Gradual Saliency Detection in Video Sequences Using Bottom-up Attributes

by

Jila Hosseinkhani, **M.Sc.**

A thesis submitted to the
Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario
January, 2020

The undersigned hereby recommends to the
Faculty of Graduate and Postdoctoral Affairs
acceptance of the thesis

Gradual Saliency Detection in Video Sequences Using Bottom-up Attributes

submitted by **Jila Hosseinkhani**, **M.Sc.**

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering

Professor Chris Joslin, Thesis Supervisor

Professor Carlos Vazquez, External Examiner

Professor Yvan Labiche, Chair,
Department of Systems and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
Department of Systems and Computer Engineering
Carleton University
January, 2020

Abstract

The demand for video streaming is growing every day which implies a higher demand for new video transmitting and compression techniques to avoid data traffics over telecommunication networks. In this dissertation, we studied saliency detection in order to apply it to video streaming problem to be able to transmit different regions of video frames in a ranked manner based on their importance (i.e., saliency). Salient areas are the regions of interest that stand out relative to their surroundings and consequently attract more attention. To determine the salient areas within a scene, visual importance and distinctiveness of the regions must be measured. The lack of a comprehensive and precise biologically-inspired study on the saliency of bottom-up stimuli prevents justifying the level of importance for different stimuli such as color, luminance, texture, and motion on the human visual system (HVS). To overcome this barrier, we investigated the bottom-up features using an eye-tracking procedure and human subjects in video sequences to provide a ranking saliency system stating the most dominant elements for each feature individually as well as in combination with other features. The experiment was performed under conditions in which we had no cognitive bias in order to speed up the video streaming procedure.

Next, we introduced a gradual saliency detection framework for both still images and video sequences using color, texture, and motion features (based on our experimental estimations). In our algorithm, we proposed new feature maps for color and texture features, and we also improved the optical flow field estimation in our motion map. Finally, different feature maps were combined and classified as different saliency levels using a Naive Bayesian Network. This work provides a benchmark to specify the gradual saliency for both static and dynamic (i.e., moving backgrounds) scenes. The main contribution of this work is the ability to assign a gradual saliency for the entirety of an image/video frame rather than simply extracting a salient object/area, which is widely performed in the state-of-the-art.

To My Parents

Acknowledgments

I have a deep debt of gratitude to everyone who supported me throughout my Ph.D. study.

Firstly, I would like to express my sincere appreciation and thanks to my Ph.D. thesis supervisor, Professor Chris Joslin for his guidance, support, motivation, and knowledge.
Thanks must also go to my committee members for their helpful suggestions and discussions.

Besides, I am mostly grateful to my wonderful parents for their unconditional love, support, and faith in me. Words cannot express how grateful I am to my mother, father, and sister for their support and encouragements. This accomplishment would not have been achievable without their constant support and love.

I also owe much to my caring and supportive friends in Canada and elsewhere who helped me stay strong and keep moving forward during these years.

Table of Contents

List of Tables

List of Figures

Acronyms

Acronym	Description
VOD	Video On Demand
IPTV	Internet Protocol Television
QoS	Quality of Service
ER	Error Resilient
LC	Layered Coding
MDC	Multiple Description Coding
HVS	Human Visual System
VAM	Visual Attention Model
2D	2-Dimensional
CIELab	International Commission on Illumination (in French) L a* b*
SMI	SensoMotoric Instrument
CVD	Color Visual Deficiency
ITU-R BT	International Telecommunication Union (Broadcasting Television series)
RGB	Red, Green, Blue
HSV	Hue, Saturation, Value/Intensity
Hright	Horizontal right side
Hleft	Horizontal left side

Acronym	Description
Vup	Vertical upward
Vdown	Vertical downward
Dup+	Diagonal of +45 degree upward
Dup-	Diagonal of -45 degree upward
Ddown+	Diagonal of +45 degree downward
Ddown-	Diagonal of -45 degree downward

Chapter 1

Introduction

An enormous stream of visual data (10^8-10^9 bits) enters human eyes every second [1]. Processing this volume of data in real-time is an extremely challenging task that demands a large volume of memory and time. All of this visual data is not informative and consequently, it is essential to preserve the more informative part and decrease the amount of insignificant visual data.

Visual attention system in human brain selects the region of interest across visual information through eye movements. Researchers studied how visual stimuli affect human eye movements to estimate the most attractive and significant regions in a scene and design saliency detection models to extract those regions. Salient regions of an image or video for human and machine are defined as visually and perceptually distinctive areas that attract the attention of the viewer immediately [2].

During the past three decades, saliency detection has gained considerable attention among the vision community, especially in the fields of image/video processing and pattern classification. This research field is widely exploited in many applications such as content-based image/video retrieval, scene understanding, video surveillance, video summarization, event detection, image/video compression, image/video retargeting, image segmentation, object detection, and fixation prediction models.

Our principal motivation for studying saliency detection returns to the idea of controlling the telecommunication network traffic and video streaming/compression. During the past decade, video streaming demand has gone through an incredible increase due to the immense expansion of multimedia communications. This evolution was driven by the development of ubiquitous applications for video streaming such as video conferencing, Video On Demand (VOD), live TV, Internet Protocol Television (IPTV), corporate webcast, and real-time surveillance as well as the increase in the

number of users, the availability of high bandwidth, and the growth of commercial of networks. This demand imposed massive network traffic, which is carrying video content mostly for video streaming applications.
According to Cisco Systems reports (released on Feburary 2019), in 2017, video data comprised 75% of all consumer traffic [3], growing to an estimated 82% by 2022. In the mobile sector, [4] this traffic volume is 60%, which is estimated to be 80% by 2022.

To tackle the resulting congestion of the enormous volume of video data over the networks, more bandwidth and reliable communications are required. For instance, network traffic issues due to bandwidth limitations can be resolved using scalable video representations. Therefore, it is important to devise video encoding/decoding schemes that make the compressed bit-stream resilient to transmission errors due to bandwidth limitations. In this way, the codec standard (encoder and decoder) can adjust its operation based on the network conditions [5].
Recently, object-based scalability is more required - which considers the region of interest within a frame (i.e., saliency).

A variety of saliency detection models exist in the computer vision field, but, they often define the saliency detection problem as a binary classification procedure to estimate the most significant and informative regions of an image/video frame, and discard the remainder of the regions. However, we consider knowing the order/gradation of the importance among different regions of a scene can be very advantageous in many applications. We call this gradation of the importance of the regions, as gradual saliency.

Gradual saliency can be used to improve Layered Coding (LC). In LC, the video data is coded into a base layer and one or more enhancement layers. The base layer can provide a low but acceptable level of quality, and enhancement layers incrementally improve the streaming video quality. This coding technique is also called scalable coding [6, 7].
The gradual saliency technique can provide guidance for the encoder to dynamically decide what regions of video data are more salient and must form the base layer. Also, based on the saliency priority, it can be determined what regions information can be dropped or form the enhancement layers to be encoded at any given time period. In this way, we can transmit the salient parts with higher quality while assigning the lower bit-rate for the non-salient regions among each layer.

Another application of gradual saliency can be the Foveated Rendering field. Foveated rendering is a rendering technique performed by incorporating an eye-tracker device and a virtual reality headset. The goal of this technique is the rendering workload reduction by decreasing the image quality in the peripheral vision - which is defined as the outside of the area gazed by the fovea [8]. The output of our gradual saliency algorithm can help foveated rendering by concentrating on the salient regions rendering and reducing the rendering burden for the non-salient areas.

1.1 Problem Statement and Research Objective

As we can see, there are many problems with using a binary saliency detection fashion. The current state-of-the-art related to saliency detection is considering the saliency as a binary segmentation problem and tries to extract the most important and attractive region of a frame without considering different levels of importance.

Two principal problems associated with the existing saliency detection methods are as follows: 1) The available methods are designed based on ground-truth data that exhibit different regions of a frame as to be true or false salient regions with no gradation, i.e., binary images. 2) The available methods produce only binary solutions for many of the applications. Our focus in this thesis is to determine what different gradations are for various visual influences, but also to develop algorithms that can (in real-time) produce the same results.

Therefore, we decided to introduce the gradual saliency concept which means the different levels of importance for the different regions of a visual scene.
For a real-time application, it is essential to have a fast and simple saliency detection method. Therefore, we decided to avoid any cognitive bias in designing our model to speed up the procedure of salient area detection (Human cognition refers to a systematic pattern of perceptual and rational judgements and decision-making actions).

In this work, we focused on the study of bottom-up attributes as visual stimuli such as color, texture, color contrast, motion direction, object velocity, and object acceleration.

1.2 Research Scope

To overcome the aforementioned problems from different perspectives, we decided to utilize a semantic video analysis mechanism known as a saliency detection technique. Salient regions are also called regions of interest (ROI) or salient regions.

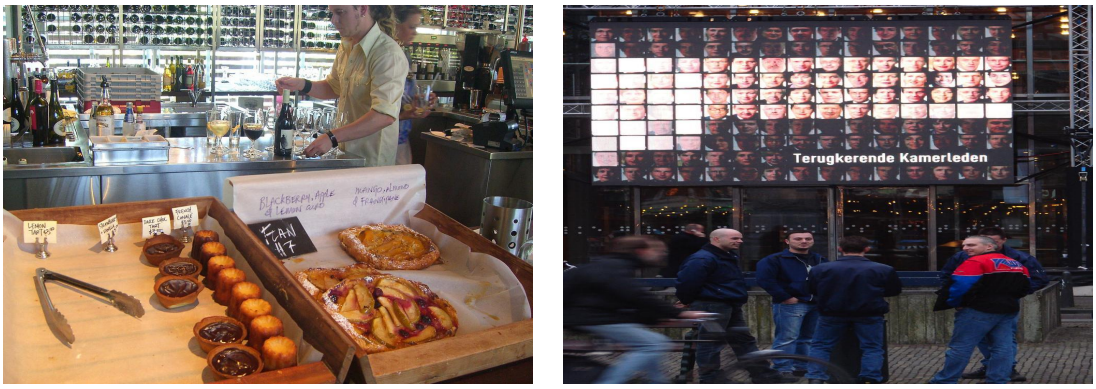
Salient areas are usually more aggregative in color or texture distribution, such as the high contrast between colors, some specific and rare colors in the scene, complicated textures/patterns, and a unique movement or different speed in a video sequence [2]. A salient region could be an object, an area, or even a pixel (for example, when we have a small white dot in a black screen), which has outstanding quality or state relative to its surrounding areas. From both human and machine perspectives, it seems that saliency arises from contrasts between items and their neighbourhood, such as a flickering panel of a store.

One challenging issue from the human being perspective refers to the fact, that salient region of a frame may result through emotional, motivational or cognitive factors, and is not necessarily related to intrinsic and statistical factors of the video sequence, such as intensity, motion, depth, clarity or size. The subjective factors such as age, culture, and experience may also influence the saliency detection procedure and may result in different saliency maps. Therefore, saliency detection system design requires matching between human and machine abilities and cognitive in the scene interpretation. For instance, a flickering region of a video can be extracted without any previous knowledge compared to selecting a diamond ring - which is influenced by human cognitive - while in both scenarios, the given scene contains a brilliant object.

Two sample images in Figure 1.1 are shown here to represent the human visual system (HVS) and cognitive diversities in saliency detection.

In this Figure, in the image (a), if subjects would be children, then the most attractive part for them will be more likely the sweets. However, for an adult, the prices or drinks at the back might be more interesting. In the image (b), depending on the subject, the monitors or people on the street might be considered as the salient regions. In summary, people see the same scene in different ways.

The main goal of related studies is to design a saliency detection system, which could be fairly comparable with HVS. These systems are usually called Visual Attention Models (VAMs) inspired by the HVS.



(a) (b)

Figure 1.1: Sample images selected from MIT300 [9] dataset to show different saliency patterns for different people.

Human visual attention contains two types of processes: pre-attentive and attentive [10]. Pre-attentive (subconscious) processing rapidly and automatically categorizes an image into regions in a spatially parallel manner to search for significant information across an image/video. The attentive (conscious) processing or focused attention incorporates the goals and desires of the viewer (i.e., cognitive bias) through the process of searching in a serial manner - which is time consuming compared to pre-attentive detection [11].

Physiological and psychological studies illustrated that the effective factors on visual attention and eye movements are categorized into bottom-up and top-down types [12]. Bottom-up factors capture pre-attentive attention very quickly and have a powerful impact on the human visual selection system. On the other hand, top-down factors capture the attention much slower and are influenced by bottom-up factors.

In the past two decades, researchers focused on designing VAMs inspired by the HVS. Saliency detection models or VAMs employ bottom-up and/or top-down factors to search for the salient part of data. Bottom-up based models use low-level attributes such as color, texture, size, contrast, brightness, position, motion, depth, orientation, and shape of objects. These attributes are rapidly scanned and detected by the human visual system. However, top-down based models exploit high-level context-dependent attributes such as a face, human, animal, vehicle, text, etc. [13]. Both bottom-up and top-down factors can be exploited to design VAMs - but because of the complexity and time limitation - a few integrated approaches have been proposed using both

factors to detect the saliency in a scene [14].

The validation of the saliency maps is usually performed by comparing them against eye-tracking datasets, which are assumed as the ground-truth data. Two kinds of eye movements are saccades and fixations. A saccade is a quick and simultaneous movement of both eyes between one fixation location to another in the same direction. A fixation is slow eye movement that preserves the visual gaze on a single location of the stimulus. Human observers can pay attention to a peripheral object without any eye movement - which is called covert attention. If the attention involves saccades it is referred to as overt [10].

Studies show that the human visual system is attracted to objects rather than locations [15]. The pre-attentive part of the HVS firstly segments the scene into objects in a rapid scan [15]. This segmentation is mostly performed based on low-level attributes.

There is a difference between the human user perception of the image/video and what extracted features indicate about the data - which is the so-called semantic gap problem. Therefore - in the feature extraction step - it is essential to achieve a feature vector (i.e., a set of features) that is capable of efficiently representing the visual content of a video [16]. For this purpose, we designed a subjective experiment to achieve a biologically plausible method to define features and obtain a reliable saliency map by filling this gap.

1.3 Research Overview

Since saliency detection is very strictly influenced by human visual perception, it is necessary to study the impact of visual stimuli on HVS. The lack of comprehensive research in this area inspired us to design an experimental study to precisely and comprehensively investigate the effect of bottom-up attributes on HVS in terms of saliency detection problem. In this way, we can design a more stable VAM to extract the most informative and attractive part of a visual scene.

The main inspiration behind this research field refers to the reduction of the massive amount of visual information received by HVS into the most informative and significant portion for the applications such as image/video summarization, semantic video analysis, network traffic control, video compression, and so on.
Our goal is to investigate how they influence the HVS and achieve a ranking system to

identify what hue range, texture pattern, motion direction, object velocity, and object acceleration are most likely to be attractive for the HVS in terms of saliency detection. We proposed a gradual saliency detection method for both static and dynamic video scenes based on our findings from an eye-tracking based experiment. Novel feature maps were introduced for color, texture, and motion speed then a Bayesian framework was used to merge all feature maps into a final gradual saliency map.

1.4 Research Contributions

The main contributions of this work can be described as follows:

- Designed a subjective eye-tracking based experiment to investigate the influence of individual as well as incorporated bottom-up stimuli on HVS that leads to achieving biologically inspired feature maps for both static and dynamic scenes (i.e., moving backgrounds as well as busy backgrounds).

- Designed saliency detection algorithms for both static and dynamic scenes using color, texture, intensity, luminance, motion direction, and motion speed/acceleration feature maps. We merged all maps by the Naive Bayesian framework.

  Our model can assign prioritized (i.e., gradual) saliency for the entire image/video frame regions rather than only the salient regions. In current literature, only the salient areas and objects are used to design a saliency detection model. Instead, we provide a comprehensive heat map for every region of an image/video frame.

- Designed a validation experiment using an eye-tracker to show how HVS grades different regions of a frame. We assumed experimental results as a ground-truth dataset to compare against our implementation results - for both static and dynamic scenes.

- Segmented salient areas with high quality and well-defined boundaries resulting in a full-resolution saliency map of the same size as the input image (without any down-sampling).

1.5 Published Papers

1. J. Hosseinkhani, C. Joslin, "Significance of Bottom-up Attributes in Video Saliency Detection Without Cognitive Bias," *18th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCICC)*, pp. 606-613, USA, 2018 (reused in Chapter 3).

2. J. Hosseinkhani, C. Joslin, "Investigating into Saliency Priority of Bottom-up Attributes in 2D Videos Without Cognitive Bias," *IEEE International Conference on Signal Processing and Information Technology (ISSPIT)*, pp. 223-228, USA, 2018 (reused in Chapter 3).

3. J. Hosseinkhani, C. Joslin, "Saliency Priority of Individual Bottom-up Attributes in Designing Visual Attention Models," *International Journal of Software Science and Computational Intelligence*, Vol. 10, No. 4, pp. 1-18, 2018 (reused in Chapter 3).

4. J. Hosseinkhani, C. Joslin, "Saliency Priority Using Bottom-up Features for Static and Dynamic Scenes Without Cognitive Bias," *IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR)*, USA, pp. 189-192, 2019 (reused in Chapter 4).

5. J. Hosseinkhani, C. Joslin, "A Biologically Inspired Saliency Priority Extraction Using Bayesian Framework," *International Journal of Multimedia Data Engineering and Management*, Vol. 10, No. 2, pp.1-20, 2019 (reused in Chapter 4).

1.6 Dissertation Organization

More detailed explanations have been provided about existing studies related to the impact of bottom-up attributes on the visual attention system in addition to our designed experiment and method in the next chapters. The rest of this dissertation was organized as follows: Chapter 2 reviews the related works and provides an overview of how the previous experimental studies have been performed to understand HVS response to the bottom-up stimuli. In addition, we reviewed and categorized the existing saliency detection models in the literature. Chapter 3 describes the characteristics of the generated dataset for our experiment and its methodology. It also

illustrates the experiment results for each individual bottom-up attribute as well as their combination. Chapter 4 contains the methodology and results of the proposed algorithm to extract salient regions within static scenes (still images). Chapter 5 explains the introduced saliency detection model and its corresponding results for dynamic scenes (video sequences), and finally, Chapter 6 concludes the work of this dissertation and discusses future work.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, we review the core empirical and analytical background of the existing visual attention models and techniques in detecting the salient regions. Investigating techniques and main modeling trends of the existing saliency detection and visual attention models leads us to be able to categorize them and have a better understanding of their algorithms. Then, we refer to the most popular existing metrics in the field with the aim of investigating a solid evaluation of the existing models. Because a solid comparison among different saliency models would be simply possible using appropriate metrics and datasets.

2.2 Related Research Areas

Some research areas are closely related to saliency detection such as fixation prediction, image segmentation, and object proposal/detection. They are sometimes highly overlapped and similar to each other, however, their fundamental differences bring them into different categories.

2.2.1 Fixation Prediction

Fixation prediction models are designed to understand human visual attention and human behaviour so that they track eye movements in order to predict the areas which are most likely to be seen by human beings. However, saliency detection models have mostly been established for content-aware applications such as image search engines,

image resizing, image/video compression, and so on.
Eye fixation models predict the points that people gaze more so that fixation maps are formed with integrated dots while saliency maps are exhibited by regions and areas [1]. They are usually used as ground-truth data to validate saliency maps. Although, they cannot be reliable ground-truth data since saliency maps consist of several regions which include a lot of pixels while fixation maps are formed based on fixation dots. In addition, they are very subject-dependant. It is necessary to convert fixation dots into regions to be comparable with saliency maps. This converted fixation map is known as fixation density map (FDM). FDMs can be extracted by post-processing of the eye gaze tracking patterns (i.e. fixation maps) to have access to the regions of interest. Therefore, the fixation points are usually convolved by a Gaussian kernel to obtain the approximate gazed areas by the human. In some studies, eye movement tracking and fixation point maps are utilized to understand the judgement of the viewers about the most salient objects in the images. Figure 2.1 shows the difference between fixation points, saliency map, and FDM resulted from gaze points.

Figure 2.1: Comparing fixation, saliency, and FDM maps. From left to right: original image, fixation map, saliency map, and FDM respectively [17].

If an image is briefly presented to an observer (i.e., 80 ms or less), the observer is able to report essential characteristics of a scene [17]. This very rough representation of a scene is called "gist" and does not contain many details about individual objects, but can provide sufficient information for coarse scene discrimination (e.g., indoor versus outdoor). It is essential to note that gist does not necessarily show the semantic information of a scene [18].

2.2.2 Image Segmentation

Image segmentation is one of the widely researched areas in computer vision that has gained scientists' attention. The goal of segmentation techniques is partitioning an

image into multiple segments and label them. In fact, the segments are sets of pixels showing an object or a part of an image which is more meaningful. Segmentation is usually used as a pre-processing step to create super-pixels in object detection. Segmentation operates as a labeling task for each pixel to indicate if it belongs to any objects or background in a scene.

In contrast, saliency detection methods try to detect the most important partitions of the image rather than specifying different existence partitions in the image. Therefore, we can state that according to the state-of-the-art, saliency detection has been considered as a segmentation with binary labeling problem to identify salient areas from non-salient parts.

Segmentation algorithms have been used as the first step of saliency detection in some VAMs and then the most salient parts of the image were chosen among the obtained partitions [16]. However, exploiting segmentation as the first step can be time-consuming while we do not have access to a very accurate segmentation algorithm yet [16]. In addition, there is a high demand to design fast VAMs because saliency detection is usually the basic step in many applications such as image/video compression, scene understanding and event summarization which needs to be real-time as much as possible. To overcome these barriers, some studies proposed super-pixels to facilitate segmentation algorithms [16].
Unfortunately, super-pixels lead to challenges such as losing useful edges and boundaries and extracting non-relevant parts of the image as the salient part.

2.2.3 Object Detection

Object detection models are more similar to saliency detection models however there are two basic differences. Firstly, object detection models assume that an object is more likely to be salient than a region on the background. Secondly, these approaches use objectness measures to assign higher saliency values to the objects in comparison with the background [1]. On the other hand, VAMs aim to extract the most distinctive parts of an image/frame no matter they belong to an object or background.

Sample images are shown in Figure 2.2 to illustrate the difference between these four research fields. In this figure, part (a) shows the original image, (b) is the result of saliency detection, the image at part (c) indicates the result of the eye fixation prediction, image segmentation method is shown in (d), and finally in (e) the

(a) (b) (c) (d) (e)

Figure 2.2: Sample images to show the difference of saliency detection, fixation prediction, segmentation, and object detection methods [1].

bounding boxes are drawn using object detection method.

2.3 Empirical Background

In this chapter, we first provide derived conclusions and results of the existing works about bottom-up attributes. Then, a brief description of the relevant experimental studies and designed VAMs will be presented.
In the literature, few experimental studies have been performed to investigate the impact of the bottom-up factors on the HVS and eye movements. This motivated us to design a comprehensive experiment in order to study the significance of bottom-up stimuli in saliency detection procedure since they are key factors in capturing human attention.

The origin of many attention models dates back to Treisman and Gelades [19] "Feature Integration Theory," in 1980, where they stated which visual features are important to direct human attention over pop-out and conjunction search tasks.
According to this theory, if conjunctions of more than one separable feature are required to distinguish objects, human attention must be directed serially to each stimulus [19]. They tested a number of predictions including visual search, texture segregation, identification, and localization. They employed both separable dimensions (shape and color) and local elements (lines, curves, etc. in letters) as the features integrated into complex regions/objects [19]. According to their findings, it seems that human can detect and identify separable features in a parallel manner across a display. This early parallel mechanism of feature registration mediates a primitive region grouping and texture segmentation task [19].

In 1994, Wolfe et al. [20] presented a model of human visual search behavior based on the existing works in the literature such as Treisman [19], Neisser [21], and

Julesz [22]. Wolfe's model known as "guided search" which distinguishes between a pre-attentive stage and a subsequent limited-capacity stage of the visual search. The pre-attentive stage is enormously parallel that processes information of low-level visual features such as color, motion, depth, etc. across large portions of the visual scope [20]. The subsequent limited stage performs more complex operations e.g., face recognition, reading, and object identification over a limited portion of the visual scope [20]. The guided search theory proposes that attention can be biased toward targets of interest by modulating the relative gains of different features that contribute to attention [23].

Most of the recent experimental research is focused on investigating differences between 2D and 3D visual data and their impact on the HVS. For example, Khaustova et al. [24] designed an experimental study to understand how texture complexity, depth quantity, and visual comfort influence the way people observe 3D content in comparison with 2D content. They utilized uncrossed disparity (i.e. all objects were behind the display plane) for the all stereoscopic content. Two experiments were performed using an eye-tracker and a 3D-TV display. In the first experiment, 51 subjects participated in the test. They found that the objects with crossed disparity are the most salient, even if observers experience discomfort due to the high-level of disparity.

The second experiment was designed with the aim of investigating whether depth is a determinative factor for visual attention. In this experiment, 28 observers watched scenes that contained objects with crossed and uncrossed disparities with different textures. They discovered that texture is more important in comparison with depth for selection of salient objects. They also found that the gaze points were concentrated in the center of the scene during the first 4 seconds of the experiment, but for the other time intervals, the gaze points were spread over the entire scene [24].

Hakkinen et al. [25] analyzed the eye movements of participants watching a six-minute movie in both stereoscopic and non-stereoscopic versions. The results indicated that viewers tend to look at the actors in the 2D version. In this test, 20 students participated. The short film (6 minutes and 20 seconds long) was presented. They used a Hyundai 46-inch polarizing stereoscopic display with a resolution of 1920 $\times$ 1080 pixels. The film was shown with a TriDef stereoscopic player and a Tobii X120 eye movement tracker was utilized. They found that eye movements of subjects are mostly concentrated on the actors and their immediate neighborhood. Based on

the eye movement patterns, they inferred that the observers are mostly looking for socially relevant information [25]. These factors are categorized as high-level and content-based features in video sequences. They reported that the eye movements spread more widely in the 3D versions. Also, the objects coming toward the observer caught more of the viewers' attention [25].

Khaustova et al. [26] in another experiment, generated six scenes with different modified parameters using Blender. The modified parameters were: texture complexity and the amount of depth changing the camera baseline and the convergence distance at the shooting side. Their experiment was performed using an eye-tracker and a 3D-TV display. They ensured that each observer had only seen the content of each scene once to avoid memory bias [26]. A Tobii x50 eye-tracker and 42 LG 42LW stereoscopic display with line interleaved technology were used as the setup for this test. The duration of the experiment was about 10 minutes for each participant. In this test, 135 people (106 males and 39 females) participated. Each image was tested on 15 observers. Their results illustrate that disparity makes saccade length shorter; however, it does not affect fixation durations [26]. They inferred that texture complexity is significant in salient area selection.

Gelasca et al. [12] designed a subjective experiment to investigate what colors attract human attention more. The goal of this experiment was to quantify the color saliency and to provide a ranking for some of the most common colors. 11 people participated in the test (3 females and 8 males, aged 19-28). They selected 12 colors including red, pink, magenta, violet, yellow, orange, green, cyan, blue, light blue, maroon, and dark green. The tested colors were chosen in the *CIE Lab* color space but there is no available information about their range. The experiment consisted of two cycles. During the first cycle, 20 synthetic images were presented to the subjects containing 12 colored disks. In the first cycle, the task was to choose at first three or four colors which each subject considered the most salient among the displayed colors. Afterward, the same images were shown, but subjects were asked to choose only one or two colored circles which attracted their attention most. They also repeated the experiment for the images containing four colored disks to confirm their results. Table 2.1 shows their results as a ranking table for 12 tested colors.

Based on the results, they divided colors into two overall groups. The colors that had much more priority were red, yellow, green and pink. The colors with lower saliency were reported as light blue, maroon, violet and dark green.

Table 2.1: Ranking for colors according to Gelasca et al. [12].

Selected Color Name	Overall sum of hits per color
Red	128
Yellow	87
Green	84
Pink	60
Orange	44
Blue	32
Cyan	32
Magenta	26
Light Blue	16
Maroon	14
Violet	11
Dark Green	10

Banitalebi-Dehkordi et al. [14] generated a 3D stereoscopic video dataset as a benchmark for saliency detection and video quality assessment purposes. They performed an eye tracking experiment to verify their designed visual attention model. They used a SensoMotoric Instrument (SMI) iView X RED system for eye tracking. A 46-inch Hyundai (S465) 3D-TV was used for displaying the test video contents. This TV set utilizes passive glasses in 3D mode. The distance between TV and the eye tracker was about 123 cm. The distance between the TV and observers was around 183 cm. In the test, 24 participants attended (12 male and 12 female). The eye tracking test contained two 12-minute sections separated by a 10-minute resting period. The test was a free viewing task.

Here, we provide more detailed findings and information from the literature about a number of bottom-up attributes consisting of color, contrast, texture, and motion.

A. Color: The color attribute has been investigated more than the other bottom-up factors in previous work [27–30] and it was shown that the warm, and saturated colors within a scene are generally more attractive for the HVS. In terms of the saliency of the colors in a scene, the human brain operation is very similar to contrast detection which means the colors that are more different from their surrounding areas

are most likely to become distinctive for our visual system [10].

B. Contrast: The contrast of a region versus its surrounding area. Some previous work [31], [32] stated that the absolute luminance value does not have an important impact on the human attention system. In fact, luminance is converted to the contrast in an early stage of HVS.

C. Texture: Many studies indicated that texture attribute does not have a significant influence on human visual attention [10]. However, rough textures are likely to be fixated more than smooth textures [33], [34].

D. Motion: Motion is one of the strongest attributes that absorbs human attention. It has been found that the influence of the motion in saliency detection is asymmetric which means it is easier to detect a moving object among stationary scene rather than identifying a stationary object amongst moving areas and objects [35]. In addition, detecting a faster object amongst slow objects is easier compared to the opposite state [36].

According to the literature [37–45], the contribution of low-level visual features is not very significant to determine fixation/gaze location in a static scene. However, transient features such as motion in dynamic scenes may result in a greater influence on gaze location with a higher consistency among viewers. For example, in 2011, Mital et al. [46] performed a study about gaze clustering during dynamic scene viewing. They investigated the impact of low-level (motion) and mid-level (corners and orientations) visual attributes on gaze location during a free-viewing task on a video dataset [46]. Their results show that mid-level visual features can distinguish between actual gaze locations and a randomly sampled baseline [46]. However, temporal features such as motion, flicker (difference-over-time), and their respective contrasts were the most informative features in predicting of gaze location [46]. It is still an open question whether this influence is involuntary or arising from the influence of high-level factors such as scene semantics.

The recent employment of dynamic visual features such as motion and flicker in designing VAMs has enhanced their ability to estimate salient regions and human gaze locations compared to models which only utilize static features, such as color, intensity, and edges [47–54]. Now, we are able to have a better understanding of designed VAMs in the literature. In the following, we briefly describe the 'state-of-the-art' of the existing main saliency detection algorithms for 2D image and video datasets.

2.4 Analytical Background

In 1998, the earliest method for saliency detection was proposed by Itti et al. [55] who is one of the pioneering researchers in the field. He established this model based on Treisman's feature integration theory [19] and Koch's biological structure [56]. This visual attention model employs bottom-up features including color, intensity (brightness), and orientation and calculates the local contrast of these features using the difference of the feature vectors in a center-surround neighborhood mechanism [55]. Then, they used a winner-take-all (WTA) network among different feature maps to obtain the ultimate saliency map.

In 1999, Rosenholtz et al. [57] proposed a simple saliency detection model based on statistical theories. They assumed that HVS is equipped with a bottom-up mechanism for detecting unusual items in a scene, e.g. items with an unusual movement compared to their neighborhood areas [57]. They defined saliency recognition as a parametric test for outliers in a statistical distribution.

Later, Itti et al. in [58] investigated the contribution of low-level saliency to human eye movements in complex dynamic scenes. They used an eye-tracker to record subjects' eye movements while watching 50 video clips. They computed instantaneous saliency using a model of bottom-up visual attention within the interval of starting each saccade and the endpoint location of that saccade [58]. Their experimental results confirmed that motion and temporal change were strong predictors for human saccades.

Itti et al. in [59] introduced a realistic avatar eye and head animation using a neurobiological model of visual attention system. In fact, their saliency detection model was extended from videos with static backgrounds to dynamic video scenes. They introduced two features such as flicker and motion. Flicker is defined as the absolute difference between the luminance of the successive frames that detects temporal change (e.g., onset and offset of lights). They used Gabor pyramids to compute the motion which was defined as the shifted differences between Gabor pyramids from successive frames. Their additional extensions concerned the generation of eye and head movements from the covert attention output of the model [59].

In fact, there is no specific boundary to discriminate among these criteria in some cases. There are many hybrid models derived by a combination of different criteria. in terms of the model of saliency detection, they can be divided into biological based methods and/or pure mathematical based algorithms, however, the hybrid algorithms

are used more. In terms of visual attention mechanism, they can be divided into bottom-up and top-down approaches. According to the processing domain, they can be split up into spatial domain and frequency domain methods.

Here we present a systematic review of major attention models. Several factors were first introduced by Borji et al. [1] to categorize these models including 1) bottom-up and top-down factors, 2) Intrinsic and extrinsic cues, 3) Pixel based, block based, or region based methods. Therefore, attention models can be classified according to these factors.

2.4.1 Bottom-up and Top-down Factors

Bottom-up based methods use low-level features such as colors, edges, contours, textures to extract the salient part of an image/ frame which are fast and pre-attentive processing task. Top-down based methods use high-level features such as a specific object or texture, human face, animal, vehicle, and so on. Top-down approaches tend to be slower than bottom-up based methods because of the complexity and computational cost of the object detection stage. To extract high-level features, we require having the previous knowledge from data or scene. Basically, top-down methods have been less explored since they are very complex due to high-level feature extraction demand and need to be aware of the data context and salient part of an image/ video.

