A COMPUTATIONAL MODEL OF VISUAL ATTENTION

Jayachandra Chilukamari

A THESIS SUBMITTED TO ROBERT GORDON UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

February 2017

Abstract

A Computational Model of Visual Attention

Jayachandra Chilukamari

Submitted to Robert Gordon University in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Visual attention is a process by which the Human Visual System (HVS) selects most important information from a scene. Visual attention models are computational or mathematical models developed to predict this information. The performance of the state-of-the-art visual attention models is limited in terms of prediction accuracy and computational complexity. In spite of significant amount of active research in this area, modelling visual attention is still an open research challenge. This thesis proposes a novel computational model of visual attention that achieves higher prediction accuracy with low computational complexity.

A new bottom-up visual attention model based on in-focus regions is proposed. To develop the model, an image dataset is created by capturing images with in-focus and out-of-focus regions. The Discrete Cosine Transform (DCT) spectrum of these images is investigated qualitatively and quantitatively to discover the key frequency coefficients that correspond to the in-focus regions. The model detects these key coefficients by formulating a novel relation between the in-focus and out-of-focus regions in the frequency domain. These frequency coefficients are used to detect the salient in-focus regions. The simulation results show that this attention model achieves good prediction accuracy with low complexity.

The prediction accuracy of the proposed in-focus visual attention model is further improved by incorporating sensitivity of the HVS towards the image centre and the human faces. Moreover, the computational complexity is further reduced by using Integer Cosine Transform (ICT). The model is parameter tuned using the hill climbing approach to optimise the accuracy. The performance has been analysed qualitatively and quantitatively using two large image datasets with eye tracking fixation ground truth. The results show that the model achieves higher prediction accuracy with a lower computational complexity compared to the state-of-the-art visual attention models.

The proposed model is useful in predicting human fixations in computationally constrained environments. Mainly it is useful in applications such as perceptual video coding, image quality assessment, object recognition and image segmentation.

Index Terms- visual saliency, saliency detection, in-focus, DCT, frequency saliency, fixation prediction, attention, visual attention models, saliency model, face saliency.

Acknowledgments

I would relish taking this opportunity to sincerely thank everyone who contributed directly or indirectly to the successful achievement of this research work.

I came to Robert Gordon University to pursue postgraduate studies. During this time I had immense number of opportunities to learn and grow. I had useful discussions with many people who did their doctorates in various fields. I did my MSc project under Dr. Yafan Zhao with whom I had discussions regarding research and also received my first recommendation letter for a Ph.D position. I would like to thank her for giving me a challenging project which availed me draw the attention of my Ph.D advisor.

I also want to thank Dr. Christopher Macleod who asked me to meet Dr. Sampath Kannangara to discuss my research interests. My discussions with Dr. Sampath motivated me to apply for a Ph.D project under him. My profound gratitude goes to him for giving me an opportunity to undertake this interesting work. I am very grateful for the freedom he has given to me in choosing my own research question, constant support, constructive discussions, encouragement and guidance during the research work. I also thank him for taking time off his diligent schedule in SriLanka to give critical feedback for my thesis chapters.

I also thank Mr. Grant Maxwell who was a key member of supervision team during the research project. Dr. Yafan Zhao also joined the supervision team during the last year of the project. I had received many useful suggestions, comments and feedback from them. I am categorically thankful to them for critically reviewing my thesis to ameliorate the quality.

I thank my past and present colleagues Akshay, Anil, James Philips, Kaushal, Thomas and Vivek for their generosity, friendship and for the help I received during my studies. I particularly thank Thomas Guibert for his contributions during the software development of the research prototype.

I also wish to thank all my MSc classmates, staff members and volunteers from the university who took part in the subjective tests during the project.

I am deeply appreciative of the IDEAS Research Institute for funding the first three years of my Ph.D. studies.

I thank my parents Sathyanarayana Chilukamari and Sarojini for their unconditional love, guidance and sacrifices they made for me to have a good

education. They always kept faith in me and allowed me to be as ambitious as I wanted. It is under their watchful eye I gained so much drive and an ability to tackle the challenges. I also thank my brothers Gopi Ramana and Jithendar for giving me inspiration and courage since my childhood.

Finally, and most importantly, I attribute much of the success of this work to my wife Manjula for constantly supporting and encouraging me. I especially thank her for staying back in India to take care of my new born daughter, which in turn gave me an opportunity to focus on my research without distractions. Her unwavering love is undeniably the bedrock upon which the past four years of my life have been built. I also thank my little daughter Sonakshi Chilukamari for the jubilance and ecstasy she brought during my PhD years.

I dedicate this work to my parents

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

2D	Two Dimensional
ACR	Absolute Category Rating
AUC	Area Under Receiver Operating Curve
BIQI	Blind Image Quality Index
BRISQUE	Blind/Reference less Image Spatial Quality Evaluator
CAS	Context Aware Saliency
CC	Correlation Coefficient
CIF	Common Intermediate Format
CPBD	Cumulative Probability of Blur Detection
CPU	Central Processing Unit
DC	Direct Current
DCT	Discrete Cosine Transform
DOG	Difference Of Gaussian
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
FFMPEG	Fast Forward Moving Pictures Expert Group
FIT	Feature Integration Theory
FMRI	Functional Magnetic Resonance Imaging
FOA	Focus of Attention
FPR	False Positive Rate
FR	Full Reference
FT	Fourier Transform
GBVS	Graph Based Visual Saliency
GTFM	Ground Truth Fixation Maps
HD	High Definition

HEVC	High Efficiency Video Coding
HGT	Human Ground Truth
HVS	Human Visual System
ICT	Integer Cosine Transform
IFC	Information Fidelity Criterion
ITU	International Telecommunication Union
IWT	Inverse Wavelet Transform
JNBM	Just Noticeable Blur Metric
JND	Just Noticeable Distortion
JPEG	Joint Photographic Experts Group
LCC	Linear Correlation Coefficient
LGN	Lateral Geniculate Nucleus
MA	Multidimensional attribute
MAE	Mean Absolute Error
MATLAB	Matrix Laboratory
MIT	Massachusetts Institute of Technology
MORSEL	Multiple Object Recognition and Attentional Selection
MOS	Mean Opinion Score
MSE	Mean Squared Error
MSSIM	Multi-Scale Structural Similarity
NIQE	Naturalness Image Quality Evaluator
NQM	Noise Quality Measure
NR	No Reference
NSS	Natural Scene Statistics
NTIA	National Telecommunications and Information Administration
NVT	Neuromorphic Vision Toolkit
OPENCV	Open Source Computer Vision

PC	Pair Comparison
PDF	Probability Density Function
PFT	Phase Spectrum of Fourier Transform
PQFT	Phase Spectrum of Quaternion Fourier transform
PSNR	Peak Signal to Noise Ratio
QP	Quantisation Parameter
RAM	Random Access Memory
RCSS	Random Centre Surround Saliency
RGB	Red-Green-Blue
RMSE	Root Mean Squared Error
ROC	Receiver operating Curve
ROI	Region Of Interest
RR	Reduced Reference
RRED	Reduced Reference Entropy Differencing
SAIM	Selective Attention for Identification Model
SC	Superior Collicus
SDSP	Saliency Detection using Simple Priors
SLAM	Selective Attention Model
SNR	Signal to Noise Ration
SR	Spectral Residual
SROCC	Spearman Rank Order Correlation Coefficient
SS	Signature Saliency
SSCQE	Single Stimulus Continuous Quality Evaluation
SSIM	Structural Similarity Index
SUN	Saliency Using Natural Scene Statistics
TPR	True Positive Rate
UQI	Universal Quality Index
V1	Visual Area One
VIF	Visual Information Fidelity
VQEG	Video Quality Experts Group

VQM	Video Quality Metric
VSNR	Visual Signal to Noise Ratio
WBSD	Wavelet Based Saliency Detection
WTA	Winner-Take-All

PART ONE: BACKGROUND

1 Introduction

1.1 Problem Statement

Human vision provides a wealth of information to the brain from the outside world. It has the ability to create a coherent global visual experience from noisy, sparse and ambiguous environments. It perceives thousands of objects, identifies hundreds of faces and appreciates beauty all around. Computer vision is a science which aims to understand the visual world that human vision perceives through processing, extracting and analysing information present in images. Computer vision is used in wide variety of applications such as object recognition, image segmentation, image/video editing and enhancement and perceptual video coding.

In the past decade, there has been a significant amount of research in the field of computer vision systems. These systems often deal with high resolution images which result in increased computational complexity and thus make it difficult for them to operate in real time. The performance of these systems can be enhanced by processing the relevant information present in the images and ignoring the irrelevant information. By selecting the relevant information in an image the amount of information that needs to be processed can be greatly reduced.

The relevant information in an image is typically detected by using a visual attention/saliency model which computationally models the important features present in images. They generally model bottom-up and top-down features for detecting salient regions. Bottom-up features are distinct and grab the viewer's attention towards them. Intensity, colour and orientation are the main bottom-up features modelled in the literature [1]. In addition to these bottom-up features, there are also top-down features that drive visual attention. These features are mainly user driven and are influenced by cognitive factors such as motivation, knowledge, desires, expectation and goals of the user [1]. Some of the key top-down features include scene context [2], [3] and task demands [4], [5].

The performance of the visual attention model greatly affects the computer vision system's operation. A highly accurate visual attention model will enable a computer vision system to achieve its desired results. Further, a low complexity visual attention model will enable the computer vision system to operate in real time with greater ease. Therefore, it is important to develop visual attention models that have the ability to detect salient regions with high prediction accuracy and have low computational complexity.

In the literature, the state-of-the-art computational models of attention have mostly focussed on modelling bottom-up features for detecting salient regions [6], [7], [8], [9], [10], [11]. They have modelled bottom-up features more than top-down features because of ease and simplicity of deriving these features from the images [1]. Furthermore, these models have used approaches with high complexity to model the salient image features [6]. Some of the models have used a greater number of feature channels with an aim of improving prediction accuracy [12]. Although these models managed to achieve better accuracy, both the higher number of features and complex approaches resulted in an increase of computational complexity of these attention models. Moreover, some of the low complexity attention models proposed in the literature have shown performance drop in prediction accuracy [8], [13]. Existing visual attention models have either achieved better prediction accuracy with high complexity or low prediction accuracy with faster operation at detecting salient regions in the images. This has limited the practical application of these models [11]. Therefore, visual attention models that have been developed to date lack the ability to achieve good prediction accuracy with low computational complexity.

Visual attention models with better accuracy and low complexity are especially required in applications such as perceptual video coding, image/video quality assessment, object recognition and image segmentation, where there is large amount of irrelevant information in images which needs to be filtered more efficiently with low computational complexity. Therefore, there is a need for a novel computational model of visual attention with better prediction accuracy and low complexity.

1.2 Research Aim

The aim of this work is to develop a novel computational model of visual attention to predict salient regions in the images. The research mainly addresses the prediction accuracy and computational complexity issues of the existing visual attention models. The developed model can be effectively used to improve the performance of computer vision systems.

1.3 Research Objectives

The research aim is achieved through a preliminary study and key objectives. During the initial study the state-of-the-art visual attention models available in the literature have been critically analysed by identifying their advantages, disadvantages and interesting aspects. The performance of these attention models has been analysed qualitatively and quantitatively with the image datasets available in the literature. Later

the state-of-the–art is advanced through some of the key objectives of this research project. A complete list of these objectives is given below.

1. Study the state-of-the-art visual attention models available in the literature to gain theoretical knowledge. Further, critically analyse them and empirically evaluate their performance.

2. Develop a novel bottom-up visual attention model. The attention model should detect salient regions in the images accurately with low computational complexity.

3. Further develop the bottom-up visual attention model (developed in the second objective). Manage the computational complexity and improve the prediction accuracy by modelling high level features present in the images.

The objectives of this project are fulfilled by developing novel approaches for computational modelling of bottom-up and top-down visual attention. Their strengths and limitations are also discussed based on empirical evaluation.

1.4 Main Contributions and Publications

During the project, a novel computational model of attention which has good prediction accuracy with low computational complexity is proposed. The main contributions of this work to the advancement of visual saliency area is summarised below.

- The development of a DCT based visual attention model for predicting salient regions in images. The model considers in-focus regions in the images as visually interesting and captivating. To develop the model, an image dataset is created which has different types of images with in-focus and out-of-focus regions. This dataset is mainly used for hypothesis generation for detecting in-focus regions in the images. The visual attention model developed detects the in-focus regions using the characteristics of DCT coefficients. This work was published in a conference paper [14].

- The computational complexity of the DCT based focus detection model is improved and combined with a location based top down feature known as image centre sensitivity to improve the overall prediction accuracy of the model.

This work was published in a conference paper [15] and also achieved the best paper award.

- The development of high prediction accuracy attention model by combining the low complexity attention model with a human face saliency map. The model is parameter tuned using a hill climbing approach to optimise its performance. Further, a dispersion measure (standard deviation) is calculated to estimate the model's performance across each image. This method is used in conjunction with the chosen quantitative analysis metrics to determine the consistency of the attention model across different image statistics within the image datasets. This work is submitted for publication in a journal.

- Investigation of the effectiveness of the objective video quality metrics such as Full Reference (FR), Reduced Reference (RR) and No Reference (NR) in detecting perceptual quality variation induced by pre-processing filters. Although this contribution is slightly outside the main research theme, this investigation has helped in analysing the ability of the existing video quality metrics. This work was published in a conference paper [16].

1.5 Organisation of the Thesis

The organisation of the thesis is as follows:

- **Chapter 2 -** This chapter provides an overview of human and machine visual attention. It provides the background knowledge related to how the visual information is captured and processed by the human brain. The fundamental concepts and terms used in the computational modelling of visual attention are introduced. Some of the early visual attention models have been explained. The main applications of computational attention models are also discussed.

- **Chapter 3** -This chapter presents the different types of computational attention systems existing in the literature. Further, it gives a critical review of the most closely related saliency models to the proposed visual saliency model.

- **Chapter 4** - The experimental methods used for the research project are explained in this chapter. The development and testing platform and image datasets used for the development of the attention models have been explained. The qualitative and quantitative assessment techniques used for

evaluating models accuracy and the methods used to measure the computational complexity are introduced. The benchmark models of visual attention chosen for the current work are outlined.

The main contributions of this research project are described in chapters 5, 6 and 7. A novel computational model of visual attention for predicting salient regions in the images is proposed in the chapter 5 and chapter 6. In the chapter 7 a study is carried out to identify suitable objective video quality metrics for the development of perceptual quality algorithms.

- **Chapter 5 -** Describes a new visual attention model for detecting salient regions in the images. It is assumed that the viewers are highly attracted towards the in-focus regions in images. Therefore, in-focus regions are detected using the characteristics of DCT coefficients. This is the main attention model developed for detecting salient regions and further developments include improving this model and integrating it with other developed algorithms.

- **Chapter 6** - The attention model in chapter 5 is further developed by incrementally innovating it by improving some of the core components of the model. Further, new algorithms are developed to detect the image centre and to generate human face maps. These are integrated with the main focus detection attention model and optimised to improve the overall model's performance in terms of prediction accuracy and computation complexity.

- **Chapter 7** – The effectiveness of the existing Full Reference (FR), Reduced Reference (RR) and No Reference (NR) video quality metrics in detecting the quality variations in pre-processed and coded videos is studied in this chapter. Although this chapter is not directly related to the main research theme, the study identifies objective video quality metrics that can be used during the development of perceptual quality algorithms.

- **Chapter 8** - This chapter summarises the main developments and experimental results related to the objectives of this research work. The conclusion of the thesis is provided and the possible future directions are indicated.

- **Appendix A –** Contains a list of publications related to this research work.

- **Appendix B –** Contains the details of software implementation of the proposed visual attention model. The software prototype is developed in C++ programming language using OpenCV and FFmpeg libraries. It can detect salient regions within the images, fed live from an external HD camera and H.264 encoded videos.

2 Visual Attention

2.1 Introduction

A rich stream of visual data (10^8-10^9 bits) enters the human eye every second [17], [18]. This process of acquiring visual information from the environment is continuous and processing this data in real time is an extremely difficult task for the human brain. To compensate for the inability of handling this enormous amount of information, the human brain classifies the information into two categories. The first one is the relevant visual information that is selected for further processing by the human brain. The latter is the irrelevant information that can be filtered out. This process of selection and prioritisation of the visual information is known as selective visual attention [19]. Therefore, human visual attention can be defined as the process of selectively reducing the incoming visual information to match the capacity of the human brain.

Machine visual attention is basically the ability of the computers to see and perceive objects in a similar way to the humans. A machine vision system or a computational attention model recovers important information from a scene from its two dimensional projections [20]. Images are usually two dimensional projections of the three dimensional world. A machine vision system creates a model of the real world from these images. Therefore, a machine visual attention is a technology that aims to imitate human visual attention. This chapter deals with the fundamentals of human and machine visual attention. Moreover, it provides some of the important findings that permitted better understanding of human attention. The study of human visual attention gives the required knowledge to develop a novel computational model of human attention which is described in chapter 5 and 6.

Section 2.2 of this chapter gives an overview of human visual attention. Later the structure of Human Visual System (HVS) is illustrated. It explains how the information is received and processed by human brain. The important areas of the human brain that play a vital role in visual perception are discussed. The eye movements and the factors that drive attention mechanism are detailed. In section 2.3 psychophysical theories on visual attention are presented. Section 2.4 of this chapter provides the background related to machine visual attention. It depicts the basic architecture of a computational model of human attention. Further, it briefly explains some of the important applications of these models in the area of computer vision and image/video compression.

2.2 Overview of Human Visual Attention

As explained earlier, in reality a very small amount of information is processed by HVS. In order to demonstrate this phenomenon, the authors of [21] performed an experiment known as ***change blindness***. In this experiment, changes are introduced in the image or visual stimulus by transforming their enduring coherent structure. In spite of these changes the viewer has failed to notice the difference as the human eye is insensitive to these areas. This experiment has shown that the visual system processes only limited visual information at any point of time. In another experiment by the authors of [3], one person approaches a pedestrian and asks about directions. During the conversation two people with a door in their hands passes in between the experimenter and the pedestrian. At the time of interruption the first experimenter is replaced by a second experimenter. Even though the second experimenter wears different clothing the pedestrian does not notice the difference. During these experiments about 50% of the subjects have failed to detect the change of person. This indicates that human visual attention is highly selective in nature.

During normal vision it is impossible to perceive two objects co-instantaneously in the same sensory act [22]. Although the tendency of the HVS is to retain a very rich representation of the visual world, in reality at each moment only a very small region under human attention is analysed [1]. During the period of attention the human eyes gets fixated over the region of interest and simultaneously many other regions of the scene are ignored. One important theory which explains how human eyes shift from one Region Of Interest (ROI) to another is the moving-spotlight theory [23]. The human visual attention is considered as a spotlight in a dark room. The spotlight illuminates the intended targets and moves on to the next region of interest in a serial fashion.

Visual attention is an interdisciplinary field of study that is closely related to psychology (a scientific study of mental functions and behaviour) and neurobiology (a branch of biology that deals with the study of the nervous system). These are the disciplines whose research is effectively focussed in this area. Psychologists develop psychophysical theories and models by an extensive investigation of human behaviour on special tasks in order to understand the internal processes of the brain [1]. These theories or models explain the relationship between the stimulus and the perceptual sensation in a quantitative way [24]. They study the subject's experience by systematically varying the stimulus properties [25]. Neurobiologists use techniques to visualize the areas of the brain which are active under certain conditions [1]. Functional Magnetic Resonance Imaging (FMRI) is one technology which is used to take a direct view of the brain. The findings from the psychology and biology can be utilised by the

researchers in the field of computer vision to develop new software, technical systems and standards. Therefore, this interdisciplinary research around visual attention has helped both human and machine vision communities.

2.2.1 The Human Visual System (HVS)

The HVS refers to the human eye and the brain working together in liaison to process visual information. The HVS performs many image processing tasks vastly superior to present day super-fast computers. To imitate the mechanisms of HVS computationally, a thorough understanding of the HVS is needed. The study of HVS helps to understand how the human eye manages to selectively attend relevant information in the visual scene.

The human eyes are the sensory organs that act as the input to the HVS. They capture the light and project it on to the retina. The visual information is then transmitted to optic chiasm through the fibres of the optic nerve. From there, there are two pathways which lead to two different brain hemispheres. These are collicular pathway leading to Superior Colliculus (SC) and the retino-geniculate pathway that leads to Lateral Geniculate Nucleus (LGN) as shown in Figure 2.1.

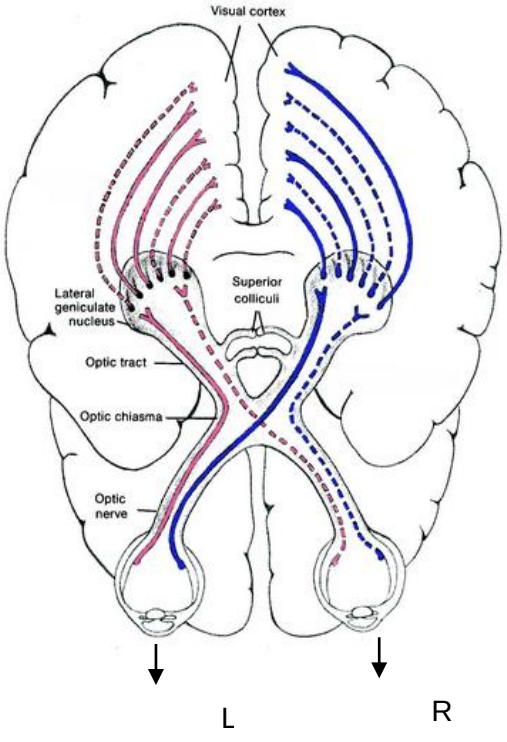


Figure 2.1: Structure of Human Visual System (HVS) (source [26])

The retino-geniculate pathway transmits 90% of the information whereas the collicular pathway is responsible for only 10% of the captured visual information [1]. The information from the LGN is transferred to the visual cortex. The superior collicus (SC) receives visual inputs from the retinal and primary visual cortex. The collicular pathway transmits very limited visual information. However, it plays a vital role in controlling visual attention and eye movements [27] .

The visual cortex is mainly responsible for processing visual information. It is a hierarchical arrangement with the primary visual cortex known as visual area one (**V1**) at the beginning of the hierarchy. Around 50% of the area of V1 processes information obtained from the fovea [28] . The other visual areas are namely visual area two (**V2**), three (**V3**), four (**V4**), and five (**V5**). The area V1 receives information from the LGN and is transmitted to two primary pathways, known as the dorsal and ventral stream. The dorsal stream as the name indicates it lies dorsally (upper side or back of an organ) and is associated with motion, depth perception and passes information to the motion sensitive parts of the visual cortex [29]. The ventral stream located at the ventral (lower side of an organ) part of the body is associated with perception of shapes and object recognition passes through V4 [29]. Each of these visual areas is sensitive to different types of visual information. All the areas and their corresponding associated functions are shown in the Table 2.1.

Table 2.1: Visual areas and functional specification

Visual Areas	Function
V1,V2 [29]	Line orientation, spatial frequency and colour
V3 [29], [30]	Global motion
V4 [29]	Shape and texture discrimination
V5 [29], [31]	Perception of motion

2.2.2 Eye movements

Humans continuously move their eyes to track the visual stimuli. The different types of eye movements are briefly explained below.

Fixation: This is achieved when the human gaze is stationary around a single location [32].

Saccade: These are fast, rapid movements of eyes [33]. The main function of these saccades is to change the fixation point from one location to another. They direct the high resolution fovea on to the region of interest for high acuity (ability of the observer to perceive high contrast spatial information) analysis. The amplitude of these movements ranges from small to large, such as reading a newspaper to gazing around the room. These saccades are either voluntary or involuntary. During a saccade, the high velocity of the retinal image leads to blurring of everything that falls in the field of vision. Therefore, the vision is usually suppressed and the information is acquired only during the fixation.

Microsaccade: These are small involuntary movements which occur when eyes fixate on a location [32], [34]. They typically occur during prolonged fixation. These are usually unnoticed and cannot be produced by an observer at will.

