A computational visual saliency model for images.

NARAYANASWAMY, M.

2021

A COMPUTATIONAL VISUAL SALIENCY MODEL FOR IMAGES

MANJULA NARAYANASWAMY

MRes 2021

A Computational Visual Saliency Model for Images

Manjula Narayanaswamy

A thesis submitted in partial fulfilment of the requirements of the
Robert Gordon University
for the degree of Master of Research

June 2021

Declaration

I hereby declare that the thesis comprises research work that has not been previously submitted for a degree at any other higher educational institution. The intellectual content of this thesis is the outcome of my own work. All the sources used in the presentation of the content have been acknowledged accordingly.

Manjula Narayanaswamy

Abstract

Human eyes receive an enormous amount of information from the visual world. It is highly difficult to simultaneously process this excessive information for the human brain. Hence the human visual system will selectively process the incoming information by attending only the relevant regions of interest in a scene. Visual saliency characterises some parts of a scene that appears to stand out from its neighbouring regions and attracts the human gaze. Modelling saliency-based visual attention has been an active research area in recent years. Saliency models have found vital importance in many areas of computer vision tasks such as image and video compression, object segmentation, target tracking, remote sensing and robotics. Many of these applications deal with high-resolution images and real-time videos and it is a challenge to process this excessive amount of information with limited computational resources. Employing saliency models in these applications will limit the processing of irrelevant information and further will improve their efficiency and performance. Therefore, a saliency model with good prediction accuracy and low computation time is highly essential. This thesis presents a low-computation wavelet-based visual saliency model designed to predict the regions of human eye fixations in images.

The proposed model uses two-channel information luminance (Y) and chrominance (Cr) in YCbCr colour space for saliency computation. These two channels are decomposed to their lowest resolution using two-dimensional Discrete Wavelet Transform (DWT) to extract the local contrast features at multiple scales. The extracted local contrast features are integrated at multiple levels using a two-dimensional entropy-based feature combination scheme to derive a combined map. The combined map is normalized and enhanced using natural logarithm transformation to derive a final saliency map. The performance of the model has been evaluated qualitatively and quantitatively using two large benchmark image datasets. The experimental results show that the proposed model has achieved better prediction accuracy both qualitatively and quantitatively with a significant reduction in computation time when compared to the existing benchmark models. It has achieved nearly 25% computational savings when compared to the benchmark model with the lowest computation time.

Index terms- visual saliency model, fixation prediction, discrete wavelet transform, image entropy.

Acknowledgements

I take this opportunity to thank everyone who helped me to successfully complete this research work. Sincere thanks to:

- My supervisor Dr. Yafan Zhao, for giving me the opportunity to carry out this research work. I express my gratitude for her dedicated support and guidance.
- Dr. Wai-keung Fung, for his thoughtful comments and recommendations. I have thoroughly enjoyed the discussions and wealth of knowledge that he has shared with me.
- Dr. Nazila Fough for feedback and support. I am thankful to her for the conference recommendations.
- The Graduate School, the School of Engineering and all its staff members who have supported me throughout this journey.
- My colleagues Yogitha, Maryam and Ruissein for their friendly discussions, help and support.

I thank my parents for giving me the financial and emotional support and my sisters for their encouragement. I cannot forget to thank my husband Dr. Jayachandra Chilukamari for his advice and unlimited project discussions and my beloved daughter Sonakshi for her motivating enthusiasm during these intensive academic years.

I dedicate this work to my parents

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem statement

Human eyes receive a rich stream of visual information every second (10^8 - 10^9 bits) [1]. It is highly difficult for the human brain to simultaneously perform the complex analysis of all the input visual information [2]. Therefore, the human visual system processes this enormous amount of information by selectively attending the relevant regions of interest in a scene. The selective attention mechanism is achieved through a sequence of saccadic eye movements called fixations [3]. Predicting the regions of human eye fixations is essential where these identified regions can be used in the intelligent processing of visual information in computer vision systems. Visual saliency models are computational models that are used to predict the regions of human eye fixations and these models have found vital importance in many areas such as image and video compression [4], image segmentation [5], remote sensing [6] and robotics [7]. In image and video compression techniques, saliency can be employed to represent the important regions with improved quality compared to the unimportant regions. It is also employed to facilitate the object segmentation tasks to separate the regions with objects, people, text and so on from the rest. Most of these saliency applications deal with high resolution images and videos. It is a challenge to process the excessive amount of visual data efficiently with limited computational resources. These applications require predicting human eye fixations accurately and efficiently. This will further improve the efficiency and computational performance of these applications by limiting the processing of irrelevant information. Therefore, a saliency model with good prediction accuracy and low computation time is highly essential.

The human visual attention mechanism is widely categorised into two types namely, bottom-up attention and top-down attention [1, 8]. Some features of a scene such as colour, luminance, motion and edges create a pop-out effect and grab our attention involuntarily. This automatic, stimulus-driven process is called bottom-up attention. Whereas, top-down attention is a goal oriented process which depends on the high-level factors such as scene context, task demand, past knowledge, user expectations and so on [1]. As an example of scene context, while walking in a street the fixations are confined to street by ignoring the sky region. Many studies have tried to simulate this

computational mechanism of visual attention [8-18]. The traditional bottom-up computational models were proposed by Itti *et al.* [19] and Harel *et al.* [14], based on three features - intensity, colour and orientation. The high-level fators such as scene context and center-bias were modelled by Oliva *et al.* [20] and Tatler *et al.* [21] respectively. The combined models were proposed by Goferman *et al.* [12], Zhang *et al.* [18] and Judd *et al.* [22]. Goferman *et al.* implemented scene context and centre-bias as top-down cues along with colour and contrast as bottom-up features. Zhang *et al.* used Shannon's self-information as bottom-up cue with prior information of a scene as top-down cue. Judd *et al.* incorporated top-down, bottom-up and mid-level cues to define saliency of an image. Hou *et al.* [15], Guo *et al.* [4] and Achanta *et al.* [9] proposed frequency models based on spectral characteristics of an image.

The existing computational models are either computationally expensive with good prediction accuracy [12, 14, 18, 19] or they are computationally efficient with low prediction accuracy [4, 9, 15] for them to be employed in practical applications. Hence there is a need for novel visual saliency model that can predict the salient regions with high accuracy and low computation time. The traditional computational models in frequency domain [4, 9, 13, 15] have employed Fourier transform (FT) to define image saliency. These models are good at estimating image saliency on a global context but are inadequate as they do not contribute to local saliency details. FT retrieves the global frequency content of a signal and is good at analysing stationary and pseudo-stationary signals [23, 24]. However, the limitations occur when it becomes necessary to locally analyse the various frequency components of a signal. An alternative solution to this limitation is to utilise the short-time FT that offers local frequency analysis of a signal. In this method, the signal is divided into several shorter space intervals and then the Fourier transform is applied to each interval which narrows down the frequency analysis with respect to the given time of a signal [23, 24]. However, the short-time FT offers local frequency analysis with constant resolution which depends on choosing the right size of spatial interval. Moreover, trying several spatial intervals on the image will increase the computational time for the application [16]. A better solution for local frequency analysis is to use Wavelet transform (WT) that offers multi-resolution analysis (MRA) [16, 23, 24]. Recent studies [16, 17, 25, 26] have exploited the MRA property of wavelets to define saliency of an image and have shown to provide improved prediction accuracy when compared to the traditional frequency models [4, 9, 15]. During MRA, the two-dimensional discrete WT (DWT) uses a set of filters (low-pass and high-pass) to decompose an image into independent frequency components which

provides localised frequency analysis at various resolutions. Therefore, the two-dimensional DWT is chosen in this research work to develop a saliency model that provides improvement in terms of prediction accuracy and computation time compared to the existing wavelet-based saliency detection models.

1.2 Research aim and objectives

The aim of this research work is to develop a bottom-up computational model of visual saliency to predict the regions of human eye fixations in images based on two-dimensional DWT. The developed model will predict the human eye fixations with improved prediction accuracy and reduced computational time when compared to the existing wavelet-based saliency detection models. It also has potential to improve the performance of visual saliency applications such as image and video compression, object segmentation and target tracking.

The aim of the project can be achieved through the following objectives:

1. **Literature review of the computational visual saliency models**

   The existing state-of-the-art computational visual saliency models are studied. The relevant benchmark models are selected and critically analysed for its advantages and disadvantages.

2. **Performance evaluation of the benchmark visual saliency models**

   The software and testing platform required for experimentation is set-up. The relevant image datasets and performance evaluation methods are identified. The performance of the benchmark methods is evaluated.

3. **Development of a bottom-up visual saliency model using DWT**

   This is the experimentation phase where the algorithm is developed and tested through iterative process. The developed algorithm will predict the regions of human eye fixations in images with improved prediction accuracy and reduced computation time when compared to the benchmark models.

4. **Performance evaluation of the proposed model**

   The performance of the proposed model is evaluated through qualitative and quantitative measurement techniques. The obtained experimental results are analysed qualitatively and quantitatively with respect to the benchmark models.

1.3 Project contributions

The key contribution of this research work is the development of a low-computation wavelet-based visual saliency model. The proposed model requires two-channel information for saliency computation when compared to the existing wavelet-based visual saliency models [16, 17, 25, 26], which require more than two colour channels. Moreover, an entropy-based feature combination scheme has been implemented to improve the prediction accuracy compared to the existing methods. The experimental results show that the proposed model has achieved significant reduction in computation time with better prediction accuracy compared to the benchmark visual saliency models [16, 17]. The proposed work has been published at the "27th IEEE International Conference on Electronics Circuits and Systems" [27].

1.4 Organisation of the Thesis

The organisation of this thesis is as follows.

- **Chapter 2 - Background**

  This chapter briefly describes the fundamentals of digital image and DWT. It also provides an overview of theoretical knowledge related to visual saliency in images, the literature of existing computational models and it discusses various saliency applications.

- **Chapter 3 - Experimental methodology**

  This chapter explains the research methodology used in conducting experiments and includes details of the software and test environments used in algorithm development. It also describes the procedure involved in the proposed model evaluation and provides the details of the image datasets, performance assessment methods and the benchmark models.

- **Chapter 4 - Low-computation wavelet-based visual saliency model**

  This chapter presents the development of the proposed bottom-up computational model of visual saliency to predict the human eye fixations in images based on DWT. The chapter provides summary of the existing wavelet-based saliency detection methods along with their advantages and disadvantages. Further, it gives an overview of the proposed model and provides the detailed explanation of the different stages in algorithm development.

- **Chapter 5 - Experimental results and analysis**

  The performance of the proposed model provided in chapter 4 is evaluated in terms of prediction accuracy and computation time using qualitative and quantitative measurement techniques. The experimental results pertaining to the model evaluation are analysed and discussed in this chapter.

- **Chapter 6 - Discussion and Conclusion**

  This chapter summarises the proposed research work, indicates the future directions and provides the thesis conclusion.

- **Appendix -** Contains information about the conference publication.

2 Background

2.1 Introduction

Understanding of how the human visual system processes the incoming information involves a high degree of complexity which has been extensively studied in recent years. Researchers have explored many approaches in modelling this process. Neurophysiologists have studied the responses of neurons during the attention process [28]. Psychologists have studied behavioural correlation of visual attention with human brain during the tasks such as change blindness [10] and attentional blink [1]. During these tasks, the viewer's experience is studied by varying the stimulus properties in a systematic setting [10]. Experiments were conducted to understand the human thought process by studying the eye movements to know what elements of a scene attracts the observer's eye [3]. These findings have been utilised by many researchers to develop novel computational systems to facilitate the computer vision tasks. This chapter presents the background knowledge related to modelling of visual saliency in images.

The chapter is divided into following sections. The fundamentals of digital image, colour spaces and DWT are discussed in sections 2.2, 2.3 and 2.4 respectively. A brief theory of human eye and image formation is discussed in section 2.5. The underlying concepts related to visual saliency are explained in section 2.6. Various approaches to computational modelling of saliency are reviewed in section 2.7. The applications of visual saliency models is discussed in 2.8. The summary of the chapter is provided in section 2.9.

2.2 Digital Image

A digital image is a two-dimensional signal with finite discrete elements called pixels. It may be defined using a function $f(x, y)$ where, x and y are spatial co-ordinates in a two-dimensional plane. The number of rows and columns indicates the respective height and width of an image. These dimensions in pixels indicates the resolution or size of an image. An image with N rows and M columns has a resolution of N × M pixels.

Grayscale image

In a monochrome or grayscale image, each pixel at a point (x, y) has a value called intensity or gray level which is represented using n bits. A n-bit grayscale image

comprises of intensity levels which ranges between 0 to $(2^n - 1)$. For an 8-bit image, the intensity level ranges between 0 (black pixel) to 255 (white pixel) with shades of gray in between.

Colour image

In a colour image each pixel is typically represented using three values. Combination of these values are used in displaying the image. Therefore, a colour image is represented using three two-dimensional matrices or colour channels in contrast to a grayscale image which is represented using single two-dimensional matrix. For an 8-bit colour image, a colour pixel is represented using $3 \times 8 = 24$ bits and the total number of possible colour combinations are: $2^{24} = 16777216$. These colour pixels are described using colour spaces.

2.3 Colour space

A colour model or colour space is used to provide the specification of colours within a standard co-ordinate system [29]. Various colour models are available to suit to the needs of various image processing tasks [30]. The RGB colour space is mostly used for image acquisition and display. The colour spaces, CMY and CMYK [29] are popularly used in colour printing. The YCbCr model [31] is used in image and video compression. The NTSC [29] and PAL [32] are used in television systems. The scope of this discussion is limited to two colour models RGB and YCbCr which are relevant to the proposed work (explained in chapter 4).

2.3.1 RGB colour space

The RGB colour space provides the specification of a colour using combination of three primary colours red (R), green (G) and blue (B). The model is defined based on the theory of trichromatic vision in humans [33]. According to the theory, the retina of human eye consists of three types of photoreceptors cells (cone cells) that are most sensitive to red, green and blue light. Due to the characteristics of the light absorption in the cone cells of human eye, any colour perceived, is seen as a combination of these primary colours [29].

2.3.2 YCbCr colour space

The bright-light vision of human eye makes it more sensitive to brightness (luminance) when compared to colour [34]. The YCbCr colour space can efficiently represent the colour information of an image when compared to RGB colour space as it separates brightness from colour information. In YCbCr colour space [31], the Y channel represents the luminance component and the Cb and Cr channels represent the chrominance components. The Y channel is obtained from weighted average of red, green and blue colour components of RGB colour space as provided in equation (2.1). The Cb and Cr channels are obtained from blue and red colour differences between the colour intensity and the mean luminance value as provided in equations (2.2) and (2.3) respectively.

$$Y = k_r R + k_g G + k_b B \tag{2.1}$$

$$C_b = B - Y \tag{2.2}$$

$$C_r = R - Y \tag{2.3}$$

where k_r, k_g and k_b are weighting factors with $k_r + k_g + k_b = 1$. According to ITU-R BT.601 conversion [31] the values of these weighting factors are: $k_r = 0.299$, $k_g = 0.587$ and $k_b = 0.114$.