According to the extraction approach for visual features including both bottom-up and top-down, existing methods can be classified in different ways such as intrinsic cues versus extrinsic cues based techniques and block based versus region based algorithms.

2.4.2 Intrinsic Cues versus Extrinsic Cues

In the existing works, two main cues are used to extract the saliency map from an image including intrinsic and extrinsic cues. Intrinsic cues are extracted only from the input image/frame itself while the extrinsic cues exploit other extra sources such as user annotation, depth map, statistical information of the similar images (i.e. learning based methods), light field images, and convexity (concavity) context [1]. Extrinsic cues are added to facilitate the process however they are more complex and time-consuming.

2.4.3 Block based versus Region based Methods

Two major tendencies of the existing algorithms for saliency detection in spatial domain can be divided into block-based and region-based algorithms. In block-based methods, the input image is first divided into blocks (patches) with the same or different sizes and then image processing algorithms are applied to each block to measure their features. In region-based methods, the image is segmented into regions first. Next, different feature sets are extracted from each region and compared to each other. Therefore, region-based methods utilize a segmentation preprocessing stage in order to reach different regions. The performance of region-based methods is basically better than block-based ones that makes them more preferable especially to calculate contrast within a frame. This is because of the following reasons: 1) The number of regions is usually smaller than blocks in an image which makes the comparison task among regions faster and easier. 2) The image is decomposed to the homogeneous areas in region-based methods that makes the saliency detection more qualified while blocks are not necessarily homogeneous [1]. However, the segmentation stage is time-consuming and makes the comparison between different regions more computationally complex.

2.5 Assortment of VAMs

In this section, VAMs are explained based on their mechanism to obtain saliency. Saliency detection was first proposed for still images, but with the progress in the relevant algorithms, researchers also started to segment and detect the distinctive regions within video sequences.

A variety of VAMs exist in the state-of-the-art for images and video in both 2D and 3D formats. In general, based on the mathematical techniques and mechanisms used to detect salient part within image/video, the existing methods can be categorized into six different groups including cognitive models, Bayesian models, information theoretic models, graphical models, spectral analysis models, and pattern classification models [1]. Spectral analysis models are in the frequency domain and the rest of these models are considered as spatial domain based VAMs. Some models may fall into more than one category.

2.5.1 Cognitive Models

Almost all attentional models are directly or indirectly inspired by cognitive concepts [1]. The ones that are more influenced by psychological or neurophysiological findings are described in this group. These models (e.g. [52,60–64]) exploit attentional concepts about a visual scene. Cognitive models take advantage of the biological base of human visual attention that helps us understand computational principles or neural mechanisms.

Itti's work [55] is a good example of this category. They generated multi-scale saliency maps by sub-sampling an input image into a Gaussian pyramid. This refers to HVS' ability to see a scene in different scales by changing the head position and consequently vision angle. Itti used center-surround method to calculate the difference of features including color, intensity, and orientation among areas. The center-surround approach is a biological mechanism because the human brain compares each area with its surrounding with a similar manner.

Another cognitive model is Le Meur's work. Le Meur et al. [52] proposed a model based on HVS structure such as contrast sensitivity, center-surround searching mechanism, hierarchical decomposition, and visual masking. They extracted three saliency maps: achromatic, chromatic, and temporal maps using low-level features and combined them into a saliency map. They later extended this model to the temporal domain. They contributed in detecting motion contrast as well as applying adaptive normalization to fuse different saliency maps [52].

Navalpakkam et al. [61] included significant aspects of biological vision in their model such as determining task-relevance of an entity and incrementally building a visual map of task-relevance at every scene location [61]. They modeled visual search as a top-down gain optimization problem by maximizing the signal-to-noise ratio (SNR) of the target versus distractors instead of learning fusion functions.

Kootstra et al. [62] developed their model based on three symmetry operators such as isotropic, radial, and color symmetry and showed that local symmetry plays an important role in guiding eye fixations by comparing it with eye tracking data. As HVS processes the scene on multiple spatial scales, therefore, they extended the operators into multi-scale symmetry saliency models by exploiting Gaussian pyramids. Finally, the feature maps were normalized and combined into a unique map [62].

Perazzi et al. [64] proposed a model based on a perceptually homogeneous decomposition of the input image and considering the uniqueness and spatial distribution of

the decomposed elements. This algorithm was classified as contrast based cognitive model. First, an input image was decomposed into compact and homogeneous elements that unnecessary details were suppressed. After, two measures of contrast for uniqueness and the spatial distribution of the elements were computed. A saliency map was estimated from the element contrast.

Achanta et al. [63] used the low-level features of color and luminance and applied a global symmetric center-surround mechanism to detect saliency map. Their saliency value was computed using the Euclidean distance between the average CIELAB vector of all pixels and each pixel of a Gaussian-blurred version of the same input image [63]. To obtain a symmetric surrounding for each pixel, they adjusted the bandwidth of the center-surround filter.

Several VAMs, especially in this group, employed a multi-scale center-surround mechanism to produce a saliency map. However, multi-scale methods either increase the computational expense or extract the backgrounds as salient areas because they often do not have previous knowledge about the size or location of the salient area/object which is a significant drawback.

2.5.2 Bayesian Models

Bayesian models such as [65–71] are based on the Bayes' rule which is a statistical model. These models probabilistically combine prior constraints/knowledge (e.g., scene context or gist) with sensory information (e.g., target features) about dataset to detect an area or object of interest [1]. Therefore, they can be applied to both bottom-up and top-down visual attention models.
A key privilege of Bayesian models is their ability to learn from data and their ability to unify many factors in a principled way [1]. Bayesian models can benefit from the statistics of scenes or other features that attract human attention.

Elazary et al. [65] proposed a top-down based model that not only can learn the preferred objects, but it can also tune feature detectors. They used the logarithm of the probability of being salient for a point in the presence of target components.

Torralba et al. [66] proposed a Bayesian framework for visual search tasks. They used context based prior on object class, location, and scale for searching salient objects. This model relays on the global scene configuration by presenting the fact that statistics of low-level features within an image can predict object location, scale, and pose before exploring the image. Therefore, visual context information should be

available early in visual processing. They formulated object detection problem as the probability of the presence of the object O, given a set of local and contextual measurements: $P(O|v_L, v_C)$, where v_L and v_C show location and contextual information respectively.

Zhang et al. [67] introduced a model known as SUN (Saliency Using Natural statistics) which is a definition of saliency by considering what the visual system is trying to optimize when directing attention. In their proposed Bayesian framework, bottom-up saliency appears as the self-information of visual features, and overall saliency emerges as the point-wise mutual information between local image features and the search target features. This model provided a general framework for many models. For example, the SUN formula for bottom-up saliency is similar to the work of Torralba [66], Oliva et al. [68], and Bruce and Tsotsos [72] such that they are based on the self-information concept (i.e., local information).

Zhang et al. [67] established their model based on the assumption that states the visual system must estimate the probability of a target at every location given the visual features observed. Assume z denotes a pixel in a visual field. Let the binary random variable C denote whether a point belongs to a target or not, let the random variable L denote the pixel coordinates, and let the random variable F denote the visual features of a pixel. Then, saliency of a pixel (S_z) can be defined as: $P(C = 1|F = f_z, L = l_z)$ where f_z and l_z indicate the feature value and the location respectively.

Using the Bayes rule and assuming that features and locations are conditionally independent given $C = 1$, then the saliency of a point is [67]:

$$\log S_z = -\log P(F = f_z) + \log P(F = f_z|C = 1) + \log P(C = 1|L = l_z) \qquad (2.1)$$

In the Equation 2.1, the first term at the right side represents the self-information (bottom-up saliency) which is determined by the visual features observed at the point z. The second term is the log-likelihood, which supports feature values that are consistent with prior knowledge about target features or objects (salient area) [67]. For example, if a target is supposed to be in red, the log-likelihood will take larger values for a red point compared to other colors. The third term represents the location prior which obtains top-down knowledge of the target location and is independent of visual features of the object [1]. For instance, this term may take knowledge of some target often located in the top-center part of a frame.

Later, Zhang et al. [69] extended their SUN model to dynamic scenes by introducing temporal filters (Difference of Exponentials). In this model, the estimated distribution for each filter response was fitted to a generalized Gaussian distribution [1]. In the first step, a bank of spatiotemporal filters was applied to each video frame. Then, features of video data were calculated and the bottom-up saliency for each point was estimated. The probability distributions of these spatiotemporal features were learned from a set of videos of natural environments [69].

Wang et al. [70] proposed a saliency detection model for 3D images using the contrast of two features of color and depth within a Bayesian framework. The depth feature map is constructed based on super-pixel contrast considering spatial priors. They estimated the density of depth-based and color-based contrast features using a Gaussian distribution in super-pixels. They used a discriminative mixed-membership naive Bayes (DMNB) model to calculate the final saliency map. The Gaussian distribution parameter can be estimated in the DMNB model by using a variational inference-based expectation maximization algorithm [70]. Their results are reliable but the complexity of the algorithm is high.

Rahtu et al. [71] introduced a salient object segmentation method based on combining a saliency measure with a conditional random field (CRF) model. This model employed a statistical Bayesian framework and local contrast of color, and motion features. To recover well-defined salient objects, a CRF model was applied to the resulting saliency map using an energy minimization-based approach. This approach is very time-consuming.

2.5.3 Information Theoretic Models

These models [72, 73, 75–82] are focused to detect the most informative part of an image/video in order to maximize the results of saliency computation. Therefore, perceptual systems are used to make the decision about the states of the surrounding environment to find the most informative part and discard the rest [1]. Therefore, these models cannot prioritize saliency of different regions.

Bruce et al. [72] designed an information theoretic model based on Shannon's self-information measure to find the most informative areas. This model known as AIM (Attention based on Information Maximization) and represents the information of a visual feature X as: $I(X) = -log(p(X))$, which inversely depends on the likelihood of observing X (i.e., $p(X)$). To estimate $I(X)$, the probability density function $p(X)$

must be estimated. They used Independent Component Analysis (ICA) to estimate probability density function of the RGB features. Then, the probability of observing the RGB values in a patch can be measured by multiplying of the likelihood of corresponding ICA basis coefficient for that patch.

Rosenholtz et al. [57,73] designed a model of visual search which could also be used for saliency detection over an image in a free-viewing task. They extracted features of each point in an appropriate uniform feature space (e.g., uniform color space). Then, distractor features parameters like mean μ, and covariance Σ, were computed from the distribution of the features. The target saliency in this model was defined as the Mahalanobis distance Δ, between the target feature vector T, and the mean of the distractor distribution as Equation 2.2.

$$\Delta^2 = (T-\mu)^{'}\Sigma^{-1}(T-\mu) \tag{2.2}$$

This model is similar to [66], [67], [74], since it estimates the rarity of a feature or self-information as $1/P(x)$ where x indicates each image location.

Li et al. [75] proposed a conditional saliency method for both image and video datasets. In this model, the saliency of a region was evaluated by minimum conditional entropy in which the uncertainty of a local region given its surrounding area was minimized to compute the saliency value [75]. To simplify the problem, they approximated the conditional entropy by the lossy coding length of multivariate Gaussian data.

Seo et al. [78] presented a unified framework for both image and video saliency detection. This method is a bottom-up based approach and computes a matrix of local regression kernels at each pixel to measure the similarity of that pixel in comparison with its surrounding area. They used a self-resemblance measure to estimate visual saliency. Matrix cosine similarity was employed to measure the resemblance of each pixel versus its surrounding.

The model proposed by Hou et al. [79] is a very well-known saliency detection method. It is a dynamic visual attention model based on the rarity of features. They introduced the Incremental Coding Length (ICL) to estimate the rarity of the features by measuring the respective entropy gain of each feature. ICL can present the energy distribution in the attention system. According to the definition of ICL, these features may evoke entropy gain in the perception state and are therefore assigned high energy. This model aims to maximize the entropy of the sampled visual features.

To optimize energy consumption, the limit of the energy is re-distributed between features according to their ICL. By selecting features with large coding length increments, the computational system can achieve attention selectivity in both static and dynamic scenes [79].

Mancas et al. [76] hypothesized that minority features in an image attract the HVS more. This method is very similar to Shannon's self-information measure. Basically, similar image areas were counted by analyzing their histograms. In this way, the spatial relationships of areas surrounding each pixel (e.g., mean and variance) were considered [76]. They introduced two types of global and local rarity models. Local contrast or rarity considers some image details that appear as salient and global rarity considers the uniqueness of features over entire image [76]. They used a multi-scale approach for the computation of local contrast.

Another example of this category is Yan et al. [81]. They proposed a saliency model based on sparse coding technique. From a cognitive science perspective, image information can be decomposed into two parts: redundancy and saliency. Redundancy part corresponds to a statistical invariant property (i.e. information with high regularities) of visual inputs and the saliency part denotes the rarity of visual inputs [81]. Yan's model consists of two steps: first, the over-complete sparse bases were learned to represent image patches; and then, direct low-rank and sparsity matrix decomposition were used to estimate the saliency information [81].

Zhu et al. [82] computed both local and global center-surround contrast between multi-size super-pixels. The super-pixels were modeled by multivariate normal distributions in CIE Lab color space. First, the global saliency was computed by probabilistically grouping similar super-pixels into the same clusters and their compactness was rated. The Wasserstein distance on the Euclidean norm (W2) was employed as a distance measure to determine local center-surround contrasts between super-pixels. Then, the local similarity between elements was considered by using a random walk technique to balance the saliency probability of probable object candidates [82].
The Wasserstein distance represents the minimum cost of transforming one distribution into another. It considers two factors: 1) the individual difference in each point of the underlying metric space, and 2) how far one has to shift probability masses [82].

2.5.4 Graphical Models

Graphical models [83–89] are constructed by incorporating graph theory, probability theory, and machine learning methods [1]. These models represent the conditional independence relationships between random variables. Attention models of this category consider eye movements as a time series. Eye movements are generated by the impact of hidden variables, therefore approaches like Hidden Markov Models (HMM), Dynamic Bayesian Networks (DBN), and Conditional Random Fields (CRF) have been unified to process them [1]. Since graphical models are the generalized version of Bayesian models, they can model complex attention mechanisms over space and time [1]. However, the complexity of these models is high especially in the training stage which is a considerable shortcoming.
Graph theory-based models always employ block-based approaches and take frame blocks as nodes. Then, graph edges are determined as the weights among blocks according to visual low-level features such as color, orientation, intensity, and so on.

Harel et al. [84] model known as graphical based visual saliency (GBVS) is a well-known example of this category. They extended Itti's model by utilizing a graph-based method to accurately measure the dissimilarity of the different regions of a scene. In this model, feature maps were extracted at multiple spatial scales. Image features like intensity, color, and orientation were used to construct a scale-space pyramid [84]. Then, a graph was built over all grid locations of each feature map. In the graph, weights between two nodes were computed based on the similarity of feature values and their spatial distance [84].

The dissimilarity between two positions (i, j) and (p, q) in the feature map, with respective feature values $F(i, j)$ and $F(p, q)$ can be defined as:

$$d((i,j) \;||\; (p,q)) = \left| log \frac{F(i,j)}{F(p,q)} \right| \tag{2.3}$$

Therefore, the assigned weight to the edge from node (i, j) to node (p, q) is proportional of their dissimilarity and their distance on lattice F:

$$w((i,j),(p,q)) = d((i,j) \;||\; (p,q)) \;.\; L(i-p, j-q) \tag{2.4}$$

$$L(a,b) = exp\left(-\frac{a^2+b^2}{2\sigma^2} \right) \tag{2.5}$$

The resulting graphs were treated as Markov chains by normalizing the weights of the edges of each node to 1 and finally, different maps were combined into a unique overall saliency map.

Salah et al. [83] proposed an approach which consists of modeling three steps in the human attention system. In the attentive level that decides where to look, a saliency map was produced using bottom-up features. In the intermediate level which analyzes the content of the fovea; the image was divided into uniform regions and supervised training was performed over each region. Finally, as the associative level this information was combined with a discrete Observable Markov Model (OMM). In fact, the observed regions by a fovea were assumed as states of the OMM. They applied this model to the handwritten digit and face recognition.

Pang et al. [85] proposed a stochastic model for visual saliency in video data using a dynamic Bayesian network. This model includes four layers: 1) A saliency map (Ittis) that shows the average saliency response in each location of a video frame. 2) A stochastic saliency map that converts the saliency map into natural human responses through a Kalman filter. 3) An eye movement pattern that predicts the human viewing patterns using a Hidden Markov Model and, 4) An eye focusing density map that predicts positions that people likely pay more attention [85].

Chikkerur et al. [86] proposed a framework that exploited both top-down and bottom-up attributes. They first developed a simplified model of attention in the brain to identify the primary attended areas and their interconnections. Then a Bayesian network was proposed where each node had direct neural correlates within the simplified biological model. Finally, the properties of the resulting model were clarified to demonstrate that the predictions are consistent with physiological and behavioral evidence. The physiological data (neural responses in the ventral stream (V4 and PIT) and dorsal stream (LIP and FEF)), as well as psychophysical data (human fixations in free viewing and search tasks), can be explained by this model [1].

Wang et al. [87] proposed a saliency detection method based on a directed graph model and multi-scale Bayesian inference. In this model, a directed graph was generated with super-pixels as the nodes of the graph. Then, a baseline node was introduced so that its saliency value is considered to be zero. They defined the saliency of each node as the shortest distance from the baseline node to that node and finally, Dijkstra's algorithm was adjusted to solve this optimization problem (Dijkstra's algorithm is an algorithm to find the shortest paths between two nodes in a graph).

Furthermore, they extended this model by developing a Bayesian inference strategy to achieve pixel-level saliency.

Cheng et al. [88] presented a region-based global contrast algorithm using graph theory and spatially weighted scores to find the salient objects. Their contrast method employed an unsupervised graph-based segmentation algorithm as the first step. Then, the color histogram was built for each region. Finally, the saliency value of a region was computed by measuring its color contrast to all other regions in the image.

Wang et al. [89] presented a spatio-temporal saliency detection method based on the gradient flow field and energy optimization. Their proposed gradient flow field incorporates two distinctive features: 1) intra-frame boundary information and 2) inter-frame motion information together to indicate the salient regions. They utilized both intra-frame and inter-frame information in the gradient flow field to estimate the object and background in a scenes and they tried to suppress the background. This method also introduces local and global contrast saliency measures using the foreground and background information estimated from the gradient flow field. They further proposed an energy function to achieve the spatio-temporal consistency of the output saliency maps [89].

They first applied the well-known SLIC method to abstract each frame into superpixels to investigate a frame region by region for both contrast and gradient flow field features. In the spatial domain, the color gradient magnitude of the abstraction frame was computed. To estimate motion, the optical flow field of [90] was used and then the magnitude of the gradient of the flow field was computed. Finally, both color gradient magnitude and optical flow gradient magnitude were combined using multiplication into a spatio-temporal gradient field. Furthermore, the motion field was emphasized by an exponential weighting function.

They further defined a global saliency measure of a superpixel as the length of its shortest distance to the virtual backgrounds. The distance between any two superpixels (i.e., regions) considers the color distance and the gradient flow field distance. In this way, the method can suppress background and detects only the salient objects which belong to the foreground.

2.5.5 Spectral Analysis Models

Spectral models [91–93, 95–97] detect the saliency in the frequency domain instead of the spatial domain. Frequency domain based VAMs use frequency spectrum to detect the relationship between saliency attributes. They are fast compared to spatial domain models which makes them suitable for real-time applications.
These models are simple from a mathematical perspective and easy to implement. However, they may lead to having strong edges in the saliency maps rather than extracting the entire area of the salient parts. Also, their biological plausibility is not still very clear [1].

The first frequency domain approach was proposed by Hou et al. [91] in 2007 named as Spectral Residual (SR). The theory behind this model indicates that image information can be distributed into two parts: saliency and redundant information. Similarities within an image indicate redundant information represented by local linear parts of log spectrum curve. Therefore, parts of the curve that jump out of the smooth manner, can be salient information and statistical singularities in the spectrum can be responsible for unique regions of the image [91].
The SR of an image was extracted using its logarithm spectrum and then the saliency map was created by transforming the spectral residual to the spatial domain. First, a two-dimensional discrete Fourier transform was applied to an input image. Then, the amplitude of the averaged spectrum was calculated and finally, the logarithm of amplitude was obtained.

Guo et al. [92] later found that the equivalent results can be achieved by computing the inverse Fourier transform of the phase spectrum alone and discarding the amplitude. Therefore, they showed that the computation of the residual is not required. This approach is known as phase spectrum of Fourier transform (PFT) and was extended to video sequences using quaternion (i.e., Hyper-complex) Fourier transform (PQFT) model by Guo and Zhang [93].

The quaternion Fourier transform (QFT) is an important tool in multi-dimensional data analysis. Four-dimensional (4D) quaternion algebra H is defined over $\Re$ with three imaginary units (i.e. basis elements i, j, k) as follows [94]:

$$ij = ji = k, \quad jk = kj = i, \quad ki = ik = j, \quad i^2 = j^2 = k^2 = ijk = -1 \tag{2.6}$$

Every quaternion (hyper-complex matrix) can be written as:

$$q(n,m) = q_r + q_i i + q_j j + q_k k \quad \in H \quad and \quad q_r, q_i, q_j, q_k \in \Re \tag{2.7}$$

The discrete version of QFT of 2.7 is given by:

$$F_H[u,v] = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\mu.2\pi \left(\frac{mv}{M} + \frac{nu}{N}\right)} f(n,m) \tag{2.8}$$

where μ is a unit pure quarternion and $\mu^2 = -1$. It should be noted that $F_H[u,v]$ is also a hyper-complex matrix. The inverse QFT is given as:

$$f(n,m) = \frac{1}{\sqrt{MN}} \sum_{v=0}^{M-1} \sum_{u=0}^{N-1} e^{\mu.2\pi \left(\frac{mv}{M} + \frac{nu}{N}\right)} F_H[u,v] \tag{2.9}$$

The only phase information of Fourier transformed video frames used to calculate saliency map. Their model is very similar to SR model [91] however, they included brightness and color features as well.

Achanta et al. [95] first filtered high-frequency components using a combination of the difference of Gaussian filters and then extracted saliency map. The resulting map had full resolution the same as the original image.
This is an advantage of the frequency domain models which can solve the resolution issues that some saliency models in a spatial domain suffer from. Visual attention models with multi-scale mechanism usually produce a saliency map with a lower resolution compared to the original image/frame.

Bian et al. [96] improved Guo's model by utilizing motion saliency factors such as discarding the redundant and frequently occurring features. They used spectral whitening as a normalization procedure in the construction of a map that only represents salient features and localized motion [96]. First, they low-pass filtered and sub-sampled a gray-scale input image. In the next step, a windowed Fourier transform of the image was calculated. The flattened or whitened spectral response was transformed into the spatial domain through the inverse Fourier transform and squared to accentuate salient regions [96]. Finally, the result was convolved with a Gaussian filter to model the spatial pooling operation of complex cells [96].

Li et al. [97] proposed a bottom-up mechanism for detecting visual saliency,

characterized by a scale-space analysis of the amplitude spectrum of input images. They showed that the convolution of the image amplitude spectrum with a low-pass Gaussian kernel of an appropriate scale is equivalent to saliency detector function [97]. A Hyper-complex (i.e. Quaternion) Fourier Transform was used to fuse multi-dimensional feature maps. The saliency map was achieved by reconstructing a 2D signal using the original phase and the amplitude spectrum, filtered at a scale selected by minimizing saliency map entropy.
The drawbacks of spatial domain-based saliency models are complexity, sensitivity to parameter selection, and high computational burden. On the other hand, they are very accurate to extract and segment the entire salient regions with well-defined boundaries, which makes them preferable models in comparison with the spectral domain models.

2.5.6 Pattern Classification Models

Pattern classification models [98, 100–106] use machine learning approaches to model the visual attention system. They train models from recorded eye fixation/tracking data or labeled salient regions by human users [1]. These models usually use both bottom-up and top-down factors to design a saliency detection algorithm. For example, they may learn faces or text which are top-down features. Recently, by increasing eye movement data and eye tracking devices, these models have been widely utilized in computer vision research community. However, data-dependency and slowness are the main drawbacks of these techniques that limit them from being a general and a real-time tool to construct a saliency map.

A model proposed by Fang et al. [98] is based on learning a set of discriminative subspaces to extract outstanding targets and suppress distractors. They use principal component analysis (PCA) on randomly selected image blocks. The candidate subspaces were constructed using selected principal components since they have impressive abilities to separate targets and distractors [98]. By projecting images onto subspaces, each image block was determined by its contrasts against randomly selected neighboring. Then, salient blocks were extracted by applying an optimization framework to learn the saliency model from subspaces that can separate salient targets and distractors [98].

Li et al. [100] introduced a probabilistic multi-task learning approach for saliency detection in dynamic visual data. They simultaneously considered bottom-up and

top-down features. In this Bayesian probabilistic framework, a stimulus-driven (i.e. low-level) components were modeled using multi-scale wavelet decomposition; while a multi-task learning algorithm employed high-level components to obtain the most interesting areas. The algorithm also presented adaptive fusion strategies to integrate the stimulus-driven and task-related components to construct the final saliency map [100].

Kienzle et al. [101] introduced a non-parametric bottom-up approach for learning attention directly from human eye tracking data. The model mapped an image block to a real value in a non-linear manner. The algorithm was trained to assign positive outputs on fixations and negative outputs on randomly selected image blocks. The saliency function was determined by its maximization of prediction performance on the observed data [101]. A support vector machine (SVM) was trained to determine the saliency using the local intensities in still images. For videos, a set of temporal filters were learned from eye-fixations to find the salient areas [101].

Peters et al. [102] presented an attention model for both static and dynamic scenes. It also incorporated both bottom-up and dynamic top-down task relevance features. They computed a bottom-up saliency map from low-level multi-scale visual features. Then, a low-level signature of the entire image was computed using top-down features. This model learns to relate different classes of signatures with the different gaze patterns recorded from human subjects performing a task of interest [102]. They demonstrated that a pixel-wise multiplication of bottom-up map with the top-down map results in a higher human gaze prediction performance [102].

Judd et al. [103] introduced a model based on training a supervised SVM from human fixation data using a set of low, mid, and high-level image features. They collected eye-tracking data of 15 subjects on 1003 images and used this database as training and testing examples. They also considered a center-bias factor in their model and showed its high performance in eye fixation prediction [103].

Ehinger et al. [104] proposed a model of search guidance (searching for people in a scene) which combines bottom-up and top-down features. They used a classifier that employed a scanning window approach to explore the image at different locations and scales. The classifier extracted a set of features from that window and applied a linear SVM to classify the window as an object or background classes. The features were a grid of Histograms of Oriented Gradients (HOG) descriptors.

Ramadan et al. [105] presented a model for spatiotemporal saliency detection.

They applied a pattern mining (PM) algorithm to construct spatiotemporal saliency map instead of combining spatial and motion cues. Discriminative spatiotemporal saliency patterns can be recognized from initial saliency maps computed in spatial and temporal domains. First, for each frame of video, a de-texturing method was applied to preserve only the general component of the image and discard undesirable edges. Next, the saliency map of those frames was computed in the spatial domain. On the other hand, the optical flow of the current frame was computed to extract its saliency flow. In addition, temporal super-pixels were used to generate input regions of the PM algorithm. According to the mined saliency patterns, seed points can be extracted for relevant background and foreground and their label information is propagated to generate the spatiotemporal saliency map. Finally, a motion of segmentation (MOS) step was performed by integrating the resulting saliency maps (with the appearance and location information) into an energy minimization framework in order to obtain the final labeling scheme for the processed frame.

In [106], Fu et al. used a co-saliency method to discover the common saliency among the multiple frames. They introduced a cluster-based algorithm for co-saliency detection. In the clustering process, global correspondence between the multiple frames was implicitly learned. Three visual attention cues: contrast, spatial, and corresponding, were devised to measure the cluster saliency. The final co-saliency maps were generated by fusing both the single image saliency and multi-image saliency. The method mostly uses bottom-up features without heavy learning procedure. They used this method for video foreground detection.

The saliency model in [106] consists of a two-layer cluster-based method to detect co-saliency on the multiple images. Given a set of images, this method starts by two-layer clustering. One layer groups the pixels on each image (i.e., single image), and the other layer associates the pixels on all images (i.e., multi-image). Then, the saliency cues are computed for each cluster, and the cluster-level saliency is measured. The measured features include the uniqueness (on single/multi-image), the distance from the image center (on single/multi-image) and the repetitiveness (on multi-image). The features were named as contrast, spatial, and corresponding cues, respectively. Based on these cues, the saliency value is computed for each pixel, which is used to generate the final saliency map.

2.5.7 Deep-Learning Based Models

With the advent of deep-learning based models since 2015, a new wave of attention models was designed that have been intensely employed to predict salient regions in scene [107–111]. These models can be categorized under the "Pattern Classification Models". These VAMs are very successful however they do not discriminate among bottom-up saliency, biases, and top-down factors that are typically used in training data. Furthermore, deep-learning based models do not explain psychophysical pieces of evidence [80]. They are like a black-box (no precise knowledge about the core algorithms), extremely data-dependent, and time-consuming.

Kruthiventi et al. [107] proposed a fully convolutional neural network known as DeepFix for predicting human eye fixations on images. In the proposed deep network, the potential of the "inception module" was employed to extract complex semantic features in a multi-scale manner. Each inception module contains a set of convolution layers with different kernel sizes operating in parallel [107]. Moreover, the ability of "filters with holes" was exploited to consider global context using large receptive fields. They introduce Location Biased Convolutional (LBC) filters as a technique to provide the deep network the ability to learn location related patterns such as center bias in a scene.

Huang et al. [108] introduced a model based on Deep Neural Network (DNN). "A DNN is a feed-forward neural network with constrained connections between layers, that take the form of convolutions or spatial pooling, besides other possible non-linearities" [108]. They fine-tuned the network with an objective function based on the saliency evaluation metrics and used information at multiple image scales.

Wang et al. work in [109] is a good example of this category which is based on a novel deep-learning video saliency model. They proposed a saliency model using a CNN-LSTM (Convolutional Neural Network-Long, Short Term Memory) network architecture to enable fast, and end-to-end saliency learning approach [109]. The attention mechanism explicitly encodes static saliency information, therefore LSTM can focus on learning more flexible temporal saliency across successive frames [109].

Li et al. [110] argued that a high-quality visual saliency model can be learned from multi-scale features extracted using deep convolutional neural networks (CNNs). A neural network architecture was used for learning this model, which has fully connected layers on top of CNNs responsible for feature extraction at three different scales. Then, a refinement method was employed to enhance the spatial coherence of

saliency results. Finally, multiple saliency maps of different levels of image segmentation were aggregated to improve the performance.

Zhu et al. [111] proposed a model via structured label prediction method. In other words, they represented region saliency with structured labels. The appearance features were learned in rectangular regions so that structural region representation can encode the local saliency distribution with a matrix of binary labels. They showed that the linear combination of structured labels can model the saliency distribution in local regions properly. To measure the consistency between a structured label and the corresponding saliency distribution, an adaptive label ranking algorithm was introduced using a Convolutional Neural Network (CNN) model. Finally, they utilized a K-Nearest Neighbor (K-NN) to improve graph representation for saliency propagation [111].

2.5.8 Other Models

There are few VAMs that took advantage of a mixture of these categories to design a saliency detection algorithm and some others do not fit into the aforementioned categorization.

For instance, Goferman et al. [112] designed a context-aware saliency algorithm by computing the contrast of color and intensity features in image blocks which is a combination of graph-based and cognitive models. They considered basic principles of HVS, supported by psychological evidence to design their algorithm. They utilized both local and global factors to be able to provide better performance. Color and contrast were used as local low-level attributes. Then, frequently occurring features were suppressed by global considerations of local extracted features. The context-aware term refers to the using of high-level factors, such as human faces [112]. Their algorithm has a promising performance, which makes it obtain the correct and precise salient areas for the hight rate of the tested images.

Zhai et al. [113] introduced a saliency detection model for video sequences. They computed the global contrast using the histogram to distinguish the most salient objects in an image. They detected feature points in consecutive frames, then they used Scale Invariant Feature Transformation (SIFT) to determine the correspondences between the interest points [113]. As the next step, Random Sample Consensus (RANSAC) algorithm was applied on those points to estimate and find the homographies of the moving areas. RANSAC is an iterative method to estimate parameters

of a mathematical model from a set of observed data including outliers. Therefore, it is considered as an outlier detection method which can be motion feature in this scenario.

Mancas et al. [114] proposed a model that only uses motion speed and direction which are dynamic features. They did not involve static features of the spatial domain in their model. Motion extraction has been performed by optical flows in Manca's model. Then, a spatiotemporal low pass filter was used to distinguish the consistent activities and discard the rest of the information. The results of the filtering stage were used to compute a histogram for each frame to find rare motions [114].

Rao et al. [115] proposed a model based on template matching mechanism. They passed a template of the desired target objects over the input image in a sliding manner. Saliency was computed as a similarity measure between the template and local image at each location [115].

Garcia-Diaz et al. [74] introduced a model known as the Adaptive Whitening Saliency (AWS). In their model, saliency measure was defined as the variability in local energy. First, an input image was converted to CIE Lab color space. The luminance (L) was decomposed into multi-oriented multi-resolution representation using Gabor filters bank [74]. They also used a multi-scale decomposition for color components $a*$ and $b*$ channels. Next, the multi-scale responses were decorrelated and a local measure of variability was extracted from them [74]. Finally, a unified measure of saliency was obtained by using a local averaging. Decorrelation was achieved by applying PCA over a set of multi-scale low-level features [1].

Ma et al. [116] proposed a model on video contents by incorporating both top-down factors and a classical bottom-up framework. They used semantic high-level cues e.g., face, speech, and camera motion. In this model, the video data was first decomposed into the primary elements of basic channels including image sequence, audio tracks, and textural informatio [116]. Then, saliency maps were generated separately using a set of attention modeling methods. Finally, different saliency maps were fused to achieve a comprehensive attention map. They applied this model to video summarization [116].