Vergence: During vergence eye movements, the fovea of both eyes is drawn to a single location. The eyes actually rotate in opposite directions (the right eye to left and the left to right) to converge on to the object or region of interest [32], [34].

Smooth pursuit: These are voluntary slow and smooth movements of the human eyes that help to keep the moving stimulus on the fovea [34].

Scanpath: A scanpath is sequence of eye movements which involve fixations, smooth pursuits and saccades [35].

2.2.3 Overt and Covert Attention

Generally there are two types of attention mechanisms, namely overt and covert attention. These attention types are explained below.

Overt attention: During overt attention the body, head and eyes are oriented to foveate (perceive with higher detail) a stimulus. This is an involuntary attention mechanism which involves eye movements.

Covert attention: Covert attention does not involve eye movements. During covert attention there is neither the movement of head nor the movement of eyes. Attention is voluntarily achieved using the peripheral part of the human eye. These fixations are not observable and are generally made using the corner of one's eye. For

example, when a football player continuously fixates his eyes on the football his covert attention may shift to a goal post. The eyes continue to remain focused on the previous object attended, yet attention is shifted. Another example is when a person drives on the road he overtly keeps his eyes on the road and simultaneously monitors the road signs and traffic lights using covert attention.

Overt attention shifts occur when eyes move overtly from one location to another location. Before the overt attention comes into play covert attention shifts to the locations that are going to be attended because of the thought process and hence covert attention drives the overt attention. Therefore, covert attention is much faster when compared to overt attention.

In the literature overt attention is more extensively studied than covert attention as it can be easily measured using eye trackers [36]. Although Posner [37] proposed few methods to compute covert attention, the behavioural mechanisms and its functions are still unknown. So far there is no proper system available for measuring covert attention [36].

2.2.4 Factors Influencing Visual Attention

As early as 1967, Yarbus studied the relationship between the saccades and visual attention [32]. These saccadic eye movements have been extensively studied in the literature [38]. The covert attention initially scans the field of view to determine an interesting location [36]. The most interesting ones among the examined targets are retained and the HVS sets up a saccade to that target using overt attention. The loss of visual acuity is compensated by a succession of rapid eye movements (saccades) [39]. While examining the targets there are two major factors that influence the human eye in selecting the targets. These are bottom-up and top-down factors [40]. These factors that drive human attention are briefly explained below.

2.2.4.1 Bottom-up Factors

Bottom-up attention is a fast, memory independent process driven by the properties of the visual stimuli. It is mostly unconscious, often reactive and comes into play during free viewing conditions. The HVS is involuntarily attracted to these regions. There will be many regions of interest which actually leap out of the scene to grab the viewer's attention. This attention mechanism is also referred as exogenous, automatic, reflexive or peripherally cued [41].

Visual search and pop out effect: In 11th century Ibn Al-Haytham found that "some of the particular properties of which the forms of visible objects are composed appear at the moment when sight glances at the object, while others appear only after

scrutiny and contemplation" [42]. This phenomenon is known as the ***pop-out*** effect. The targets during the visual search leap out to grab the viewer's attention.

Visual search can be classified as a bounded or unbounded search. In a bounded search the target to be searched is known in advance whereas it is unknown in an unbounded search. The search process in both of these classifications can be either efficient or inefficient. The efficiency is usually measured as a function of reaction time. It is the time a viewer needs to detect the target among the distracters. The lower the time, the better is the efficiency.

A bounded visual search example is shown in Figure 2.2 (a). The pop-out effect occurs in this example as the distracters are homogeneous in nature (The black vertical line pops out among all the white vertical lines). Therefore, the visual search in this example is significantly efficient because of a lower reaction time.

In the real world scenario searches for stimuli are not defined by a single property. They are usually defined by conjunction of two or more properties. Therefore, this type of visual search is known as conjunctive visual search. A conjunctive visual search example is shown in Figure 2.2 (b). The visual search in this example is less efficient when compared to the previous example (Figure 2.2 (a)). In the Figure 2.2 (b), it can be seen that the target is black vertical line and it is searched among the distracters that are both black horizontal and white vertical lines. This is in contrast to the pop out effect where the distracters are completely homogeneous (white vertical lines). As the target is defined by more properties the viewer needs more time in detecting the black vertical line among the distracters. Therefore, in this scenario the search process is less efficient when compared to the pop-out effect.

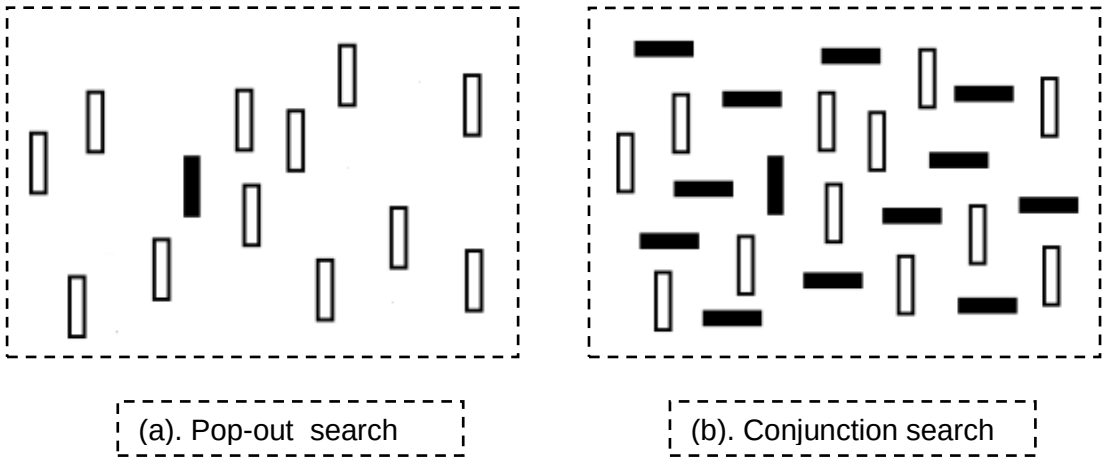


Figure 2.2: Visual search. (a) The black vertical line pops among the distracters (white vertical lines) by a unique visual property. (b) The black vertical line differs from the distracters (black horizontal and white vertical lines) by a conjunction of properties (source [43])

The pop-out search and conjunctive visual search are two different scenarios of bottom-up processing. In psychological experiments it is shown that, during a bounded bottom-up visual search, the search time is mostly linear and not exponential [1].

2.2.4.2 Top-down Factors

Top down attention is a slow, memory dependent process driven by cognitive factors such as knowledge, expectations and goals of the user [44]. This is also called voluntary [45], endogenous [37] or centrally cued attention [1]. For example, whenever a person is driving and wants to find a petrol station then only petrol stations on the way are going to attract his attention. An important feature of top down attention is given the same scene; the attended regions change depending upon the observer's tasks. Yarbus [32] in one of his famous experiments recorded fixations and saccade patterns of observers while viewing objects and scenes. One of the examples of his work is shown in Figure 2.3.

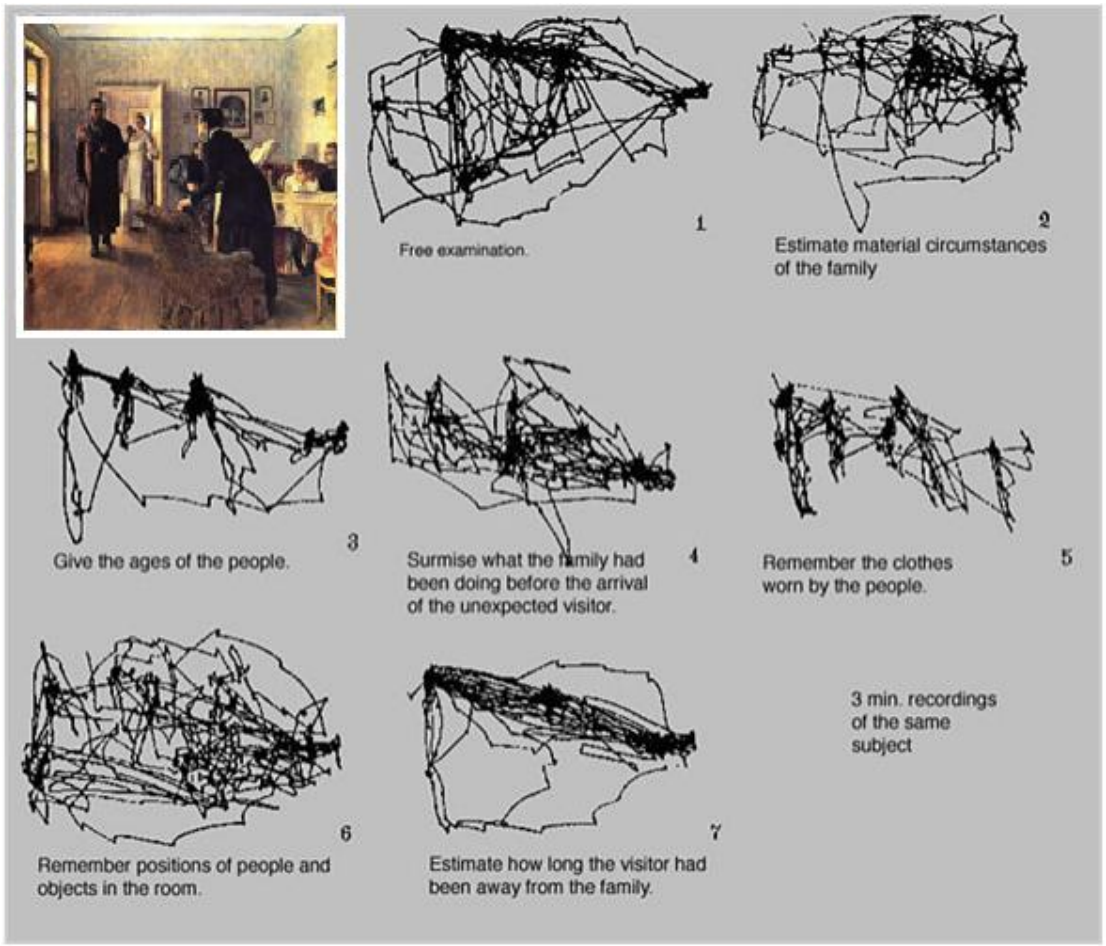


Figure 2.3: The effect of task on human scan path (source [32])

Initially an image is shown to the viewers. Then the viewers were asked different questions to study the impact of task on human attention. For each different type of task the scan paths were recorded and analysed. He showed that saccade patterns varied for different types of questions viewers were asked prior to viewing. Generally there are many different kinds of top down influences such as prior knowledge about the target [1], expectations [46], emotions [46], desires [46] and motivations [1]. Wells and Matthews [47] studied attention and emotions from a psychological perspective. Based on the interplay of attention and emotions in the human brain, Fragopanagos and Taylor [48] developed a neurobiological model of attention.

As discussed in an earlier section, the visual attention effect is observed in all the different areas of visual cortex. There is significant evidence which has shown that top-down signals are generated outside the visual cortex and are transmitted via feedback connections to visual cortex [49].

The neurophysiological studies have indicated that there are two independent but interacting brain areas associated with these two kinds of attention mechanisms [50]. However, very little is known regarding the interaction between these two kinds of mechanisms [1].

2.2.4.3 Bottom-up vs. Top-down Attention

When someone is fully engrossed in reading a newspaper, and if someone walks beside him, the attention is immediately shifted to the walking person. Similarly, when an emergency bell rings in a shopping mall the attention is immediately shifted in spite of the individual top-down influences. According to Theeuwes [51] bottom-up influences are not voluntarily suppressible. In his experiments, he gave the participants a task of searching a diamond shape in two different displays as shown in Figure 2.4. Although the participants knew colour had no significance in the search task, the red colour circle slowed down the visual search of the participants by about 65ms (885 vs. 950 ms) [51]. This clearly indicates that colour pop-out captures the visual attention independent of the top-down influences such as task.

The authors of [52] critiqued Theewes assertion and clarified that this automaticity does not apply to all stimuli impinging upon the retina, but only those that fall inside an attention window. All objects within the window compete among themselves and the most important target receives the attention. Objects outside the window, however, do not necessarily compete for selection and hence can be ignored. Another stipulation of Theewes' theory is the attention window is dynamic over the visual field. When an observer initiates a very difficult search task, the attention window is very small encompassing only two to three objects at a time. Therefore, though a

distracter is present in the display, it would not receive a noticeable degree of attention because

1) It would not impinge on the retina when it is outside the window,
2) It would fall inside the window only for a fraction of trials, because in some cases the visual search may be terminated before this small window is moved onto the distracter,
3) Even if it falls inside the window it may not cause interference as the window is small consisting of only one object. When there is only one object inside the window, there is no possibility of competition for the distracter to win.

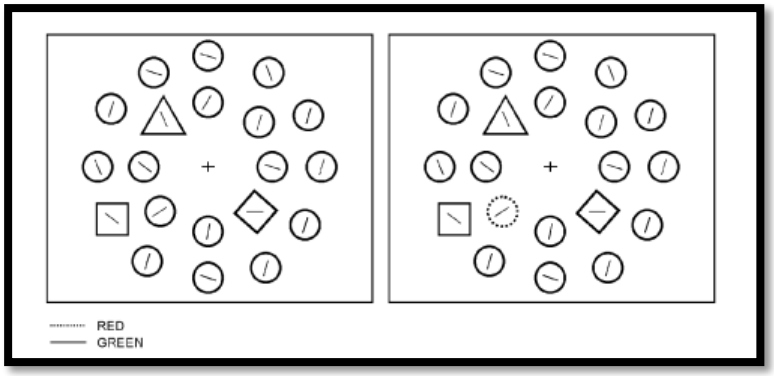


Figure 2.4: Attention capture (source [1])

2.3 Psychophysical Models of Attention

In psychology, many psychophysical theories and models of attention are proposed. These theories are built on psychological concepts. The aim of these models is to better understand and explain visual attention in humans [19]. In this section some essential background related to psychophysical models is provided. Later two main theories that greatly influenced computational visual attention models are briefly explained.

2.3.1 Features

Features are the fundamental attributes that are used to recognise the attended regions in a visual scene [53]. Imagine, for example, a situation where a person is being searched for in a crowd. The visual search here can be done in two ways. Firstly, each and every person can be visually scanned to determine the person. However, this is too costly procedure as it introduces a significant delay. In the second instance, if we possess some prior knowledge (e.g., you might know that the person is wearing a green sweater) about the person, then these attributes can guide our attention. These

attributes that play an important role in improving the detection performance in real world scenarios are known as features. Some of the basic features that make stimulus “pop-out” from its surroundings are colour, orientation and motion etc. [43].

2.3.2 Feature Integration Theory (FIT)

The FIT of Triesman [54] has been one of the influential theories in the field of visual attention. The theory was introduced in 1980 and it claims that “features are registered early, automatically, and in parallel across the visual field, while objects are identified separately and only at a later stage, which requires focussed attention”. The features from the visual scene such as colour, orientation, spatial frequency, brightness and motion are represented in individual topographical feature maps. These individual feature maps are combined into a master map. This master feature map is then scanned serially using focussed attention to provide the data for higher perception tasks.

Triesman mentions that targets differentiated from the distracters with more unique features are easier to search. If the target has no unique features but it still differs from the distracters, then it results in a longer search process. Triesman theory states that the attended regions are searched by either focussed attention or through top-down processing. In any case, it is impossible to predict which has contributed to what we see. Focussed attention is directed serially to all the locations. For example, in proofreading a document and instrument monitoring, focussed attention is needed.

The other way in which the attended regions are identified is by top-down processing. If the features of the target are known in advance, the search time is reduced. In highly redundant and familiar environments in which humans operate, top–down processing is usually much faster. However, when the environment is less predictable then humans are less efficient. For example, searching for the face of one’s own child in a school photograph is a very inefficient visual search, in spite of complete knowledge about the target.

2.3.3 Guided Search Model

The Guided Search Model [55] was developed by Wolfe as an answer to the criticism of FIT. Over a period of time because of huge competition between Triesman and Wolfe’s work, it resulted in many improved versions of the models.

The model shares some of the concepts of FIT: however, it is more comprehensive for computer implementations. It considers both bottom-up and top-down influences in predicting the results of visual search experiments. The authors have chosen colour and orientation as the basic bottom-up features in their

implementation. Unlike FIT, the model generates maps for each feature dimension (colour, orientation,...) rather than for each feature type (red, green,...). A master map of location has been used in FIT. Unlike FIT, an activation map is generated by summing up all of the feature dimensions in the Guided Search Model. As the model also considers top-down information, for each bottom-up feature map, there is a corresponding top-down map that is used to distinguish the target from the distracters. Mimicking the convention of numbered software upgrades, Wolfe has contributed many versions of his model.

2.3.4 Units of Attention

The units of attention refer to the regions that are fixated by the human eye in a scene. Whether a human eye attends to locations, to features, or to objects is a question of debate. The majority of the studies from psychophysics and neurobiology is about space based attention (location based attention) [56], [57]. There is also strong evidence for both feature [58], [59] and object based attention [60], [61]. Today the research community believes that these are not mutually exclusive and humans attend to any of these candidate units [62], [63]. Humans have the ability to attend multiple regions, usually between four to five regions of interest. This has been verified by many psychological [64], [65] and neurobiological experiments [66]. Some of the recent models have used hybrid approaches in which the visual sensitivity information related to space, object and location is fused for predicting the human visual attention [12].

2.4 Machine Visual Attention

In a real time scenario, a computer vision system deals with many pixels in images. Pixels are the smallest addressable units of an image. Each image is composed of several thousand pixels and dealing with all of them in reality results in increased computational complexity. This additional complexity makes the computer vision systems extremely difficult to operate in real time. To address this issue, computer vision scientists have developed many attention models [67], [68], [69], [70]. These are saliency or mathematical models which predict the attended information in the images in free viewing conditions. The need for computational models and better understanding of the HVS led to many attention models over the past two decades. Most of the early models were built around the psychophysical models [71]. As already discussed, one of the most influential theories for computational models is Feature Integration Theory (FIT).

2.4.1 Koch and Ullman

A number of psychophysical theories suggested a two stage attention theory for human perception [72]. The first stage is the pre-attentive mode where the features are processed in parallel over the visual field. In the next stage known as the attentive mode, a specialised processing mechanism known as Focus of Attention (FOA), moves serially to the conspicuous locations in the visual field. Based on this concept, in 1985, Koch and Ullman [72] developed a neurally plausible model.

The model considers only data driven stimuli (bottom-up features) for developing the model. The main idea of Triesman theory is the computation of attention as a ***master feature*** map. The master feature map is developed using individual features such as colour and orientation. Koch and Ullman, in their computational approach defined the master feature map as a ***saliency map*** derived from various elementary feature maps. Therefore, the term saliency map introduced by Koch and Ullman corresponds to Triesman's master feature map. It extracts bottom-up features such as colour and orientation and combines them into a two dimensional grey scale saliency map. A sample input image and its corresponding visual saliency map is shown in the Figure 2.5. This map indicates the importance or conspicuity of every

(a) Input image (b) Visual saliency map

Figure 2.5: Saliency map generation. (a) Input image. (b) Visual saliency map

pixel in the image. According to the authors, this saliency map is either within the LGN or striate cortex (V1). However, there is no significant evidence supporting the claim that only a unique saliency map exists in the brain that guides visual attention.

In the first stage, elementary features are computed in parallel across the visual field and are represented in a set of topographical maps. These maps are then combined into a saliency map. In the second stage, a ***Winner-Take-All*** (WTA) which is analogous to a maximum finding operator scans the saliency map, to select the most

active unit or conspicuous location. The WTA sets all the active units in the saliency map to zero except the one that corresponds to the most active or conspicuous unit. This selected region is considered as the most salient part of the image (winner). The attention is shifted to this location. The ***inhibition-of-return*** mechanism is activated in the saliency map. This makes the current winner to be inhibited and the attention moves towards the next winner in the saliency map. This inhibition-of-return mechanism prevents the FOA returning to the previous winner. WTA shows how the selection of the maximum is implemented using a neural network. This is biologically motivated and explains how the mechanism is realised in the human brain. Finally, the properties of the selected location using WTA are routed to the central representation. The central representation at any instant contains only the properties of the single location in the visual scene [73]. An illustration of the model is shown in the Figure 2.6.

The parallel (pre-attentive mode) and serial (attentive mode) visual search which is described at the beginning of the section can be explained with this model. For

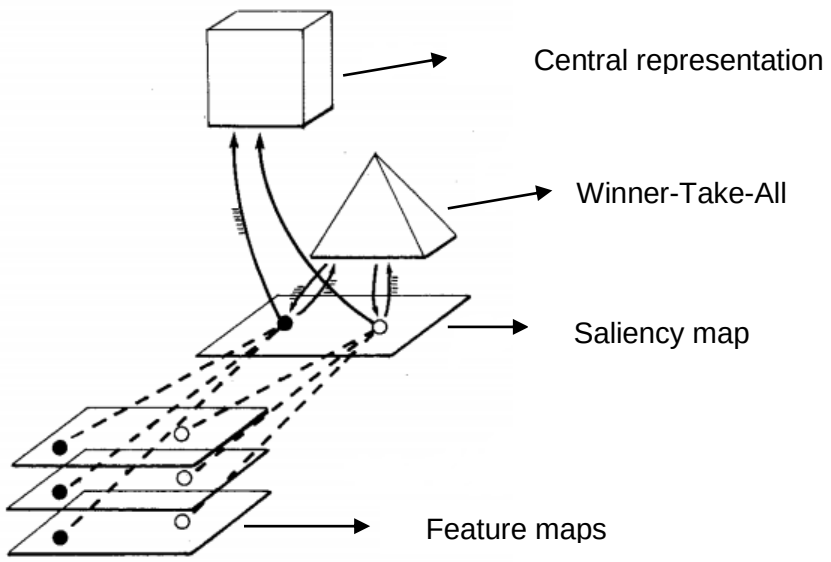


Figure 2.6: A schematic diagram of Koch and Ullman model. (source [72])

example, imagine a scenario where in a target object has to be detected among an array of objects. The model detects the target object's features in its corresponding feature and saliency map. Now, if there are no other distracters in the vicinity of the salient object, then WTA will immediately detects this by inhibiting all other regions in the saliency map. In other words, the target immediately pops out of the scene with homogeneous distracters. The majority of the models have followed the basic idea of Koch and Ullman [28, 71]. They only vary in the types of features, the different

normalisation strategies used to combine the individual feature maps and the weights given to each map.

After Koch and Ullman, Milanese [74] used a new term known as ***conspicuity*** maps for generating the saliency map. In this theory, feature dimensions are subdivided into feature types. For example, if the feature dimension is colour, then the feature types are red, green, blue and yellow. The feature types are represented in feature maps and are then summed up to obtain feature dependent maps known as conspicuity maps. These conspicuity maps are finally fused into a saliency map.

2.4.2 Neuromorphic Vision Toolkit (NVT)

Pioneering work has been done by the authors of [6] in modelling visual saliency. Very good source code and the documentation are provided by the authors. This led to the model serving as the basis for many research groups working in the visual saliency area. In their work, the authors presented a comprehensive implementation and verification of Koch Ullman's and Milanese's theory. The ideas of feature maps, saliency maps, WTA and inhibition-of-return are obtained from Koch-Ullman model. Similarly, the ideas of using linear filters for computation of features, using centre surround difference and the conspicuity maps are adopted from Milanese's work. The main contribution of his work is the realisation of the theoretical concepts, implementation of their work and its application for synthetic patterns and natural scene images.

NVT is a bottom-up visual saliency model built on three features. These are intensity, colour and orientation. The process of obtaining the saliency map consists of five major steps. The first is the construction of an image pyramid by linear filtering at 8 different scales. The second is the computation of colour, intensity and orientation channels from the image pyramid and generating their respective Gaussian pyramid. The third is the generation of feature maps. The fourth is the computation of conspicuity maps. The fifth is the normalisation and summation of the conspicuity maps to obtain the final saliency map. The model is very good at predicting human gaze and it served as the benchmark model for many early visual saliency models. Some of the major drawbacks of this model are significant complexity which limits its practical application and low resolution saliency maps.