2.4 Image processing in frequency domain

The Fourier Transform (FT) is the most widely used method for signal analysis. It is useful in analysis of stationary and pseudo-stationary signals and provides description of frequency information in a global context. The two-dimensional Discrete Fourier Transform (DFT) [35] applied to an image $f(x, y)$ with size N × N is given by:

$$F(u, v) = \frac{1}{N} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) e^{-j\left(\frac{2\pi}{N}\right)(ux+vy)} \tag{2.4}$$

where $F(u, v)$ is transformed image and $u, v = 0, 1, 2, \ldots, (N-1)$ are coefficients in frequency domain. The two-dimensional inverse DFT is given by:

$$f(x, y) = \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} F(u, v) e^{-j\left(\frac{2\pi}{N}\right)(ux+vy)} \tag{2.5}$$

where $x, y = 0, 1, 2, \ldots, (N-1)$. Equations (2.4) and (2.5) are computationally complex for large image sizes. A Fast Fourier Transform (FFT) replaces the DFT which is computationally more effective method. Another important transform is Discrete Cosine Transform (DCT) which has been widely employed in image and video compression [35]. The two-dimensional DCT [35] applied to an image $f(x, y)$ with size N × N is given by:

$$D(u,v) = \begin{cases} \frac{1}{N^2} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) & if \;\; u = 0 \;\; and \;\; v = 0 \\ \frac{2}{N^2} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) \times cos\left(\frac{(2x+1)u\pi}{2N}\right) \times cos\left(\frac{(2x+1)v\pi}{2N}\right) & otherwise \end{cases} \tag{2.6}$$

where $D(u, v)$ is a transformed image in DCT domain. The two-dimensional inverse DCT is given by:

$$f(x,y) = \frac{1}{N^2} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} D(u,v) \times cos\left(\frac{(2x+1)u\pi}{2N}\right) \times cos\left(\frac{(2x+1)v\pi}{2N}\right) \tag{2.7}$$

Like FFT, a fast version of DCT is available to improve the computational speed. In practice, FT is not an optimal technique for image coding as DCT can provide a higher compression rate with the same image quality with reduced storage and transmission requirements. Similar to DCT, Discrete Wavelet Transform (DWT) which is discussed in the next section, has been a popular technique employed in various image processing tasks such as de-noising, compression, feature detection [36] and so on.

2.4.1 Wavelet Transform (WT)

The real-time signals are typically non-stationary in nature. The limitations of FT occur when it is required to perform a localised time-frequency analysis of a signal. A continuous WT (CWT) uses a short waveform called 'wavelet' which allows to examine the signal at various frequencies and resolutions [24]. The transitions of a signal are better analysed by scaling (dilation) and shifting (translation) of the wavelet throughout the entire signal duration. This property of wavelet analysis is known as multi-resolution analysis (MRA). The wavelet consists of oscillations that exists for finite duration and has zero mean. An illustration of different types of wavelets is provided in Figure 2.1.

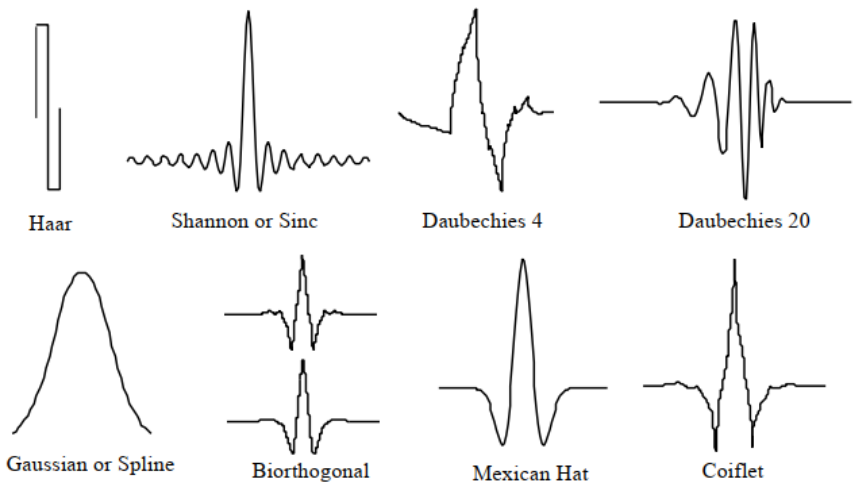

Figure 2.1 Illustration of types of wavelets [23].

2.4.2 Discrete wavelet transform (DWT)

The DWT uses a set of filters (low-pass and high-pass) in the process of MRA, which decomposes the signal into independent frequency components. During the process of MRA, the scaling operation is performed using a wavelet $\psi(t)$ in powers of 2, also called as dyadic scaling which is defined as:

$$\psi_{j,m}(t) = \frac{1}{\sqrt{2^j}}\ \psi\left(\frac{t - 2^j m}{2^j}\right) \tag{2.8}$$

where, j and m are scaling and shifting parameters respectively with both being integers. A two-dimensional DWT applied to an image *f(x, y)* at 2-levels is shown in Figure 2.2. At level-1, the rows of input signal are convolved with the separable filters and the resulting coefficients are down-sampled by a scale 2. The process is further repeated with the columns at level-2. The resulting filtered outputs *a(x, y)*, *h(x, y)*, *v(x, y)* and *d(x, y)* are called as approximation, horizontal, vertical and diagonal details respectively. The next level of details are derived by further decomposing the approximation details depending on the desired resolution. The reconstruction is an inverse operation to decomposition process which is achieved by applying inverse DWT (IDWT) to the required details. This involves up-sampling of details by a scale 2 and combining the filter outputs.

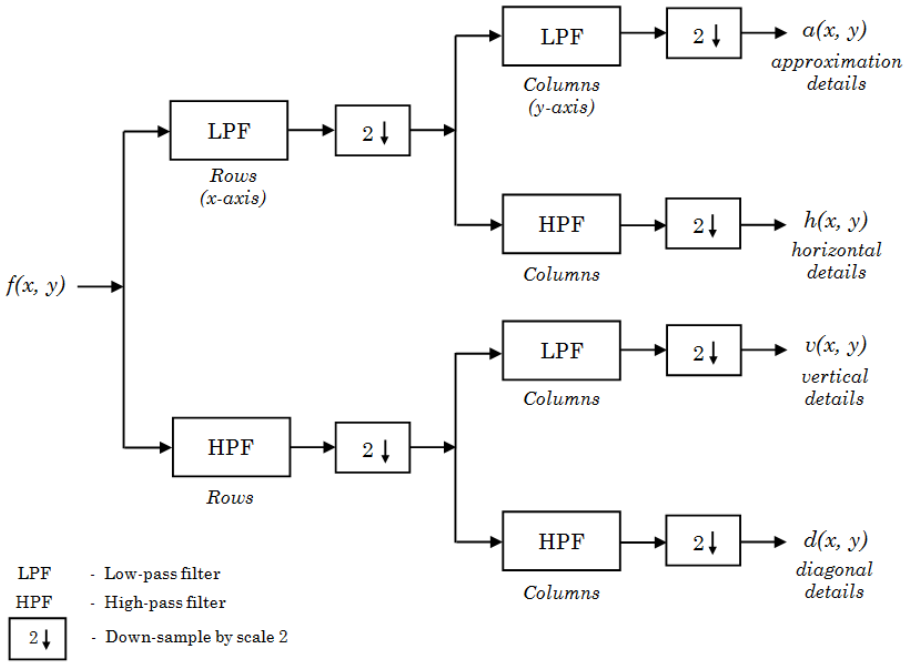

Figure 2.2 An illustration of two-level two-dimensional DWT of a signal.

The human visual system is sensitive to the contrast variations which are better represented in the horizontal, vertical and diagonal details of DWT of an image [37]. When combined at multiple scales, these details forms the feature maps comprising of contrast variations from edge to texture [16]. The proposed model (provided in chapter 4) focuses on extracting the contrast variations by applying IDWT to the horizontal, vertical and diagonal details at multiple scales while ignoring the approximation details.

2.5 The human eye and image formation

As light enters the human eye, the image formation occurs through mapping of spatial pattern in the optic array (intensity and spectral composition of light) onto retina [33]. The optimal mapping of an image depends on visual acuity of an eye, which comprises of central field and peripheral field. The central field or *fovea* samples smaller segments of optic array with high visual acuity whereas peripheral field samples larger segments of optic array with low visual acuity [33]. The retina is composed of two types of photo-light receptors called rod cells and cone cells. The distribution of rod cells and cone cells in the eye's retina is shown in Figure 2.3. The rod cells are mainly responsible for dim-light or *scotopic* vision whereas cone cells are responsible for bright-light or *photopic* vision. These photo-receptors convert the spatial pattern into electrical activity [33]. These electrical signals are further processed in the human brain.

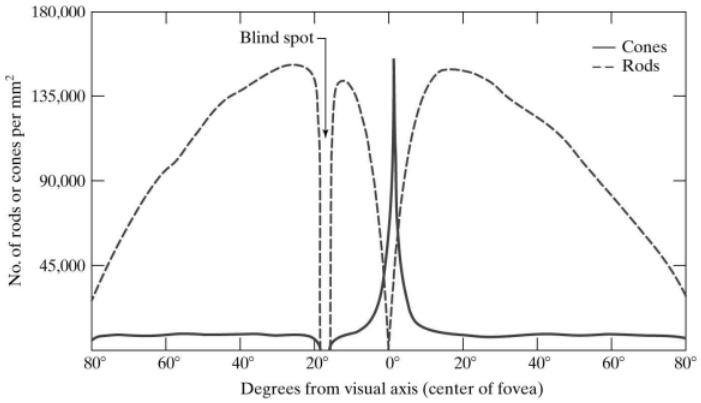

Figure 2.3 Distribution of rod cells and cone cells in the retina [29].

2.6 Visual saliency

A visual scene comprises of enormous amount of information which is typically complex in nature. When inspecting such a complex scene, the human visual system seems to employ a serial computational strategy through selectively attending the relevant regions of interest [8]. Visual saliency characterises some regions of a scene such as features and objects that appears to stand out relative to its neighbouring regions [1]. Figure 2.4 provides an illustration of saliency of an image. The enclosed region stands out from the rest of the image and is visually salient. Identifying these regions through visual saliency models facilitates in intelligent processing of visual information in computer vision systems.

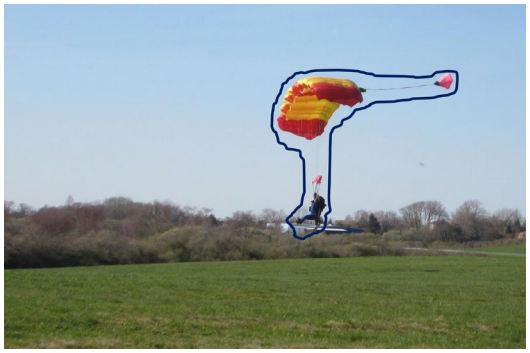

Figure 2.4 Illustration of visual saliency of an image. Sample image (obtained from MIT300 dataset [38]) with enclosed region indicating visual saliency.

2.6.1 Selective attention mechanism

The human visual attention can be compared with a moving-spotlight, where the intended regions are attended as an illuminated spotlight moving in a serial fashion

[39]. The attention process is achieved through a sequence of saccadic eye movements called fixations [3]. The objects in a scene are identified using eye fixations to bring the fovea of high visual acuity onto to the object [8]. During which, the human eyes in conjunction with brain plays a decisive role in selecting the relevant information to attend, for further processing. This process of *selective attention* mainly depends on the type of information, the user is interested in. An illustration is provided in Figure 2.5. It can be observed in image (b) that the eye fixations of many viewers (indicated using red markers [22]) are highly concentrated at the regions with human faces which are most relevant when compared to the rest. These eye fixations are obtained using an eye tracker where the user sits in front of a computer screen and the eye tracker records the eye movements [22].

(a) **(b)**

Figure 2.5 Illustration of human eye fixations. (a) Sample image obtained from MIT dataset [22], (b) the corresponding eye fixations are indicated using red markers [22].

2.6.2 Bottom-up versus top-down attention

In 1980, Treisman and Gelade introduced a two-stage framework of attention mechanism called Feature Integration Theory (FIT) [40]. According to FIT, over a visual field, the features of a scene are registered automatically at an early stage while the objects are identified separately at a later stage which requires focussed attention [40]. The focussed attention achieved through eye fixations are used to integrate the initially separable features into unitary objects. This theory has led to the development of many computational models of visual attention [1, 8, 41]. Two types of attention processes have been widely modelled in literature, which are categorised as, bottom-up attention and top-down attention as shown in Figure 2.6.

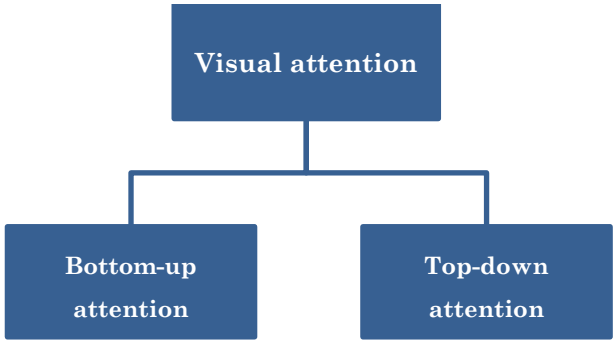

Figure 2.6 Classification of visual attention mechanism.

'Salient' is a term that is often considered in the process of bottom-up computation [1, 19, 42]. The bottom-up attention is a data-driven process that relies on the sensory information of the input image. It is termed as fast, exogenous and involuntary process [1]. During a free viewing condition, the low-level features of a scene, such as intensity, colour, orientation, edges and so on, automatically stand out and become salient from the rest. In the due course, our attention is unconsciously drawn towards these regions. As an example, in a visual scene with only one horizontal bar among several vertical bars, our attention is immediately drawn to the horizontal bar in a bottom-up manner [1, 40].

The later-stage of attention process is a slow, voluntary and driven by task called as top-down attention. Modelling this process mainly depends on understanding of behavioural aspects of human cognition and is based on high-level factors such as search task, scene context, past knowledge and user expectations [1]. Given a task, a search process is initiated and the scene is attended based on demands of the assigned task [3]. When assigned a task to find red balls from an assorted pool, the attention is confined to the red regions. Likewise, in the case of scene context, the regions are attended based on the given context while ignoring the rest. When searching for cars on the street, the attention is focussed on to the street, while ignoring the sky region [10, 42]. Many computational models have been developed to simulate these attention processes which are discussed in the next section.

2.7 Computational models

Over the past two decades, modelling of saliency-based attention has been a very active research area [1]. There is an increasing demand in visual saliency models, that can accurately predict the most relevant information within large amount of visual data [42]. Many researchers have contributed their knowledge through psychological,

neurobiological, and computational perspectives, to simulate the attention mechanism. In literature, several computational visual saliency models with various approaches have been proposed. The main idea behind these models is to compute several feature maps in parallel and combine them according to their saliencies to result in a topographical representation called *saliency map* [42]. The models differ in the way the feature maps are computed and combined. Some of the important computational models of visual attention available in literature are discussed below. Here the classification is based on the attention process.