Rosin [117] proposed an edge-based scheme (EDS) for saliency detection over grayscale images. A Sobel edge detector was used to extract the edge features from an input image. Then, a thresholding method was applied to the gray level edge image at multiple levels to achieve a set of binary edge images [117]. Next, a distance

transform was employed to propagate the edge information from each of the binary edge images [117]. Finally, the overall saliency map was obtained by summation of the gray level distance transforms. This model has not been successful over color images in the extraction of the correct salient areas.

Chen et al. [118] introduced a spatiotemporal saliency detection method using low-level color and motion features. They first applied the SLIC clustering algorithm over input frames to produce super-pixels. Both local and global contrasts were computed to extract color and motion gradient feature maps among all frames. They used RGB color space and optical flow to compute super-pixel feature distance [118]. Next, the motion gradient and color gradient were combined to obtain the spatial-temporal gradient map to guide the low-level contrast computation. An element-wise Hadamard product was used to fuse color and motion saliency. However, the resulting saliency map not only contains many false detections but also the saliency distributions are not temporally consistent [118]. Hence, Chen et al. proposed a low-rank coherency analysis to boost the saliency map accuracy by keeping its temporal smoothness [118].

Nikitha et al. [119] introduced a saliency detection model for video data based on a pixel-wise difference within spatio-temporal neighborhoods. They first examined on spatio-temporal eye fixation information to understand a typical element in human visual consideration. The proposed algorithm was formed according to spatial, temporal, and both the spatio-temporal neighborhoods [119]. Next, a weight value was added to each pixel to produce the final saliency map. Unfortunately, there is no more description and information about how they performed their method, and which algorithms were used for each step.

Based on the aforementioned descriptions, it is essential to possess a solid knowledge about HVS properties and reaction in response to the bottom-up stimuli in order to design an efficient saliency detection algorithm for a multimedia dataset with the aim of video streaming. Therefore, we decided to compensate for the lack of a comprehensive investigation about the influence (i.e. in a ranking manner) of bottom-up stimuli on HVS by proposing this work.

2.6 Evaluation Metrics

Resulting saliency map S, from a VAM, must be quantitatively evaluated by comparing it against a ground-truth G, which is usually eye movement data, click positions, or drawn objects/areas by human subjects. It is necessary to have standard and solid metrics to compare these. There are different ways to perform this comparison and assess the performance of the saliency detection algorithms. For example, S and G can be assumed as probability distributions to be able to use Kullback-Leibler (KL) or Percentile metrics to measure the distance between two distributions. As another way, S can be considered as a binary classifier and a signal detection theory such as Area Under the ROC Curve (AUC) metric is used to assess the performance of this classifier. Moreover, S and G can be assumed as random variables and their statistical relationship can be measured using the Correlation Coefficient (CC) or Normalized Scan-path Saliency (NSS).

From another perspective, evaluation metrics for assessing the performance of saliency detection algorithms can be classified into three categories: 1) point-based, 2) region-based, and 3) subjective evaluation [1]. In point-based measures, salient points from estimated saliency maps are compared to ground-truth saliency maps made by combining eye fixations [136]. In region-based metrics, the saliency maps and labeled salient regions which are annotated by human subjects are compared against each other [100]. Subjective scores on estimated saliency maps were reported on three levels: "Good," "Acceptable," and "Failed" by Zhai et al. [113]. The main problem of subjective evaluation method are the difficulties and barriers for extending it to large-scale datasets.

In the following sections, we focus on explaining these metrics in more details.

Kullback-Leibler (KL) divergence

The KL divergence is a general information theoretic measure used to calculate the distance between two probability distributions. In the context of saliency, it is used to measure the distance between distributions of saliency values at estimated saliency maps and ground-truth maps [120]. To avoid future confusion about the KL implementation used in this work, we can refer to this as KL-Judd (Judd's version) as used on the MIT Benchmark [9, 121].

This metric takes an estimated saliency map P and a ground-truth fixation map

Q^D as inputs and evaluates the loss of information when P is used to approximate Q^D [121].

$$KL = \sum_{i=1}^{n} Q_i^D \ log\left(\epsilon + \frac{Q_i^D}{\epsilon + P_i}\right) \tag{2.10}$$

where ϵ is a regularization constant and i shows i^{th} pixel in a map. KL-Judd is an asymmetric dissimilarity metric, while a lower score indicates a better prediction of the ground-truth by a resulting saliency map of an attention model.

In fact, a resulting saliency map by an attention model is sampled (or averaged in a small vicinity) at both human saccadic and random points. The saliency magnitude at the sampled locations is then normalized to the range of $[0, 1]$. The histogram of these values in n bins across saccadic points is then calculated [120]. Finally, the difference between these histograms is computed.

The advantage of KL divergence over other metrics refers to its invariant property to re-parameterizations. Therefore, this metric is reliable and unaffected in response to applying continuous monotonic non-linearity operations over the saliency measure S such as $S^3, \sqrt{S}, e^S$ etc. [120]. One disadvantage of the KL divergence is the lack of a well-defined upper bound where the KL divergence approaches infinity if the two histograms never overlap [121].

Normalized Scan-path Saliency (NSS)

The Normalized Scan-path Saliency is a simple measure between saliency maps and ground-truth, computed as the average normalized saliency at fixated locations. It is defined as the response value at the human eye position, (x_h, y_h) in an estimated saliency map that has been normalized to have zero mean and unit standard deviation. NSS can be shown as [1, 120].

$$NSS = \frac{1}{\sigma_S}\big(S(x_h, y_h) - \mu_S\big) \tag{2.11}$$

where σ_S and μ_S indicate the standard deviation and the mean of the saliency map, respectively.

NSS is computed for each saccade, then the mean and standard error are computed across the set of NSS scores. The metric value of 1 indicates that the subjects' eye positions located in a region with a predicted density of one standard deviation above average [121]. On the other hand, $NSS = 0$ demonstrates that the model performs

no better than random. Unlike KL, NSS is not invariant to re-parameterizations. However, it is invariant to linear transformations like contrast offsets since the mean saliency value (μ_S) is subtracted. [121].

Area Under Curve (AUC)

The area under Receiver Operating Characteristic (ROC) curve is known as AUC. The ROC is the most popular measure in the saliency community and is used for the evaluation of a binary classifier system with a variable threshold. In the definition of this metric, the estimated saliency map is treated as a binary classifier on every pixel in the image. The pixels with higher saliency values compared to a threshold are classified as fixated while the rest of the pixels fall into the non-fixated class [121]. Human fixations are then used as ground truth. The ROC curve is drawn as the false positive rate versus true positive rate, by varying the threshold. The area under this curve represents how well the saliency map predicts actual salient areas and human eye fixations [120]. Perfect prediction corresponds to a score of 1.
This metric has the advantage of being transformation invariant so that applying monotonically increasing function to the saliency value does not force the area under the ROC curve to be changed [1].

2.7 Conclusion

In this chapter, previous experimental studies, as well as saliency detection techniques, were briefly reviewed. A classification of the VAMs was provided based on different criteria that aimed to categorize and determine their main trends.
In general, there is no specific boundary to discriminate among the criteria and there are many hybrid models derived by a combination of them to achieve better performance.

Usually, saliency detection methods are divided into two main directions from the feature extraction perspective including bottom-up and top-down based methods. In terms of the model of saliency detection, they can be divided into biological based methods and/or pure mathematical based algorithms however the hybrid algorithms have better performance. In terms of the processing domain, they can be designed in the spatial and/or frequency domain.

The main challenges in designing visual saliency models can be described as the

lack of a comprehensive study to investigate the significance bottom-up stimuli (i.e., color, texture, motion, speed, etc.) in a ranking system over HVS especially in dynamic scenes. Therefore, we designed a set of eye-tracking based experiments to compensate for this gap and designed a saliency detection model for 2D images based on a Bayesian framework. Since the main application of this work will be video compression, a fast and reliable model is required to be able to extract and segment the entire salient areas with high quality and well-defined boundaries.

In the next chapters, we will describe our designed experiment and saliency detection algorithm in details.

Chapter 3

Bottom-up Stimuli Test

3.1 Introduction

In this chapter, we explain our designed set of experiments involving bottom-up stimuli in detail.
Since saliency detection is very closely influenced by human visual perception, it is necessary to study the impact of visual stimuli on HVS. The lack of a comprehensive study in this area inspired us to design an experimental study in order to precisely and completely investigate the impact of bottom-up attributes including color, texture, and motion on HVS in terms of saliency detection problem. The main purpose of this work is to achieve a ranking system among different bottom-up stimuli to be able to design saliency detection maps and algorithms. Furthermore, our goal is to achieve the ability to assign the order of saliency for the entirety of an image/video frame rather than simply focusing on the most salient object/area.

3.2 Experimental Methodology

We hypothesized that color, texture, and motion influence visual attention as the Equation 3.1.

$$S \quad = \quad \alpha C \quad \Theta \quad \beta T \quad \Theta \quad \gamma M \tag{3.1}$$

where S indicates the saliency value, C, T, and M present color, texture, and motion respectively. Also, α, β, and γ are the weights that show the amount of importance of each stimulus in absorbing attention. It should be noted that $C = \{c_1, c_2, c_3, ...\}$,

$T = \{t_1, t_2, t_3, ...\}$, and $M = \{m_1, m_2, m_3, ...\}$, are sets of features in a ranked manner to prioritize each individual stimulus. On the other hand, the relative importance among colors, textures, and motions i.e. α, β, and γ can be computed based on the combined experiments.

It is expected that these will be a complicated and non-linear relationship between these three feature types, however, we show a general formulation on Equation 3.1 to be able to explain every possible circumstance. Θ represents the operator that determines how to combine or decide among different feature maps. This operator can be a summation, a multiplication, an averaging, an optimization problem, etc. We selected a Bayesian framework to combine/decide different bottom-up feature maps which will be described in Chapters 4 and 5. Since we must assign the priority of importance to different regions of a frame, a statistical framework (such as Bayesian) is expected to be the best candidate because it conditionally decides across various states.

Therefore, the designed experiment aimed to find those weights and orders for both individual assessments of stimuli as well as their combined states. We asked human subjects to watch image/video content datasets and their eye movements were recorded using an eye-tracking device. Then, we studied and analyzed the fixation and saccade points pattern and duration to estimate the pattern of the HVS to discover which stimuli are more effective on the attention system and how they change its operation. We obtained a ranking system to explain which colors, textures, motion directions, and movement speeds are most likely to be attractive for human vision.

We found a priority quantity for different ranges of each individual feature. In this way, we will know which features to focus on more on designing our VAM. We designed a set of experiments to investigate each attribute individually. Then, we tested a limited number of the permutations of the attributes.

In this study, we had 26 participants (12 females and 14 males) within ages 18-35. Participants viewed a 55-inch LG 3D-TV screen, and a tripod mounted eye-tracking device (an SMI eye-tracker iView 120 Hz) recorded their eye movements. Participants were expected to watch 39 images and 94 short videos, while their eye movements were recorded. These 2D images and video sequences were created using Adobe Illustrator and Adobe After Effects. This experiment was a free-viewing task and participants did not need to perform other tasks at the same time. We combined all images and videos together to generate a unified video with a duration of 16 minutes. Figure 3.1

shows a general schematic of the test set-up.

Figure 3.1: A general schematic of the designed test set-up.

We assessed the participants' vision in advance to avoid having vision issues, such as color blindness or low visual acuity. Also, we provided instructions to participants on how to act during the experiment. We trained them by performing a sample test to avoid misunderstandings and false data recording (as much as possible).
Our designed experiment contains five stages for each participant:

1) A Visual Test to Check any Vision Issues.

All the subjects were checked by pretests to assess their:

- Visual acuity using a Snellen chart
- Color blindness using Ishihara graphs

A Snellen chart was prepared to check participants visual acuity. We used an online Ishihara test website (i.e. Ishihara 38 Plates CVD Test) including 38 plates available at [122].
In our study, the subjects who did not pass either of the two vision pretests were not allowed to continue to the eye-tracking experiment.

2) A Verbal Explanation of the Instruction.

We instructed each participant on how to behave during the calibration stage, the training stage, and the actual test itself. We emphasized that they should keep their position without any movement during the test because their distance from the eye-tracking device and the TV must be fixed during the entire experiment. In addition, we asked them not to rotate their head while watching the video as well as being concentrated on watching the video. In this way, we attempted to avoid many problems that might happen during the test and make it inefficient by producing invalid data.

3) A Calibration Procedure.

The eye-tracker requires calibration for each participant in order to learn the characteristics of their eyes. The SMI iView 120 Hz has an automatic calibration software. During the calibration stage, a participant should look at and track a dot that appears and moves to different positions of the screen. We used a 9-fixation point calibration setting. The calibration will fail if participants' gaze points and eye movements cannot be recorded by the eye-tracker. Participants who did not provide proper data at the calibration stage were not allowed to participate in the test.

In our experience, people with high prescription glasses or droopy eyelids are most likely to have difficulty passing the calibration stage.

4) A Training Period.

We trained each participant before starting the main visual attention test to avoid facing misunderstanding during the actual test as much as possible. We prepared different image/video sets for the training stage and showed them to participants to help them get familiar with the test practically. These images or videos were not considered in our final results.

Based on our experience, only providing the instruction to the participants may not be enough. Some participants need to practice the actual test to know how to act during the test. Therefore, we decided to train them by showing a few sample images and video data in advance to have more valid and reliable recorded data.

5) A Visual Attention Test.

At this stage, we presented our produced datasets to the participants and asked them to do a free-viewing task by simply observing the images or videos. Then, we recorded their eye gaze and eye movements using the eye-tracker. The distance between subjects and the screen was approximately 218 cm (217.6 cm) which is around 3.2 times the display height based on the recommendations listed in ITU-R BT.2022 which is a standard guide on general viewing conditions for subjective assessment of image/video dataset on flat displays [123]. The distance between the eye-tracker device and subjects was 60 cm.
We followed ITU-R BT.2022 document in designing our experiment because it has been widely used in saliency detection literature for many eye-tracking based experiments.

The rest of this section provides a detailed explanation of different parts of our experiment. This experiment was broken down into both individual and combined feature tests to study bottom-up stimuli including a color test, texture test, motion test, contrast test, color and texture test, color and motion test, texture and motion test, and finally all three features together.

3.2.1 Color Test

In order to obtain reliable results, ideally, a large number of different colors should be considered. However, this would make the task too complicated with a huge number of different permutations of the color positions and causes eye fatigue in participants. Therefore, we selected 12 main colors in the HSV color space that is more compatible with the human vision and perceptual system.

We created 10 two dimensional images using Adobe Illustrator for our color test. Each image consists of 12 colored disks of the same size located on the circumference of a circle like a clock dial. The background was gray with the luminance of 120 in HSV. Since gray is a neutral color and has an average intensity difference with most of the colors compared to white and black, we chose it as the background color for our generated images. In this way, the high contrast occurrence can be avoided to decrease any possible contrast bias in color saliency.

To introduce our selected colors, we illustrated their HSV characteristics in Table 3.1. The selection of the named colors follows the previous research on color names by

Lindsey et al. [124]. We selected only pure and saturated colors - with the saturation of 100 - for our test because we needed to test the chroma without affecting it by any black, white, or gray tones - different levels of saturation. Since we considered the intensity feature within our texture test, therefore, saturation is out of the scope of the color test.

It is worthwhile to mention that experiments showed saturated colors are generally salient [125,126]. Tian et al. argued that highly bright and saturated colors are salient regardless of their associated hue value [126]. The reason is due to the sensitivity of the human eye photoreceptors to these types of colors.

Table 3.1: HSV Color Table.

Color Name	$H°$(Hue)	S%(Saturation)	V%(Value)
Dark Blue	240	100	50
Orange	30	100	100
Green	120	100	100
Blue	240	100	100
Cyan	180	100	100
Red	0	100	100
Turquoise	180	100	50
Pink	300	100	100
Purple	300	100	50
Yellow	60	100	100
Dark Green	120	100	50
Dark Red	0	100	50

We utilized disk shapes for our objects because of their simplicity and symmetry. Polygonal shapes may cause bias in choosing salient color since they have different angles and the viewer may see color in different tones due to vision error. However, disk shapes do not contain all angles and have a uniform shape. A sample of our created image with 12 colored disks is shown in Figure 3.2.

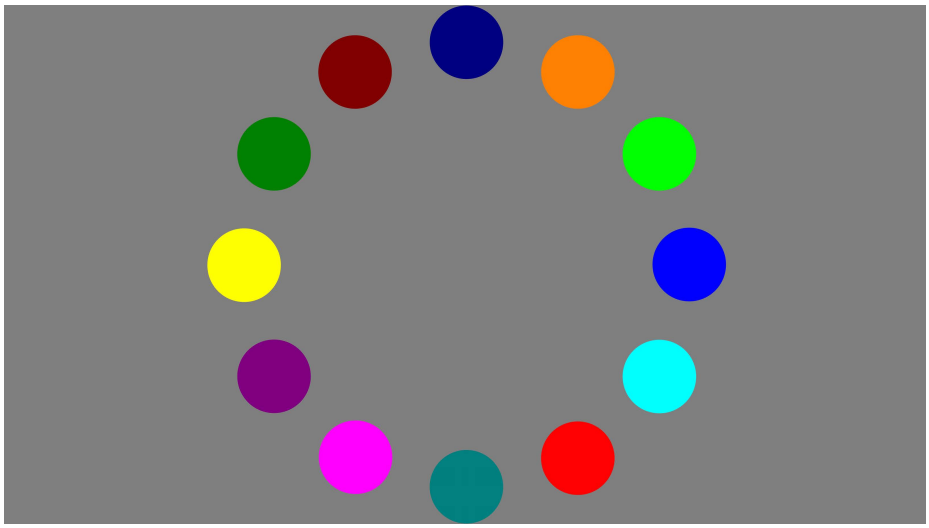


Figure 3.2: A sample image for experimental color test.

The colored disks were placed along a circle's circumference across 10 images in a way that different colors can be scattered along the circle and no high contrast happens because the contrast itself may cause some bias in absorbing attention which we will examine later in a different test. To avoid categorizing colored disks based on their HSV characteristic, we made sure not to have very similar colors beside each other. For example, we ensured not to have different tones of a similar color close to each other such as red and dark red. We ensured not to put all warm colors or all cold colors beside each other as much as possible.

The human visual system usually pays more attention to the center of the scene which is known as center bias. This fact is used in photography and film making strategies. According to this fact, we located the colored disks close to the center of the image with equal distance from the center that leads us to reach a circular circumference. Both displacement of the colored disks and having the same distance from the center of the display screen will provide an equal chance to each color to be selected as the salient one by human subjects.

As per the instruction of ITU-R BT.2022 [123], each image was presented to participants for 10 seconds and we embedded a plain gray image between two consecutive images for 3 seconds to clear the participants' gaze point.

3.2.2 Texture Test

For this test, we selected three different levels of texture from a complexity perspective (usually textures with too many mixed edges and different angles are considered as complex) including low, medium, and high complexity. The low complexity texture is known as the absence of a pattern on objects and low contrast. The medium-level contains simple geometrical patterns. Finally, the high-level appears with more complex geometrical patterns, higher contrast, and dense edges.

We created 10 two dimensional images for our texture test in a similar manner to the color test. Each image consists of 12 textured disks located on the circumference of a circle and the background is gray. Textures were selected from gray-scale images within similar intensity ranges to achieve more similar texture patterns from an intensity/brightness perspective. Otherwise, a contrast may happen that biases the human subjects in a specific textured disk and eventually saliency detection procedure. Therefore, we performed low-pass filtering (i.e., Gaussian) and histogram equalization as a preprocessing step to generate an approximately similar range of intensity for different textures. A sample image for the texture test is shown in Figure 3.3.

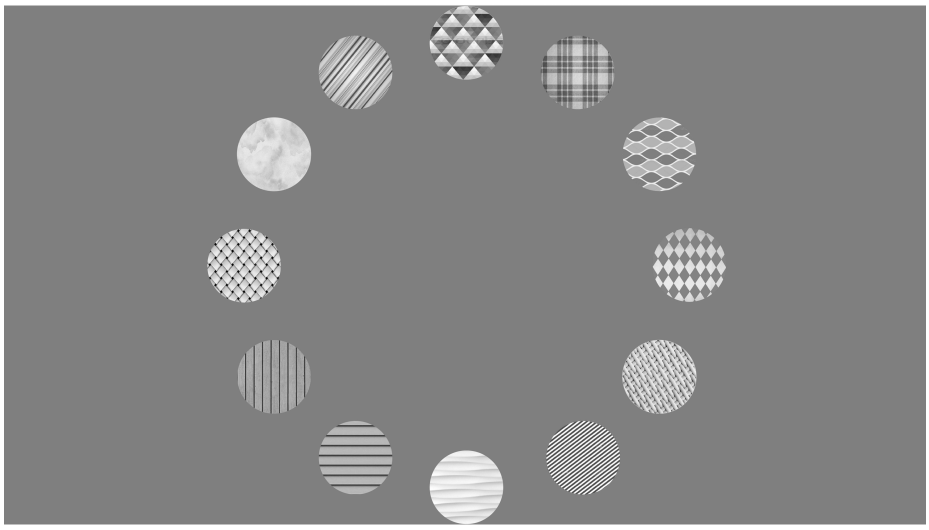


Figure 3.3: A sample image for experimental texture test.

In our produced images, each textured disk was located in different positions on the perimeter of a circle within 10 different images. According to the ITU-R BT.2022 document, we presented each image to participants for 10 seconds and a plain gray

image was embedded between two consecutive images for three seconds to avoid having the same gaze points from previous images.

3.2.3 Motion Direction and Velocity Test

We utilized four main motion directions - from both sides, i.e., eight in total - including horizontal on both sides (toward left and right), vertical on both sides (upward and downward), diagonal with an angle of 45 degrees in both sides (toward up and down), and diagonal of 135 (i.e. -45) degrees in both directions (upward and downward). We only used linear directions and no curvature movement was considered to avoid encountering complexity and having a large number of variations in our experimental study.

We exploited velocity and acceleration in our motion test to investigate their influence on the human attention system. To this end, we used two different motion patterns such as motion with the constant speed, and motion with acceleration. In addition, three different levels of velocity i.e. slow, medium, and fast were utilized to render 2D video sequences as the data for this section of the test.
We used white circles as the moving objects in a gray background. All the circles have the same size but different movement directions and speeds.

Rendered videos for this test are divided into three main groups: 1) moving circles with the same speed in different directions, 2) moving circles in the same direction with different speeds and different accelerations, and 3) moving circles in different directions with different speeds/accelerations.

We rendered 14 video sequences for the first group and then 7 and 5 video sequences for the second and third groups respectively. The average time duration of each video lasts almost 5 seconds. The average motion speed in our rendered videos was 960 pixels per seconds (24 degrees per second). The maximum and minimum speeds were 2203 (55 degrees per second) and 734 (18 degrees per second) pixels per seconds respectively.

Figure 3.4 shows a sample frame of the produced video for a motion test while objects have the same velocity with different directions in horizontal left and right, vertical upward and downward, and diagonal of 45 degrees upward and downward.

Circle shaped objects were also used in this test for their simplicity and help

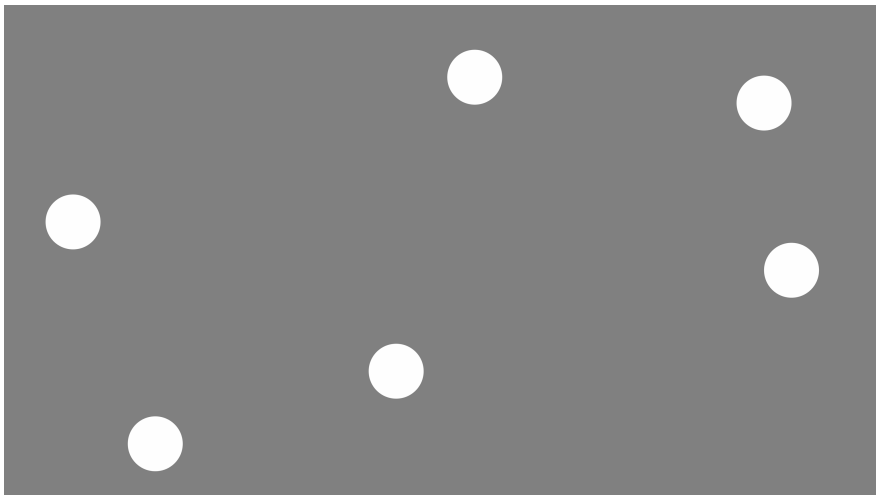


Figure 3.4: A sample frame of experimental 2D motion dataset.The objects have the same velocity but different directions in horizontal, vertical, and diagonal of 45 degrees, all three in both directions.

in preventing participants' confusion that can result from polygonal or other shapes. The white color was selected for objects because it provides a medium contrast related to the gray background.

3.2.4 Contrast Test

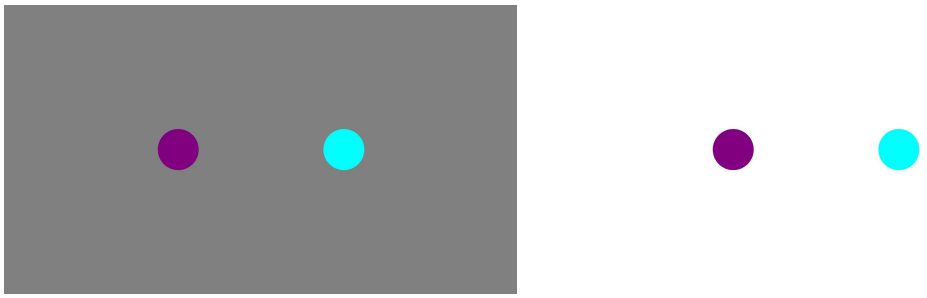
According to the saliency detection literature (as discussed in Chapter 2), contrast generally plays an important role for all kinds of attributes in absorbing human visual attention and consequently makes the areas with higher difference stand out and become more salient.

We designed a test for color contrast to investigate and demonstrate this hypothesis. For this purpose, we produced nine images. Each image contains two colored disks so that the background of five of them is gray and the rest four have a white colored background. The reason for creating images with two different background colors is to be able to test and consider the contrast between two colored disks as well as colored disks and background.

We noticed that the contrast between colored disks and background is more significant in the selection of the salient areas. Based on our hypothesis, when the luminance difference (i.e. contrast) between colored disks and background is similar and two disks simply have different colors, then viewers tend to see their favorite color

as more salient while both disks can be salient from contrast perspective.
On the other hand, if the contrast between disks and background is different, it is assumed that the most salient disk should be the one with higher contrast with respect to the image background even though it does not contain the viewer's favorite color.

Two samples of our rendered images for contrast test are shown in Figure 3.5. In this figure, part (a) illustrates a gray background consisting of two disks with purple and cyan colors located in the horizontal direction. In part (b), another image can be seen with the same colored disks but a white background.
In the image with the gray background, it is expected that cyan disk will be fixated on more and the purple disk in the white background image that will be discussed on the results section.



(a) (b)

Figure 3.5: A sample image for experimental contrast test.

Here, the combined tests for these stimuli are described in more detail. These tests include 1) color and texture, 2) color and motion, 3) texture and motion, and 4) color, texture, and motion.

3.2.5 Color and Texture Test

We created 10 two-dimensional images to combine different selected colors and textures. Each image consists of 12 colored and textured disks on the gray background with the same size located on the circumference of a circle similar to both color and texture tests. In order to achieve reliable and comprehensive results, we combined color and texture attributes to generate six different types of the disks including 1) salient colors and salient textures, 2) salient colors with non-salient textures, 3) salient textures with non-salient colors, 4) salient colors with average salient textures,

5) salient textures with medium-level salient colors, and 6) medium-level salient colors with medium-level salient textures.
The salient colors and textures are determined in section 3.3.
The disks were displaced along the circle circumference in a way that different colors and texture patterns can be scattered. A sample image of combined color and texture test is shown in Figure 3.6.

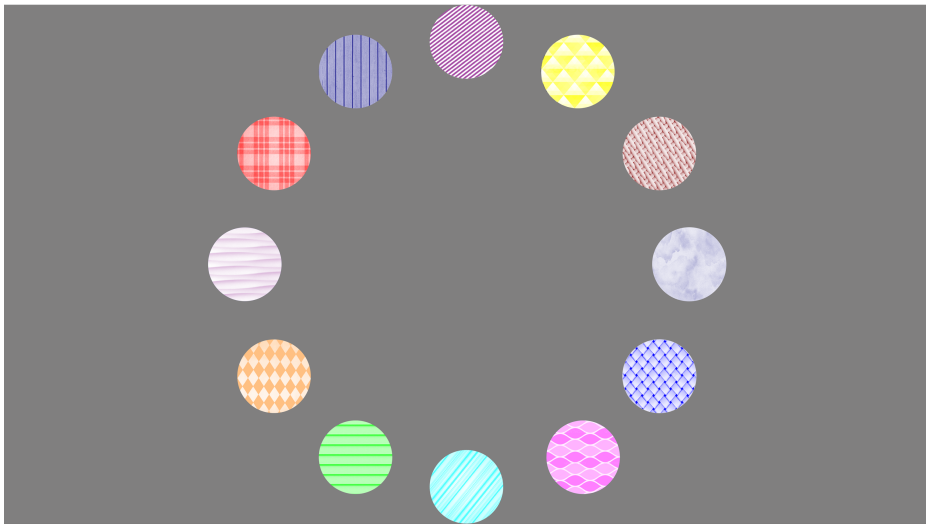


Figure 3.6: A sample image of color and texture test.

According to the ITU-R BT.2022 document [123], each image was presented to each participant for 10 seconds and we embedded a plain gray image between two consecutive images for 3 seconds.

3.2.6 Color and Motion Test

In this test, we rendered 24 video sequences with all moving objects/disks. In these videos, 12 of them incorporated colors and motion directions but the same speed and 12 others were a combination of different colors with different motion directions and speeds/accelerations divided into two groups: 1) colored disks with different motion directions and speeds/accelerations, 2) colored disks with the same motion directions but different speeds/accelerations. Moreover, eight videos were generated to test the movement of only one object among static objects. The time duration of the generated videos were 5-10 seconds. Color and motion attributes

were combined to generate six different types of moving disks in a similar manner of color and texture combination including 1) salient colors and salient directions, 2) salient colors with non-salient directions, 3) salient directions with non-salient colors, 4) salient colors with medium-level salient directions, 5) salient directions with medium-level salient colors, and 6) medium-level salient colors with medium-level salient directions. Figure 3.7 shows a sample frame of rendered video for this test.

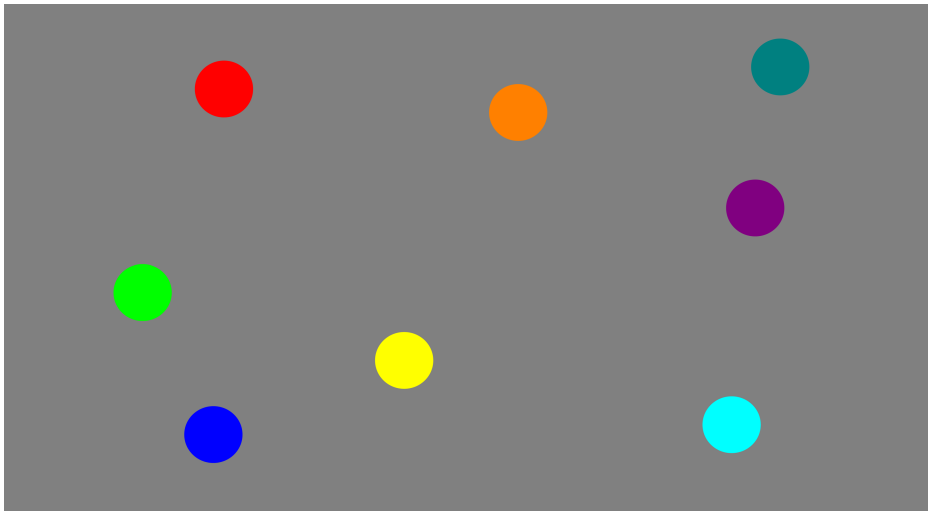


Figure 3.7: A sample created frame for color and motion test.

3.2.7 Texture and Motion Test

We rendered 20 video sequences with all moving disks for this test. Among these videos, 10 sequences combined different textures and motion directions but the same speed. The other 10 videos are divided into two groups: 1) textured disks with different movement directions and speeds/accelerations, and 2) textured disks with the same movement directions but different speeds/accelerations. Also, 8 videos were rendered to consider the circumstances that there is only one moving object in a scene among static objects.

Six different types of moving disks were created by a combination of texture and motion in a similar manner that was explained in sections 3.2.5 and 3.2.6. A sample frame of the rendered dataset is presented in Figure 3.8.

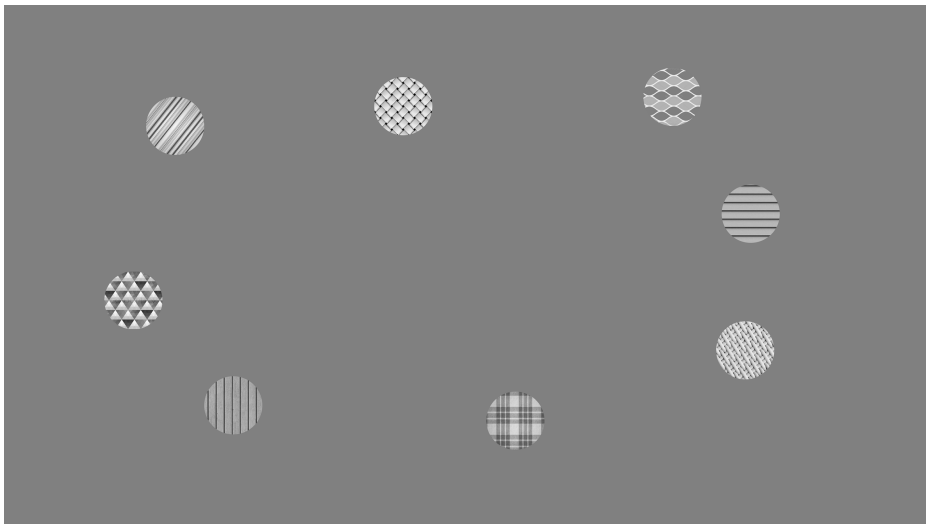


Figure 3.8: A sample frame of texture and motion test.