2.4.3 Normalisation and Feature Map Combination

The general structure of bottom-up attention system is shown in the Figure 2.7.

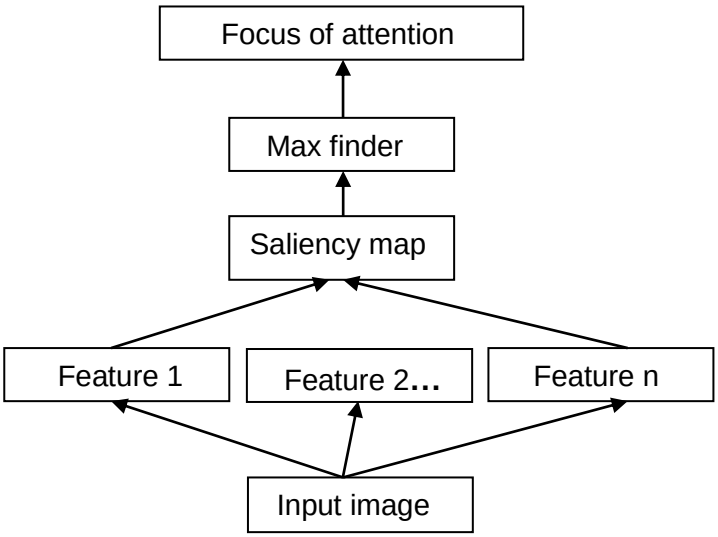


Figure 2.7: General structure bottom-up visual saliency model

From the previous sections, it is evident that individual feature maps are combined to obtain the saliency map.

Before summing up the feature maps, these maps are normalised to the same dynamic range. Therefore, the two important operations that are usually done before the output saliency map is obtained are normalisation and summation. These two operations are briefly explained below.

Normalisation: This is the procedure to transform the grey scale image with intensity values in the range (Min, Max), into a new image with intensity values in the range (newMin, newMax). Normalisation can be both linear and non-linear. In the literature different normalisation schemes are proposed for normalising the feature maps. A straightforward approach is to normalise all the maps to a fixed range [6]. However, in this approach as the magnitude of the maps is lost, it results in a problem if one particular feature is relatively more important. One approach to solve this problem is to determine a maximum M of all the maps and then normalise each of the maps to the range [0...M] [73]. An alternative method is proposed in the work by [75]. In this method the maps are scaled with respect to the long term estimate of its maximum.

Summation: A very important aspect of the saliency model is the summing up of the feature maps. So far it is not clear how this operation is achieved in the brain. In [76] a uniqueness weight is applied to each map before adding the maps. The weighting function determines the uniqueness of features. For example, if there is only one conspicuous bright location in one feature, then a higher weight is given to this

when compared to other maps. If the bright location is surrounded by several other bright location or regions then a lower weight is applied. To achieve this, the authors determine the number of local maxima (m) in each feature map and then divide each pixel with square root of m. Some other different types of solutions are also proposed in the works of Itti *et.al* [6], Itti and Koch [77] (a review of feature map combination is provided in this paper) and Harel *et.al* [7]. Another two interesting normalisation strategies are content-based global amplification normalization and Iterative non-linear normalization which are proposed in [78].

2.5 Applications of Saliency Models

The major goal of psychophysical models is to better understand and interpret human visual perception. However, computational models of saliency improve technical systems. The applications of saliency models are in the areas of computer vision and graphics, robotics and others [36]. The target application of the proposed saliency model falls in image/video compression and computer vision area where predicting the human fixations is very important. Consequently only applications pertinent to these areas are briefly explained.

2.5.1 Perceptual Video Coding

During the process of video coding, the redundancies present in the video data are removed while preserving the video quality to achieve video compression. The existing video coding standards such as H.263 [79], H.264 [80] and its latest successor HEVC [81] achieve video compression by eliminating spatial (similarities between adjacent pixels), temporal (similarities between adjacent frames in the video) and entropy redundancy (similarities between coded symbols in the videos). They do not consider the perception of HVS towards different regions when allocating the resources. This is called as visual or perceptual redundancy. The information obtained from the saliency map of an attention model can be used to reduce the perceptual redundancy present in image/video.

In the saliency map approach, a visual attention model is used to output a saliency map which represents the importance of regions in a video frame. Based on the distinctive features, the saliency map highlights the regions that are relatively more important when compared to others. This relatively less important information is visually redundant and irrelevant to the HVS during interpretation of the image. The bit allocation is done based on the information from the saliency map. The salient regions are allocated more number of bits whereas for the non-salient regions less number of bits are allocated and thereby successfully eliminating the perceptual redundancy.

The HVS is less sensitive to colour information than luminance information. Therefore, during video coding, the colour information is represented with lower resolution than luminance information to achieve better compression and is one of the good examples of eliminating perceptual redundancy. Moreover, during quantisation which is one of the main steps of video compression, the continuous range of values is converted to a finite range of discrete values. Lossy compression is achieved by quantisation. It typically involves dividing the transform coefficient value using a quantisation step and rounding it to the nearest integer. During this process, the frequencies of the video data which are of variable interest to the HVS are eliminated.

The authors of [82] perform image compression using a visual attention system. A colour image compression method adaptively determines the number of bits to be allocated for coding image regions according to their saliency. Regions with high saliency have a higher reconstruction quality than less salient regions. Itti [6] uses his attention system to perform video compression by blurring every frame, increasingly with distance from salient locations. Taking advantage of the multiresolution representation of the wavelet, Guo *et.al* [10] also proposed a foveation approach to improve coding efficiency in video compression.

The two main problems that limited the use of saliency for perceptual video coding are saliency accuracy and complexity [83]. As an ideal saliency map should provide extensive information regarding human perception, achieving this in real time needs a significant amount of computational resources. Further, the mechanisms underlying human attention for predicting the human gaze accurately are not yet fully discovered. Therefore, a novel computational model of attention is proposed in this thesis that has better accuracy with lower computational complexity when compared to the state-of-the-art visual attention models.

2.5.2 Image/Video Quality Assessment

The image/video quality assessment is usually done either using subjective or objective video quality assessment techniques. As video quality is a subjective notion, subjective video quality assessment is generally the best way to assess the video quality. However, due to several limitations such as involvement of human subjects and time complexity, subjective assessment is impractical for most of the applications. In order to address the issues related to subjective testing procedures, several objective video quality assessment techniques were developed.

During objective video quality assessment a mathematical model is utilised to predict the quality of the pictures in a similar way to the humans. Most of the state-of-the-art video quality metrics do not take saliency into account for assessing the

perceived video quality. Some of the works which contributed image/video quality assessment metric based on saliency map are discussed here. The authors of [70] proposed an efficient approach based on the phase spectrum of the Fourier Transform (FT). The indexes of Mulit-Scale Structural Similarity (MSSIM) and Visual Information Fidelity (VIF) are modified by treating the saliency map as a weighting function. The results show that the saliency based strategy improved the original image quality assessment. In the work by [6] based on the assumption that an artefact is more annoying when it falls in a salient region, the visual attention information is employed for image quality prediction. The attention model in [84] uses colour, motion, location, foreground/background, people and context for obtaining the importance maps. The visible errors are then weighted according to the perceptual importance of regions shown in the saliency maps. Their work showed a high correlation with subjective data when compared to widely used Peak Signal to Noise Ratio (PSNR). Similarly in [85] a perceptual importance map is used for assessing the quality of compressed (Joint Photographic Experts Group) JPEG 2000 images.

In this thesis, the effectiveness of the state-of-the-art objective video quality metrics such as Full Reference (FR), Reduced Reference (RR) and No Reference (NR) in detecting perceptual quality variations induced by pre-processing filters have been investigated. This investigation has shown that existing video quality metrics are not good at detecting the quality variations. Therefore, it is indicated that a new image/video quality metric based on saliency map approach is needed for effectively detecting the quality variations.

2.5.3 Object Recognition

The aim of an object recognition system is to find objects of the real world from captured images. Humans perform this task effortlessly and instantaneously. In humans, attending the objects and recognising them is an associated task. Object recognition can be considered as the most relevant application of a saliency model. The main reason being the two stage approach of a pre-processing attention system and classifying recogniser is analogous to the way in which human recognise the objects in their surroundings. The authors of [86] proposed an integrated vision system to detect persons in natural scenes. Their system has two processing stages of which the first stage is a visual saliency front-end and the second stage is the object recognition back-end. They have used HMAX model which is inspired from the neurobiology of inferotemporal cortex. Although it is biologically plausible it is not robust with natural images. To improve the performance they have used a support vector machine algorithm which is highly reliable way of recognising the pedestrians in

images. Walther *et.al* [87] proposed a similar technique which combines an attentional system with an object recogniser based on SIFT features. They have shown that the performance of object recognition results improve with the help of visual attention model. Salah *et.al* [88] uses a saliency based attention system with a neural network which sends observations to Markov models to do handwritten digit recognition and face recognition tasks.

In the above discussed approaches, both the visual attention part and the object recognition are separate entities. However, these are strongly intertwined processes in human perception. Some of the authors have proposed approaches which both processes share resources. The authors of [87] suggested a unifying framework where the HMAX object recogniser is modified to suppress or enhance the regions during spatial attention. Furthermore, the visual attention and object recognition are brought together by using saliency model with object detectors. Some of these models [89], [12] include Viola Jones face detection [90], Felzenshwalb person detector [91], car and other object detectors. As more powerful object recognition systems are developed in the future, the usage of saliency model as attentional front end will be a promising direction in terms of time saving.

2.5.4 Image Segmentation

During image segmentation the image is partitioned into segments that are more meaningful and easier to analyse. In this process the selection of the seed points is an important step. Seed points are the image pixels that represent a particular characteristic or property such as intensity, colour, texture etc. Saliency models are generally used to select or detect these seed points based on some important features. The other pixels in the image which share similar properties with that of these seed points are used to segment the image. The authors of [92] presented a colour image segmentation method based on seeded region growing technique and visual saliency model. The candidates for the seeds are initially selected using the saliency model and then the authors have used the seeded region growing technique to segment conspicuous parts of the image based on a colour homogeneity criterion to discriminate the regions to be segmented from the surrounding regions. Ma and Zhang [93] proposed local contrast based method for detecting salient regions in images. Their model operates on colour quantised CIELuv image which is divided into pixel blocks. The saliency is obtained by summing up of the pixel differences with their respective surrounding pixels within a small neighbourhood. The authors then use a fuzzy growing method that segments salient regions from the visual saliency map. The saliency maps from Itti's model [6] have been used by other researchers for unsupervised object

segmentation. Ko and Nam [94] proposed object of interest segmentation algorithm based on visual saliency and semantic region clustering. The authors initially segment an image into regions and they are then merged as a semantic object. During the process an attention window is created based on the saliency map. A support vector machine is trained on the window to select the salient regions. These regions are clustered together to form an object of interest. The authors of [95] create saliency maps at different scales and combine them using pixel-wise addition to obtain the final saliency map. The input image is over segmented and the corresponding saliency value per segmented region is calculated by averaging the saliency values from the final saliency map. A threshold based method is used in which segments with an average saliency greater than threshold T are retained while the rest of the segments are discarded. The output contains the segments that constitute the salient object. In the work by [96] the authors proposed Conditional Random Field (CRF) model for segmenting objects in images and videos based on the information in the saliency map.

2.6 Summary

In this chapter, the background knowledge related to the visual attention was provided. The different types of eye movements were explained, the bottom-up and top-down factors that drive attention were discussed. The chapter has shown that many different disciplines have been involved in attention research. The psychophysical models which were the basic foundation of today's modern saliency models are detailed. The basic structure of the saliency model that serves as the basis for understanding many saliency models is presented. Lastly, some of the important applications of saliency models in computer vision were briefly explained.

3 Literature Review of Visual Attention Models

3.1 Introduction

Better understanding of the HVS and the improved processing power of computers has led to the invention of a wide variety of visual saliency models in the last two decades. In the previous chapter, some of the early visual attention models have been explained. Their aim was to better understand and explain the underlying principles of human visual attention. Although these models were implemented computationally they had a psychological perspective. However, there may be an overlap of objectives between computational and psychological models. Some of the psychological models are also used to interpret the psychophysical data in computer applications.

In this chapter the literature review related to computational attention systems is provided. The purpose of these attention systems is to improve the computer vision systems. Moreover, the algorithms of some of the state-of-the-art visual saliency models in terms of predicting human gaze are discussed. The underlying mechanism of each of these models is clearly described. Later the strengths and the potential problems related to each of these models are discussed. Further, the interesting aspects of each of the models in terms of future implications in attention research are also outlined. These interesting things are generally novel aspects which give rise to later inventions.

Section 3.2 discusses the different types of computational attention systems. The different types of features used in the literature to develop the bottom-up and top-down saliency models are outlined. Section 3.3 discusses some popular saliency models and the most closely related approaches to the proposed visual saliency model. Section 3.4 provides a conclusion to the chapter.

3.2 Computational Attention Models

In the literature, computational attention models are classified as filter based and connectionist models [1]. The filter based models are further divided as bottom-up and top-down visual attention models. The characteristics of these different types of attention models are discussed in the following sub-sections.

3.2.1 Connectionist Models

Connectionist models are based on neural networks. These models are biologically plausible and they have units corresponding to neurons in the human brain. However, very little is still known regarding the processes in the human brain. Some of the connectionist models are dynamic routing circuit [97], Multiple Object Recognition and Attentional Selection (MORSEL) [98], Selective Attention Model (SLAM) [99] and Selective Attention for Identification Model (SAIM) [99]. Further, many of the psychophysical models proposed in the literature also fall into this category. Connectionist models are discussed very briefly here as they are beyond the scope of this thesis. Major emphasis is given to filter based models as the proposed computational model in this thesis is a filter based model. However, some of the recent advancements in saliency research which are based on deep learning (e.g. [100], [101], [102], [103], [104]) are discussed at the end of the chapter.

3.2.2 Filter Based Models

Filters models generally use the linear filtering operations to compute the features of an image. Some of the examples of filter based visual saliency models are presented in [74], [6], [105]. The filter based models can be further classified as bottom-up and top-down visual attention models.

3.2.2.1 Bottom-up Visual Saliency Models

In the case of images, similar to the psychological patterns there will be regions of interest that pop-out when the background is homogeneous in nature. An example of this is a football on the ground; here the ground (which is green in colour) is homogeneous whereas the football is something which pops out from the ground being extremely different as shown in the Figure 3.1.

Figure 3.1: Bottom-up region demonstration (Football on the ground) (source [106])

These regions which pop out from the homogeneous background are known as bottom-up regions that attract human attention.

The three basic bottom-up features used to detect the bottom-up regions as proposed by psychological and biological work are intensity, colour and orientation [6, 105, 107]. The other simple bottom-up features are curvature [74], spatial resolution [108], corners [109], regions with good edges, optical flow [110] and flicker [111]. In addition, some of the complex bottom-up features that were modelled are entropy [112], eccentricity [113], Shannon self-information measure [114], ellipses [115] and symmetry [113]. All these are bottom-up features related to images.

Each feature has its own associated computational complexity depending upon how a feature is computed and implemented. By increasing the number of bottom-up features the regions that pop out from the visual scene can be accurately detected. However, too many bottom-up features may introduce a significant amount of processing overhead [1]. Hence, there should always be a compromise between the number of features and the expected processing speed. Usually three to four feature channels will be an ideal choice to achieve a trade-off between accuracy and processing speed [1].

3.2.2.2 Top-down Visual Saliency Models

Top-down saliency is generally user driven. Research has found three major sources of top-down attention. In the first instance the models address visual search in determining how the attention is drawn towards the targets. Another type of information that is used to determine the top-down regions are scene context. These models investigate the role of scene context or gist in deriving the saliency. There are also scenarios in which it is extremely difficult to predict the human gaze as there could be a complex task (task demands) that govern eye fixations that play a role in visual attention.

Scene context: Humans highly depend on scene context for facilitating object detection in natural scenes [2], [3]. For example in a scene, searching for cars on the street, the search process is confined to the street and the sky region is ignored. As humans have many experiences in similar environments from the past related to a street scene (context), whenever this kind of scene is encountered, the search process begins from the street ignoring the sky region. Other examples include a computer on a desk, a plate on the table. The other type of context that can be used to determine the salient regions in the scene is gist. When an observer briefly looks at an image (80 ms or less), he or she is able to report the essential characteristics of the visual scene.

This rough interpretation known as gist or scene essence does not contain individual object details but can provide sufficient information for coarse scene discrimination. Gist is a semantic category which consists of scenes such as an office scene or forest. Gist guides the eye movements and it is calculated from the feature channels. Gist representations have useful applications in computer vision such as searching for objects of interest [3], [116] and scene completion [117].

Task demands: An introduction to task demands was given in the previous chapter, in which one of the famous examples of Yarbus related to task demands was explained. The authors of [1] found a strong relationship between visual cognition and eye movements when dealing with complex tasks. In visually guided tasks, the majority of the human eye fixations are towards task relevant regions [118]. The top down influences are studied in natural scenes during tasks such as walking, playing cricket, sandwich making and driving [119], [5], [120], [4]. Eye movements during activities such as answering a phone call, during driving and adjusting the radio have also been studied in [121].

There are also some other learning based approaches to determine the top down information. This includes detection of the object, detection of face, etc and the other high level features such as hand, text and gesture detection. This thesis explores these learning based approaches in chapter 6. In order to imitate human visual behaviour both bottom-up and top-down saliency has to be fused to obtain the focus of attention [1]. Bottom-up mechanisms have been thoroughly studied compared to top-down mechanisms. The fundamental reason is that bottom-up features are easier to control and model than cognitive features [71].

The advantage of connectionist models lies in its ability to show different behaviour for each neuron. However, in a real time scenario this is computationally expensive and so a group of units exhibit similar behaviour [1]. In contrast to the connectionist models, in the case of filter based models each pixel in the map is given equal importance. Moreover, these models are well suited to real world image applications and profit from approved techniques in computer vision. Therefore, these models are given more emphasis and are discussed in detail in the following sections.

3.3 Related Work

This section reviews some representative saliency models available in the literature followed by discussions on frequency domain approaches as the proposed saliency model in this thesis is a frequency based model. The performance of these

models in terms of prediction accuracy and complexity is compared with the proposed visual saliency model in the later chapters.

3.3.1 Visual Saliency Detection

The authors of [7] proposed a bottom-up visual saliency model known as Graph Based Visual Saliency (GBVS) in the year 2006. It highlights the regions that are more informative in an image according to specified criterion. The authors achieved this in a three step process: Extraction of feature maps, forming activation maps, normalization and summation process. During the process of extracting feature maps, three important features, namely intensity, colour and orientation are extracted at multiple scales. The activation maps are generated for all the feature maps of the given input image. The authors wanted the pixels or nodes that are highly dissimilar to the surrounding nodes to be given a higher value or to be shown as highly important in the corresponding activation map

A fully connected graph of dissimilar regions is built across all nodes within each feature map. The resulting graphs are treated as Markov chains by normalising the weights of the outbound edges of each node to 1. This equilibrium distribution is treated as an activation map. To normalise an activation map $A:[n]^2 \to \mathbb{R}$ they propose another Markovian algorithm. They construct another graph and for each node (i, j) and node (p, q) to which it is to be connected, and the edge weights are defined. After defining the edge weights, the edge weights are normalised to unity and then the resulting graph is treated as another Markov chain. The Markov chain computes the equilibrium distribution over the nodes and mass will flow preferentially to nodes with high activation. The resulting map is a normalised activation map. Finally these normalised activation maps are fused using additive summation to obtain the final saliency map. GBVS achieved high prediction accuracy in detecting human fixations and it is also widely cited in the literature. However, according to a recent review paper [36] in visual saliency, the major drawback of graph based models is computational complexity.

The authors of [70] proposed SUN model in 2009. It derives saliency based on both bottom-up features and prior knowledge about the target. The prior knowledge about the scene is the top down information regarding the visual environment. Both of these pieces of information are combined probabilistically according to Bayes' rule. The authors considered local or self-information of the target as the bottom-up saliency in their model. This local information is distinct or different to the background information and is rarely seen in the image. Therefore, this rare information indicates the target.

Let Z denote a pixel in an image, L represents the location of the pixel (pixel coordinates) and F be the features corresponding to the pixel. C indicates whether or not the pixel or the point belongs to the target class. If C=1 the pixel belongs to the target to which the human eye is interested or else it is a pixel or point of no interest to the HVS. The probability of all the locations in an image can be estimated by using a log scale. Therefore, the saliency is referred as $\log S_z$

$$\log S_z = -\log p(F = f_z) + \log p(F = f_z | C = 1) + \log p(C = 1 | L = l_z) \qquad (3\text{-}1)$$

The first term in the equation (3-1) represents self-information of the target. The negative sign indicates that an increase in the self-information results in the decrease of the feature probability. It also means that rare features are more informative and is referred to as the bottom-up component. The second term denotes features related to the targets that are consistent with human knowledge. It means that if the observers are already familiar or possess some prior knowledge about the target that is being searched, then the log likelihood increases only if the expected target is in the image and decreases if an unexpected target is present. Therefore, this term is the top-down component of the model. The third term is prior knowledge regarding where the target is likely to be present in an image. It is the knowledge regarding the location of the target and is independent of the visual features. The interesting aspect of these models is the use of Bayesian framework for detecting salient regions. The key advantage of the Bayesian models is their ability to learn from the given data. However, similar to GBVS model computational complexity is the main drawback of these models.

Goferman *et al.* [9] introduced Context Aware Saliency (CAS) in 2012. The CAS model detects not only the objects but also the context or the background which is just immediate to the object that describe the purpose of the object being there. They detect salient regions based on four principles of human attention. The model considers low level factors such as colour and contrast. Frequently occurring features are suppressed while maintaining the features that deviate from the norm. Visual forms possess one or more centres of gravity about which the forms are organised. Finally the top down factors such as human faces are considered to attract human attention.

According to the first two principles, a pixel will be salient if it is distinct from the surrounding pixels. The authors, instead of considering isolated pixels, considered surrounding patches of scale r at each pixel in the image to determine the pixel saliency. Therefore, a pixel i is considered to be salient only if the patch p_i centred at pixel i is distinct with respect to all other patches. The Euclidean distance between two

vectorized patches p_i, p_j in CIE $L*a*b$ colour space is defined as $d_{colour}(p_i, p_j)$. The pixel i is salient only if $d_{colour}(p_i, p_j)$ is very high $\forall j$. The third principle is achieved by considering the positional distance between the patches. The dissimilarity measure between the two patches p_i and p_j as a function of $d_{colour}(p_i, p_j)$ and $d_{colour}(p_i, p_j)$ is defined as

$$d(p_i, p_j) = \frac{d_{colour}(p_i, p_j)}{1 + c.d_{position}(p_i, p_j)} \tag{3-2}$$

The value of c is assumed to be 3. Therefore, the single scale saliency at a scale r is defined as

$$S_i^{\,r} = 1 - \exp\left(-\frac{1}{k} \sum_{k=1}^{k} d(p^r{}_i, q^r{}_k \right) \tag{3-3}$$

They also use four different scales and the saliency $\overline{S_i}$ is obtained by taking the mean of the saliency at different scales.

$$\overline{S_i} = \frac{1}{M} \sum_{r \in R} S^r{}_i \tag{3-4}$$

To determine this immediate context all the salient regions are initially extracted from the saliency map by defining a threshold of $\overline{S_i}$ > 0.8. All the regions above this threshold are considered to be salient regions. The pixels outside the attended regions are weighed according to the Euclidean distance to the closest attended pixel. Let $d_{foci}(i)$ be the Euclidean distance between pixel i and the closest focus of attention pixel. Then the saliency of pixel is defined as

$$\overline{S_i} = \overline{S_i}(1 - d_{foci}(i)) \tag{3-5}$$

Finally the saliency map is improved by detecting human faces in the images.

The ability to detect context around the salient pixels in the image is the major advantage of this work. However, low prediction accuracy and significant complexity

when compared to the other state-of-the-art models is the disadvantage of this saliency model.

