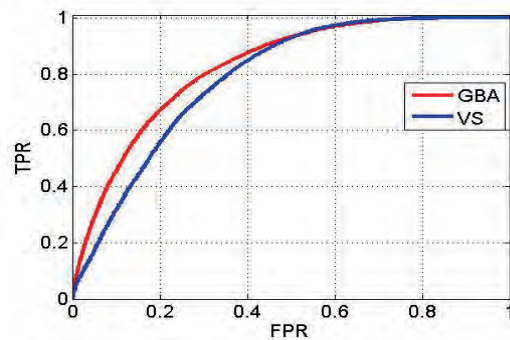
(c) Binary map.

**Fig. 3.** From the fixations map and the image, we can see that the region containing the road sign received a significant number of fixations. Figure 3(c) shows the binary map obtained by thresholding the fixations map by 20 percent.

In order to evaluate, how well the visual saliency models predict the fixations maps for different images, we use a receiver operating characteristic(ROC) curve [25] which requires that a fixations map is thresholded to yield a binary map with two classes – the positive class consisting of fixated regions, and the negative class consisting of non-fixated regions. As an example, figure 3 shows the binary map obtained by thresholding the fixations map by 20 percent. This procedure is in keeping with the study by Judd et al. [26]. The ROC curve evaluates how well the visual saliency algorithm predicts the two classes [25]. For plotting the ROC curve we randomly select 500 pixels from the positive class and an equal number of pixels from the negative class. The area under the ROC curve (AUC) is used as a measure of the performance of a classifier. AUC gives a scalar value in the interval  $[0,1]$  where larger the area, better is the performance [25].

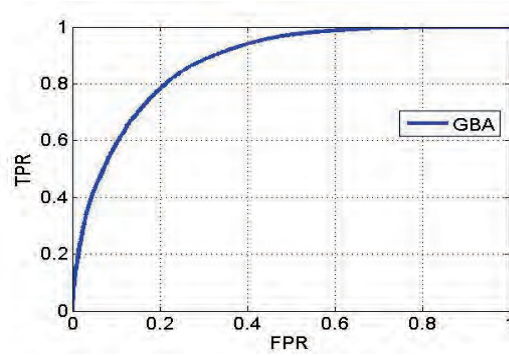
**Results.** We plot the ROC curves for the visual saliency algorithm proposed by Itti et al. [3], and the proposed group based asymmetry algorithm(GBA). Results in figure 4 show that the GBA algorithm results in an AUC value of 0.81 which is better than that achieved with the visual saliency algorithm by Itti et al. [3] which gives AUC of 0.77.

In order to measure the similarity between the proposed group based asymmetry algorithm and the visual saliency algorithm by Itti et al. [3] we calculated another ROC curve. In this case, we use the saliency maps from the visual saliency algorithm [3] as the ground truth maps. By following the procedure described in section 2.3, we evaluated how well the maps obtained from the GBA algorithm predict the maps obtained from the visual saliency algorithm [3]. Figure 5 shows the ROC curve for the proposed GBA algorithm which gives an AUC of 0.88 indicating that the prediction of the saliency values obtained by the proposed algorithm is indeed close to that of the visual saliency model.



**Fig. 4.** Figure shows the ROC curves for the visual saliency(VS) model by Itti et al. [3](AUC = 0.77), and the proposed group based asymmetry (GBA) model (AUC = 0.81). The x-axis shows the false positive rate(FPR) and the y-axis shows the true positive rate(TPR).

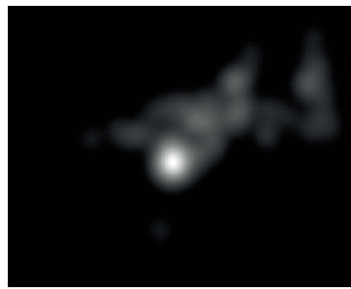




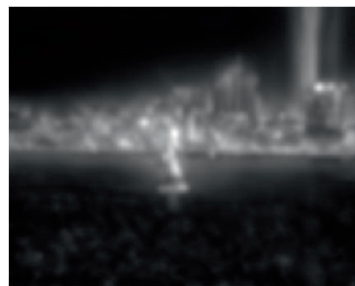
**Fig. 5.** Figure shows the ROC curves for the proposed group based asymmetry (GBA) model,  $AUC = 0.88$ . The x-axis shows the false positive rate(FPR) and the y-axis shows the true positive rate(TPR). Here we use the maps from the visual saliency algorithm [3] as the ground truth.



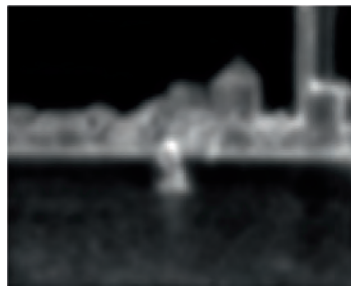
(a) Image from database [26]



(b) Fixations Map



(c) Saliency Map [3]



(d) Group based Asymmetry Map(GBA)

**Fig. 6.** Comparison of visual saliency algorithms, both algorithms return the region containing the boat at the center as salient, which is also in agreement with the fixations map obtained from the eye fixations data

To offer a visual comparison between the two methods we show the fixations map, and the saliency maps obtained from the GBA algorithm and the visual saliency algorithm [3] for an example image. In figure 6, we can see that the maps from both the algorithms are quite similar. In fact both of them return the region containing the boat at the center as salient, which is also in agreement with the fixations map.

### 3 Discussion

In this study, we set about unifying the mathematical description of saliency in a single metric. Backed by the knowledge gained from research in image processing where it has been shown that the dihedral group  $D_4$  can be used to encode edges and contrast which are the main current descriptions of saliency we chose to devise an algorithm that represents the level of saliency in an image region by virtue of the transformations of  $D_4$ .  $D_4$  is the symmetry group of the square image grid and includes two types of symmetries, i.e., rotation and reflection.

In our implementation, we chose to describe the symmetry of an image region at three different scale, however, we didn't perform any center surround operations by taking the differences between the scales. In this view, what we have presented in this study is a new unified metric together with a new description of saliency where we define saliency as the combined level of asymmetry at different image scales.

In our experiment, we used a receiver operating characteristic(ROC) curve to compare the performance of the proposed method with that of Itti et al. [3]. Here we used 200 images and fifteen observers and found that the new method results in a predication of fixations that is better than that achieved with the saliency algorithm. We thus concluded that the transformations of the dihedral group  $D_4$  are a good metric to estimate salient image regions which if backed by further studies can represent a mathematically sound method to define a salient image region.

### References

1. Suder, K., Worgotter, F.: The control of low-level information flow in the visual system. *Reviews in the Neurosciences* 11, 127–146 (2000)
2. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology* 4, 219–227 (1985)
3. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 1254–1259 (1998)
4. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* 40, 1489–1506 (2000)
5. Gorsuch, R.L.: *Factor Analysis*. Lawrence Erlbaum Associates, LEA (1983)
6. Jolliffe, I.T.: *Principal component analysis*. Springer (2002)
7. Braun, J., Sagi, D.: Vision outside the focus of attention. *Perception and Psychophysics* 48, 45–58 (1990)

8. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. *Annual Reviews in the Neurosciences* 18, 193–222 (1995)
9. Steinman, S.B., Steinman, B.A.: Vision and attention. i: Current models of visual attention. *Optometry and Vision Science* 75, 146–155 (1998)
10. Itti, L., Koch, C.: Computational modelling of visual attention. *Nature Reviews Neuroscience* 2, 194–203 (2001)
11. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2049–2056 (2006)
12. Mozer, M.C., Sitton, M.: 9. In: Computational modeling of spatial attention, pp. 341–393. Psychology Press (1998)
13. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 185–207 (2013)
14. Lenz, R.: Using representations of the dihedral groups in the design of early vision filters. In: ICAASP, pp. 165–168 (1993)
15. Lenz, R.: Investigation of receptive fields using representations of the dihedral groups. *Journal of Visual Communication and Image Representation* 6, 209–227 (1995)
16. Foote, R., Mirchandani, G., Rockmore, D.N., Healy, D., Olson, T.: A wreath product group approach to signal and image processing. i. multiresolution analysis. *IEEE Transactions on Signal Processing* 48, 102–132 (2000)
17. Chang, W.Y.: Image processing with wreath products. Master’s thesis, Harvey Mudd College (2004)
18. Lenz, R., Bui, T.H., Takase, K.: A group theoretical toolbox for color image operators. In: IEEE International Conference on Image Processing, ICIP 2005, vol. 3, pp. 557–560 (2005)
19. Dummit, D.S., Foote, R.M.: *Abstract Algebra*. John Wiley & Sons (2004)
20. Walther, D., Koch, C.: Modeling attention to salient proto-objects. *Neural Networks* 19, 1395–1407 (2006)
21. Cerf, M., Harel, J., Einhauser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: *Advances in Neural Information Processing Systems (NIPS)*, vol. 20, pp. 241–248 (2007)
22. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: *Proceedings of Neural Information Processing Systems (NIPS)* (2006)
23. Cerf, M., Frady, E.P., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision* 9, 1–15 (2009)
24. Duchowski, A.T.: *Eye Tracking Methodology: Theory and Practice*. Springer, Heidelberg (2007)
25. Fawcett, T.: Roc graphs: Notes and practical considerations for researchers. *Pattern Recognition Letters* 27, 882–891 (2004)
26. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: *International Conference on Computer Vision (ICCV)* (2009)

## A.6 Calculating saliency using the dihedral group $D_4$

**Authors:** Ali Alsam, Puneet Sharma, and Anette Wråsen.

**Full title:** Calculating saliency using the dihedral group  $D_4$ .

**Accepted for publication in:** Journal of Imaging Science & Technology.

## Calculating saliency using the dihedral group $D_4$

Ali Alsam,<sup>1</sup> Puneet Sharma,<sup>1, a)</sup> and Anette Wrålsen<sup>1</sup>

*Department of Informatics & e-Learning (AITeL),  
Sør-Trøndelag University College (HiST),  
Trondheim, Norway*

(Dated: 10 February 2014)

A salient image region is a part of the scene that stands out relative to neighboring regions. By that we mean that a human observer would experience a salient region as being more prominent. It is, however, important to quantify saliency in terms of a mathematical quantity that lends itself to measurements. Different metrics have been shown to correlate with human fixation data. These include contrast, brightness and orienting gradients calculated at different image scales.

In this paper, we show that saliency can be measured by the transformations pertaining to the dihedral group  $D_4$ , which is the symmetry group of the square image grid. Our results show that salient features can be defined as the image features that are most asymmetric in their surrounds.

Keywords: Saliency, eye fixations, visual attention

### I. INTRODUCTION

From psychophysical experiments<sup>16,19,26</sup>, it is well established that our visual system is selective, i.e. we only see parts of the visual scene where the eyes fixate. In other words, we are selective because our brains do not process all the visual information that surrounds us but rather integrates a seamless sensation from a smaller set of regions. In this view, visual selection, or attention, is an information reduction method that builds a continuous image of the world based on fragments.

Mathematically, information reduction methods start with a process of identifying the most important aspects of the data, i.e., the parts of the data that cannot be disregarded. As an example both factor analysis and principal component analysis are based on the idea that multi-dimensional data can be represented with a set of limited bases that account for them with limited information loss<sup>12,17</sup>. Based on this mathematical analogy we might wonder how the reduction of visual information is achieved.

In the literature, two main methods have been proposed: Top-down, also known as attention, and bottom-up or pre-attention visual information reduction<sup>4,7,15,23–26</sup>. As an example of top-down we might consider the problem of locating an item such as the red book on the bookshelf. Here our visual system would be trying to quickly browse the scene, disregarding any other color. As such, top-down visual reduction is task-driven and voluntary<sup>26</sup>, where we would be looking for an aspect in the scene that matches a mental representation. Bottom-up methods on the other hand are involuntary, faster than top-down and not task-driven<sup>26</sup>. Instead they are driven by the identification of a new, unknown, visual experience. The question that arises in bottom-up approaches is: How do we reduce the visual data of an arbitrary scene?

Most of the bottom-up, pre-attention models share the same basic elements. The basic assumption is that the different regions of the visual information field differ in their visual content<sup>19</sup>. Based on that, an area of the scene that is clearly different from its surround, salient, is thought to represent an anchor point for data reduction. In other words, the visual reduction task is similar to statistical methods such as principal component analysis, where the most salient features of the scene represent the set of bases around which the rest of the scene is arranged. To measure the difference between a center and its surround, a number of stimulus characteristics have been proposed. These include color difference, contrast and orientation<sup>16</sup>. For a given scene, these differences are measured and the results stored in so-called feature maps which are then combined in a so-called saliency map<sup>16</sup>.

While salient feature detection algorithms are progressively more efficient at predicting where a person might look under free viewing conditions, the actual definition of a salient feature and thereby the mechanism of selecting such regions is still debatable. Generally, a salient feature is defined as a region in the scene that is different from its surround<sup>16</sup>. The nature of this difference is, however, loosely defined. As previously mentioned, the difference is measured in terms of a number of metrics pertaining to contrast and gradients or orientation at different spatial scales commonly implemented by means of image pyramid decomposition.

The question addressed in this paper is mathematical, namely, we ask if the differences used in estimating the level of saliency at a given scene location can be grouped in a unified mathematical definition. By examining the metrics used to construct the feature maps, we observe that this can be accounted for by transformations described by the dihedral group  $D_4$ . This is the symmetry group of the square image grid and includes two types of symmetries- rotation and reflection. The transformations based on the elements of  $D_4$  have exhibited immense power in image processing operations including image compression, denoising, and indexing<sup>6,10,20–22</sup>.

---

<sup>a)</sup>Electronic mail: er.puneetsharma@gmail.com

To test the usefulness of the dihedral group in describing salient image features, we constructed a saliency map based on the elements of  $D_4$ , namely, rotation by 90, 180 and 270 degrees (excluding rotation by 0 degrees) and reflection about the horizontal, vertical and two diagonal axes. These transformations were performed on the blocks obtained by decomposing the image into square regions. The results at different image resolutions were calculated and stored in separate feature maps (details in the theory section).

Finally, the feature maps were grouped into a saliency map in a linear manner without the use of center surround operations. We then evaluated the correspondence between the proposed saliency map and human fixation data. Our results show that a saliency map derived based on the transformations of the dihedral group  $D_4$  matches well with human fixation data, and has very high correspondence with the existing saliency maps. We further performed an experiment where we attempted to validate the claim that eye fixations correspond to salient image regions, here we divided images from a large database into regions which have received fixations and others that did not. We then collected the associated saliency values from the feature maps corresponding to the saliency algorithm and the proposed metric into two matrices and used linear discrimination analysis to find a dimension which separates the two sets optimally. Our findings indicate that both the proposed method and the saliency algorithm by Itti et al.<sup>16</sup> are efficient at predicting regions which human observers deem as non-salient and also regions which are highly salient, i.e., receiving many fixations. We, however, found that regions in the middle range of saliency are less predictable by the two algorithms. Furthermore, when compared with a large set of algorithms using a large database of images and associated fixations, we found that the proposed method ranked as the second best algorithm for predicting human fixations.

Based on these results and the knowledge that the  $D_4$  transformations represent a mathematical measure of symmetry, we conclude with the hypothesis that a salient image feature is a part of the scene that is highly asymmetric compared to its surround, and the more asymmetric a feature is the more salient it is. This hypothesis is strengthened by the knowledge that the transformations of  $D_4$  are extremely fast. The latter aspect of the transformations is in agreement with the knowledge that bottom-up operations are fast, in the order of 25 to 50 ms<sup>15</sup>.

The rest of this article is organized as follows: In Section II, we discuss the theory behind the dihedral group  $D_4$  and the implementation of the proposed algorithm in detail. In Section III, we examine the results obtained from the evaluation of saliency algorithms.

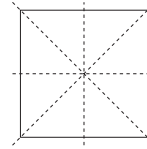


FIG. 1: The four axes of reflection symmetries of the square.

## II. THEORY

### A. Mathematical background

Mathematically, the symmetries of geometric objects can be defined by group theory, and in particular the symmetries of the square are encoded in the dihedral group  $D_4$ . In this section we briefly define and describe this group and then show how it can be applied to detect asymmetry in an image.

#### 1. The group $D_4$

One of the most basic structures studied in abstract algebra is *groups*. A group is a set  $G$  together with a binary operation  $*$  on its elements. This operation  $*$  must behave in a very specific way:

- i)  $G$  must be *closed* under  $*$ , that is, for every pair of elements  $g_1, g_2$  in  $G$  we must have that  $g_1 * g_2$  is again an element in  $G$ .
- ii) The operation  $*$  must be *associative*, that is, for all elements  $g_1, g_2, g_3$  in  $G$  we must have that

$$g_1 * (g_2 * g_3) = (g_1 * g_2) * g_3.$$

- iii) There is an element  $e$  in  $G$ , called the *identity element*, such that for all  $g \in G$  we have that

$$e * g = g = g * e.$$

- iv) For every element  $g$  in  $G$  there is an element  $g^{-1}$  in  $G$ , called *the inverse of  $g$* , such that

$$g * g^{-1} = e = g^{-1} * g.$$

Groups appear in many places in mathematics. For instance, the integers form a group with the operation  $+$ , and the real numbers become a group under multiplication. For a good introduction to group theory, see<sup>8</sup>.

One of the origins of group theory was the study of *symmetry* in various settings. In particular we can express the symmetries on a geometric object as a group.

FIG. 2: Rotation of the square by  $90^\circ$  counterclockwise.

If the geometric object is a regular polygon with  $n$  sides, i.e., all its sides are of equal length and all its angles are the same, its symmetry group is called *the dihedral group*  $D_n$ . In this paper we are interested in  $D_4$ , the symmetry group of the square. The ease of computational complexity associated with dividing an image grid into square regions, and the fact that the  $D_4$  group has shown promising results in various computer vision applications<sup>6,10,20–22</sup>, motivated us to use this group for our proposed algorithm.

The group  $D_4$  has eight elements, four rotational symmetries and four reflection symmetries. The rotations are  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$  and  $270^\circ$ , and the reflections are defined along the four axes shown in Figure 1. We refer to these elements as  $\sigma_0, \sigma_1, \dots, \sigma_7$ . Note that the identity element is rotation by  $0^\circ$ , and that for each element there is another element that has the opposite effect on the square, as required in the definition of a group. The group operation is composition of two such transformations. As an example of one of the group elements, consider Figure 2, where we demonstrate rotation by  $90^\circ$  counterclockwise on a square with labeled corners.

## 2. The average asymmetry matrix

The elements of  $D_4$  can be viewed as transformations that act on a square. Such an action on a set which respects the group operation is called a *group action* on the set. We will not define this formally here, just note that this means that we can define the action of  $D_4$  on the entries of a real square matrix in a natural way by letting the group elements rotate or reflect the entries according to the corresponding group elements. We will denote such an action by  $\sigma_i M$ , where  $\sigma_i$  is the element of  $D_4$  acting on a square matrix  $M$ .

Let  $M$  be an  $n \times n$ -matrix and  $\sigma_i$  some element of  $D_4$ . We define the *asymmetry of  $M$  by  $\sigma_i$* , denoted by  $A_i(M)$ , to be the matrix

$$A_i(M) = |M - \sigma_i M|. \quad (1)$$

We note that if  $M$  has a strong degree of the symmetry preserved by  $\sigma_i$ , the entries of this matrix will be close to zero.

Now we are ready to define the *average asymmetry of  $M$* , denoted by  $A(M)$ . Let  $M$  be an  $n \times n$  matrix. Then we define the average asymmetry matrix  $A$  of  $M$ , denoted

$A(M)$ , as the matrix

$$A(M) = \frac{1}{8} \sum_{i=0}^7 A_i(M). \quad (2)$$

The more symmetries a matrix has, the smaller the entries of  $A(M)$  will be, and in this way we can say that  $A(M)$  provides a way to measure the degree of asymmetry of  $M$ .

We demonstrate the results of calculating the average asymmetry matrix for a number of example images. Consider Figure 3. Note how the asymmetries of the images are detected in different ways, depending on the image. For instance, we note that  $M_4$ , which has a mixed edge results in the strongest (highest in intensity) average asymmetry matrix  $A(M_4)$ . The calculated intensities of  $M_1$  and  $M_3$  are ranked second in strength. Thus by calculating the asymmetry of the matrix we seem to be able to detect gradients in different orientations. We further note that the intensity encoded in the average asymmetry matrix is a function of the contrast between the image pixels. As an example, the intensity of  $A(M_1)$  would be reduced should the contrast between the top and bottom of the image be dampened. The only example image which gives zero average asymmetry is  $M_5$ —it is completely symmetric around its center.

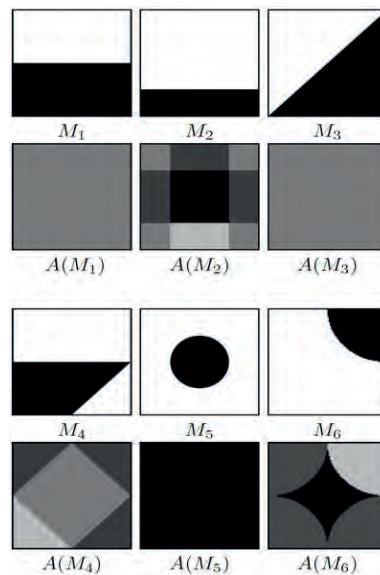


FIG. 3: Images together with the images representing their average asymmetries.

We note, however, that if we further divide  $M_5$  into four equally sized squares the asymmetry is detected. As an example, we consider  $M_6$  of Figure 3 which is a rep-



resentation of the third quadrant of  $M_6$ . Here, we note that the average asymmetry matrix  $A(M_6)$  is non-zero.

### B. Proposed group based asymmetry algorithm

In this section, we outline the implementation of the proposed group based asymmetry algorithm. From the color image, we calculate three channels i.e., luminance channel, red-green and blue-yellow color opponency channels as described by<sup>30</sup>. In order to calculate a feature map, we decompose the channel into square blocks. In the general case when the image dimensions are not perfectly divisible by the selected block size we padded the image borders with neighboring pixels. For example, in our experiments we used a block size of 24 by 24 pixels for an image of size 1024 by 768 pixels, thus after padding the image size become 1032 by 768 pixels. For each block, we calculate the absolute difference between the block itself and the result of the  $D_4$  group element acting on the block. We take the mean of the absolute difference for each block, which is taken as a measure of asymmetry for the block and has a scalar value in the range  $[0,1]$ . The asymmetry values for all the blocks are then collected in an image matrix and scaled up to the size of the original image using bilinear-interpolation. In the resultant feature map the saliency of a location is represented by its scalar value, where a greater value represents a higher saliency. From the  $D_4$  group elements i.e., rotations by 90, 180 and 270 degrees, and reflections along the four axes of a square, we get seven feature maps. In order to capture both the local and the global salient details in a channel, we use three scales: the original, 1/2 and 1/4. This gives three scales which, combined with the seven  $D_4$  group elements, give 21 feature maps, i.e., from the three channels we get a total of 63 feature maps which are combined linearly to get a single saliency map.

As an example, an image and the associated feature maps for its blue-yellow channel are shown in Figure 4. The columns 1 to 3 show the feature maps for the  $D_4$  group rotation elements i.e., 90, 180, and 270 degrees. The columns 4 to 7 show the feature maps for the reflection elements i.e., horizontal, vertical, and the two diagonals. Here, we can see the different gradients detected by the different group elements. The rows show the different scales of the image where the first is calculated based on the original image resolution, the second at 1/2 of the original and third at 1/4. As the saliency map is calculated based on both the local and global image gradients, therefore, a smaller block size would capture local gradients more, and vice versa. In principle, the saliency map can be calculated by combining the feature maps obtained by using either different block sizes, or, a fixed block size and different image scales. For our implementation, we used a fixed block size and three different image scales. The block size and the number of scales were determined based on an initial evaluation of

the algorithm on 50 test images by using the *AUC* metric discussed in Section III C 2.

## III. EXPERIMENTS AND RESULTS

### A. Data set

The images and the associated fixations data used in the analysis were obtained from the comprehensive study by Judd et al.<sup>18</sup>. The data-set<sup>18</sup> includes 1003 images which were shown to 15 different observers with normal vision under free viewing conditions, i.e., the observers viewed the images without a specific task such as searching for an object. The images were viewed by the observers for a period of 3 seconds each.

### B. Validating the saliency theory

In this experiment we set about validating the claim that the eye fixates on regions in the image that are salient or different with respect to their surround. To achieve an objective validation we chose to divide each image into two different sets of regions, in the first we have image regions which have attracted observers fixations and in the second set we have image regions that did not attract fixations. The data were based on a subset of the images and corresponding fixations obtained by<sup>18</sup> where we used 200 landscape images and all fifteen observers. The images were 1024 by 768 pixels in dimension and a fixated area was defined as a square region of dimensions 100 by 100 pixels where the center was located at the fixation point. Non fixated areas were chosen randomly from parts of the image that had a region of a 100 by 100 pixels without any fixations. As an example, the fixated and the non-fixated regions for an image and the corresponding feature map obtained by the proposed group algorithm are shown in Figure 4. The fixated regions for an observer are marked as blue and the non-fixated regions are marked as red. For each observer, the fixated and non-fixated regions were calculated separately. In order to simplify the calculations, we represent the fixated and non-fixated locations as square regions.