2.7.1 Bottom-up models

In 1998, Itti *et al.* [19] proposed Neuromorphic Vision Toolkit (NVT). It is one of the earliest bottom-up computational model of visual attention and most commonly used as benchmark. The model was designed to mimic the visual system of early primates. An image with resolution of 640×480 pixels was given as input to the model. Three image features, namely intensity, colour and orientation are used to define feature maps at eight spatial scales. The centre-surround differences of features at various spatial scales were derived to obtain several feature maps in parallel. Summation of these feature maps results in conspicuity maps which are normalised and linearly combined to obtain a final saliency map. The saliency map drawn from the model is a grayscale image with brightness of each pixel corresponding to its saliency [42]. The model has shown to be successful in predicting fixations [18], but is still considered complex and involves many parameters which needs to be tuned as per the requirement.

Harel *et al.* in [14] proposed a Graph-based visual saliency model. Similar to Itti *et al.*, the features, intensity, colour and orientation are modelled in defining feature maps at multiple scales. These feature maps are used to build Markov chains on fully connected graphs. The equilibrium of dissimilar regions of these graphs forms the normalised activation maps which are linearly combined to obtain a final saliency map. The model still outperforms many state-of-the-art saliency models in predicting fixations. However, it is computationally complex to be employed in practical applications.

2.7.2 Top down models

In 2003, Oliva *et al.* [20] proposed a probabilistic model with scene context as a top-down cue. In a visual search task, the scene context plays an important role in directing the

attention process to relevant regions of the image. It provides a shortcut for the identification of locations in object detection. Based on this theory, the model uses local and global image statistics to derive contrast features. These features are used to detect the salient target object in a visual search task. The probability of features being frequent is more likely to belong to the background. The presence of an object in the target's location is identified based on the joint probability distribution of features of the target which is conditioned on the scene gist. The model is good at estimating saliency on a global context than local context.

In 2011, Borji *et al.* [43] proposed a top-down model based on attentional control used in goal-oriented tasks. The authors consider the perspective of human behavior with respect to the learning of offline data. When learning how to drive, a subject familiarises himself with offline data such as traffic signs and their associated meanings and applies the knowledge in real-time driving where attention is controlled based on top-down influences. The salient object candidates are recognised by intelligent selection of feature channels, scales of the saliency model and heuristic approach which results in the reduction of computational resources.

2.7.3 Combined models

In contrast to [20], Zhang *et al.*, [18] proposed SUN (Saliency Using Natural statistics) model. The model is implemented based on Bayesian framework which incorporates high-level information with bottom-up saliency. The Shannon's self-information of the visual features is used as low-level information with prior knowledge of the target as high-level information. This prior knowledge is obtained from the natural statistics of collection of natural images. Similar to [20], the target information of a scene is seen as distinctive information from the background. The overall saliency is derived based on the point-wise mutual information between the visual features and the desired target.

Goferman *et al.* [12] proposed Context Aware Saliency (CAS) where scene context is considered as essential information just as salient objects. The colour and contrast are considered as low-level features. According to their concept, the salient regions should not only contain the salient object but also the parts of scene that convey the context. A pixel which is consistently distinct from its surrounding pixels both locally and globally is defined as salient. The regions of immediate context of these pixel regions is extracted by applying a threshold value. Moreover, a center prior which is a top-down feature, is

incorporated to the estimated saliency by applying a two-dimensional Gaussian filter positioned at the center. The authors also suggested to incorporate the top-down features such as detection of human face and objects for further improvement of the model.

2.7.4 Frequency based models

The popular frequency-based approach of bottom-up attention was proposed by Hou *et al.* in [15]. The model uses spectral characteristics of a grayscale image in Fourier transform domain to compute image saliency. Saliency is defined as a spectral residue obtained from difference between log amplitude spectrum of a grayscale image and its average. The model requires only few lines of code for implementation with basic operations involved and is fast compared to many existing models. However, the limitations of the model occur from the generated saliency maps which are of low resolution (64 × 64 pixels) when compared to the size of an original image.

In contrast, Guo *et al.* in [13] obtained saliency using phase spectrum of Fourier transformed image. The model has shown to be faster than [15] and better in terms of performance. The work was further extended in [4] called Phase spectrum of Quaternion Fourier Transform (PQFT). The model included temporal aspects by introducing a motion feature. The four low-level features, two colour channels, one intensity channel and one motion channel were employed to derive the saliency map. These features were computed in a parallel fashion to enhance the computational speed of the model.

Achanta *et al.* in [9] proposed a frequency model based on contrast differences of an image in *CIE* Lab colour space. The difference between arithmetic mean vector and Gaussian blurred image is used to define the image saliency which is further processed using mean-shift segmentation algorithm to improve the performance. The model suffers from limitations of being applicable to images with large and homogenous objects.

2.7.5 Machine learning models

Many authors have explored machine learning approaches in modeling visual attention. In 2009, Judd *et al.* [22] employed Support Vector Machine (SVM) learning to predict the regions of salient fixations. The authors collected eye tracking data of 15 viewers on 1003 natural images to train the model based on bottom-up, top-down and mid-level cues. Liu *et al.* in [44] employed a multi-resolution Convolution Neural Network (*Mr*-

CNN) to learn the top-down and bottom-up visual features from raw image data. The model uses raw image pixels as inputs and eye fixation attributes as labels. Fixation information is predicted using logistic regression by integrating the top-down and bottom-up information. These approaches have provided improved performance compared to the existing methods. However, they depend on prior knowledge of image data and require massive size of image datasets for training accurate models.

2.8 Visual saliency applications

Many computer vision systems deal with high resolution images and videos. In real time it is a challenge to process excessive visual information with limited computational resources [10]. One of the important computer vision tasks is to detect the regions of interest (ROI) [42]. Detecting the ROI information through saliency models will further eliminate the need of processing redundant information while limiting the computational requirements. Visual saliency models have applications in the areas such as image and video compression [4], image segmentation [5], remote sensing [6], and robotics [7].

In image and video compression algorithms, given a saliency model it highlights the relevant regions of human interest, these image regions can be compressed by adaptively allocating the number of bits for coding image regions according to their saliency [42]. For example, the salient regions can be allocated with higher number of bits when compared to other regions which can be compressed using lower number of bits. This will further avoid the loss of relevant information and enhance the image quality and performance of the compression algorithm. In image segmentation techniques, a saliency map can be used to separate the regions of interest such as objects, human faces and so on from the image background. The model proposed by Achanta *et al.* [9] generates a saliency map with well-defined boundaries of salient objects with homogenous regions, which facilitates the process of object segmentation tasks. Analogously, saliency is also employed in driver assistance systems to provide necessary traffic information to ensure safe driving by detecting the regions of pedestrian and traffic light information [45]. In conclusion, prediction accuracy and computation time of a saliency detection model is critical and has great impact on these applications in terms of efficiency and performance.

2.9 Summary

This chapter presented the background knowledge related to the digital image processing in frequency domain, visual saliency detection in images and its applications. In the earlier sections, the fundamental knowledge of digital image and frequency-based image processing techniques with emphasis on DWT were discussed. In the later sections, the underlying concepts of visual saliency including types of visual attention mechanisms, existing literature in saliency modelling and applications of visual saliency were discussed. In the next chapter, the experimental methodology used for the development of the proposed model is presented.

3 Experimental methodology

3.1 Introduction

This chapter presents the experimental methodology used in algorithm development and evaluation of the saliency detection model proposed in chapter 4. The chapter is organised into the following sections. The details of the experimental set-up used in software development and testing are discussed in section 3.2. The testing methods used in model evaluation including details of image datasets, performance assessment methods and benchmark models are discussed in section 3.3. The summary of the chapter is provided in section 3.4.

3.2 Experimental set-up

The details including the choice of software used for algorithm development and the system specifications of the test environment are briefly discussed in this section. The experiments were conducted using the following experimental set-up.

3.2.1 Development environment

The algorithm development and testing were carried out using MATLAB (MATrix LABoratory) software, version R2019a. MATLAB is a high-level programming language that is extensively used in various numerical calculations, modelling and simulations, graphical visualisation and algorithm development [1]. It facilitates the exploration process of algorithmic design and allows fine parameter tuning with minimum coding effort. In addition, MATLAB was beneficial in this project for the following reasons. The image data can be represented in the form of a matrix (rows and columns) for detailed analysis and processing. The essential toolboxes such as Image processing, Parallel computing and Wavelet toolboxes with built-in functions are available that eliminate the need for additional coding requirements. The integrated development environment (IDE) provides simultaneous access to editor, debugger, command and workspace windows which allows easy debugging of the source code. The plotting and graphics tools serve in graphical visualisation of experimental data. Hence the use of MATLAB enables efficient design, development and testing of the proposed model.

3.2.2 Test environment

The experiments were conducted on a high-performance computer with the system specifications provided in Table 3.1. One of the challenges encountered during this project was to test the hypothesis on the entire image dataset with 1000 - 2000 images with maximum resolution of 1920×1080 pixels that required few hours of processing time. During this, the parallel computing toolbox was useful to assign computations of numerous images across available multi-core processor resources which further accelerated the program execution.

Table 3.1 System specification

Processor	Speed	RAM	Operating System	Architecture
Intel core i7 2600K	3.4 GHz	16 GB	Microsoft Windows 7	64-bit

3.3 Model evaluation

The proposed model (presented in chapter 4) is evaluated across two large public image datasets with widely used testing methods. The choice of image datasets and evaluation metrics are discussed in this section. Figure 3.1 shows a block diagram of the model evaluation. The images from dataset are given as input to the model and the corresponding saliency maps are obtained as output. The prediction accuracy of the model is evaluated using qualitative and quantitative analysis techniques. The computation time of the model is measured over a set of images. Finally, the obtained results are compared with the benchmark models.

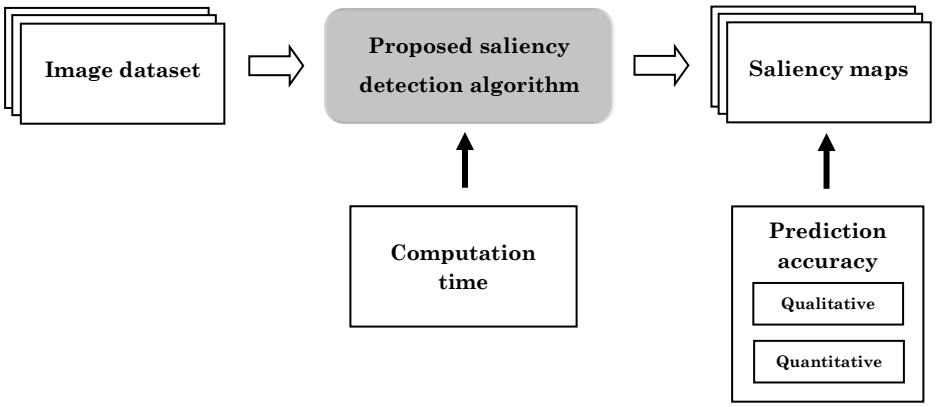

Figure 3.1 Block diagram of model evaluation.

3.3.1 Image dataset

An image dataset consists of numerous images under test and corresponding human ground truth (HGT) maps. The HGT map consists of eye fixation information in binary format obtained from human subjects through eye tracking system. These maps are widely used as reference for model evaluation. Figure 3.2 shows a sample image from MIT dataset [22] provided in column (a) with its corresponding eye fixations indicated using red markers in column (b). The HGT map with continuous distribution of fixations is provided in column (c). The authors obtained the continuous HGT map by convolving the binary map of eye fixations with a Gaussian filter [22].

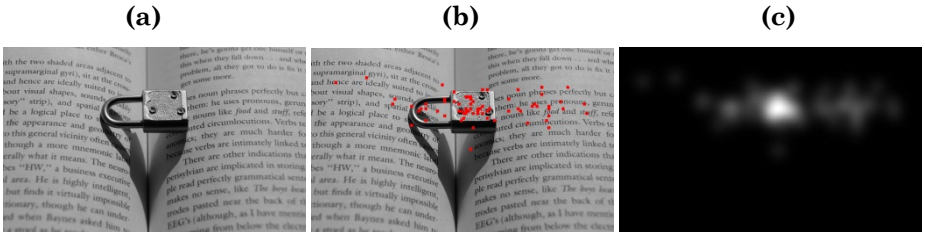

Figure 3.2 Sample image from MIT dataset [22]. (a) Input image, (b) Image with eye fixations indicated using red markers, (c) Corresponding HGT map indicating continuous distribution of fixations.

Testing a saliency model on large multiple image datasets ensures validating the model for prediction accuracy over various combination of images under test. This ensures that the model is not biased to a single dataset. The developed algorithm (presented in chapter 4) is tested on two large public image datasets namely, MIT [22] and CAT2000 [46]. In both datasets, the authors have employed eye-tracking devices to capture the eye fixations, represented as discrete points.

MIT dataset

The MIT dataset [22] consists of 1003 natural images collected from LabelMe and Flickr creative commons. The images are mainly comprised of indoor and outdoor scenes which include objects, people, faces and text. The resolution varies between 405 to 1024 pixels with a maximum dimension of 1024 pixels for all images and consists of 779 landscape and 228 portrait images. An illustration of sample images from MIT dataset is shown in Figure 3.3. The authors collected the eye fixation data from 15 observers with each image viewed for 3 seconds in a free viewing state. This dataset has been widely used as benchmark for evaluation of many saliency models.

Figure 3.3 Illustration of sample images from MIT dataset [22].

CAT2000 dataset

The CAT2000 dataset [46] consists of 2000 images (natural and artificial) with resolution of 1920×1080 pixels. These are categorised as 20 different types of scenes which include images of Action, Affective, Art, Black & White, Cartoon, Fractal, Indoor, Inverted, Jumbled, Line drawings, Low resolution, Noisy, Object, Outdoor manmade, Outdoor natural, Pattern, Random, Satellite, Sketch, and Social. An illustration of sample images from CAT2000 dataset is shown in Figure 3.4. The HGTs were obtained from eye fixation data of 120 observers with each image viewed for 5 seconds in a free viewing state. To address the challenges of dataset bias in saliency modelling, the authors introduced this large scale fixation dataset over several categories of images. This dataset mainly aims to overcome the shortcomings of small datasets which consists of small number of scenes shown to few observers.