Vikram *et al.* [122] proposed Random Centre surround Saliency (RCSS) in 2012. In this method the input image is initially Gaussian filtered to remove noise and abrupt onsets. The filtered image is then transformed into $CIE1976L^*a^*b^*$ [123] colour space. The authors use this colour space as this has many similarities to human psycho visual colour space. Random sub-windows are generated over each individual L^*, a^*, b^* channels. The co-ordinates of the random windows are generated using discrete uniform probability function, as it helps in placing windows without any bias towards specific region of an image or size of the window. According to the authors this is a very important step as the salient regions can occur at arbitrary scales and positions of the image. The saliency at a point or pixel in the individual channel is defined as the sum of absolute differences of pixel intensity values to mean intensity values of random sub windows. The final saliency map is obtained as the pixel wise Euclidean norm of all the saliency maps generated across all the channels. The normalised saliency map is median filtered because of its ability to preserve edges while eliminating noise from the map. The contrast of the map is then increased by histogram equalisation. This is done as HVS enhances the perceptual contrast of the salient stimulus in the visual scene. RCSS does not have any parameters that need to be tuned. In spite of the model achieving good prediction accuracy there is significant amount of processing overhead in computing the saliency maps of the images.

3.3.2 Frequency Based Models

Xiodi and Liqing [8] developed Spectral Residual (SR) model based on frequency domain characteristics. For an input image $I(x)$ the amplitude $A(f)$ and phase spectrum $P(f)$ are calculated as

$$A(f) = R(F[I(x)]) \tag{3-6}$$

$$P(f) = \gamma(F[I(x)]) \tag{3-7}$$

Where F denotes the Fourier transform. Then the image is down sampled and the log-spectrum $L(f)$ is computed. $L(f)$ is then multiplied with an N X N averaging filter $h_n(f)$ to obtain the averaged spectrum.

$$L(f) = \log(A(f)) \tag{3-8}$$

$$A(f) = h_n(f) * L(f) \tag{3-9}$$

The *spectral residual* (SR) $R(f)$ is obtained by subtracting the result from the log spectrum itself.

$$R(f) = L(f) - A(f) \tag{3-10}$$

Finally the spectral residual obtained is smoothed using a Gaussian filter for better visual effect. The entire process, as described by the authors, can be realised in a short (approximately six lines) MATLAB code. However, the implementation does involve complex functions such as Fourier and inverse Fourier transforms embedded in MATLAB. The main advantage of this model is the speed at which it derives the saliency of an image. Moreover, SR is a very simple model to explain and is easy to implement.

The authors of [124] came up with another model known as Phase Spectrum of Fourier Transform (PFT) in 2008 after careful analysis of spectral Residual (SR). They found that the amplitude spectrum is not fully successful in obtaining an accurate saliency map. A better saliency map can be obtained by using the phase spectrum of the Fourier transform. For a given input image $I(x, y)$ initially the Fourier transform and the phase is computed.

$$f(x, y) = F(I(x, y)) \tag{3-11}$$

$$p(x, y) = P(f(x, y)) \tag{3-12}$$

Later the saliency map $sM(x, y)$ is obtained using the equation (3-13)

$$sM(x, y) = g(x, y) * \left\| F^{-1}\left[e^{i.p(x,y)}\right] \right\|^2 \tag{3-13}$$

Where $g(x, y)$ is a 2D Gaussian filter with sigma=8. PFT is much faster when compared to SR. It saves one third of the computational complexity compared to SR and is better in terms of accuracy. The authors further extended this model to the

Phase Spectrum of Quaternion Fourier transform (PQFT) [10] to obtain a spatio-temporal saliency map. Initially the image is represented as a quaternion image using four features. Later the PQFT is computed to generate the spatio-temporal saliency map. The authors obtain four features of the image, namely two colour channels, one intensity channel and one motion channel. These features are processed in parallel fashion and thus save processing time. By setting the motion channel to zero, the PQFT model can also be used to work with static images. Unlike PFT and SR it is independent of the parameters and prior knowledge. However, there are also some limitations to the PQFT model. Although the model is robust to white coloured noise, if the noise is similar to the salient region in the image, the models fails to detect it. In a conjunctional search (refer to section 2.2.4.1), humans are extremely good at detecting the salient regions. The PQFT model and many other models failed to perform well at conjunctional search patterns.

Achantha *et al.* [125] proposed frequency tuned salient region detection using low level features such as colour and luminance. Initially the image from RGB colour space is converted to CIE colour space [123]. The saliency map S corresponding to an image I of width W and height H is defined as

$$S(x,y) = \left| I_{\mu} - I_{\omega hc}(x,y) \right| \tag{3-14}$$

I_{μ} is the arithmetic mean pixel value of the image. $I_{\omega hc}$ is the Gaussian blurred version of the original image. The blurred version eliminated fine texture details, noise and coding artefacts. Some saliency maps have badly defined object boundaries [6] that limit their usage in certain applications. This happens due to the downsizing the image to a greater extent before computing the saliency map. When an image is downsized the spatial frequency content that is present in the original image is lost. For example, Itti's model outputs saliency maps that are 1/256th of the original image resolution. Similarly the SR model outputs saliency maps of 64x64 pixel size. In contrast to Itti's model and SR, the algorithm proposed by Achantha outputs saliency maps that are of same size as the input image.

In 2009 a spectral whitening model was proposed by Bian and Zhang [126]. The authors came up with this model as a refinement of the early spectral based models such as SR, PFT and PQFT. The model is based on the idea that the visual system attends to rare informative features while ignoring irrelevant or redundant non informative features. The given input image I is initially resized. The ratio of image

width to height is retained while the maximum length of the image is set to 64px. Next the windowed Fourier transform of the resized image is calculated.

$$f(u,v) = F\left[w(I(x,y))\right] \tag{3-15}$$

$$n(u,v) = f(u,v) / \left\| f(u,v) \right\| \tag{3-16}$$

F denotes the Fourier transform and W is the windowing function. The normalised (flattened or whitened) spectral response $n(u,v)$ is converted to spatial domain using an inverse Fourier transform. The result obtained is squared to emphasise the salient regions and is then convolved with a Gaussian low pass filter to generate the final saliency map as shown below.

$$S(x,y) = g(u,v) * \left\| F^{-1}\left[n(u,v\right]\right\|^2 \tag{3-17}$$

The interesting aspect of the SR is generating the saliency in the spectral domain. This novel and interesting aspect gave rise to many spectral domain approaches such as PFT, PQFT and spectral whitening method.

A novel visual saliency model was proposed by Lin *et al.* [13] based on three simple priors (features). The authors have developed a model that has the ability to obtain better prediction accuracy simultaneously with low complexity so that it is suitable for real time applications. This is an important model for the benchmark in this project as it is the only model which has considered both accuracy and complexity as criteria for developing the model. The three simple priors or components of the saliency model proposed by the *a*uthors are frequency, colour and location prior. They use $CIEL*a*b$ colour space for deriving these features. There are some studies which have shown that warm colours such as red and yellow are more pronouncing to HVS when compared to cold colours such as green and blue. The authors derive colour saliency based on this concept. They model location prior using a Gaussian map. Finally, the frequency, colour and location prior saliency maps are fused to obtain the final saliency map.

Xiodi *et al.* [11] in 2012 proposed image signature which highlights the sparse salient regions. The image signature is defined as

$$\text{Imagesignature}(x) = (sign(DCT(x)) \tag{3-18}$$

The reconstructed image $(\bar{x})$ is generated by applying inverse discrete cosine transform as shown

$$\bar{x} = IDCT(sign(DCT(x)) \tag{3-19}$$

The saliency map m is then obtained by smoothing the squared reconstructed image.

$$m = g * (\bar{x} \circ \bar{x}) \tag{3-20}$$

Where $\circ$ indicates Hadamard product operator, * is the convolution operator and g is the Gaussian kernel. It is an extremely fast model.

Imamoglu *et.al* [127] proposed Wavelet Based Saliency Detection (WBSD) in 2013. According to the authors the models using FT may encounter difficulties and lead to unsatisfactory results when there are non-stationary or periodic signals. For example, the high level down sampling using FT resulted in spatial information loss in both SR and PQFT models. Moreover, the global irregularities of the visual scene are more emphasised when compared to the local irregularities. This is due to the analysis of frequency components using FT in a global context. In the literature it is shown that multi-scale wavelet transform does better local frequency analysis as the input signal is examined carefully at different bandwidths. The authors use wavelet transform in their work because it has the ability to provide multi-scale spatial and frequency analysis simultaneously. Furthermore, it is very easy to account for both global and local features in the wavelet domain. The authors obtain global and local saliency maps in $CIE\ Lab$ colour space and fuse them to obtain final saliency map. The interesting aspect of WBSD is the computation of saliency map as a combination of both global and local saliency maps. However, the complexity of the model is very high as it uses several scales to derive the feature maps. The complexity issue limits its practical application.

The primary benefit of these spectral based models lies in their simplicity of generating the saliency maps. They are extremely fast at predicting human attention. Moreover, they can be very easily implemented within few lines of MATLAB code. However, the biological plausibility of these models is still not clear. Models which

replicate some of the known properties of physiology or biological vision are generally considered as biologically plausible.

In the last five years two sub-fields of research have emerged in visual saliency area. These are salient object detection and object proposal generation models. As already discussed the fixation prediction models pop out visually salient regions corresponding to human eye movement. This is the major challenge which has been tackled in this work. Whereas, the salient object detection models detect the most attention grabbing objects and segments them. The intensity of the pixel in the output saliency map represents its probability of belonging to the salient object. The object proposal generation models detect image regions that may contain objects from any object category [128]. They differ from traditional object detectors which are class specific dealing with only one object class (e.g. cars). These object proposal models in contrast define generic objectness measure over all the classes. They quantify how likely an image window consists of an entire object. These objects may belong to any class e.g. car, swan etc. These models instead of dealing with pixels that belong to the objects, deal with windows containing objects. A complete review of the recent salient object detection models and object proposal generation models can be found in [129].

3.4 Machine Learning Approaches for Saliency Detection

Machine learning approaches have been successfully used to detect salient fixations in the images. Judd *et.al* [12] proposed a supervised learning model of saliency based on bottom-up and top-down features. They use low, mid and high level features to define the salient locations and use linear support vector machine to train their saliency model.

Ensemble methods were also used to detect salient regions. The authors of [130] combine low level features such as intensity, colour, orientation, saliency maps of previous best model and the top down features such as faces, human, cars etc in deriving the saliency. They learned from these features using learning algorithms like regression, SVM and Adaboost. The authors found that the boosting model out-performs several state-of-the-art visual saliency models. They also show that their model is able to detect most salient object in a scene without region segmentation. In [131], the authors consider visual saliency computation as a regression problem. They proposed DRFI model which uses a random forest regressor that maps feature vector of each region to a saliency score. They use this model for salient object detection. According to the review paper in [129], which compared the performance of 29 salient object detection model has found that DRFI model gives best performance for detecting salient objects. Li *et.al* [132] proposed a salient object detection model by

combining existing segmentation techniques and fixation based saliency. They initially generate set of object candidates and then use a fixation algorithm to rank these regions based on their saliency. The authors use a random forest with 30 trees for the datasets and use a random regression forest to quantify the saliency of an object mask. In the work of [133], several instances of bio-inspired hierarchical model family are selected and combined using hyper parameter optimisation. The final model is an ensemble of individual models. A simple linear classifier trained on this mixture out-performs the state-of-the-art.

In the last three years due to the overwhelming performance of deep learning in other vision tasks such as image classification and object detection, some of the deep learning algorithms were used to develop the saliency models. Deep learning methods such as deep convolution neural networks (CNN), boltzmann machine, deep belief networks and auto encoders have been used to study their performance in predicting fixations and salient object detection. The authors of [134] proposed an unsupervised three layer deep learning network to learn from mid and high-level features that attract attention. It learns mid and high level features such as junctions, textures, parallelism and faces, text respectively. Finally they also contribute a unified feature integration framework that integrates low, mid and high level features in a biologically plausible way. The authors of [100] proposed a salient object detection model based on multiscale deep CNN features. They use both low level i.e., visual contrast combined with high-level semantically meaningful features in extracting salient objects. They perform feature extraction using a CNN trained on ImageNet dataset. To compute the contrast, they extract multiscale CNN features for every region using three nested and increasing larger rectangular windows which encloses the current region, immediate region and the entire image. The penultimate fully connected layer of their neural network becomes a high-level feature vector for saliency detection. Wang *et.al* [104] computes saliency based on local features and global search. They use two deep CNN's, the first one uses supervised learning scheme to capture local contrast, texture and shape information. The second deep CNN is trained to predict the saliency of object region based on global features such as global contrast and geometric information. The saliency is finally obtained as weighted sum of individual saliency maps. The authors of [103] consider saliency as a high-level task in their work and they model it based on local and global context. A deep CNN with multiple contexts is designed for salient object detection. The global context is used to detect saliency in the entire image, while the local context is used for meticulous areas. The global and local context are integrated into a multi-context deep learning framework and finally optimised for detecting objects. The authors have also evaluated contemporary deep

structures such as AlexNet, Clarifai, OverFeat and GoogleNet on image datasets. Nian *et.al* [67] proposed multiresolution CNN (Mr-CNN) for learning both bottom-up and top-down features from raw image data for predicting fixations. Mr-CNN is trained on fixated and non-fixated locations over multiple resolutions using raw image pixels as the input and eye fixation attributes as labels. Bottom-up information is obtained by combining information at multiple layers while the top-down features are learned using the higher layers. The bottom-up and top-down information is integrated in the final logistic regression layer to predict fixations. A deep CNN architecture using mid-level features based on low-level k-means filters has been proposed in [135]. This model generates multi-scale and multi-level saliency maps and fuses them to obtain final saliency map.

Xia *et.al* [136] proposed a deep autoencoder based centre surround inference saliency model to estimate bottom-up saliency. They explore an adaptive centre surround comparison scheme by taking global competition into non local centre surround reconstruction framework scheme. Their deep network is trained using global data to detect the central patch from the neighbouring patches. The saliency is estimated by taking the residual of reconstructed and original central patches. In the research work of [102], a two stage deep learning network has been proposed which learns from raw input pixels in an unsupervised manner. In the first stage, an unsupervised stacked denoising autoencoder (SDAE) is developed to learn robust representative features to capture patterns of image patches. The second learning stage jointly learns optimal mechanisms to detect the feature contrast and integrates them for predicting human eye fixations. The authors of [101] proposed a super pixel wise CNN approach known as SuperCNN. It learns hierarchical contrast features from two superpixel sequences instead of raw pixels. SuperCNN recovers contextual information from these sequences and the saliency is detected by using multiscale network architecture.

3.5 Conclusion

This chapter initially explained the need for and importance of computational attention systems. The different types of attention systems (filter and the connectionist approaches) were briefly described. The bottom-up and the top-down approaches for modelling the visual attention models were explained in detail. The different types of features utilised in the literature to model the bottom-up and top-down attention models were discussed. Later some of the important saliency models in the literature and the most closely related attention models to the proposed saliency model in this thesis are critically analysed by discussing their strengths, weaknesses and their interesting

aspects. Moreover, the recent advancements and machine learning approaches including deep learning methods used by research community are discussed.

In the following chapters the methodology used to validate the saliency model is described. The quantitative and qualitative techniques used to analyse the saliency maps will be explained. Moreover, the characteristics of the benchmark datasets used for analysing the models performance will be discussed.

4 Experimental Methodology

4.1 Introduction

The performance of the proposed visual saliency model is validated by comparing its performance in terms of prediction accuracy and computational complexity with the state-of-the-art saliency models. Image datasets are used in this work to evaluate the performance. Qualitative and quantitative approaches are used to analyse the model's performance against the chosen image datasets.

Section 4.2 provides the information about the development environment and the testing platform used in the research work. In section 4.3 a brief overview of the different types of image datasets used by research community is given. The strengths and weaknesses of each of the dataset are discussed. Moreover, the characteristics of the datasets that are employed in this work are explained in detail. The process of obtaining a human attention map from the eye tracking information is given in section 4.4. In section 4.5, the empirical validation of the saliency model is explained. It involves the qualitative and quantitative assessment of the model's performance. The different types of quantitative metrics chosen for quantifying the model's performance are elucidated. In section 4.6, the approach followed for measuring computational complexity of the saliency models is given. The important factors considered during the measurement are briefly outlined. Further, the criteria employed for the state-of-the-art model comparison is also provided. Section 4.6 discusses the benchmark models of saliency. The important factors that are considered for the selection of models in the benchmark comparison are given. Moreover, the different approaches undertaken for obtaining these models are also discussed. Section 4.8 summarises the chapter.

4.2 Software Testing and Implementation

The programming software and the testing platform used in this work are provided in this section.

4.2.1 Development Environment

The software environment chosen for the development of models is MATLAB (version R2013b). It is used for the development and testing of the models. A software tool enabling faster coding and easy bug handling is required as it speeds up the development process. MATLAB offers reduced coding time with flexible error management. The MATLAB computer vision toolbox is used during the development of

the attention models. It includes many in built computer vision functions related to object detection and tracking, camera calibration and video processing. These can be easily utilised and there is no need to reinvent existing functionality. Furthermore concise code, excellent plotting tools and good Integrated Development Environment (IDE) favours the usage of MATLAB for developing the algorithms. Therefore, the proof of concept is developed in MATLAB in this project. Later as MATLAB is very slow at execution time the models which have been developed are implemented in Microsoft Visual C++ programming language. Further, OpenCV, FFmpeg libraries and QT project are used during the implementation. These are libraries of functions aimed at real time operation. C++ offers processing speed which is much faster compared to MATLAB. Therefore, in a performance critical scenario, or for the development of any prototype or product, it is better to implement in C++. This enables the algorithms to work in real time. Please refer to the Appendix-B of this thesis for complete details regarding the implementation process.

4.2.2 Testing Platform

A computer with following specifications is used for developing and testing of the attention models.

Operating System	: Microsoft Windows 7 Ultimate.
Processor	: Intel core I7-2600K CPU @3.40 GHz
Installed Memory (RAM)	: 16.00 GB
System type	: 64-bit Operating System.
Display screen	: Intel(R) HD graphics
Display mode	: 1920x1080 (32 bit) (59 Hz)

As indicated in section 4.2.1 the MATLAB has been chosen for the development of attention models. One of its major problems is slower execution time. During the testing phase of saliency model, the performance of the developed models is evaluated across large image datasets. Moreover, simultaneously several applications need to run in parallel. Therefore, this demands more RAM and a faster processor. The above given specification for the computer was chosen keeping in view of all this.

4.3 Image Datasets

An image dataset is basically a collection of images and human ground truth (eye tracking maps). The human ground truth is the information obtained from the eye tracker related to where a human eye is interested in the images. In the current project datasets that obtained ground truth using eye tracking devices are used as the goal is to predict human fixations. The effectiveness of visual saliency models is computed on different image datasets. In the literature several datasets are proposed. Each of them varies in terms of the number of images in the dataset, the resolution of the images used, the number of subjects used to collect the eye-tracking data, the viewing time per image, the subject's distance from the screen and the stimulus variety [137]. Some of the popular eye tracking datasets are Bruce and Tsotsos [138], Kootstra Shomacker [139], Le Meur [140], Judd *et.al* [12] and DUT-OMRON [141]. The Bruce and Tsotsos dataset has 120 images. The image content is mostly indoor and city scenes. This dataset was published during the early stages of the research area and is most widely cited in the literature. However, its weaknesses are the smaller number of images and stimulus variety. The Kootstra and Shomacker dataset has a wide variety of images; however, it has only 100 images. The Le Meur dataset has 27 images with the highest number of eye tracking subjects. The state-of-the-art related to datasets has also changed significantly in the last few years. The Judd *et.al* dataset was published in the year 2009. According to a review paper [46] published in January 2013, it has indicated that Judd *et.al.* is the largest dataset with respect to eye tracking fixations. However, in late 2013 another dataset DUT-OMRON was published and it has even more images than Judd *et.al* and a better variety of images with both bottom-up and top-down image content. Therefore, these two public datasets are selected for the development of visual saliency model. Moreover, apart from these two public datasets another small self-dataset is also proposed. The details of these datasets are given below.

Judd *et.al* : The dataset consists of 1003 images that are collected from Flickr creative commons and Label Me dataset. The eye tracking data is collected from fifteen users under free viewing conditions. In free viewing conditions the viewers are not given any instructions before viewing the images. Instead, they are simply asked to watch the images. These are the situations in which humans observe the world without any specific goal [71]. This process negates the top-down influences that exist in viewers' minds. The majority of the images in the dataset are 768x1024 or 1024x768 pixels in size and a few other images have different dimensions. Sample images from the dataset are shown in the Figure 4.1. Both male and female viewers between the

age of 18 and 35 viewed the images. Among the fifteen users two were researchers who belong to the project and others were non experts. A 19 inch computer screen of

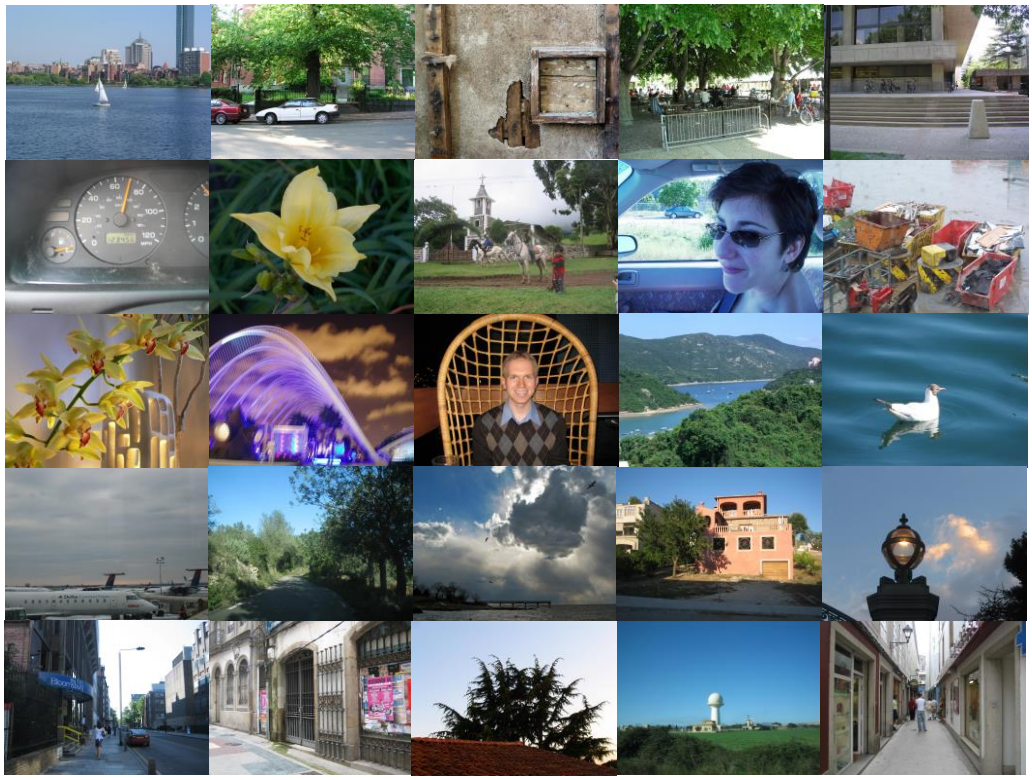


Figure 4.1: Sample images from Judd's dataset

resolution 1280x1024 was used to show the images. The viewers sat at a distance of 2 feet using a chin rest to stabilize the head. Each image is shown at a full resolution for a period of 3 sec with an interval of 1 second. The table mounted ETL 400 ISKAN eye-tracker is used to record the scan path of the viewers. The first fixation from all the scanpaths of the viewers was discarded to avoid the centre bias problem. The authors obtain raw eye tracking maps from the eye tracker. Later, in order to obtain a continuous human saliency map they use a Gaussian filter to convolve the fixated locations.