3.2.8 Color, Texture, and Motion Test

To test all selected bottom-up stimuli together, we rendered six video sequences in a way that all colored and textured disks are moving at the same speed but different directions. Color, texture, and motion features were combined to create six different types of disks similar to the previous tests including 1) salient colors, textures, and directions at the same time 2) salient colors and textures with non-salient directions, 3) salient directions with non-salient colors and textures, 4) salient colors with non-salient directions and textures, 5) salient directions with medium-level salient colors and textures, and 6) non-salient colors, textures, and directions. Figure 3.9 presents a sample rendered frame of this dataset.

3.3 Results and Data Analysis

In this section, we consider the results of each part of the test individually. The scan-path, fixation points, and saccade paths of all participants were analyzed during the time to extract the areas gazed at more as the salient parts. We implemented a mechanism to map the fixation points of each participant to the corresponding locations on the video sequence. Then, we created a matrix for all participants indicating their gaze points by the sequential order. Each row of this matrix belongs

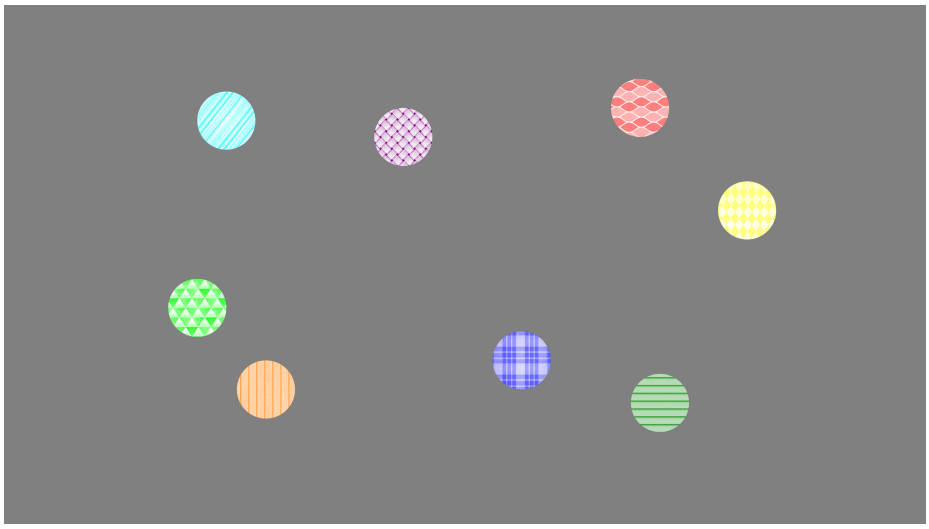


Figure 3.9: A sample frame to test all attributes.

to one participant and the columns of the matrix represents the sequential ordered locations of the gaze points by mentioning their corresponding feature. Besides, we watched the recorded eye-movements several times to match the matrix and make sure to eliminate any potential invalid mapping points. It should be noted that at the analysis stage, we used the following computational equations to obtain the number of fixations A for N participants. If we assume f_i implies i^{th} feature such as color, texture, and motion at the j^{th} gazing order, then vector F can be stated as:

$$F = [f_1, f_2, f_3, ..., f_P]^T \in \{C, T, M\}. \tag{3.2}$$

where $C = [c_1, c_2, c_3, ..., c_Q]^T$, $T = [t_1, t_2, t_3, ..., t_H]^T$, and $M = [m_1, m_2, m_3, ..., m_O]^T$. In our study 12 colors, 12 texture patterns, and 8 motion directions were used. After analyzing participants' scan-path, fixation, and saccade points, we determined a matrix of features for all participants. In this matrix, S_j shows the summation of fixations/gaze G_{i_j} on a feature in j^{th} column and each column indicates the order of attention. Therefore, the total number of fixations for each feature A can be defined as Equation 3.4.

$$S_j = \sum_{i=1}^{N} G_{i_j}. \tag{3.3}$$

$$A = \sum_{j=1}^{L} e^{(-\sigma.j)} \cdot S_j$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{L} e^{(-\sigma.j)} \cdot G_{i_j}. \quad (3.4)$$

where N and L represent the number of participants and the number of order of the attention during the time respectively. We assigned weights for each order based on an exponential function to emphasize the importance of the order of fixation. We selected $\sigma = 0.06$ in our experiment to have a smoothly descending order for the weights.

3.3.1 Color Test

Based on the fixation maps of all 26 participants, the graph of Figure 3.10 was extracted. This graph indicates that red, yellow, dark red, pink, and cyan are the most salient colors. Paying attention to the ranking table of the colors made us conclude that warm colors such as red, dark red, and pink and bright colors such as yellow and cyan are usually more salient for the HVS. Dark blue, dark green, purple, and turquoise are the least salient colors because they were fixated on less and mostly at the end of the watching time duration for each image.

We investigated the influence of luminance in color saliency as well. Therefore, we selected a few colors with different luminance because it can be considered as one of the features in designing VAMs. According to our results, the colors with lower luminance such as dark red, dark blue, and dark green usually attract HVS less.

The main differences between our designed color experiment and Gelasca et al. work in [12] can be summarized as follows: 1) They did not use the eye-tracker device; they simply asked subjects to choose the most interesting and salient colors instead. 2) At the analysis stage, they summed up overall hits for each color while we considered the order of gaze by weighting it rather than the total number of fixations per color. 3) The selected colors are different. 4) They used numbers over the disks which we believe it may cause bias in choosing disks. 5) They used an ellipse circumference to locate the colored disks while we used a circle by influencing

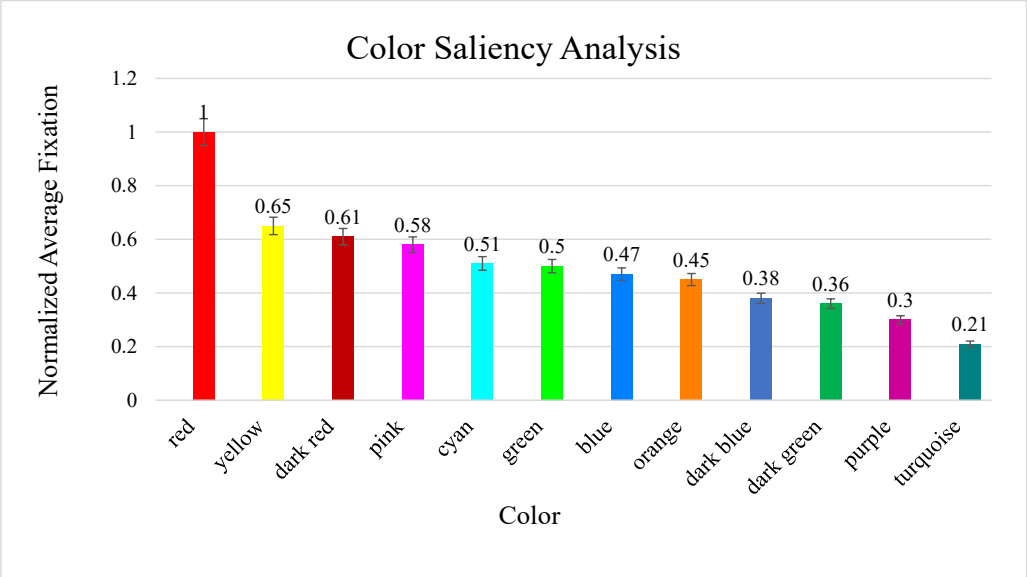


Figure 3.10: Color saliency graph. It shows the order of importance among 12 selected colors.

of center bias phenomena.

3.3.2 Texture Test

We provided a look-up table (Table 3.2) to show the labels of each texture used in our test.

Figure 3.11 shows the ranking of saliency for 12 chosen texture patterns. According to this graph, DM8, SQ3, spl2, and DM2 absorbed more attention across all 26 participants (12 females and 14 males). Therefore, more complex textures which contain dense edges such as DM8, SQ3, and spl2 and patterns with higher contrast such as DM8 will become more interesting and outstanding for human vision. We concluded that areas including compact edges are more important than the areas with high distinction intensity inside textures. However, high contrast parts with smaller areas of the intensities which means an entirely homogeneous and repeated pattern such as D7, cannot stand out as a salient pattern.

On the other hand, Spl12, V1, H2, and D7 are the least salient patterns in our experiment which are more simple and ordinary patterns. Based on our observations and pieces of evidence, we concluded that simple patterns with lower edges and lower

Table 3.2: Texture Look-up Table.

Texture Label	Sample Pattern	Sample Pattern	Texture Label
SQ3		D4	
DM2		spl12	
DM1		spl14	
spl2		D7	
DM8		H2	
D3		V1	

intensity variation are not attractive to the human attention system while dense and compact edges are distinctive.

3.3.3 Motion Test

We tested four main motion directions including horizontal, vertical, diagonal of 45 and 135 degrees by rendering 2D videos with Adobe Illustrator and Adobe After Effects. We rendered 26 videos with time durations of 3-7 seconds each by incorporating both movement direction and movement speed. Generated videos for investigating different directions can be divided into four different categories: 1) movements in one main direction with two different sides, 2) movements in two main directions which contain both sides and result in four directions in total, 3) movements in three main

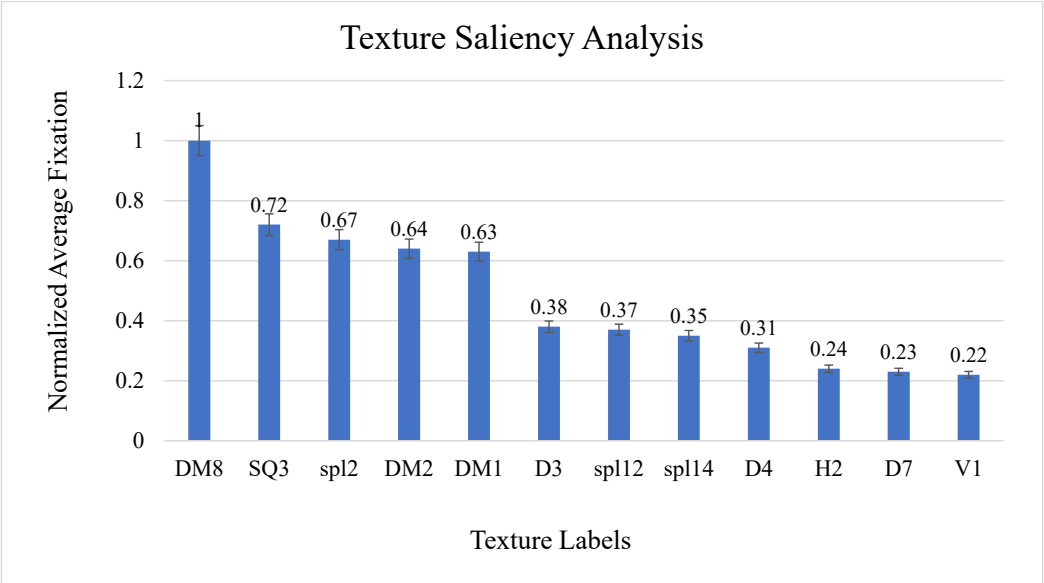


Figure 3.11: Texture saliency graph. It shows the order of importance among 12 selected texture patterns.

directions which result in six different directions considering both sides, and 4) movements in all directions.

Four main directions on both sides will give eight different directions. To this end, four videos were rendered using Adobe After Effects. In this group, each main direction is compared in two statuses like upward and downward to consider which orientation of these directions are more significant for human subjects.

According to our results among 26 participants, horizontal movement toward the right side is more attractive than the left side. Also, vertical movement downward seems to be more salient than upward. Diagonal movements with 45 degrees downward is more salient compared to upward orientation. The contrary circumstance happened for 135 degrees, i.e., upward orientation became more salient.

Table 3.3 shows the results of the tests for the second category of movement direction. We rendered six video sequences to compare pairs of the main movements on both sides including vertical and horizontal, vertical and diagonal of 45 degrees, vertical and diagonal of 135 degrees, horizontal and diagonal of 45 degrees, horizontal and diagonal of 135 degrees, and finally diagonal of 45 degrees with diagonal of 135 degrees.

According to this table, vertical movements are more interesting than horizontal

ones, however, both vertical and horizontal directions stand out compared to diagonal directions of both 45 and 135 degrees. Among different diagonal orientations, the following directions absorbed more attention: upward of 45, downward of 135, upward of 135, and downward of 45 degrees respectively.

Table 3.3: Ranking for two different directions.

Compared Directions	Salient Direction
Horizontal vs. Vertical	Vertical
Horizontal vs. Diagonal 45	Horizontal
Horizontal vs. Diagonal 135	Horizontal
Vertical vs. Diagonal 45	Vertical
Vertical vs. Diagonal 135	Vertical
Diagonal 45 vs. 135	Almost same

We rendered four videos to compare between three main movement directions. The content of these videos includes the following combination of the directions: (horizontal, diagonal of 45 and 135 degrees), (vertical, diagonal of 45 and 135 degrees), (horizontal, vertical, diagonal of 45 degrees), and (horizontal, vertical, diagonal of 135 degrees). Figures 3.12- 3.15 represent the saliency priority graphs for the mentioned four videos analyzed across all participants.

Finally, one video was rendered to compare all four main directions (i.e. 8 directions considering both sides) together. The graph in Figure 3.16 shows the results for the ranking of all directions. Based on this chart, the vertical downward orientation is the most prominent direction followed by the horizontal moving the right side and the vertical upward compared to the tested moving directions. Horizontal toward the left side and diagonal of 135 downward are the least salient items.

The rest of this section assesses the saliency of movement direction for the rendered video which included both changes in velocity and acceleration. For this purpose, we rendered 12 different videos. These videos are divided into two different groups: 1) objects with the same movement direction, but different speed which some of the

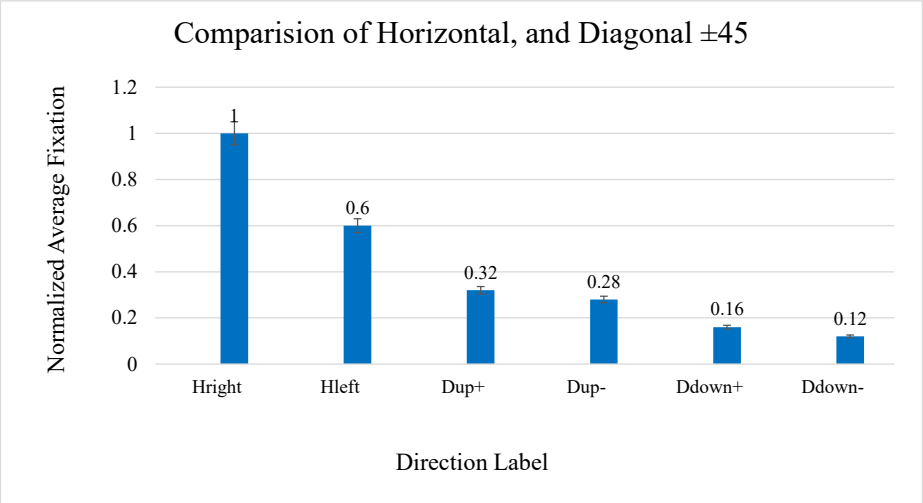


Figure 3.12: Resulting graph to compare motion directions. Directions of horizontal, diagonal of 45 degrees, and 135 degrees.

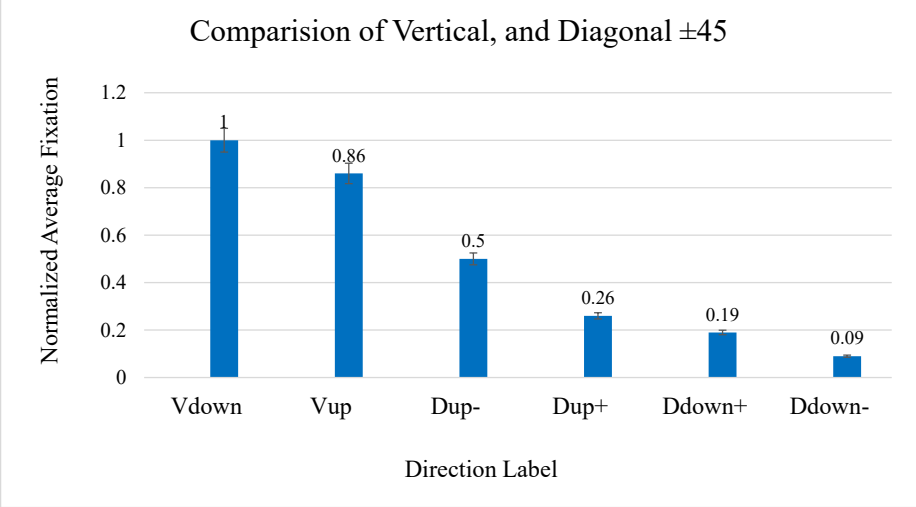


Figure 3.13: Resulting graph to compare motion directions. Directions of vertical, diagonal of 45 degrees, and 135 degrees.

objects move with acceleration and 2) objects with different directions and speeds simultaneously.

According to our results, we observed that the fastest objects usually stand out relative to other objects in a scene. However, if the speed became very fast, subjects could not follow and track those objects because they did not have enough time to

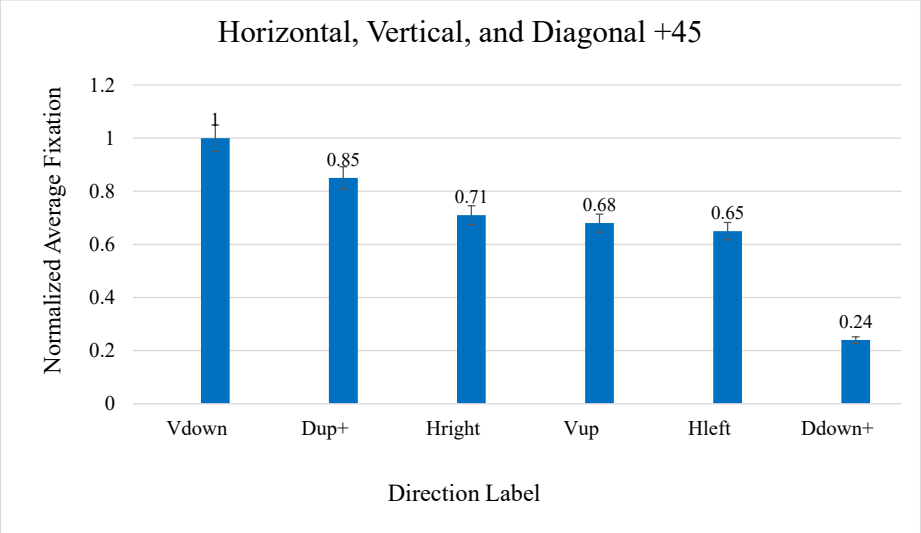


Figure 3.14: Resulting graph to compare motion directions. Directions of horizontal, vertical, and diagonal of 45 degrees.

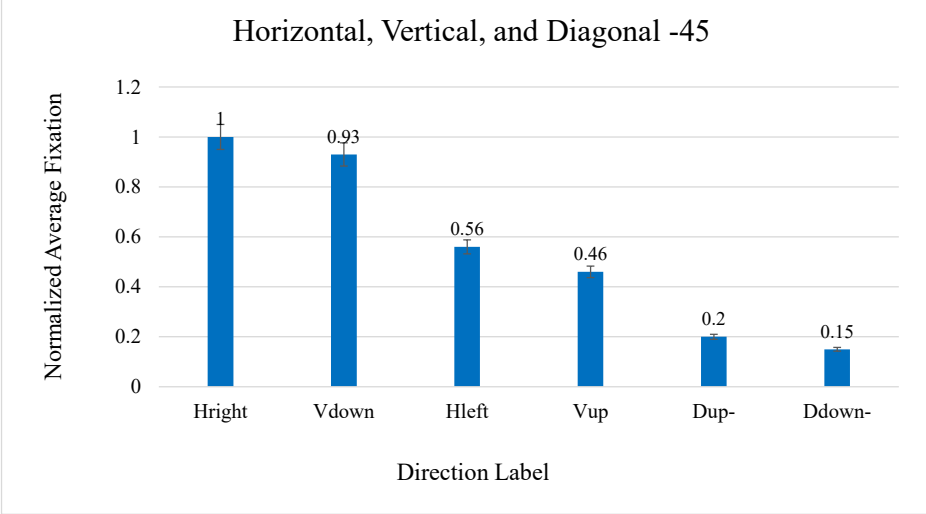


Figure 3.15: Resulting graph to compare motion directions. Directions of horizontal, vertical, and diagonal of 135 degree.

focus on them. Consequently, they ignore objects which move too fast. On the other hand, the slowest objects can be attractive to the HVS too because subjects have more time to see and track them within a scene and those objects are displayed for a longer time. Furthermore, any rare movement which includes abrupt acceleration

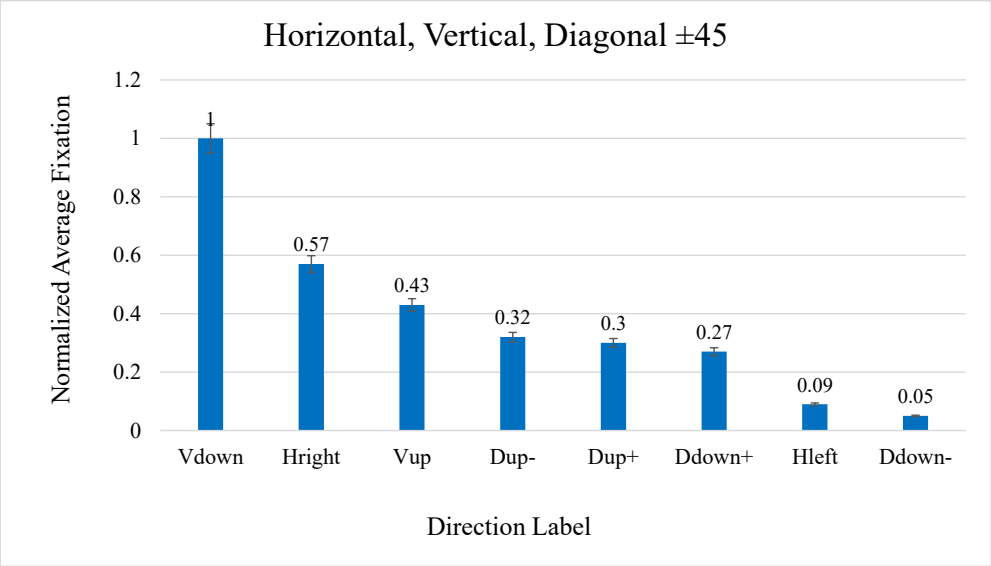


Figure 3.16: A graph to compare motion in all selected directions.

absorbs more attention than other movement types.

It should be remarked that movement speed and acceleration are more significant than the movement direction. In all videos, participants fixated on the very fast, very slow, and any rare moving objects regardless of their movement direction. Based on the eye tracking results, we can conclude that motion speed outweighs motion direction in saliency detection.

For example, Figure 3.17 is the result of moving in only the horizontal direction with different velocities and accelerations.
The graph resulted from the situation that four objects were moving in the horizontal directions, but different velocities and not having acceleration.

In the motion test, the objects which are moving vertically across the display may be selected as the major salient items because their pathway approaches the center of the display (i.e. due to the center-bias phenomena).
In the scenario in which moving disks have different velocities/accelerations, objects become salient candidates when they have different motion compared to others, i.e., moving faster or slower than other objects or having abrupt acceleration, especially when changing their direction. Additionally, objects moving in different directions from the majority of other objects will receive more attention.

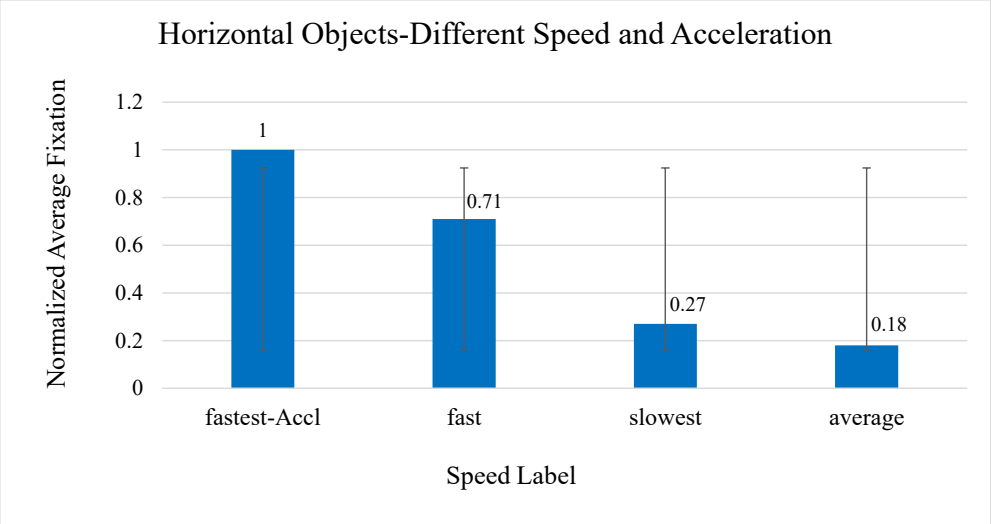


Figure 3.17: Graph shown a comparison of motion speed for four objects in the horizontal, direction with different speeds and accelerations.

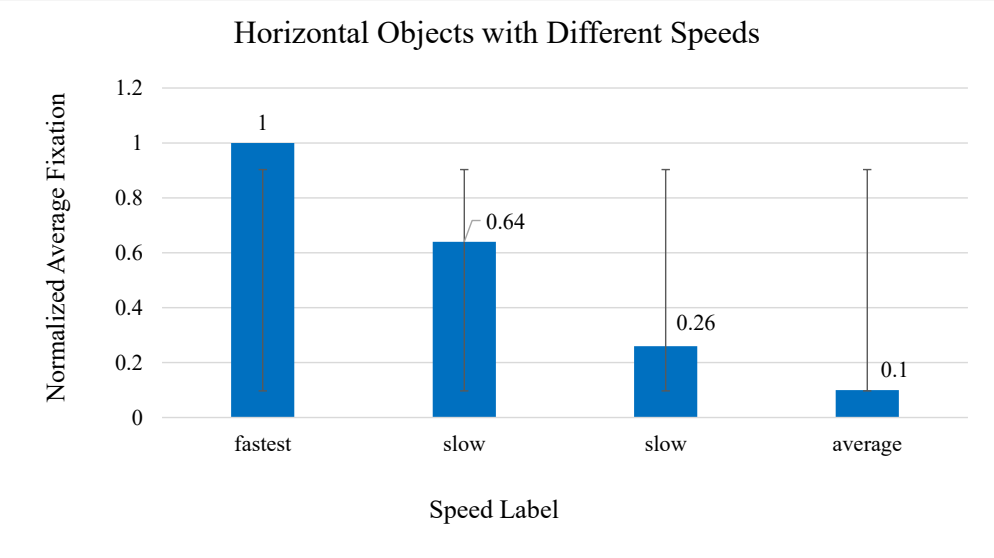


Figure 3.18: Graph shown a comparison of motion speed for four objects in the horizontal, direction with different speeds.

Note that smooth pursuit eye movements were not considered in our experiment because the eye-tracker device was only recording and producing the fixations points and saccades as output information. Additionally, smooth pursuit eye movements involve cognitive bias [99].

There are two types of eye movements for voluntarily gaze shift in HVS known as

smooth pursuit and saccades - that can be important in motion saliency. Smooth pursuit allows the eyes to strictly follow a moving object and is modified by continuous visual feedback [99].

3.3.4 Contrast Test

Our contrast test indicates that the contrast between colors within a scene plays a very significant role in absorbing human attention. The contrast between colors in the produced test images leads participants to be attracted to the colors which are more outstanding from the contrast perspective rather than being salient in color by itself. The consistency among participants' fixation areas in the recorded eye movements data proves this. We computed correlation coefficients for each participant in comparison with all other human subjects and results indicated a high consistency among participants as the fixated salient objects.

Figure 3.19 shows a graph of inter-subject consistency indicating correlation coefficients among 26 participants.

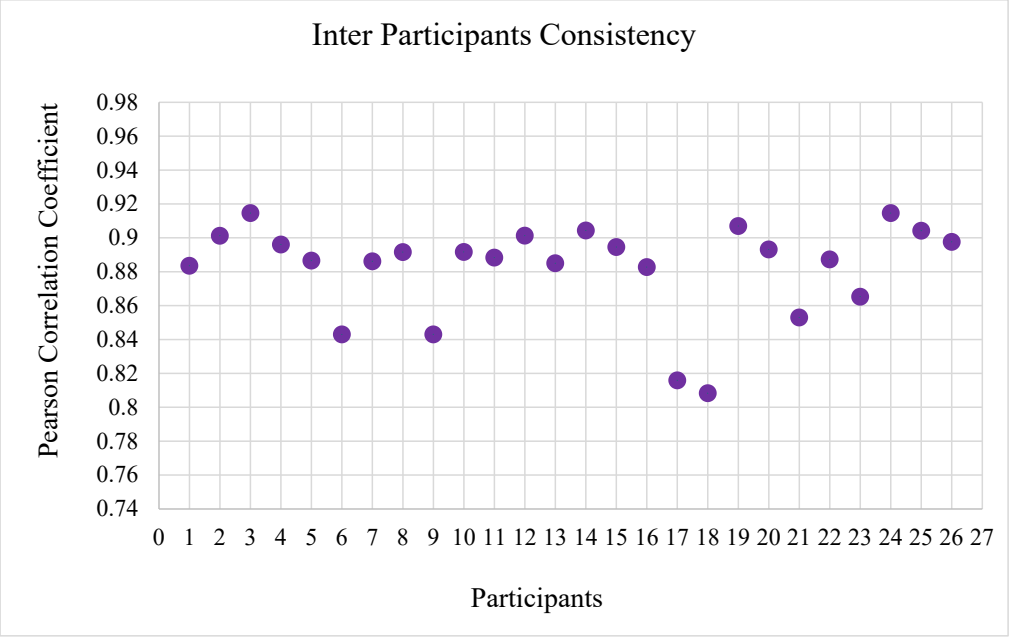


Figure 3.19: Inter participant consistency for color contrast test.

According to the Figure 3.19, the contrast between colors produces a more consistent response in HVS and is not a very subjective reaction. This confirms the previous studies about the contrast of colors. Therefore, we can use this fact in designing a saliency detection algorithm.

3.3.5 Color and Texture Test

To analyze the results of the combined tests, we obtained the ranking charts as described at the beginning of section 3.3 for each created image/video individually. Next, we computed the effect of each feature in the combined data using a weighted summation of the feature. Assume W_f shows the resulting overall weight for each feature within each ranking chart for a created image/video. The component f_i indicates the normalized average number of fixations on each disk which is the weight on the top of each bar in the chart e.g. Figure 3.20. For this graph f_i can take one of the following weight values: 1, 0.88, 0.74, 0.69, 0.67, and so on.

In addition, w_i denotes the weights resulting from each individual feature tests for each specific tested color, texture, or motion. For example, in the combined color and texture test, as shown in Figure 3.20, w_i can be shown as $w_i = \{c_i, t_i\}$ where c_i and t_i represent the resulting importance priority from each individual color test (i.e., section 3.3.1) and texture test (i.e., section 3.3.2) respectively. To make this more clear, assume the first component of the chart i.e., "yellow-DM8", then w_i for "yellow" color and "DM8" texture can be found from Figure 3.10 as 0.65 and Figure 3.11 as 1 respectively.
Therefore, the overall weight for each feature in an image can be computed by Equation 3.5.

$$W_f = \frac{1}{R}\sum_{i=1}^{R} w_i \cdot (f_i) \tag{3.5}$$

where, R represents the number of disks in each image or video (i.e., the number of bars in each chart). For the color and texture test, we computed W_C and W_T for each image. Obviously, it should be assumed that $w_i = c_i$ in the computation of W_C and for computing W_T the assumption is $w_i = t_i$. Finally, the average of these measures were calculated over all created images or videos for the motion involving tests.

Figure 3.20 shows the analysis of the saliency priority for one sample image.

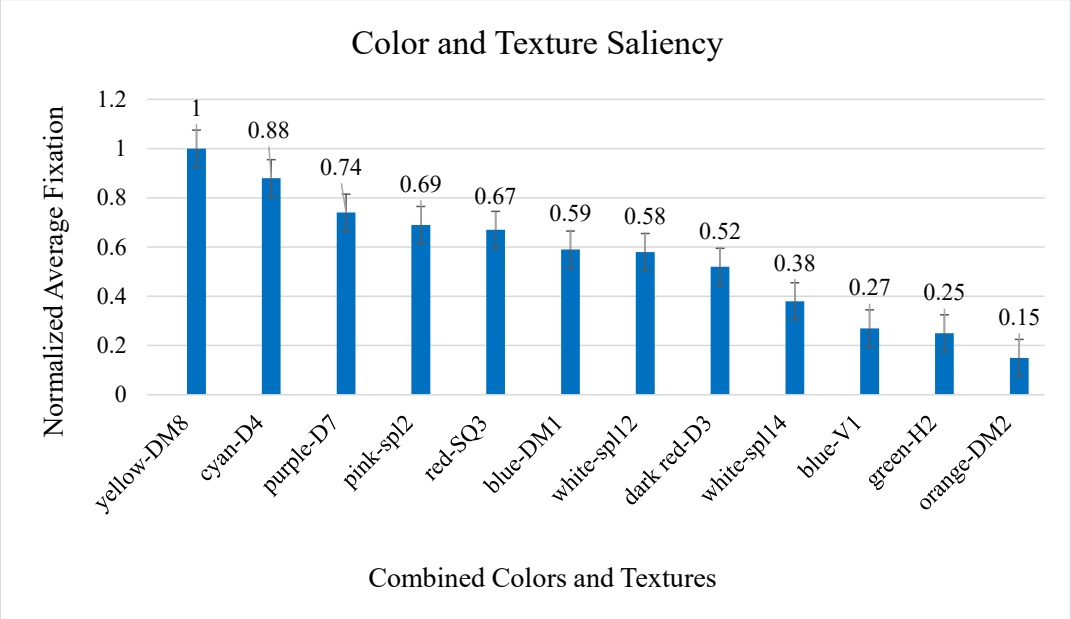


Figure 3.20: Color and texture saliency graph.