By dividing the image into square regions that are classed as either fixated or not fixated we were able to assign a value to each square part that corresponds to the average of the intensity of the corresponding pixels in the saliency map obtained by<sup>16</sup> and the proposed group algorithm. In so doing we obtained two matrices,  $F$  and  $N_f$  where the elements in the vectors of  $F$  were the values of the averages of the feature maps (from the saliency algorithm or the proposed group method) based on the square regions centered at the fixation points while the vectors of  $N_f$  were the average values for non-fixated areas. Further we chose the number of non-fixated areas to be equal to that of the fixated regions, thus, the size of



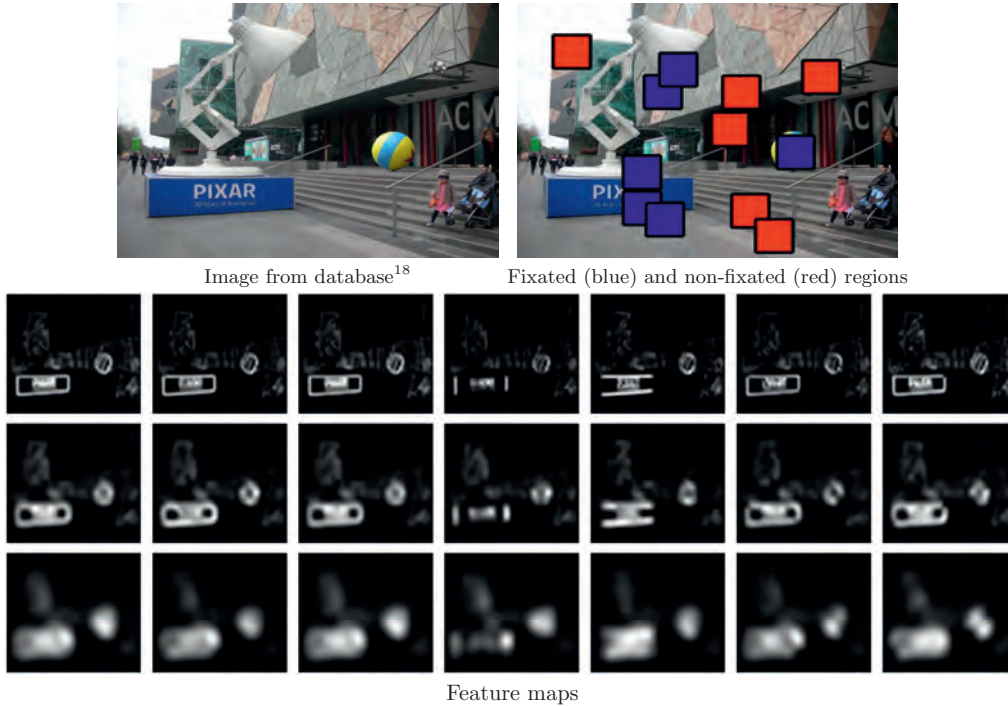


FIG. 4: An image and the associated feature maps for its blue-yellow channel are shown. The columns 1 to 3 show the feature maps for the  $D_4$  group rotation elements i.e., 90, 180, and 270 degrees. The columns 4 to 7 show the feature maps for the reflection elements i.e., horizontal, vertical, and the two diagonals. Here, we can see the different gradients detected by the different group elements. The rows show the different scales of the image where the first is calculated based on the original image resolution, the second at  $1/2$  of the original and third at  $1/4$ .

$F$  was  $n \times k$  where  $n$  was the number of fixations in all the 200 images and  $k$  was the number of feature maps which, in case of the saliency algorithm proposed by<sup>16</sup> was defined to be three maps pertaining to intensity, color and orientation, while in the group based method the number of channels was defined by three channels and three scales, this combined with seven group elements give 63 feature maps.

Our main objective with the creation of the matrices  $F$  and  $N_f$  was to determine whether or not we can separate the data of the two matrices using discrimination analysis. The basic idea was that being able to separate the data would constitute a proof that the fixations are indeed driven by low level features such as contrast and lightness as is the claim by researchers supporting the bottom up attention model. We further believe that the level of separation achieved between the fixated and non-fixated regions would offer us a clear view of the efficacy the saliency and the group based algorithms in predicting the fixations. Thus if the prediction is random we can conclude, based on the available data set, that the idea that salient regions attract attention is false while a per-

fect separation would indicate that salient image regions dictate our visual attention.

As a by product of the examination using linear discrimination, we can find the optimal weights, given the technique, to combine the feature maps into a unified saliency map. We chose a simple discrimination method which involves calculating the difference vector between the averages of  $F$  and  $N_f$  and then projecting the vectors of  $F$  and  $N_f$  onto the difference vector to judge whether the data is separated along that vector, or, in other words, whether  $F$  and  $N_f$  are significantly different. Mathematically, the operations are:

$$w = \mu_F - \mu_{N_f}, \quad (3)$$

where the size  $w$  corresponds to the number of feature maps (63 for the group method and 3 for the saliency algorithm), and  $\mu_F$  and  $\mu_{N_f}$  are the means along the columns of  $F$  and  $N_f$ .

$$p_F = wF \quad (4)$$

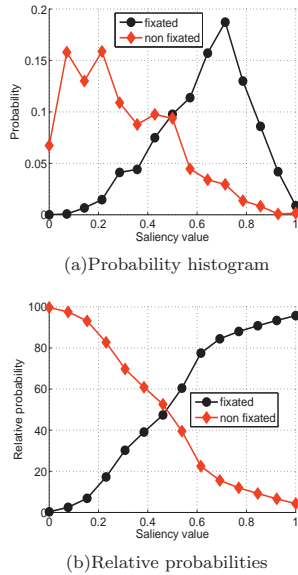


FIG. 5: Probability histograms and relative probabilities for the fixated and the non-fixated regions for observer no 1. X-axis shows the saliency values obtained by using the the proposed group algorithm.

$$p_{N_f} = wN_f, \quad (5)$$

where the number of elements of the vectors  $p_F$  and  $p_{N_F}$  are 1 by  $k$ . The distribution of  $p_F$  and  $p_{N_F}$  provides a mathematical description of whether the fixated and non-fixated regions are indeed different as predicted by the saliency and group based algorithms.

In figures 5 and 6, we plotted the probability histograms of  $p_F$  and  $p_{N_f}$  for the proposed group method and the saliency algorithm, respectively. The histograms were normalized such that the area under the curve is one. We note that the separation between the two sets is not ideal in either method where we find a considerable overlap between the two histograms specifically in the middle range. We further note that there is a clear separation between the two sets for regions of the images that received no fixations indicating that the methods are good at predicting non-salient regions of the images. At a value of 0.5 the classification of the two sets based on the proposed method is random. This value is lower i.e., 0.3, for the saliency algorithm.

To gain better insight into the performance of the two methods and their ability to separate the image regions into fixated and non-fixated, we plotted the relative probabilities of the histograms. For the non-fixated histogram, the relative probabilities were obtained by dividing the area under the non-fixated probability histogram curve of a specific bin  $i$  by the area under the fixated

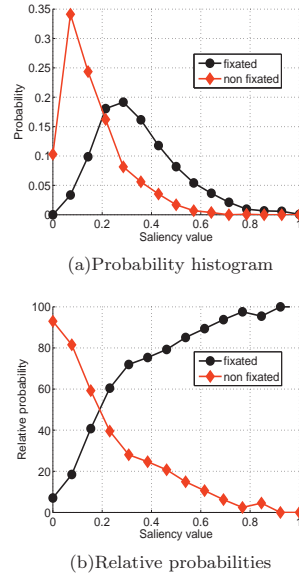


FIG. 6: Probability histograms and relative probabilities for the fixated and the non-fixated regions for observer no 1. X-axis shows the saliency values obtained by using the visual saliency algorithm<sup>16</sup>.

histogram curve for the same bin. For the relative probability of the fixated histogram the reciprocal value was calculated. Based on the fixation data of observer number one, these curves are plotted in the second column of figures 5 and 6 where we observe that for low saliency values from the proposed group algorithm, the separation of non-fixated regions is ideal and that the goodness of the separation declines to a level that is random in the middle range of the saliency value. We also note that the separation of the highly salient regions, i.e., high saliency value, is nearly ideal. Based on this we can conclude that both algorithms are good in predicting non-salient and highly salient regions and their performance drops in the middle range. Assuming that the algorithms are a good representation of the way in which the human vision system functions we can state that flat regions which are almost never fixated while middle range contrast attracts fixations though not in every part and regions with very high saliency almost always attract fixations. This interpretation is of course dependent on the total number of fixations and the spatial distribution of the salient regions. Finally, figures 5 and 6 indicate that the proposed group based asymmetry algorithm performs better than the well established saliency method.

To generalize the analysis for the other observers, we performed the same calculations for all the observers and found similar trends in all cases. The results for observers two, three and four and shown in figures 7, 8, 9, 10, 11,

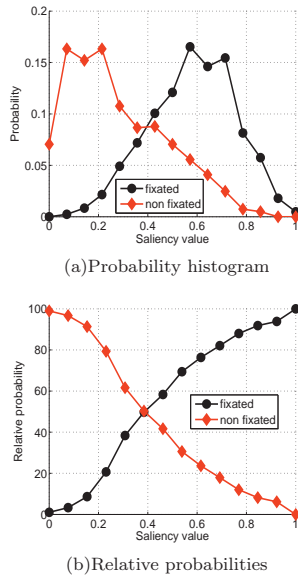


FIG. 7: Probability histograms and relative probabilities for the fixated and the non-fixated regions for observer no 2. X-axis shows the saliency values obtained by using the proposed group algorithm.

and 12 respectively; and similar results were obtained for the fifteen individual observers. The results for the average observer based on all fifteen observers are shown in figures 13, and 14.

To gain deeper insight into the correspondence between the value returned by the group transformations as a measure of saliency and the fixations we looked at the histogram of the group values for the 200 images. The results are shown in Figure 15. Here we note that around 10% of the values are in the region 0.65 to 1 where 1 is the highest value and most of those are predicted correctly by the algorithm. We also note that around 33% of the values are between 0 and 0.19 and those are also predicted well by the group algorithm. In the region between 0.19 and 0.65 we have nearly 57% of the image regions and that is the region where we have most overlap between the histograms of the fixated and non-fixated regions. We thus remark, that both the saliency algorithm and the proposed group method offer good prediction of the low saliency and high saliency regions while the prediction of the middle saliency region is more ambiguous. This in turn can be explained by the fact that the number of regions in the middle range is higher than that in the low and high ends of the histograms and that the number of fixations is always less than the number of regions in that range.

Finally, in figures 16-17 we plotted the elements of the discrimination vector  $w$  for both the proposed algorithm

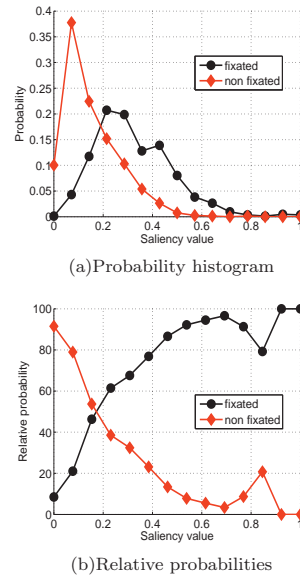


FIG. 8: Probability histograms and relative probabilities for the fixated and the non-fixated regions for observer no 2. X-axis shows the saliency values obtained by using the visual saliency algorithm<sup>16</sup>.

and the saliency maps described by<sup>16</sup>. Here we observe that the values are higher for the luminance channel and the third resolution indicating that most of the discrimination is based on the luminance rather than the opponent color channels and that the distinction between the fixated and non-fixated regions is determined to a large degree by the low resolution image, a finding which is in keeping with the results published by<sup>18</sup>. For the group algorithm we find that the transformations pertaining to rotation by 90 and 270 degrees are slightly but consistently more significant than those associated with the other transformations.

### C. Comparing the proposed group algorithm with state-of-the-art saliency algorithms

#### 1. Saliency models

The performance of the proposed **GBA** model is compared with seven state of the art saliency models, namely, **AIM** by Bruce & Tsotsos<sup>5</sup>, **AWS** by Garcia-Diaz et al.<sup>11</sup>, **SUN** by Zhang et al.<sup>31</sup>, **Hou** by Hou & Zhang<sup>14</sup>, **1**, **GBVS** by Harel et al.<sup>13</sup>, **Itti** by Itti et al.<sup>16</sup>, **Judd** by Judd et al.<sup>18</sup>. Figure 18 illustrates the saliency maps obtained from the different saliency models used in this paper. The saliency maps are normalized in the range [0,1]. In line with the study by Borji et al.<sup>3</sup>, we used

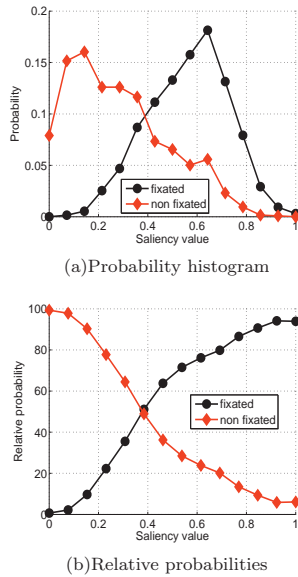


FIG. 9: Probability histograms and relative probabilities for the fixated and the non-fixated regions for observer no 3. X-axis shows the saliency values obtained by using the proposed group algorithm.

**Gauss** and **IO** to provide a baseline for the evaluation. **Gauss** is defined as a two-dimensional Gaussian blob at the center of the image. This model corresponds well with the fixations falling at the image center. **IO** model is based on the fact that an observer’s fixations can be predicted best by the fixations of other observers viewing the same image. In this model the map for an observer is calculated as follows: first, the fixations corresponding to a given image from all the observers except the one under consideration are averaged into a single two-dimensional map. Having done that the fixations are spread by smoothing the map using a Gaussian filter. The **IO** model gives us an upper bound on level of correspondence that is expected between the saliency models and the fixations.

## 2. Evaluation metrics

To evaluate the saliency models, we employed two metrics: an area under the receiver-operating-characteristic curve (*AUC*), and a shuffled *AUC* discussed by<sup>31</sup>.

### *AUC* metric

In order to calculate the  $AUC^{2,9}$ , the fixations pertaining to a given image are averaged into a single two dimensional map which is then convolved with a two-dimensional Gaussian filter. The resultant fixations map

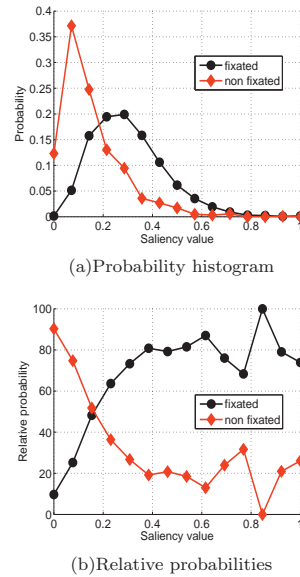


FIG. 10: Probability histograms and relative probabilities for the fixated and the non-fixated regions for observer no 3. X-axis shows the saliency values obtained by using the visual saliency algorithm<sup>16</sup>.

is then thresholded to yield a binary map with two classes—the positive class consisting of fixated locations, and the negative class consisting of non-fixated locations. Next, from the two dimensional saliency map, we obtain the saliency values associated with the positive and the negative classes. Using the saliency values, a receiver-operating-characteristic (*ROC*) curve is drawn that plots the true positive rate against the false positive rate. For a detailed description of *ROC*, see the study by<sup>9</sup>. The area under the *ROC* curve gives us a measure of the performance of the classifier. *AUC* gives a scalar value in the interval [0,1]. If *AUC* is 1 then it indicates that the saliency model is perfect in predicting fixations. An *AUC* of 0.5 implies that the performance of the saliency model is not better than a random classifier or by chance prediction.

### *Challenges associated with using the AUC metric*

While viewing images, observers tend to look at the center regions more as compared to peripheral regions. As a result of that a majority of fixations fall at the image center. This effect is known as center bias and is well documented in vision studies<sup>27,28</sup>. The two main reasons for this are: first, the tendency of photographers to place the objects at the center of the image. Second, the viewing strategy employed by observers, i.e., to look at center locations more in order to acquire the most information about a scene<sup>29</sup>. The presence of center bias in

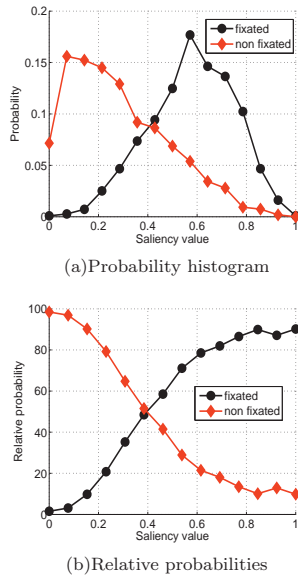


FIG. 11: Probability histograms and relative probabilities for the fixated and the non-fixated regions for observer no 4. X-axis shows the saliency values obtained by using the proposed group algorithm.

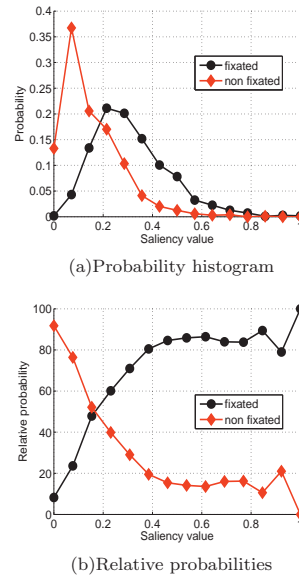


FIG. 12: Probability histograms and relative probabilities for the fixated and the non-fixated regions for observer no 4. X-axis shows the saliency values obtained by using the visual saliency algorithm<sup>16</sup>.

fixations makes it difficult to analyze the correspondence between the fixated regions and the salient image regions. In a study by<sup>18</sup>, it was observed that a dummy classifier consisting of a two-dimensional Gaussian shape drawn at the center of the image outperformed all saliency models. The center bias is implicitly linked with a so-called edge effect<sup>3</sup>. Edge effect<sup>31</sup> is defined as adding a varied image border of zeros to a saliency map, as a result of which it can yield different values from the evaluation metrics. For instance, in the study by<sup>31</sup>, it was observed that a dummy saliency map consisting of all ones with a four-pixel image border consisting of zeros gave an *AUC* value of 0.62. Meanwhile, an *AUC* of 0.73 was obtained with a dummy saliency map using eight-pixel border. In the presence of center bias and edge effect, a fair comparison of the performance of the saliency algorithms becomes a challenging task. To alleviate the influence of the center bias and the edge effect, a shuffled *AUC* metric was proposed in the study by<sup>31</sup>.

#### Shuffled *AUC* metric

To calculate the shuffled *AUC* metric for a given image and one observer, the locations fixated by the observer are associated with the positive class in a manner similar to the regular *AUC*, however, the locations for the negative class are selected randomly from the fixated locations of other unrelated images, such that they do not coincide with the locations from the positive class.

### 3. Analysis using the *AUC* metric

Figure 19 shows the ranking of the visual saliency models obtained by using the ordinary *AUC* metric. We observed that all saliency models used in this paper perform above chance. We also observed that **SUN**, **GBA**, **AWS**, **Hou**, **AIM**, and **Itti** perform worse than the **Gauss** model, with **GBVS**, and **Judd** being the two best models. This finding can be explained by the fact that the center regions are weighted more in both the **GBVS**, and **Judd** models.

### 4. Analysis using the shuffled *AUC* metric

The results obtained by employing the shuffled *AUC* metric are shown in Figure 20. We note that as compared to the ordinary *AUC*, this metric changes the ranking of the saliency models significantly. As an example, the **Gauss** classifier changes from being one of the best to being clearly the worst. Further, the **GBVS**, and **Judd** models drop significantly in the rankings. In fact in this case, **AIM**, **Hou**, **GBA**, and **AWS** models are the four best models. In-line with the study by Borji et al.<sup>3</sup>, our results show that the **AWS** model is the best among all. The results suggest that, first, the shuffled *AUC* metric is robust to the influence of the fixations associated with the center-bias compared to the ordinary *AUC* metric.

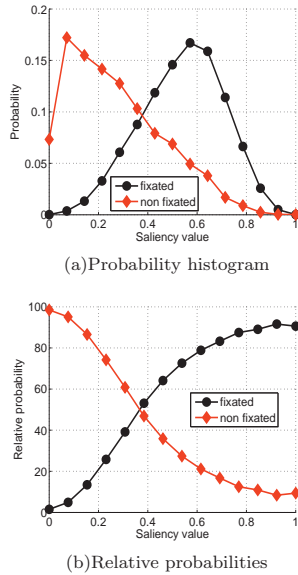


FIG. 13: Probability histograms and relative probabilities for the fixated and the non-fixated regions for an average observer. X-axis shows the saliency values obtained by using the proposed group algorithm.

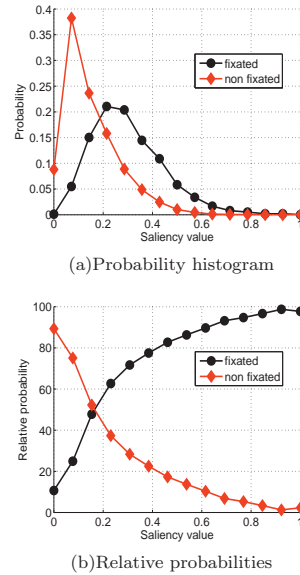


FIG. 14: Probability histograms and relative probabilities for the fixated and the non-fixated regions for an average observer. X-axis shows the saliency values obtained by using the visual saliency algorithm<sup>16</sup>.

Second, the proposed **GBA** model is among the four best saliency models.

#### IV. DISCUSSION

In this study, we set about unifying the mathematical description of saliency in a single metric. Backed by the knowledge gained from research in image processing where it has been shown that the dihedral group  $D_4$  can be used to encode edges and contrast which are the main current descriptions of saliency we chose to devise an algorithm that represents the level of saliency in an image region by virtue of the transformations induced by the dihedral group  $D_4$ . The dihedral group  $D_4$  is the symmetry group of the square image grid and includes two types of symmetries, i.e., rotation and reflection.

In our implementation, we chose to describe the symmetry of an image region at three different scales, however, we did not perform any center surround operations by taking the differences between the scales. In this view, what we have presented in this study is a new unified metric together with a new description of saliency where we define saliency as the combined level of asymmetry at different image scales.

To test the usefulness of the proposed method we constructed two experiments: in the first, we performed a study to validate the claim that human eye fixations cor-

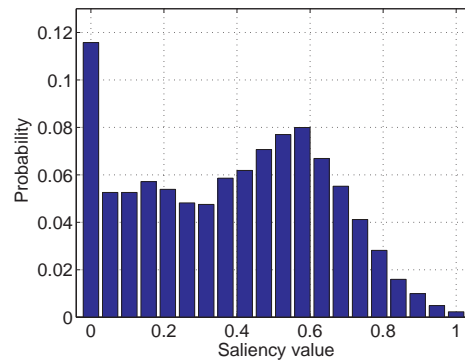


FIG. 15: Figure shows the probability histogram for the saliency values obtained from the proposed group algorithm for the 200 images. We note that nearly 10 percent of the saliency values lie between 0.65 and 1, approximately 33 percent of the saliency values lie in the range between 0 and 0.19, and nearly 57 percent of all the saliency coefficients lie in the range between 0.19 and 0.65.

respond to salient image features. In this study, we divided the image into regions which attracted fixations and others that were deemed by the observers as non-



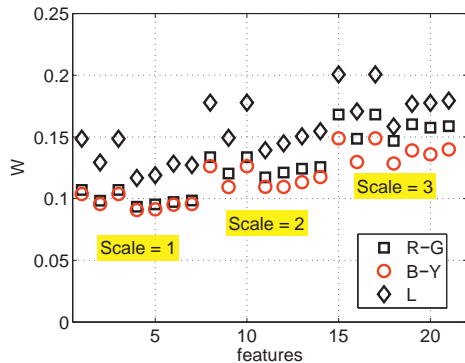


FIG. 16: The elements of the discrimination vector  $w$  for the 63 feature maps obtained from the proposed group algorithm. Here we observe that the values are higher for the luminance channel and the third resolution indicating that most of the discrimination is based on the luminance rather than the opponent color channels and that the distinction between the fixated and non-fixated regions is determined to a large degree by the low resolution image a finding which is in keeping with the results published by<sup>18</sup>. Further, we find that the transformations pertaining to rotation by 90 and 270 degrees are slightly but consistently more significant than those associated with the other transformations.

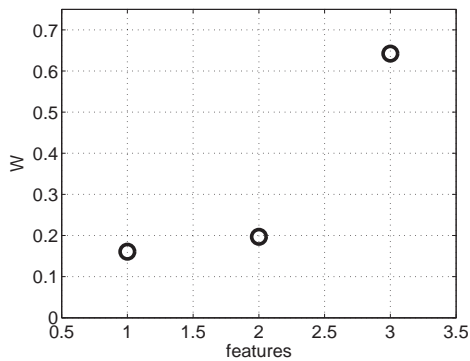


FIG. 17: The discrimination vector  $w$  for the visual saliency algorithm by<sup>16</sup>. The three features are: color, intensity, and orientation respectively. Here the orientation feature is more significant.

salient. By grouping the associated values for the feature maps obtained from the visual saliency algorithm by<sup>16</sup> and the proposed method into two matrices one pertaining to the fixated regions and the other to the non-fixated areas we were able to use linear discrimination to separate the regions optimally. Our working hypothesis was that being able to distinguish between the local values of

the feature maps at fixated and non-fixated regions would indicate that the algorithms are indeed useful in predicting eye fixations. Our findings indicate that both the visual saliency algorithm by<sup>16</sup> and the proposed group metric are nearly ideal at predicting non-salient and highly salient regions with a considerable confusion in the mid saliency region. Further more, test results on 200 images and fifteen observers indicate that the proposed metric is better at separating salient and non-salient regions than the saliency algorithm by<sup>16</sup>. Here we emphasize that in the case of the group algorithm we did not make use of the center surround aspect of the saliency algorithm but rather added the feature maps obtained at different scales linearly. Thus, our first experiments shows that salient image features can be predicted by the proposed dihedral group  $D_4$  transformations.