DUT-OMRON Image dataset : Chuan *et.al* [141] introduced the DUT-OMRON dataset. According to the authors, in recent years the experimental results on the existing datasets have reached a very high level. They are hardly of any use for the advancement of current research in visual attention models. The reason behind this is that the images in the existing datasets are much simpler when compared to the real life images. As a result, the authors have proposed a dataset consisting of 5168 images. Sample images from the dataset are shown in Figure 4.2. These images are of

high quality and manually selected from more than 140,000 images. The dimensions of the images are resized to 400*x or x*400. The value of x is less than 400 pixels. These images consist of one or more salient regions with relatively complex background. For collecting the ground truth from the eye tracker the authors used 5 participants. They have used a Tobbii X1 Light Eye tracker to record eye fixations. Each image was displayed for a period of 2 seconds with no interval between successive images. All of the viewers have normal or corrected to normal vision. Similar to Judd *et.al* the fixation maps are obtained by convolving the raw eye tracking using a Gaussian filter. Further, the first fixation which has the highest probability to be at the centre of image is also removed. It is the only dataset that has eye fixations from an eye tracker, bounding box and pixel-wise ground truth. When compared to other datasets the images are more difficult and challenging and thus provide space for improvement of visual saliency models. In this research project Judd *et.al* dataset is used in the training phase. DUT-OMRON is used during the testing phase of the saliency model development.

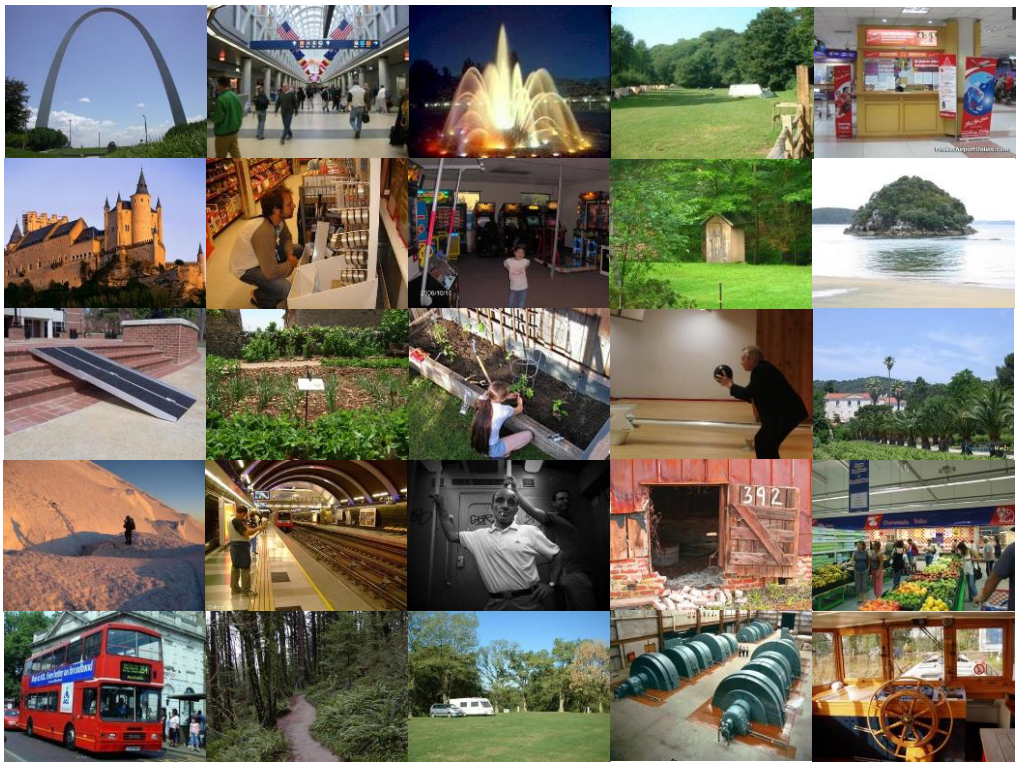


Figure 4.2: Sample images from DUT-OMRON dataset

Self-dataset: This is a self-dataset proposed during the research work. Its main objective is for generating the hypothesis and testing purpose. This dataset consists of 50 images. The images are collected from the internet, and some of the images are

captured using high resolution Digital Single-lens Reflex (DSLR) camera. The images that are shot include images that are captured by alternating the focus regions, images with multiple in-focus regions and very high resolution images. Sample images from the dataset are shown in Figure 4.3. The ground truth (in-focus regions) related to these images is selected manually. Three people participated in the collection of the ground truth. The main purpose of this dataset is to develop an in-focus region detection algorithm.

In the literature, the majority of the saliency models have used only one dataset to validate their saliency models. Due to this the models are mostly tuned towards the dataset and are not independent of the characteristics of the dataset. Therefore, to make the model less dependent of the dataset characteristics, three datasets have been used in the project.

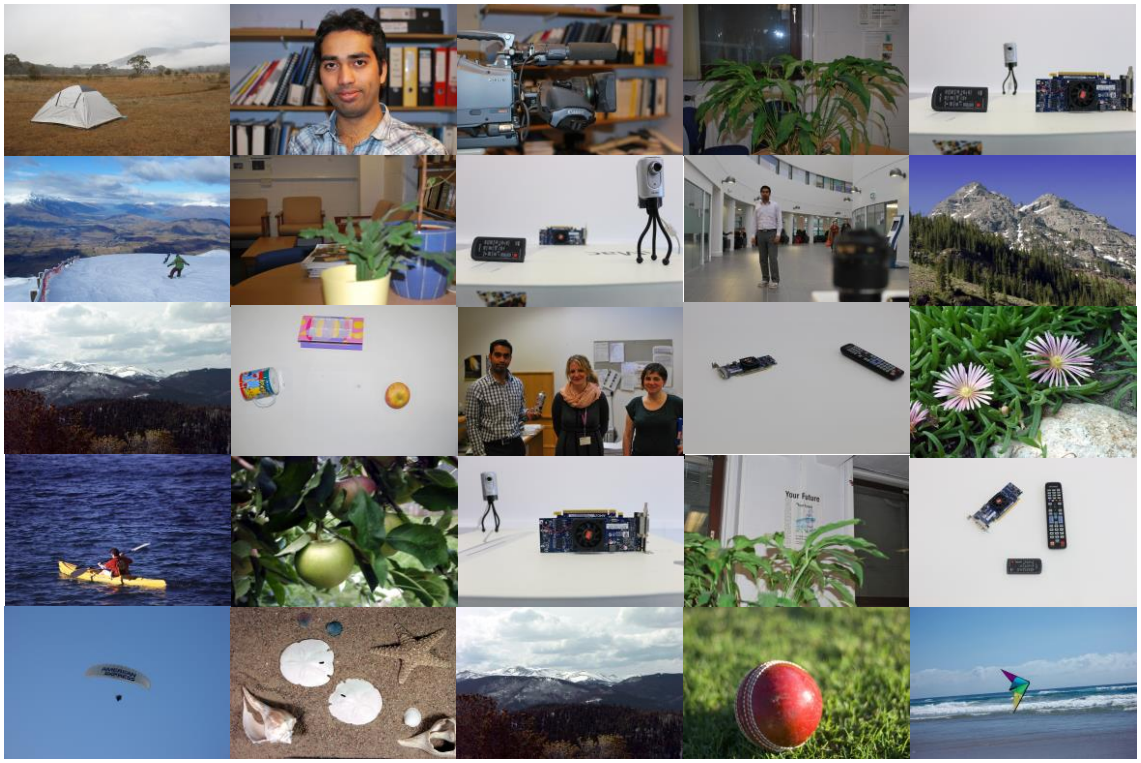


Figure 4.3: Sample images from self-dataset

4.4 Human Attention Map

A human attention map is a smooth, distributed and continuous representation of eye tracking information. All the datasets come with eye tracking data in the form of binary MATLAB format files. It is also referred to as raw eye tracking data. The pixel co-ordinates are extracted from these files to develop raw eye-tracking maps. The fixation map algorithms utilise these eye-tracking maps as the input to generate the

output human attention map. The saliency maps from different models are compared with these human attention maps to calculate their performance.

A sample output of the eye tracker when viewing an image for 4 seconds is shown in the Figure 4.4. From the figure it can be seen that the eye tracker captures the pixel co-ordinates of the attended units, the start time, end time and the overall duration of the fixation. Further, it also provides the average and the overall duration of the fixations.

Fixation parameters:
Maximum gap between gazepoints (seconds): 0.500
Minimum fixation time (seconds): 0.200
Minimum fixation diameter (pixels): 50

Number of fixations: 7

Fixation Listing: (fixation number, x position, y position, begin time, end time, duration)

1. 636, 458, 0.030, 0.250, 0.220
2. 648, 370, 0.280, 0.881, 0.601
3. 648, 330, 0.881, 1.282, 0.401
4. 457, 530, 1.412, 1.712, 0.300
5. 983, 441, 1.983, 2.454, 0.471
6. 640, 362, 2.834, 3.165, 0.331
7. 641, 460, 3.205, 3.966, 0.761

Average fixation duration (seconds): 0.441
Total fixation duration (seconds): 3.085

Figure 4.4: An example of eye-tracker data for one human subject

Based on this information a raw grey scale eye-tracking map can be created by marking the fixated locations on the images as ones and the unattended pixels as zeros. A raw eye tracking map (binary image) is a union of all fixations by all the observers across the image. A natural scene image with a single human subject's eye tracked fixations and scan path overlaid is shown in Figure 4.5. In the right image of the figure the fixated regions are numbered according to their order. The dots between the fixated regions show the scan path.

Figure 4.5: A natural scene image (left) and fixations overlaid (right) of one viewer

The fixation map algorithms take the raw eye tracking information as input to generate the output fixation or human attention map. The human attention map is also referred to as fixation density map. In the literature several strategies are used to obtain human attention maps. In the work of [142], the authors use fixation duration as parameter for generating the fixation maps. Kootstra and Shoemacker [143] transformed eye tracking information into fixation-distance maps. In their approach, the inverse distance transform of the fixation data is calculated. The distance transform approach computes the distance to all pixels from a fixation. It gives the probability that a fixation will be at a certain location of the image based on the eye tracking data. Recently Judd *et.al* [12] convolved the fixated locations using a Gaussian filter to obtain the human ground truth. A similar approach is followed by the authors of [141] to generate the human fixation maps. There is much variability among the methods used by the researchers to obtain these fixation maps. In the research community there is still no unique consensus regarding the fixation map algorithm to be used for generating fixation maps.

In this project, datasets from different authors are used to evaluate the performance of the attention models. The raw eye tracking information from all the

chosen datasets is extracted and a common fixation map generating strategy is used to generate the fixation maps. Similar to the authors of [12] the fixations of raw eye tracking map are convolved to obtain the human attention maps. Using a common fixation map generating scheme is an important step when testing a model performance across several datasets. For the same dataset and a saliency model, the fixation maps obtained with different strategies will show a variation in the results produced. A sample image and its corresponding fixation map from Judd *et.al* dataset is shown in Figure 4.6. The left image in the figure shows fixations collected from 15 viewers. The right image is the fixation map obtained using Gaussian convolution.

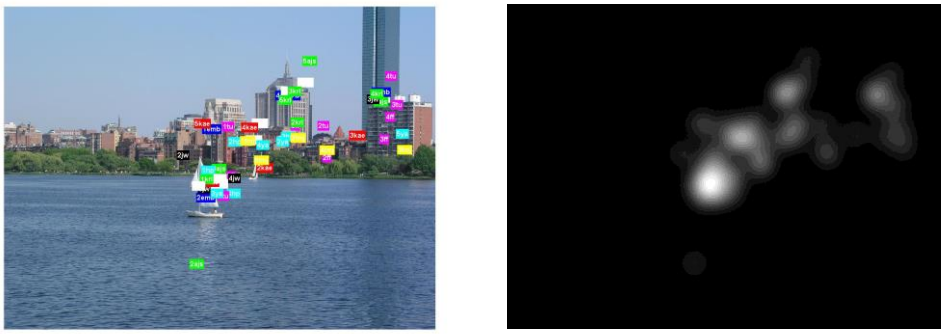


Figure 4.6: Image with fixations overlaid from 15 users (left) and fixation map (right)

4.5 Empirical Validation of Visual Saliency Model

During the process of validation of a visual saliency model the saliency map produced by the visual saliency model is compared with the human attention or fixation density map derived from the eye tracking information. Empirical evidence which involves direct observation of model's performance is analysed through quantitative and qualitative techniques. These two types of techniques are explained below.

4.5.1 Qualitative Analysis

The qualitative analysis is based on the subjective appreciation of the correlation between human attention map and computational attention map. During this process the viewer analyses both the human and computational attention map through visual comparison. It gives an approximate idea about the map correlations. In this process the image, the human map of attention and the computational attention map are put side by side to observe the correlations. Apart from visually measuring the correlation the qualitative analysis also helps in generating or improving the hypothesis for developing novel saliency models. During the visual comparison new information

emerges that helps in revising the research direction. Moreover, by observing different human scan path patterns and their corresponding saliency maps, the limitations of the saliency model in terms of predicting the human gaze can be easily identified. A sample qualitative analysis for four images with four computational attention models is shown in Figure 4.7. In the figure the first row contains the images chosen for predicting the salient locations. The second row is the Human Ground Truth (HGT) fixation maps. The remaining rows are the saliency maps under qualitative test from different models. A saliency map is a grey map and falls in the dynamic range [0, 255]. A higher grey value is treated as highly salient. Visually a higher grey value pixel appears with higher intensity. Therefore, pixels with higher intensity on the fixation map indicate that they are more salient when compared to the rest of the pixels that correspond to the image. It can be seen from the figure that PQFT has least amount of correlation to the HGT maps as it mostly detects edges and maps are mostly sparse. The SUN model although it is more smooth and contagious detects background and few salient locations in the image.

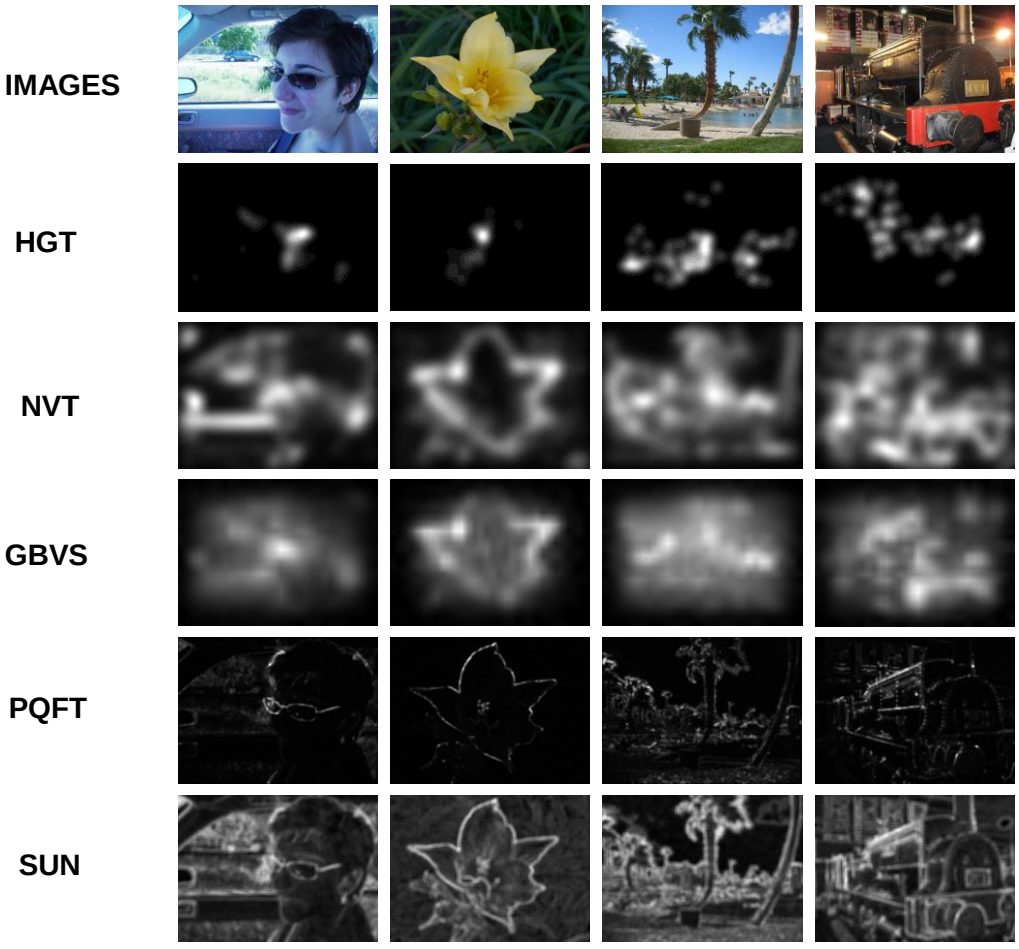


Figure 4.7: An example for Qualitative analysis

The maps from NVT highlight the majority of the saliency locations. The locations are mostly localised, continuous and smooth similar to the fixation map. However, the only drawback is the false detections (non-salient background regions) being detected as visually salient similar to the true detections (salient regions). In the case of GBVS model the authors tried to limit the false detections by ignoring the periphery of the image and gave higher weights as they approach towards to the centre of the image. By doing this the false detections are reduced and thereby higher correlation with ground truth human maps is achieved.

Qualitative analysis provides more depth and detail regarding the saliency maps that are studied and is helpful for hypothesis generation. However, there are also some limitations for this approach. It is complex and time consuming if this entire process needs to be implemented across a dataset of few hundreds of images. As only few images under test are generally studied it is not possible to generalise results to that of the entire dataset. The decisions made through qualitative analysis are heavily dependent on the skills of the researcher and hence can be biased towards his or her tendencies.

4.5.2 Quantitative Analysis

The objective comparison of the computational attention map and the human attention map involves the usage of mathematical metrics which determines how far an attention map correlates with the human attention map. A block diagram for quantitative analysis of the visual saliency map is shown in Figure 4.8. A quantitative analysis metric generally takes two inputs, namely the human fixation map or raw eye tracking map and the saliency map of an image generated by the attention model under test. The metric outputs a numerical value as measure of comparison between the two inputs.

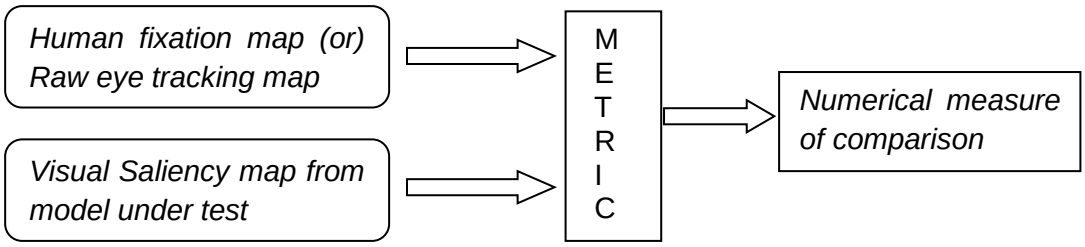


Figure 4.8: A Block diagram of the quantitative analysis of the visual saliency map

To make a fair comparison, to provide diverse picture of the overall model's performance and to make the conclusions less dependent of the choice of the metric

three metrics are used in this project. The metrics used in the work are Correlation coefficient (CC), Normalised Scanpath Saliency (NSS) and Area Under Curve (AUC). These metrics are explained below.

Linear Correlation Coefficient (CC)- CC is a statistical measure of the linear relationship between two variables. The authors of [46], [144] used it to evaluate the performance of their saliency models. Let G represent Ground truth fixation map, S represents saliency map from an attention model. Then the correlation between these two maps is defined as

$$CC(G,S) = \frac{\sum_{x,y}(G(x,y)-\mu_G)\cdot(S(x,y)-\mu_s)}{\sqrt{\sigma_G^2 \cdot \sigma_S^2}} \quad (4\text{-}1)$$

Where (x, y) represents image pixel coordinates, μ and σ^2 are the mean and variance of the values in the corresponding fixation and saliency maps. For the given two input maps the metric outputs a single scalar value which has an upper bound of 1. The value of CC lies in the [-1, 1] interval. When the value is close to +1/-1 there is a perfect linear relationship between the Ground truth fixation map and computation attention map. A value of 1 indicates that both the maps are similar. A value of 0 indicates that both the maps are totally different. A value of -1 indicates that both the maps are anti-correlated, i.e. a salient feature in one of the maps is completely non salient in the other map.

Normalised Scanpath Saliency (NSS): The authors of [46], [145] and [146] used NSS for validating their saliency models. In this approach, initially the computational attention map is linearly normalised to have zero mean and unit standard deviation. Next the normalised values at each point in the saliency map (S) that correspond to the fixated locations in the human attention map (x_H^i, y_H^i) are extracted. Finally, the NSS is calculated by taking the average or mean of all the extracted values.

$$NSS = \frac{1}{N}\sum_{i=1}^{N}\frac{S(x_H^i, y_H^i)}{\sigma_s} \quad (4\text{-}2)$$

Where, N is the total number of fixations for each image. An NSS value greater than or equal to one indicates that there are significantly higher saliency values at the human fixated locations in the saliency map. The higher the value, the better is the performance of the saliency model in predicting human fixations. An NSS value of zero indicates that the model performs no better than the random model and it mostly predicts the salient locations by chance. A value less than zero indicates that the model is predicting non-salient locations as salient.

Area Under Curve (AUC): AUC is the acronym for **Area Under** receiver operating characteristic **Curve** (AUC). This metric is widely used in the research community for validating the saliency models [138], [36], [147]. The metric is explained in two parts. Initially the process of drawing the ROC curve is explained and then the AUC is discussed.

In the process of evaluation both human attention map and saliency map are needed as inputs to draw an ROC. First all the pixels attended in the human map are considered as *fixated* pixels and the unattended pixels as *non-fixated* pixels. Next the saliency map from the saliency model is normalised between the range [0, 255]. Then the saliency map is treated as a binary classifier and the threshold is varied on all the probable thresholds ranging in [0,1,....254, 255]. On each threshold the saliency maps are binarised into foreground and background and the number of True Postives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) are calculated.

True Positive (TP) (*foreground and fixated*): A point or location is fixated in the eye tracking fixation map. The saliency model also predicts the same point or location as salient in the saliency map.

False Positive (FP) (*foreground and non-fixated*): A negative response is falsely predicted as Positive. A point or location is not fixated in the eye tracking map. The fixated point is not salient but the model in its saliency map detects it as salient.

False Negative (FN) (*background and fixated*): The pixel in the image is actually attended by the human viewer; however, the saliency model considers it as a non-salient pixel.

True Negative (TN) (*background and non-fixated*): The pixel in the image is not fixated in the human ground truth map. The saliency model also predicts that the pixel is not salient in the visual saliency map.

Consequently at each threshold the True Positive Rate (TPR) and False Positive Rate (FPR) is calculated as:

$$FPR = \frac{FP}{FP + TN} \tag{4-3a}$$

$$TPR = \frac{TP}{TP + FN} \tag{4-3b}$$

The ROC is a two-dimensional curve with the Y-axis showing TPR and X-axis FPR. These pairs (TPR, FPR) are plotted at all thresholds. Therefore, by plotting true positive rate vs. false positive rate an ROC curve is achieved. It is also referred to as the average sensitivity (True positives) over the entire range of possible specificities (False positives) or the average specificity (False positives) over the entire range of possible sensitivities (True positives). A sample receiver operating curve is shown in Figure 4.9.