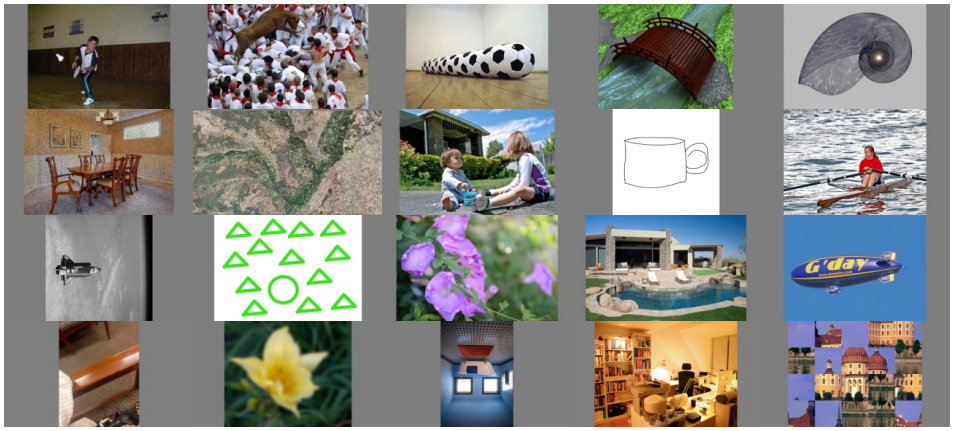

Figure 3.4 Illustration of sample images from CAT2000 dataset [46].

3.3.2 Qualitative analysis

The performance of the proposed model is qualitatively evaluated through a subjective visual comparison of the final saliency map with respect to HGT map. This comparison is used to visually measure the consistency of predictions in the final saliency map with the corresponding HGT map through human observation. Here the qualitative analysis is performed in two ways described as follows:

i. Analysis based on continuous distribution

In this method of comparison, both the reference HGT map and final saliency map consists of a continuous distribution of fixations [22]. The intensity values in the HGT and saliency maps ranges between (0, 255). Visually, the most salient pixels are represented with higher intensity values, in a decreasing order corresponding to the least salient pixels [10]. Figure 3.5 shows a qualitative analysis based on the continuous distribution of saliency maps obtained from different methods.

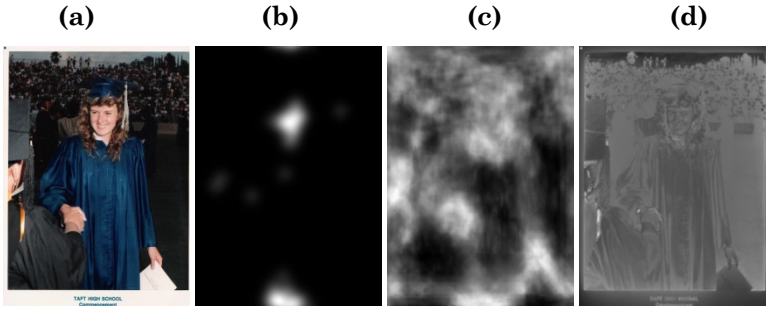

Figure 3.5 Qualitative analysis based on continuous distribution. (a) Input image [22], (b) HGT map, (c) Saliency map of model in [17], (d) Saliency map of model in [16].

ii. Analysis based on thresholding

In this method, the similarity of the saliency map and the HGT map in binary form, obtained through thresholding are visually compared. The binary map consists of foreground (salient) regions represented using white pixel information separated from background (non-salient) regions represented using black pixel information. Thresholding is performed based on Otsu's global threshold method, as it provides distinguishable foreground and background regions [47]. This method is user-independent and automatically selects the threshold value for binary map generation and has been widely used in the research community. The similarity of predictions is measured by placing the eye fixations in conjunction with the thresholded maps. Figure 3.6 shows a qualitative analysis of saliency maps based on thresholding.

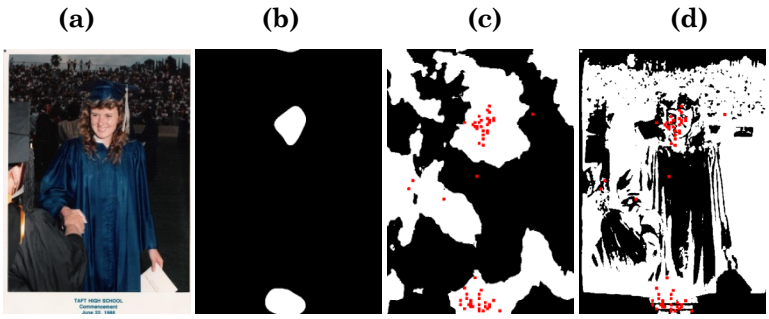

Figure 3.6 Qualitative analysis based on thresholding. The maps are thresholded using Otsu's global threshold method [47]. The red markers indicate the eye fixations. (a) Input image [22], (b) HGT map after thresholding, (c) thresholded saliency map of model in [17], (d) thresholded saliency map of model in [16].

3.3.3 Quantitative analysis

The prediction accuracy of the proposed model is quantitatively evaluated using three metrics - the Area under the ROC Curve (AUC), Pearson's Correlation Coefficient (CC) and Normalised Scanpath Saliency (NSS) which are most commonly used in fixation prediction. The saliency map and HGT map are given as inputs to the metrics. AUC and NSS are location-based metrics which considers saliency map values and HGT map as discrete fixation locations (binary format) whereas CC is a distribution-based metric which considers saliency map and HGT map as a continuous distribution [48]. Recently, the authors of [48] have indicated that either CC or NSS metric is sufficient to evaluate the model performance. However, both the metrics are included here, as they have been widely used in the literature (refer to TABLE 1 of [48]).

Area under ROC Curve (AUC)

AUC is a single scalar value representing the performance of the model's saliency map with respect to the corresponding HGT map [49]. It is the portion of the area under unit square obtained from ROC (receiver operating characteristics) curve and ranges between $[0, 1]$ with a high value of AUC is preferred. Given a saliency map, it is considered as a discrete classifier of fixated and non-fixated pixel predictions [48]. At each threshold, there are four possible outcomes from the matrix. If the actual pixel is positive and it is predicted as positive, it is counted as TP; if it is predicted as negative, it is counted as FN. If the actual pixel is negative and it is predicted as negative, it is counted as TN; if it is predicted as positive, it is counted as FP. Figure 3.7 shows a confusion matrix for prediction analysis [49].

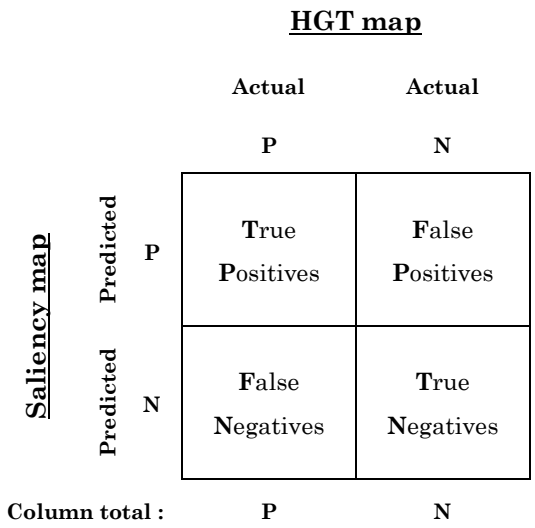

Figure 3.7 Confusion matrix for prediction analysis.

The TP rate and FP rate are obtained from equations (3.1) and (3.2) respectively.

$$TP\ rate = \frac{TP}{TP + FN} \tag{3.1}$$

$$FP\ rate = \frac{FP}{FP + TN} \tag{3.2}$$

An illustration of the ROC plot is shown in Figure 3.8. The ROC curve represents a two-dimensional graph obtained from the classification of true positive (TP) and false positive (FP) instances at various discriminative thresholds. At each threshold value, the TP rate (percentage of true salient pixels identified correctly) and the FP rate (percentage of non-salient pixels identified incorrectly) are obtained. The TP rate and FP rate at various thresholds form the ROC curve. The value of AUC is obtained from area under the ROC curve using the trapezoidal numerical integration method [49].

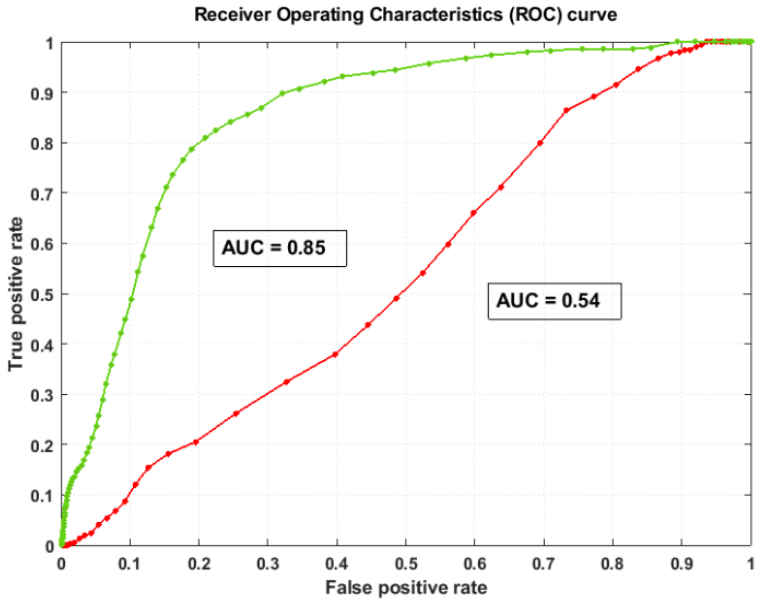

Figure 3.8 ROC plot for saliency map obtained from two different methods. A saliency map with green ROC curve indicating higher AUC score is preferred when compared to the other map with red ROC curve.

Pearson's Correlation Coefficient (CC)

CC is a measure of the linear relationship between the saliency map and its corresponding HGT map [25, 48]. The metric takes both the maps as inputs in the form of continuous distribution. It provides the similarity index of correlation between two maps. The value ranges between -1, representing anti-correlation, to 1, representing perfect correlation. CC is calculated as

$$CC(S, G) = \frac{cov(S, G)}{\sigma(S) * \sigma(G)} \tag{3.3}$$

where cov is the covariance between saliency map (S) and HGT map (G). $\sigma(S)$ and $\sigma(G)$ represents the standard deviations of respective maps.

Normalised Scanpath Saliency (NSS)

Given a saliency map that is normalised to have a zero mean and unit standard deviation, NSS is a measure of average values obtained from the saliency map that corresponds to fixated values in the HGT map [10]. The value of NSS is obtained by:

$$NSS = \frac{1}{N}\sum_{i=1}^{N}\frac{S(x_h^i, y_h^i)}{\sigma(S)} \tag{3.4}$$

where N is the total number of fixated pixels, $S(x_h^i, y_h^i)$ indicates fixated values at the pixel (x_h^i, y_h^i) in the HGT map, $\sigma(S)$ is the standard deviation of saliency map. The positive value of NSS indicates that the predictions corresponds to fixations, while the negative value of NSS indicates anti-correspondence. If the value of NSS is greater than 1, it indicates significant correspondence with fixations.

3.3.4 Evaluation of computation time

The computation time of the model is an important parameter that needs to be evaluated to make a model to be deployed in real time conditions. The average time cost per image required by a model under test is evaluated over a test set of 100 images with resolution of 768x1024 pixels. These times are obtained using MATLAB software with built-in stopwatch timer functions - *tic* and *toc* [50]. The timer is initiated by the *tic* command followed by the program execution and displays the elapsed time with the *toc* command in seconds with 6 decimal digits of precision. The average computation time (in seconds rounded to 2 decimal digits) over 100 images is obtained for different saliency methods and is used for comparison.

3.3.5 Benchmark saliency detection models

The proposed work explained in chapter 4 is compared for efficiency in terms of prediction accuracy and computation time with respect to the existing methods. Two popular bottom-up computational models proposed by Murray *et al.* [17] and Imamoglu *et al.* [16] are chosen for comparison. Both the models detect saliency in images using wavelet coefficients, developed in MATLAB with available open-source code. The respective saliency maps were obtained for both the models using the images of MIT [22] and CAT2000 [46] datasets which are used in qualitative and quantitative analysis of the proposed model.

3.4 Summary

The experimental methodology involved in the research project is explained in this chapter. The MATLAB software with a relevant test environment is set up for experimentation. The performance of the model is evaluated using qualitative and quantitative analysis techniques using two large public image datasets which consist of various natural and artificial images under test. The qualitative analysis involves a comparison of saliency maps obtained from different methods. The quantitative analysis

involves validating model prediction accuracy using three metrics - AUC, CC and NSS and the computation time of the model is evaluated over 100 images of 768x1024 resolution. The details of two relevant benchmark models used for comparison with the proposed model is provided.

4 Low-computation wavelet-based visual saliency model

4.1 Introduction

This chapter presents the low-computation visual saliency model in the WT domain. The model aims to predict the regions of human eye fixations in images using bottom-up features. The proposed work requires two-channels, luminance (Y) and chrominance (Cr) in YCbCr colour space for saliency computation. The local contrast features of these channels are used to construct the multi-level feature maps in the wavelet domain, which are combined using a two-dimensional entropy scheme and enhanced using natural logarithm transformation to obtain a final saliency map. This unique saliency map highlights the regions of human eye fixations. The proposed model has been tested on common benchmark datasets. The experimental results (provided in chapter 5) show that the proposed model has achieved a significant reduction in computation time with better prediction accuracy compared to the benchmark models.

The chapter is divided into following sections. The existing wavelet-based saliency detection models are discussed in section 4.2. An overview of the proposed algorithm is provided in section 4.3. The stages involved in algorithm development are discussed in section 4.4. The summary of chapter is provided in section 4.5.

4.2 Related work

In literature, computational models in frequency domain [4, 9, 13, 15] (discussed earlier in section 2.7.4) estimate image saliency on a global context and lack in prediction accuracy as they do not contribute to local saliency details. As discussed in chapter 2, WT is a better tool for local signal analysis as it offers multi-scale spatial and frequency analysis at the same time. Recent studies [16, 17, 25, 26] have exploited the MRA property of WT and have shown to provide improved prediction accuracy compared to the traditional frequency-based methods [4, 9, 13, 15].

Murray *et al.* in [17] proposed a low-level vision model which uses biologically inspired Gabor-like wavelets to detect saliency at multiple scales. The model estimates saliency in the high-frequency directional details (horizontal, vertical and diagonal) of three colour-opponent channels [11] (p.17-18) which are convolved with a set of centre-

surround filters at decomposition phase. The saliency details are further integrated using Euclidean norm during the reconstruction phase, by applying scale-weights defined by Extended Contrast Sensitivity Function (ECSF). The model has shown improved prediction accuracy compared to other existing methods due to the local saliency contribution. However, it has certain limitations. The window-sizes of centre-surround filters involved in the filtering process are not well defined which induces noise in the salient regions. This effect is reflected at multiple scales which makes it difficult to distinguish between true salient pixels and noisy pixels in the final saliency map. As a result, spatial information loss is unavoidable in this method.

In contrast to [17], Imamoglu *et al.* [16] proposed a bottom-up model in wavelet domain, which incorporates local saliency along with global saliency information. The local saliency map is obtained from the maximum of directional details between the channels of *CIE* Lab colour space (CIE standard D65 illumination) [29] at multiple scales in the wavelet domain. A corresponding global saliency map is derived based on the distribution of local features. The maps are combined non-linearly by multiplying the local saliency map with exponential value of the global saliency map. The combined map is normalised and enhanced to derive a final saliency map. The model lacks in prediction accuracy with local saliency solely when compared to [17]. The overall model out-performs the model in [17] but is computationally expensive. The additional time cost is mainly associated with the computation of global saliency map and enhancement operation of the final saliency map.