In the second experiment, we compared the proposed algorithm with the best current methods and found that it ranked as second best for the given data. We thus concluded that the transformations of the dihedral group  $D_4$  are a good metric to estimate salient image regions which if backed by further studies can represent a mathematically sound method to define a salient image region.

<sup>1</sup>Ali Alsam, Puneet Sharma, and Anette Wrånsen. Asymmetry as a measure of visual saliency. In *SCIA 2013, Lecture Notes in Computer Science (LNCS)*, volume 7944, pages 591–600. Springer-Verlag Berlin Heidelberg, 2013.

<sup>2</sup>A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

<sup>3</sup>Ali Borji, Dicky N. Sihite, and Laurent Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2013.

<sup>4</sup>Jochen Braun and Dov Sagi. Vision outside the focus of attention. *Perception and Psychophysics*, 48(1):45–58, 1990.

<sup>5</sup>Neil D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. In *NIPS'05*, pages 155–162, 2005.

<sup>6</sup>William Y. Chang. Image processing with wreath products. Master's thesis, Harvey Mudd College, 2004.

<sup>7</sup>Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual Reviews in the Neurosciences*, 18:193–222, 1995.

<sup>8</sup>David S. Dummit and Richard M. Foote. *Abstract Algebra*. John Wiley & Sons, 2004.

<sup>9</sup>Tom Fawcett. Roc graphs with instance-varying costs. *Pattern Recognition Letters*, 27(8):882–891, 2004.

<sup>10</sup>Richard Foote, Gagan Mirchandani, Daniel N. Rockmore, Dennis Healy, and Tim Olson. A wreath product group approach to signal and image processing. i. multiresolution analysis. *IEEE Transactions on Signal Processing*, 48(1):102–132, 2000.

<sup>11</sup>Anton Garcia-Diaz, Xose R. Fdez-Vidal, Xose M. Pardo, and Raquel Dosl. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, 2012.

<sup>12</sup>Richard L. Gorsuch. *Factor Analysis*. Lawrence Erlbaum Associates (LEA), 1983.

<sup>13</sup>Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 545–552. MIT Press, 2006.

<sup>14</sup>Xiaodi Hou and Liqing Zhang. Computer vision and pattern recognition, 2007. cvpr '07. ieee conference on. In *Saliency Detection: A Spectral Residual Approach*, pages 1–8, 2007.

<sup>15</sup>Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001.

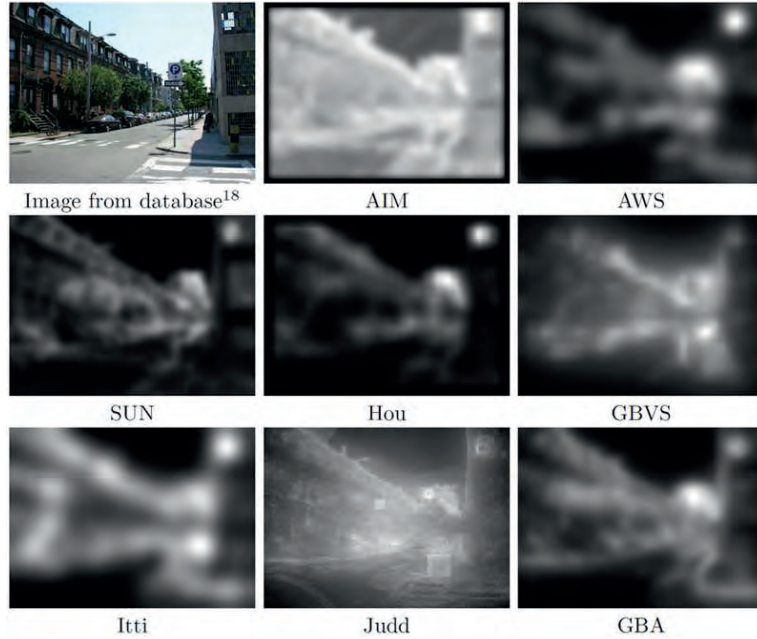


FIG. 18: Figure shows a given image and the associated saliency maps obtained from the different saliency models used in this paper. The saliency maps are normalized in the range [0,1].

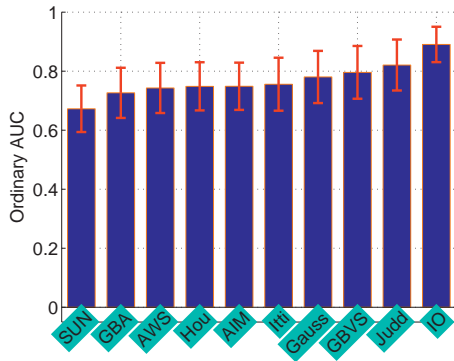


FIG. 19: Ranking of visual saliency models using the ordinary AUC metric. The results are obtained from 1003 images and fifteen observers. Error bars show the uncertainty of the mean (defined by 1 standard deviation).

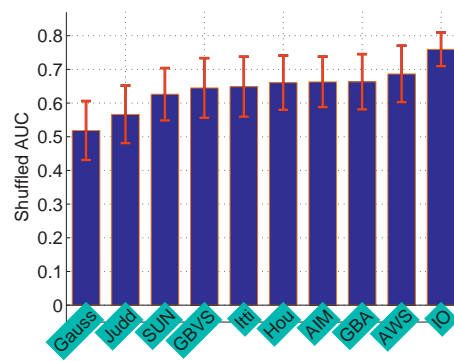


FIG. 20: Ranking of visual saliency models using the shuffled AUC metric. The results are obtained from 1003 images and fifteen observers. Error bars show the uncertainty of the mean (defined by 1 standard deviation).

<sup>16</sup>Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

<sup>17</sup>I. T. Jolliffe. *Principal component analysis*. Springer, 2002.

<sup>18</sup>Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Tor-

ralba. Learning to predict where humans look. In *International Conference on Computer Vision (ICCV)*, pages 2106–2113. IEEE, September 2009.

<sup>19</sup>C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4:219–227, 1985.



- <sup>20</sup>Reiner Lenz. Using representations of the dihedral groups in the design of early vision filters. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*, pages 165–168. IEEE, 1993.
- <sup>21</sup>Reiner Lenz. Investigation of receptive fields using representations of the dihedral groups. *Journal of Visual Communication and Image Representation*, 6(3):209–227, 1995.
- <sup>22</sup>Reiner Lenz, Thanh Hai Bui, and Koichi Takase. A group theoretical toolbox for color image operators. In *ICIP 2005. IEEE International Conference on Image Processing*, volume 3, pages 557–560, 2005.
- <sup>23</sup>Michael C. Mozer and Mark Sitton. *Computational modeling of spatial attention*, chapter 9, pages 341–393. Psychology Press, 1998.
- <sup>24</sup>Vidhya Navalpakkam and Laurent Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2:2049 – 2056, 2006.
- <sup>25</sup>Scott B. Steinman and Barbara A. Steinman. Vision and attention. i: Current models of visual attention. *Optometry and Vision Science*, 75(2):146–155, 1998.
- <sup>26</sup>Katrin Suder and Florentin Worgotter. The control of low-level information flow in the visual system. *Reviews in the Neurosciences*, 11:127–146, 2000.
- <sup>27</sup>Benjamin W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7:1–17, 2007.
- <sup>28</sup>Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643–659, March 2005.
- <sup>29</sup>Po-He Tseng, Ran Carmi, Ian G. M. Cameron, Douglas P. Munoz, and Laurent Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7):1–16, 2009.
- <sup>30</sup>Dirk Walthner and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, 2006.
- <sup>31</sup>Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20, 2008.

## **A.7 What the eye did not see—a fusion approach to image coding**

**Authors:** Ali Alsam, Hans Jakob Rivertz, and Puneet Sharma.

**Full title:** What the eye did not see—a fusion approach to image coding.

**Published in:** ISVC 2012, Advances in Visual Computing, Lecture Notes in Computer Science (LNCS), Springer-Verlag Berlin Heidelberg.

# What the eye did not see—a fusion approach to image coding

Ali Alsam, Hans Jakob Rivertz, and Puneet Sharma

Department of Informatics & e-Learning(AITeL)  
Sør-Trøndelag University College(HiST)  
Trondheim, Norway  
email: er.puneetsharma@gmail.com

**Abstract.** The concentration of the cones and ganglion cells is much higher in the fovea than the rest of the retina. This non-uniform sampling results in a retinal image that is sharp at the fixation point, where a person is looking, and blurred away from it. This difference between the sampling rates at the different spatial locations presents us with the question of whether we can employ this biological characteristic to achieve better image compression. This can be achieved by compressing an image less at the fixation point and more away from it. It is, however, known that the vision system employs more than one fixation to look at a single scene which presents us with the problem of combining images pertaining to the same scene but exhibiting different spatial contrasts. This article presents an algorithm to combine such a series of images by using image fusion in the gradient domain. The advantage of the algorithm is that unlike other algorithms that compress the image in the spatial domain our algorithm results in no artifacts. The algorithm is based on two steps, in the first we modify the gradients of an image based on a limited number of fixations and in the second we integrate the modified gradient. Results based on measured and predicted fixations verify our approach.

## 1 Introduction

From the very beginning of photography, cameras were designed and iteratively improved with the aim of mimicking the human visual system. From this perspective, a camera is thought of as a machined eye—a device that is sensitive to illumination. Equally, we normally think of algorithms such as white-balancing [1], adaptation [2] and tone mapping [3] as being similar to the biological processes of the vision system.

A camera is of course not a human visual system. The two are different in a number of ways some of which are relevant to the work presented in this article. Primarily, while digital camera manufacturers are striving to produce devices with progressively higher resolution, the human brain has evolved to be efficient, i.e. use less information to reach greater conclusions. Thus while the camera sensor has a uniform number of pixels per unit area, the human eye has

a much higher resolution in the fovea which is the center part of the retina [4]. It is well known that the fovea is responsible for our central, sharpest vision while the cone distribution in the rest of the retina results in blurred vision [4].

In the process of exploring a scene, the brain directs the eyes to different spatial locations. At those locations, known as fixations the eyes pause and gather the visual information [5]. Due to the concentration of photo-receptors at the fovea, we can think of each pause as the time taken to capture an image that is sharp at the fixation point and blurred away from it. Given that the average distribution per unit area and spatial location of the cones in the retina is known, it is possible to model the spatial contrast of the retinal image at each fixation.

For a given scene, the number of fixations and their locations vary. The question of whether fixations are guided by image features has been addressed extensively in vision research; and some conclusions are widely accepted. Specifically, experiments have shown that for a given image, people tend to look at the same regions [6, 7], they tend to look at the central part [8, 7] and that certain image attributes such as luminance and colour contrasts tend to attract fixations [9, 10]. Furthermore, fixations can be measured using eye trackers and the experimental data shows conclusively that for a general image the human visual system employs more than one fixation [6].

Based on a given digital image and a number of measured or predicted fixations, we can model the foveation effect, i.e a sharp region at the fixation point and blurring away from it. The result of such a model would be a number of images with different spatial contrast. As an example, see figure 1 where we have modeled the foveation effect based on 3 different fixations. Given such an image series we might wonder how the vision system integrates the different foveation results into a seamless visual experience; and subsequently how we can design signal processing algorithms that offer such functionality.

In this article, we present an algorithm which integrates a number of differently foveated images in the gradient domain. The algorithm starts by calculating the gradients of the input image. Having done that a number of fixation locations are used to calculate the corresponding foveated gradients. Here we use the foveation function described by Geisler and Perry [11]. As a second step, the gradients are combined using the fast colour to gray algorithm by Alsam and Drew [12]. The Alsam and Drew algorithm [12] combines the gradients from  $n$  channels into a single gradient by arguing that the maximum horizontal and vertical differences over all the channels result in the maximum contrast. Thus the gradient fusion step is guaranteed to result in a gradient where the maximum differences pertaining to the fixations locations are maintained. As a final step the resultant gradient is integrated using the modified Frankot-Chellappa-algorithm [13] proposed by Alsam and Rivertz [14].

The need for a fast algorithm to combine foveated images is best motivated in the image compression domain where improvements in statistically based image compression, i.e. methods that are based on data analysis have long slowed down. The use of human vision steered compression is seen by researchers as the

most promising path toward further improvements. In this regard, the algorithm presented in this article can be used as part of an image compression pipeline with very promising results. From our initial tests, we have noticed that the algorithm results in reduced storage requirements without the added artifacts associated with frequency based compressions in the wavelets domain.

Like other foveation driven algorithms, our method is dependent on accurate estimation of the fixation points. Thus in our experimental section, we present results based on measured fixation data as well as predictions based on the visual saliency algorithm by Itti et al. [15].



**Fig. 1.** Figures show the foveated images for three fixations, here the fixation points are represented as red dots.

## 2 The filter and the integration.

Experiments for measuring the contrast sensitivity of the human eye have been carried out [16, 17]. Based on these experiments, the contrast threshold has been modeled through the function

$$CT(f, \theta) = CT_0 \exp\left(\alpha f \frac{\theta + \theta_2}{\theta_2}\right).$$

Here,  $f$  is the spatial frequency measured in degrees,  $\theta$  is the retinal eccentricity.  $CT_0$  is the minimal contrast threshold,  $\theta_2$  is the half-resolution eccentricity constant, and  $\alpha$  is the spatial frequency decay constant. The values used in [18] are  $\alpha = 0.106$ ,  $\theta_2 = 2.3$ , and  $CT_0 = 1/64$ .

Given a normalized gray scale image  $z_0 : \Omega \rightarrow [0, 1]$ . Denote its width by  $w$ , measured in pixels. An observer views the image from a distance  $d$ , measured in

pixels. The maximal spatial frequency of the image is given by  $f_d = \frac{w}{4 \arctan \frac{w}{2d}}$ . If  $r$  is the distance measured in pixels from a fixation point, then  $\theta(r) = \arctan \frac{r}{d}$ .

The gradient  $\nabla z_0$  is modified by setting its magnitude to zero if it is less than  $CT(f_d, \theta)$  for some of the fixation points.

We make a new contrast threshold function based on  $f = f_d$  and the fixation points,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .

$$CT(x, y) = \min(CT_1(x, y), CT_2(x, y), \dots, CT_n(x, y)),$$

where  $CT_k(x, y) = CT\left(f_d, \theta\left(\sqrt{(x - x_k)^2 + (y - y_k)^2}\right)\right)$ ,  $k = 1, 2, \dots, n$ . This step is equivalent to the Alsam and Drew method [12].

The direction of the original and modified gradients are  $\hat{u} = \nabla z_0 / |\nabla z_0|$ . The length of the new gradient is  $|\nabla z| = |\nabla z_0|$  if  $CT(x, y) < |\nabla z_0|$ , otherwise  $|\nabla z| = 0$ . We now reconstruct the contrast by using the integration method of Alsam and Rivertz [14] where we minimize the functional:

$$W(z) = \lambda \int_{\Omega} |z - z_0|^2 dx dy + \int_{\Omega} (|z_x - p|^2 + |z_y - q|^2) dx dy.$$

This minimization results in an image whose gradients are as close as possible to  $(p, q)$ , under the constraint that the luminance is close to the original image. The image  $z$  in the Fourier domain can be taken as

$$Z(u, v) = \frac{\lambda Z_0 - i(uP + vQ)}{\lambda + u^2 + v^2},$$

where  $P$  and  $Q$  correspond to the Fourier transforms of  $p$ , and  $q$ .

### 3 Results

To test the proposed method, we used images and corresponding fixations data from the study by Judd et al. [6]. The results for two images and the associated fixations are shown in figures 2 to 3. In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. In agreement with the predicted results for the application of the contrast function by Wang and Bovik [18], we notice that the regions around the fixation points are sharper than the rest. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. Here we remark that the difference between the original and the result can be optimized by controlling the  $\lambda$  parameter defined in the previous section.

In figure 4, the left column contains the foveated images obtained by using the first three salient points from the visual saliency algorithm by Itti et al. [15]

and the right column contains the original image, the result obtained by using the proposed method, and the difference between the result and the original image. For this experiment, we notice that the results are very similar to those obtained for the first test image. We underline, however, that the choice of fixation locations and the number of salient regions is clearly related to the results that we obtain, where the higher the number of fixations and the more spread they are in the image plane the closer the result is going to resemble the original.

Finally, in figures 5(a) to 5(f), we show the bitrates obtained by saving the original image and corresponding result image in JPEG format with different quality values, ranging from 10 to 100 based on six different images. Here we notice that for the same compression quality the new images require lower storage space. Given that the foveation function reduces the high frequency elements of the original image, we can argue that this result is not surprising. The advantages of this approach are, however, more subtle than a simple removal of high frequency elements- we have removed high frequencies locally- in regions where the foveation function predicts that the observer couldn't see with the sharp part of their vision.

## 4 Conclusion

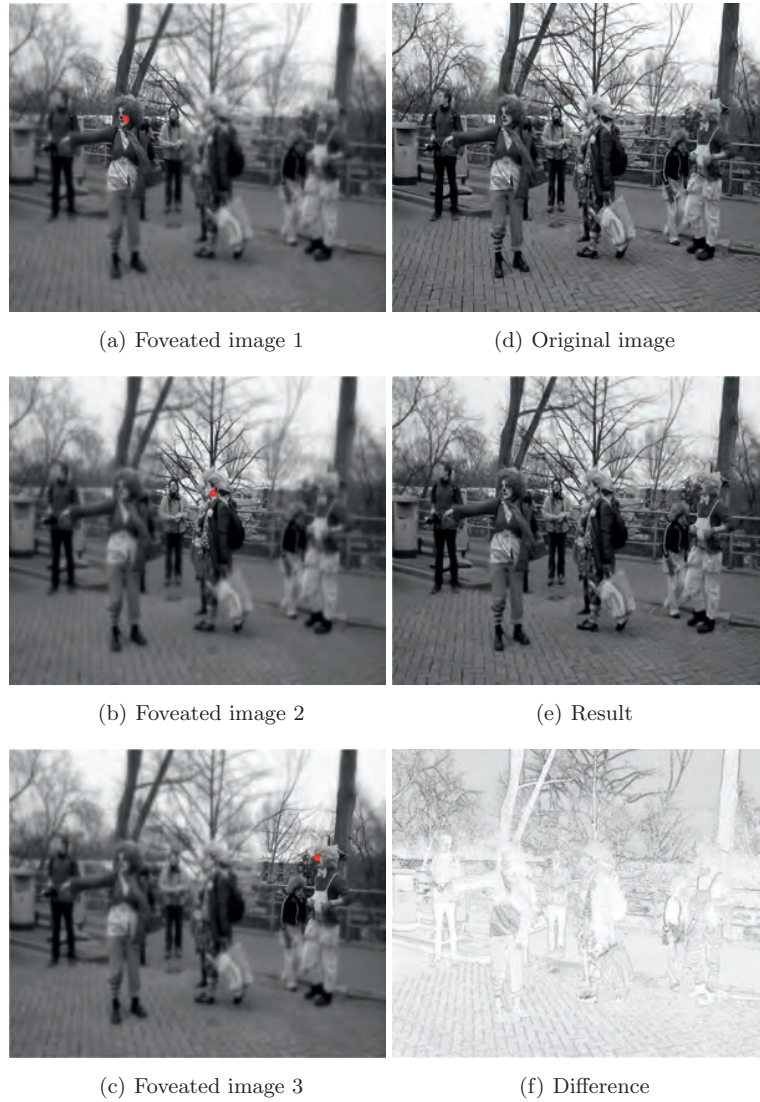
This article presents an algorithm to combine a series of differently foveated images pertaining to an identical scene. This is achieved by using image fusion in the gradient domain. The advantage of the algorithm is that unlike other algorithms that compress the image in the spatial domain our algorithm results in no artifacts. The algorithm is based on two steps, in the first we modify the gradients of an image based on a limited number of fixations and in the second we integrate the modified gradient. Results based on measured and predicted fixations verify our approach. The need for a fast algorithm to combine foveated images is best motivated in the image compression domain where improvements in statistically based image compression, i.e. methods that are based on data analysis have long slowed down. The use of human vision steered compression is seen by researchers as the most promising path toward further improvements. In this regard, the algorithm presented in this article can be used as part of an image compression pipeline with very promising results. From our initial tests, we have noticed that the algorithm results in reduced storage requirements without the added artifacts associated with frequency based compressions in the wavelets domain.

## References

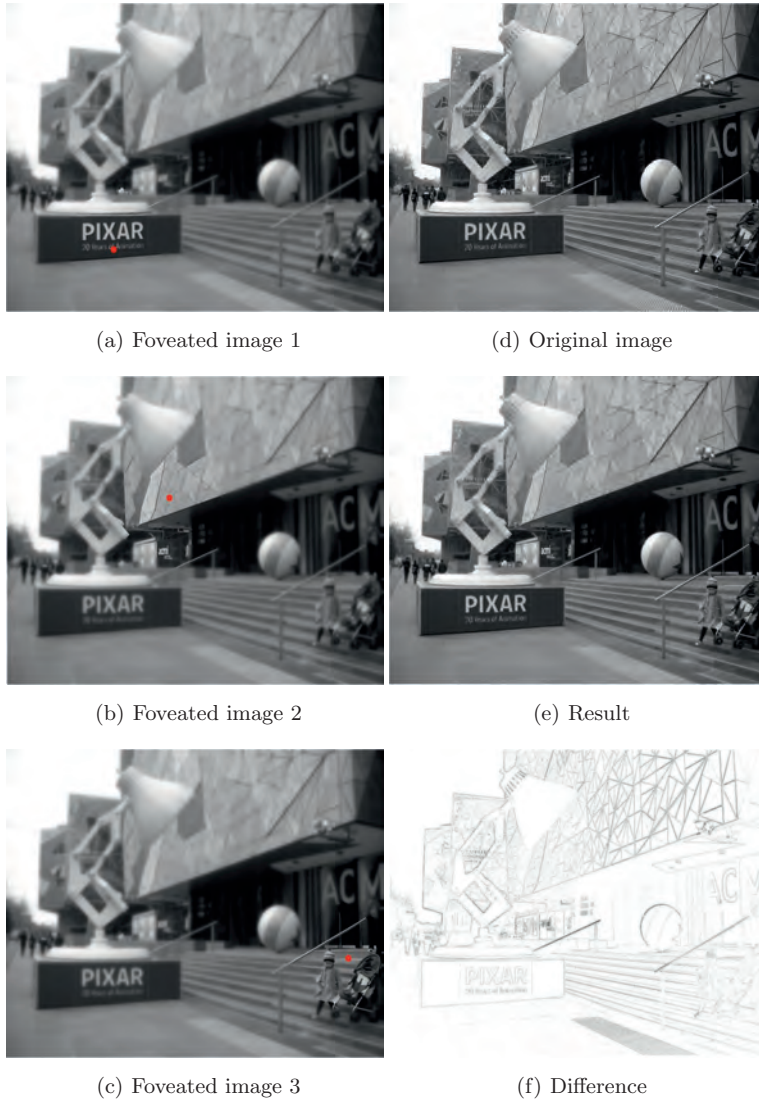
1. Chikane, V., Fuh, C.S.: Automatic white balance for digital still cameras. *Journal Of Information Science and Engineering* **22** (2006) 497-509
2. Hurley, J.B.: Shedding light on adaptation. *Journal of General Physiology* **119** (2002) 125-128

3. Qiu, G., Guan, J., Duan, J., Chen, M.: Tone mapping for hdr image using optimization a new closed form solution. In: ICPR 2006. 18th International Conference on Pattern Recognition. Volume 1. (2006) 996–999
4. Cormack, L.K.: Computational models of early human vision. In: Handbook of Image and Video Processing. Elsevier Academic Press (2005) 325–345
5. Rajashekar, U., van der Linde, I., Bovik, A.C., Cormack, L.K.: Gaffe: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing* **17** (2008) 564–573
6. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: International Conference on Computer Vision (ICCV). (2009)
7. Alsam, A., Sharma, P.: Analysis of eye fixations data. In: Proceedings of the IASTED International Conference, Signal and Image Processing (SIP 2011). (2011) 342–349
8. Tatler, B.W.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision* **7** (2007) 1–17
9. Itti, L., Koch, C.: Computational modelling of visual attention. *Nature Reviews Neuroscience* **2** (2001) 194–203
10. Meur, O.L., Callet, P.L., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28** (2006) 802–817
11. Geisler, W.S., Perry, J.S.: A real-time foveated multiresolution system for low-bandwidth video communication. In: SPIE Proceedings. Volume 3299. (1998) 1–13
12. Alsam, A., Drew, M.S.: Fast colour2grey. In: 16th Color Imaging Conference: Color, Science, Systems and Applications, Society for Imaging Science & Technology (IS&T)/Society for Information Display (SID) joint conference. (2008) 342–346
13. Frankot, R.T., Chellappa, R.: A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10** (1988) 439–451
14. Alsam, A., Rivertz, H.J.: Constrained gradient integration for improved image contrast. In: Proceedings of the IASTED International Conference, Signal and Image Processing (SIP 2011). (2011) 13–18
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (1998) 1254–1259
16. Banks, M., Sekuler, A., Anderson, S.: Peripheral spatial vision: limits imposed by optics, photoreceptors, and receptor pooling. *J. Opt. Soc. Am. A* **8** (1991) 1775–1787
17. Arnou, T.L., Geisler, W.S.: Visual detection following retinal damage: Predictions of an inhomogeneous retino-cortical model. In: Human Vision and Electronic Imaging. Proceedings of SPIE. Volume 2674. (1996)
18. Wang, Z., Bovik, A.C.: Embedded foveation image coding. *IEEE Transactions on Image Processing* **10** (2001) 1397–1410