In this test, the fixation maps of all participants for each created image were analyzed and the comparative average of importance was computed for color and texture features among 10 created images. We found that color feature outweighs texture in general. The significance of color feature is approximately 60% in comparison with the texture which is 40% within the created dataset.

3.3.6 Color and Motion Test

In this test, W_C and W_M were computed using Equation 3.5 for each rendered video. The average of these measures were computed among all videos to obtain the importance weight of each feature.

Analysis of the fixation maps among all participants indicates that when there is only one moving object in a scene, the HVS detects and tracks that object as the most salient item regardless of its color or movement direction. Therefore, we can conclude that the motion stimulus outweighs color in this scenario.
On the other hand, for the scenarios in which all disks are moving with different colors

and directions but the same speed, the results indicate that the color feature can be seen and fixated in approximately 70% of the cases compared to the motion direction with a quantity of 30% on average.
Moreover, in the scenarios of all moving disks in a scene with different colors and speeds, we found that color is again more salient than speed/acceleration with the weights of approximately 61% and 39% respectively.

This is an interesting finding because according to the literature described in Chapter 2 [35, 36], the motion should be the most important feature in terms of visual attention and saliency detection perspective. However, it seems that for a dynamic scene other features such as color can be as significant as motion and perhaps sometimes more salient.
In Figure 3.21, a resulting analysis of a sample rendered video was presented where all disks move horizontally with different colors and speed/acceleration.

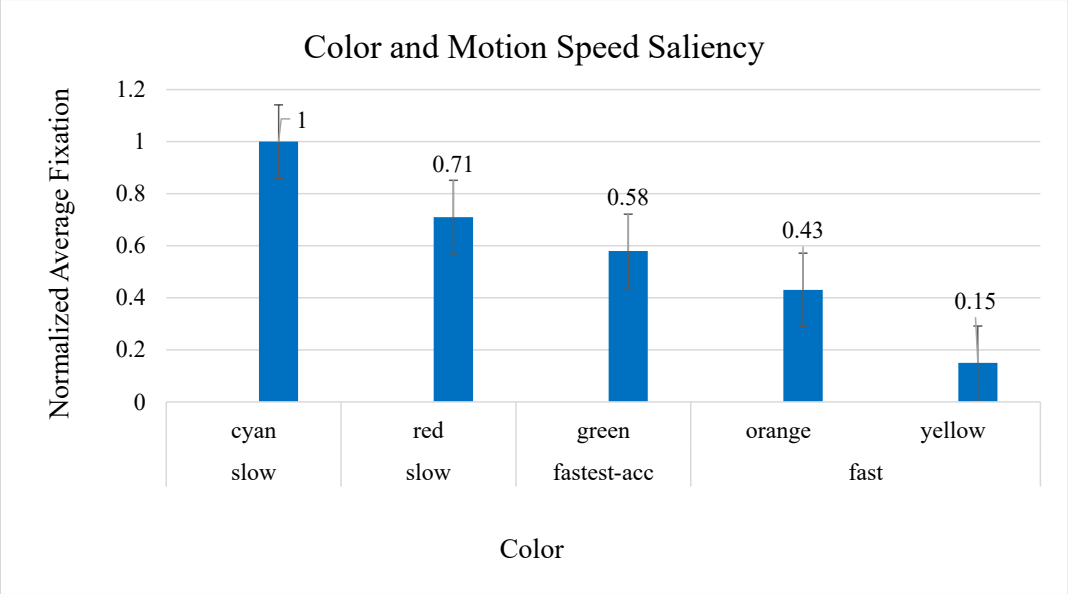


Figure 3.21: Color and motion speed saliency graph.

3.3.7 Texture and Motion Test

Analysis of the fixation maps resulting from the eye-tracking procedure leads us to conclude that having only one moving disk in a scene results HVS to detect it as the most salient object. In this case, texture or movement direction of the disk does not

make any change which demonstrates that the motion stimulus outweighs texture within similar scenarios.

According to our results, in the scenarios where all disks move with different textures and directions but the same speed, texture feature absorbed participants' attention in approximately 65% of the tested cases compared to the motion direction which was 35% on average. Furthermore, the results of the tested scenarios with all moving disks possessing different textures and speeds illustrated that texture attribute is more salient than movement speed/acceleration with the quantitative weights of 68% and 32% respectively.

In Figure 3.22, we presented a resulting saliency ranking analysis of a sample rendered video for texture and motion test.

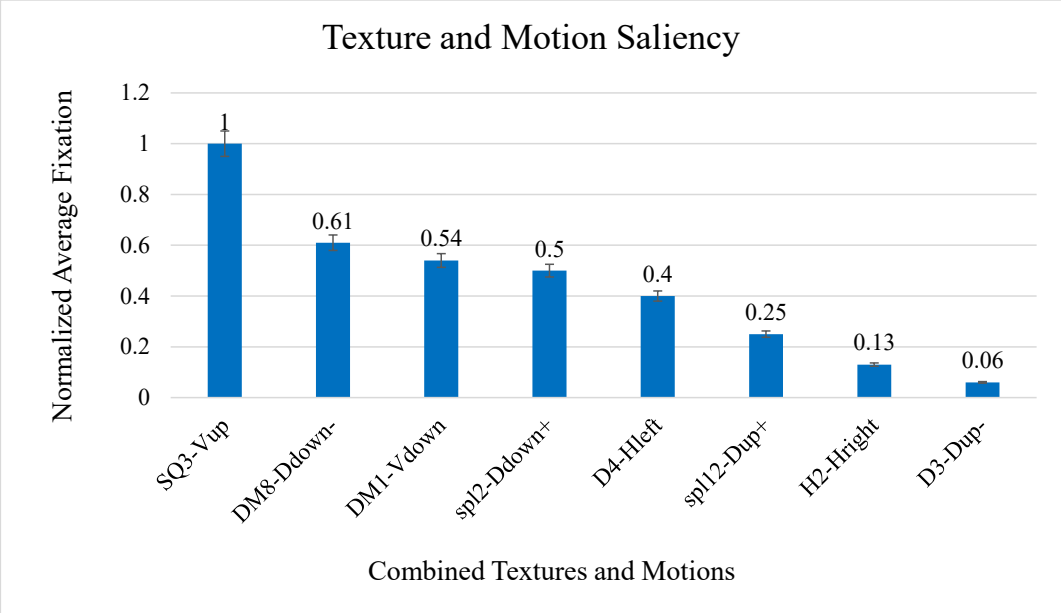


Figure 3.22: Texture and motion saliency graph.

3.3.8 Color, Texture, and Motion Test

In this test, W_C, W_T, and W_M were computed in a similar manner as the previously described combined tests for each rendered video.

After analyzing the saliency priority among all videos one-by-one and eventually averaging the results, we found that the ranking of saliency among bottom-up features

for a dynamic scene was color ($\sim$ 40%), texture ($\sim$ 31%), and motion ($\sim$ 29%) respectively.

We tested all moving objects in order to understand the importance of motion stimulus in dynamic scenes. We observed that color and texture can stand out as significant as motion in these scenarios. This leads us to consider all bottom-up attributes carefully in designing visual attention models with the aim of saliency detection. A resulting saliency ranking analysis of a sample rendered video is shown in Figure 3.23.

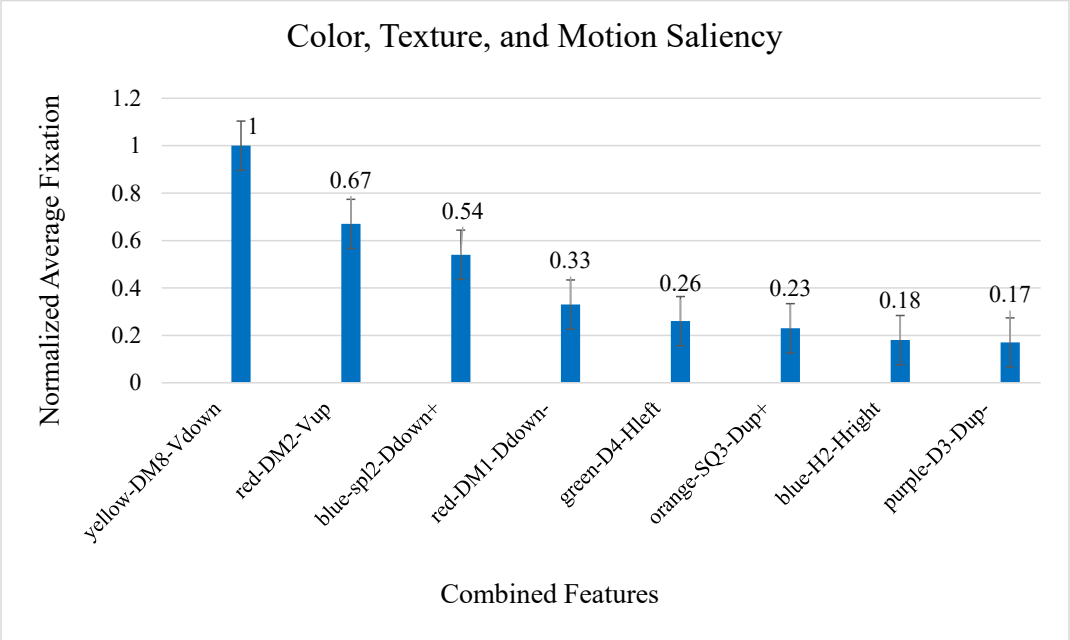


Figure 3.23: Color, texture and motion saliency graph.

3.4 Conclusion

In this chapter, we explained our experimental study which avoided cognitive bias to model the salient areas within a scene. For this purpose, we investigated bottom-up features including color, texture, motion direction, motion speed, and color contrast as stimuli for the HVS. This experiment was designed to understand the order of importance (i.e., saliency priority) among bottom-up features in absorbing human attention while observing a scene.

According to our results, warm colors such as red and pink, and bright colors such as yellow and cyan are more salient. Textures with dense and compact edges are more outstanding. In addition, textures with obvious high contrast within their pattern are likely to be salient. Vertical movements are more likely to be fixated on and the motions with high (not very high) or very low speed or any unique acceleration are more salient for human subjects.

We concluded that color contrast is as important as color and texture stimuli. Our test for contrast illustrated high consistency among participants while saliency within the individual color and texture tests can be a subjective decision.
We found that texture is an important feature in determining the salient area in a scene unlike what has been mentioned in the literature. Especially, in the absence of cognitive bias texture might be much more important to guide human attention. Color and texture are very important features in dynamic scenes compared to motion in spite of previous studies that claimed motion should be the most significant feature. Motion is very significant feature if we have few moving objects in a scene. Our results also confirmed the center bias factor since the majority of the participants' fixation points are concentrated at the center of the display.

According to our results, we observed that the fastest objects usually stand out more. However, if the speed exceeded a threshold and became too fast, subjects could not follow and track those objects. On the other hand, the slowest objects can be attractive to the HVS because subjects have more time to see and track them. In addition, any unexpected movements which include abrupt acceleration absorb more attention. If the speed of an object changes from slow to fast, the object will be more salient compared to the opposite change in speed.

It should be identified that smooth pursuit type of human eye movements is outside of our scope in this dissertation, but it might be necessary to be addressed for motion saliency in future works.

Chapter 4

Static Saliency Detection Model

4.1 Introduction

In this chapter, we described our introduced algorithm to extract saliency for 2D images. The introduced method to obtain the saliency map is based on a Bayesian framework. We used this rigorous statistical formulation to compute a robust saliency measure relied on feature level information fusion. We selected a probabilistic technique because human visual attention is not deterministic and people may attend to different locations on the same visual input at the same time.

This saliency detection framework was established based on our experimental findings. The model used color, luminance, intensity, and texture features within 2D images. To extract color saliency, we used a feature contrast in *CIE Lab* color space as well as a k-nearest neighbor search based on k-d tree search technique [127] to assign a ranking system into different colors. To find the salient textured regions, we employed contrast-based Gabor energy features and then we added a new feature as intensity variance map to support the results of our empirical work. Finally, we combined different feature maps and classified different saliency using a Naive Bayesian Network.

We compared our model to four related and well-known state-of-the-art methods using publicly available ground-truth data and their source code. The four methods are as follows: Context-Aware saliency method (CA) [112], Segmenting salient objects (SEG) [71], Maximum Symmetric Surrounding method (MSS) [63], and Learning Discriminative Sub-spaces on random contrast (LDS) [98].

We briefly explain these methods here. Goferman et al. [112] designed a context-aware saliency algorithm by computing the contrast of color and intensity features

in image blocks which is a combination of graph-based and cognitive models. They considered basic principles of HVS, supported by psychological evidence to design their algorithm. They utilized both local and global factors to be able to provide better performance. Color and contrast were used as local low-level attributes. Then, frequently occurring features were suppressed by global considerations of local extracted features. The context-aware term refers to the using of high-level factors, such as human faces [112]. Their algorithm has a promising performance.

Rahtu et al. [71] introduced a salient object segmentation method based on combining a saliency measure with a conditional random field (CRF) model. This model employed a statistical Bayesian framework and local contrast of color, and motion features. To recover well-defined salient objects, a CRF model was applied to the resulting saliency map using an energy minimization-based approach. This approach is very time-consuming.

Achanta et al. [63] used the low-level features of color and luminance and applied a global symmetric center-surround mechanism to detect saliency map. Their saliency value was computed using Euclidean distance between the average *CIE LAB* vector of all pixels and each pixel of a Gaussian-blurred version of the same input image [63]. To obtain a symmetric surrounding for each pixel, they adjusted the bandwidth of the center-surround filter.

A model proposed by Fang et al. [98] that was based on learning a set of discriminative subspaces to extract outstanding targets and suppress distractors. They used Principal Component Analysis (PCA) on randomly selected image blocks. The candidate subspaces were constructed using selected principal components since they have impressive abilities to separate targets and distractors [98]. By projecting images onto subspaces, each image block was determined by its contrasts against randomly selected neighboring. Then, salient blocks were extracted through applying an optimization framework to learn the saliency model from subspaces that can separate salient targets and distractors [98].

4.2 Proposed Static Model

The inspiration of the introduced model in this work has derived from the Bayesian framework and the bottom-up attention model resulted in our human subjective experiment. The simplicity and speed of this approach make it an appropriate candidate

for a real-time system. The proposed framework includes color, luminance, intensity, and texture features as bottom-up features. We extracted two different color maps from color stimuli: 1) a color contrast map using chroma and luminance in *CIE Lab* color space, and 2) a color importance ranking map using the results of our experiment from chapter 3. Moreover, two texture maps were extracted: 1) a texture map using the contrast of Gabor energy features at different frequencies and orientations according to the experiment results. 2) a variance of intensity feature was introduced to extract the contrast of intensity to support the results of our experiment.

Next, the resulting feature maps were merged through a Bayesian framework. Additionally, we exploited the human visual acuity factor to enhance the final saliency map based on the characteristics of the HVS.

Figure 4.1 shows a block diagram of our model to help a better understanding of different steps of that. Each step of the framework was explained in detail in the following subsections.

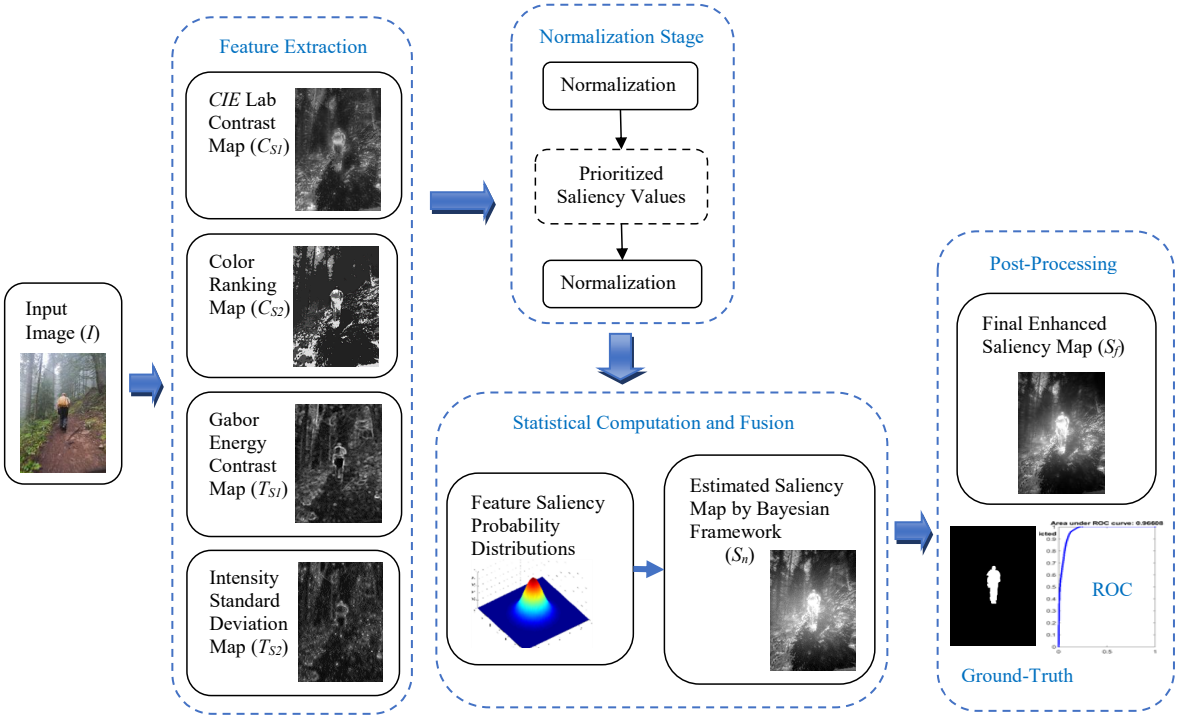


Figure 4.1: Block diagram of our framework for salient region detection.

4.2.1 Color Saliency Map

To extract a color map based on our experimental findings, a saliency probability was assigned to each pixel using the resulting order of color saliency. The saliency of various colors which are not included in the experiment was estimated based on the closest numerical distance of the RGB color value of that pixel and the proportion of fixation points resulting from the eye-tracking stage. For this purpose, we employed a K-Nearest Neighbor (KNN) search using a k-d-tree search method to find the most similar color ranges to our selected colors in the experiment. The KNN-search method finds the nearest color of the pixels within an input image for each given color value and classifies the closest colors to the corresponding colors of the look-up table for our 12 selected colors in Chapter 3. Hence, the ranking weights i.e., $W^T = [w_1, w_2, ..., w_{12}]$ resulting from the experiment was used to train the KNN-search method. A three-dimensional tree was used to partition color space in an image because it is a very useful implementation to classify and organize different color points as well as color ranges in a three-dimensional color space. Here, we explained the procedure of obtaining the color map which is C_{S1}. Assume the saliency probability of each selected color of our empirical test is shown as Equation 4.1.

$$P_S = \frac{W^T}{\sum_{r=1}^{12} w_r} \tag{4.1}$$

Then, the KNN-search method finds the nearest color of the pixels within an input image for each given color value and classifies the closest colors to the corresponding colors of the look-up table for our 12 selected colors according to Equation 4.2 [127]:

$$A \triangleq argmax_{C_j}(\sum_{d_j \in KNN^{dt}} sim(p_t, p_i) I_n(p_i, p_j) \quad) \tag{4.2}$$

where, KNN^{dt} has the set of k-nearest neighbors of test colors, $sim(p_t, p_i)$ is the similarity between p_t and p_i which are sample pixels of the training and testing data respectively. $I_n(p_i, p_j)$ is the indicator function that can be 1 or 0 depending on whether the d_i belongs to class [127]. The similarity is determined using Euclidean distance and a 3-d tree search technique.

Next, for each pixel p_i the saliency probability $C_{S1}(p_i)$ is obtained by Equation 4.3 which is the product of the $P_S(A(p_i))$ indicating the saliency probability of the corresponding color among 12 classes c_j for pixel p_i, which was weighted by the

Euclidean distance between the color of p_i and the color holding class c_j.

$$C_{S1}(p_i) = P_S(A(p_i)) \times Euclidean(RGB_{c_j} - RGB_{p_i}) \tag{4.3}$$

4.2.2 Contrast Saliency Map

We also extracted another map in *CIE Lab* color space using the center-surround contrast of L, a^*, and b^* features. Then, we computed the Euclidean distance of the resulting contrast maps to amplify the luminance effect and generate a chroma-luminance saliency map. This color space has been utilized because of its uniform chromaticity properties.

We formulated the resulting color-luminance map C_{S2} in Equation 4.4.

$$C_{S2} = \sqrt{(C_a*)^2 + (C_b*)^2 + (C_L)^2} \tag{4.4}$$

where, C_a*, C_b*, and C_L indicate the contrast of a^*, b^*, and L channels respectively. Since the color contrast is defined as the luminance difference between colors, hence, *CIE LAB* color space is becoming the best candidate to show the contrast changes. In other words, the most contrast information exists in the L channel as its variation is more than a^* and b^* channels in the Lab cylinder, consequently, we decided to use Lab color space. The Euclidean distance was used to emphasize this information among C_a*, C_b*, and C_L maps.

We computed the contrast of *Lab* features using a center-surround difference mechanism weighted by a "fovea mask", introduced by Banitalebi-Dehkordi et al. [128]. The center-surround difference is computed based on a sliding window technique indicating the absolute value subtraction between the average of the pixel values within the inner window and a collar window.
The fovea mask is a circular mask based on fovea photoreceptor concentration. This mask models the spatial distance by assigning different weights to different pixels within the mask based on their distance from the central pixel. Applying fovea mask makes our model more compatible with HVS. We also tested the common approach of center-surround difference based on Gaussian kernels (i.e., Gaussian mask) by Itti et al. [55] and found that the fovea mask provides better performance with respect to the accuracy in detecting the most salient regions.

To understand the definition of the fovea mask, suppose α is half of the angle of

the viewer's eye at the highest visual acuity. The range of 2α is between $0.5°$ and $2°$ [130]. The sharpness of vision declines very fast beyond this range. The mask radius is defined by Equation 4.5 [130]:

$$r = Z \times tan(\alpha)_{[cm]} \quad = \frac{Z \times tan(\alpha) \times R_H}{H}_{[pixel]} \tag{4.5}$$

where, H and R_H are the vertical height and resolution of the display, and Z is the distance of the viewer to display [130]. In our implementation, we selected an angle of $\alpha = 1°$, HD resolution video at resolution 1080×1920, a viewing distance of 217.5 cm, and display height of 68 cm.

According to Jonas et al. [129] and Banitalebi-Dehkordi et al. [128], the original fovea mask is defined based on the density of the photoreceptors from 0 to 1-degree visual angle is defined as $x = [197, 161, 125, 89, 74, 65, 58]$. By extrapolating the $1D$ vector to $2D$ and normalizing it into a $[0, 1]$ interval it can be resized to different fovea radius as it is shown in Figure 4.2 for radii of 21.

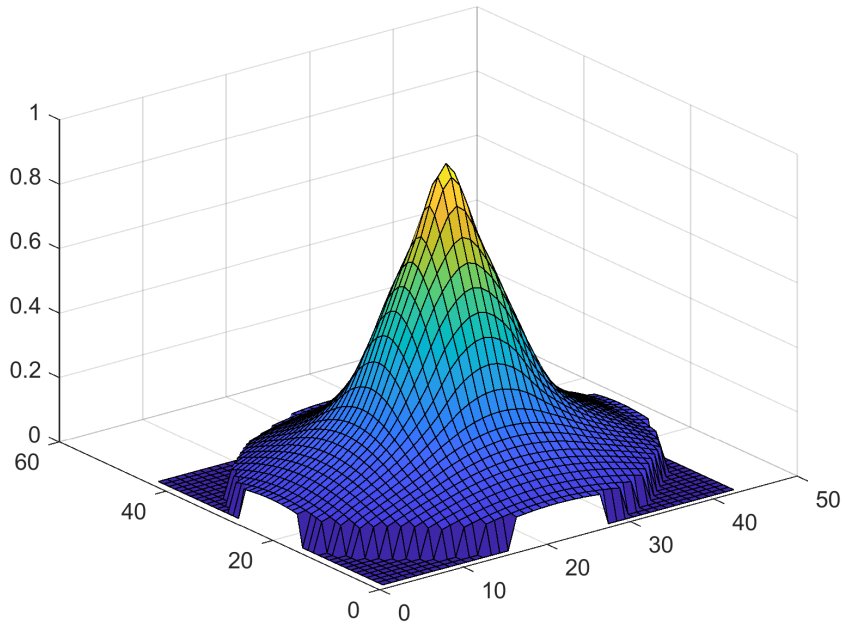


Figure 4.2: A sample Fovea mask computed and depicted for fovea radii of 21 as an example.

Accordingly, the contrast of each feature can be formulated as Equation 4.6.

$$C_f = \frac{1}{n_k}(\sum_i \sum_j W_{fovea}(i', j') \times |f_C(i, j) - f_C(i', j')|) \tag{4.6}$$

where, C_f is the feature contrast assuming $C_f = [C_a*, C_b*, C_L]$ and $f_C = [a^*, b^*, L]$; n_k is the number of pixels within a neighborhood around each pixel (i.e., collar window), $W_{fovea}(j)$ shows the weight of the fovea function for a pixel of the surrounding area with the coordination of (i', j'). Also, (i, j) and (i', j') indicate the center pixel location and a surrounding pixel location, respectively.

4.2.3 Texture Saliency Map

Salient textures within a scene stand out due to their edge orientation and intensity contrast from their surrounding regions. We calculated the feature contrast between each image pixel and its surrounding neighborhood as a texture saliency map. We used Gabor energy features extracted from the Gabor filter bank as the texture features. Because they are capable to efficiently extract different orientations and spatial frequencies which provides a complete feature space for a texture. It has been proven that the Gabor energy operator is a more efficient discriminant component than the Gabor filter [131]. Therefore, we used its energy to construct the feature vectors for textures of the images.

We applied a dimension of 1×49 to the feature vectors for each pixel resulting from seven different preferred spatial frequencies (wavelength) and seven different equidistant orientations of [0, 30, 45, 60, 90, 120, 135] degrees.
Then, we computed the contrast of Gabor energy features using the center-surround mechanism weighted by a fovea mask similar to the color map. Therefore, we estimated the texture saliency value of each image pixel based on the weighted average difference between its Gabor features and the Gabor features of the surrounding pixels inside a fovea mask using Equation 4.7.

$$T_{S_1} = \frac{1}{n_k}(\sum_i \sum_j W_{fovea}(i', j') \times |f_T(i, j) - f_T(i', j')|) \tag{4.7}$$

where, T_{S_1} is the resulting Gabor feature contrast as a texture map, n_k is the number of pixels within a neighborhood around each pixel, $W_{fovea}(i', j')$ shows the weight of the fovea function for a pixel of surrounding area located at (i', j'), and $f_T(i, j)$ is the texture feature vector for a pixel located at the (i, j) coordinate.

We introduced a second texture feature map as the local standard deviation of the image intensity. We employed this feature map to compute the contrast of the intensity in textures. It was computed by estimating the average standard deviation within a circular neighborhood surrounding each pixel. This filtering method can be expressed as Equation 4.8.

$$T_{S_2} = sd_{map} = \frac{1}{n_S - 1} \sum_{S} (I_C - I_S)^2 \tag{4.8}$$

where sd_{map} denotes the map of intensity standard deviation, n_S is the number of pixels in the surrounding area of each pixel, I_C, and I_S represent the intensity of the center and surroundings respectively. Normalization was applied to both resulting maps.

4.2.4 Saliency Map Estimation and Enhancement

To obtain the ultimate saliency map, color and texture maps should be fused. For this purpose, we used a naive Bayesian network where the salient regions model descriptor can be built by computing the likelihood probability distributions of the resulting color and texture feature maps. The probability density function of each individual feature was modeled using a Gaussian distribution, where both the mean and variance were learned. This distribution was selected because of its simplicity and efficiency in obtaining the model parameters i.e. mean and variance.

Before learning the probability functions of the extracted maps explained in Sections 4.2.1 and 4.2.2, the feature maps are weighted and normalized based on the results of our empirical study [133, 134] indicating that the significance of color feature compared to the texture features were estimated as 60% versus 40% in our designed test and dataset. Then, both the prior and likelihood probabilities were learned from the extracted feature maps and targeted feature ranges resulting from our experiment [132].

Assume a random variable X_S denotes saliency value of a pixel in an image/frame. Given the observed color-based and texture-based features of that point C_{S1}, C_{S2}, T_{S1}, T_{S2}; we formulated the saliency detection as a Bayesian inference problem to estimate the posterior probability at each pixel of the image:

$$p(X_S|C_{S1}, C_{S2}, T_{S1}, T_{S2}) = \frac{p(X_S, C_{S1}, C_{S2}, T_{S1}, T_{S2})}{p(C_{S1}, C_{S2}, T_{S1}, T_{S2})} \tag{4.9}$$

where $p(X_S|C_{S1}, C_{S2}, T_{S1}, T_{S2})$ indicates the probability of predicting whether a pixel belongs to a predefined saliency class, $p(C_{S1}, C_{S2}, T_{S1}, T_{S2})$ is the likelihood of the observed color-based and texture-based defined features, and $p(X_S, C_{S1}, C_{S2}, T_{S1}, T_{S2})$ is the joint probability of the saliency value and observed features. To simplify Equation 4.9, a feature set F was assumed including all feature maps, where $F = \{C_{S1}, C_{S2}, T_{S1}, T_{S2}\}$. Thus, Equation 4.9 can be rewritten as Equation 4.10.

$$p(X_S|F) = \frac{p(F|X_S) \times p(X_S)}{p(F)} \tag{4.10}$$

In this type of network, the features can be assumed conditionally independent given the saliency classes, therefore, the computational burden for classification is reduced. Furthermore, the probability of the feature maps was normalized (i.e. the probabilities all added up to 1.0), the term $p(F)$ can be neglected from the equation. Therefore, the computation of the saliency probability simplifies to multiplying the likelihoods together ($n = 4$) as can be seen in Equation 4.11.

$$p(X_S|F) = p(X_S)(\prod_{i=1}^{n} p(F_i|X_{s_i})) \tag{4.11}$$

In this study, the prior $p(X_S)$ was assumed as a uniformed distribution between classes which can be defined as $1/C$, where C is the number of classes including strongly salient, average salient, weakly salient, non-salient regions. We defined these classes to be able to prioritize the saliency within entire of an image/frame. It should be mentioned that, computing the product of many probabilities can encounter numerical instability because they can be very small [65]. To solve this problem the log of the likelihood was applied [65, 67, 71] which transformed the likelihood products into likelihood summations to obtain the saliency value S_n on Equation 4.12.

$$S_n = p(X_S|F) = p(X_S)(\sum_{i=1}^{n} log(p(F_i|X_{s_i}))) \tag{4.12}$$

Furthermore, we used a characteristic of the HVS to enhance the resulting saliency map known as eccentricity sensitivity [135]. According to [136, 137], human visual acuity decreases with increased eccentricity from a fixation point.
The areas with higher saliency are seen more precise compared to the farther surrounding areas that seem slightly blurry.

We exploited this property to enhance our saliency map by using human visual sensitivity $E_{VS}(f, e)$ computed in [135] as Equation 4.13.

$$E_{VS}(f, e) = \frac{1}{C_0[exp(\tau f \times (e + e_2)/e_2)]} \tag{4.13}$$

$$e = tan^{-1}(d'/v) \tag{4.14}$$

where, f is the spatial frequency (cycles/degree); e is the retinal eccentricity (in degree); C_0 is the minimum contrast threshold; τ is the spatial frequency decay constant; e_2 is the half resolution eccentricity. In the Equation 4.14, v is the viewing distance and d' is the spatial distance between each image pixel and the nearest salient pixel. Based on the experimental results in [137], the best parameter values are tuned as follows: $\tau = 0.106$, $e_2 = 2.3$, $C_0 = 1/64$ [135].

The enhancement step emphasizes the importance of foveal vision compared to peripheral vision in our designed algorithm. Foveal vision is used for accurately inspecting detailed objects, whereas peripheral vision is used for organizing the broad spatial scen [138]. Human foveal vision is optimized for fine details of a scene, while the peripheral vision is optimized for coarser information [138].

Lastly, the final enhanced saliency map by applying the normalized visual sensitivity can be calculated as [135]:

$$\mathbf{S}_f = S_n \times E_{VS}(f, e) \tag{4.15}$$

where $\mathbf{S}_f$ shows the final saliency map of our model.

4.3 Experimental Results

In this section, the results of different parts of our work are presented and evaluated by comparing them with other existing related work. We conducted implementational experiments to demonstrate the performance of the proposed saliency detection model. The implementations were performed using MATLAB R2017b.

As mentioned before, we compared the performance of our saliency model with following existing saliency detection methods: Learning Discriminative Sub-spaces on random contrast (LDS) [98], Maximum Symmetric Surrounding method (MSS) [63], Segmenting salient objects (SEG) [71], and Context-Aware saliency method (CA)

[112].