Ma *et al.* [25], proposed a bottom-up wavelet model to handle various contrast variations using different colour spaces. The wavelet transform is applied to channels of RGBYI, *CIE* Lab and YCbCr colour spaces to obtain respective saliency maps using directional details. These saliency maps are Gaussian smoothed and combined to obtain a final saliency map. The reciprocal value of two-dimensional entropy of a saliency map is used as weight during combination. The final map is derived during wavelet decomposition while the reconstruction is ignored to ensure reduction in time cost of the model. The model considers information at three decomposition scales which contributes to local saliency details and requires nine channels from three colour spaces for saliency computation. The model uses large filter sizes (0.1 times the size of image) in the process of Gaussian smoothing of saliency maps which causes heavy blurring and results in spatial information loss.

Scharfenberger *et al.* [26] proposed a statistical model in wavelet domain which defines saliency as non-redundant pixels at multiple scales. The model obtains salient regions in *CIE* Lab colour space using un-decimated wavelet transform with directional details processed for non-redundancy at three scales. The model is mainly designed to deal with noisy images. Each pixel is processed for noise which estimates the probability of pixel information being a noisy observation to its neighbouring pixel at each scale.

In summary, the aforementioned wavelet-based saliency detection methods have considerably achieved good prediction accuracy, but have not considered the performance evaluation with respect to computation time of the model. As discussed earlier in chapter 2, many of the saliency applications deal with high-resolution images and real-time videos. It is a challenge to process excessive visual information with limited computational resources. Hence achieving good prediction accuracy while maintaining low computation time is critical for a saliency detection model which has a great impact on these applications in terms of accuracy and efficiency.

4.3 Overview of the proposed model

The proposed model will focus on two main objectives. Firstly, it focuses on reducing the computation time of the model by limiting the number of colour channels required for saliency computation. Secondly, it aims to improve the prediction accuracy by combining the feature maps obtained from high-frequency directional details through an entropy-based feature combination scheme.

The mechanism is to create feature maps at multiple scales by extracting the local contrast features of luminance (Y) and chrominance (Cr) channels in YCbCr colour space, which are combined using two-dimensional entropy scheme and enhanced using natural logarithm transformation. Initially, an input RGB image is colour transformed to YCbCr colour space. A two-dimensional DWT is employed to decompose the luminance (Y) and chrominance (Cr) channels at multiple dyadic scales. The feature maps are constructed by applying IDWT to the high-frequency directional details (horizontal (h), vertical (v) and diagonal (d)) of luminance and chrominance channels. The obtained feature maps at multiple levels are combined using two-dimensional entropy scheme. The combined map is enhanced using natural logarithm transformation to obtain a final saliency map. The detailed development of the proposed model is presented in section 4.4.

4.4 Algorithm development

The development of the proposed model is divided into four stages: a) colour transformation, b) multi-scale feature extraction, c) entropy-based feature combination, d) normalisation and enhancement. A schematic representation of the model is depicted in Figure 4.1. The four stages of algorithm development is explained in detail in the following sub-sections.

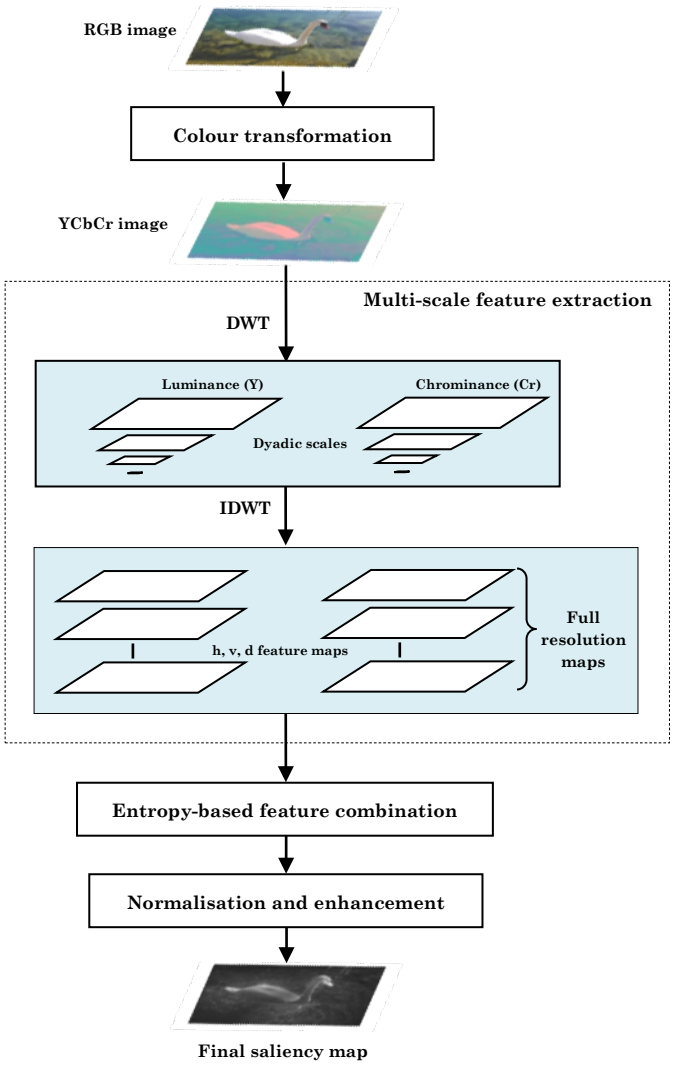

Figure 4.1 Schematic representation of the proposed model.

4.4.1 Colour transformation

In the proposed work, the YCbCr colour space is preferred over RGB colour space as it can represent luminance and chrominance channels separately (refer to the details of YCbCr colour space provided in chapter 2). Therefore, an input image in RGB format is

converted to YCbCr colour space (refer to colour conversion equations provided in chapter 2) to obtain separate luminance and chrominance channels. The MATLAB function 'rgb2ycbcr' is employed for the conversion [50]. The obtained Y, Cb and Cr channels are convolved using a two-dimensional Gaussian low-pass filter. This filter is rotationally symmetric and produces uniform smoothing of image details. A small filter size (3×3) (with σ = 0.5; the default MATLAB configuration) is chosen for the smoothing operation as in [11, 16]. This 3×3 filter is used to eliminate very high-frequency noise present in the channels due to the colour conversion. The discrete approximation of a sample 3×3 Gaussian filter is presented in Figure 4.2.

$$\frac{1}{16} \times \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

Figure 4.2 Discrete approximation of a sample 3 × 3 Gaussian filter

Equation (4.1) represents the smoothing process:

$$f(x, y) = I(x, y) * G_{s \times s} \tag{4.1}$$

where $I(x, y)$ represents input image channel with (x, y) being pixel co-ordinates in two-dimensional space, $G_{s \times s}$ is a two-dimensional Gaussian low-pass filter with filter size s = 3, '$*$' is the convolution operator and $f(x, y)$ represents Gaussian smoothed channel.

Colour channel selection

To identify the effective channels required for saliency computation, a set of channel combinations were used in obtaining the final saliency scores using two large image datasets, MIT [22] and CAT2000 [46] (refer to the details of image datasets explained in chapter 3). The experimental results obtained for different channel combinations are provided in Table 4.1. The first column shows the different channel combinations considered. The second and third columns (main) show the results obtained for MIT [22] and CAT2000 [46] datasets respectively. The corresponding AUC, CC and NSS metric scores are indicated in respective sub-columns (refer to chapter 3 for more details related to the metrics used in this work). Higher scores of AUC, CC and NSS indicate good prediction accuracy of a saliency model.

As shown in Table 4.1 the model has achieved better saliency scores for the channel combinations {Y, Cb, Cr} and {Y, Cr} on two large image datasets when compared to the other channel combinations. The model has achieved the lowest performance when Y only is used whereas it has shown improved performance when Y is combined with the other channels. It is notable from the results of channel combinations {Y, Cb, Cr} and {Y, Cr} that both have provided similar performance. This shows that channel Cb has provided an insignificant contribution to the model performance and can be eliminated. This achieves approximately one-third of reduction in computation time. Therefore, in the proposed work, the channel combination {Y, Cr} is chosen as it provides good performance with reduced computation time.

Table 4.1 Experimental results for different channel combinations

Channel combination	MIT dataset [22]			CAT2000 dataset [46]		
	AUC	CC	NSS	AUC	CC	NSS
{Y, Cb, Cr}	0.70	0.24	0.82	0.70	0.31	0.78
{Y, Cb}	0.69	0.22	0.76	0.69	0.29	0.74
{Y, Cr}	**0.70**	**0.24**	**0.82**	**0.70**	**0.31**	**0.77**
Y	0.67	0.20	0.67	0.61	0.26	0.37

4.4.2 Multi-scale feature extraction

Edges are the bottom-up features that highlight image contrast or step-change in intensities defining boundaries of image objects [35]. The human visual system is sensitive to these contrast variations of an image [37] and can draw human attention in a bottom-up manner. The multi-resolution analysis of DWT is employed to extract local contrast variations in Y and Cr image channels. DWT uses a set of filters which decomposes the signal into independent frequency components (low-pass and high-pass). The local contrast variations are better represented in the high-pass frequency components of DWT which consists of details oriented in horizontal, vertical and diagonal directions [37] (refer to the fundamentals of DWT provided in chapter 2). Furthermore, the experiments conducted by authors in [51] showed that relevant information can exist at different scales (from fine to coarse). Therefore, the local contrast features of Y and Cr channels are extracted by decomposing them at a maximum number of scales obtained by:

$$N = log_2(D_{max}) \quad (4.2)$$

where N is an integer (obtained by using MATLAB 'round' function [50]) representing the number of decomposition scales and D_{max} being the maximum dimension of the input image.

4.4.2.1 Wavelet decomposition

The Y and Cr channels are decomposed at N scales, obtained using equation (2.4). The biorthogonal wavelet 'bior4.4' with symmetrical nature of its wavelets and scaling functions are chosen for decomposition because human vision is more tolerant of symmetrical errors [23]. In addition, it has also provided better saliency results (which are preliminary results obtained using MIT dataset) when compared to other wavelets of the family as shown in Table 4.2.

Table 4.2 Experimental results for biorthogonal wavelet family

Channel combination - {Y, Cr} Dataset - MIT [22]			
Wavelet	***AUC***	***CC***	***NSS***
'bior1.1'	0.6779	0.2101	0.7224
'bior2.2'	0.6986	0.2377	0.8083
'bior3.3'	0.6949	0.2344	0.8016
'bior4.4'	**0.7014**	**0.2415**	**0.8206**

Equation (4.3) shows DWT applied to the colour channel f(x, y) at the i^{th} scale:

$$(f_i^a(x_a, y_a),\ f_i^s(x_s, y_s)\ ) = DWT(f_i(x, y)) \quad (4.3)$$

where $f(x, y) \in$ {Y, Cr}, $i \in$ {1, 2, ..., N}, f^a represents transformed matrix that consists of low-frequency approximation (a) coefficients (x_a, y_a), and f^s with $s \in \{h,\ v\ ,d\}$ represents individual transformed matrices that consist of high-frequency horizontal (h), vertical (v) and diagonal (d) coefficients (x_s, y_s), respectively.

4.4.2.2 Feature map generation

The Y and Cr feature maps are generated by applying IDWT to the high-frequency detailed coefficients at N scales to derive feature maps at N levels as in [16]. These feature maps can represent contrast details from edge to texture. The approximation (a)

details are omitted during reconstruction as these are of purely low-pass details. At each level, the feature maps are derived based on the equation:

$$c_i(x,y) = IDWT\big(c_{i+1}(x,y),\, f_i^s(x_s,y_s)\big) \tag{4.4}$$

where $c_i(x,y)$ represents the feature map obtained at the i^{th} level with f^a is set to zero. The feature map $c_i(x,y)$ is constructed using f_i^s details and details in the preceding feature map $c_{i+1}(x,y)$. At level N, the feature map consists of details from the N^{th} scale (the coarsest scale) to the 1st scale (finest scale) and so on.

An illustration of the process of feature construction at 3 levels is shown in Table 4.3. At level 1, the feature map is obtained using the finest scale wavelet coefficients by setting the approximation details to zero; at level 2, the feature map is constructed using 1st and 2nd scale wavelet coefficients and so on. The obtained feature maps at each level are in full resolution (same size of the input image) and represent the contrast details from edge to texture.

Table 4.3 Illustration of a 3-level feature map construction.

Level 1	Level 2	Level 3
$c_1(x,y) = IDWT\big(0,\, f_1^s(x_s,y_s)\big)$	$c_2(x,y) = IDWT\big(0,\, f_2^s(x_s,y_s)\big)$ $c_1(x,y) = IDWT\big(c_2(x,y),\, f_1^s(x_s,y_s)\big)$	$c_3(x,y) = IDWT\big(0,\, f_3^s(x_s,y_s)\big)$ $c_2(x,y) = IDWT\big(c_3(x,y),\, f_2^s(x_s,y_s)\big)$ $c_1(x,y) = IDWT\big(c_2(x,y),\, f_1^s(x_s,y_s)\big)$

4.4.3 Entropy-based feature combination

The entropy of an image can be defined as a statistical measure of information content present in the image. Equation (4.5) represents entropy *end* of an image I with n gray levels:

$$en(I) = -\sum_{k=1}^{n}(h_k)\log_2(h_k) \tag{4.5}$$

where h_k is normalised histogram count of k^{th} gray level. The feature maps of luminance (Y) and chrominance (Cr) channels are combined at each level using two-dimensional entropy weights. Each feature map is multiplied with its two-dimensional entropy value as provided in equation (4.6)

$$c'(x,y) = c(x,y) \times en_c \tag{4.6}$$

where $c'(x,y)$ is weighted feature map, en_c is two-dimensional entropy of feature map $c(x,y)$. Equation (4.7) represents the feature combination:

$$C(x,y) = \sum_{i=1}^{N} (|y'_i(x,y)| + |cr'_i(x,y)|) \tag{4.7}$$

where $|y'(x,y)|$ and $|cr'(x,y)|$ correspond to respective Y and Cr weighted feature maps with absolute values and $C(x,y)$ represents the combined map at N levels. The combined map defines the most conspicuous regions of the image. An illustration is provided in Figure 4.3.