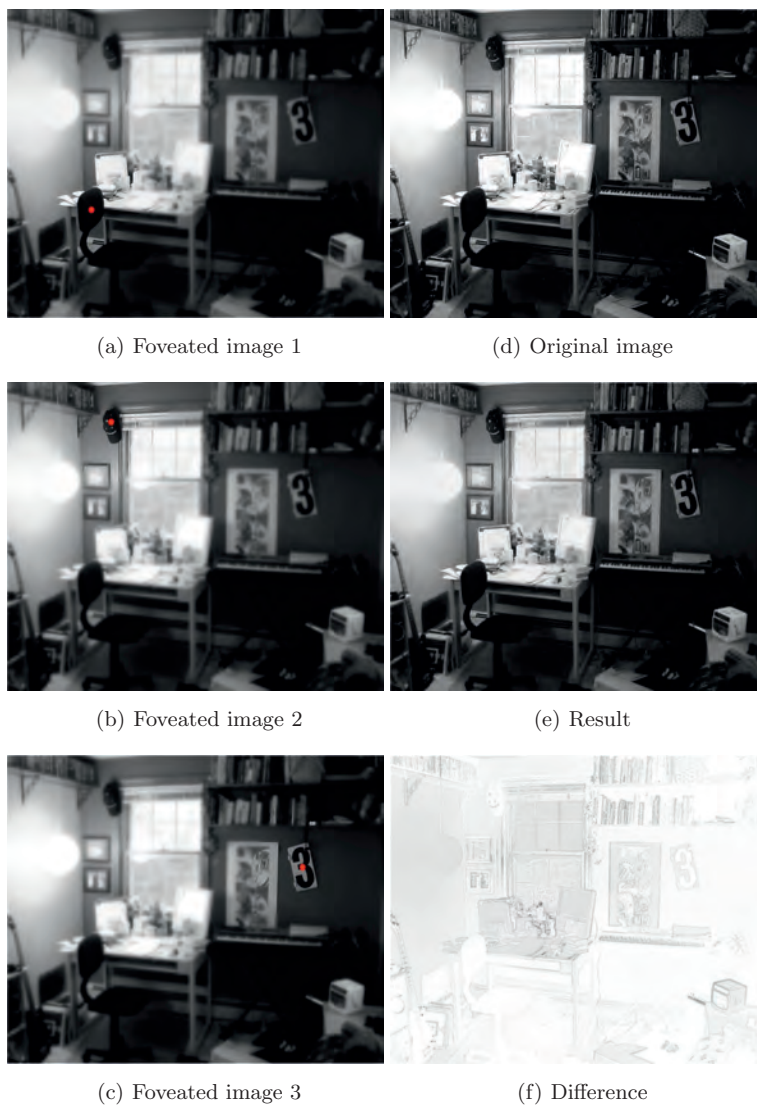




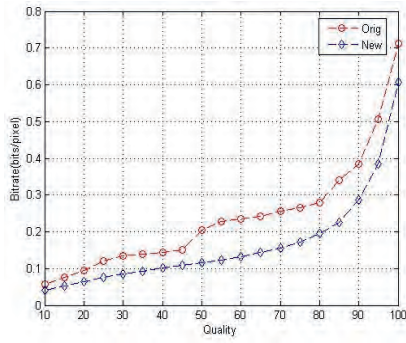
**Fig. 2.** In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. In the difference image, the dark regions indicate the locations where the differences are higher.



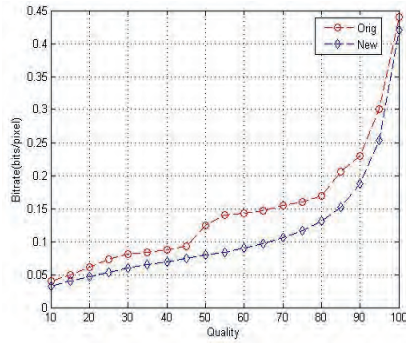
**Fig. 3.** In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. In the difference image, the dark regions indicate the locations where the differences are higher.



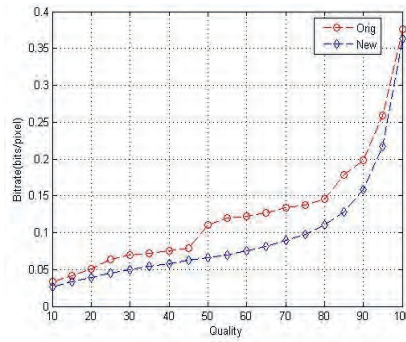
**Fig. 4.** In the left column the foveated images obtained by using first three salient points from the visual saliency algorithm by Itti et al. [15] are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. In the difference image, the dark regions indicate the locations where the differences are higher.



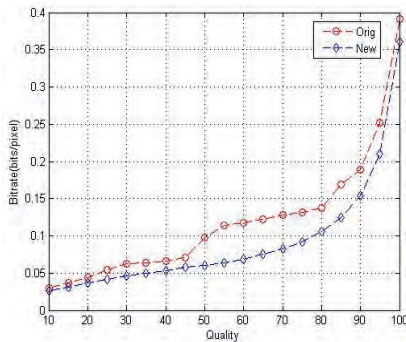
(a) image 1



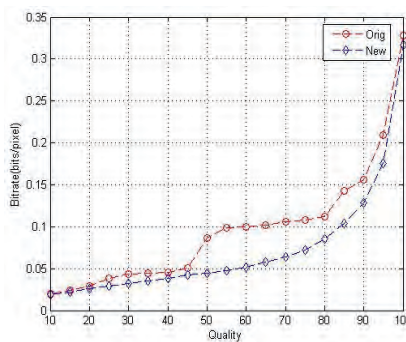
(b) image 2



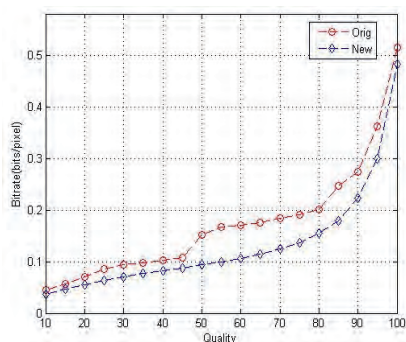
(c) image 3



(d) image 4



(e) image 5



(f) image 6

**Fig. 5.** Figures show the bitrates for saving the original image and corresponding result image in JPEG format with different quality values, ranging from 10 to 100 based on six different images. Here we notice that for the same compression quality the new images require lower storage space.



## **A.8 What the eye did not see—a fusion approach to image coding (extended)**

**Authors:** Ali Alsam, Hans Jakob Rivertz, and Puneet Sharma.

**Full title:** What the eye did not see—a fusion approach to image coding.

**Published in:** International Journal on Artificial Intelligence Tools.

## WHAT THE EYE DID NOT SEE — A FUSION APPROACH TO IMAGE CODING

ALI ALSAM, HANS JAKOB RIVERTZ and PUNEET SHARMA\*

*Department of Informatics & e-Learning (AITeL)  
Sør-Trøndelag University College (HiST)  
Trondheim, Norway*

*\*er.puneetsharma@gmail.com*

Received 15 January 2013

Accepted 14 July 2013

Published 20 December 2013

The concentration of the cones and ganglion cells is much higher in the fovea than the rest of the retina. This non-uniform sampling results in a retinal image that is sharp at the fixation point, where a person is looking, and blurred away from it. This difference between the sampling rates at the different spatial locations presents us with the question of whether we can employ this biological characteristic to achieve better image compression. This can be achieved by compressing an image less at the fixation point and more away from it. It is, however, known that the vision system employs more than one fixation to look at a single scene which presents us with the problem of combining images pertaining to the same scene but exhibiting different spatial contrasts. This article presents an algorithm to combine such a series of images by using image fusion in the gradient domain. The advantage of the algorithm is that unlike other algorithms that compress the image in the spatial domain our algorithm results in no artifacts. The algorithm is based on two steps, in the first we modify the gradients of an image based on a limited number of fixations and in the second we integrate the modified gradient. Results based on measured and predicted fixations verify our approach.

*Keywords:* Foveation; image compression; eye fixations.

### 1. Introduction

From the very beginning of photography, cameras were designed and iteratively improved with the aim of mimicking the human visual system. From this perspective, a camera is thought of as a machined eye — a device that is sensitive to illumination. Equally, we normally think of algorithms such as white-balancing,<sup>7</sup> adaptation<sup>11</sup> and tone mapping<sup>17</sup> as being similar to the biological processes of the vision system.

\*Corresponding author

A camera is of course not a human visual system. The two are different in a number of ways some of which are relevant to the work presented in this article. Primarily, while digital camera manufacturers are striving to produce devices with progressively higher resolution, the human brain has evolved to be efficient, i.e., use less information to reach greater conclusions. Thus while the camera sensor has a uniform number of pixels per unit area, the human eye has a much higher resolution in the fovea which is the center part of the retina.<sup>8</sup> It is well known that the fovea is responsible for our central, sharpest vision while the cone distribution in the rest of the retina results in blurred vision.<sup>8</sup>

In the process of exploring a scene, the brain directs the eyes to different spatial locations. At those locations, known as fixations the eyes pause and gather the visual information.<sup>18</sup> Due to the concentration of photo-receptors at the fovea, we can think of each pause as the time taken to capture an image that is sharp at the fixation point and blurred away from it. Given that the average distribution per unit area and spatial location of the cones in the retina is known, it is possible to model the spatial contrast of the retinal image at each fixation.

For a given scene, the number of fixations and their locations vary. The question of whether fixations are guided by image features has been addressed extensively in vision research; and some conclusions are widely accepted. Specifically, experiments have shown that for a given image, people tend to look at the same regions,<sup>14,3</sup> they tend to look at the central part<sup>21,20,3</sup> and that certain image attributes such as luminance and colour contrasts tend to attract fixations.<sup>22,15,19,12,16,4</sup> Furthermore, fixations can be measured using eye trackers and the experimental data shows conclusively that for a general image the human visual system employs more than one fixation.<sup>14</sup>

Based on a given digital image and a number of measured or predicted fixations, we can model the foveation effect, i.e., a sharp region at the fixation point and blurring away from it. The result of such a model would be a number of images with different spatial contrast. As an example, see Fig. 1 where we have modeled the foveation effect based on 3 different fixations. Given such an image series we might wonder how the vision system integrates the different foveation results into a seamless visual experience; and subsequently how we can design signal processing algorithms that offer such functionality.

In this article, we present an algorithm which integrates a number of differently foveated images in the gradient domain. The algorithm starts by calculating the gradients of the input image. Having done that a number of fixation locations are used to calculate the corresponding foveated gradients. Here we use the foveation function described by Geisler and Perry.<sup>10</sup> As a second step, the gradients are combined using the fast colour to gray algorithm by Alsam and Drew.<sup>1</sup> The Alsam and Drew algorithm<sup>1</sup> combines the gradients from  $n$  channels into a single gradient by arguing that the maximum horizontal and vertical differences over all the channels result in the maximum contrast. Thus the gradient fusion step is guaranteed to result in a gradient where the maximum differences pertaining to the fixations



Fig. 1. (Color online) Figures show the foveated images for three fixations, here the fixation points are represented as red dots.

locations are maintained. As a final step the resultant gradient is integrated using the modified Frankot-Chellappa-algorithm<sup>9</sup> proposed by Alsam and Rivertz.<sup>2</sup>

The need for a fast algorithm to combine foveated images is best motivated in the image compression domain where improvements in statistically based image compression, i.e., methods that are based on data analysis have long slowed down. The use of human vision steered compression is seen by researchers as the most promising path toward further improvements. In this regard, the algorithm presented in this article can be used as part of an image compression pipeline with very promising results. From our initial tests, we have noticed that the algorithm results in reduced storage requirements without the added artifacts associated with frequency based compressions in the wavelets domain.

Like other foveation driven algorithms, our method is dependent on accurate estimation of the fixation points. Thus in our experimental section, we present results based on measured fixation data as well as predictions based on the visual saliency algorithm by Itti *et al.*<sup>13</sup>

## 2. The Filter and the Integration

Experiments for measuring the contrast sensitivity of the human eye have been carried out.<sup>6,5</sup> Based on these experiments, the contrast threshold has been modeled through the function

$$CT(f, \theta) = CT_0 \exp\left(\alpha f \frac{\theta + \theta_2}{\theta_2}\right).$$

Here,  $f$  is the spatial frequency measured in cycles per degrees,  $\theta$  is the retinal eccentricity.  $CT_0$  is the minimal contrast threshold,  $\theta_2$  is the half-resolution eccentricity constant, and  $\alpha$  is the spatial frequency decay constant. We use  $\alpha = 0.106$ ,



$\theta_2 = 2.3$ , and  $CT_0 = 1/64$  obtained from the experiments done by Geisler and Perry in Ref. 10. Other experiments give slightly different values. This is discussed in Ref. 23.

Given a normalized gray scale image  $z_0 : \Omega \rightarrow [0, 1]$ . Denote its width by  $w$ , measured in pixels. An observer views the image from a distance  $d$ , measured in pixels. The maximal spatial frequency of the image is given by  $f_d = \frac{w}{4 \arctan \frac{w}{2d}}$ . If  $r$  is the distance measured in pixels from a fixation point, then  $\theta(r) = \arctan \frac{r}{d}$ .

We make a new contrast threshold function based on  $f = f_d$  and the fixation points,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .

$$CT(x, y) = \min(CT_1(x, y), CT_2(x, y), \dots, CT_n(x, y)),$$

where  $CT_k(x, y) = CT\left(f_d, \theta\left(\sqrt{(x - x_k)^2 + (y - y_k)^2}\right)\right)$ ,  $k = 1, 2, \dots, n$ . This step is equivalent to the Alsam and Drew method.<sup>1</sup>

The gradient  $\nabla z_0$  is modified by setting its magnitude to zero if it is less than  $CT(x, y)$  for a fixation point. The new gradient is  $\nabla z = \nabla z_0$  if  $CT(x, y) < |\nabla z_0|$ , otherwise  $\nabla z = \mathbf{0}$ . This step removes high and low frequency contrasts away from a fixation point. The low frequency contrast should, however, be preserved.

We now reconstruct the low frequency contrast by using the integration method of Alsam and Rivertz<sup>2</sup> where we minimize the functional:

$$W(z) = \lambda \int_{\Omega} |z - z_0|^2 dx dy + \int_{\Omega} (|z_x - p|^2 + |z_y - q|^2) dx dy.$$

This minimization results in an image whose gradients are as close as possible to  $(p, q)$ , under the constraint that the luminance is close to the original image. The  $\lambda$  parameter controls how close the pixel values are to the original image. This value of  $\lambda$  is set by the observer such that the resultant image and the original image are visually the same when viewed from the same distance and using the same fixation points.

The image  $z$  in the Fourier domain can be taken as

$$Z(u, v) = \frac{\lambda Z_0 - i(uP + vQ)}{\lambda + u^2 + v^2},$$

where  $P$  and  $Q$  correspond to the Fourier transforms of  $p$ , and  $q$ .

### 3. Results

To test the proposed method, we used images and corresponding fixations data from the study by Judd *et al.*<sup>14</sup> The results for four images and the associated fixations are shown in Figs. 2–5. In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. In agreement with the predicted results for the application of the contrast function by Wang and Bovik,<sup>23</sup> we notice that the regions around the fixation points are sharper than the rest. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference

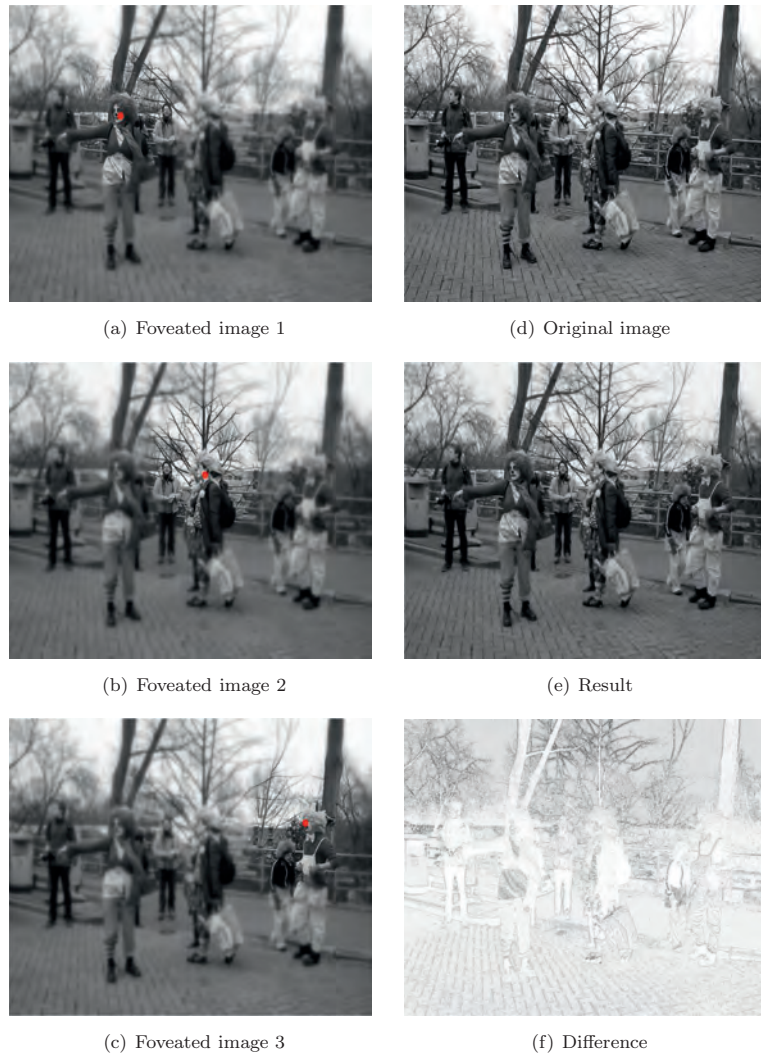


Fig. 2. (Color online) In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. For the result  $\lambda = 0.50$  is used. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. In the difference image, the dark regions indicate the locations where the differences are higher.

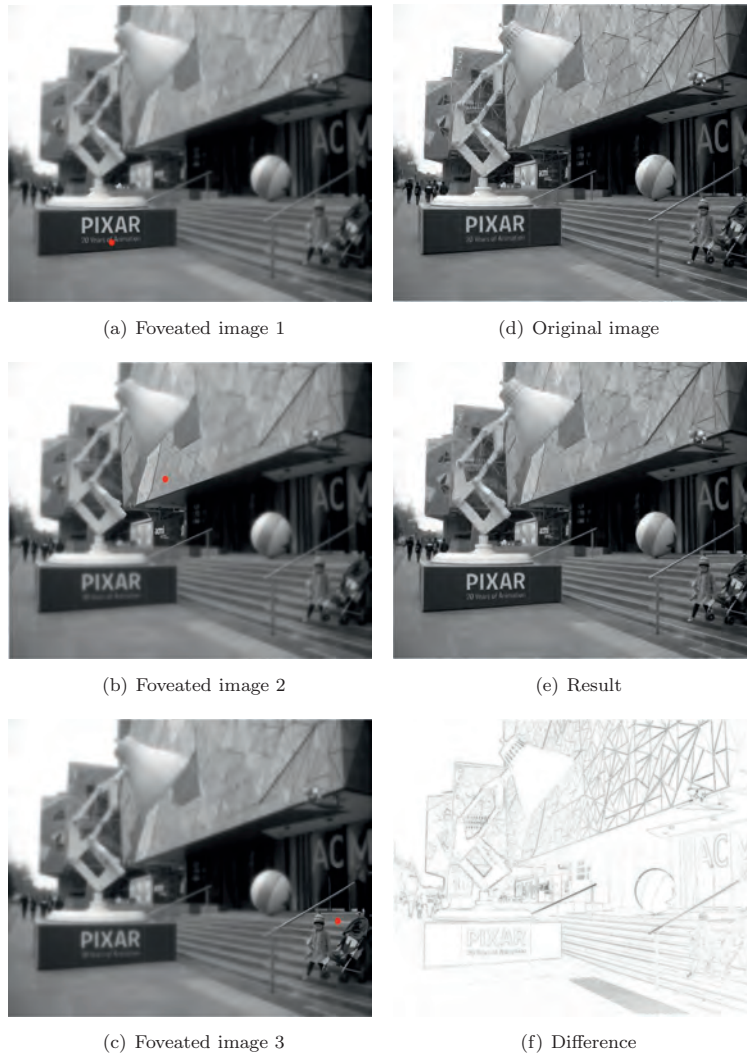


Fig. 3. (Color online) In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. For the result  $\lambda = 0.55$  is used. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. In the difference image, the dark regions indicate the locations where the differences are higher.

1360014-6



Fig. 4. (Color online) In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. For the result  $\lambda = 0.45$  is used. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. In the difference image, the dark regions indicate the locations where the differences are higher.

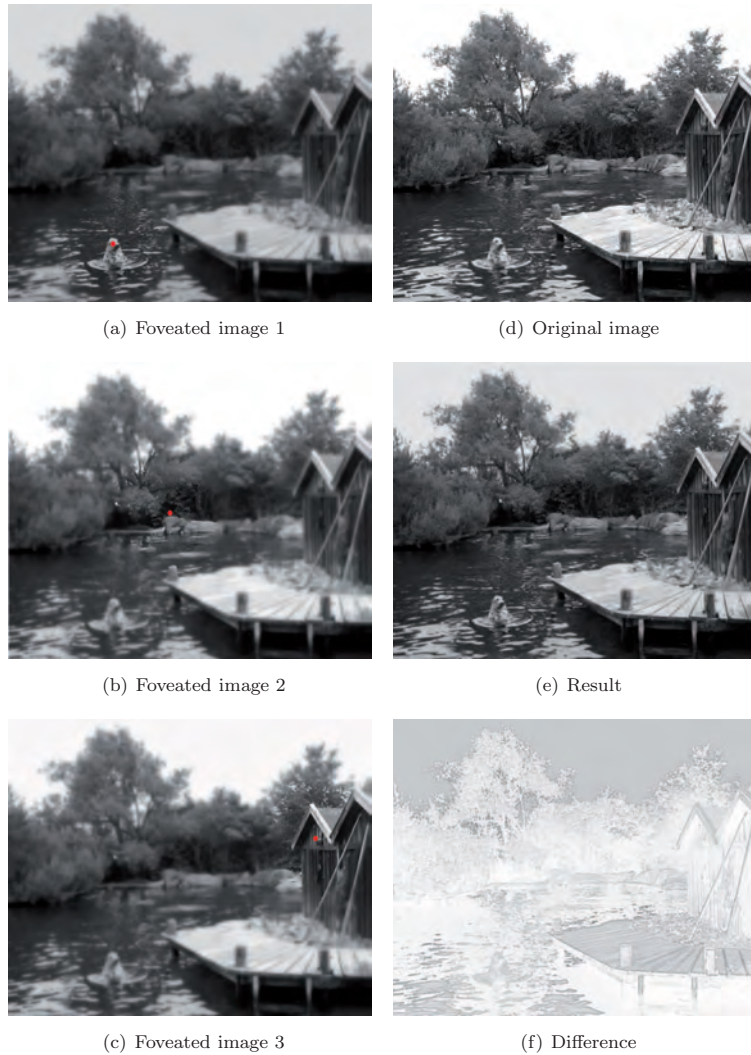


Fig. 5. (Color online) In the left column the foveated images for three fixations are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. For the result  $\lambda = 0.50$  is used. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. In the difference image, the dark regions indicate the locations where the differences are higher.