In this experiment, we used three well-known benchmark datasets consisting of the Complex Scene Saliency Dataset (CSSD) consisting of 200 images [139] (provided by Yan et al. [140]), Extended Complex Scene Saliency Dataset (ECSSD) consisting of 1000 images [140], and MSRA-B dataset consisting of 5000 images (provided by Liu et al. [141]) in order to evaluate the performance of the proposed model. We selected those complex datasets to test the ability and acuity of our model in detecting salient areas.

The performance measurement was assessed by comparing the ground-truth resulting from an eye-tracking mechanism and the saliency map from the saliency detection model as the standard comparison approach in the saliency extraction field.

It should be mentioned that we used the three most common and well-known metrics in the saliency detection literature consisting of AUC (Area Under the Receiver Operating Characteristics Curve), KLD (Kullback-Leibler Divergence), and NSS (Normalized Scan-path Saliency) metrics to evaluate the quantitative performance of the proposed saliency detection model.

There are different versions of AUC used for evaluating of the saliency detection models such as AUC-Judd, AUC-Borji, and Shuffled AUC. We employed AUC-Judd because it covers more general and accurate circumstances. AUC-Borji uses a uniform random sample of image pixels as negatives and defines the saliency map values above the threshold at these pixels as false positives. The false positive calculation in AUC-Borji is a discrete approximation of the calculation in AUC-Judd. Shuffled AUC is very similar to AUC-Borji but it specifically penalizes models that include the center bias.

Since a few approximations in the AUC-Borji implementation can lead to suboptimal behavior, we report AUC scores using AUC-Judd in the rest of this dissertation.

The employed version of AUC metric has been proposed by Judd [142] known as AUC-Judd [143]. In this method, the saliency map is treated as a binary classifier to separate positive and negative samples at various thresholds [143].

The true positive (TP) rate is defined as the proportion of saliency map values located inside the salient regions of both the saliency map and the ground-truth map which are above the threshold. The false positive (FP) rate is the proportion of the saliency map values above the threshold at non-fixated pixels [143]. The thresholds were sampled from saliency map values. This operation is repeated 100 times [142].

Then, the ROC curve can be drawn and AUC computed. An ideal score is 1 while random classification provides a score of 0.5.

The KLD is a commonly used metric to estimate an overall dissimilarity between two distributions. It measures the divergence between the saliency map and the fixation map assumed as distributions. This metric is a non-symmetric measure for information lost when the saliency map is used to estimate the fixation map [143]. Therefore, lower KLD values demonstrate better performance.

The NSS metric introduced by Peters and Itti [144] that quantifies the normalized scan-path saliency between a saliency map and its corresponding fixation map. It is measured as the mean value of the normalized saliency map at fixation locations [143].

We calculated these metrics over each image in the CSSD, ECSSD, and MSRA-B datasets and then obtained the average among all images for each dataset. Tables 4.1, 4.2, and 4.3 show the achieved average AUC, KLD, and NSS metrics for different tested methods over different datasets in this study.

It can be observed from Figures 4.3, 4.4, and 4.5 that the AUC and NSS values of the proposed model are, in general, larger than those of the other compared models and that the KLD value of our model is, in general, lower than other models.

According to these tables, our method outperforms other compared methods. The visual comparison of the estimated saliency maps has been provided in Figures 4.6 and 4.7 for CSSD/ECSSD, and MSRA-B datasets respectively.

Table 4.1: Performance comparison across different methods for CSSD dataset (200 images).

Methods	AUC(%)	KLD	NSS
LDS [98]	86.38	1.69	1.09
MSS [63]	81.44	1.62	1.00
SEG [71]	86.73	1.53	1.10
CA [112]	82.69	1.19	1.08
Ours	88.17	1.16	1.12

Table 4.2: Performance comparison across different methods for ECSSD dataset (1000 images).

Methods	AUC(%)	KLD	NSS
LDS [98]	85.59	1.82	1.11
MSS [63]	74.92	1.95	0.77
SEG [71]	83.43	1.56	0.98
CA [112]	78.96	1.25	0.95
Ours	86.89	1.21	1.09

Table 4.3: Performance comparison across different methods for MSRA-B dataset (5000 images).

Methods	AUC(%)	KLD	NSS
LDS [98]	88.62	1.49	1.25
MSS [63]	82.79	1.53	1.08
SEG [71]	88.06	1.39	1.21
CA [112]	85.91	1.10	1.19
Ours	89.46	1.14	1.24

In general, it is agreed that for good saliency detection a model should meet at least the following three criteria: 1) good detection: the probability of missing real salient regions and falsely marking the background as a salient region should be low, 2) high resolution: saliency maps should have high or full resolution to accurately locate salient objects and retain original image information, and 3) computational efficiency: as front-ends to other complex processes, these models should detect salient regions quickly [1].

In comparison with Goferman et al. [112], Rahtu et al. [71], and Fang et al. [98], our model is simple as well as low computationally complex. Although Achanta et al. [63] provide a fast method, we could obtain more accurate resulting saliency maps. In addition, we only exploited low-level features in designing our model while in [112]

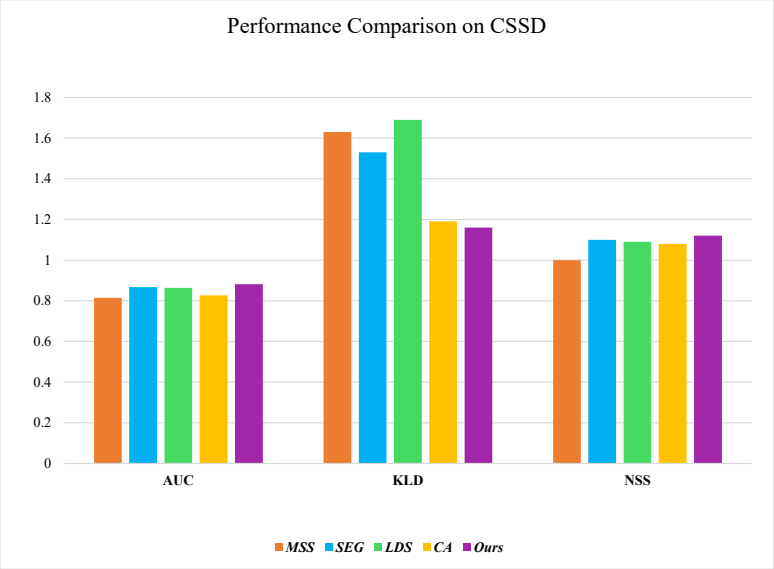


Figure 4.3: Performance comparison across different methods for CSSD dataset (200 images).

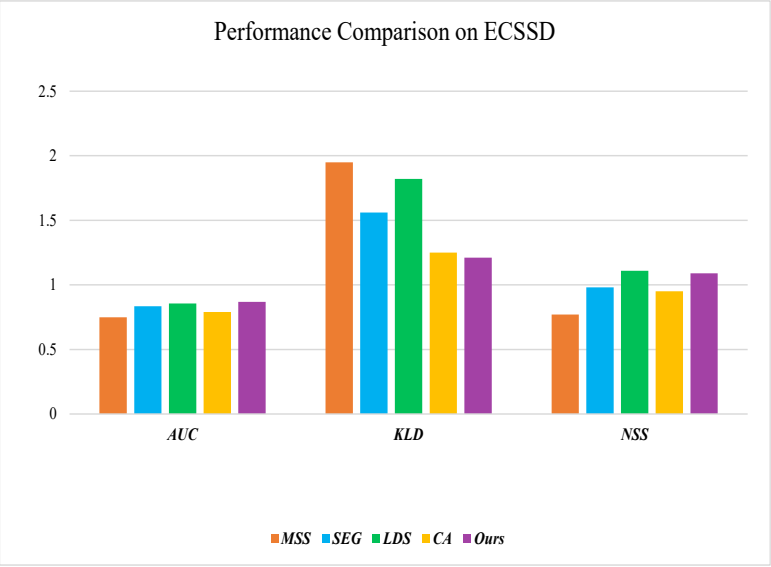


Figure 4.4: Performance comparison across different methods for ECSSD dataset (1000 images).

both low-level and high-level features have been used in which the performance of our method is superior.

Based on the visual results, LDS [98] can detect the location of the salient areas

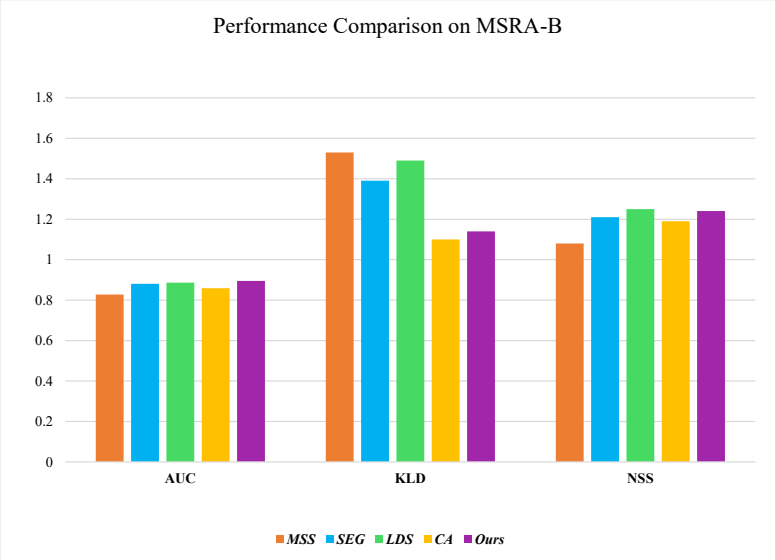


Figure 4.5: Performance domparison across different methods for MSRA-B dataset (5000 images).

correctly however it provides a very blurry saliency map in which objects and boundaries cannot be identified. Moreover, the numbers of pixels that are estimated as the salient area are less than the real areas. The MSS [63] method returns a very accurate saliency map, but on the other hand, it is often unable to extract the entire salient regions properly. The SEG [71] method can extract the salient areas with well-defined boundaries and high quality however the detected areas are not correct sometimes.

The CA [112] method is one of the benchmark methods in the visual saliency field. In general, its performance is good and reliable, but the salient areas are slightly blurry and low-quality.

As seen in Figures 4.6 and 4.7, our proposed method can estimate high-quality saliency map with well-defined boundaries for the salient objects. In addition, we can assign the saliency value for the entire image while most of the existing models in the literature are only able to find the most salient area and discard the rest of an image. This is one of the main advantages that our proposed model provides.

Figures 4.8 and 4.9 show the Precision-Recall (PR) curve for all the compared methods tested on CSSD, ECSSD, and MSRA-B datasets respectively. The PR curves demonstrate that the proposed saliency detection model performs better than the compared models. As shown in the Precision-Recall curve, our saliency model gets the best results for each dataset. This clearly demonstrates the generalization

performance of our proposed method and robustness to dataset biases. Our model explores biologically inspired feature maps and fuses them using a Bayesian framework to investigate visual attention in response to the scene content and has much stronger generalization capability.

To compute Precision and Recall for a saliency map S, it is first converted to a binary mask M and then it is compared with ground-truth map G:

$$Precision = \frac{|M \cap G|}{|M|} \tag{4.16}$$

$$Recall = \frac{|M \cap G|}{|G|} \tag{4.17}$$

Based on aformentioned definition, it can be seen that the binarization of the saliency map (S) is the key step in the evaluation. According to the literature, there are three well-known ways to perform the binarization step [1].

- The first solution proposed by Achanta et al. [95] is based on the image-dependent adaptive thresholding method which is computed as twice as the mean of the saliency value of S.

- The second method, employs a fixed threshold which changes from 0 to 255. A pair of precision and recall scores are computed on each threshold, and they are ultimately combined to form a precision-recall curve to describe the model performance [121,145]. This method is usually applied for large datasets because of its efficiency.

- The third method of binarization uses the SaliencyCut algorithm [88]. In this solution, a loose threshold is used to generate the initial binary mask. Then the method iteratively uses the GrabCut segmentation method [146] to gradually refine the binary mask. The final binary mask is used to re-compute the precision-recall value. This solution is time consuming.

We used the second method for our experimental results. This method introduced by Bylinski et al. [121], uses 256 thresholds to apply over the saliency map and creates a pair of precision and recall vectors with the size of 1×256 for each image within a dataset. Then, the average is computed for each of the 256 corresponding vector

elements among all N images belonging to a specific dataset. Therefore, we can plot a PR curve for all N images of a dataset into a graph.

In terms of speed comparison among the methods, it should be mentioned that our method is capable of providing a well-defined saliency map with high-quality within a short time period compared to the most of graph-based and learning-based models such as CA [112] which are very reliable from saliency quality perspective. For example, we tested all five methods with a sample image with a resolution of 400×266 pixels and obtained the average run-time over 10 iterations as seen in Table 4.4. The test was performed using a computer with Core (TM) i7-2600, CPU 3.4 GHz, and RAM 12GB.

Table 4.4: Speed Comparison Across Different Methods.

Methods	MSS [63]	SEG [71]	CA [112]	LDS [98]	Ours
Time (sec)	2.87	6.72	37.34	0.89	6.42

According to Table 4.4, LDS and MSS methods are faster however the quality of their extracted saliency map is lower than other models in this study. In general, the proposed method has a reasonable speed considering its high quality of performance and, when optimized, a good choice for real-time applications.

4.3.1 Gradual Saliency Validation

In the field of saliency detection, the available ground-truth datasets are divided into two types: 1) fixation density maps resulting from an eye-tracking procedure, and 2) object segmented binary masks (used in this chapter). Therefore, it is difficult to evaluate gradual saliency maps which we have proposed in this dissertation relying on those ground-truth data. To overcome this limitation, we performed an eye-tracking based experiment using human subjects in order to validate our created gradual saliency maps.

In this validation test, 14 subjects (5 females and 9 males) within age range 18-35 participated. We used an SMI eye-tracker iView 120 Hz device and a 55-inch TV screen in a similar manner as described in Chapter 3. It is impracticable to perform the test over a large number of images, due to time limitations for an eye-tracking

test, because it will cause eye fatigue and invalid recorded data. Therefore, we could only perform the test on a limited number of images, i.e., the CSSD dataset with 200 images. Each image was shown for 4 seconds with a gray frame in between for 3 seconds to discard the previous fixation points.

We showed the results of our test as the eye fixation path for several sample images in the Figures 4.10, 4.11, and 4.12. A fixation path is one type of eye-tracker output which consists of several dots connected by straight lines indicating the points that are fixated by participants in their temporal order. The fixation points/dots are numbered sequentially, and each point is shown by a circle in which its radius represents the fixation duration. Furthermore, a larger radius denotes a longer gaze on a point and vice versa. The points are connected by straight lines showing the path that human eyes have moved while watching a scene.

Furthermore, we showed the object ranking analysis of our test for several sample images on CSSD dataset on Table 4.5. To compute the number of fixations on each object within an image, we counted the number of fixations among all participants over the assumed object seen with the same priority. For instance, on the image "Boat", the ranking of the boat was obtained based on how many times participants looked at the boat as the main salient object which was 13 (out of 14).

According to our analysis, our gradual saliency maps were generally correct for approximately 81% of the cases.

4.4 Conclusion

In this chapter, we introduced a saliency detection framework for 2D still images using color, luminance, intensity, and texture attributes. The algorithm extracts the corresponding colors and texture patterns resulting from our eye-tracking based subjective study. Four feature maps were extracted including color, color contrast, texture, and intensity maps. A naive Bayesian network was employed to estimate an ultimate saliency map using the extracted feature maps. Furthermore, human visual acuity factor was exploited to enhance the saliency map.

The ability to assign the prioritized saliency for the entirety of an image/video is the main goal of this work. Another advantage of this model is its ability to extract

the salient areas with well-defined boundaries and high accuracy with the entire area of an of interest object.

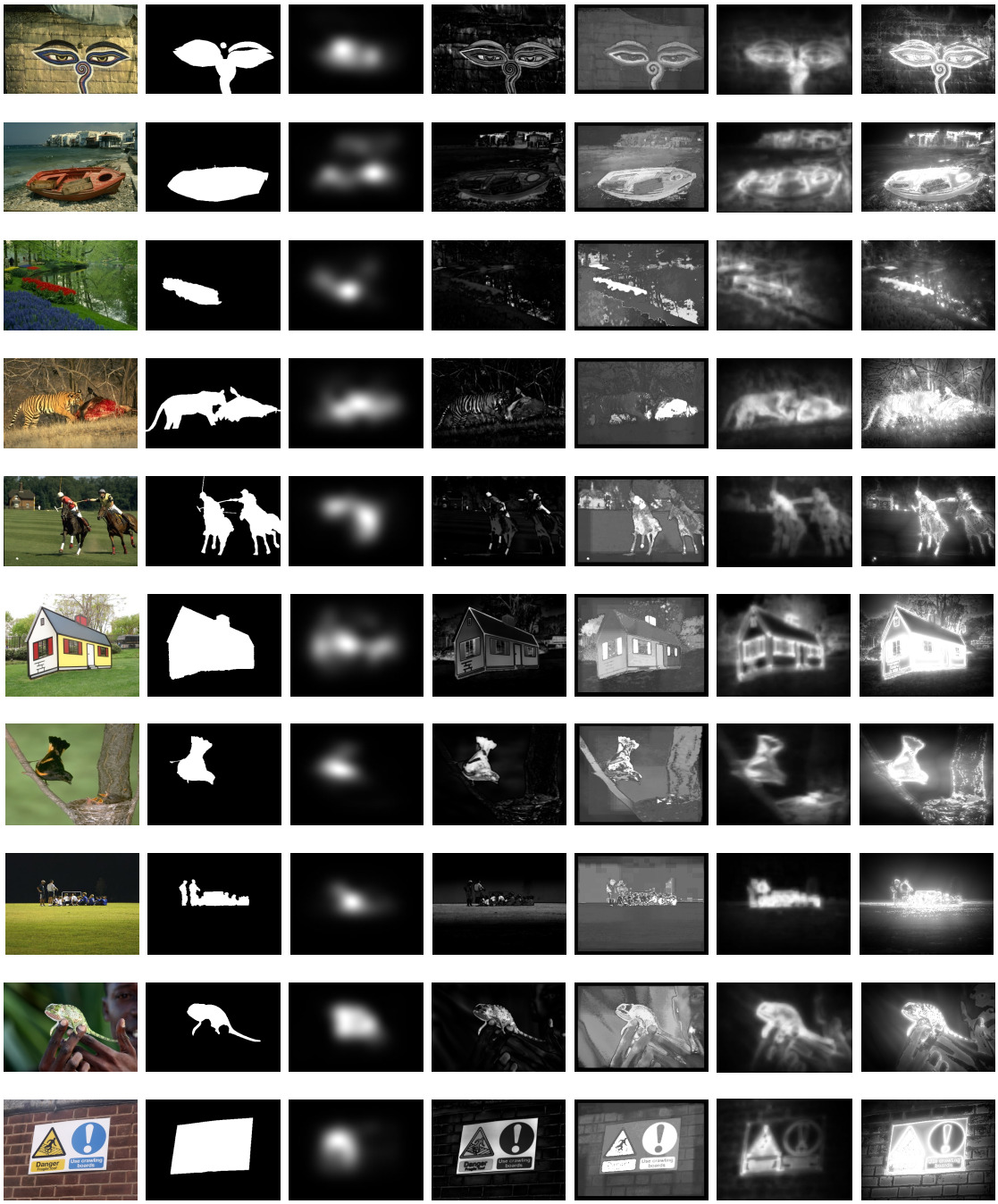


Figure 4.6: Visual comparison of saliency estimation from different models over CSSD, and ECSSD datasets: (a) input image; (b) ground truth map; (c) Learning discriminative sub-spaces method (LDS) [98]; (d) Maximum symmetric surrounding method (MSS) [63]; (e) Segmentation salient object method (SEG) [71]; (f) Context-aware saliency method (CA) [112]; (g) our method.

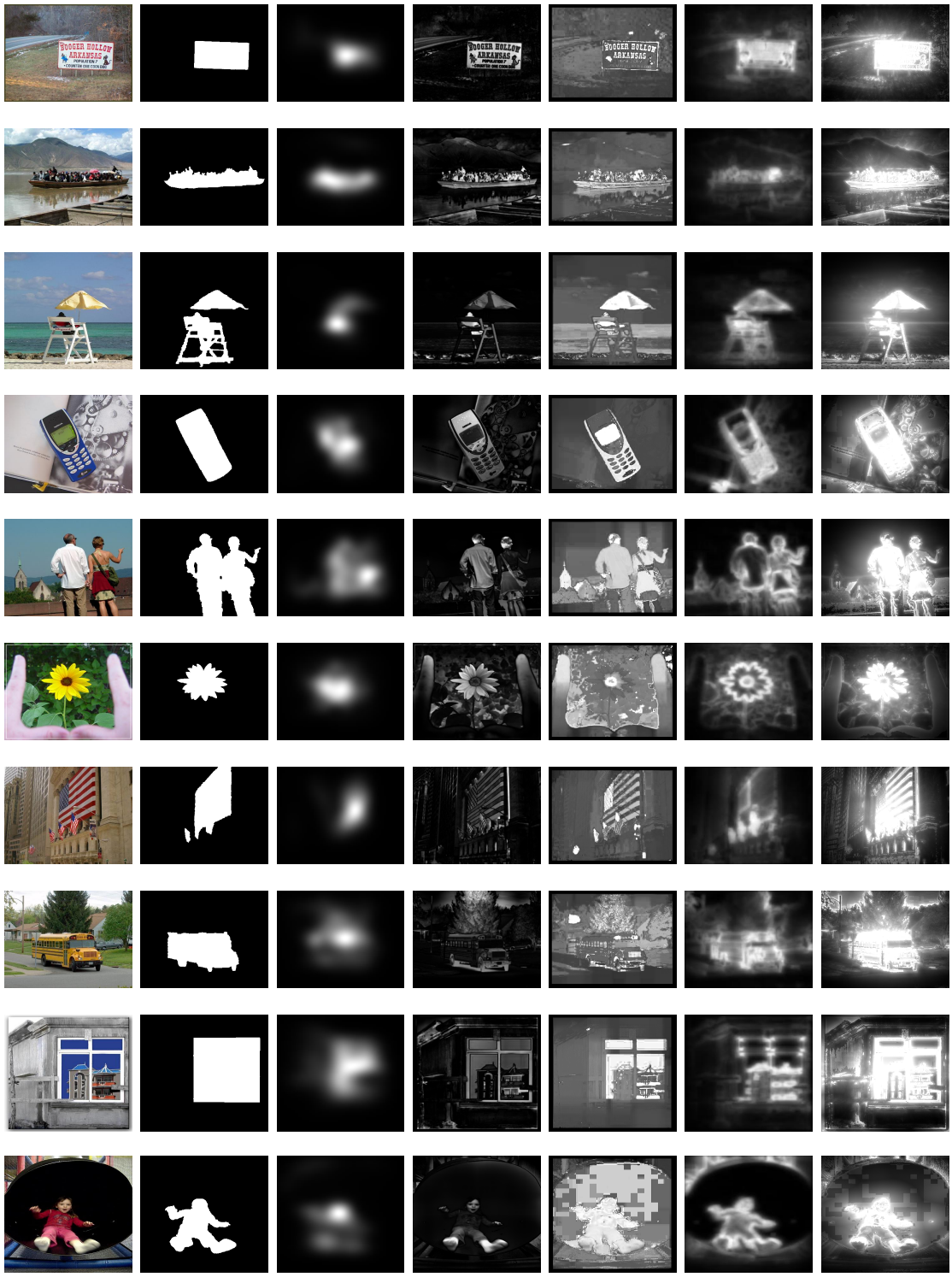
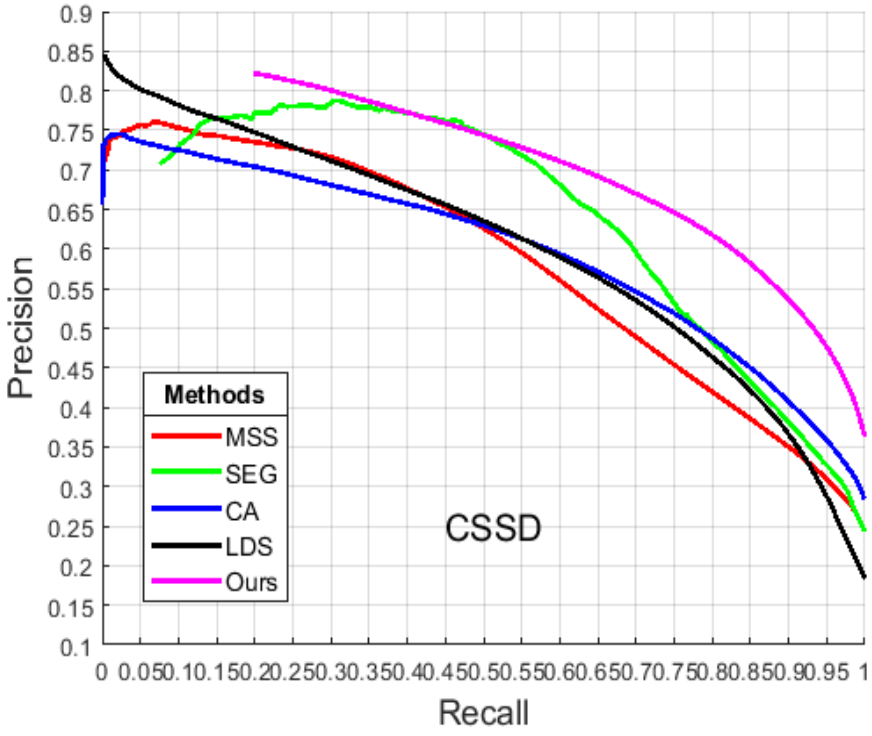
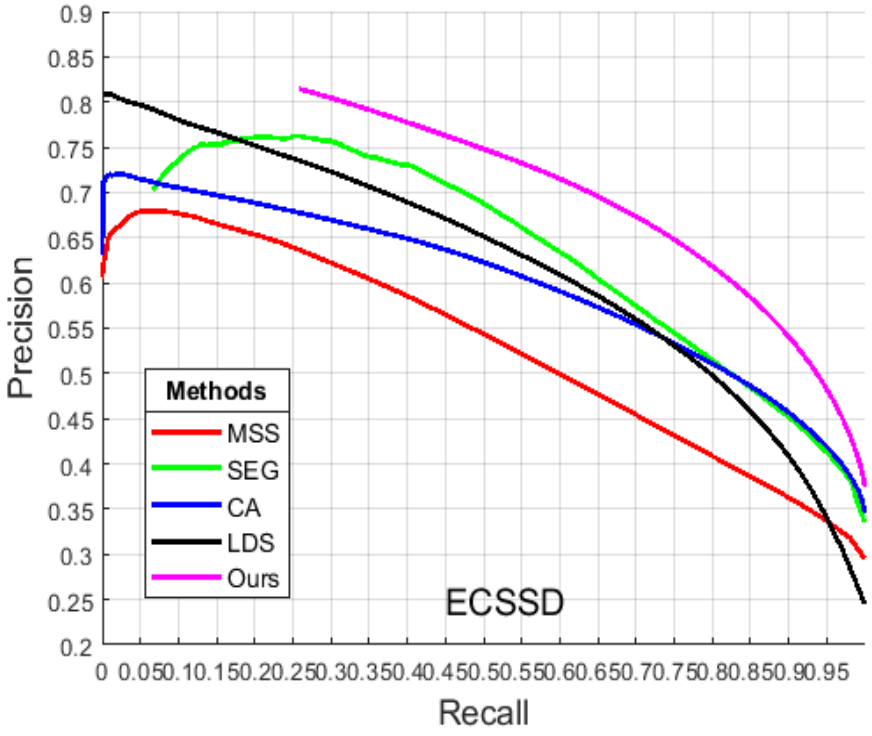


Figure 4.7: Visual comparison of saliency estimation from different models over MSRA-B dataset: (a) input image; (b) ground truth map; (c) Learning discriminative sub-spaces method (LDS) [98]; (d) Maximum symmetric surrounding method (MSS) [63]; (e) Segmentation salient object method (SEG) [71]; (f) Context-aware saliency method (CA) [112]; (g) our method.



(a)



(b)

Figure 4.8: Quantitative Precision-Recall performance of all compared models over different datasets: (a) the CSSD with 200 images; (b)the ECSSD with 1000 images.

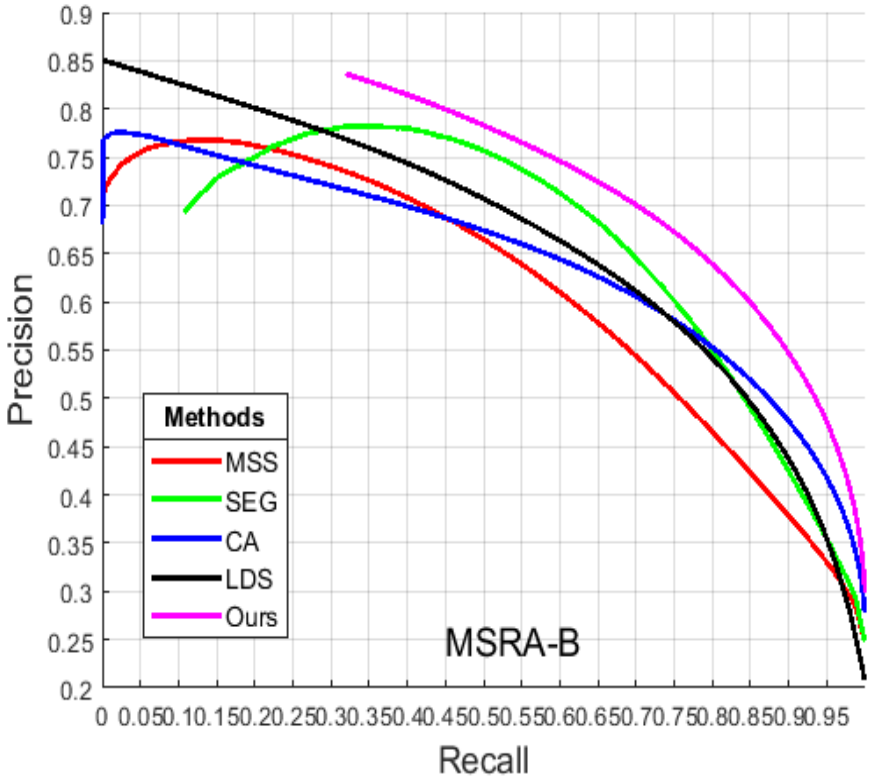
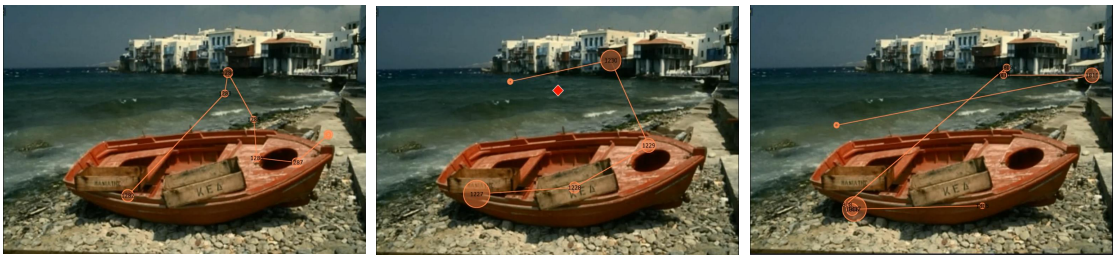


Figure 4.9: Quantitative Precision-Recall performance of all compared methods on the MSRA-B dataset with 5000 images.

Table 4.5: Saliency Ranking for Different Objects of Sample Images (CSSD Dataset) Resulting in our Eye-tracking Experiment.

Ranking	Object	Fixation # (out of 14)	Fixation % (divided by 14)
		Image: Boat	
1	Boat	13	92.86
2	Building	11	78.57
3	Water	8	57.14
4	Ground	2 (out of 2)	14.29
		Image: Horsing	
1	Red horseman	13	92.86
2	Yellow horseman	11	78.57
3	Ball	4 (out of 5)	28.57
4	House	0 (no fixation)	0
		Image: Lizard	
1	Lizard	14	100
2	Man's face	10	71.43
3	Man's hand	8	57.14
		Image: Bird	
1	Bird	12	85.72
2	Chicks	12	85.72
3	Tree	6 (out of 6)	42.85
		Image: Sign	
1	Yellow sign	13	92.86
2	Blue sign	14	100
3	Background	5 (out of 6)	35.72
		Image: Garden	
1	Red flowers	13	92.86
2	Yellow flowers	9	64.28
3	Water/Shine	10	71.43
4	People	7	50
5	Blue flowers	2 (out of 5)	14.29

a) From left to right: original image, ground-truth, and our saliency map.



b) Three sample fixation-paths from three different participants.

c) From left to right: original image, ground-truth, and our saliency map.

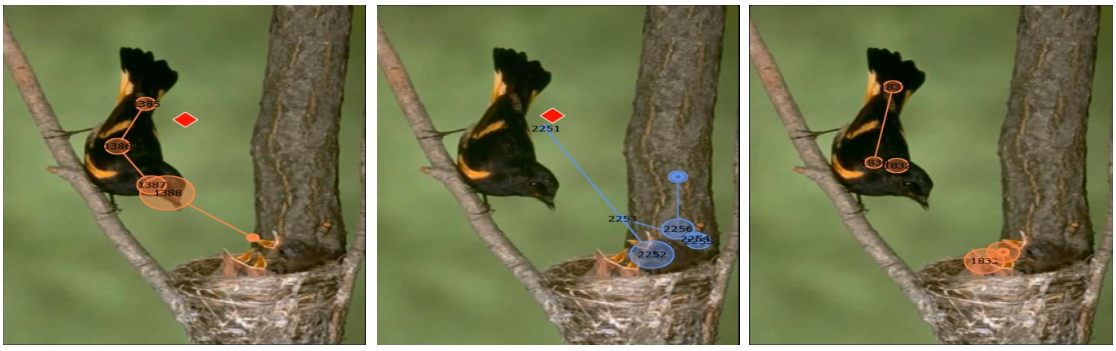
d) Two sample fixation-paths from two different participants.