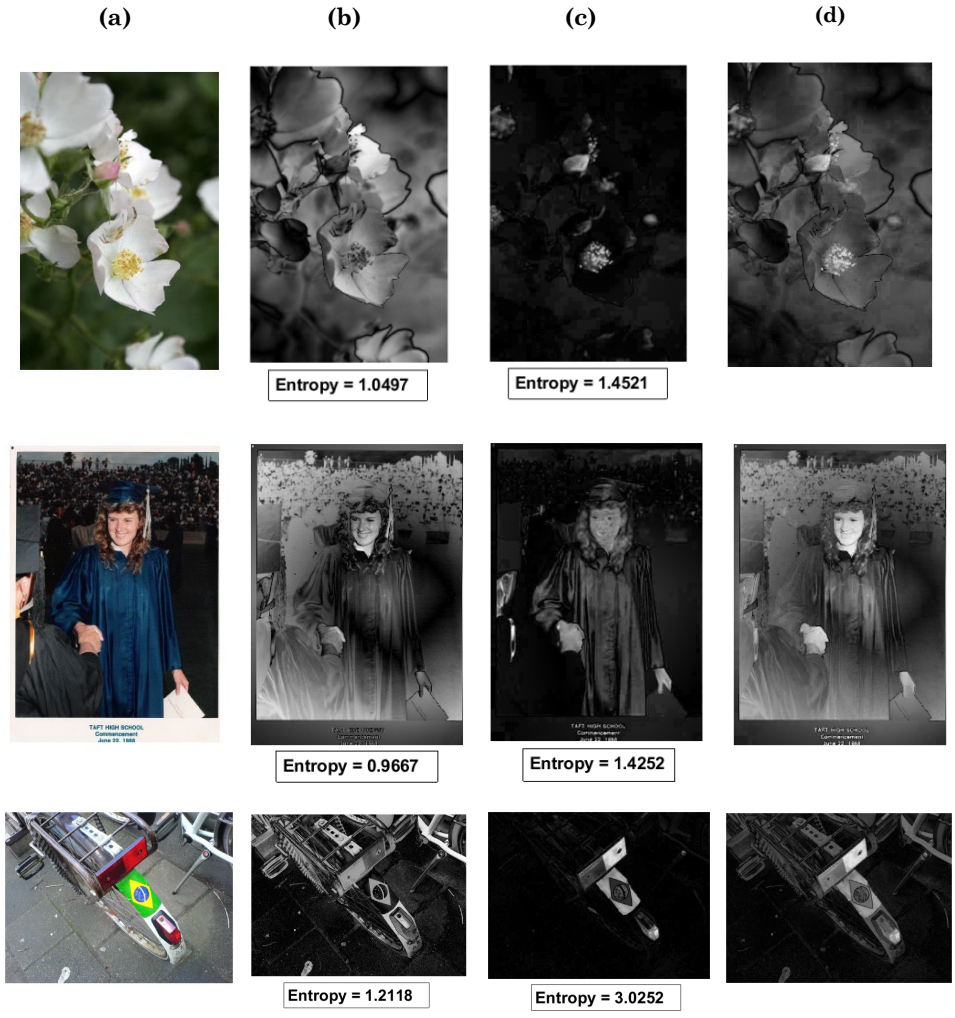

Figure 4.3 Illustration of entropy-based feature combination (a) Input image, (b) Y channel feature map, (c) Cr channel feature map, and (d) the combined map obtained from Y and Cr feature maps using entropy as weights.

Here, the two-dimensional entropy of the feature map is used as a weight to prioritise the feature combination. A feature map with high entropy value indicates a high

saliency content and gets higher priority compared to the feature map with a low entropy value. For instance, if the entropy value of chrominance is high, then these features get high priority than luminance and vice versa.

4.4.4 Normalisation and enhancement

The details in the combined map are smoothed using a two-dimensional Gaussian low-pass filter. The filtering process eliminates high-frequency noise due to wavelet processing. To ensure there is minimal information loss, the common filter size 5 (with σ = 0.5; the default MATLAB configuration) is considered for smoothing as in [9, 11, 16]. Furthermore, the combined map is normalised to a range $[0, 1]$ which will ensure the details lie within the same range. Finally, the normalised map is enhanced using natural logarithm transformation to obtain the final saliency map.

Equation (4.8) represents the normalisation operation:

$$C_M(x, y) = M(C(x, y) * G_s) \tag{4.8}$$

where $C_M(x, y)$ is a normalised map, $M(.)$ is normalisation operator and $'*'$ is a convolution operator and G_s represents a two-dimensional Gaussian low-pass filter with filter size $s = 5$. Equation (4.9) represents the enhancement operation:

$$fmap(x, y) = ln(C_M(x, y) + 1) \tag{4.9}$$

where $fmap(x, y)$ is a final saliency map and operator $ln(.)$ is a natural logarithm transformation. The value of '1' is added in the operator, as the logarithm of '0' is undefined. The logarithm transformation maps the narrow range of low intensity values to a wider intensity range [29]. This transformation will enhance the salient regions embedded in the darker regions of the saliency map and compresses the higher-level values.

An illustration of the enhancement operation is provided in Figure 4.4. It can be observed in Figure 4.4, that the darker pixel regions in the saliency maps of row (b) such as the regions of flower in the first and third image and the regions of human face in the second image have been enhanced as shown in row (c). Further, the enhancement operation has also improved the visual appearance of the respective saliency maps in the row (c) when compared to row (b).

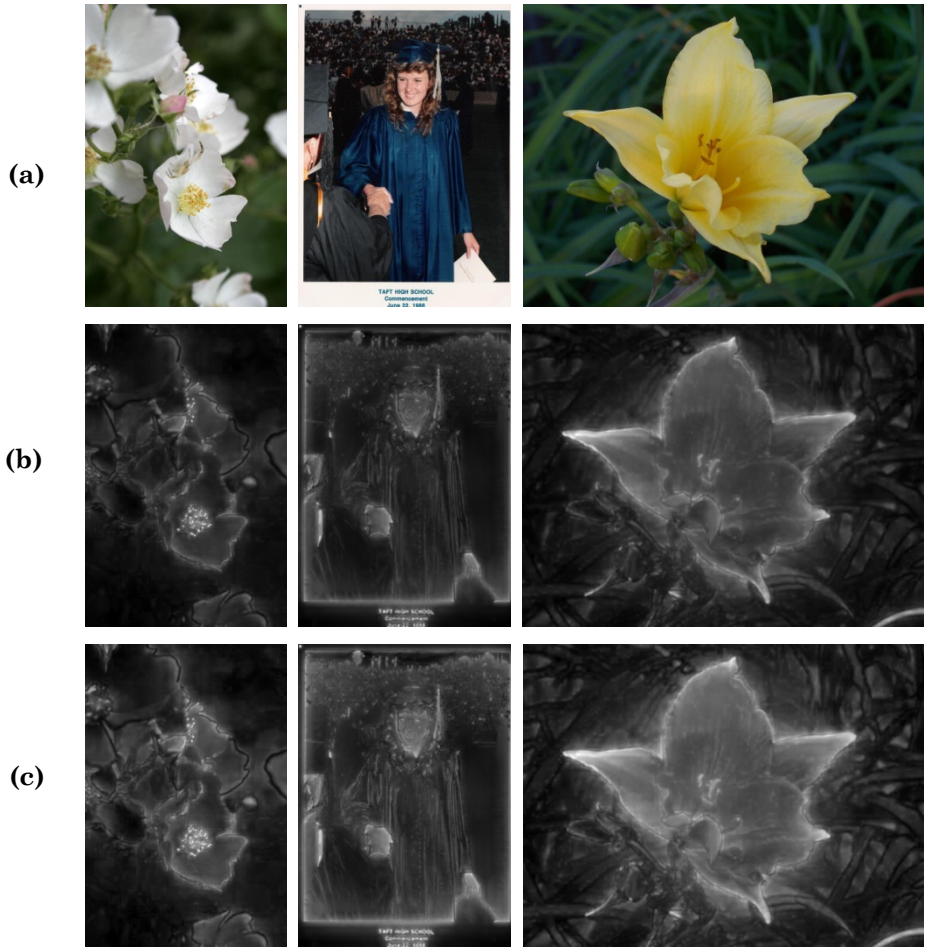

Figure 4.4 Illustration of the enhancement operation. (a) Input images, (b) Saliency maps before enhancement, (c) Saliency maps after enhancement.

4.5 Summary

The proposed low-computation wavelet-based saliency detection model is explained in this chapter. The model aims to improve prediction accuracy with low computation time when compared to the existing wavelet-based saliency detection models. Unlike the existing wavelet-based saliency detection methods, the proposed model requires only two colour channels, luminance (Y) and chrominance (Cr) in YCbCr colour space for saliency computation as these two channels are sufficient to provide good performance with one-third of computational savings (from the results provided in Table 4.1). The identified colour channels of an image are decomposed using two-dimensional DWT at multiple scales to extract local contrast variations from the high-frequency directional details. These high-frequency directional details are selected as they can represent the contrast variations of an image, which draws human attention. The extracted feature maps are combined based on two-dimensional entropy scheme. This combination scheme

is used to prioritise the feature combination between Y and Cr feature maps which has shown improvement in prediction accuracy. The combined map is normalised using natural logarithm transformation to derive a final saliency map. The experimental results of the proposed model are discussed in chapter 5.

5 Experimental results and analysis

5.1 Introduction

This chapter presents the analysis based on the experimental results of the proposed low-computation wavelet-based saliency detection model explained in chapter 4. The procedure related to the model evaluation has been explained in chapter 3. The prediction accuracy of the model is evaluated using qualitative and quantitative measurement techniques using two public image datasets - MIT [22] and CAT2000 [46]. Both the datasets consist of images of various categories such as natural and artificial, indoor and outdoor scenes including human faces, people and text (refer to the details of image datasets provided in chapter 3). Finally, the model is evaluated for computation time. The experimental results are compared with the results of two bottom-up state-of-the-art wavelet-based saliency detection models Murray *et al.* [17] and Imamoglu *et al.* [16]. The authors in [16] have included focus of attention as in [12] to enhance the performance of their model. Including focus of attention can influence the performance of the model greatly by either increasing or decreasing its prediction accuracy. Therefore, to obtain a fair comparison among the methods, the results are computed without including focus of attention for the model in [16]. The rest of the chapter is organised into the following sections. The qualitative results are discussed in section 5.2. The quantitative results are discussed in section 5.3. The results pertaining to model's computation time are discussed in section 5.4. The summary of the chapter is provided in section 5.5.

5.2 Qualitative results

The proposed model is qualitatively evaluated by visually comparing the final saliency map of an input image with the corresponding human ground truth (HGT) map and saliency maps of benchmark models. This subjective observation will help in evaluating the consistency between the human ground truth map and the corresponding saliency maps of different models. The qualitative results of saliency maps based on continuous distribution is provided in Figure 5.1 and the results based on thresholded saliency maps are provided in Figure 5.2. The input images are provided in the first column with corresponding HGT maps, saliency maps of Murray *et al.* [17], Imamoglu *et al.* [16] and proposed model are provided in second, third, fourth and fifth columns respectively.

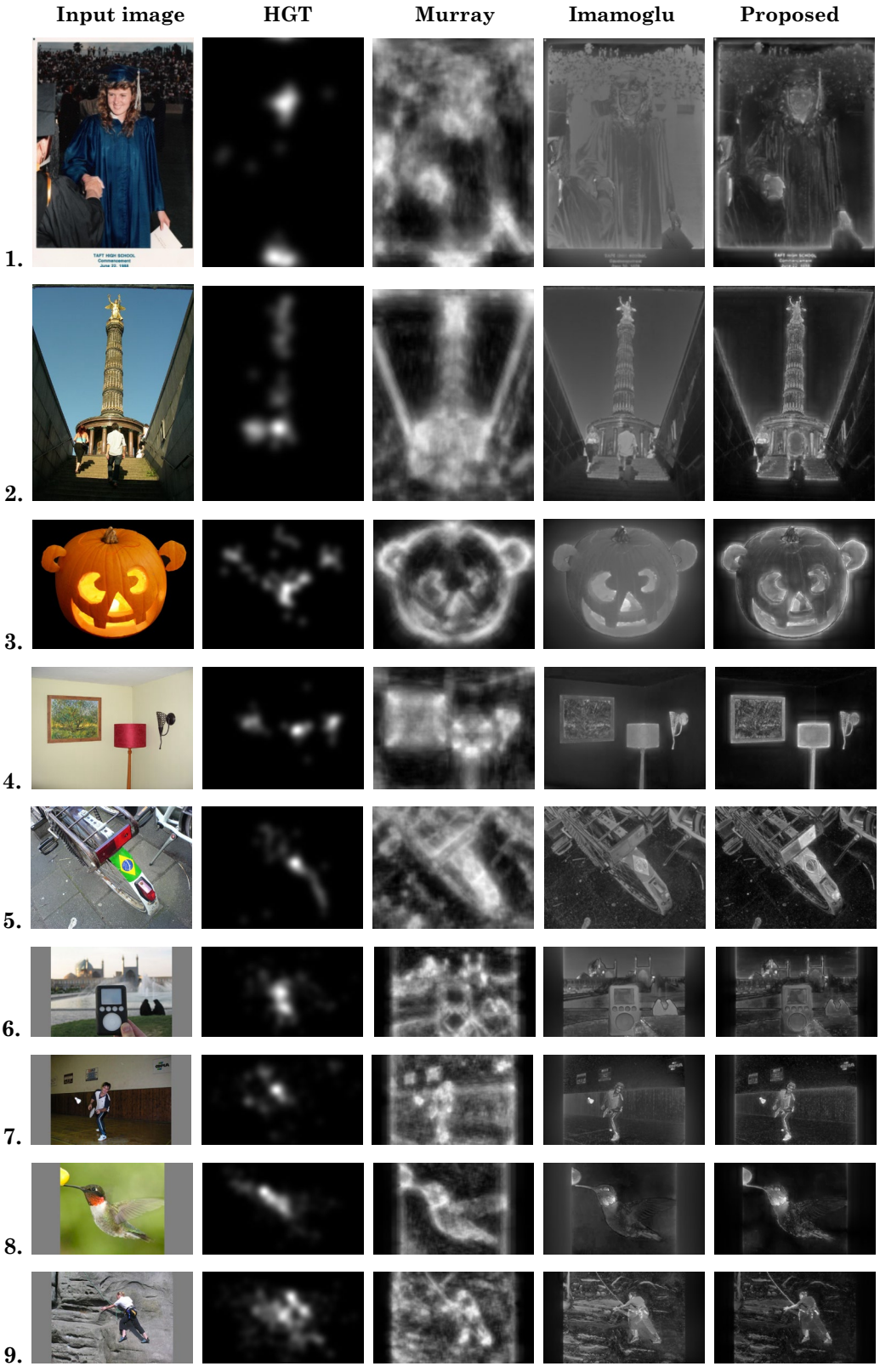

Figure 5.1 Qualitative comparison of saliency maps based on continuous distribution. Images 1 to 5 are obtained from MIT dataset [22] and 6 to 9 are obtained from CAT2000 dataset [46].