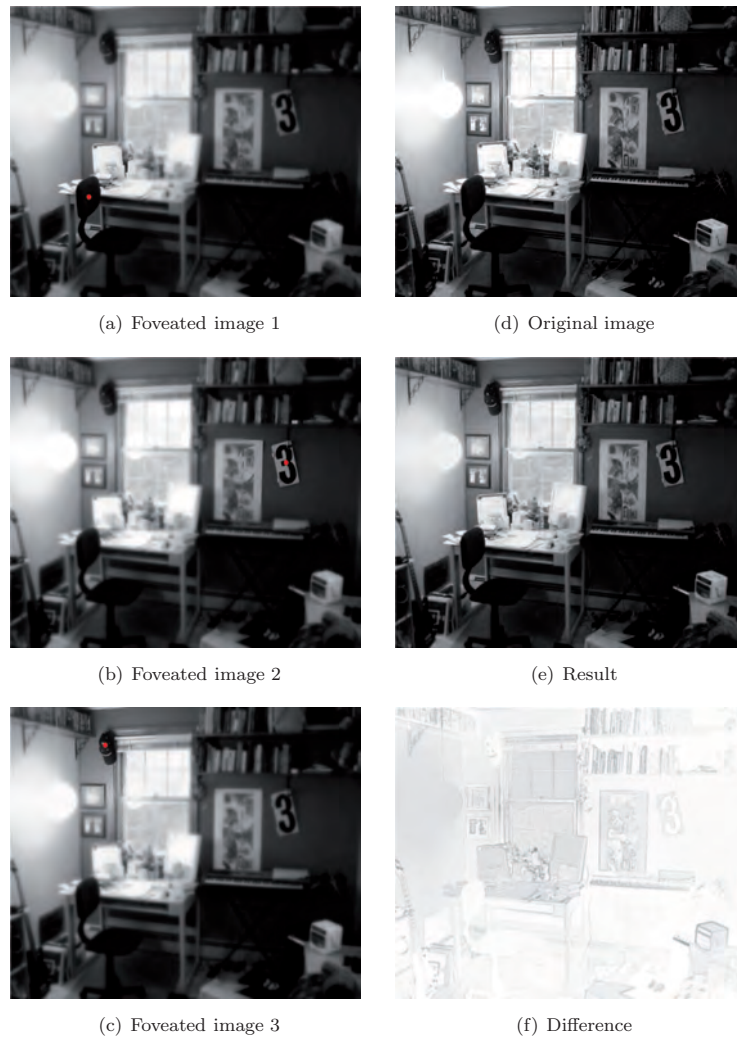


Fig. 6. (Color online) In the left column the foveated images obtained by using first three salient points from the visual saliency algorithm by Itti *et al.*<sup>13</sup> are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. For the result  $\lambda = 0.50$  is used. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. In the difference image, the dark regions indicate the locations where the differences are higher.

between the result and the original image. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. Here we remark that the difference between the original and the result can be optimized by controlling the  $\lambda$  parameter defined in the previous section.

In Figs. 6 and 7, the left column contains the foveated images obtained by using the first three salient points from the visual saliency algorithm by Itti *et al.*<sup>13</sup>

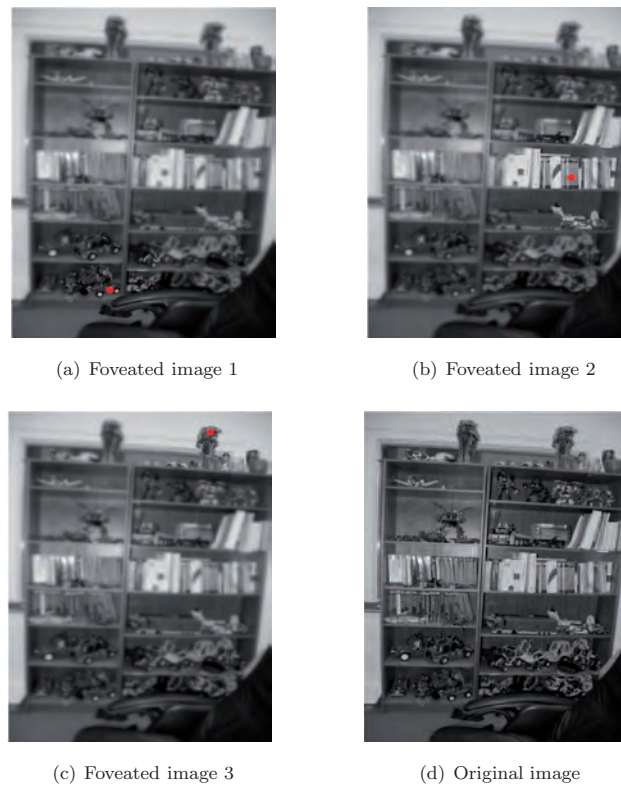


Fig. 7. (Color online) In the left column the foveated images obtained by using first three salient points from the visual saliency algorithm by Itti *et al.*<sup>13</sup> are shown. Here, the fixation points are represented as red dots. The images in the right column show the original image, the result obtained by combining the foveated images using the proposed method, and the difference between the result and the original image. For the result  $\lambda = 0.55$  is used. We notice that the result image is sharp in the regions corresponding to the three fixation points, we further notice that the image represents a good approximation of the original with greater differences in the parts that the observer deemed to be less salient. In the difference image, the dark regions indicate the locations where the differences are higher.

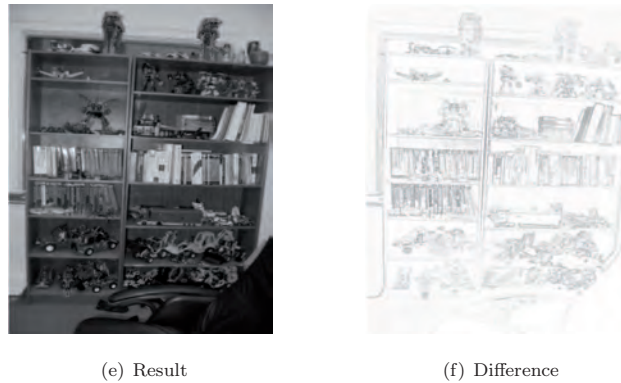


Fig. 7. (Continued)

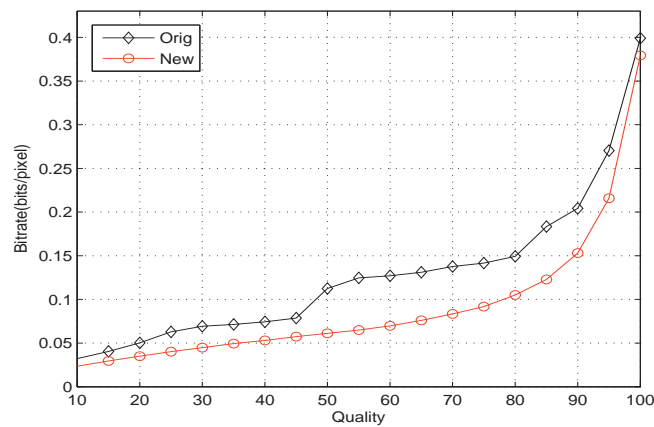


Fig. 8. (Color online) Figure shows the average bitrates for 200 different images obtained by saving the original image and corresponding result image in JPEG format with different quality values, ranging from 10 to 100. Here we notice that for the same compression quality the new images require lower storage space.

and the right column contains the original image, the result obtained by using the proposed method, and the difference between the result and the original image. For this experiment, we notice that the results are very similar to those obtained for the first test image. We underline, however, that the choice of fixation locations and the number of salient regions is clearly related to the results that we obtain, where the higher the number of fixations and the more spread they are in the image plane the closer the result is going to resemble the original.



Finally, in Fig. 8, we show the average bitrates for 200 different images obtained by saving the original image and corresponding result image in JPEG format with different quality values, ranging from 10 to 100. Here we notice that for the same compression quality the new images require lower storage space. Given that the foveation function reduces the high frequency elements of the original image, we can argue that this result is not surprising. The advantages of this approach are, however, more subtle than a simple removal of high frequency elements — we have removed high frequencies locally — in regions where the foveation function predicts that the observer could not see with the sharp part of their vision.

#### 4. Conclusion

This article presents an algorithm to combine a series of differently foveated images pertaining to an identical scene. This is achieved by using image fusion in the gradient domain. The advantage of the algorithm is that unlike other algorithms that compress the image in the spatial domain our algorithm results in no artifacts. The algorithm is based on two steps, in the first we modify the gradients of an image based on a limited number of fixations and in the second we integrate the modified gradient. Results based on measured and predicted fixations verify our approach. The need for a fast algorithm to combine foveated images is best motivated in the image compression domain where improvements in statistically based image compression, i.e., methods that are based on data analysis have long slowed down. The use of human vision steered compression is seen by researchers as the most promising path toward further improvements. In this regard, the algorithm presented in this article can be used as part of an image compression pipeline with very promising results. From our initial tests, we have noticed that the algorithm results in reduced storage requirements without the added artifacts associated with frequency based compressions in the wavelets domain.

#### References

1. Ali Alsam and Mark S. Drew, Fast colour2grey in *16th Color Imaging Conference: Color, Science, Systems and Applications, Society for Imaging Science & Technology (IS&T)/Society for Information Display (SID) Joint Conference (2008)*, pp. 342–346.
2. Ali Alsam and Hans Jakob Rivertz, Constrained gradient integration for improved image contrast, in *Proc. of the IASTED Int. Conf. on Signal and Image Processing (SIP 2011) (2011)*, pp. 13–18.
3. Ali Alsam and Puneet Sharma, Analysis of eye fixations data, in *Proc. of the IASTED Int. Conf. on Signal and Image Processing (SIP 2011) (2011)*, pp. 342–349.
4. Ali Alsam and Puneet Sharma, Validating the visual saliency model, in *SCIA 2013, Lecture Notes in Computer Science (LNCS)*, Vol. 7944 (Springer-Verlag Berlin Heidelberg, 2013), pp. 153–161.
5. Thomas L. Arnow and Wilson S. Geisler, Visual detection following retinal damage: Predictions of an inhomogeneous retino-cortical model, in *Human Vision and Electronic Imaging. Proc. of SPIE*, Vol. 2674 (1996).

6. M. Banks, A. Sekuler and S. Anderson, Peripheral spatial vision: Limits imposed by optics, photoreceptors, and receptor pooling, *J. Opt. Soc. Am. A* **8** (1991) 1775–1787.
7. Varsha Chikane and Chiou-Shann Fuh, Automatic white balance for digital still cameras, *Journal of Information Science and Engineering* **22** (2006) 497–509.
8. Lawrence K. Cormack, Computational models of early human vision, *Handbook of Image and Video Processing* (Elsevier Academic Press, 2005), pp. 325–345.
9. Robert T. Frankot and Rama Chellappa, A method for enforcing integrability in shape from shading algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**(4) (1988) 439–451.
10. Wilson S. Geisler and Jeffrey S. Perry, A real-time foveated multiresolution system for low-bandwidth video communication, in *SPIE Proc.*, Vol. 3299 (1998), pp. 1–13.
11. James B. Hurley, Shedding light on adaptation, *Journal of General Physiology* **119** (2002) 125–128.
12. Laurent Itti and Christof Koch, Computational modelling of visual attention, *Nature Reviews Neuroscience* **2** (2001) 194–203.
13. Laurent Itti, Christof Koch and Ernst Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(11) (1998) 1254–1259.
14. Tilke Judd, Krista Ehinger, Fredo Durand and Antonio Torralba, Learning to predict where humans look, in *Int. Conf. on Computer Vision (ICCV)* (2009).
15. C. Koch and S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, *Human Neurobiology* **4** (1985) 219–227.
16. Olivier Le Meur, Patrick Le Callet, Dominique Barba and Dominique Thoreau, A coherent computational approach to model bottom-up visual attention, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(5) (2006) 802–817.
17. Guoping Qiu, Jian Guan, Jian Duan and Min Chen, Tone mapping for HDR image using optimization a new closed form solution, in *18th Int. Conf. on Pattern Recognition (ICPR 2006)*, Vol. 1 (2006), pp. 996–999.
18. Umesh Rajashekar, Ian van der Linde, Alan C. Bovik and Lawrence K. Cormack, Gaffe: A gaze-attentive fixation finding engine, *IEEE Transactions on Image Processing* **17**(4) (2008) 564–573.
19. Katrin Suder and Florentin Worgotter, The control of low-level information flow in the visual system, *Reviews in the Neurosciences* **11** (2000) 127–146.
20. Benjamin W. Tatler, The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions, *Journal of Vision* **7**(1) (2007) 17.
21. Benjamin W. Tatler, Roland J. Baddeley and Iain D. Gilchrist, Visual correlates of fixation selection: effects of scale and time, *Vision Research* **45** (2005) 643–659.
22. Anne Treisman and Garry Gelade, A feature-integration theory of attention, *Cognitive Psychology* **12**(1) (1980) 97–136.
23. Zhou Wang and Alan C. Bovik, Embedded foveation image coding, *IEEE Transactions on Image Processing* **10**(10) (2001) 1397–1410.

## **A.9 Evaluation of geometric depth estimation model for virtual environment.**

**Authors:** Puneet Sharma, Jan H. Nilsen, Torbjørn Skramstad and Faouzi A. Cheikh.

**Full title:** Evaluation of geometric depth estimation model for virtual environment.

**Published in:** NIK-2010, Tapir Academic Press.

# Evaluation of Geometric Depth Estimation Model for Virtual Environment

Puneet Sharma<sup>1</sup>, Jan H. Nilsen<sup>1</sup>,

Torbjørn Skramstad<sup>2</sup>, Faouzi A. Cheikh<sup>3</sup>

<sup>1</sup>Department of Informatics & E-Learning (AITeL),

Sør Trøndelag University College(HiST), Trondheim, Norway

<sup>2</sup>Department of Computer & Information Science (IDI),

Norwegian University of Science and Technology (NTNU), Trondheim, Norway

<sup>3</sup>Faculty of Computer Science and Media Technology,

Gjøvik University College(HiG), Gjøvik, Norway

## Abstract

3-D virtual environment is a computer generated experience which gives us a feeling of presence in the environment. Objects displayed in virtual environment unlike the real world have no physical depth. Due to the distance between the eyes, the images formed on the retina are different, this facilitates our perception of depth. In the range of personal space, eyes converge at different angles to look at objects in different depth planes, known as convergence angle. Since we cannot get images of the scene viewed by the two eyes, the convergence angle cannot be calculated by standard photogrammetry principles such as triangulation. However, we can measure the point of focus(fixations) of the eyes on 2-D display plane, by using eye tracker. Each eye gets a different view of the virtual scene. Knowing the physical location of both eyes and their corresponding fixations, we can calculate the estimated depth using geometry. In this paper, first, we discuss the experiment setup and 3-D virtual scene used for depth estimation. Second, we evaluate the performance of the geometric model for depth estimation. Third, we discuss a histogram based filtering approach, for improving the performance of the geometric model. Results show that histogram based filtering improves the performance of the geometric model.

## 1 Introduction

A Virtual environment is a computer-generated three-dimensional visual experience displayed either on a computer screen or a stereoscopic display. Due to their low cost and the ability to simulate any real or imagined scenario, virtual environments have assumed a leading role in training personnel such as pilots and fire-fighters to tackle hazardous situations without risking their safety. Though the rendition of virtual

---

*This paper was presented at the NIK-2010 conference; see <http://www.nik.no/>.*

environments can be highly realistic, objects displayed in virtual environment are different from their real world counterparts i.e., objects in 3-D virtual environments have no physical depth. When optimizing virtual environments it is important to be able to measure the user's perceived depth of an object and correct for any discrepancy between the measured value and that specified by the environment's designers. This need presents us with the challenge of measuring a non-physical quantity namely: perceived depth.

In real or 3-D virtual environments, the two eyes view two different images of the same scene and the brain fuses these images to give a perception of depth. Depth perceived in virtual environment can be reported by the observer verbally. The experimental evidence provided by Waller [11] states that verbal feedback from the observer improves the accuracy of the depth in virtual environments. However, in the absence of verbal feedback, we can estimate depth by other means, for instance, eye tracking.

In personal space(see section 2), eyes converge to focus on the object at a certain depth plane. It can be compared to a two camera system viewing a real world scene, giving us two images, where a point in the first image has a corresponding point in the second image. In photogrammetry the 3-D location of the correspondence point can be calculated from orientation, focal length and location of the cameras. The solution relies on the concepts of epipolar correspondence, for details on epipolar geometry see Zhang [12], for details on triangulation see Hartley & Sturm [2].

Unlike the two camera system we cannot get the images captured by the two eyes for correspondence. However, we can measure the point of focus of the eye on the 2-D display, that is called fixation, using an eye tracker. Knowing the physical location of eyes, the intersection of lines connecting the left and right eyes to their fixations, extended behind the display can give us the estimated depth.

Estimated depth is calculated by intersection of two lines in 3-D. However, experimental data shows that these lines do not intersect. In this paper, the method used to resolve this issue is elaborated and its performance is measured.

The cues that influence our depth perception in both real and virtual world can be classified as binocular and monocular. Binocular cues are: accommodation, disparity and vergence. Monocular cues are: shading, shadow, linear perspective, relative height, relative size, texture gradient, and motion perspective [9, 4, 1].

## 2 Depth Perception Cues

Table 1 gives definitions of cues for depth perception. The effectiveness of the above mentioned cues varies with space. The space around the observer can be divided into three egocentric regions: personal space, action space, and vista space [1]. Personal space is the zone surrounding the observer's head, within an arms reach( $\approx 1$  m). Action space is the circular region beyond the personal space and extending upto 30 meter. Vista space is the region beyond 30 meter. Cues that are effective in personal space are: occlusion, binocular disparity, relative size, vergence, accommodation, and motion perspective [1].

### Depth Estimation using Vergence

Vergence, the simultaneous movement of eyes in opposite directions gives us precise depth perception. In virtual environment we cannot track the 3-D gaze behind the

Table 1: Cues for depth perception

Cue	Definition
Accommodation	Ciliary muscles adjust the curvature of the lens, and hence its refractive power, to bring images of objects at a particular distance into clear focus [4, 3].
Aerial Perspective	It is determined by the relative amount of moisture, pollutants, or both in atmosphere through which one looks at a scene. When air contains high degree of either, objects in the distance becomes bluer, decreased in contrast, or both with respect to objects in foreground [1].
Binocular Disparity	Eyes are about 6.5 cm apart, which gives two vantage points. This causes the optic arrays and images of 3-D object to differ in two eyes [5].
Linear Perspective	It combines different cues like relative size, relative height, and texture gradient, Parallel lines that recede into the distance appear to converge [1].
Motion Perspective	Relative motion of images of the object points at different distances that is caused by motion of the observer or of the object points [5].
Occlusion	When one object hides, or partially hides, another from the view. This cue offers information on depth order but not about the amount of depth [1, 3].
Relative Size	Size of any 2-D or 3-D object lying at a fixed angle to line of sight varies inversely with distance of the object along that line of sight [5].
Relative Height	Objects farther from the ground appear to be far as compared to objects near the ground [5].
Shading	Variations in the irradiance from surface due to changes in the orientation of the surface to incident light or variations in specularly [5].
Shadow	Variations in the irradiance from surface caused by obstruction by an opaque or semi-opaque object [5].
Texture Gradient	Images of textured elements become more densely spaced with increasing distance along the surface [5].
Vergence	Movement of eyes through equal angles in opposite directions to produce a disjunctive movement. Horizontal vergence occurs when a person changes fixation from an object in one depth plane to one in another depth plane [4].

display, so the problem becomes estimation of 3-D fixations based on the geometry of two 2-D images of the virtual environment. 3-D fixation can be calculated if the observer is looking at the virtual object in the scene. Observer was instructed to look at the object during the experiment.

Figure 1 shows the scheme for estimation of depth based on vergence. In symmetrical convergence, the angle of horizontal vergence,  $\phi$  is related to the interocular distance,  $a$ , and distance of the point of fixation,  $d$ , as in the following expression [4],

$$\tan(\phi/2) = \frac{a}{2d} \quad (1)$$

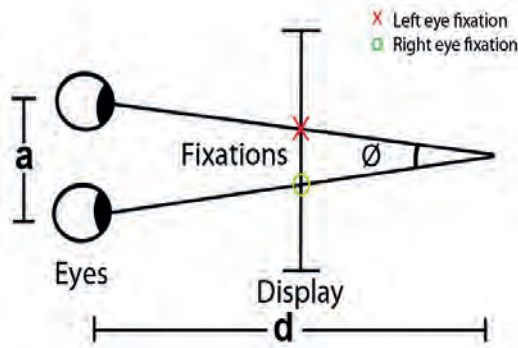


Figure 1: Vergence

Fixation marked by cross correspond to the left eye and fixation marked by circle correspond to the right eye. Eye tracking cameras measure the fixations with respect to the display screen. The lines from left and right eyes passing through the fixations are extended behind the display. The intersection obtained by these lines is the 3-D fixation [8]. The intersection of two lines in space is a trivial problem, However, in the presence of noise these lines do not intersect. Therefore, we should find a solution to intersect the lines.

Assuming two 3-D line segments  $P_1P_2$  and  $P_3P_4$  are joined by the shortest line segment  $P_aP_b$ .  $P_m$  is the mid point of the shortest line between two lines, as shown in figure 2. A point  $P_a$  on the line  $P_1P_2$  and point  $P_b$  on the line  $P_3P_4$  is given by the following line equations:

$$P_a = P_1 + \mu(P_2 - P_1) \quad (2)$$

$$P_b = P_3 + \eta(P_4 - P_3) \quad (3)$$

Truong et al. [10] states that the shortest line between the two lines can be found by minimizing  $|P_b - P_a|$ .  $\eta$  and  $\mu$  can be arbitrary real numbers.

$$P_b - P_a = P_3 - P_1 + \eta(P_4 - P_3) - \mu(P_2 - P_1) \quad (4)$$

$P_aP_b$  is the shortest line segment between two lines, so it should be perpendicular to the two lines  $P_1P_2$  and  $P_3P_4$ . Hence their dot product is zero.

$$(P_b - P_a) \cdot (P_2 - P_1) = 0 \quad (5)$$

$$(P_b - P_a) \cdot (P_4 - P_3) = 0 \quad (6)$$

Using equations 4-6 we get

$$[P_3 - P_1 + \eta(P_4 - P_3) - \mu(P_2 - P_1)] \cdot (P_2 - P_1) = 0 \quad (7)$$

$$[P_3 - P_1 + \eta(P_4 - P_3) - \mu(P_2 - P_1)] \cdot (P_4 - P_3) = 0 \quad (8)$$

Expanding equations 7, 8 in (x,y,z) gives  $\mu$  and  $\eta$ . After calculating  $P_a$  and  $P_b$ . We get  $P_m$ , the mid point of shortest line by the following equation,

$$P_m = (P_a + P_b)/2 \quad (9)$$

$P_m$  is the estimated 3-D fixation and the z-component of euclidean point  $P_m$  gives us the estimated depth.

An experiment was designed for testing the personal space, the details of which are discussed in section 3.

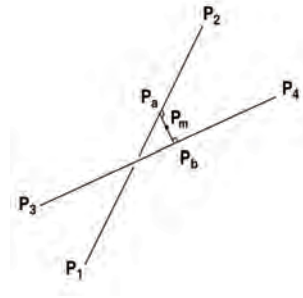


Figure 2: 3-D Intersection

### 3 Experiment

Figure 3(a) shows the 3-D virtual scene created by using Coin3d [6] library. Figure 3(b) shows side view of the same scene. The dimensions of the scene are 472\*296\*400 millimeters (width\*height\*depth). In each virtual scene a spherical object is displayed at a specific location in front of the checkerboard pattern. The pattern is 400 mm behind the display. Sphere is moved to different locations in 3-D to give 27 virtual scenes. These virtual scenes are shown to the observer, one at a time. Observer is at a distance of 600 mm from the display. Since, the majority of a person's normal range of vergence is used within one meter from the eyes [4], the sum of distance from the eye and maximum extent of virtual scene is fixed at (600+400) 1000 mm.

A real world model similar to the virtual scene was created and digital camera was placed at a distance of 600 mm from the model. Digital images of the real world model were used as a reference to accurately position the spheres in the virtual environment similar to Pfeiffer et al. [8].

Figure 4(a) shows the observer wearing NVidia 3D Glasses, Arrington Research's eye tracking cameras are mounted below the glasses. Figure 4(b) shows the experiment setup, head of the observer is fixed by using a chin rest. Samsung 2233RZ 3d display and NVidia Quadro FX 3800 graphics card are used for presenting the 3-D virtual scene.

5 Observers with no prior experience of 3-D environments performed the experiment. Mean age of the group was 38 years, written consent was taken for using the eye data. The experiment was performed in the following steps: First, the observer's head is fixed by using a chin rest such that the distance of the eyes from display is 600 mm and the distance between the eyes is measured. Second, the eyes of the observer are calibrated to the display using a standard calibration procedure. In this procedure, the observer looks at a series of 16 points and eye tracker records the value that corresponds to each gaze point. Third, observer is shown the virtual scene and after viewing the scene, observer reports the depth of the sphere. This



procedure is followed for all 27 virtual scenes. As a reference measure, the depth of the first sphere is told to the observer. Observer controls the switching of the virtual scenes by pressing a key. The task of finding depth of the object forces the observer to maintain the gaze on virtual object and thus maximizing the number of fixations on the virtual object. The results from the experiment are discussed in the next section.