Figure 4.10: Visual representation of validation for gradual saliency estimation over sample images of CSSD dataset. The fixation-path resulting from eye-tracking procedure was shown for several participants: (on part (d), the 2nd and 3rd images are consecutive over the time for the same participant).

a) From left to right: original image, ground-truth, and our saliency map.

b) Three sample fixation-paths from three different participants.

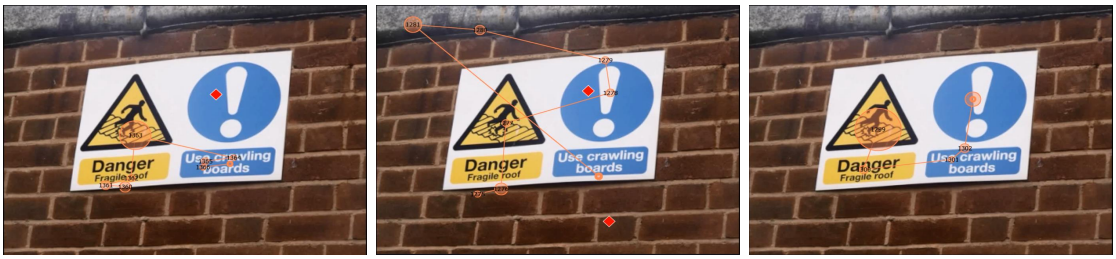
a) From left to right: original image, ground-truth, and our saliency map.



b) Three sample fixation-paths from three different participants.

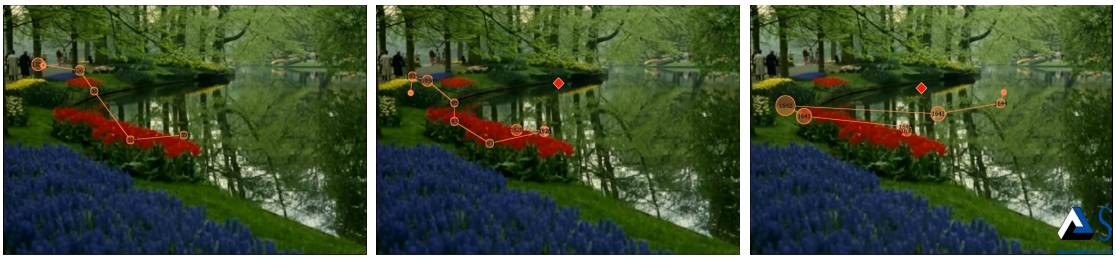
Figure 4.11: Visual representation of validation for gradual saliency estimation over sample images of CSSD dataset. The fixation-path resulting from eye-tracking procedure was shown for three different participants.

a) From left to right: original image, ground-truth, and our saliency map.



b) Three sample fixation-paths from three different participants.

a) From left to right: original image, ground-truth, and our saliency map.



b) Three sample fixation-paths from three different participants.

Figure 4.12: Visual representation of validation for gradual saliency estimation over sample images of CSSD dataset. The fixation-path resulting from eye-tracking procedure was shown for three different participants.

Chapter 5

Dynamic Saliency Detection Model

5.1 Introduction

In this chapter, we explain our spatio-temporal algorithm used to detect saliency in 2D video sequences. The spatial domain-based saliency detection method was previously described in Chapter 4. The temporal saliency was extracted using the optical flow field and energy minimization problem. Then, spatio-temporal information - including all feature maps (i.e., color, texture, and motion maps) - was combined using a Bayesian framework, as explained in Chapter 4.

This saliency detection framework was established based on bottom-up features: color, luminance, intensity, texture, and optical flow field. The model used energy optimization of the flow field to extract the more effective movements among video frames. Then, all the feature maps were normalized and weighted according to our experimental study on human eye tracking and fixations described in Chapter 3. Finally, different feature maps were combined using a Naive Bayesian Network to form an ultimate gradual saliency map.

We used both static and dynamic features in our proposed model because motion information alone is insufficient for identifying the salient regions since the moving objects may have a very small optical flow, or the background may be dynamic.

We compared our model to three related state-of-the-art methods using publicly available ground-truth data and their source code. The three methods are as follows: Consistent video saliency using local gradient flow optimization and global refinement (LFGR) [89] by Wang et al., Segmenting salient objects (SEG) [71] by Rahtu et al., and Itti's model [55]. The aforementioned models are briefly explained here.

Wang et al. [89] presented a spatio-temporal saliency detection method based on

the gradient flow field and energy optimization. Their proposed gradient flow field incorporates two distinctive features: 1) intra-frame boundary information and 2) inter-frame motion information together to find the salient regions. They utilized both intra-frame and inter-frame information in the gradient flow field to estimate the object and background in a scene, and they tried to suppress the background. This method also introduces local and global contrast saliency measures using the foreground and background information estimated from the gradient flow field. They further proposed an energy function to achieve the spatio-temporal consistency of the output saliency maps [89].

Wang et al. [89] first applied the well-known SLIC method to abstract each frame into superpixels to investigate a frame region by region for both contrast and gradient flow field features. In the spatial domain, the color gradient magnitude of the abstraction frame was computed. To estimate motion, the optical flow field of [90] was used, and then the magnitude of the gradient of the flow field was computed. Finally, both color gradient magnitude and optical flow gradient magnitude were combined using multiplication into a spatio-temporal gradient field. Furthermore, the motion field was emphasized by an exponential weighting function.
They further defined a global saliency measure of a superpixel as the length of its shortest distance to the virtual backgrounds. The distance between any two superpixels (i.e., regions) considers the color distance and the gradient flow field distance. In this way, the method can suppress background and detects only the salient objects which usually belong to the foreground.

Rahtu et al. [71] introduced a salient object segmentation method based on combining a saliency measure with a Conditional Random Field (CRF) model. This model employed a statistical Bayesian framework and local contrast of color, and motion features. They used optical flow to extract motion information. To recover well-defined salient objects, a CRF model was applied to the resulting saliency map using an energy minimization-based approach.
Rahtu's saliency measure is based on applying a sliding window to the input image. Next, the contrast between the distributions of certain features is comparing in each window which is a comparison in an inner window to the distribution in the collar of the window [71]. They used the features like intensity, the value of different color channels, and motion information. This saliency measure defines a pixel as salient if the feature at that pixel is similar to the features at points of the inner window and

different from points in the border of the window [71].

The earliest method for saliency detection was proposed by Itti et al. [55], who is one of the pioneer researchers in the field. He established this model based on Treisman's feature integration theory [19] and Koch's biological structure [56]. This visual attention system was inspired by behaviour and the neural architecture of the early primate visual system. Itti's visual attention model employs bottom-up features, including color, intensity, and orientation, to calculate the local contrast of these features using the difference of the feature vectors in a center-surround neighborhood mechanism [55].
They exploited a multi-scale mechanism based on Gaussian pyramids to create different resolutions of an input image (i.e., down-sampling). Then, the feature vectors were extracted from each pyramid to compute their differences within a neighborhood and finally, the resulting feature vectors were normalized. At the next step, they used a dynamical neural network called the Winner-Take-All (WTA) network among different feature maps to select the important locations and obtain the ultimate saliency map.

5.2 Proposed Dynamic Model

We extracted different feature maps including color contrast map, color importance ranking map, texture contrast map, a variance of intensity map, and motion map (which contains optical flow field, velocity, and acceleration information). The feature maps related to color and texture attributes have been previously explained in detail in Chapter 4. The motion map will be described in the current Chapter.

All resulting feature maps were weighted based on our HVS based experiment from Chapter 3 and then normalized. Later, all maps were merged through a Bayesian framework. Additionally, we exploited the human visual acuity factor to enhance the final saliency map based on the characteristics of the HVS. Figure 5.1 shows a block diagram of our model to help a better understanding of different steps of that. This framework was explained in detail in the following subsections.

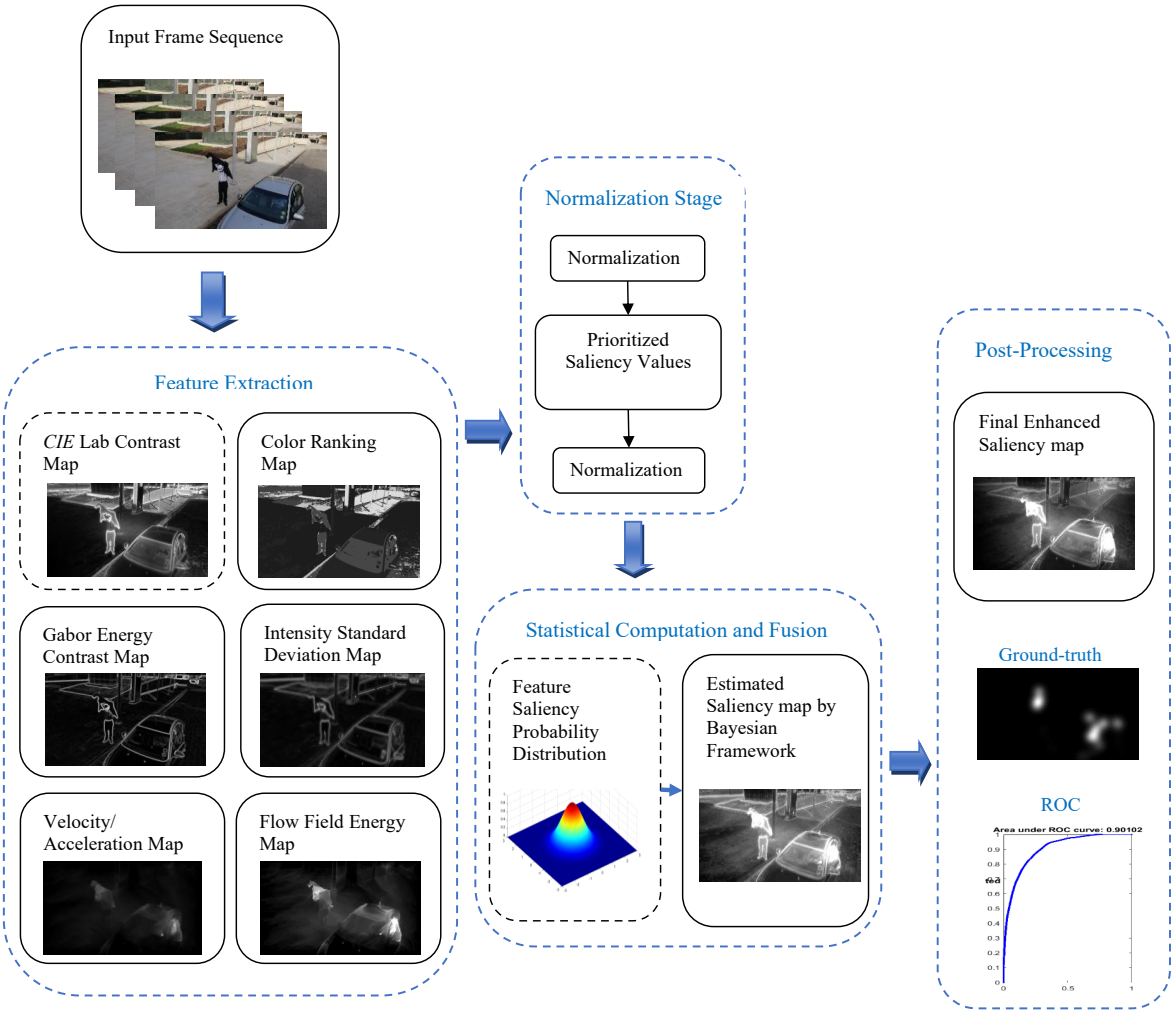


Figure 5.1: Block diagram of our framework for salient region detection.

5.2.1 Motion, Velocity, and Acceleration maps

As mentioned in Chapter 3, due to humans' biological instincts, moving objects attract human attention.

We weighted them based on the importance of each attribute using our experimental results (Chapter 3).

In order to extract the motion map for the video sequences, we used a combination of the local and global optical flow methods because local methods are usually more robust under noise, while global techniques yield dense flow fields.

We incorporated the gradient flow field algorithm introduced by Liu [147] as the

base of our motion detection since this method has shown a promising performance on various datasets and is publicly available. Furthermore, we improved its performance by combining it with the Brox's [90] optical flow method to have more accuracy.

We considered three different motion directions: horizontal (dx), vertical (dy), and diagonal $[dx, dy]$. Note that dx and dy are calculated using the optical flow algorithm which are usually shown by (u, v) indicating the displacement of a pixel. In other words, $(i_x + dx, i_y + dy)$ in the current frame is the approximate location of (i_x, i_y) in the previous frame.

Once the motion map was generated, it was normalized. In addition to the mentioned map, more motion maps were created from the velocity vectors and used in our scheme as follows:

Velocity in different directions is defined as Equation 5.1.

$$Vx = fr \times u \quad , \quad Vy = fr \times v \tag{5.1}$$

where, fr is the frame rate of the video and Vx, and Vy indicate the velocity on the horizontal and vertical directions, respectively. The velocity vector magnitude is evaluated as Equation 5.2.

$$V = \sqrt{Vx^2 + Vy^2} \tag{5.2}$$

In our implementations, the velocity was considered across multiple frames, so that, it was computed every 10 frames.

According to our experimental results, objects or areas with relatively high acceleration are generally considered to be salient. We included a map of relative acceleration to our set of motion saliency maps by computing the acceleration using the Equation 5.3, [130].

$$A = \Delta V / \Delta t = fr \times (V_{Currentframe} - V_{Referenceframe}) \tag{5.3}$$

where Δ shows the variations of the velocity and the time. Velocity was calculated using 5.2 for the current and reference frames. We considered multiple frames to compute the acceleration (i.e., every 10 frames). Also, the reference frame indicates a frame at the start of every 10 frames time interval.

5.2.2 Optical Flow Field Algorithm

In order to extract the motion map for the video sequences, we used a combination of the local and global optical flow methods because local methods are usually more robust under noise, while global techniques yield dense flow fields. We incorporated the gradient flow field algorithm by Liu [147], which is a combination of Lucas/Kanade [148] (local) and Horn/Schunck [149] (global) algorithms.

They used the smoothing/regularisation processes that are required in local and global differential methods for optical flow computation. They combined the advantages of local and global approaches by obtaining dense flow fields that are robust against noise. This hybrid method can be applied for spatio-temporal and non-linear extensions as well as multi-resolution frameworks which are very helpful for saliency detection models. The method minimizes the energy functions in the optical flow field to sparsify a dense flow field and consequently creates a reliable flow field for our motion map.

The core of the optical flow field algorithm that we used in this dissertation, is based on [147], [150], and [151]. The major difference between Liu's algorithm [147] as the main method and [150, 151] is that Liu used the conjugate gradient to solve large linear systems. The reason behind using a conjugate gradient is because this makes the solver very simple and easy to implement.
Assume if we need to solve a large linear equation $Ax = b$ in every inner step (where x is the flow field), then it may be found that matrix A can be decomposed into concatenations of filtering and weighting. However, in a conjugate gradient solver, it is not required to formally write down matrix A, but it is only needed to write a function that consists of filtering and weighting to apply A to x [147].

Finally, they used an iterative re-weighted least square (IRLS) method instead of Euler-Lagrange to derive an optical flow solver. Both IRLS and Euler-Lagrange end up with the same equations, however, IRLS is much easier to understand and simpler to derive.

Applied Constraints

Here, we explained the optical flow optimization problem formulated by Liu [147], therefore we provided an intuitive idea of which constraints are required to be included in such a model.

In the optical flow estimation state-of-the-art [152, 153], it has been assumed that

the grey value of a pixel is not changed by the displacement which results in Equation 5.4 known as *grey value constancy assumption* [150].

$$I(x, y, t) = I(x + u, y + v, t + 1) \tag{5.4}$$

where I represents an image/video frame, and $w = (u, v, 1)$ denotes the optical flow field which is the displacement vector between two frames at time t and $t + 1$ respectively. By linearising the grey value constancy assumption, the famous optical flow constraint is obtained [90] as follows:

$$I_x u + I_y v + I_t = 0 \tag{5.5}$$

In the Equation 5.5, subscripts of x, y, and t indicate partial derivatives. However, the linearization is only valid under the assumption that the video frames change linearly along the displacement [150], which cannot cover all kinds of movements especially along with curvature or large displacements. In order to have a general optical flow algorithm, most of the existing models including Liu's model employ the original non-linearised grey value constancy assumption mentioned in Equation 5.4. Since the grey value constancy assumption i.e., Equation 5.4 is sensitive to slight changes in brightness in a scene, usually optical flow estimation algorithms must exploit another assumption to be able to determine the displacement vector which is invariant under grey value changes.

Such a criterion is the gradient of the image grey value that is known as *gradient constancy assumption* and allows small variations in the grey value which often appear in natural scenes [153]. It can be written as Equation 5.6.

$$\nabla I(x, y, t) = \nabla I(x + u, y + v, t + 1) \tag{5.6}$$

where $\nabla = (\partial x, \partial y)^T$ denotes the spatial gradient. This can be suited for more complicated motion patterns [147].

In our employed optical flow field algorithm, two further assumptions were used as the smoothness of the flow field and multi-scale approach introduced by Brox et al. [150]. The smoothness constraint can either be applied solely to the spatial domain (for only two frames), or the spatio-temporal domain (for a sequence of frames). Since the displacement field encounters discontinuities at the boundaries of objects within a scene, it is helpful to generalize the smoothness assumption. It should be noted

that if the displacements are larger than one pixel per frame, then a minimization algorithm could easily be trapped in a local minimum [147]. Therefore, multi-scale approaches must be applied to find the global minimum in the final optimization problem - which describes the energy of the objective function.

Objective Function

After all the constraints description, now we can represent the energy functional that has been used in the current dissertation.
We combined the objective functions used in [147], and [90] for optical flow estimation in our saliency detection model. In this way, the data term, smoothness term, gradient flow, and symmetric flow computation were incorporated to improve accuracy.

Suppose that $I_1, I_2 : (\Omega \subset \Re^2) \rightarrow \Re^d$ be the first and the second frame of an input video sequence [150] (for a gray scale image d = 1 and for color images d = 3) and also let $\chi := (x, y, t)^T$ and $w := (u, v, 1)^T$. Then, the global deviations from the grey value constancy assumption and the gradient constancy assumption are measured by the energy function as described in the following.

First, the data term can be formulated by Equation 5.7.

$$E_{data}(w) = \int_{\Omega} \Psi\Big(g * |I_1(x+u, y+v) - I_2(x, y)|\Big) d\chi \tag{5.7}$$

where g is a Gaussian filter. The L_1 norm was used here to take into account the outliers matching [147]. Moreover, $\Psi(s^2) = \sqrt{s^2 + \epsilon^2}, \epsilon = 0.001$ is the robust function which makes the objective function be able to deal with difficult matching situations such as occlusions and non-Gaussian deviations [150]. This non-linear robust function corresponds to a Laplace distribution which is more preferable than Gaussian distribution because it has a longer tail and high convergence chance.

Second, smoothness term is imposed by Equation 5.8.

$$E_{smooth}(w) = \int_{\Omega} \Psi\Big(|\nabla u|^2 + |\nabla v|^2\Big) d\chi \tag{5.8}$$

The smooth term E_{smooth} provides spatio-temporal consistency constraint indicating that all the spatio-temporally adjacent regions of the whole video sequence should have the same saliency when they are similar [147].

Third, we decided to supplement the constraint in the data term by a constraint on the gradient and make it more reliable. Due to illumination effects, matching the

color or gray value is not always reliable [150] therefore we used gradient term to improve it. This constraint is invariant to additive brightness change [150].

$$E_{grad}(w) = \int_{\Omega} \Psi\left(|\nabla I_2(x + w(x)) - \nabla I_1(x)|^2\right) d\chi \tag{5.9}$$

The gradient constraint provides solid invariance properties without being as sensitive to noise as second-order constraints.

Lastly, symmetric matching is required because we want to consider the local properties of the optical flow among video sequences. The symmetric term can be defined as follows:

$$E_{sym}(w) = \int_{\Omega} \Psi\left(|u(x,y) + u(x+u, y+v)| + |v(x,y) + v(x+u, y+v)|\right) d\chi \tag{5.10}$$

Finally, the objective function used in this dissertation can be obtained by the sum of the above four terms, as shown in Equation 5.11.

$$E(w) = E_{data} + \alpha E_{smooth} + \beta E_{grad} + \gamma E_{symtr} \tag{5.11}$$

where α, β, and γ are tuning parameters that can be determined and tuned using Liu's model.

As the next step, IRLS was used to the outer and inner fixed point iterations proposed in Liu's model [147], combined with a coarse-to-fine search to optimize this objective function.

It is important to note that this model is extremely general from the modeling perspective which makes it capable of dealing with all kinds of deformations, motion discontinuities, occlusions, and large displacements. However, due to the approximative optimization of this energy model, the optical flow estimation is not a perfectly solved problem yet [147]. Therefore, the problems in the optical flow estimation area should not reflect any shortcomings related to the model itself.

5.2.3 Saliency Map Estimation and Enhancement

To obtain the ultimate saliency map for dynamic scenes, color, texture, and motion maps are fused in a similar manner described in Chapter 4. For this purpose, we used

a Naive Bayesian Network where the salient regions model descriptor can be built by computing the likelihood probability distributions of the resulting feature maps. The probability density function of each individual feature map was estimated as a Gaussian distribution and the model parameter i.e., the mean and the variance were learned.

First, the feature maps were weighted and normalized based on the results of our empirical study [132–134] indicating that the significance of color, texture, and motion features with respect to each other. Next, the probability functions of the extracted feature maps were learned. Then, both the prior and likelihood probabilities were learned from the extracted feature maps and the posterior probabilities were estimated. Therefore, the computation of the saliency probability simplifies to multiplying the likelihoods together ($n = 6$) as can be seen in Equation 5.12.

$$p(X_S|F) = p(X_S)(\prod_{i=1}^{n} p(F_i|X_{s_i})) \tag{5.12}$$

In this study, the prior $p(X_S)$ was assumed as a uniform distribution between classes which can be defined as $1/C$, where C is the number of classes including strongly salient, average salient, weakly salient, non-salient regions. We defined these classes to be able to prioritize the saliency within the entire image/frame.
To solve this problem the log of the likelihood was applied [65, 67, 71] which transformed the likelihood products into likelihood summations to obtain the saliency value S_n in Equation 5.13.

$$S_n = p(X_S|F) = p(X_S)(\sum_{i=1}^{n} log(p(F_i|X_{s_i}))) \tag{5.13}$$

As the final step, similar to Chapter 3 (and our works in [156,157]), we employed a characteristic of the HVS to enhance the resulting saliency map known as eccentricity sensitivity [135] which emphasizes on the salient areas and removes their farther surrounding areas by blurring them.

Therefore, the final enhanced saliency map by applying the normalized visual sensitivity can be calculated as [135].

$$\mathbf{S}_f = S_n \times E_{VS}(f, e) \tag{5.14}$$

where $\mathbf{S}_f$ shows the final saliency map of our model.

5.3 Experimental Results

In this section, the results of our work have been presented and evaluated by comparing them with other existing related work. We conducted implementational experiments to demonstrate the performance of the proposed saliency detection model. The implementations have been performed using MATLAB R2019a and C++.

We evaluated and compared the performance of our saliency model with following related existing saliency detection methods: Consistent Video saliency using Local gradient Flow optimization and Global Refinement (LFGR) by Wang et al. [89], Segmenting salient objects SEG by Rahtu et al. [71], and a modified version of Itti's model [55].

In this experiment, we used two sets of well-known publicly-available benchmark datasets in order to evaluate the performance of the proposed model. These datasets are EyeTrackUAV (with dynamic background) and SAVAM (with static background).

EyeTrackUAV

The EyeTrackUAV dataset (provided by Mueller et al. [154]), was used for the scenarios with the dynamic backgrounds. It contains different clips that have been captured using moving cameras. Therefore, there is more than one moving object in each clip, where one of them is the main object, and also, their backgrounds are not simple. We selected one of the difficult clips with a cluttered background, including 1393 frames with a resolution of 1280×720 pixels and a rate of 30 frames per second.

Figures 5.2 and 5.3 show the visual comparison of the results of our method compared with three different mentioned models.

Moreover, we calculated three main metrics in the saliency detection field known as AUC, KLD, and NSS metrics to evaluate the performance of our method compared to the others. The metrics were computed over each frame of the dataset and then obtained the average among all frames. Table 5.1 shows the achieved average AUC, KLD, and NSS metrics for EyeTrackUAV.

SAVAM

To test the datasets with the static backgrounds, we used SAVAM (Semiautomatic Visual-Attention Modeling) dataset (provided by Gitman et al. [155]) and selected different video clips with total 1125 frames. The frame resolution of this dataset is

1920×1080 pixels, and the rate is 30 frames per second.

Although the selected clips are considered static scenes, there are slight movements in objects belonging to the background. For example, the leaves of the tree displaced by the wind. The visual comparison of the results for this clip across different methods was shown in Figure 5.4. Also, the average AUC, KLD, and NSS metrics were computed over the SAVAM dataset among all frames, and are shown in Table 5.2.

We selected complex datasets with both static and dynamic backgrounds to test the ability and acuity of our model in detecting salient areas.
The performance measurement was assessed by comparing the ground-truth resulting from an eye-tracking mechanism and the saliency map from the saliency detection model as the standard comparing approach in the saliency extraction field.

It should be mentioned that since the available ground-truth datasets are usually binary masks it is difficult to compare a gradual saliency map concerning them. To overcome this barrier, we used different levels of thresholds to create different levels of saliency across ground-truth datasets to be able to evaluate our model. We applied this to compute the AUC metric as it does not apply to the two other metrics. Finally, we computed the average of resulting AUC values among different thresholds.

The model proposed by Chen et al. [118] is also similar to our model in terms of employing the low-level features and using the optical flow gradient to create the saliency map. However, our model includes further comprehensive feature maps and more reliable extracted saliency results because it can segment the entire object with a fixed level of saliency. Chen's model does not produce a homogeneous area as a salient object.

Table 5.1: Performance Comparison across Different Methods for EyeTrackUAV Dataset (1393 frames).

Methods	AUC(%)	KLD	NSS
LFGR [89]	81.38	2.05	1.56
SEG [71]	74.72	2.35	1.48
Itti [55]	72.86	1.99	1.04
Ours	87.51	1.46	1.71

Table 5.2: Performance Comparison across Different Methods for SAVAM Dataset (1125 frames).

Methods	AUC(%)	KLD	NSS
LFGR [89]	83.98	1.98	1.25
SEG [71]	81.75	1.86	1.37
Itti [55]	71.04	1.92	0.87
Ours	86.02	1.17	1.32

5.3.1 Gradual Saliency Validation

The available ground-truth datasets in the saliency detection research area, are divided into two types: 1) fixation density maps resulting from an eye-tracking procedure (used in this chapter), and 2) object segmented binary masks. However, we have proposed the gradual saliency concept, meaning it is difficult to evaluate using such ground truth datasets.

To overcome this limitation, we decided to design an eye-tracking based experiment to record human subjects' eye movements while watching sample standard publicly available video sequences. For example, we used the SAVAM dataset in our test. Therefore, we can validate our created gradual saliency maps.

In this validation test, we had 14 human participants and used an SMI eye-tracker iView 120 Hz device as well as a 55-inch TV screen as Chapter 3. Due to time limitation for an eye-tracking test (to avoid eye fatigue), we were only able to perform the test over a limited number of frames, therefore, we only tested the SAVAM dataset.

We presented the resulting eye fixation path from our test for different sample frames in Figure 5.5. We have showed the accumulated gaze points over all participants (which is available in the SAVAM package) indicating the fixation density areas as well as two sample fixation paths from two different participants.

A fixation path is one type of eye-tracker output that consists of several dots connected by straight lines indicating the points that are fixated by participants in their order. The fixation points/dots are numbered sequentially, and each point is shown by a circle in which its radius represents the fixation duration. Furthermore, a larger radius denotes a longer gaze on a point and vice versa. The points are connected by straight lines showing the path that human eyes have moved while watching the

scene.

Moreover, we showed the object ranking analysis of our test for a sample video clip on Table 5.3. This analysis is slightly different for video sequences compared to still images. Each video clip is usually divided into different conceptual parts that play an important role in guiding the human visual system. Therefore, there is more consistency among participants' eye movements compared to still images because of the semantic information of a video.
This video clip contains four main parts showing that a man is first standing and then opens an umbrella, rotates it, and finally moves it above his head.

In Table 5.3, the number of fixations among all the 14 participants for each object was shown. These quantities were computed with the assumption that participants have seen an object with a specific level of importance/priority. For instance, in part 3 of the sample video, we computed how many times participants looked at the umbrella as a first priority.

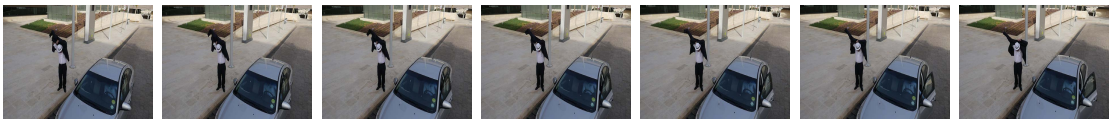
According to our analysis, our gradual saliency maps were generally correct for approximately 88% of the cases.

5.4 Conclusion

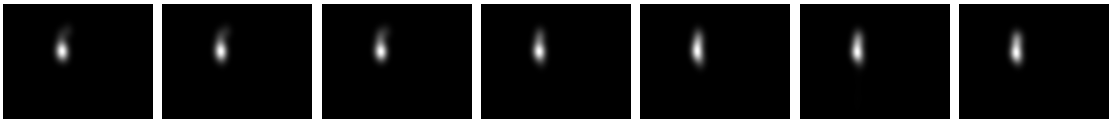
In this chapter, we introduced a saliency detection framework for 2D video sequences using color, texture, and motion as bottom-up attributes. The color and texture saliency maps are obtained as described in Chapter 4, and motion maps are computed using the optical flow field. All feature maps are weighted based on our eye-tracking based experimental results and then combined using a Naive Bayesian Network. Furthermore, the human visual acuity factor was exploited to enhance the saliency map. Our results indicate that this model is reliable for both static and dynamic scenes, and it is capable of extracting a reliable gradual saliency map even for the moving and cluttered backgrounds.

Table 5.3: Saliency Ranking for Different Objects of a Sample Video Clip from SAVAM Dataset Resulting in our Eye-tracking Experiment.

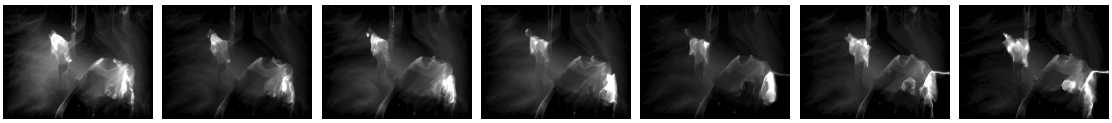
Ranking	Object	Fixation # (out of 14)	Fixation % (divided by 14)
Part 1: Man Standing			
1	Man's face	12	85.72
2	Umbrella	12	85.72
3	Leaves	3	21.43
Part 2: Opening Umbrella			
1	Umbrella	13	92.86
2	Man's face	12	85.72
3	Leaves	2 (out of 4)	14.29
Part 3: Rotating Umbrella			
1	Umbrella	14	100
2	Leaves	7 (out of 8)	50
Part 4: Umbrella on the Top			
1	Man's face	12	85.72
2	Umbrella	8	57.14
3	Leaves	6 (out of 8)	42.86



a) The original video frame sequence.



b) The ground-truth frame sequence.



c) The motion saliency map of our model.

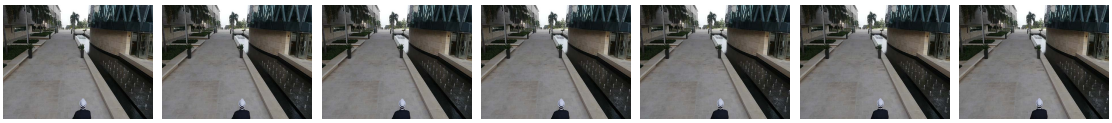
d) The full saliency map of our model.

e) The saliency map of Wang et al. [89] method.

f) The saliency map of Rahtu et al. [71] method.

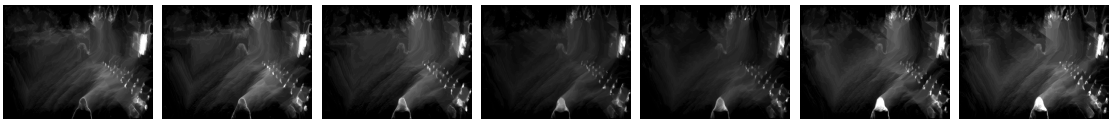
g) The saliency map of Itti et al. [55] method.

Figure 5.2: Visual comparison of saliency estimation from different models over EyeTrackUAV dataset (HD-video) with a dynamic background: (a) input frame; (b) ground truth map; (c) Our motion saliency map; (d) Our final saliency map; (e) Consistent video saliency using local gradient flow and global refinement (LFGR) by Wang et al. [89]; (f) Segmentation salient object method (SEG) by Rahtu et al. [71]; (g) Itti's model [55].