The sample images provided in the figure have been randomly selected from MIT and CAT2000 datasets. From the results shown in Figure 5.1, it is observed in the HGT map that the fixations in the first image are concentrated at the human face and text. The corresponding saliency map of the proposed model highlights most of the fixated regions such as face and text and it has a clear distinction between salient and non-salient regions with less false detections when compared to the models of Murray and Imamoglu. In the second image, the fixations lie vertically along with the structure of the monument and around the people. These regions are clearly highlighted in the saliency map of the proposed model when compared to the maps of Murray and Imamoglu. In the third image, the fixations are dispersed due to the shape of the object and illumination from inside the object. The proposed model highlights the regions around the object and illumination. The map of Murray highlights the edges and that of Imamoglu highlights illumination. The fixations for the fourth image are concentrated around the objects. The regions of these objects are clearly highlighted in the maps of the proposed model and Imamoglu when compared to Murray with non-distinguishable predictions. With the fifth image, the fixations are concentrated at the details of the object. The predictions are close in the map of Imamoglu when compared to the maps of the proposed model and Murray. The sixth image has fixations concentrated at the center due to the object in focus. The proposed model highlights the object in the center with other contrast variations. There is a clear distinction between salient and non-salient regions with the maps of the proposed model and Imamoglu. The seventh image has fixations concentrated at the face, text and object in action. This has been clearly highlighted in the maps of Imamoglu and the proposed model when compared to Murray. For the eighth image, the fixations lie around the structure of the bird. The model of Murray has provided close predictions when compared to the Imamoglu and proposed model. However, it still highlights noise. Finally, for the ninth image, fixations are concentrated around the person in action and around variations of the rock. The proposed model has achieved reduced false detections when compared to the models of Murray *et al.* and Imamoglu *et al.*

The results indicate that the saliency maps of the proposed model can highlight clear and distinguishable salient regions with reduced false detections compared to the saliency maps of Murray and Imamoglu. The results are further evaluated by comparing the saliency maps of different methods based on thresholding, obtained using Otsu's global threshold method [47]. This is illustrated in Figure 5.2. The threshold or binary map consists of salient regions (shown with white pixels) separated from the background

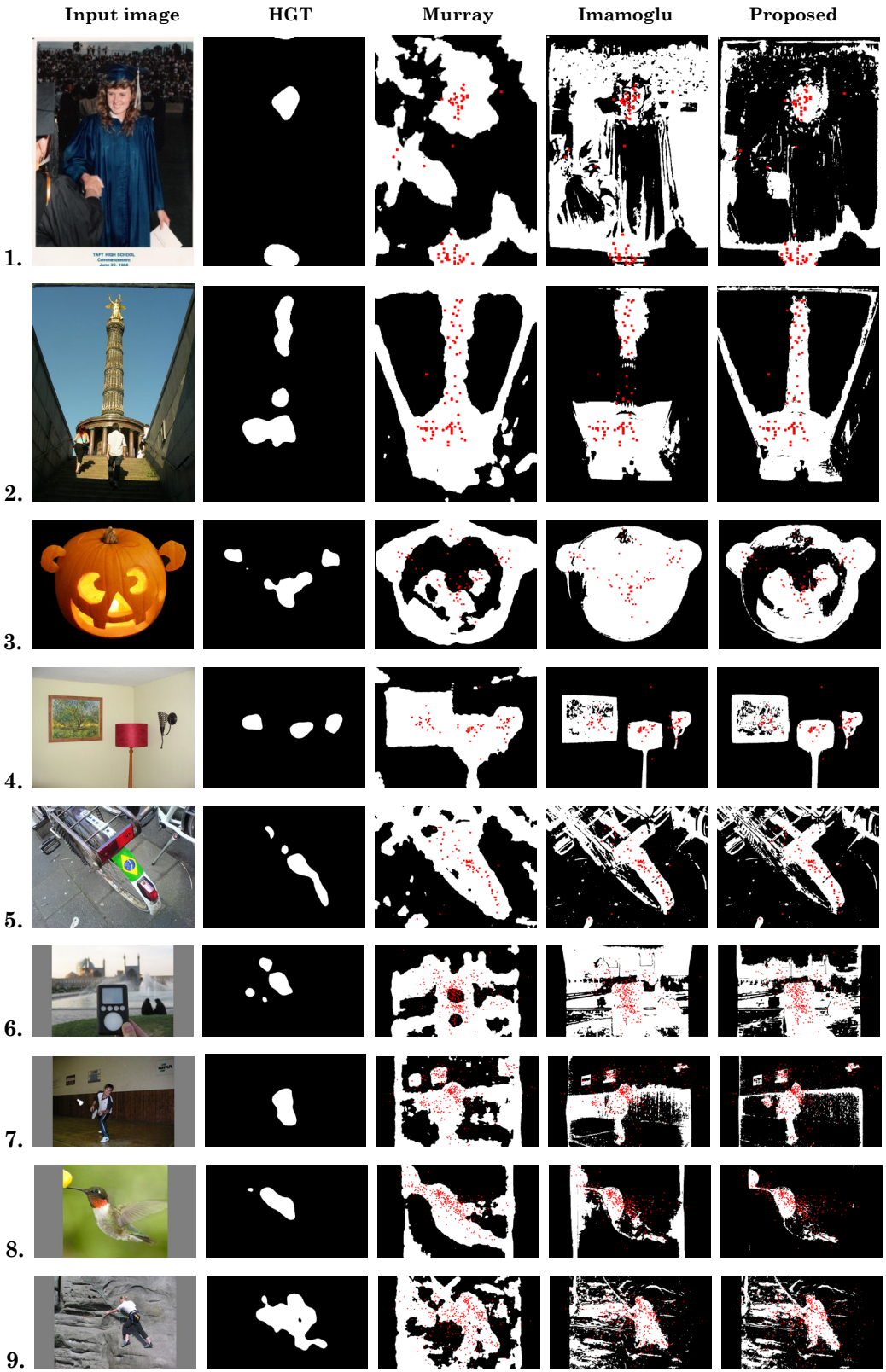

Figure 5.2 Qualitative comparison of saliency maps based on thresholding. The threshold maps are obtained using Otsu's global threshold method [47]. The red markers indicate the eye fixations.

regions (shown with black pixels) along with eye fixations indicated using red markers.

From the threshold maps of the first image, it is observed that the eye fixations coincides with all the models. However, the proposed model has provided reduced false detections. In the second image, the maps of Murray and the proposed model have achieved better predictions than the map of Imamoglu. For the third image, the map of Imamoglu covers all the fixations when compared to the maps of Murray and the proposed model but it highlights much of the non-salient regions. In the fourth and fifth images, all the models have predicted well with the maps of Imamoglu and the proposed model has highlighted reduced non-salient regions. With the sixth image, the predictions are better achieved in the maps of Imamoglu and the proposed model when compared to Murray. For the seventh and ninth images, the regions in the maps of Murray covers all the fixations but highlights non-salient regions when compared to the maps of Imamoglu and the proposed model. For the eighth image, only fewer fixations coincide with the maps of the proposed model and Imamoglu when compared to the map of Murray which covers all the fixations. The overall results show that the proposed algorithm is able to detect the relevant regions of fixations with respect to the HGT map. It can be observed from the results that the threshold maps from all three methods have correlation with the human eye fixations. But, the threshold maps of benchmark models highlight most of the non-salient regions (false detections) while this is reduced in the case of threshold maps of the proposed method. This ensures improvement in the quality of saliency maps obtained using the proposed method. In the next section, the proposed model is quantitatively evaluated.

5.3 Quantitative results

The proposed model is quantitatively evaluated using three metrics- Area under the ROC Curve (AUC), Pearson's Correlation Coefficient (CC) and Normalised Scanpath Saliency (NSS). Further, the performance of the models is measured by plotting ROC curves. A detailed description of the metrics is provided in chapter 3. These metrics are used to measure the prediction accuracy between fixations in the human ground truth (HGT) maps and the corresponding saliency maps of different models. The high scores of AUC, CC and NSS indicate the better performance of a model. The quantitative comparison of model saliency scores is illustrated in the bar graph as shown in Figure 5.3. The blue bar, red bar and green bar indicate the results of Murray *et al.,* Imamoglu *et al.* and proposed model respectively. For the MIT dataset, the proposed model has

achieved scores of 0.70 for AUC, 0.24 for CC and 0.83 for NSS; and for the CAT2000 dataset, it has achieved scores of 0.70 for AUC, 0.31 for CC and 0.77 for NSS. When compared with Murray *et al.*, the model has shown an improvement of 0% in AUC, 1% in CC and 5% in NSS and improvement of 1% in AUC, 3% in CC and 5% in NSS over MIT and CAT2000 datasets respectively. Similarly, when compared with Imamoglu *et al.,* the model has shown improvement of 3% in AUC, 4% in CC and 12% in NSS and improvement of 3% in AUC, 4% in CC and 8% in NSS over MIT and CAT2000 datasets respectively. The results show that for both MIT and CAT2000 datasets, the proposed model has outperformed in terms of CC and NSS while achieved similar or better scores in terms of AUC with respect to the benchmark methods.

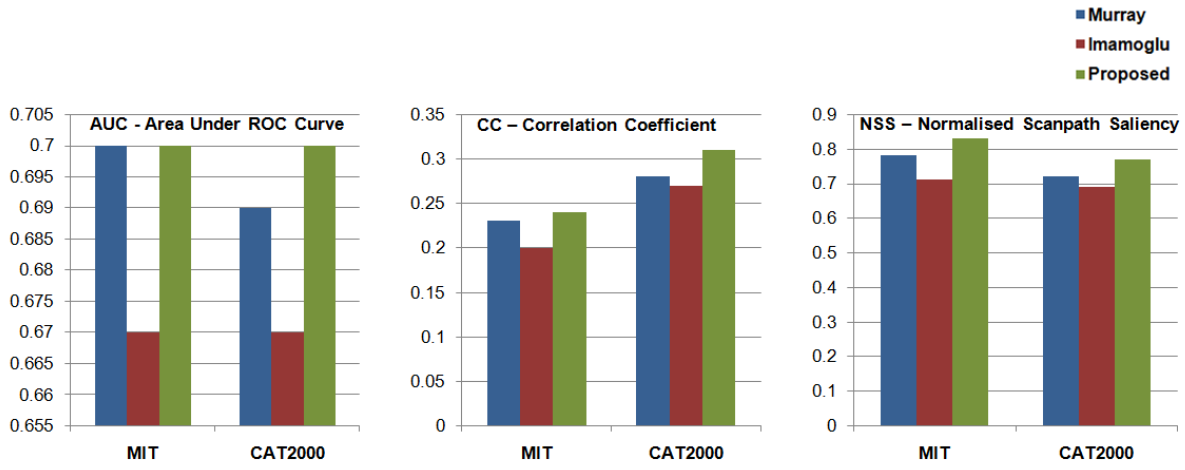

Figure 5.3 Quantitative comparison of saliency models.

Receiver Operating Characteristics (ROC) curve

The ROC curve is a two-dimensional graph obtained from the classification of true positive (TP) and false positive (FP) instances at various discriminative thresholds (refer to section 3.3.3). Figure 5.4 and Figure 5.5 shows the ROC plots for MIT and CAT2000 datasets indicating the performance of the saliency models. The curve with a blue line, red line and green line indicate the performance of Murray *et al.*, Imamoglu *et al.* and the proposed model respectively. The higher portion of the area under the curve indicates the better performance of the model. The ROC curve indicated using the green line shows that the proposed model has achieved similar AUC with respect to the model of Murray *et al.* and has achieved slightly better AUC with respect to the model of Imamoglu *et al.*

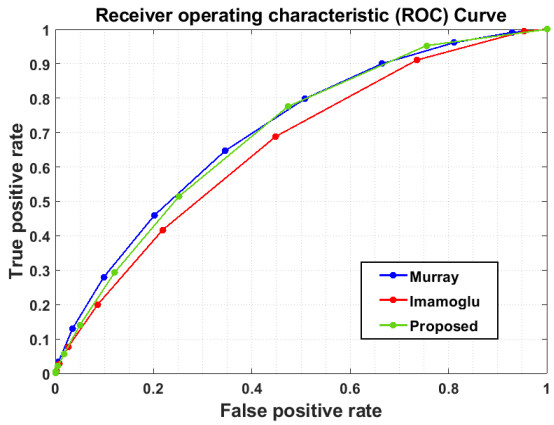

Figure 5.4 ROC plot for MIT dataset [22]

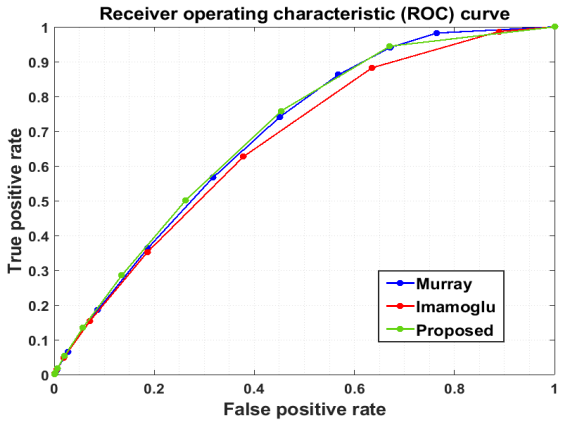

Figure 5.5 ROC plot for CAT2000 dataset [46]

5.4 Computation time results

As discussed in chapter 4, when compared to the existing wavelet-based saliency detection models, the proposed model requires only two colour channels, luminance (Y) and chrominance (Cr) for saliency computation. These channels were chosen because the channel combinations {Y, Cb, Cr} and {Y, Cr} have both achieved similar prediction accuracy results (explained in section 4.4.1), with {Y, Cr} saving nearly one-third of computational time by eliminating the chrominance channel Cb. Further to this, the average computational time per image required by different wavelet models has been evaluated over a test set of 100 random images obtained from MIT dataset [22] with resolution of 768x1024 pixels as shown in Table 5.1. However, the time cost calculation is independent of the dataset. The available open source MATLAB code for benchmark methods was obtained from online. The test environment with system specifications provided in Table 3.1 within MATLAB environment (with no parallel pool) was used. All the background applications were put to halt during evaluation. The average execution time over 25 iterations was obtained for each model using the MATLAB built-in timer functions - *tic* and *toc* [50], as explained in section 3.3.4. The results provided in Table 5.1 clearly indicates that the proposed model has achieved significant reduction in computational time when compared to the benchmark models. The proposed model contributes to a reduction of 91% computational time when compared to Imamoglu *et al.* [16] and nearly 25% when compared to Murray *et al.* [17].

Table 5.1 Evaluation of computation time of the saliency models. The results indicate the average computation time over 100 images with resolution of 768x1024 pixels.

Model	Time (sec)
Murray *et al.*[17]	2.45
Imamoglu *et al.*[16]	20.61
Proposed method	**1.84**

5.5 Summary

The experimental analysis of the proposed low-computation wavelet-based saliency detection algorithm has been presented in this chapter. The model is qualitatively and quantitatively evaluated over two public image datasets and compared with related benchmark models. The computation time of the proposed model has been evaluated and compared with the benchmark models. The experimental results show that the proposed model has achieved significant reduction in computation time (91% when compared to the model in [16] and 25% when compared to the model in [17]) and it has outperformed in terms of CC and NSS with similar or better performance in terms of AUC when compared to the benchmark models in [17] and [16]. This proves that the proposed model is efficient in terms of prediction accuracy and computation time compared to the benchmark models.