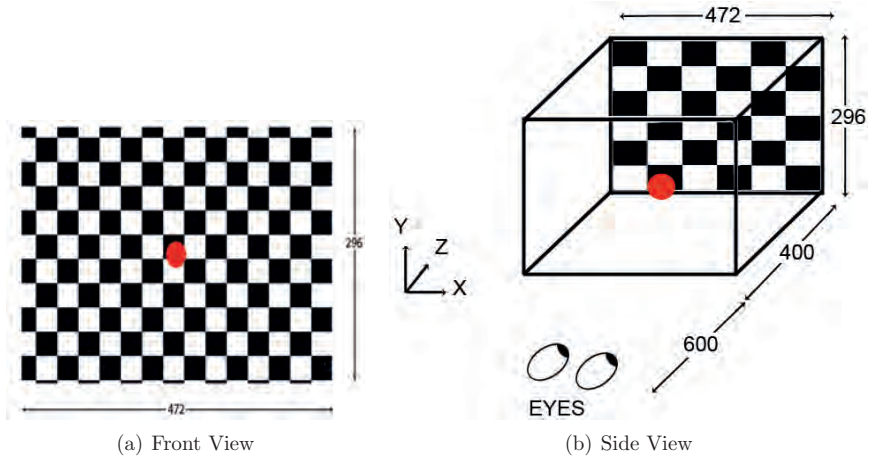


Figure 3: Virtual Scene



Figure 4: Experiment Setup

## 4 Results

27 virtual scenes were shown to 5 observers. Mean estimated depth(MED) for an observer and a virtual scene is calculated by the following expression,

$$MED_s = \frac{\sum_{j=1}^N G(j)s}{N} \quad (10)$$

$N$  is the number of estimated fixations,  $G$  is the 3-D estimated fixation for a virtual scene,  $S$  is virtual scene number.

Figure 5(a) shows the estimated 3-D fixations obtained by the geometric method discussed in section 2. The actual location of the object is represented by the red sphere, estimated gaze is represented by asterisk. Vergence eye movements result in a number of depth estimates. The histogram of the depth estimates (in figure 5(b)) shows that mean value of depth estimates lies around -61 mm. Object is at a depth of -100 mm, as specified by the design of virtual environment. So, there is difference of 39 mm in the mean of estimated depth and actual position of depth. Figure 5(a) shows the volume of data over which the mean is calculated. Noise in the data can be contributed by a number of factors: inaccurate measurement of distance between the eyes, device errors, slight head movements, inaccurate design of environment. In addition to noise there are external factors, for instance, when observer looks at parts of the scene, which do not contain the spherical object. A filter is implemented to reduce such isolated 3-D fixations, as a result of noise and external factors. It is assumed that observer spends most of the time looking at the virtual object. Considering this, the region with maximum number of the estimated fixations corresponds to the location of virtual object.

The filter operates as follows: First, the filter divides the virtual space into cuboids of equal size and records the population of data for each cuboid region. Second, the cuboid region with maximum population size is selected. Third, the cuboid regions with size more than half the maximum size are selected and their mean is calculated which gives us the estimated depth. Figure 6(a) shows the distribution of the estimated gaze after filtering. Histogram of the filtered data now lies around -64 mm. It represents one of the typical cases, the improvement in accuracy is considerable as discussed in next section.

## Comparison of Performance of Geometric Depth Estimates

The performance of the estimated depth can be measured by comparing it with the depth specified by the virtual environment. Mean absolute depth error(MDE) is calculated by subtracting the depth of the object specified by the design of virtual environment from the mean of the estimated depth as follows,

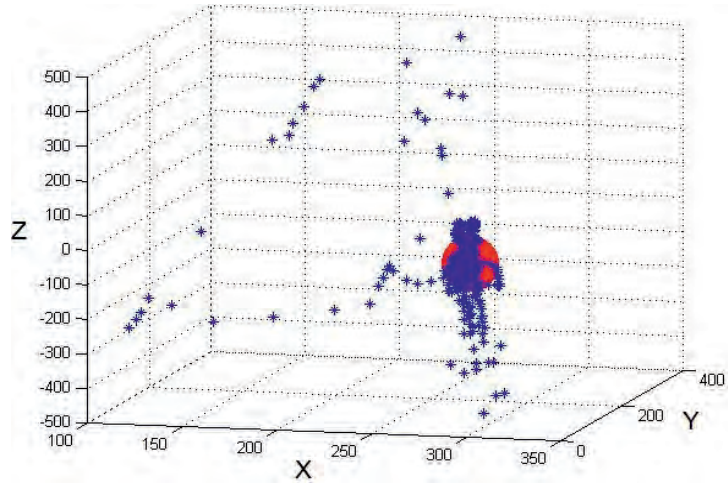
$$DE_S = MED_S - AD_S \quad (11)$$

$S$  is the virtual scene number,  $AD$  is the euclidean  $Z$  value of virtual object defined by the environment. Depth errors(DE) corresponding to all the virtual scenes are calculated. Small depth error indicates a good correspondence between the estimated depth and the depth specified by the design of the virtual environment, whereas, a large depth error indicates otherwise.

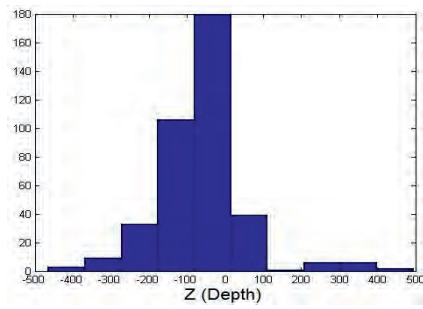
Figure 7- 11 show the absolute depth errors for filtered data and unfiltered data. The x-axis represents the virtual scene number. Clearly on the average, the depth errors are reduced for the filtered data as compared to the unfiltered data. Thus, histogram based filtering reduces the depth errors.

$$MDE = \frac{\sum_{S=1}^{27} |(DE)_S|}{27} \quad (12)$$

MDE gives the absolute error between the actual  $z$  position of the object and estimated  $z$  position for all 27 scenes. Table 2 shows the mean depth errors for 5



(a) Estimated Gaze



(b) Histogram of Depth(Z)

Figure 5: Estimated Depth

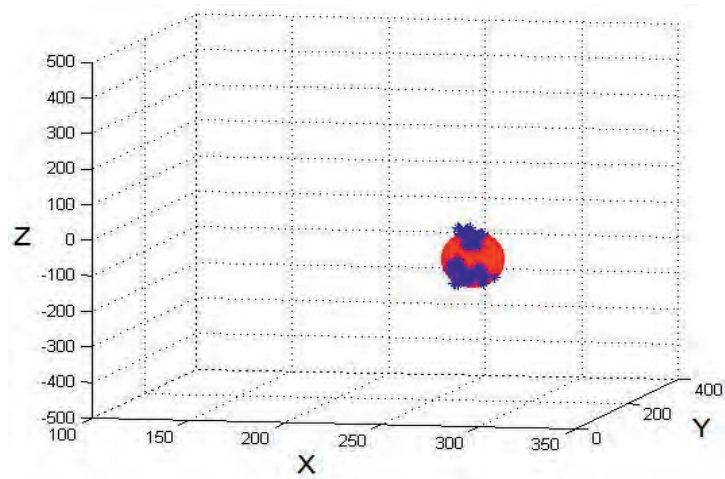
observers. Histogram filtered data clearly reduces the depth errors, hence improving the depth estimates for vergence.

Table 2: Comparison of Mean Depth Errors

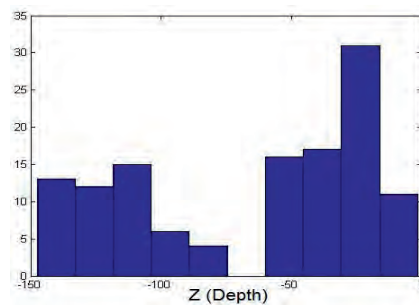
Observer	MDE(in mm) of Complete Data	MDE(in mm) of Histogram Filtered Data
1	127.4	107.9
2	224.7	102.9
3	114.6	69.4
4	154.7	127.1
5	131.6	104.8

## 5 Conclusions

Depth estimation via vergence for a virtual environment is possible, given that the virtual environment is designed within the range of the personal space. The depth estimate is calculated by taking the mean of estimated 3-D fixations. The



(a) Estimated Gaze



(b) Histogram of Depth(Z)

Figure 6: Estimated Depth after filtering

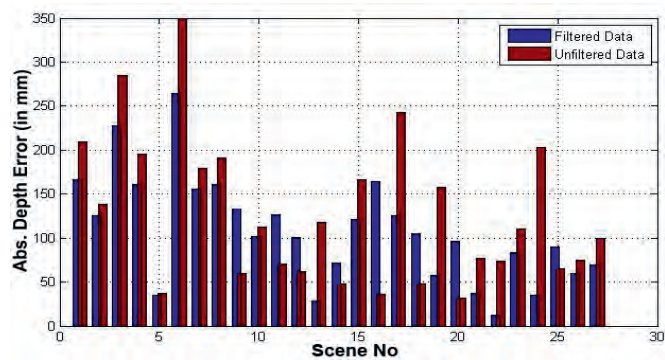


Figure 7: Absolute Depth Error for Observer 1

results obtained from the evaluation of the geometric depth estimation algorithm are discussed in section 4, Results in table 2 show that histogram based filtering

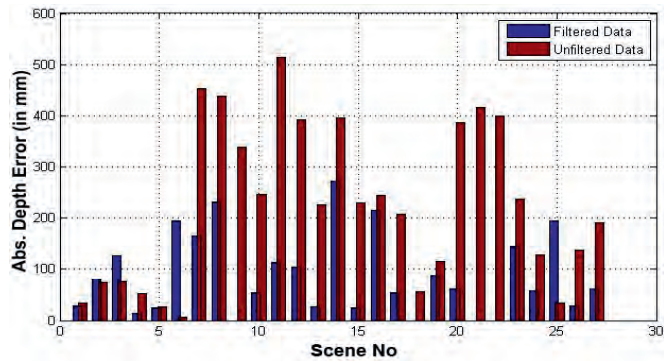


Figure 8: Absolute Depth Error for Observer 2

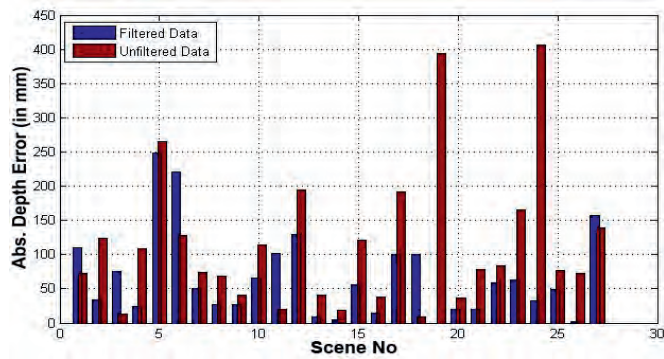


Figure 9: Absolute Depth Error for Observer 3

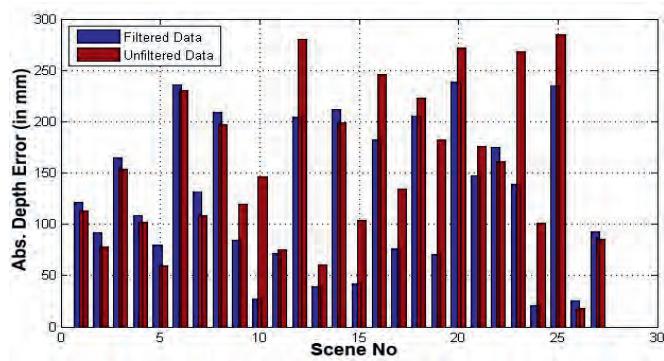


Figure 10: Absolute Depth Error for Observer 4

improves the performance of the depth estimates. Mon-Williams et al. [7] suggests that stereoscopic depth can be estimated by a combined signal provided by disparity and vergence with weighting attached to either varying as a function of availability. In future, we intend to investigate this issue.



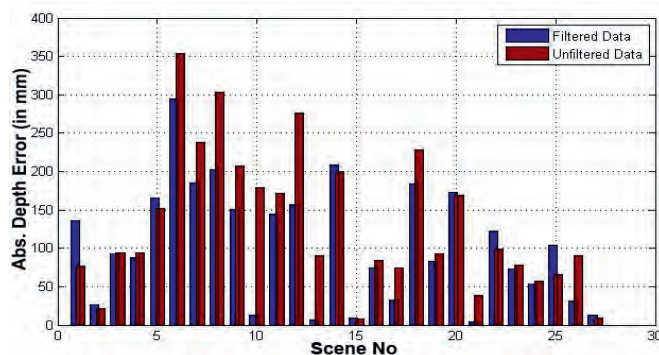


Figure 11: Absolute Depth Error for Observer 5

## 6 Acknowledgments

Authors wish to thank Ali Alsam, HiST for his tremendous support. We are grateful to Thies Pfeiffer, Bielefeld University and Andrew T. Duchowski, Clemson University for their valuable suggestions. Last but not the least, we would like to thank our colleagues at Department of Informatics & E-Learning (AITeL), Sør Trøndelag University College(HiST).

## References

- [1] James E. Cutting and Peter M. Vishton. *Perceiving layout: The integration, relative dominance, and contextual use of different information about depth*, volume 5, chapter 3, pages 69–117. New York: Academic Press, 1995.
- [2] Richard I. Hartley and Peter Sturm. Triangulation. In *Proceedings of ARPA Image Understanding Workshop*, pages 957–966, 1994.
- [3] Maurice Hershenson. *Visual Space Perception : A Primer*. The MIT Press, 2000.
- [4] Ian P. Howard. *Seeing in Depth : Volume 1 Basic Mechanisms*. I Porteous, Toronto, 2002.
- [5] Ian P. Howard and Brian J. Rogers. *Seeing in Depth: Volume 2 Depth Perception*. I Porteous, Toronto, 2002.
- [6] <http://www.coin3d.org/> (Last Visited on 21-05-2010).
- [7] Mark Mon-Williams, James R. Tresilian, and Andrew Roberts. Vergence provides veridical depth perception from horizontal retinal image disparities. *Exp Brain Res*, 133(3):407–413, 2000.
- [8] Thies Pfeiffer, Marc E. Latoschik, and Ipke Wachsmuth. Evaluation of binocular eye trackers and algorithms for 3d gaze interaction in virtual reality environments. *Journal of Virtual Reality and Broadcasting*, 5(16), December 2008. urn:nbn:de:0009-6-16605, ISSN 1860-2037.

- [9] R. Troy Surdick, Elizabeth T. Davis, Robert A. King, and Larry F. Hodges. The perception of distance in simulated visual displays: A comparison of the effectiveness and accuracy of multiple depth cues across viewing distances. *Presence*, 6(5):513–531, October 1997.
- [10] Hung Q. Truong, Sukhan Lee, and Seok-Woo Jang. Model-based recognition of 3d objects using intersecting lines. *Multisensor Fusion and Integration for Intelligent Systems*, 35:289–300, 2009.
- [11] David Waller. Factors affecting the perception of interobject distances in virtual environments. *Presence*, 8(6):657–670, 1999.
- [12] Zhengyou Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27(2):161–198, 1998.

## **A.10 Estimating the depth in three-dimensional virtual environment with feedback.**

**Authors:** Puneet Sharma and Ali Alsam.

**Full title:** Estimating the depth in three-dimensional virtual environment with feedback.

**Published in:** Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012), ACTA Press.



# ESTIMATING THE DEPTH IN THREE-DIMENSIONAL VIRTUAL ENVIRONMENT WITH FEEDBACK

Puneet Sharma and Ali Alsam  
Department of Informatics & e-Learning(ATeL)  
Sør-Trøndelag University College(HiST)  
Trondheim, Norway  
email: er.puneetsharma@gmail.com

## ABSTRACT

Visual interaction in 3-D virtual space can be achieved by estimating objects depth from the fixations of the left and right eyes. Training a PSOM neural network to estimate depth, from eye fixations, has been shown to result in good level of accuracy. Instead of training a neural network we postulate that it is possible to improve the accuracy of the fixation data by providing the observer with feedback. In order to test this hypothesis we introduce a closed-loop feedback in the environment. When the user's visual axes intersect, within a range of the correct depth, a sound is produced. This mechanism trains the users to correct their fixations in a fashion that results in improved depth estimation. Our results show that indeed the accuracy of depth estimation improves in the presence of feedback.

## KEY WORDS

Eye fixations, depth estimation, virtual environment

## 1 Introduction

Our perception of the layout of the world around is three-dimensional. The eyes represent the centroid of our perceived world with objects scattered to their left, right, nearer or farther away from them. From a computer vision point of view, the mechanism which enables us to see in three-dimensions can be explained by means of stereovision [9]. The basic idea is that the images formed on the retinas of the left and right eyes represent two different three-dimensional planes that are merged into a three-dimensional scene based on the principles of epipolar geometry [17, 6, 2].

Research in human vision shows, however, that the explanation provided by the epipolar geometry is only part of a more complex perception-mechanism. Indeed we can simply verify that the world appears three dimensional even when one eye is shut—a fact that is readily used in fine art and visual illusions [9]. Extensive research in layout perception indicates that our vision system makes use of a wealth of information sources which are fused to render the final perception. Among these sources, or cues, are: accommodation, aerial perspective, binocular disparity, convergence, height in visual field, motion perspective, occlusion, shading, shadow, relative size, and relative den-

sity [8, 9, 3, 13, 16].

When designing a three-dimensional virtual environment, it's important that the resultant layout is realistic. It is, however, implausible to incorporate all the visual cues into the design. Assuming that there are fifteen cues [3], there would be 105 possible pairs of information sources to take into account, 455 possible triples and 1365 possible quadruples, not to mention higher order combinations [3]. Clearly, no realistic design process can take such a high order of variables into account.

Accurate depth perception in virtual environments would enable users to visually interact with objects embedded therein [14]. By visual interaction we mean that a match between the three-dimensional coordinates of a human fixation point and those of an object in the environment would trigger a predefined action. Here, we envisage a scenario where the user's eye movements are recorded using a calibrated high frequency eye-tracker. The question that we need to answer is whether the perceived depth can accurately be estimated from the user's eye locations. A number of researchers [4, 5, 11, 12, 1] have endeavored to answer this question. The basic method employed is based on the assumption that the lines emerging from the centers of the two eyes to the fixation points on the screen, as recorded by the eye-tracker, intersect at the perceived depth. In other words, it is assumed that convergence is sufficient to estimate depth. Unfortunately, this assumption suffers from a number of drawbacks. Firstly, the empirical lines defined by the centers of the eyes and the fixation points almost never intersect, thus, some optimization method such as the shortest distance between the lines is normally employed [15]. The second problem is more fundamental in that the assumption that the intersection provides an accurate depth does not incorporate any of the aforementioned visual cues. Some level of accuracy has been achieved by employing a POSM neural network that is trained to the individual user [5].

In this paper, we take a novel approach to the estimation of perceived depth in virtual environments. Specifically, we postulate that users can be trained to move their eyes in a fashion that would result in an accurate depth estimation based on the line-intersection method [15]. This is done by providing the user with a compensatory cue which is assumed to compensate for the lack of visual cues avail-

able in natural settings. To test the effectiveness of the compensatory cue we constructed a simple three-dimensional virtual environment with a checkerboard background and a spherical object that was located at different depth values ranging from 50 to 300mm behind the screen. Users were provided with shutter glasses and their eye movements were recorded with an eye-tracker. Furthermore, we calculated the depth estimated by the intersecting lines method in real-time. When the correct depth was estimated by the model, a sound was played by the system. Thus, in this experiment, the compensatory cue was audible rather than visual. The choice of an audible cue is motivated by the thought that providing a visual cue such as a change in the color or intensity of the object would alter the settings of the environment in an unpredictable fashion. Our experiments show that incorporating a compensatory cue does indeed result in a significantly improved depth estimation. In fact, we observed that even when slight head movements were allowed users could quickly train their eyes to fixate on the region of the scene associated with the sound cue.

## 2 Theory

### 2.1 Line-intersection method

In the line-intersection method, three-dimensional fixation is estimated using triangulation [7]. Two lines are defined as originating from the left and right eyes, passing through their respective fixation points, and extending into infinity. The intersection of two lines in space is a well defined problem where the solution is obtained by solving the simultaneous equations that describe the lines. Due to noise, however, the lines defined based on real data do not intersect. Thus, to estimate a representation of the intersection point, a cost function is defined and the estimation is obtained by optimization. In previous studies [5, 12], the optimization problem was defined as the search for a point with minimum Euclidian distance to the two fixation lines. Geometrically, there is a unique shortest line segment that joins two lines in three dimensions [15]. Thus the mid-point of the shortest line is assumed to represent the best estimate of the three-dimensional fixation.

Figure 1 represents the top and side views of a scenario where the lines do not intersect. Assuming that the shortest line segment that joins the two lines in three dimensions is  $\vec{G}_l\vec{G}_r$ , as shown in figure 1(b). A point  $\vec{G}_l$  on line  $\vec{P}_l\vec{F}_l$  and a point  $\vec{G}_r$  on line  $\vec{P}_r\vec{F}_r$  can be defined by the following line equations:

$$\vec{G}_l = \vec{P}_l + \mu(\vec{F}_l - \vec{P}_l) \quad (1)$$

$$\vec{G}_r = \vec{P}_r + \eta(\vec{F}_r - \vec{P}_r) \quad (2)$$

where  $\vec{P}_l$  and  $\vec{P}_r$  are left and right eye locations,  $\vec{F}_l$  and  $\vec{F}_r$  are left and right eye fixations on the display, and  $\eta$  and  $\mu$  are arbitrary real numbers. The shortest distance between

the two lines can be calculated by minimizing  $|\vec{G}_l - \vec{G}_r|$  as,

$$|\vec{G}_l - \vec{G}_r| = |\vec{P}_l - \vec{P}_r - \eta(\vec{F}_r - \vec{P}_r) + \mu(\vec{F}_l - \vec{P}_l)| \quad (3)$$

The equations 1- 3 can be solved for  $\vec{G}_l$  and  $\vec{G}_r$ , the points on both visual axes. Mid-point of the line segment  $\vec{G}_l\vec{G}_r$  is assumed as the three-dimensional fixation  $\vec{F}$ .

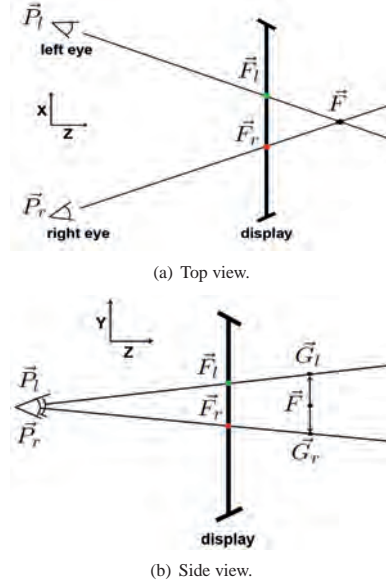


Figure 1. When the two lines do not intersect mid-point of the shortest line is assumed to represent the best estimate of 3-D fixation;  $\vec{F}$  is the mid-point of shortest line segment between the two visual axes.

## 3 Experiment

In this study, observer's left and right eye fixations were recorded by using Arrington Research's eye-tracker. Shutter glasses for viewing the three-dimensional scene, were mounted above the eye tracker as shown in figure 2(a). Figure 2(b) shows the side view of the experiment setup. Head movements were minimized using a chin rest. Samsung 2233RZ 3D display and NVidia Quadro FX 3800 graphics card were used for presenting the three-dimensional virtual scene.

### 3.1 Three-dimensional virtual scene

The three dimensional scene was created by using Coin3d [10] graphics library. The dimensions of the scene were 472\*296\*400 millimeters (width\*height\*depth). Figure 3 shows the front and side views of the three-dimensional virtual scene. In the virtual scene, a spherical

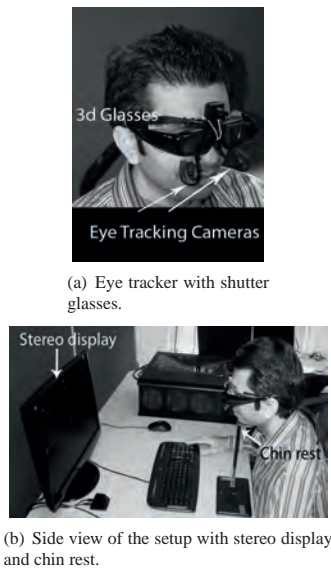


Figure 2. Experiment setup with eye tracker, stereo display, and chin rest.

object was displayed at different depths(-200,-100,-300,-50,-150,-250 mm) with a checkerboard background. The checkerboard background was 400 mm behind the display.

To create a realistic virtual scene, an identical real world model was constructed and a digital camera was used to image a spherical object at different depth values. The camera was placed at a distance of 600 mm, which is the same distance as that of the observer in the experiment setup. Using the digital images of the real world model, a scale measure was computed and later used in the design of virtual scene. The procedure followed in this study is in keeping with the method described by Pfeiffer et al. [12].

## 4 Results

To test our hypothesis that the introduction of a compensatory cue improves observers' estimated depth perception two experiments were performed. In the first, the observers viewed the scene without a compensatory cue. In the second experiment, the audible cue was included into the environment. The observers who performed the first experiment were instructed to fixate on the spherical objects. In the second experiment, the observers were presented with an identical scene. The instructions were, however, different. In this case, the observers were informed that maneuvering their eyes, as they gaze at the object, could produce a sound. They were, further, instructed to try and prolong the duration of the sound. As mentioned in the introduction section, the sound was produced when the correct depth, within some error range, was estimated by the

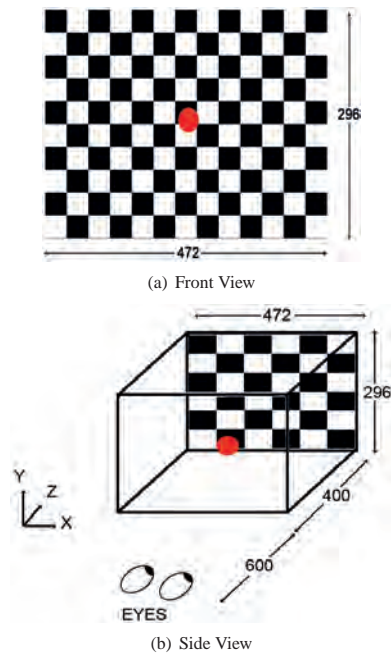


Figure 3. Front and side view of the virtual scene with the checkerboard background. The dimensions of the scene are 472x296x400 mm<sup>3</sup>. Distance between the observer and the display is 600 mm.

line-intersection method based on the eye fixation data obtained by the eye-tracking system. In keeping with Cutting & Vishton [3] we define accurate perceived depth as being within 15 percent of the actual depth. Three observers took part in the experiments which were separated by a period of two weeks. The two weeks period was introduced to avoid the possibility of observer's adaptability to the virtual environment.