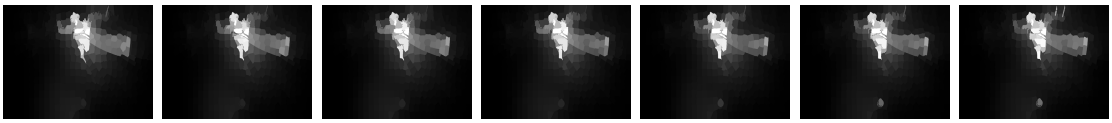
a) The original video frame sequence.

b) The ground-truth frame sequence.



c) The motion saliency map of our model.

d) The full saliency map of our model.



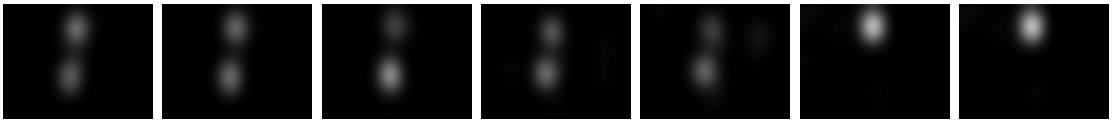
e) The saliency map of Wang et al. [89] method.

f) The saliency map of Rahtu et al. [71] method.

g) The saliency map of Itti et al. [55] method.

Figure 5.3: Visual comparison of saliency estimation from different models over EyeTrackUAV dataset (HD-video) with a dynamic background: (a) input frame; (b) ground truth map; (c) Our motion saliency map; (d) Our final saliency map; (e) Consistent video saliency using local gradient flow and global refinement (LFGR) by Wang et al. [89]; (f) Segmentation salient object method (SEG) by Rahtu et al. [71]; (g) Itti's model [55].

a) The original video frame sequence.



b) The ground-truth frame sequence.

c) The frames with eye-tracking fixation points over all participants.

d) The full saliency map of our model.

e) The saliency map of Wang et al. [89] method.

f) The saliency map of Rahtu et al. [71] method.

g) The saliency map of Itti et al. [55] method.

Figure 5.4: Visual comparison of saliency estimation from different models over SAVAM dataset (HD-video) with a static background: (a) input frame; (b) ground truth map; (c) Our motion saliency map; (d) Our final saliency map; (e) Consistent video saliency using local gradient flow and global refinement (LFGR) by Wang et al. [89]; (f) Segmentation salient object method (SEG) by Rahtu et al. [71]; (g) Itti's model [55].

Figure 5.5: Visual representation of validation for gradual saliency estimation over sample frames of SAVAM dataset: (a) input frame; (b) ground-truth map; (c) our saliency map; (d) fixation dots over all the participants; (e) fixation path of participant #1; (f) fixation path of participant #2; (g) input frame; (h) ground-truth map; (i) our saliency map; (j) fixation dots over all the participants; (k) fixation path of participant #1; (l) fixation path of participant #2.

Chapter 6

Conclusion and Discussion

6.1 Conclusions

In this dissertation, we proposed a biologically-inspired saliency detection model indicating a reduction in the huge volume of visual information into the most important and informative part. This can be used for many applications such as image/video summarization, image/video retrieval, video streaming traffic control, and video compression. First, we designed an experimental study without cognitive bias to formulate the saliency within a scene. For this purpose, we investigated bottom-up features such as color, texture, motion direction/speed, and color contrast as stimuli for the HVS. We introduced this experiment to understand the order of importance among bottom-up attributes in absorbing human attention while observing a scene. We tested each individual stimulus as well as their different combinations to be able to provide a comprehensive search and understanding on which colors, textures, and motions can attract human attention more.

According to our results, warm colors such as red and pink, and bright colors such as yellow and cyan are more salient. Textures with dense and compact edges are more outstanding. In addition, textures with obvious high contrast within their pattern are most likely to be salient. Vertical movements are more fixated on and the motions with high (not very high) or very low speed or any unique change in the direction/speed are more attractive for human subjects.

We concluded that color contrast is as important as the color and texture stimuli. Our test for contrast illustrated high consistency among participants.

In terms of combined stimuli, we found that texture is an important feature in determining the salient area in a scene opposing previous work in literature. Especially,

if we have a cognitive bias then texture might be more important to guide human attention. In fact, color and texture features can be as important as motion feature in dynamic scenes. Our results also confirmed the center bias phenomenon.

It was noticed that the more attractive features accumulated in an area/object, the higher the chance that the area will stand out as a salient region. For example, if an object/area possess red color, textures with dense edges, and vertical movement, it will be one of the most salient regions in a frame.

Next, we introduced a saliency detection framework for both static and dynamic scenes using color, luminance, texture, intensity, and motion features. This algorithm prioritizes the corresponding colors, texture patterns, and motion directions resulting from our eye-tracking based study. To extract color saliency, a k-nearest neighbor search technique based on a k-d tree search was applied to assign a ranking system to different colors according to the resulting ranking system from the recorded eye fixations. Moreover, feature contrast in *CIE Lab* color space was used to extract the color contrast.

To find the salient textured regions, the densest edges with distinctive orientations were extracted using the contrast of Gabor energy features. An intensity variation map was added to the texture map to support the results of our subjective study.

Motion maps were created based on optical flow fields and energy optimization. We incorporated both local and global algorithms to compute the optical flow field and we improved the flow computation by combining smoothness, gradient constraint, and symmetric matching to the main data term in the optical flow energy information. Then, we computed the motion velocity and acceleration maps.

All resulting feature maps fused into an ultimate saliency map using a Naive Bayesian network. A human visual acuity factor was employed to enhance the saliency map. The Bayesian network provides the ability to assign the gradual saliency for the entire of an image which is the main goal of this dissertation. Another advantage of this model is its ability to segment full salient areas with well-defined boundaries and high accuracy.

Lastly, we performed a second validation stage for our resulting saliency maps in order to evaluate the gradual saliency since the available ground-truth datasets have limitations. For this purpose, we performed another eye-tracking test using human subjects and recorded their eye movements while watching the publicly available benchmark datasets from the real world. Therefore, we could estimate the gradual

saliency of the objects within a scene by analysing the recorded eye movements to compare with our results.

6.2 Future Work

As we have described in this dissertation, we simulated our saliency detection model and compared it with the related existing works in the state-of-the-art. However, we noticed that the visual saliency detection research field suffers from two main shortcomings like the lack of appropriate ground-truth data and metrics.

Ground-truth Dataset

A variety of visual attention models has been proposed in the saliency detection area and there are many image and video datasets publicly available to test those methods. However, there is a lack of appropriate and meaningful ground-truth dataset to evaluate the existing models.

Currently, the performance of the VAMs is compared with the datasets captured using eye-tracking procedure and human subjects. These available datasets provide very little information about the salient areas as well as they might change from a group of participants to another. Therefore, they cannot be considered as a reliable source to be called ground-truth datasets.

In addition, datasets created using eye-tracking, are usually helpful for the fixation prediction methods but not for the segmentation oriented algorithms. In other words, if a saliency detection model tries to segment the entire area of a salient object then we will encounter difficulties to compare it with the ground-truth datasets created by the eye-tracker devices.

Evaluation Metrics

Recently, many studies has been done to design different metric to evaluate saliency detection models and compare them with each other. However, we still do not have a reliable and proper metric to compare the performance of the existing models toward each other fairly. Since, most of the existing metrics first perform a pre-processing stage on the resulting saliency maps and convert them into binary maps therefore, some parts of the maps are eliminated. In this way, the accuracy of the produced maps are reduced and we believe this cannot lead us to a very accurate comparison.

Furthermore, our proposed saliency detection algorithm provides a gradual saliency map which is difficult to be compared using binary masking based metrics. We used different thresholds to create different masks for each map but the metric issue is still left as an open problem.

To tackle these barriers, we recommend designing new ways for ground-truth data capturing/creating as well as new metrics that can be tuned for different levels of saliency.

List of References

[1] A. Borji, and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185-207, 2013.

[2] Y. Wo, X. Chen, G. Han, "A Saliency Detection Model Using Aggregation Degree of Color and Texture," *Signal Processing: Image Communication*, vol. 30, pp. 121-136, 2015.

[3] Cisco.com, "Cisco Visual Networking Index: Forecast and Methodology, 2017-2022 White Paper," 2019. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf. [Accessed: 1- Apr- 2019].

[4] Cisco.com, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017-2022 White Paper," 2019. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf. [Accessed: 1- Apr- 2019].

[5] S. Taksande, K. Joshi, V. Chikaraddi, S. Raksha, "Video Streaming Techniques and Issues," *International Journal of Advanced Research in Computer Science and Technology*, vol. 3, no. 1, pp. 141-144, 2015.

[6] Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos1, "Review of Error Resilient Coding Techniques for Real-Time Video Communications," *IEEE Signal Proc. Magazine*, vol. 17, pp. 61-82, 2000.

[7] P. Pinol, M. Martinez-Rach , P. Garrido , O. Lopez-Granado, and M. P. Malumbres, "Error Resilient Coding Techniques for Video Delivery over Vehicular Networks," *Sensors*, vol. 18, no. 10, pp. 1-25, 2018.

[8] I. Hwang, E. Kim, H. J. Kwon, Y. Kim, C. G. Kim, S. K. Lim, W. C. Park, "A Hardware-Oriented Fast Foveated Rendering Algorithm for Mobile Real-Time Ray Tracing," *International Conference on Computational Science and Computational Intelligence(CSCI)*, pp.466-468, 2018.

[9] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT Saliency Benchmark," 2012. Available: http://saliency.mit.edu/. [Accessed: 10- Jan- 2015].

[10] W. Osberger, Perceptual Vision Models for Picture Quality Assessment and Compression Applications, Ph.D. thesis, Queensland University of Technology, Australia, 1999.

[11] C. G. Healey, and J. T. Enns, "Attention and Visual Memory in Visualization and Computer Graphics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 7, pp. 1170-1188, 2012.

[12] E. D. Gelasca, D. Tomasic, and T. Ebrahimi, "Which Colors Best Catch Your Eyes. A Subjective Study of Color Saliency," *SPIE International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pp. 1-6, 2005.

[13] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 438-445, 2012.

[14] K. Duncan, and S. Sarkar, "Saliency in Images and Videos: a Brief Survey," *IET Computer Vision*, vol. 6, no. 6, pp. 514-523, 2012.

[15] A. Banitalebi-Dehkordi, E. Nasiopoulos, M. T. Pourazad, and P. Nasiopoulos, "Benchmark Three-dimensional Eye Tracking Dataset for Visual Saliency Prediction on Stereoscopic Three-dimensional Video," *SPIE Journal of Electronic Imaging*, vol. 25, no. 1, pp. 1-20, 2016.

[16] A. Borji, M. M. Cheng, Q. Hou, H. Jiang and J. Li, "Salient Object Detection: A Survey," *Journal of Computational Visual Media*, vol. 5, no. 2, pp. 117-150, 2019.

[17] O. Le Meur, "Methods for comparing scan-paths and saliency maps: strengths and weaknesses," *Behavior Research Methods*, vol. 45, no. 1, pp. 251-266, 2013.

[18] J. Hosseinkhani, H. Soltanian-Zadeh, M. Kamarei, "Ball Tracking in Soccer Video Using Kalman Filter," *IEEE Joint International Conference on Computer Science and Information Technology Conference*, pp. 596-599, 2011.

[19] A. M. Treisman, and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97-136, 1980.

[20] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic bulletin and review*, vol. 1, no. 2, pp. 202-238, 1994.

[21] U. Neisser, Cognitive psychology, New York: Appleton, Century, Crofts, 1967.

[22] B. Julesz, "A brief outline of the texton theory of human vision," *Trends in Neuroscience*, vol. 7, no. 2, pp. 41-45, 1984.

[23] E. N. Dzhafarov, R. Sekuler, and J. Allik, "Detection of changes in speed and direction of motion reaction time analysis," *Perception and Psychophysics*, vol. 54, no. 6, pp. 733-750, 1993.

[24] D. Khaustova, J. Fournier, E. Wyckens, "An Investigation of Visual Selection Priority of Objects with Texture and Crossed and Uncrossed Disparities," *SPIE Human Vision and Electronic Imaging*, vol. 90, no. 14, pp. 1-13, 2014.

[25] J. Hakkinen, T. Kawaid, J. Takataloc, R. Mitsuyad, and G. Nymanc, "What Do People Look at When They Watch Stereoscopic Movies?," *SPIE Stereoscopic Displays and Applications*, vol. 75, no. 24, pp. 1-11, 2010.

[26] D. Khaustova, J. Fournier, and E. Wyckens, "How Visual Attention is Modified by Disparities and Textures Changes?," *SPIE HVEI*, vol. 86, no. 51, pp. 1-15, 2013.

[27] P. Correia, and F. Pereira, "Video Object Relevance Metrics for Overall Segmentation Quality Evaluation," *EUSIPCO Conference on Estimation of Video Objects Relevance*, pp. 1-11, 2000.

[28] W. Osberger, and A. M. Rohaly, "Automatic Detection of Regions of Interest in Complex Video Sequences," *in Proceedings of Human Vision and Electronic Imaging*, vol. 6, pp. 361-372, 2001.

[29] F. Birren, Le Pouvoir de la Couleur. Les Editions de Homme., 1998.

[30] J. Hosseinkhani, H. Soltanian-Zadeh, M. Kamarei, O. Staadt, "Ball detection with the aim of corner event detection in soccer video," *IEEE International Symposium on Parallel and Distributed Processing with Applications Workshops (ISPAW)*, pp. 147-152, 2011.

[31] K. F. Van Orden, J. Divita, and M. J. Shim, "Redundant Use of Luminance and Flashing with Shape and Color as Highlighting Codes in Symbolic Displays," *Human Factors*, vol. 35, no. 2, pp. 195-204, 1993.

[32] W. Osbergera, A. Maederb, and N. Bergmanna, "A Technique for Image Quality Assessment Based on a Human Visual System Model," *IEEE Signal Processing Conference*, PP. 1-4, 1998.

[33] N. H. Mackworth, and A. J. Morandi, "The Gaze Selects Informative Details Within Pictures," *Perception and Psychophysics*, vol. 2, pp. 547-552, 1967.

[34] A. L. Yarbus, Eye Movements and Vision. New York: Springer Plenum Press., 1967.

[35] M. Dick, S. Ullman, and D. Sagi, "Parallel and Serial Processes in Motion Detection," *Science.*, vol. 237, pp. 400-402, 1987.

[36] R. B. Ivry, "Asymmetry in Visual Search for Targets Defined by Differences in Movement Speed," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, no. 4, pp. 1045-1057, 1992.

[37] J. M. Henderson, "Regarding scenes," *Current Directions in Psychological Science*, vol. 16, no. 4, pp. 219-222, 2007.

[38] J. M. Wolfe, and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?," *Nature reviews. Neuroscience*, vol. 5, pp. 495-501, 2004.

[39] J. M. Henderson, "Visual saliency does not account for eye movements during visual search in real-world scenes," *Elsevier: Eye movements*, pp. 537-562, 2007.

[40] G. Krieger, "Object and scene analysis by saccadic eye movements: an investigation with higher-order statistics," *Spatial Vision*, vol. 13, no. 2, pp. 201-214, 2000.

[41] S. K. Mannan, KH. Ruddock, and D. S. Wooding, "Automatic control of saccadic eye movements made in visual inspection of briefly presented 2D images," *Spatial Vision*, vol. 9, no. 3, pp. 363-386, 1995.

[42] S. K. Mannan, KH. Ruddock, and D. S. Wooding, "The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images," *Spatial Vision*, vol. 10, no. 3, pp. 165-188, 1996.

[43] S. K. Mannan, KH. Ruddock, and D. S. Wooding, "Fixation sequences made during visual examination of briefly presented 2D images," *Spatial Vision*, vol. 11, no. 2, pp. 157-178, 1997.

[44] D. J. Parkhurst, and E. Niebur, "Scene content selected by active vision," *Spatial Vision*, vol. 6, no. 2, pp. 125-154, 2003.

[45] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643-659, 2005.

[46] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive Computation*, vol. 3, no. 1, pp. 5-24, 2011.

[47] D. J. Berg, S. E. Boehnke, R. A. Marino, D. P. Munoz, and L. Itti, "Free viewing of dynamic stimuli by humans and monkeys," *Journal of Vision*, vol. 9, no. 5, pp. 1-15, 2009.

[48] R. Carmi, and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," *Vision Research*, vol. 46, no. 26, pp. 4333-4345, 2006.

[49] R. Carmi, and L. Itti, "The role of memory in guiding attention during natural vision," *Journal of Vision*, vol. 6, no. 9, pp. 898-914, 2006.

[50] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Vision Cognition*, vol. 12, no. 6, pp. 1093-1123, 2005.

[51] L. Itti, "Quantitative modelling of perceptual salience at human eye position," *Vision Cognition*, vol. 14, no. 4, pp. 959-984, 2006.

[52] O. Le Meur,P. Le Callet, and D. Barba, "Predicting Visual Fixations on Video Based on Low-Level Visual Features," *Vision Research*, vol.47, no.19, pp. 2483-2498, 2007.

[53] BM. t'Hart, J. Vockeroth, F. Schumann, K. Bartl, E. Schneider, P. Konig, and W. Einhauser, "Gaze allocation in natural stimuli: comparing free exploration to head-fixed viewing conditions," *Vision Cognition*, vol. 17, no. 6, pp. 1132-1158, 2009.

[54] C. Vazquez, A. Mitiche, R. Laganiere, "Joint Multi-region Segmentation and Parametric Estimation of Image Motion by Basis Function Representation and Level set Evolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 782-793, 2006.

[55] L. Itti, C. Koaggch, and E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[56] C. Koch, and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Matters of intelligence. Springer*, pp. 115-141, 1987.

[57] R. Rosenholtz, "A simple saliency model predicts a number of motion pop-out phenomena," *Vision research*, vol. 39, no. 19, pp. 3157-3163, 1999.

[58] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition*, vol. 12, no. 6, pp. 1093-1123, 2005.

[59] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *In Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*, pp. 64-79, 2003.

[60] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Gue rin-Dugue, "Modeling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos," *Int'l J. Computer Vision*, vol. 82, pp. 231-243, 2009.

[61] V. Navalpakkam, and L. Itti, "Modeling the Influence of Task on Attention," *Vision Research*, vol. 45, no. 2, pp. 205-231, 2005.

[62] G. Kootstra, A. Nederveen, and B. de Boer, "Paying Attention to Symmetry," *Proc. British Machine Vision Conf.*, pp. 1115-1125, 2008.

[63] R. Achanta, and S. Susstrunk, "Saliency detection using maximum symmetric surround," *IEEE Conference on Image Processing (ICIP)*, pp. 2653-2656, 2010.

[64] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 733-740, 2012.

[65] L. Elazary, and L. Itti, "A Bayesian Model for Efficient Visual Search and Recognition," *Vision Research*, vol. 50, pp. 1338-1352, 2010.

[66] A. Torralba, "Modeling Global Scene Factors in Attention," *Journal of Optical Society of America*, vol. 20, no. 7, pp. 1407-1418, 2003.

[67] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," *Journal of Vision*, vol. 8, no. 32, pp. 1-20, 2008.

[68] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-Down Control of Visual Attention in Object Detection," *IEEE Conference on Image Processing (ICIP)*, pp. 253-256, 2003.

[69] L. Zhang, M. H. Tong, and G. W. Cottrell, "SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes," *Proc. Cognitive Science Conf.*, pp. 2944-2949, 2009.

[70] S. Wang, Zh. Zhou, H. Qu, and B. Li, "RGB-D Saliency Detection under Bayesian Framework," *International Conference on Pattern Recognition (ICPR)*, pp. 1881-1886, 2016.

[71] E. Rahtu, J. Kannala, S. Mikko, and J. Heikkil, "Segmenting Salient Objects from Images and Videos," *European Conference on Computer Vision (ECCV)*, pp. 366-379, 2010.

[72] N. D. B. Bruce, and J. K. Tsotsos, "Saliency Based on Information Maximization," *Proc. Advances in Neural Information Processing Systems*, pp. 155-162, 2005.

[73] R. Rosenholtz, A. L. Nagy, and N. R. Bell, "The Effect of Background Color on Asymmetries in Color Search," *J. Vision*, vol. 4, no. 3, pp. 224-240, 2004.

[74] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Decorrelation and Distinctiveness Provide with Human-Like Saliency," *Proc. Advanced Concepts for Intelligent Vision Systems*, pp. 343-354, 2009.

[75] Y. Li, Y. Zhou, J. Yan, and J. Yang, "Visual Saliency Based on Conditional Entropy," *Proc. Ninth Asian Conf. Computer Vision*, pp. 246-257, 2009.

[76] M. Mancas, Computational Attention: Modelisation and Application to Audio and Image Processing, PhD thesis, 2007.

[77] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating Human Saccadic Scanpaths on Natural Images," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 441-448, 2011.

[78] H. J. Seo, and P. Milanfar, "Static and Space-Time Visual Saliency Detection by Self-Resemblance," *J. Vision*, vol. 9, no. 12, pp. 1-27, 2009.

[79] X. Hou, and L. Zhang, "Dynamic Visual Attention: Searching for Coding Length Increments," *Proc. Advances in Neural Information Processing Systems*, pp. 681-688, 2008.

[80] V. Leboran, A. Garcia-Diaz, XR. Fdez-Vidal, and XM. Pardo, "Dynamic whitening saliency," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, no. 5, pp. 893-907, 2017.

[81] J. Yan, J. Liu, Y. Li, Z. Niu, and Y. Liu, "Visual saliency detection via rank-sparsity decomposition," *IEEE Conference on Image Processing (ICIP)*, pp. 1089-1092, 2010.

[82] L. Zhu, D. A. Klein, S. Frintrop, Z. Cao, and A. B. Cremers, "A Multisize Superpixel Approach for Salient Object Detection based on Multivariate Normal Distribution Estimation," *Image Processing, IEEE Transactions on*, vol. 23, no. 12, pp. 5094-5107, 2014.

[83] A. Salah, E. Alpaydin, and L. Akrun, "A Selective Attention-Based Method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 420-425, 2002.

[84] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *Neural Information Processing Systems*, pp. 545-552, 2006.

[85] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino, "A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network," *IEEE Conf. Multimedia and Expo*, pp. 1073-1076, 2008.

[86] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and Where: A Bayesian Inference Theory of Visual Attention," *Vision Research*, vol. 55, pp. 2233-2247, 2010.

[87] Sh. Wang, Sh. Yang, Zh. Liu, and L. Jiao, "Saliency generation from complex scene via digraph and Bayesian inference," *Neurocomputing*, vol. 170, pp. 176-186, 2015.

[88] M. M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu , "Global contrast based salient region detection," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 409-416, 2011.

[89] W. Wang, J. Shen, and L. Shao, "Consistent Video Saliency Using Local Gradient Flow Optimization and Global Refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185-4196, 2015.

[90] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500-513, 2010.

[91] X. Hou, and L. Zhang, "Saliency Detection: A Spectral Residual Approach," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[92] C. Guo, Q. Ma, and L. Zhang, "Spatio-Temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2008.

[93] C. Guo, and L. Zhang, "A Novel Multi-resolution Spatio-temporal Saliency Detection Model and Its Applications in Image and Video Compression," *IEEE Trans. Image Processing*, vol. 19, no. 1, pp. 185-198, 2010.

[94] S. J. Sangwine, "Fourier transforms of colour images using quaternion, or hypercomplex, numbers," *Electronics Letters*, vol. 32, no. 21, pp. 1979-1980, 1996.

[95] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Susstrunk, "Frequency-Tuned Salient Region Detection," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1597-1604, 2009.

[96] P. Bian, and L. Zhang, "Biological Plausibility of Spectral Domain Approach for Spatio-temporal Visual Saliency," *Conf. Advances in Neuro-Information Processing*, pp. 251-258, 2009.

[97] J. Li, and M. D. Levine, "Visual Saliency Based on Scale-Space Analysis in the Frequency Domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 996-1010, 2013.

[98] S. Fang, J. Li, Y. Tian, T. Huang, and X. Chen, "Learning Discriminative Subspaces on Random Contrasts for Image Saliency Analysis," *IEEE Trans. Neural Networks and Learning Systems*, vol. 28, no. 5, pp. 1095-1107, 2016.

[99] S. G. Lisberger, "Visual guidance of smooth pursuit eye movements: sensation, action, and what happens in between," *in PMC: Neuron.*, vol. 66, no. 4, pp. 477-491, 2010.

[100] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic Multi-Task Learning for Visual Saliency Estimation in Video," *J. Computer Vision*, vol. 90, no. 2, pp. 150-165, 2010.

[101] W. Kienzle, M. O. Franz, B. Scholkopf, and F. A. Wichmann, "Center-Surround Patterns Emerge as Optimal Predictors for Human Saccade Targets," *J. Vision*, vol. 9, no. 5, pp. 1-15, 2009.

[102] R. J. Peters and L. Itti, "Beyond Bottom-up: Incorporating Task dependent Influences into a Computational Model of Spatial Attention," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[103] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," *IEEE Conf. Computer Vision*, pp. 2106-2113, 2009.

[104] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modeling Search for People in 900 Scenes: A Combined Source Model of Eye Guidance," *Visual Cognition*, vol. 17, no. 6, pp. 945-978, 2009.

[105] H. Ramadan, H. Tairi, "Pattern mining-based video saliency detection: Application to moving object segmentation," *Journal of Computers and Electrical Engineering*, Vol. 1, No. 13, pp. 567-579, 2018.

[106] H. Fu, X. Cao, and Z. Tu, "Cluster-Based Co-Saliency Detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766-3778, 2013.

[107] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A fully convolutional neural network for predicting human eye fixations," *IEEE Trans. Image Processing*, vol. 26, no. 9, pp. 4446-4456, 2015.

[108] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," *IEEE Conf. Computer Vision*, pp. 262-270, 2015.

[109] W. Wang, J. Shen, F. Guo, M. M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4894-4903, 2018.

[110] G. Li, Y. Yu, "Visual Saliency Based on Multi-scale Deep Features," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-13, 2015.

[111] L. Zhu, H. Ling, J. Wu, H. Deng, and J. Liu, "Saliency Pattern Detection by Ranking Structured Trees," *International Conference on Computer Vision (ICCV)*, pp. 5468-5477, 2017.

[112] S. Goferman, Z. Lihi, T. Ayellet, "Context-Aware Saliency Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915-1926, 2012.

[113] Y. Zhai, M. Shah, "Visual Attention Detection in Video Sequences Using Spatio Temporal Cues," *ACM Multimedia*, pp. 815-824, 2006.

[114] M. Mancas, N. Riche, B.G. Leroy, "Abnormal Motion Selection in Crowds Using Bottom-up Saliency," *IEEE Conference on Image Processing (ICIP)*, pp. 229-232, 2011.

[115] R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard, "Eye Movements in Iconic Visual Search," *Vision Research*, vol. 42, pp. 1447-1463, 2002.

[116] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A Generic Framework of User Attention Model and Its Application in Video Summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907-919, 2005.

[117] P. L. Rosin, "A Simple Method for Detecting Salient Regions," *Pattern Recognition*, vol. 42, no. 11, pp. 2363-2371, 2009.

[118] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video Saliency Detection via Spatial-Temporal Fusion and Low-Rank Coherency Diffusion," *IEEE Trans. Image Processing*, vol. 26, no. 7, pp. 3156-3170, 2017.

[119] R. Nikitha, R. Vedhapriyavadhana, and V.P. Anubala, "Video Saliency Detection using Weight Based Spatio-Temporal Features," *International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 343-347, 2018.

[120] L. Itti, and P. Baldi, "Bayesian Surprise Attracts Human Attention," *Proc. Advances in Neural Information Processing Systems*, pp. 1295-1306, 2005.

[121] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740-757, 2019.

[122] Color-blindness.com, "Ishihara 38 Plates CVD Test," 2017. Available: http://www.color-blindness.com/ishihara-38-plates-cvd-test/. [Accessed: 3-Nov- 2017].

[123] Recommendation ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays," ITU, 2012.

[124] D. T. Lindsey, and A. M. Brown, "Universality of color names," *Proceedings of the National Academy of Sciences*, vol. 103, no. 44, pp. 16608-16613, 2006.

[125] M. Baik, H. J. Suk, J. Lee, and K. A. Choi, "Investigation of eye-catching colors using eye tracking," *SPIE Electronic Imaging*, vol. 8651, pp. 1-6, 2013.

[126] M. Tian, S. Wan, and L. Yue, "A Color saliency model for salient objects detection in natural scenes," Advances in Multimedia Modelling, Lecture Notes in Computer Science, vol. 5916, pp. 240-250, 2010.

[127] T. Priyanka, N. Narasimha, "KNN Based Document Classifier Using K-d Tree: An Efficient Implementation," *Journal of Computer Science and Communication Networks*, vol. 5, no. 5, pp. 270-274, 2013.

[128] A. Banitalebi-Dehkordi, M. T. Pourazad, and P. Nasiopoulos, "A Learning-Based Visual Saliency Prediction Model for Stereoscopic 3D Video," *Multimedia Tools and Applications*, vol. 76, no. 22, pp. 23859-23890, 2017.

[129] J. B. Jonas, U. Schneider, and G. O. H. Naumann, "Count and density of human retinal Photoreceptors," *Graefe's Archive Ophthalmology*, vol. 230, pp. 505-510, 1992.

[130] A. Banitalebi-Dehkordi, 3D Video Quality Assessment, PhD thesis, University of British Colombia (UBC), Canada, 2015.

[131] I. Kokkinos, G. Evangelopo, "Texture Analysis and Segmentation Using Modulation Features, Generative Models, and Weighted Curve Evolution," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 142-157, 2009.

[132] J. Hosseinkhani, and C. Joslin, "Significance of Bottom-up Attributes in Video Saliency Detection Without Cognitive Bias," *IEEE Conf. on Cognitive Informatics and Cognitive Computing*, pp. 606-613, 2018.

[133] J. Hosseinkhani, C. Joslin, "Investigating into Saliency Priority of Bottom-up Attributes in 2D Videos Without Cognitive Bias," *IEEE Symposium on Signal Processing and Information Technology*, pp. 223-228, 2018.

[134] J. Hosseinkhani, C. Joslin, "Saliency Priority of Individual Bottom-up Attributes in Designing Visual Attention Models," *Journal of Software Science and Computational Intelligence*, Vol. 10, No. 4, pp. 1-18, 2018.

[135] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency Detection for Stereoscopic Images," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2625-2636, 2014.

[136] T. Judd, K. Ehinger, F. Durand, A. Torralba, "Learning to predict where humans look," *IEEE conference on Computer Vision*, pp. 2106-2113, 2009.

[137] W. S. Geisler and J. S. Perry, "A real-time foveated multi-resolution system for low-bandwidth video communication," *SPIE, Human Vision and Electronic Imaging III*, vol. 3299, no. 294, pp. 1-13, 1998.

[138] J. P. Harris, and M. Fahle, "Differences Between Fovea and Periphery in the Detection and Discrimination of Spatial Offsets," *Vision Research*, vol. 36, no. 21, pp. 3469-3477, 1996.

[139] Q. Yan, L. Xu, J. Shi, J. Jia, "Extended Complex Scene Saliency Dataset (ECSSD)," 2013. Available: http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/dataset.html.

[140] Q. Yan, L. Xu, J. Shi, J. Jia, "Hierarchical Saliency Detection," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1155-1162, 2013.

[141] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H. Y. Shum, "Learning to Detect A Salient Object," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2007.

[142] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," MIT computer science and artificial intelligence Lab tech. report, 2012.

[143] N. Riche, M. Duvinage, M. Mancas, B. Gosselin and T. Dutoit, "Saliency and Human Fixations: State-of-the-art and Study of Comparison Metrics," *ICCV*, pp. 1153-1160, 2013.

[144] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397-2416, 2005.

[145] X. Li, Y. Li, C. Shen, A. Dick, A. Hengel, "Contextual Hypergraph Modeling for Salient Object Detection," *IEEE Conference on Computer Vision*, pp.3328-3335, 2013.

[146] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts," *ACM TOG*, vol. 23, no. 3, pp. 309-314, 2004.

[147] C. Liu, Beyond Pixels: Exploring New Representations and Applications for Motion Analysis, Doctoral Thesis, Massachusetts Institute of Technology, 2009.

[148] B. Lucas, and T. Kanade, "An iterative image registration technique with an application to stereo vision," *In Proc. Joint Conference on Artificial Intelligence*, pp. 674-679, 1981.

[149] B. Horn, and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1, pp. 185-203, 1981.

[150] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," *European Conference on Computer Vision (ECCV)*, pp. 25-36, 2004.

[151] A. Bruhn, and J. Weickert, "Lucas/Kanade meets Horn/Schunk: combining local and global optical flow methods," *J. Compute Vision*, vol. 61, no. 3, pp. 211-231, 2005.

[152] M. Tistarelli, "Multiple constraints for optical flow," *European Conference on Computer Vision (ECCV)*, pp. 61-70, 1994.

[153] S. Uras, F. Girosi, A. Verri, and V. Torre, "A computational approach to motion perception," *Biological Cybernetics*, vol. 60, pp. 79-87, 1988.

[154] M. Mueller, N. Smith, and B. Ghanem, "A Benchmark and Simulator for UAV Tracking," *European Conference on Computer Vision (ECCV)*, pp. 1-17, 2016.

[155] Y. Gitman, M. Erofeev, D. Vatolin, B. Andrey, F. Alexey, "Semi-automatic Visual Attention Modelling and its Application to Video Compression," *IEEE Conference on Image Processing (ICIP)*, pp. 1105-1109, 2014.

[156] J. Hosseinkhani, C. Joslin, "Saliency Priority Using Bottom-up Features for Static and Dynamic Scenes Without Cognitive Bias," *IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 189-192, 2019.

[157] J. Hosseinkhani, C. Joslin, "A Biologically Inspired Saliency Priority Extraction Using Bayesian Framework," *Journal of Multimedia Data Engineering and Management*, Vol. 10, No. 2, pp. 1-20, 2019.