6 Conclusion and future work

6.1 Introduction

During visual scene analysis, the human visual system limits the processing of information by attending only the relevant regions of interest. Predicting these regions of human relevance through visual saliency models is important in many saliency applications. In real time, the saliency applications deal with high resolution images and videos. By employing the saliency models in these applications will limit the amount of data to be processed and reduces the need for additional computational resources. This will further improve the efficiency and computational performance of these applications. Therefore, a saliency detection model with good prediction accuracy and low computation time is desired.

This chapter provides the conclusion of the thesis, summarises the advantages and limitations of the proposed low-computation wavelet-based visual saliency model and indicates the future directions. The conclusion of thesis with a discussion on the achievement of the project objectives is provided in section 6.2. The advantages and limitations of the model are critically analysed in section 6.3. The future work of the project is discussed in section 6.4.

6.2 Conclusion

The aim of this research work was to develop a bottom-up computational visual saliency model to predict the regions of human eye fixations in images based on DWT. This developed algorithm predicts the regions with reduced computation time and improved prediction accuracy compared to the existing wavelet-based saliency detection models. The developed model has been presented in chapter 4 with corresponding experimental analysis has been discussed in chapter 5. Reflecting on the project objectives described in chapter 1, the achieved objectives are summarised as follows.

1. **Literature review of the computational visual saliency models**

   The literature of different approaches in computational modelling has been studied and critically reviewed. The study has been briefly described along with the theoretical and background knowledge provided in chapter 2.

2. **Performance evaluation of the benchmark visual saliency models**

   The relevant benchmark models have been identified as the works proposed in [17] and [16] which are tested using two large public image datasets MIT [22] and CAT2000 [46]. The models are qualitatively evaluated by analysing the saliency maps. The models are quantitatively evaluated using three metrics AUC, CC and NSS. Further, the execution speed of the models is evaluated over 100 images of resolution 768 × 1024 pixels. The experimental methodology used for model evaluation was presented in chapter 3.

3. **Development of a bottom-up visual saliency model using DWT**

   The proposed bottom-up visual saliency model based on DWT has been presented in chapter 4. The proposed work is different from the existing methods as it computes saliency using two-channel information (Y and Cr) in YCbCr colour space. This has achieved a significant computational savings when compared to the benchmark methods in [17] and [16]. Moreover, the model has introduced a two-dimensional entropy-based feature combination scheme which has provided better prediction accuracy when compared to the benchmark methods.

4. **Performance evaluation of the proposed model**

   The performance of the proposed model has been evaluated in terms of prediction accuracy and computation time with widely used benchmark image datasets and testing methods. The experimental results obtained are compared with the benchmark methods [17] and [16] and the results are discussed in chapter 5.

The aim and objectives of the project have been fulfilled through this thesis. The main contribution of this project is the development of a low-computation visual saliency model based on DWT. The experimental results obtained in chapter 5 shows that the model can operate on a significantly reduced time cost by using Y and Cr channels in YCbCr colour space while achieves similar or better prediction accuracy when compared to the benchmark models using a two-dimensional entropy based feature combination scheme. This work has been published as a conference paper [27]. The model has a potential to improve the performance of saliency applications in real-time. It can improve the efficiency of edge computing by processing only the relevant information and highly reducing the communication cost in vision based applications such as video surveillance and automated vehicle systems [52].

6.3 Advantages and limitations

The advantages and limitations of the proposed work are summarised as follows.

Advantages

- The proposed model operates with significantly reduced computation time and achieves similar or better prediction accuracy when compared to the benchmark saliency detection methods.
- It requires only two-channel information (Y and Cr) in YCbCr colour space for saliency computation when compared to the existing methods [16, 17, 25, 26] which require more than two colour channels. This saves approximately one-third of the computation time of the proposed model in wavelet domain. The model has shown to be consistent with the performance over two large image datasets MIT [22] and CAT2000 [46] which covers a wide range of image types such as natural and artificial, indoor and outdoor scenes consisting of people, objects, faces and text.
- The proposed model has no data or operation dependencies. Hence the model has a possibility of parallelisation through optimised DWT and IDWT implementation [53] in computing the saliency map and can be employed in real-time applications.

Limitations

- The qualitative results provided in chapter 5 of Figure 5.2 indicate that the saliency maps of the proposed model has provided reduced false detections (non-salient regions) compared to the benchmark methods. This ensures improvement in the quality of saliency maps obtained using the proposed method. However, the model still highlights some of the non-salient regions as salient. For example, in the first image, the hand shake is highlighted in the saliency map but the viewers have not fixated at these regions.
- In scenes with cluttered regions, fixations are large and distributed across the image as they require the most scrutiny by the observer [21]. These scenes are likely to have the highest feature content for which the proposed model is less accurate as it highlights all the regions with features as salient. This can be overcome by limiting the feature detection by including top-down influences such as centre-bias and fixation behaviour in search task.

- Although the proposed model has exploited local saliency information to provide satisfactory results, the performance can further be improved by including the global saliency information [12, 16].

6.4 Future work

The following future works are suggested for this project.

1. The number of wavelet scales employed for saliency analysis should be optimised as this directly influences the time cost of the model. This investigation was not carried out in this work as it was beyond the scope of this project. The optimisation of wavelet scales is assumed to greatly reduce the computation time of the model.
2. The future work should consider incorporating the global saliency information. This work should investigate the priorities of local and global saliency influences based on the context. The context of the local saliency is confined to a specific location while the context of the global saliency depends on the whole scene [2].
3. The model can be further extended to videos by including temporal correlation and motion cues to dynamically predict the saliency in video sequences.
4. The optimised implementation of the proposed method using GPU computing will enable the model to be used in real-time applications, like driver assistance systems [45].
5. The general directions for this work is to consider incorporating location based features such as centre-bias and top-down influences such as prior knowledge and task demands.
6. Deep learning based methods such as Convolution Neural Networks (CNN's) should be investigated for developing efficient visual attention models.

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, pp. 185-207, 2013.

[2] J. Li and W. Gao, *Visual Saliency Computation: A Machine Learning Perspective*. Switzerland: Springer International Publishing 2014.

[3] A. L. Yarbus, *Eye Movements and Vision*. New York: Plenum, 1967.

[4] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," *IEEE Transactions on Image Processing,* vol. 19, pp. 185-198, 2010.

[5] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient Object Segmentation via Effective Integration of Saliency and Objectness," *IEEE Transactions on Multimedia,* vol. 19, pp. 1742-1756, 2017.

[6] L. Zhang, Q. Sun, and Y. Sun, "Visual Saliency Analysis for Common Region of Interest Detection in Multiple Remote Sensing Images," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2316-2320.

[7] X. Yuan, J. Yue, and Y. Zhang, "RGB-D Saliency Detection: Dataset and Algorithm for Robot Vision," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2018, pp. 1028-1033.

[8] L. Itti, "Models of Bottom-Up and Top-Down Visual Attention," Doctor of Philosophy, California Institute of Technology, 2000.

[9] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597-1604.

[10] J. Chilukamari, "A computational model of visual attention," Doctor of Philosophy, Robert Gordon University, February 2017.

[11] S. Frintrop, *VOCUS: A visual attention system for object detection and goal-directed search* vol. 3899: Springer, 2006.

[12] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-Aware Saliency Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 1915-1926, 2012.

[13] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal Saliency detection using phase spectrum of quaternion fourier transform," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[14] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Neural Information Processing Systems,* pp. 545-552, 2006.

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

[16] N. Imamoglu, W. Lin, and Y. Fang, "A Saliency Detection Model Using Low-Level Features Based on Wavelet Transform," *IEEE Transactions on Multimedia,* vol. 15, pp. 96-105, 2013.

[17] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Computer Vision and Pattern Recognition*, 2011, pp. 433-440.

[18] 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, pp. 32-32, 2008.

[19] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1998.

[20] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, 2003, pp. I-253.

[21] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of vision,* vol. 7, pp. 4-4, 2007.

[22] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2106-2113.

[23] D. L. Fugal, "Conceptual Wavelets in Digital Signal Processing," ed: Space & Signals Technical Publishing, 2009, pp. 5-5.

[24] R. J. E. Merry, "Wavelet Theory and Application: A Literature Study, DCT 2005.53," Eindhoven University of Technology, 2005.

[25] X. Ma, X. Xie, K.-M. Lam, and Y. Zhong, "Efficient saliency analysis based on wavelet transform and entropy," *Journal of Visual Communication and Image Representation,* vol. 30, pp. 201-207, 2015.

[26] C. Scharfenberger, A. Jain, A. Wong, and P. Fieguth, "Image saliency detection via multi-scale statistical non-redundancy modeling," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 4294-4298.

[27] M. Narayanaswamy, Y. Zhao, W. K. Fung, and N. Fough, "A Low-complexity Wavelet-based Visual Saliency Model to Predict Fixations," presented at the IEEE International Conference on Electronics Circuits and Systems, Glasgow, United Kingdom, 2020.

[28] S. Treue, "Neural Correlates of Attention in Primate Visual Cortex," *Trends in Neurosciences,* 2001.

[29] R. C. Gonzalez and R. E. Woods. (2018). *Digital Image Processing (Fourth ed.).*

[30] D. Sundararajan. (2017). *Digital Image Processing: A Signal Processing and Algorithmic Approach*.

[31] *ITU-R BT.601-5: Studio Encoding Parameters of Digital Television for Standard 4:3 and Wide-screen 16:9 Aspect Ratios*, (1982-1986-1990-1992-1994-1995).

[32] *Rec. ITU-R BT.470-6 Conventional Television Systems*, (1970-1974-1986-1994-1995-1997-1998).

[33] V. Bruce, P. R. Green, and M. A. Georgeson, *Visual Perception: Physiology, Psychology and Ecology*: Psychology Press, 2003.

[34] I. E. Richardson, *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*: John Wiley & Sons, 2004.

[35] M. S. Nixon and A. S. Aguado, *Feature Extraction and Image Processing*: Newnes. Reed Elsevier PLC, Group, 2002.

[36] M. G. E. Schneiders, "Wavelets in control engineering," Masters, Eindhoven University of Technology, 2001.

[37] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 11, pp. 674-693, 1989.

[38] Dataset. [Online]. Available: http://saliency.mit.edu/results_mit300.html

[39] G. L. Shulman, R. W. Remington, and J. P. Mclean, "Moving attention through visual space," *Journal of Experimental Psychology,* vol. 5, p. 522, 1979.

[40] A. M. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cognitive Psychology,* vol. 12, pp. 97-136, 1980.

[41] A. Borji, "Interactive learning of task-driven visual attention control," Ph. D. thesis, Institute for Research in Fundamental Sciences (IPM), School School of Cognitive Sciences (SCS), Tehran, Iran, 2009.

[42] T. Judd, "Understanding and predicting where people look in images," Ph.D thesis, Massachusetts Institute of Technology, 2011.

[43] A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Machine Vision and Applications,* vol. 22, pp. 61-76, 2011/01/01 2011.

[44] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting Eye Fixations using Convolutional Neural Networks," *Computer Vision and Pattern Recognition,* 2015.

[45] J. Kim, S. Kim, R. Mallipeddi, G. Jang, and M. Lee, "Adaptive driver assistance system based on Traffic Information Saliency Map," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 1918-1923.

[46] A. Borji and L. Itti, "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency," *Computer Vision and Pattern Recognition,* 2015.

[47] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 9, pp. 62-66, 1979.

[48] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What Do Different Evaluation Metrics Tell Us About Saliency Models?," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 41, pp. 740-757, 2019.

[49] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters,* vol. 27, pp. 861-874, 2006/06/01/ 2006.

[50] MathWorks(online). Available: https://uk.mathworks.com/

[51] L. Bonnar, F. Gosselin, and P. G. Schyns, "Understanding Dali's Slave Market with the Disappearing Bust of Voltaire: A case study in the scale information driving perception," 2001.

[52] F. Al-Turjman and Al-Turjman, *Edge Computing*: Springer, 2019.

[53] H. Sava, M. Fleury, A. C. Downton, and A. F. Clark, "Parallel pipeline implementation of wavelet transforms," *IEE Proceedings - Vision, Image and Signal Processing,* vol. 144, pp. 355-360, 1997.

Bibliography

[1] R. C. Gonzalez and R. E. Woods. (2018). *Digital Image Processing (Fourth ed.).*

[2] V. Bruce, P. R. Green, and M. A. Georgeson, *Visual Perception: Physiology, Psychology and Ecology*: Psychology Press, 2003.

[3] M. Nixon and A. Aguado, *Feature extraction and image processing, Newnes*, 2002.

[4] L. Debnath and F. A. Shah, "Lecture Notes on Wavelet Transforms, Compact Textbooks in Mathematics," ed: Springer International Publishing AG 2017, 2017.

[5] S. Frintrop, *VOCUS: A visual attention system for object detection and goal-directed search* vol. 3899: Springer, 2006.

[6] D. L. Fugal, "Conceptual Wavelets in Digital Signal Processing," ed: Space & Signals Technical Publishing, 2009, pp. 5-5.

[7] F. Al-Turjman and Al-Turjman, *Edge Computing*: Springer, 2019.

[8] I. E. Richardson, *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*: John Wiley & Sons, 2004.

[9] Y. Zhao, "Complexity Management for Video Encoders," Doctor of Philosophy, Robert Gordon University, March 2004.

[10] F. Wai-keung, "Robot Behavior Learning with Adaptive Categorization in Logical Perceptual Space," Doctor of Philosophy, The Chinese University of Hong Kong, February 2001.

[11] N. Fough, "Design and Analysis of RTP Circuit Breaker for Multimedia Applications," Doctor of Philosophy, University of Aberdeen, December 2015.

[12] J. Chilukamari, "A computational model of visual attention," Doctor of Philosophy, Robert Gordon University, February 2017.

[13] A. S. Nagaraghatta, "Algorithms and Methods for Video Transcoding," Doctor of Philosophy, Robert Gordon University December 2018.

[14] J. Chilukamari, S. Kannangara, and G. Maxwell, "A low complexity visual saliency model based on in-focus regions and centre sensitivity," in 2014 IEEE

Fourth International Conference on Consumer Electronics Berlin (ICCE-Berlin), 2014, pp. 411-414.

[15] J. Chilukamari, S. Kannangara, and G. Maxwell, "A DCT based in-focus visual saliency detection algorithm," in 2013 IEEE Third International Conference on Consumer Electronics Berlin (ICCE-Berlin), 2013, pp. 1-5.

Appendix

Conference publication

M. Narayanaswamy, Y. Zhao, W. K. Fung, and N. Fough, "A Low-complexity Wavelet-based Visual Saliency Model to Predict Fixations," published at the IEEE International Conference on Electronics Circuits and Systems, Glasgow, United Kingdom, 23-25th November 2020.