### 4.1 Average depth for line-intersection

For each object in the environment, the fixation-data obtained from an observer were used to estimate the associated depth values. Given that our eyes are in constant movement, the estimated depth values represent a distribution that varies in time. An example of the depth distribution for a single object is shown in figure 4. In this case, the actual depth of the object, as specified in the environment, is 150mm behind the screen. We notice, however, that the depth obtained from the fixation data varies from zero, which is the plane of the screen to 300mm behind it indicating that the observer is continuously browsing the scene. Knowing that this browsing mechanism is a natural aspect of the our vision system we expect that the depth perception data obtained from any given method

would exhibit variations. Thus, to judge the goodness of a method compared to another we need to examine the local rather than the global statistics of the distribution. Having said that, we start our comparisons by considering the average values obtained from the estimated depth data with and without the compensatory cue based on the whole distribution. Tables 1- 3 show the results of the two experiments for three different observers. In case of the experiment performed without compensatory cue, the average values of the distribution for the line-intersection method exhibit little accuracy. In case of data obtained with compensatory cue we notice, however, that the average values of the distributions offer clearer discrimination making it better suited for visual interaction.

Table 1. Average depths for the first observer obtained from line-intersection(LI) method for two experiments: without compensatory cue, and with compensatory cue. All units are in millimeters

Object no.	Actual depth	Without compensatory cue	With compensatory cue
1	-200	-81	-162
2	-100	-151	-99
3	-300	-50	-194
4	-50	-61	-87
5	-150	-56	-151
6	-250	-66	-136

Table 2. Average depth for the second observer obtained from line-intersection(LI) method for two experiments: without compensatory cue, and with compensatory cue. All units are in millimeters

Object no.	Actual depth	Without compensatory cue	With compensatory cue
1	-200	-212	-239
2	-100	-264	-220
3	-300	-292	-219
4	-50	-297	-95
5	-150	-291	-139
6	-250	-275	-174

Table 3. Average depth for the third observer obtained from line-intersection(LI) method for two experiments: without compensatory cue, and with compensatory cue. All units are in millimeters

Object no.	Actual depth	Without compensatory cue	With compensatory cue
1	-200	-162	-179
2	-100	-170	-133
3	-300	-158	-162
4	-50	-97	-69
5	-150	-82	-159
6	-250	-69	-204

#### 4.2 Variation of local means over time for line-intersection method

To examine the local variations in the data we sub-sampled the distribution into twenty regions. For each sub-sample we calculated the average values of the depth obtained by employing the line-intersection method. Figures 5- 7 show the variation over time of the local average values for different depths -50, -200 and -150mm. From these figures we notice that the introduction of the compensatory cue is indeed improving the estimated depth over time. Further, the comparison of the histograms, figures 5b- 7b, for the two experiments reflects that the introduction of the compensatory cue results in a higher frequency of depth estimates that are in the vicinity of the actual depth. Similar results were obtained for the other depth values.

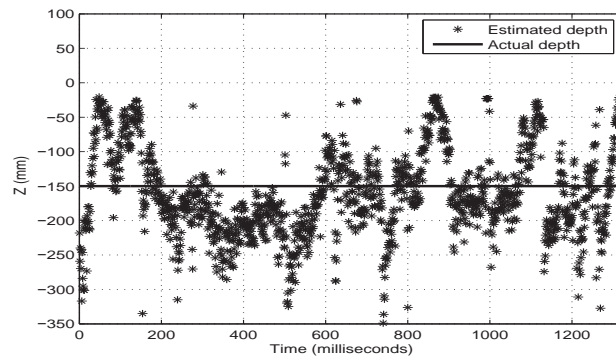
## 5 Conclusion

In this paper, we show that the introduction of a closed loop feedback in the form of a compensatory cue improves the estimation of perceived depth in virtual environments. The depth range used in the experiment varied from 50 to 300mm behind the screen. This range corresponds to the users personal space which is believed to be the range in which convergence is a significant cue. Furthermore, we included an audible cue into the design of the environment. The audible cue was provoked when the fixation-data obtained from the eye tracker resulted in a depth estimate that was within a predefined error value. Here the calculations were based on the line-intersection method. Our intuition in the design of the experiment was that providing the observers with feedback would stimulate them to correct their fixations in a manner that improves the obtained depth values. Our results show that indeed the estimated depth in the presence of the compensatory cue represents a clear improvement. Here we underline that improving the depth estimation allows visual interaction with the virtual environment. Thus our goal in the experiment was not to improve perceived depth but rather to improve the estimation of depth in a fashion that results in improved interaction.

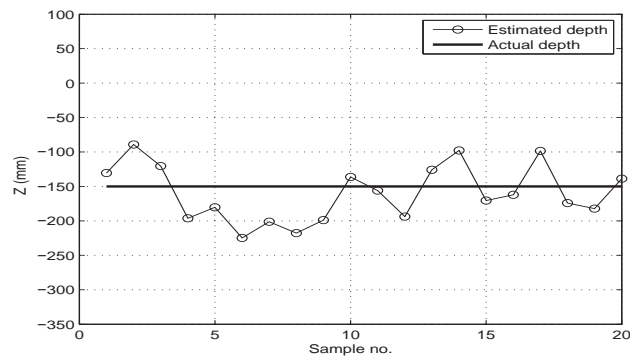
## References

- [1] Gunnar Blohm, Aarlenne Z. Khan, Lei Ren, Kai M. Schreiber, and J. Douglas Crawford. Depth estimation from retinal disparity requires eye and head orientation signals. *Journal of Vision*, 8(16):1–23, December 2008.
- [2] Roberto Cipolla and Peter Giblin. *Visual motion of curves and surfaces*. Cambridge University Press, 2000.
- [3] James E. Cutting and Peter M. Vishton. *Perceiving layout: The integration, relative dominance, and contextual use of different information about depth*, vol-

- ume 5, chapter 3, pages 69–117. New York: Academic Press, 1995.
- [4] Andrew Duchowski, Eric Medlin, Nathan Cournia, Hunter Murphy, Anand Gramopadhye, Santosh Nair, Jeenal Vorah, and Brian Melloy. 3d eye movement analysis. *Behavior Research Methods, Instruments, and Computers (BRMIC)*, 34(4):573–591, 2002.
- [5] Kai Essig, Marc Pomplin, and Helge Ritter. A neural network for 3d gaze recording with binocular eye trackers. *The International Journal of Parallel, Emergent and Distributed Systems*, 21(2):79–95, April 2006.
- [6] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2003.
- [7] Richard I. Hartley and Peter Sturm. Triangulation. In *Proceedings of ARPA Image Understanding Workshop*, pages 957–966, 1994.
- [8] Ian P. Howard. *Seeing in Depth : Volume 1 Basic Mechanisms*. I Porteous, Toronto, 2002.
- [9] Ian P. Howard and Brian J. Rogers. *Seeing in Depth: Volume 2 Depth Perception*. I Porteous, Toronto, 2002.
- [10] <http://www.coin3d.org/> (Last Visited on 15-02-2012).
- [11] Yong-Moo Kwon, Kyeong-Won Jeon, Jeongseok Ki, Qonita M. Shahab, Sangwoo Jo, and Sung-Kyu Kim. 3d gaze estimation and interaction to stereo display. *The International Journal of Virtual Reality*, 5(3):41–45, 2006.
- [12] Thies Pfeiffer, Marc E. Latoschik, and Ipke Wachsmuth. Evaluation of binocular eye trackers and algorithms for 3d gaze interaction in virtual reality environments. *Journal of Virtual Reality and Broadcasting*, 5(16), December 2008. urn:nbn:de:0009-6-16605, ISSN 1860-2037.
- [13] R. Troy Surdick, Elizabeth T. Davis, Robert A. King, and Larry F. Hodges. The perception of distance in simulated visual displays: A comparison of the effectiveness and accuracy of multiple depth cues across viewing distances. *Presence*, 6(5):513–531, October 1997.
- [14] Vildan Tanriverdi and Robert J. K. Jacob. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '00*, pages 265–272, 2000.
- [15] Hung Q. Truong, Sukhan Lee, and Seok-Woo Jang. Model-based recognition of 3d objects using intersecting lines. *Multisensor Fusion and Integration for Intelligent Systems*, 35:289–300, 2009.
- [16] Bob G. Witmer and Paul B. Kline. Judging perceived and traversed distance in virtual environments. *Presence*, 7(2):144–167, March 1998.
- [17] Zhengyou Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27(2):161–198, 1998.

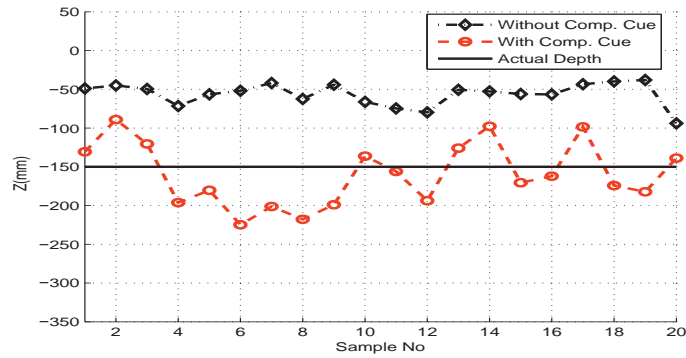


(a) Depth estimates obtained from left and right eye fixations by line-intersection method for an object 150mm behind the screen. The distribution of estimated depth varies from zero, which is the plane of the screen to 300mm behind it indicating that the observer is continuously browsing the scene.

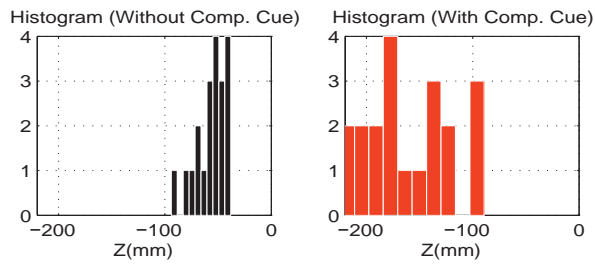


(b) Distribution of depth estimates for the sub-sampled data over twenty samples of the total time.

Figure 4. Distributions of estimated depth for raw data and sub-sampled data using line-intersection method.

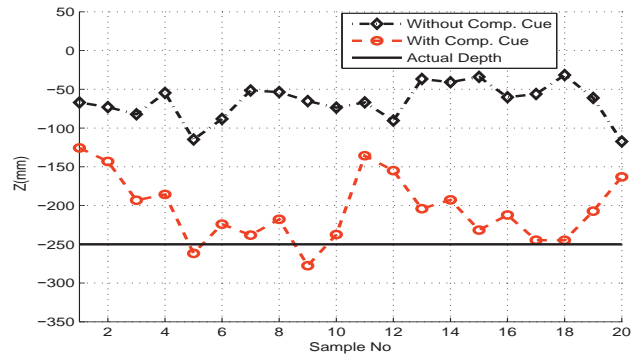


(a) Distributions of depth estimates for the sub-sampled data of two experiments over twenty samples of the total time. In the experiment with compensatory cue we see a clear convergence towards the actual depth of the object, that is 150 mm behind the screen.

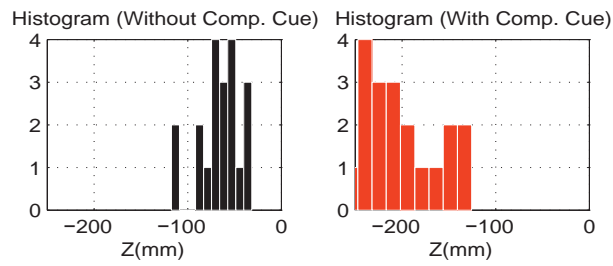


(b) Histograms of the sub-sampled data for two experiments.

Figure 5. Distributions and histograms of depth estimates for two experiments: without compensatory cue, and with compensatory cue. Depth estimates were calculated using the line-intersection method.



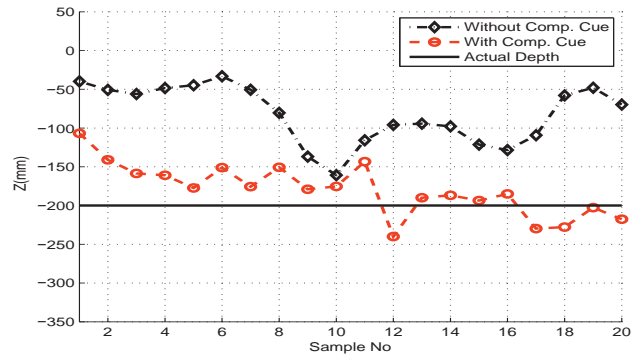
(a) Distributions of depth estimates for the sub-sampled data of two experiments over twenty samples of the total time. In the experiment with compensatory cue we see a clear convergence towards the actual depth of the object, that is 250 mm behind the screen.



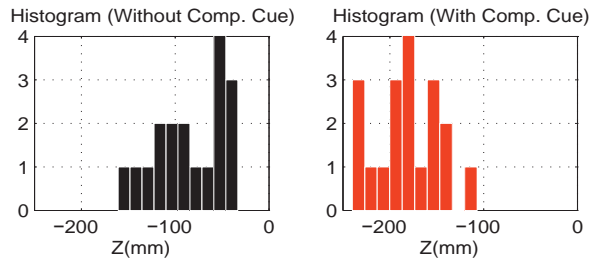
(b) Histograms of the sub-sampled data for two experiments.

Figure 6. Distributions and histograms of depth estimates for two experiments: without compensatory cue, and with compensatory cue. Depth estimates were calculated using the line-intersection method.





(a) Distributions of depth estimates for the sub-sampled data of two experiments over twenty samples of the total time. In the experiment with compensatory cue we see a clear convergence towards actual depth of the object, that is 200 mm behind the screen.



(b) Histograms of the sub-sampled data for two experiments.

Figure 7. Distributions and histograms of depth estimates for two experiments: without compensatory cue, and with compensatory cue. Depth estimates were calculated using the line-intersection method.

## **A.11 Estimating the depth uncertainty in three-dimensional virtual environment**

**Authors:** Puneet Sharma and Ali Alsam.

**Full title:** Estimating the depth uncertainty in three-dimensional virtual environment.

**Published in:** Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012), ACTA Press.

# ESTIMATING THE DEPTH UNCERTAINTY IN THREE-DIMENSIONAL VIRTUAL ENVIRONMENT

Puneet Sharma and Ali Alsam  
Department of Informatics & e-Learning(ATeL)  
Sør-Trøndelag University College(HiST)  
Trondheim, Norway  
email: er.puneetsharma@gmail.com

## ABSTRACT

Visual interaction in 3-D virtual space can be achieved by estimating objects depth from the fixations of the left and right eyes. The current depth estimation methods, however do not account for the presence of noise in the data. To address this problem we note that any measured fixation point is a member of a statistical distribution defined by the level of noise in the measurement. We thus propose a new numerical method that provides a range of depth values based on the uncertainty in the measured data.

## KEY WORDS

Eye fixations, depth estimation, virtual environment

## 1 Introduction

Our perception of the layout of the world around is three-dimensional. The eyes represent the centroid of our perceived world with objects scattered to their left, right, nearer or farther away from them. From a computer vision point of view, the mechanism which enables us to see in three-dimensions can be explained by means of stereovision [9]. The basic idea is that the images formed on the retinas of the left and right eyes represent two different three-dimensional planes that are merged into a three-dimensional scene based on the principles of epipolar geometry [15, 6].

Research in human vision shows, however, that the explanation provided by the epipolar geometry is only part of a more complex perception-mechanism. Indeed we can simply verify that the world appears three dimensional even when one eye is shut—a fact that is readily used in fine art and visual illusions [9]. Extensive research in layout perception indicates that our vision system makes use of a wealth of information sources which are fused to render the final perception. Among these sources, or cues, are: accommodation, aerial perspective, binocular disparity, convergence, height in visual field, motion perspective, occlusion, shading, shadow, relative size, and relative density [9, 3, 11, 7].

When designing a three-dimensional virtual environment, it's important that the resultant layout is realistic. It is, however, implausible to incorporate all the visual cues into the design. Assuming that there are fifteen cues [3],

there would be 105 possible pairs of information sources to take into account, 455 possible triples and 1365 possible quadruples, not to mention higher order combinations [3]. Clearly, no realistic design process can take such a high order of variables into account.

Accurate depth perception in virtual environments would enable users to visually interact with objects embedded therein [13]. By visual interaction we mean that a match between the three-dimensional coordinates of a human fixation point and those of an object in the environment would trigger a predefined action. Here, we envisage a scenario where the user's eye movements are recorded using a calibrated high frequency eye-tracker. The question that we need to answer is whether the perceived depth can accurately be estimated from the user's eye locations. A number of researchers [4, 5, 10, 12, 2] have endeavored to answer this question. The basic method employed is based on the assumption that the lines emerging from the centers of the two eyes to the fixation points on the screen, as recorded by the eye-tracker, intersect at the perceived depth. In other words, it is assumed that convergence is sufficient to estimate depth. Unfortunately, this assumption suffers from a number of drawbacks. Firstly, the empirical lines defined by the centers of the eyes and the fixation points almost never intersect see figure 1, thus, some optimization method such as the shortest distance between the lines is normally employed [14]. The second problem is more fundamental in that the assumption that the intersection provides an accurate depth does not incorporate any of the aforementioned visual cues. Some level of accuracy has been achieved by employing a PSOM neural network that is trained to the individual user [5].

In this paper, we propose a definition of a new numerical method which allows us to estimate the depth uncertainty that arises from non-intersecting visual lines. The lines emerging from the centers of the eyes to the fixation points on the screen define the normal vectors of the visual planes. Theoretically, these lines intersect at the depth of the fixation point [8]. Due to the presence of noise the lines obtained from real data do not intersect. Among the sources of noise in the measurements are: the accuracy of the measuring device, slight head movements, errors in measuring the exact distance between the eyes as well as biological factors such as adaptation. The existence of noise

in the measurements means that any optimization method would result in an approximation of the real fixation point rather than its exact location. In this paper we take a different approach—instead of estimating a unique point we estimate the depth range or uncertainty in the depth estimation. We start by considering a number of consecutive fixations, for each eye, these fixations define a distribution, of points, on the plane. Given that these points are close to each other, we assume that the distribution measured corresponds to noise in the data. This distribution together with the center of the eye, defines a volume in space. We note that assuming that there is no noise in the data, these volumes would converge to the visual lines. For real data, however, these volumes represent two cones. By intersecting the two cones, we arrive at an intersection region rather than a unique point. Thus we define the depth uncertainty as the length of the intersection volume in the  $z$ -direction. This approach offers a number of advantages: Firstly, given that the method incorporates a noise distribution it is more robust. Secondly, the depth uncertainty is a measure of the goodness of the estimate where the more uncertainty the less we would trust the depth value. Thirdly, in terms of optimization methods, the intersection volume represents the feasible solution space where any point is likely to be the actual depth. Thus the method allows us to study which statistical representation is most likely to represent the actual depth.

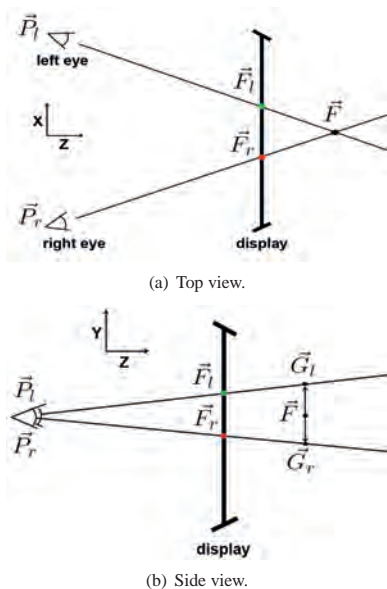


Figure 1. When the two lines do not intersect mid-point of the shortest line is assumed to represent the best estimate of 3-D fixation;  $\vec{F}$  is the mid-point of shortest line segment between the two visual axes.

## 2 Theory

### 2.1 Cone-intersection method

The estimation of uncertainty in measurements is of considerable importance in all engineering disciplines. In the case of perceived depth estimation, we wish to know the level of uncertainty associated with our estimate. To achieve this goal, we need to remember that the measurement obtained from the eye-tracker is an element in a statistical distribution. That is to say that if the measurement is repeated several times then the location of the fixation point will not be the same but rather it will form a cloud of points. This basic statistical knowledge is at the heart of the method proposed in this section.

We start with an assumption that the two fixation points obtained from the eye tracker are accurate—in this case the lines result in perfect intersection. On the other hand, if we were to consider that each point is a member of a statistical distribution, then the result is not two lines but rather a distribution of lines. The question that arises is: How can we intersect a distribution of lines?

The solution proposed in this paper is based on the assumption that the statistical distribution of the fixation points is convex. This assumption is motivated by the observation that noise distributions can be modeled using a Gaussian function. Thus, if we were to consider the center of the eye together with the corresponding set of fixation points then the result is a convex volume. Convex sets can be represented in two distinct ways. The first is based on the extreme points. In this case the fixation distribution plus the center of the eye. The second method is to represent the convex set using a set of half-planes that enclose the convex solid. A half-plane is plane which is defined by three extreme points and has the form  $ax \leq b$ . Based on this, we define two sets:

$$C_{LE} = \{x \in X : A_{LE}x \leq b_{LE}\} \quad (1)$$

and

$$C_{RE} = \{x \in X : A_{RE}x \leq b_{RE}\} \quad (2)$$

where  $A_{LE}x \leq b_{LE}$  and  $A_{RE}x \leq b_{RE}$  are the half-planes that enclose the convex solids obtained from the left and right eye respectively. A point  $x$  is in both sets if and only if it is in the intersection region, i.e.,

$$C = C_{LE} \cap C_{RE} \quad (3)$$

To solve for the set of all three-dimensional points  $x$  that would satisfy the system in Equation (3), we note that each inequality in Equations (1) and (2) defines a hyperplane. A hyper-plane, defined by an inequality of the form  $ax \leq b$ , divides the space into three parts: the first contains the vectors  $x$  that satisfy the inequality, i.e.,  $ax < b$ ; the second is the space of all the weights that violate the inequality, i.e.,  $ax > b$ ; and the third satisfies the equality, i.e.,  $ax = b$ . For a linear system of equalities and inequalities, such as the one defined in Equations (1) and (2), intersecting all the

hyper-planes results in a closed and convex region, which is the space of all feasible solutions, depth values, that satisfy the system. Using computational algorithms such as quick hull [1], it is possible to solve for the region of all feasible solutions.

Finally, Figure 2 shows a two-dimensional schematic representation of the cone-intersection method.

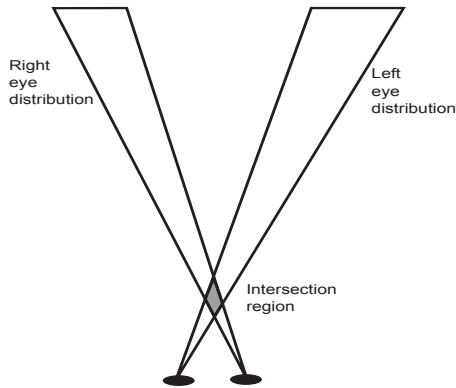


Figure 2. A schematic representation of the cone-intersection method. We note that the area of triangles emerging from the eyes is defined by the size of the statistical distribution of the fixation points. The intersection area, represented in gray, is the region defined by the intersection of the two convex sets where the larger this area is the more uncertain we are of the estimated depth value.

## 2.2 Note on practical implementation

The fixation points are extended along the lines that go from the center of the eye to the respective fixation point on the screen. Theoretically, the points are extended to infinity, for practical reasons the points are, however, extended to 50m behind the screen. Having done that we use the points together with the location of the eye as the input to quick hull algorithm. The output of the algorithm is a set of extreme points and the half-planes  $Ax \leq b$  that enclose the extremes. The process is repeated for both eyes and the set of half-planes are intersected to arrive at the intersection region which defines the feasible space of all the likely depth values.

## 3 Results

In this section, we discuss the results obtained from the proposed cone-intersection method. As previously discussed, the method is designed to estimate the uncertainty in the predicted depth. This uncertainty can be represented in a number of different ways such as the volume of the intersection region or maximum and minimum values. To constrain the values to the most likely region we have, however, chosen to represent the depth uncertainty using three standard deviations.

In order to test our method the fixations data are generated as follows: first, it is assumed that the observer is at 400 mm in front of the display screen and the virtual object is at depth  $d$  mm behind the screen. In the absence of noise for a given depth  $d$ , each of the left and right eye fixations have only one unique value. Second, random noise is added in horizontal and vertical directions to the left and the right eye fixations. This generates a distribution of left and right eye fixations for a given depth.