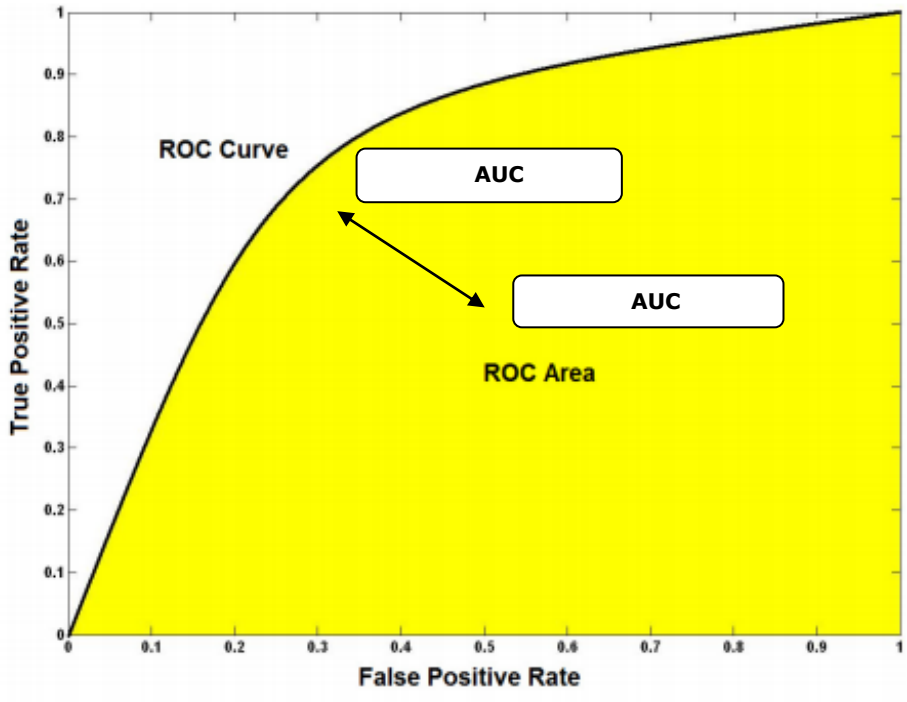


Figure 4.9: An example of Receiver Operating Characteristic (ROC) curve

The area underneath the ROC curve gives the AUC of the saliency model as shown in the figure. When the ROC curve moves towards the top left, the AUC increases and it decreases when the curve moves towards bottom right. The AUC on a dataset of images can be usually calculated in two different ways.

1) The AUC is calculated for each image of the dataset. Then the mean of all AUC scores is considered as AUC of the saliency model on the dataset.
2) At each threshold the TP's, FP's, FN's and TN's of all the images in the dataset are summed up and a unique ROC curve is drawn. The area under this final ROC curve gives the AUC score for the entire dataset. In simple terms it is taking average of all ROC curves. Both of these ways are used in the literature. However in recent works, the first approach is preferred.

In this project, the Area Under Receiver operating Characteristic (AUC) curve is used as the global indicator of the model's performance by considering all images of the dataset. A perfect prediction of the salient locations gives a score of 1 for AUC. A score of 0.5 indicates a chance level. A good saliency model AUC should lie between 0.5 and 1 and close to 1 for better performance. AUC is chosen as it quantifies the ROC and helps in finely comparing the models' performance when models with overlapping ROC curves or visually indiscriminative ROC curves are present in the final ROC graph with many models. For some models the relative difference in the performance is very small. Therefore, providing the numerical AUC values will help in identifying minor performance variations and also enable other researchers to effectively compare new saliency models against the model proposed in this work.

In this project metrics such as CC, NSS and AUC are used in evaluating the saliency model's performance due to their wide usage in the research community. Moreover, still there is no unique consensus among the researchers regarding the validating metrics [148], [130] to be used during the performance evaluation. There are many advantages of quantitative analysis over visual comparison of saliency maps. A metric can be used to study the model's performance over a dataset of several hundreds of images. Therefore, it helps in generalisation of results and allows for broader study with greater objectivity and accuracy of results. As it is a numerical output, it helps in identifying the fine differences in the model's performance. Some of the disadvantages include less elaborate human perception because of numerical descriptions.

The image datasets in the section 4.3, the fixation map strategy explained in the section 4.4, qualitative and quantitative methods described in this section are used to develop a novel visual attention model in the next chapter. Initially a self-dataset of images is used to generate the hypothesis. Later to validate and compare the prediction accuracy of the model, qualitative and quantitative analysis of saliency models is done across the chosen image datasets. Using blended qualitative and

quantitative approaches is generally referred as methodological triangulation. This process helps in verifying (confirming/rejecting and reinforcing) results from qualitative data using quantitative data or vice versa. In the next section, the procedure followed for measuring and comparing computational complexity of saliency models is explained.

4.6 Computational Complexity

In the following sub sections different kinds of methods and the important factors that are taken into account during the measurement of computational complexity are explained.

4.6.1 Methods for Measuring Complexity

Software profiling: The majority of the visual saliency models are developed in MATLAB. Therefore, the time complexity of all the models in literature along with the proposed work in the next few chapters is measured in MATLAB. During the development and testing phase of the model complexity, the MATLAB in-built profiler is utilised to identify the functions that are spending more time within the model. The profiler summary report with a graphical interface gives various details related to number of calls, total time and self-time of the functions of the model under test. This helps in identifying the time consuming functions or models that needs to be improved in efficiency.

MATLAB Commands: MATLAB has provided three in-built functions for measuring program time complexity. These are *cputime*, *timeit* and *tic/toc* functions.

Cputime: Returns the total CPU time used by MATLAB application since it was started. It does not take any first time costs into account.

tic/toc: tic starts a stopwatch timer to measure performance. Tic records the internal time of the execution of tic command and the elapsed time is displayed using the toc command. As it starts a timer, it is beneficial to use this for measuring time of portions of code within a function. As it is a timer, it does not take first time cost into account.

Timeit: It measures the total time required to run the function. It calls the function several times and then computes the median. The first time costs are taken into account during the computation of time. Therefore, the timeit function is used in

this research project to calculate the complexity of the model. When needed even tic/toc functions are also used for calculating the time of the portions of code within the functions. Precautions such as shutting down of background programs that will have an influence on complexity measurement are also considered.

4.6.2 Factors Considered During Complexity Measurement

Unoptimised versions: The majority of the saliency models are implemented in MATLAB. However, some of the models are also implemented in optimised MATLAB code (MEX code versions) and the C programming language. To ensure a fair comparison of time complexity unoptimised MATLAB implementations (without MEX code) are used for all the chosen models in the benchmark.

Resolution and Number: Image resolution is one of the important parameters which affects the computational complexity of the saliency model. In general a higher resolution usually requires higher time for computing saliency. In our comparison the average time required to compute a saliency map is calculated over 100 images with resolution 1024x768.

Content independency: The complexity measurement can also be influenced by the type of the image content such as contrast and detail. To make the complexity measurement content independent the complexity of the model is calculated across 100 images chosen from two different datasets.

Common conditions: A common environment and testing platform is used for measuring complexity of all the models. The details regarding these are provided as a separate section at the beginning of this chapter.

Criteria for Comparison: In a time constrained scenario, a saliency model should be fast enough to meet the real time performance requirements. In the literature, the majority of the saliency models targeted at achieving a very high performance in predicting human fixations. This is generally referred to as the prediction accuracy of the saliency model. There are other models whose main aim was to achieve real time performance. In this work, both complexity and accuracy are chosen as the criteria for the development and comparison with the state-of-the-art.

4.7 Benchmark Visual Saliency Models

The model which will be proposed in the next few chapters is compared with 10 state-of-the-art visual saliency models. In this area as the state-of-the-art is changing very rapidly, care has been taken to ensure that very recent visual attention models are included in the benchmark. All of these models are explained in detail in chapter 3. The majority of the models used for the benchmark were proposed in the last five years. The complete selection of models for comparison is based on wide citation, popularity, recency and model characteristics. The model characteristics include biologically plausible models, frequency based approaches, spatial, models that consider bottom-up and top-down elements in detecting saliency. These models are collected in different ways. Some of the saliency models are shared online. A few of the models were collected by contacting the creators. The authors sent us the source code for us to compile or the executables.

Table 4.1: Chronological listing of visual saliency models

Year	Models
1998	NVT [6]
2006	GBVS [7]
2007	SR [8]
2008	SUN [70]
2010	PQFT [10]
2010	CAS [9]
2012	SS [11]
2012	RCSS [122]
2013	SDSP [13]
2013	WBSD [127]

However, some of the authors preferred to run their model on the images and sent us the results. One model was implemented directly by reading the author's published paper. All of these models are listed here chronologically as shown in Table 4.1. The Neuromorphic Vision Toolkit (NVT) popularly known as Itti's model was introduced in 1998. This is a widely cited model based on bottom-up characteristics with biological plausibility. Graph Based Visual Saliency (GBVS) proposed in 2006 is a bottom-up attention model similar to NVT. This model is highly popular, widely cited in the literature and very good at predicting human fixations. Spectral Residual (SR)

proposed in 2007 is a spectral based approach for predicting human fixations. This is a very fast model and is based on bottom-up characteristics. Saliency Using Natural Statistics (SUN) (2008) is model based on both bottom-up and top-down characteristics. It is a popular and widely cited saliency model. Phase Quaternion Fourier Transform (PQFT) proposed in 2010 is a popular, widely cited and one of the recent bottom-up visual saliency models. The model is also fast at predicting fixations. Context Aware Saliency (CAS) proposed in 2010 is a very popular and widely cited model. It incorporates top-down features for detecting salient regions. Signature saliency (SS) a very recent model proposed in 2012. It retrieves the salient regions in the frequency domain. Random Centre Sorround Saliency (RCSS) is another recent model proposed in 2012. It is chosen for the benchmark comparison as the authors achieved high prediction accuracy on two large datasets. Saliency Detection using Simple Priors (SDSP) has considered both prediction accuracy and complexity as the main constraints in the development of the model. It is proposed in 2013 and is an important choice for the benchmark as the model in the current research work also considers both accuracy and complexity. Wavelet Based Saliency Detection (WBSD) (2013) is a frequency based approach based on bottom-up features. All of these models will be used as the benchmark for proposing a novel saliency model in this research project.

4.8 Conclusion

This chapter presented the experimental methods used in the research project. Based on the objectives and requirements of the project a suitable development environment and testing platform is chosen. The performance of the state-of-the-art models is used as the benchmark for the development of the models. Publicly available image datasets in which the human eye analysis studied using eye trackers are selected for the project. A self-dataset is also created for developing novel attention model. All these datasets are used to evaluate the performance of the models. The models performance is evaluated with respect to both prediction accuracy and computational complexity. To empirically validate the model qualitative and quantitative approaches have been utilised. The qualitative analysis includes visual comparison and the quantitative analysis includes the usage of mathematical metrics like CC, NSS and AUC. For assessing and comparing the time complexity, MATLAB's in-built profiler and in-built functions are utilised.

PART TWO: EXPERIMENTAL WORK

5 A DCT Based In-Focus Bottom-up Visual Attention Model

5.1 Introduction

This chapter presents a novel low complexity visual attention model for predicting regions of interest in images. The model detects visually salient regions based on camera focus. The salient frequencies present in the in-focus areas are detected using the characteristics of Discrete Cosine Transform (DCT) coefficients. The saliency map is developed using a mathematical model developed by observing the amplitudes of 8x8 image block DCT coefficients. The DCT based focus maps are convolved and contrast stretched to obtain the salient regions of an image. The performance of the developed model is evaluated against popular visual attention models. The results indicate that the model achieves better prediction accuracy in saliency detection at significantly lower computational complexity compared to some of the benchmark attention models.

In section 5.2 the hypothesis behind the proposed visual saliency model is explained. The directly related work to the visual saliency model based on DCT and image focus detection is discussed in the section 5.3. Section 5.4 describes the fundamentals of DCT. The different types of frequencies in the DCT transformed image and its relation with the visual content (image detail) is explained. It also provides the important properties of DCT that are relevant to the current development. The development of the attention model is provided in two phases in section 5.5. The experimental results related to prediction accuracy and complexity of the model is given in section 5.6. Finally, the important aspects of the attention model, the advantages and disadvantages of the developed model are discussed in section 5.7.

5.2 Hypothesis

One of the primary ways to lead the viewer's attention to a specific region of an image is to bring the region in to focus. Photographers adjust the camera's focus into the regions of interest rather than the background when capturing images. An "in-focus" region in an image contains contents which are of interest to the human eye than an "out-of-focus or blurred" region [149]. Objects which are in focus are sharper and appealing to the viewer and these regions are bottom-up in nature. Therefore, in-focus region selected by the photographer is a good candidate for saliency detection. Typically, the camera is focussed into different regions in images so that the human visual system (HVS) follows these regions to understand the information content.

These in-focus regions may constitute either bottom-up or top-down features of an image. For example, the most common top down features are the human faces, animals and cars. The other top down features are hands and eyes of the people. Whereas, the bottom-up features are specific objects which pop out from the image to grab the attention of the viewer. Usually they pop out because of their own attributes such as intensity, colour, orientation etc. (Please refer to section 3.2.2.1).

In the literature bottom up features are investigated more when compared to top down features. The main reason behind this is that the bottom-up features are easier to control when compared with cognitive factors such as expectation and knowledge. There are significant differences between individual human subjects and hence they have their own strategies of directing their gaze from one point to another and this gives rise to extreme difficulty in modelling the top down features [150]. If the very nature of the HVS has to be imitated, then both bottom up and top down saliency characteristics have to be fused together in order to obtain the region of interest. A saliency model with a mixture of several top down and bottom up features will result in a complex and time consuming attention model.

In most cases, a photographer focuses on regions that convey information and these regions may have either bottom-up or top-down features. The advantage of detecting these in-focus regions is that it reduces the complexity to detect both top-down and bottom-up regions separately to some extent.

5.3 Directly Related Work

The visual attention models in the literature that are based on detecting salient regions using image focus and DCT domain are briefly discussed in the following sections.

5.3.1 Focus

Ki Tae *et al.* [151] used the sharpness of the regions in the image as a measure to differentiate in-focus and out-of-focus regions. The DCT is performed on each of the Y, C_b and C_r channels of an image and later, by calculating Bayes entropy of the DCT coefficients, three focus maps are generated. A saliency map is then generated by combining these three focus maps. The intensity of focus in the images is inversely related to the degree of the blurriness in an image. Based on this concept, most of the existing algorithms compute blurriness as a way to detect the in-focus regions of an image. The authors of [152] calculate the spatially varying amount of blur by estimating the blur kernel at the image edges. They propagate this defocus measure over the entire image using non homogeneous optimisation. However, their idea is limited to

regions with smooth interiors. Zhuo and Sim [153] used an image matting to compute the blurriness. However, their blur estimation cannot tell whether the blur edge is caused by defocus or blur texture which includes soft shadows and blur texture. Moreover, the canny edge detection and joint bilateral filtering increased the computational complexity of their method. More recently the authors of [154] detect salient regions by measuring blurriness using scale space analysis. In their work for each edge pixel Difference of Gaussian (DOG) responses are calculated at different scales. Later on the degree of blur is estimated and pixel level focus is approximated. Although as the authors claimed, their approach has a solid mathematical proof, it resulted in an increased complexity because of several DOG operations.

5.3.2 DCT based Attention Models

In the research by Yiwei and De [155] the properties of DCT coefficients are used to obtain the bottom-up features of visual attention. The bottom-up intensity and colour features are obtained from the DC coefficients in the luminance and chrominance images respectively. The first (DCT (1, 0)) and the second AC coefficients (DCT (0,1)) in zig-zag scanning order from the DCT transformed block are used to obtain the orientation feature. Later the saliency map is generated by combining the feature maps. The authors of [156] used quaternion DCT signatures and face detection to detect the salient regions of an image. They have transferred scalar and real DCT signatures to quaternion images. Since faces are the prominent top-down feature, they are detected by using a face detection algorithm and later combined with the generated saliency map to improve the overall saliency accuracy. The authors of [157] calculated the saliency of videos by computing luminance, colour, texture and motion features from DCT coefficients and motion vectors in the video stream. They generated a static saliency map from luminance, colour and texture features and dynamic saliency map from motion feature. These two saliency maps are later fused to get the final saliency map.

The properties of the DCT and its effective utilisation in the works discussed above have motivated us to use DCT in the development of current attention model. The DCT and its properties that are relevant to the current development of the model are briefly explained in the next section.

5.4 Discrete Cosine Transform (DCT)

The DCT transforms an N x N square matrix of pixel values to an N x N square matrix of frequency coefficients. The two dimensional DCT applied on an image is defined as

$$C(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{Y=0}^{N-1} f(x,y)\cos\left[\frac{\prod(2x+1)u}{2N}\right]\cos\left[\frac{\prod(2y+1)v}{2N}\right] \qquad (5\text{-}1)$$

Where $u,v = 0,1,2,3,.....$N-1 and $x,y = 0,1,2,3,.....$N-1.

In the above two equations $\alpha(u)$ is defined as

$$\alpha(u) = \sqrt{\frac{1}{N}} \qquad (5\text{-}2a)$$

Where $u = 0$

$$\alpha(u) = \sqrt{\frac{2}{N}} \qquad (5\text{-}2b)$$

Where $u \neq 0$

$C(u,v)$ represents the frequency coefficients of DCT transformed block and $f(x,y)$ represents the pixel values of the input image data respectively. Generally in most of the cases N is 8. A bigger block results in a better image compression however, it takes more computation time in performing DCT calculations. As a tradeoff, during DCT implementations the image is broken down into manageable 8x8 blocks.

It is clear from equation (5.1) that when u and v equals zero then $C(0,0) = \frac{1}{N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)$. Therefore, the first coefficient is the average of all the image data. It is referred to as the DC coefficient and all other coefficients are AC coefficients.

In the DCT domain the entries are organised according to the sensitivity of HVS. Low frequency coefficients are placed in the top left corner of the 8x8 matrix. Similarly high frequency coefficients are arranged in the bottom right corner of the matrix. The DC component corresponds to the average of all the input image data. The DC component in a Y-channel extracted from YC_bC_r colour space is related to the brightness of the image. The low frequency components of Y-channel represent general luminance whereas high frequency components represent contours and drastic

changes of luminance [158]. Large high frequency coefficients indicate that the information is changing rapidly on a very short distance scale. For example, in an image of newspaper, the text keeps changing rapidly and when transformed, such detail is retained in the large value of high frequency coefficients. Whereas, large low frequency coefficients indicate large scale features of a picture are more important. Objects that are homogeneous and occupy most of the image area tend to be retained in large magnitude low frequency coefficients.

The properties of DCT that are relevant to the current development of the visual saliency model are briefly explained below.

5.4.1 Data De-correlation

In an image there exists a high level of redundancy between the neighbouring pixels. This is referred to as the data correlation. The DCT transforms spatially correlated image data to uncorrelated frequency coefficients. These transformed coefficients can be dealt independently of one another during the development of the attention model.

5.4.2 Energy Compaction

Energy compaction is the ability to transform the data into a few large valued coefficients. Due to this there will be fewer coefficients in the DCT domain that are sensitive to the HVS. The other coefficients can be discarded as they are least important.

5.5 Development of Focus Detection Algorithm

The focus detection algorithm is developed in two phases. In the first phase, an initial hypothesis for detecting in-focus regions in the images is constructed. The hypothesis is then manually verified across a range of test images. In the second phase an algorithm is developed to realise this hypothesis.

5.5.1 Hypothesis

The in-focus regions in the images are sharp and are of interest to the HVS. It is assumed that in-focus regions will contain some frequency coefficients that hold large values when compared to out-of-focus regions.

5.5.2 Development Phase I

In chapter 4 section 4.3 a self-dataset has been described. The dataset consists of images with in-focus regions. Some pictures are randomly selected from the dataset to test the hypothesis. The luminance component of YC_bC_r image is extracted.

The Y-channel is mainly selected because of the higher sensitivity of HVS to image brightness when compared to colour information. The image is partitioned into a number of 8x8 distinct square blocks of pixels. To avoid the intense computation of taking DCT over the entire image, the image is divided into 8x8 blocks and the DCT is applied to those blocks.

(a) Sheep Y-picture **(b) Sheep - Focus map**

Figure 5.1: Sheep (a) Y-picture (b) Focus map

The frequency coefficients in the 8x8 DCT transformed blocks corresponding to the in-focus and out-of-focus are visually analysed. This analysis revealed that there are frequency coefficients in the in-focus DCT blocks whose amplitude is very high when compared to out-of-focus frequency coefficients. These frequency coefficients from all the 8x8 blocks of the image are extracted, summed (by taking their absolutes) and displayed as a focus map. A sample Y-picture and its focus map is shown in the Figure 5.1. In Figure 5.1 (a) the enclosed region marked by an outline is in focus. It can be seen in Figure 5.1(b) that a sparse focus map is generated by taking the sum of the frequency coefficients which has very high magnitude, which indicates the in-focus region in the Y-picture.

Later this is verified across many images by manually locating the frequency coefficients that have higher amplitude than out-of-focus coefficients. To apply this to several images an algorithm which can automatically locate the frequency coefficients is needed. Therefore, this development is further carried out in phase II.

5.5.3 Development Phase II

The main objective of phase II is to develop an algorithm which can automatically locate frequency coefficients that correspond to in-focus regions. To achieve this during the second phase the in-focus region of the image is considered as the foreground region and the out of focus as background region. The foreground superblock $FS_{8\times8,}$ background superblock $BS_{8\times8}$ and full image superblock $IS_{8\times8}$ are calculated as follows:

$$[FS]_{8x8} = \frac{1}{M}\sum_{c=1}^{M}\left|[f_c]_{ij}\right| \tag{5-3}$$

$$[BS]_{8x8} = \frac{1}{N}\sum_{c=1}^{N}\left|[b_c]_{ij}\right| \tag{5-4}$$

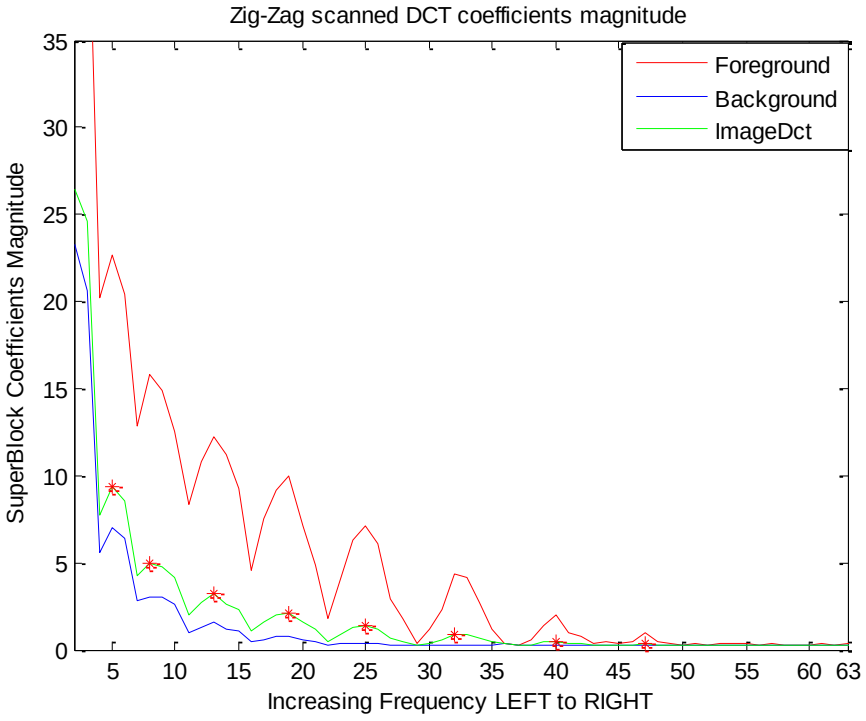
$$[IS]_{8x8} = \frac{1}{(M+N)}\sum_{c=1}^{M+N}\left|[I_c]_{ij}\right| \tag{5-5}$$

Where, $i \in (0,1,..,7)$, $j \in (0,1,..,7)$. f_c, b_c and I_c denote foreground, background and image (all) DCT blocks respectively. M is the number of foreground 8x8 DCT blocks and N is the number of background blocks. The (M+N) accounts for the total number of 8x8 blocks of an image. To obtain the foreground superblock, initially the in-focus region is manually selected and then the foreground superblock is calculated by taking the mean of the absolute of 8x8 DCT blocks that correspond to the selected in-focus region. Similarly the background superblock is calculated. Finally the image superblock is calculated by taking the sum of foreground and background superblock and dividing it by the total number of 8x8 blocks in the image. These DCT coefficients in all the three superblocks are selected using a zigzag scanning method so that the low frequency components precede the high frequency components. These DCT coefficients when scanned in a zig-zag manner are subsequently converted into a one dimensional vector. To determine the relationship between in-focus and out-of-focus region the frequency coefficients that are zigzag scanned from all the superblocks are plotted against each other on a graph. The DCT frequency coefficient amplitude patterns of all the three images are shown in the Figure 5.2, Figure 5.3 and Figure 5.4. In Figure 5.2 the sheep's head is highly in-focus relative to the background which is out-of-focus. This difference can also be seen in the Figure 5.2 (c) where at

each frequency coefficient on x-axis the difference between the amplitude of an in-focus and out-of-focus frequency coefficient tends to be significant. This difference is more significant at the peaks (peak is indicated by star on the ImageDCT waveform of Figure 5.2 (c), 5.3 (c) and 5.4 (c)) of the in-focus and out-of-focus amplitude waveform). Moreover the peaks of in-focus, out-of-focus and entire image almost coincide with each other. However, this difference tends to decrease towards the very high

(a) Sheep-RGB picture

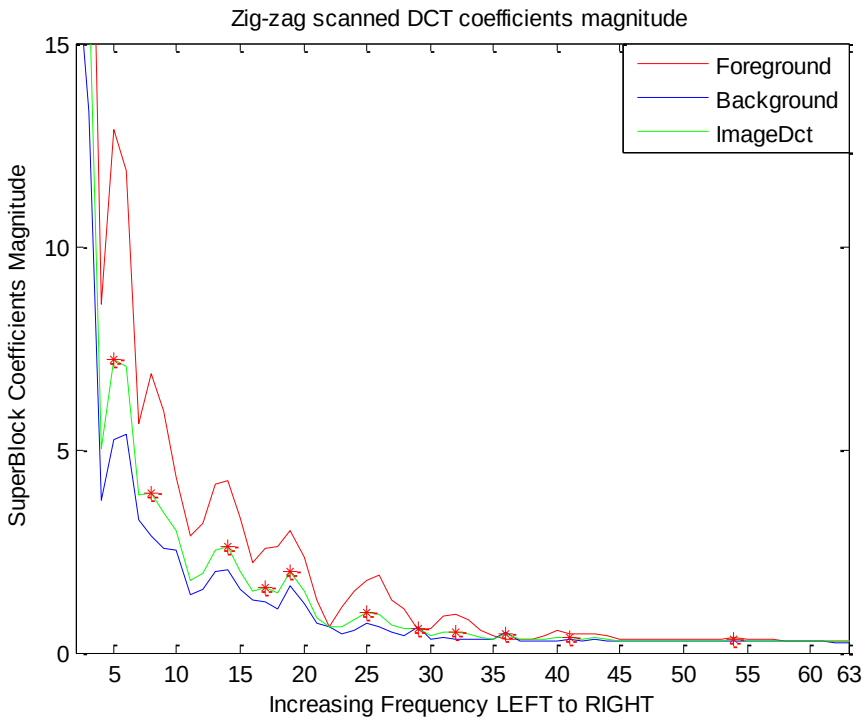
(b) Sheep-Y picture



(C) Sheep - Zigzag scanned DCT coefficients

Figure 5.2: Sheep (a) RGB picture (b) Y-picture (c) Zig-zag scanned DCT coefficients

(a) Cricket ball- RGB picture **(b) Cricket ball-Y picture**

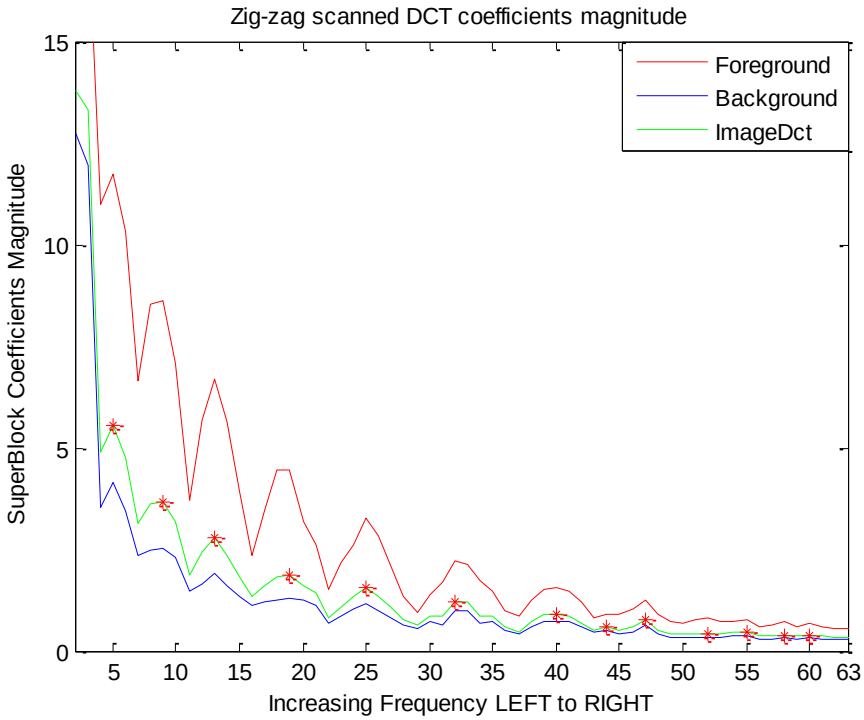


(C) Cricket ball Zig-Zag scanned DCT coefficients

Figure 5.3: Cricket ball (a) RGB picture (b) Y-picture (c) Zig-zag scanned DCT coefficients

(a) Face-RGB picture

(b) Face-Y picture



(c) Face-Zigzag scanned DCT coefficients

Figure 5.4: Face (a) RGB picture (b) Y-picture (c) Zig-zag scanned DCT coefficients

frequency range (51-63). In the very low frequency range very minimal variations are observed. In Figure 5.3 (a) the cricket ball and the grass in the bottom right corner of the image is highly in-focus and rest of the image is relatively blurred or out-of-focus. Similar to the Figure 5.2 (c) very minimal variations are observed in the very low

frequency range. However, in this image the difference falls off from the 45th frequency coefficient on the x-axis. Further as the difference falls off drastically, there are no peak frequencies observed. In the Figure 5.3 (a) the face and area around the shirt region is highly in-focus when compared to the background which consists of books. This image (similar to the images in Figure 5.2 (a) and Figure 5.3 (a)) exhibits too many frequency variations (many peaks) in range of 5-50. After the 50th coefficient there is a difference; however, there are no variations at all.

These graphs in the Figure 5.2 (c), 5.3 (c) and 5.4 (c) reveal that

i) The peaks of spatial frequency amplitude waveform of in-focus, out-of-focus and the entire superblock almost coincide.
ii) The maximum magnitude difference occurs at these peaks.
iii) This difference tends to be significant within a band of frequencies which excludes both very low and high frequency DCT coefficients.
iv) The peak magnitude frequencies in the in-focus regions are absent in the out-of-focus regions. Therefore, these peak frequencies can be used to calculate the in-focus regions in an image.

Noise is very high frequency information which occurs during image capture. At very low frequency including the DC coefficient, gradual changes are observed in both in-focus and out-of-focus regions. Therefore, a band pass filter can be used to eliminate these frequencies by inhibiting some of the very high and low frequency irrelevant DCT coefficients. The band pass filter selects zig-zag scanned frequencies between 5 and 60 ignoring the beginning 1-4 (very low frequency coefficients) and the last 51-64 (very high frequency coefficients). These frequencies are chosen based on the experimental results shown in the Figure 5.2 (c), 5.3 (c) and 5.4 (c). Later all the frequency positions that correspond to the peaks indicated by stars within the band of frequencies of the image superblock coefficients are identified and stored. A peak frequency consists of points that are lower by a value of x (peak threshold) on either of the sides. Empirically, by testing across different images and values, the peak threshold is determined as 0.1. Therefore, we need a difference of at least 0.1 between a peak and its surrounding for it to be declared as a peak frequency. These peak frequencies are plotted as red colour stars on the graphs.

Later the sum of corresponding absolute DCT coefficients of identified frequency positions in all of the 8x8 blocks of the image is calculated and stored. As discussed already a sparse focus map is obtained by this stage. These sparse focus

maps are shown in Figure 5.5 (a), Figure 5.5 (b) and Figure 5.5 (c). This sparse focus map is convolved using a Gaussian kernel (size $n \times n$) and contrast stretched by multiplying each pixel in the focus map with a factor k. The smoothing filter is used to generate connected regions from the sparse salient frequency plots. The contrast is stretched to improve the overall intensity of the focus map. The final focus maps are shown in Figure 5.6 (a), Figure 5.6 (b) and Figure 5.6 (c).

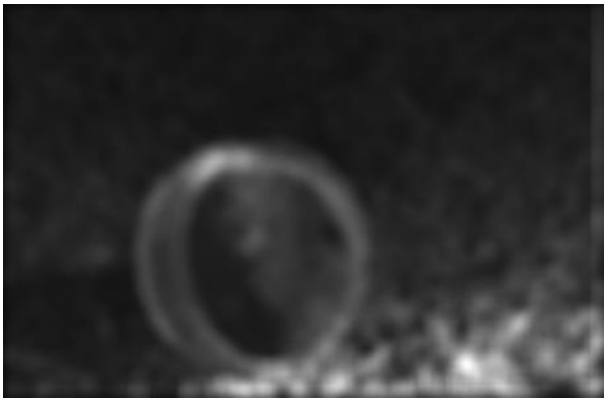
(a) Sheep- sparse focus map

(b) Cricket ball- sparse focus map

(c) Face-sparse focus map

Figure 5.5: Sparse focus map (a) Sheep (b) Cricket ball (c) Person face

(a) Sheep Saliency map



(b) Cricket ball Saliency map

(c) Face Saliency map

Figure 5.6: Saliency map (a) Sheep (b) Cricket ball (c) Face

5.5.4 The Complete Attention Model

The complete model based on in-focus regions is summarised below

1. Extract the Y component of YUV image.
2. Divide the luminance image into 8x8 blocks.
3. Perform DCT on all 8x8 blocks of the image.
4. Calculate image superblock according to equation **(5-5)**.
5. The image superblock frequency coefficients are selected using the zig-zag scan method and then the image superblock coefficients magnitude is plotted against the zig-zag scanned frequencies.
6. The Peaks of the above plotted graph are computed and the corresponding frequency positions are identified and stored.
7. Display the focus map by summing all the identified frequency position frequencies in all the 8x8 blocks.
8. Convolve and contrast stretch the map obtained.
9. Display the visual saliency map.

5.6 Experimental Results

The proposed visual saliency model is evaluated using the proposed self-dataset and Judd's public dataset which contains 1003 images. For more details

regarding these datasets refer to chapter 4 section 4. The quantitative and qualitative analysis with the chosen datasets is given in this section.

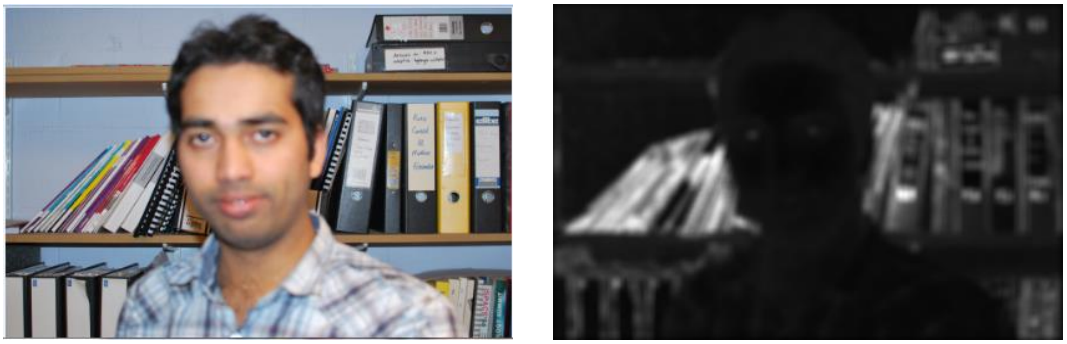
5.6.1 Qualitative Analysis of the Focus Map

In this section the performance of the focus detection algorithm is evaluated across the images in the proposed self-dataset. It can be seen in Figure 5.7 (a) that the person's face is in-focus and there exists many salient frequencies around the eyes, hair, around the shirt area and the edges. The algorithm detects the in-focus face region which can be seen in the Figure 5.7(b) focus map. The focus map highlights the salient frequency information.

(a) Face- (Face in-focus) **(b) Face - Focus map**

Figure 5.7: Face (a) In-focus (b) Focus map

The same image in Figure 5.8 (a) is captured by making the face out of focus. In Figure 5.8 (b) the person's face is out-of-focus and the salient frequency content exists in the background area. The algorithm detects this in-focus background region which can be clearly seen in the (Figure 5.8 (b)) focus map.



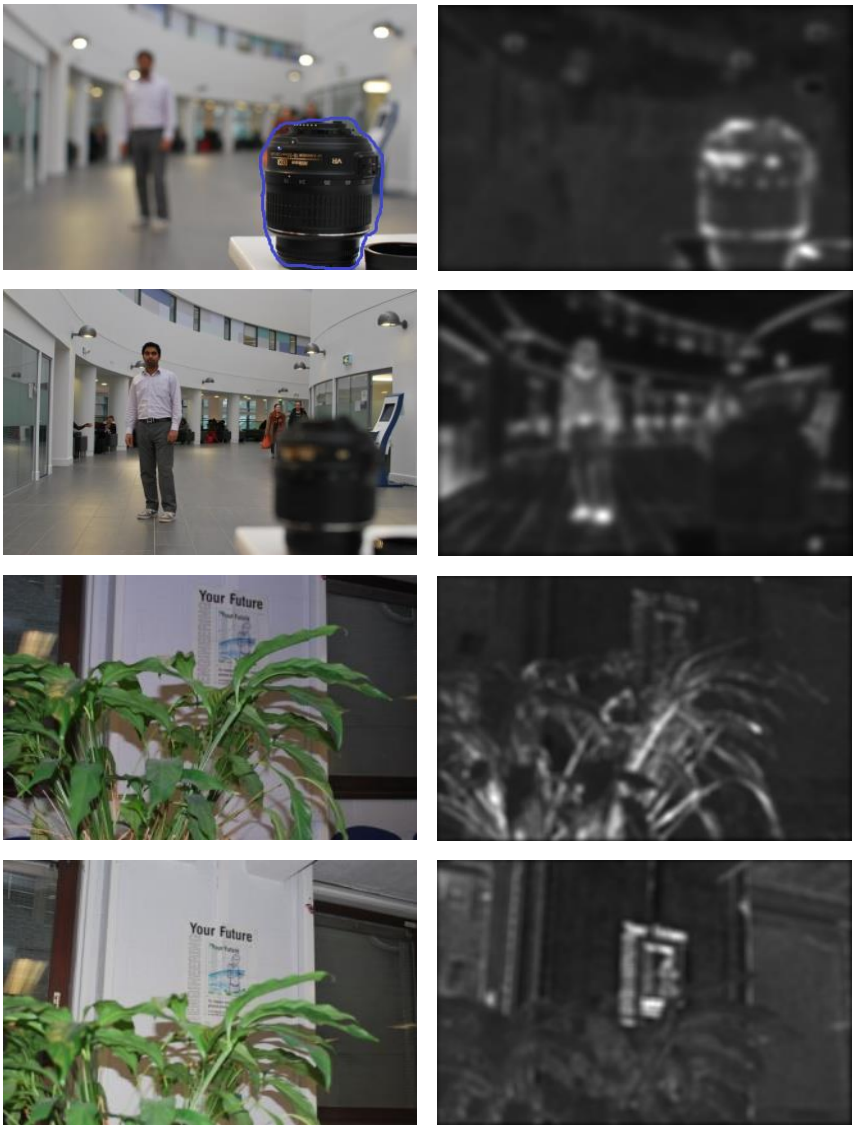
(a) Face-(Face out-of-focus) **(b) Face-visual saliency map**

Figure 5.8: Face (a) Out-of-focus (b) Visual saliency map

This shows the effectiveness of the algorithm in differentiating in and out-of-focus areas in the image. In the Figure 5.9 the pair of images is captured by alternating the focus regions between foreground object and the background region. The first image focuses on the video camera. The second image focuses on the background (books and shelf). The focus map can distinguish between in-focus and out-of-focus regions in these two images.

Similarly, in the Figure 5.10 the first image focuses on the camera lens and the person standing is out of focus. In the second image the focus is shifted on to the standing person by making the lens out-of-focus (blurred). The focus maps can distinguish this focus shift in both the images. In the last pair of images the first image focuses on the plant and the poster on the wall is out-of-focus. In the second image the poster on the wall is in-focus and the plant is out-of-focus. In all the images in Figure 5.10 the algorithm is able to detect the focus shift.

Figure 5.9: Images captured by making video camera in and out-of-focus



.

Figure 5.10: Images captured by alternating focus regions

The first image (top left) in the Figure 5.11 shows sparsely spaced multiple objects. In order to test the performance of the algorithm it is made sure during the image capture that only the enclosed regions marked by an outline are in focus. Similarly, the second image shows closely spaced multiple objects that are in focus. In both cases the algorithm correctly detects the objects that are in focus. In these images the algorithm is tested with multiple regions that are in focus. Moreover, it can be seen that the algorithm is also robust in terms of distance between in-focus regions.

In the Figure 5.12 the top left image focuses on the middle object leaving out the extreme end objects as out-of-focus. In contrast, the bottom left image focuses on

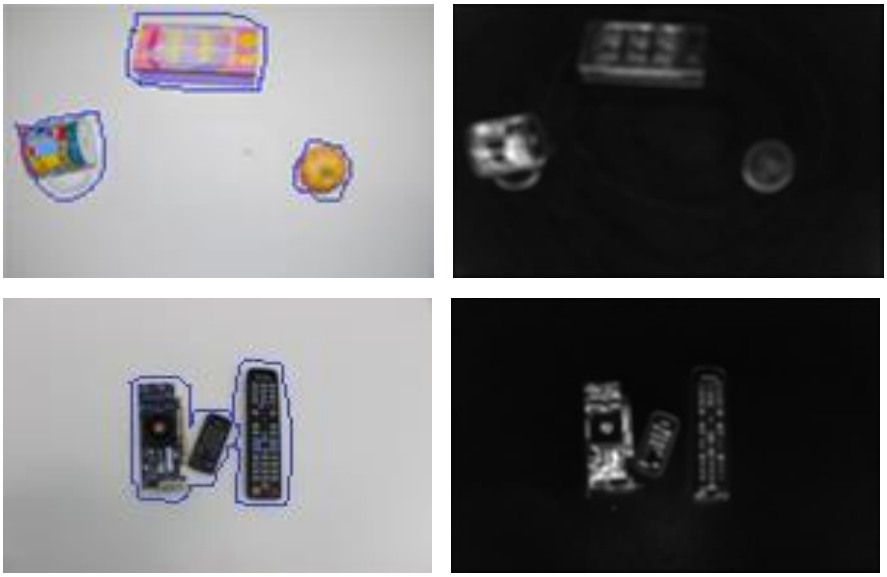


Figure 5.11: Images captured with multiple focus regions

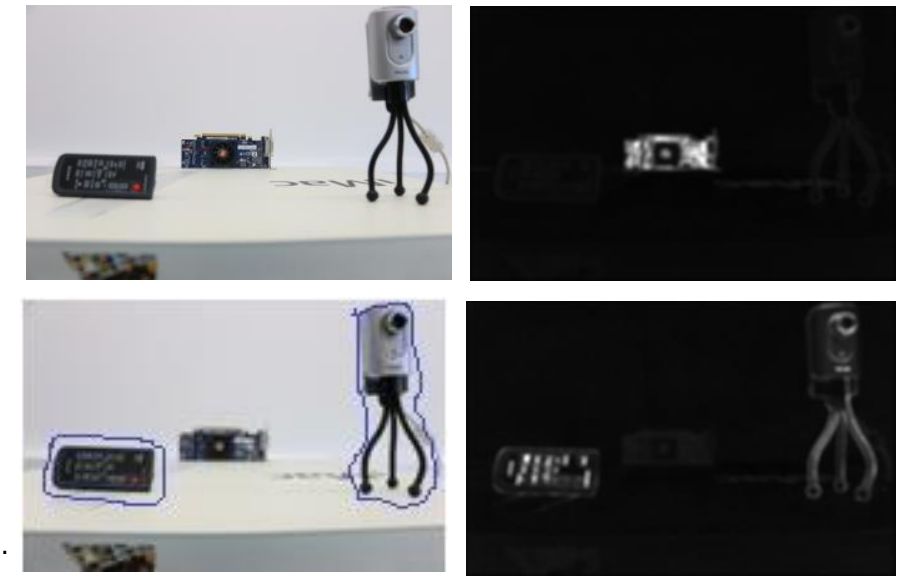


Figure 5.12: Images captured with single and multiple focus regions

extreme end objects leaving the middle object out-of-focus. The proposed algorithm correctly maps the objects that are in focus in these cases as well. Therefore, the algorithm has good detection ability irrespective of the distance between the in-focus regions and also the number of in-focus regions.

In the Figure 5.13, the image contains highly textured background (sand) which has high spatial frequencies throughout. However, the camera is clearly focused

around the foot imprint. The entire background with hill, sky and the bottom left corner in the image is out of focus. The proposed model clearly distinguishes the dominant frequencies within the in-focus region from the noisy frequencies contained in the background as evidenced by the focus map. In the next image (bottom-left) the person riding the cycle and the adjacent surrounding area is highly in focus and the background is out of focus. The algorithm is able to detect the in-focus regions irrespective of complex backgrounds.

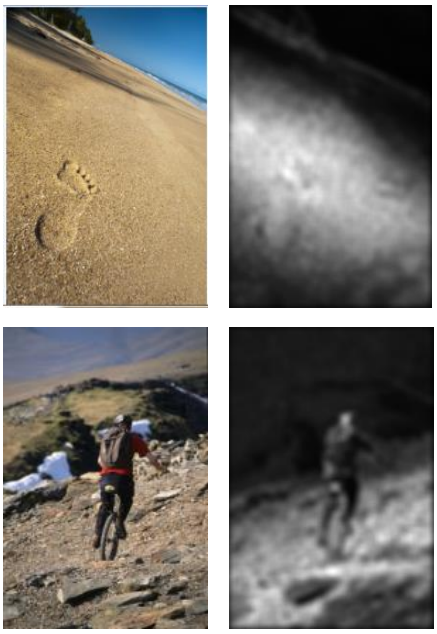


Figure 5.13: Focus regions with complex image background.

In the Figure 5.14 the star mark in the first image (top left) to the upper right of the picture is highly in-focus and the background sand is marginally in focus. The corresponding focus map reflects this by showing a clear variation in the intensity. Similarly, in the second image the canoe sailing in the water is in-focus and the adjacent surrounding water to the left and right are in-focus. The image tends to become out-of-focus towards the periphery. In the focus map it can be clearly seen that the canoe and the surrounding water is shown with higher grey scale intensity values and the intensity falls off towards the periphery of the image.

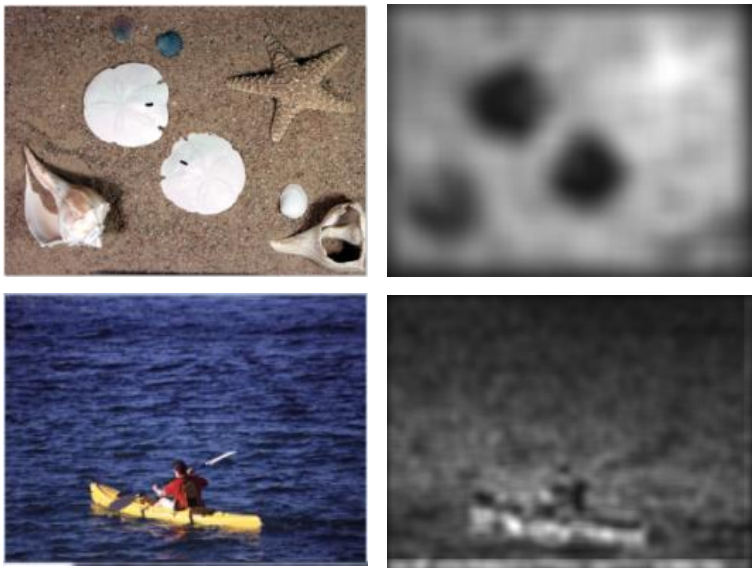


Figure 5.14: Images with random regions in focus

5.6.2 Quantitative Analysis

A quantitative evaluation of the proposed model is carried out against popular Judd's dataset of 1003 images. The performance of the model is compared with three popular visual saliency models. Moreover, three quantitative metrics, namely, Correlation Coefficient (CC), Normalised Scanpath Saliency (NSS) and Area Under Receiver operating Curve (AUC) (please refer to section 4.5.2 for more details regarding these metrics) are used to evaluate the performance. The higher the score achieved with respect to these metrics, the better is the prediction accuracy of the saliency model. The performance evaluation of the state-of-the-art saliency models is shown in Table 5.1. In the table, the proposed model is highlighted in bold letters. It can be seen that the model performs better than the SUN saliency model which is popular and widely cited. This model computes saliency based on bottom-up and top-down features. The fact that the proposed model outperforms the SUN model clearly indicates that focus is a key component that plays a significant role in attracting human gaze irrespective of bottom-up and top-down features. Further, the model also outperforms PQFT which is a popular model built using purely bottom-up features such as colour, intensity and motion channels. Inspite of the model using several bottom-up features, it achieves lower prediction accuracy when compared with the proposed model. In the table WBSD is a recent model built using bottom-up features. The proposed model achieves the same accuracy using two metrics and a slightly lower score with respect to NSS.