Figures 3 to 6 show the results obtained based on four depth values, namely, 100, 150, 200, and 300mm behind the screen. For all these depth values we note that the average values of the cone intersection region are a fair representation of the actual depth, the uncertainty depicted by the error-bars offers a more comprehensive view into the estimation. We observe that the real depth is almost always within the uncertainty region.

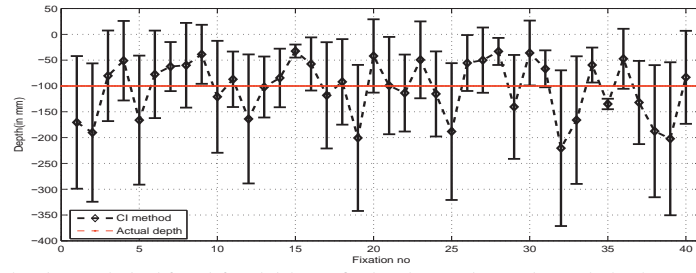
## 4 Conclusion

The contribution of this paper is the introduction of a numerical method that allows designers of virtual environments to estimate the uncertainty in the measured depth value. The proposed method is based on the principle of intersection of convex sets where two sets are defined. The first set is defined by the statistical distribution of the left eye fixations together with the center of the eye. A corresponding set is defined for the right eye. In an ideal situation i.e., when no noise is present in the data these two sets are reduced to the visual lines and the method is identical to the line-intersection method. When noise is present, however, the sets represent conical volumes and their intersection is the feasible solution space where any point is likely to be the actual depth. Based on that we represented the uncertainty in the estimate by means of three standard deviations from the average value. Our results show that the actual depth as specified in the environment is almost always within the uncertainty range.

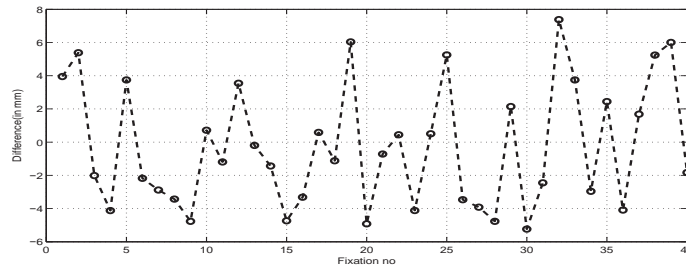
## References

- [1] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM TRANSACTIONS ON MATHEMATICAL SOFTWARE*, 22(4):469–483, 1996.
- [2] Gunnar Blohm, Aarlenne Z. Khan, Lei Ren, Kai M. Schreiber, and J. Douglas Crawford. Depth estimation from retinal disparity requires eye and head orientation signals. *Journal of Vision*, 8(16):1–23, December 2008.
- [3] James E. Cutting and Peter M. Vishton. *Perceiving layout: The integration, relative dominance, and contextual use of different information about depth*, volume 5, chapter 3, pages 69–117. New York: Academic Press, 1995.

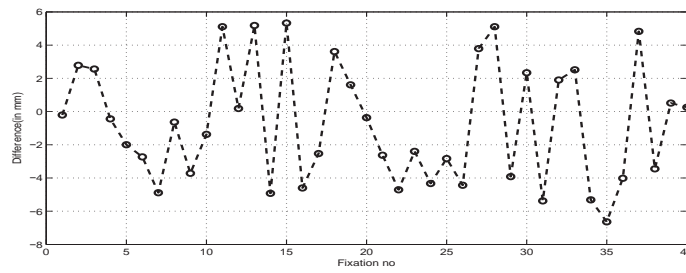
- [4] Andrew Duchowski, Eric Medlin, Nathan Cournia, Hunter Murphy, Anand Gramopadhye, Santosh Nair, Jeenal Vorah, and Brian Melloy. 3d eye movement analysis. *Behavior Research Methods, Instruments, and Computers (BRMIC)*, 34(4):573–591, 2002.
- [5] Kai Essig, Marc Pomplin, and Helge Ritter. A neural network for 3d gaze recording with binocular eye trackers. *The International Journal of Parallel, Emergent and Distributed Systems*, 21(2):79–95, April 2006.
- [6] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2003.
- [7] Maurice Hershenson. *Visual Space Perception : A Primer*. The MIT Press, 2000.
- [8] Ian P. Howard. *Seeing in Depth : Volume 1 Basic Mechanisms*. I Porteous, Toronto, 2002.
- [9] Ian P. Howard and Brian J. Rogers. *Seeing in Depth: Volume 2 Depth Perception*. I Porteous, Toronto, 2002.
- [10] Yong-Moo Kwon, Kyeong-Won Jeon, Jeongseok Ki, Qonita M. Shahab, Sangwoo Jo, and Sung-Kyu Kim. 3d gaze estimation and interaction to stereo display. *The International Journal of Virtual Reality*, 5(3):41–45, 2006.
- [11] Jonathan David Pfautz. *Depth Perception in Computer Graphics*. PhD thesis, Trinity College, University of Cambridge, 2000.
- [12] Thies Pfeiffer, Marc E. Latoschik, and Ipke Wachsmuth. Evaluation of binocular eye trackers and algorithms for 3d gaze interaction in virtual reality environments. *Journal of Virtual Reality and Broadcasting*, 5(16), December 2008. urn:nbn:de:0009-6-16605, ISSN 1860-2037.
- [13] Vildan Tanriverdi and Robert J. K. Jacob. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '00, pages 265–272, 2000.
- [14] Hung Q. Truong, Sukhan Lee, and Seok-Woo Jang. Model-based recognition of 3d objects using intersecting lines. *Multisensor Fusion and Integration for Intelligent Systems*, 35:289–300, 2009.
- [15] Zhengyou Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27(2):161–198, 1998.



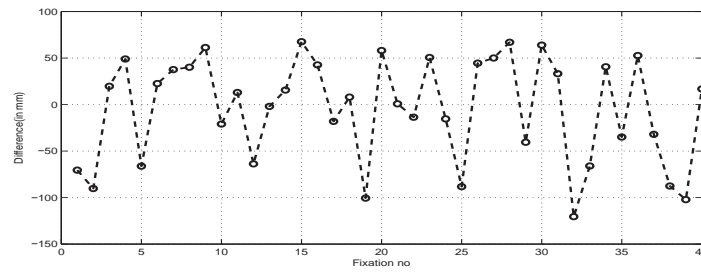
(a) Depth estimates obtained from left and right eye fixations by cone-intersection method. The average of depth estimates is 115 mm when the actual depth is 100 mm behind the screen. We notice that the actual depth is almost always within the uncertainty range.



(b) Difference between actual and estimated values of X.

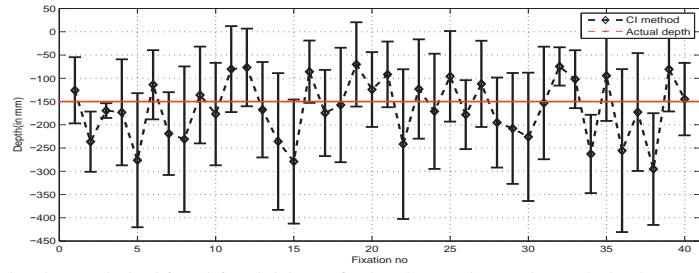


(c) Difference between actual and estimated values of Y.

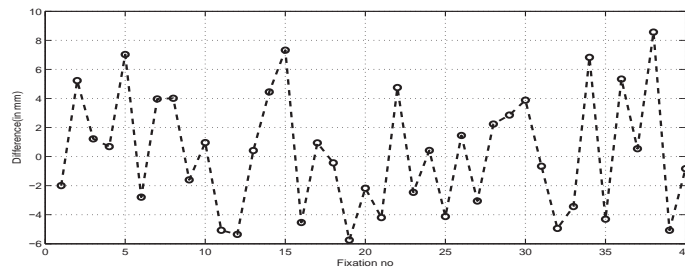


(d) Difference between actual and estimated values of Z.

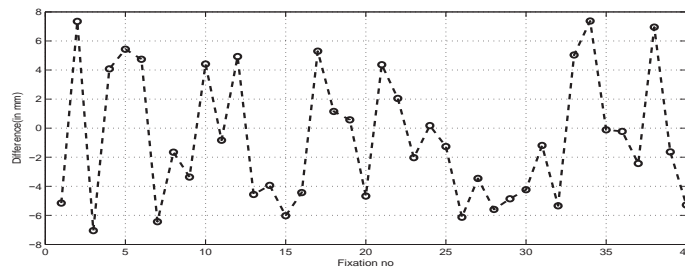
Figure 3. Depth estimates obtained from left and right eye fixations by cone-intersection method. Differences in actual and estimated values of X, Y, and Z.



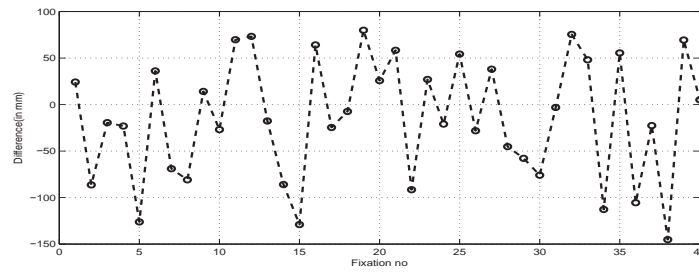
(a) Depth estimates obtained from left and right eye fixations by cone-intersection method. The average of depth estimates is 164 mm when the actual depth is 150 mm behind the screen. We notice that the actual depth is almost always within the uncertainty range.



(b) Difference between actual and estimated values of X.



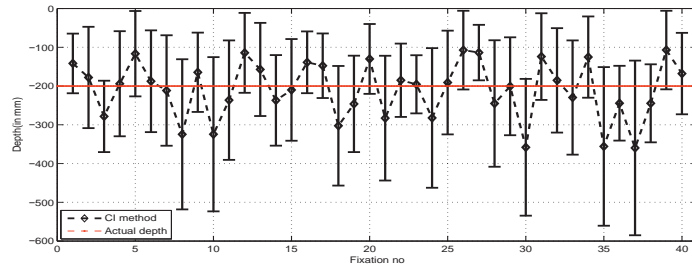
(c) Difference between actual and estimated values of Y.



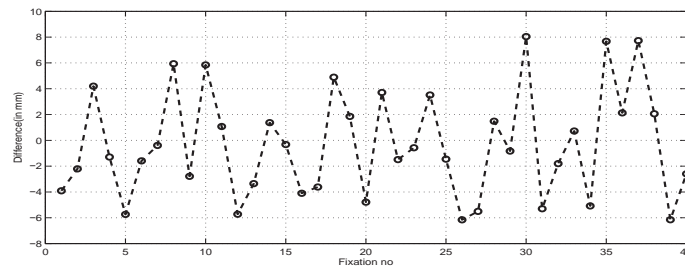
(d) Difference between actual and estimated values of Z.

Figure 4. Depth estimates obtained from left and right eye fixations by cone-intersection method. Differences in actual and estimated values of X, Y, and Z.

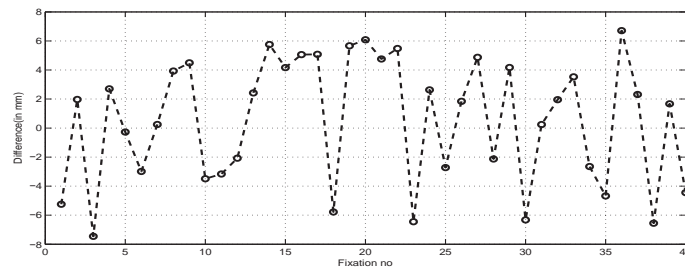




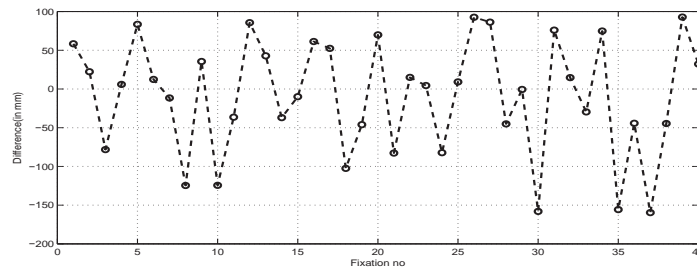
(a) Depth estimates obtained from left and right eye fixations by cone-intersection method. The average of depth estimates is 219 mm when the actual depth is 200 mm behind the screen. We notice that the actual depth is almost always within the uncertainty range.



(b) Difference between actual and estimated values of X.

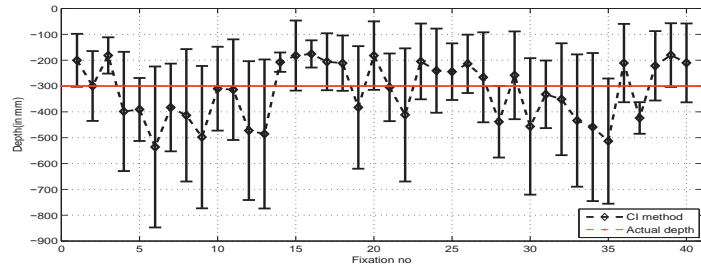


(c) Difference between actual and estimated values of Y.

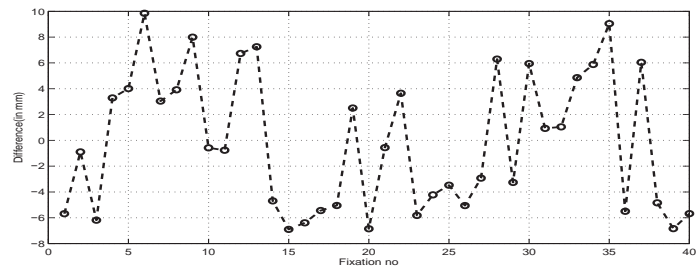


(d) Difference between actual and estimated values of Z.

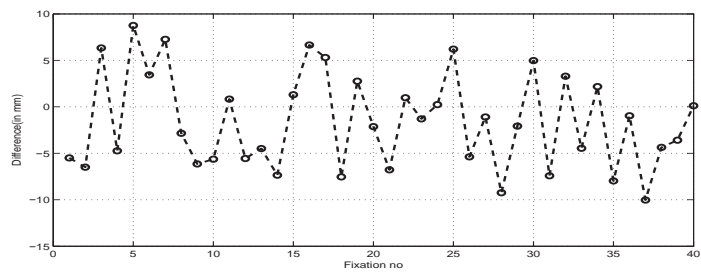
Figure 5. Depth estimates obtained from left and right eye fixations by cone-intersection method. Differences in actual and estimated values of X, Y, and Z.



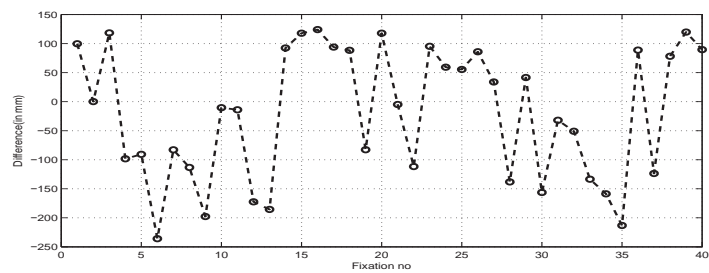
(a) Depth estimates obtained from left and right eye fixations by cone-intersection method. The average of depth estimates is 329 mm when the actual depth is 300 mm behind the screen. We notice that the actual depth is almost always within the uncertainty range.



(b) Difference between actual and estimated values of X.



(c) Difference between actual and estimated values of Y.



(d) Difference between actual and estimated values of Z.

Figure 6. Depth estimates obtained from left and right eye fixations by cone-intersection method. Differences in actual and estimated values of X, Y, and Z.

# Appendix **B**

## Statements of co-authorship

Statements of co-authorship from:

1. Ali Alsam
2. Anette Wrålsen
3. Faouzi Alaya Cheikh
4. Hans Jakob Rivertz
5. Jan Harald Nilsen
6. Torbjørn Skramstad

To whom it may concern

## **Statement of co-authorship on joint publications to be used in the PhD-thesis of Puneet Sharma**

(cf. the PhD regulations, section 10.1)

As co-author on the following joint publication in the PhD thesis “Towards three-dimensional visual saliency” by Puneet Sharma:

Alsam, A., & Sharma, P. (2011). Analysis of eye fixations data. In Proceedings of the IASTED International Conference, Signal and Image Processing (SIP 2011), (pp. 342-349).

*Alsam proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Alsam and Sharma contributed equally in the writing process.*

Sharma, P., & Alsam, A. (2014 (accepted)). A robust metric for the evaluation of visual saliency models. In International Conference on Computer Vision Theory and Applications (VISAPP 2014).

*Alsam and Sharma proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Sharma wrote the paper. Alsam provided feedback during the writing process.*

Alsam, A., & Sharma, P. (2014). Robust metric for the evaluation of visual saliency algorithms. Journal of the Optical Society of America A (JOSA A), 31 (2), 1-9.

*Alsam and Sharma proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Alsam and Sharma co-wrote the paper. Sharma lead the writing process.*

Alsam, A., & Sharma, P. (2013). Validating the visual saliency model. In SCIA 2013, Lecture Notes in Computer Science (LNCS), vol. 7944, (pp. 153-161). Springer-Verlag Berlin Heidelberg.

*Alsam proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Alsam and Sharma contributed equally in the writing process.*

Alsam, A., Sharma, P., & Wrålsen, A. (2013b). Asymmetry as a measure of visual saliency. In SCIA 2013, Lecture Notes in Computer Science (LNCS), vol. 7944, (pp. 591-600). Springer-Verlag Berlin Heidelberg.



*Alsam and Sharma proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Alsam, Wrålsen and Sharma co-wrote the paper. Sharma lead the writing process.*

Alsam, A., Sharma, P., & Wrålsen, A. (2014). Calculating saliency using the dihedral group d4. *Journal of Imaging Science & Technology*, accepted.

*Alsam and Sharma proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Alsam, Wrålsen and Sharma co-wrote the paper. Sharma lead the writing process.*

Alsam, A., Rivertz, H. J., & Sharma, P. (2012). What the eye did not see--a fusion approach to image coding. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, C. Fowlkes, S. Wang, M.-H. Choi, S. Mantler, J. Schulze, D. Acevedo, K. Mueller, & M. Papka (Eds.) *Advances in Visual Computing*, vol. 7432 of *Lecture Notes in Computer Science*, (pp. 199-208). Springer.

*Alsam proposed the method, Alsam, Rivertz, and Sharma designed and implemented the solution. Sharma conducted the experiments. Alsam, Rivertz and Sharma contributed equally in the writing process.*

Alsam, A., Rivertz, H. J., & Sharma, P. (2013a). What the eye did not see--a fusion approach to image coding. *International Journal on Artificial Intelligence Tools*, 22 (6), 13.

*Alsam proposed the method, Alsam, Rivertz, and Sharma designed and implemented the solution. Sharma conducted the experiments. Alsam, Rivertz and Sharma co-wrote the paper. Sharma lead the writing process.*

Sharma, P., & Alsam, A. (2012a). Estimating the depth in three-dimensional virtual environment with feedback. In *Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012)*, (pp. 9-17).

*Alsam and Sharma proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Alsam and Sharma co-wrote the paper. Sharma lead the writing process.*

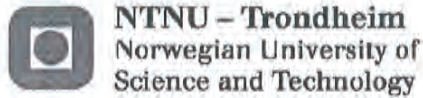
Sharma, P., & Alsam, A. (2012b). Estimating the depth uncertainty in three-dimensional virtual environment. In *Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012)*, (pp. 18-25).

*Alsam and Sharma proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Alsam and Sharma co-wrote the paper. Sharma lead the writing process.*

I declare that the contributions to the papers are correctly identified, and I consent to this work being used as part of the thesis.

Date: 11/02/2014

Ali Alsam 



To whom it may concern

**Statement of co-authorship on joint publications to be used in the PhD-thesis of Puneet Sharma**

(cf. the PhD regulations, section 10.1)

As co-author on the following joint publication in the PhD thesis “Towards three-dimensional visual saliency” by Puneet Sharma:

Alsam, A., Sharma, P., & Wrålsen, A. (2013b). Asymmetry as a measure of visual saliency. In SCIA 2013, Lecture Notes in Computer Science (LNCS), vol. 7944, (pp. 591-600). Springer-Verlag Berlin Heidelberg.

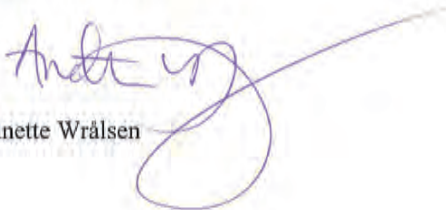
*Alsam and Sharma proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Alsam, Wrålsen and Sharma co-wrote the paper. Sharma lead the writing process.*

Alsam, A., Sharma, P., & Wrålsen, A. (2014). Calculating saliency using the dihedral group d4. Journal of Imaging Science & Technology, accepted.

*Alsam and Sharma proposed the idea, Sharma designed and implemented the solution, conducted the experiments. Alsam, Wrålsen and Sharma co-wrote the paper. Sharma lead the writing process.*

I declare that the contributions to the papers are correctly identified, and I consent to this work being used as part of the thesis.

Date: 11.02.14.....

  
Anette Wrålsen



**NTNU – Trondheim**  
Norwegian University of  
Science and Technology

To whom it may concern

**Statement of co-authorship on joint publications to be used in the  
PhD-thesis of Puneet Sharma**

(cf. the PhD regulations, section 10.1)

As co-author on the following joint publication in the PhD thesis “Towards three-dimensional visual saliency” by Puneet Sharma:

Sharma, P., Nilsen, J. H., Skramstad, T., & Cheikh, F. A. (2010). Evaluation of geometric depth estimation model for virtual environment. In Norsk informatikkonferanse (NIK-2010)

Sharma formulated the problem, designed and conducted the experiments and wrote the paper. Nilsen, Skramstad, and Cheikh provided feedback during the design and writing process.

I declare that the contributions to the papers are correctly identified, and I consent to this work being used as part of the thesis.

Date: *.11.02.2014*

Faouzi Alaya Cheikh





**NTNU – Trondheim**  
Norwegian University of  
Science and Technology

**To whom it may concern**

**Statement of co-authorship on joint publications to be used in the PhD-thesis of Puneet Sharma**

(cf. the PhD regulations, section 10.1)

As co-author on the following joint publication in the PhD thesis “Towards three-dimensional visual saliency” by Puneet Sharma:

Alsam, A., Rivertz, H. J., & Sharma, P. (2012). What the eye did not see--a fusion approach to image coding. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, C. Fowlkes, S. Wang, M.-H. Choi, S. Mantler, J. Schulze, D. Acevedo, K. Mueller, & M. Papka (Eds.) *Advances in Visual Computing*, vol. 7432 of *Lecture Notes in Computer Science*, (pp. 199-208). Springer.

*Alsam proposed the method, Alsam, Rivertz, and Sharma designed and implemented the solution. Sharma conducted the experiments. Alsam, Rivertz and Sharma contributed equally in the writing process.*

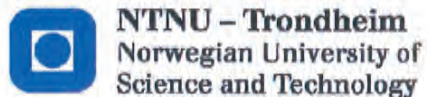
Alsam, A., Rivertz, H. J., & Sharma, P. (2013a). What the eye did not see-a fusion approach to image coding. *International Journal on Artificial Intelligence Tools*, 22 (6), 13.

*Alsam proposed the method, Alsam, Rivertz, and Sharma designed and implemented the solution. Sharma conducted the experiments. Alsam, Rivertz and Sharma co-wrote the paper. Sharma lead the writing process.*

I declare that the contributions to the papers are correctly identified, and I consent to this work being used as part of the thesis.

Date: 17/2 2014

Hans Jakob Rivertz



To whom it may concern

**Statement of co-authorship on joint publications to be used in the  
PhD-thesis of Puneet Sharma**

(cf. the PhD regulations, section 10.1)

As co-author on the following joint publication in the PhD thesis “Towards three-dimensional visual saliency” by Puneet Sharma:

Sharma, P., Nilsen, J. H., Skramstad, T., & Cheikh, F. A. (2010). Evaluation of geometric depth estimation model for virtual environment. In Norsk informatikkonferanse (NIK-2010)

Sharma formulated the problem, designed and conducted the experiments and wrote the paper. Nilsen, Skramstad, and Cheikh provided feedback during the design and writing process.

I declare that the contributions to the papers are correctly identified, and I consent to this work being used as part of the thesis.

Date: *2014-02-25*

*Jan H. Nilsen*

Jan H. Nilsen



**NTNU – Trondheim**  
Norwegian University of  
Science and Technology

To whom it may concern

**Statement of co-authorship on joint publications to be used in the  
PhD-thesis of Puneet Sharma**

(cf. the PhD regulations section 10.1))

As co-author on the following joint publication in the PhD thesis “Towards three-dimensional visual saliency” by Puneet Sharma:

Sharma, P., Nilsen, J. H., Skramstad, T., & Cheikh, F. A. (2010). Evaluation of geometric depth estimation model for virtual environment. In Norsk informatikkonferanse (NIK-2010)

Sharma formulated the problem, designed and conducted the experiments and wrote the paper. Nilsen, Skramstad, and Cheikh provided feedback during the design and writing process.

I declare that the contributions to the papers are correctly identified, and I consent to this work being used as part of the thesis.

Date: ..... 10. February 2014 .....

Torbjørn Skramstad