Table 5.1: Quantitative comparison of saliency models on Judd dataset

Judd's dataset (1003 images)				
Year	**Models**	**CC**	**NSS**	**AUC**
2008	SUN [70]	0.15	0.75	0.67
2010	PQFT [10]	0.12	0.59	0.56
2013	WBSD [127]	0.18	0.88	0.71
2013	**Proposed [14]**	**0.18**	**0.84**	**0.71**

5.6.3 Qualitative Analysis of the Saliency Map

The saliency maps for sample images from Judd's database are shown for the proposed model and three state-of-the-art saliency models in Figure 5.15. The Ground Truth Fixation Maps (GTFM) is also shown in the figure. In the first image (flower), the ground truth indicates that users mostly gazed at the centre of image. Moreover, in the image the flower is highly in-focus compared to the background. In the saliency map it is evident that the focus detection algorithm successfully extracted the in-focus region with slightly higher intensity at the centre of the image. It also has lower false detections when compared to other models which detects unimportant areas as salient. In the second image the ground truth shows that viewers mostly gazed along the tower with higher number of fixations around the person. The proposed model gave closer performance to the ground truth. The model extracted the tower from the image with higher intensity making it a qualitatively better saliency map than SUN, PQFT and WBSD.

In the Figure 5.16, the ground truth map in the first image shows that viewers mainly gazed at the swing seat. For this image the proposed model exhibits high intensity in the saliency map around the swing seat. However, the model also detects other unimportant areas in the image as salient. Moreover, both PQFT and WBSD obtain a very sparse saliency map and they also detect the background as visually salient. In the second image ground truth fixation maps clearly indicates that viewers are more interested at the centre of the image. The proposed model (similar to the eye tracking fixation map) exhibits relatively higher intensity at the centre.

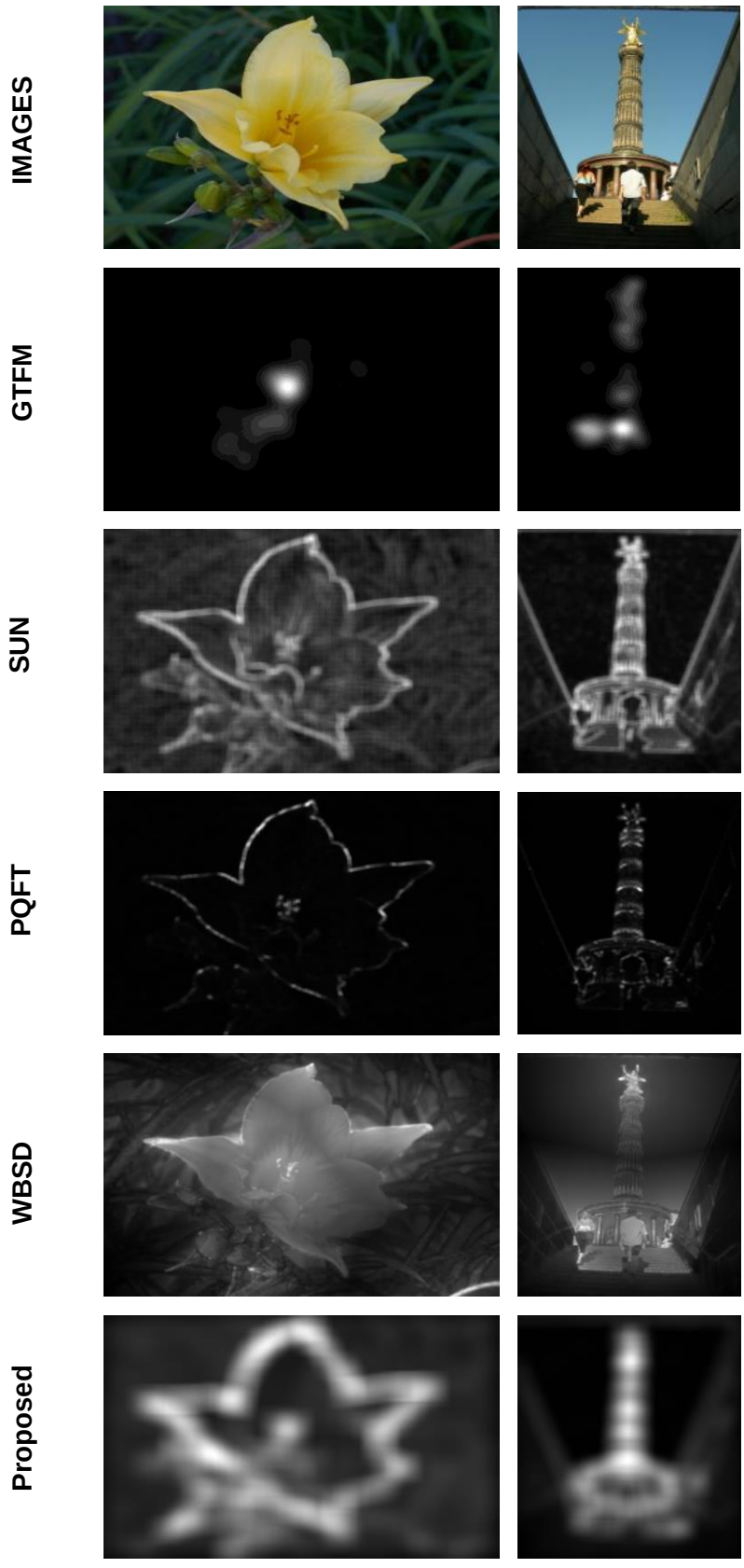


Figure 5.15: Qualitative comparison of sample images from Judd's

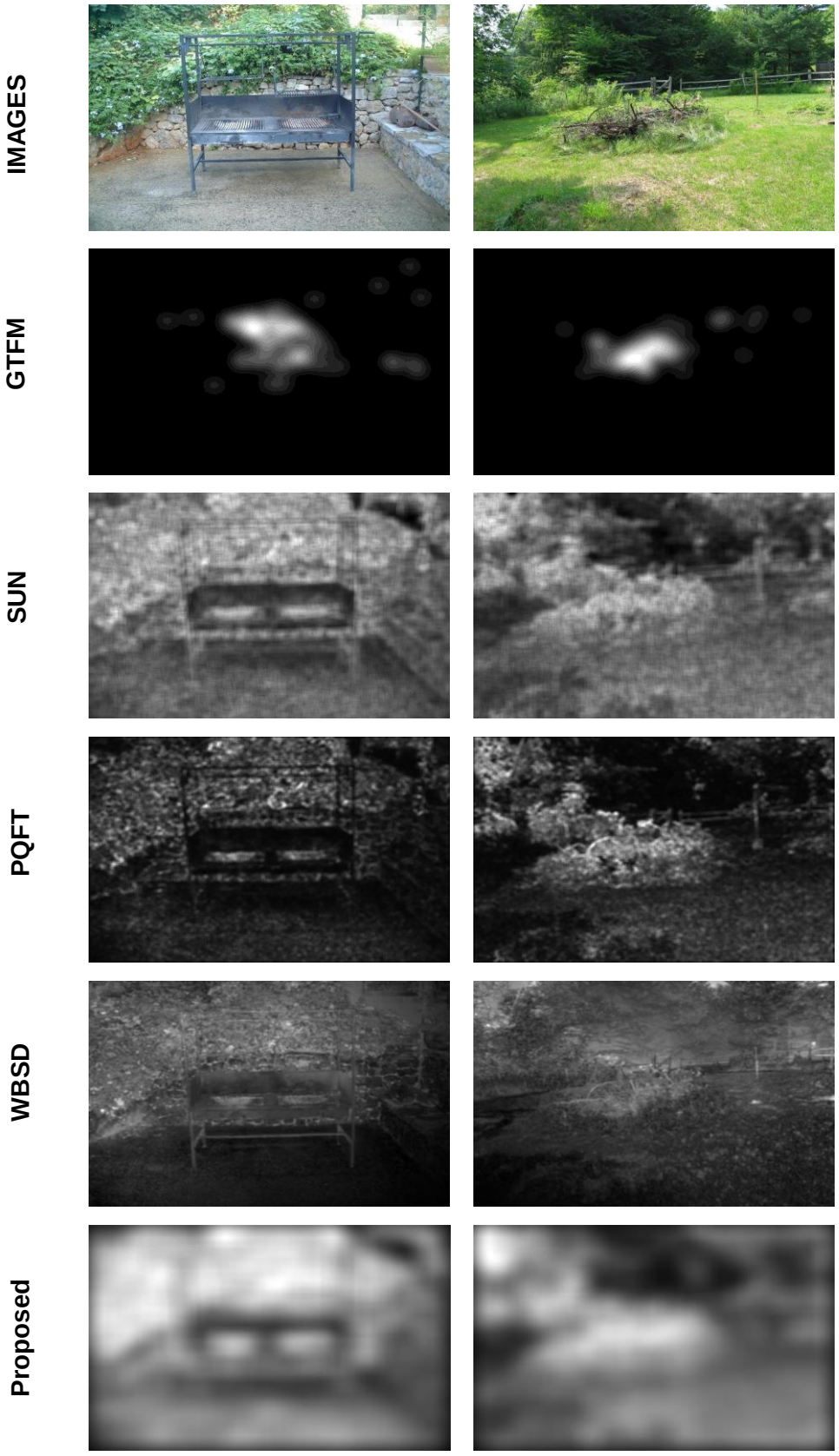


Figure 5.16: Qualitative comparison of sample images from Judd's

5.6.4 Complexity Analysis

The main criteria used for comparison with the state-of-the-art are prediction accuracy and complexity. During the complexity optimisation of the model the main constraint is the prediction accuracy. Therefore, it is a constrained optimisation problem in which the complexity of the model has to be reduced by using low complexity features that have the ability to achieve a higher accuracy. It means that in a time constrained scenario, a saliency model should be fast enough to meet the real time performance requirements while meeting the accuracy requirements. In our current scenario the proposed model should achieve prediction accuracy better than or equal to WBSD with lower complexity. It is already shown in the earlier section that the proposed model has achieved better accuracy than SUN and PQFT and an almost equal score when compared with WBSD saliency model.

The complexity of the proposed model is compared with the state-of-the-art saliency models in the Table 5.2. We use un-optimised MATLAB code (without MEX code) for all of these saliency models in order to ensure a fair comparison. The average time required to compute a saliency map is calculated over 100 images with resolution 1024x768 from Judd's database. It is evident that, the proposed model is the fastest among the saliency models. Compared to our model, WBSD achieves similar performance in terms of prediction accuracy. However, it is still significantly more complex than the proposed model. The complexity of the SUN model is increased as it uses both bottom-up and top-down features.

Table 5.2: Complexity comparison of the state-of-the-art saliency models

Judd's dataset (100 images)		
Year	**Models**	**Complexity(secs)**
2008	SUN [70]	9.36
2010	PQFT [10]	0.83
2013	WBSD [127]	24.28
2013	**Proposed [14]**	**0.80**

The low complexity nature of the proposed model can be attributed to DCT properties such as energy compaction and data correlation as indicated in section 5.4.2. The proposed model achieves complexity reduction with prediction accuracy similar to WBSD model and hence outperforms this model.

5.7 Discussion

During the first phase of model development an image dataset with in-focus/out-of-focus images is created. Through qualitative investigation on these images the DCT coefficients characteristics of in-focus and out-of-focus regions are studied. During the investigation, it has been discovered that there are a few large valued frequencies which contribute to the most of the energy in the in-focus regions, whereas these frequency coefficients hold low magnitude values in the out-of-regions. In the second phase an in-focus visual saliency model has been proposed based on the peak frequency components. These peak frequencies are identified by zig-zag scanning a mean absolute 8x8 superblock of luminance channel in YC_bC_r colour space. The summation of peak frequency magnitudes across all image blocks provides an in-focus visual saliency map.

The results demonstrate that the proposed model achieves similar prediction accuracy (CC=0.18, NSS= 0.84 and AUC=0.71) (refer to section 5.6.2) when compared with state-of-the-art saliency models [70], [10], [127] at a significant reduction of computational complexity. In the literature there are very few works [152], [153], [154] which have utilised focus detection as the core element for deriving image saliency. As already discussed these works estimate blurriness as way to detect in-focus regions. In contrast to these work the proposed focus model does not depend on image blur and derives in-focus regions based on DCT coefficients. It is a novel contribution as it detects the in-focus regions using the peak frequencies present in the DCT domain. Moreover, a mathematical model is developed to detect these peak frequencies. The computational complexity of the model is calculated over 100 images with resolution 1024x768 on Intel core I7-2600K CPU operating at 3.40 GHz. The model takes an average time of 0.80 seconds (refer to section 5.6.4) to compute the focus map. The advantages and disadvantages of this model are summarised as follows.

Advantages

- Significant reduction in the computational complexity whilst achieving the prediction accuracy similar to or better than some other benchmark models [70],

[10], [127]. It has been already discussed that the model takes an average time of 0.80 seconds for calculating the saliency map of an image with 1024x768 resolution. With the same image resolution and testing platform the WBSD [127] model achieves similar prediction accuracy as the proposed model; however, it requires 24.28 seconds. This indicates that the model is extremely fast at detecting salient regions in images.

- The focus detection algorithm has the ability to detect multiple in-focus regions.

- The saliency detection using in-focus regions has achieved better prediction accuracy compared to the PQFT model [10]. PQFT uses four channels for detecting attended regions namely two colour channels, one intensity channel and one motion channel. The proposed model based on a single channel using in-focus region detection outperforms the PQFT model. This indicates that focus plays an important role in attracting human gaze.

Disadvantages

- The model achieves prediction accuracy similar to the WBSD model [127] and better than SUN [70] and PQFT [10] models. In chapter four several models were considered for comparison. The accuracy of the saliency model developed is still not better than some of the benchmark models specified in chapter 4 such as GBVS, SR, NVT etc.

- When an image is made completely out-of-focus the entire image looks visually blurred. The attention model when used to detect in-focus regions across such images produces incorrect results as the difference between the peak frequencies may not be significant enough. This results in false detections (detecting non-salient regions as visually salient). Therefore, the model produces inaccurate results when an image is completely in-focus or out-of-focus.

As already discussed in the advantages section, detecting image focus alone outperforms other models that considered several features [70], [10] for predicting salient regions. Therefore, focus detection can be considered as a promising idea for further development. Therefore, this model forms the foundation for further research that resulted in an improved model which is described in the next chapter.

5.8 Conclusion

A novel visual attention model has been developed to detect salient regions in images based on camera focus. The model considers in-focus regions in the images to derive the visual saliency. The characteristics of DCT coefficients are used in modelling the focus map. It outperforms some of the state-of-the-art visual attention models in saliency detection performance with respect to low complexity. Note that, in terms of prediction accuracy and complexity this model is still behind some of the benchmark models such as NVT, GBVS, SR, CAS, SS, RCSS and SDSP. Therefore, future research involves improving the accuracy and complexity to outperform the chosen benchmark models. To achieve this, in the next chapter the complexity of the model is further reduced by replacing the DCT with the Integer Cosine Transform (ICT). Moreover, the prediction accuracy of the model is improved by considering some of the prominent top-down features. Further, the saliency model will be tested across bigger and challenging datasets.

6 Visual Attention Model: Top Down Extension

6.1 Introduction

In this chapter the DCT based in-focus visual attention model described in the previous chapter is improved in terms of both prediction accuracy and complexity. To achieve this, the traditional DCT is replaced with Integer Cosine Transform (ICT) and HSV colour space is used instead of YC_bC_r colour space. Moreover, top-down features are detected and combined with in-focus map to improve the overall prediction accuracy. The model developed predicts human fixations based on in-focus regions and top-down components such as image centre and human faces. Similar to the focus detection algorithm in chapter 5, the in-focus regions in the images are detected using the magnitudes of frequency coefficients in the Integer Cosine Transform (ICT) domain. The centre sensitivity maps are constructed by placing anisotropic 2D Gaussian distribution at the centre of image with standard deviation as a function of the image resolution. The human face map is generated by using a face detection algorithm. The ICT based focus maps are convolved, contrast stretched and combined with centre and face maps to obtain the salient regions of the image. A hill climbing approach is used to tune the model parameters. The performance of the model in predicting human fixations is evaluated qualitatively and quantitatively against ten state-of-the-art visual saliency detection algorithms. The results demonstrate that the proposed algorithm achieves higher prediction accuracy at significantly lower computational complexity compared to the state-of-the-art visual saliency detection models. Further, the saliency model's performance was measured using a new evaluation method known as dispersion measure. It measures the consistency of a saliency model in predicting fixations across each image in the dataset. The existing models are compared using the proposed measure and it is shown that the proposed model achieves the best dispersion measure compared to the existing models.

In section 6.2 the hypothesis used for improving the visual saliency model is described. The directly related work to the proposed visual saliency model based on centre sensitivity and human faces is explained in the section 6.3. In section 6.4 the proposed model is described which includes algorithms for the development of focus, centre and face saliency map. The experimental results related to prediction accuracy and computational complexity of the model is given in section 6.5. The model is critically discussed in section 6.6 and concluded in section 6.7.

6.2 Hypothesis

A significant amount of research on human eye saccades and fixations has clearly demonstrated that humans are highly sensitive towards the centre of the image compared to the periphery [159, 160]. The key factors that are responsible for this phenomenon are photographer bias [160-162] and viewing strategy [146]. Photographer bias is a natural tendency of the photographer to put the objects at the centre to emphasise their importance. They usually tend to put the most interesting objects at the centre of the image. As these are at the centre, viewers automatically move on to these locations. The other important factor that makes a viewer highly sensitive towards the centre is viewing strategy. Viewing strategy is the by-product of photographer bias where in the viewers repeatedly re-orient themselves to the image centre. Whenever a viewer looks at an image, they have a natural assumption that an important or interesting object will lie at the centre of the image and they initiate their search process from the centre. In datasets like Judd *et al.* [12] and DUT-OMRON [141] the first eye fixation during eye tracking is eliminated to avoid the viewing strategy as the first fixation is generally at the centre of the screen. In addition to these two important root causes, some other less-influential factors include, orbital reserve (straight head position of the viewer) [161, 163, 164], motor bias (tendency to use short viewing saccades) [159, 165], screen centre [166], low sensitivity of the HVS towards the periphery of the human eye [167-169] and some other high level influences [167-169].

Humans have a tendency to gaze at faces irrespective of other visual stimuli [147, 170, 171]. The evidence of sensitivity of humans towards faces collected from infants as young as 6 weeks suggests that faces are visually captivating [172]. It has been found that, in free-viewing conditions, people are 16.6 times more likely to look at faces compared to other similar regions [147]. Moreover, it is also shown that facial expression and gaze direction have an in-built capacity to attract attention [173]. Therefore, in this work human faces were considered to be a predominant top down feature, irrespective of whether they are in-focus or out-of-focus.

6.3 Directly Related Work

In chapter 3 the literature related to the state-of-the-art visual attention models was provided. These models were built using different kinds of approaches to detect salient regions in the images. The current work is compared with these models to show the effectiveness of the proposed approach. In the following sub sections the directly related work of the top-down aspects such as centre sensitivity and using human faces

which are the key components of the proposed model for improving the prediction accuracy are analysed. The main purpose of providing the directly related work is to indicate the novelty of the proposed model.

6.3.1 Centre Sensitivity

In the literature, centre sensitivity has been used by the authors with an aim of improving prediction accuracy. Although by considering centre bias the model's prediction accuracy can be improved, a performance drop can be seen if the key components of the attention model are not well innovated architecturally (architectural innovation refers to different types of strategies used to combine the key components of the developed attention model). Another reason that can affect the performance is the response of other feature maps towards the very way in which the centre map has been modelled. In the literature the centre map is obtained using different methods. For instance a very recent model SDSP [13] considers centre as salient by modelling the location prior. Despite considering centre sensitivity the model has achieved low prediction accuracy when compared with the state-of-the-art as shown in Table 6.3 (section 6.5.2). In the work by [174] it is also shown that adding centre bias to some of the best models has resulted in performance drop. The models in the literature are optimised towards centre either explicitly [12] or implicitly [175]. The GBVS [175] model is implicitly centred through activation and normalisation. It is one of the best examples of implicit optimisation; however; their approach resulted in an increased computational complexity. Judd *et.al* [12] has explicitly biased towards their saliency maps towards the centre by adding a uniform Gaussian blob. Although it has achieved good prediction accuracy the complexity of the model is very high due to the calculation of several complex channels such as person, face, car, horizontal line, gist, etc [46].

6.3.2 Human Faces

The authors of [9] indicated that face detection can be used as a posteriori refinement of the saliency map. Their approach is to use all 1's for the face regions and 0 otherwise. However, this is a naive approach as it ignores user sensitivity distribution across the human face. Similarly Judd *et al.* [12] detected human faces in her saliency model for improving prediction performance. The authors, through their qualitative analysis, indicated that users have different sensitivity to different parts of the human face. However, they still used simple bounding boxes leaving modelling of the face saliency map to future research. In the work by [89] the facial centres are convolved using 2D Gaussians. Although a better face map is built by this approach, taking standard deviations equal to facial radius limits the accuracy. Moreover in their work,

face map importance is not preserved as it is equally weighted as the other bottom-up features (intensity, colour and orientation).

6.4 Proposed Model

The main components of the model are generation of a focus map using the salient frequency coefficients present in in-focus areas, generation of a centre sensitivity map, face detection and generation of a face map and the integration of focus, face and centre maps to obtain the final visual saliency map at the original image resolution.

6.4.1 Focus Map

The process is similar to the DCT based in-focus attention model described in chapter 5, however, with the following improvements. In this work the Integer Cosine Transform (ICT) [176] is used because of its low complexity compared to DCT (see

Figure 6.1: Image with enclosed region in-focus

section 6.5 for complexity comparison). The Value channel of the perceptual HVS colour space is used to calculate the ICT. Earlier work in chapter 5 employed the Y component of the YUV color space to calculate the DCT. In this work, HSV colour space is used to improve the prediction accuracy of our earlier focus detection algorithm. Similar to the procedure followed for developing the in-focus detection algorithm in Chapter 5, random images were selected and converted to HSV color space. In-focus regions of the images were manually observed and identified as shown in Figure 6.1. The value component (V of HSV) of the entire image is extracted and divided into 8x8 blocks and the ICT of each 8x8 block is calculated. The 8x8 Integer Cosine Transform is defined as

$$W = C_8 X C_8^{\ T} \tag{6-1}$$

Where X is 8x8 image block as input to the transform, C_8 is the core transform matrix and W is the output Integer transform of X. The core transform matrix implementation is obtained from [176]. The ICT transformed 8x8 blocks revealed that in-focus blocks have significantly large coefficients compared to out-of-focus blocks. The foreground superblock, background superblock and full image superblock are calculated as they were calculated in the earlier work. The spatial frequency composition of in-focus (foreground) and out-of-focus (background) regions were analysed by selecting the ICT coefficients in all three superblocks using the zig-zag scanning method. The average frequency coefficient magnitudes of foreground, background and the overall image superblocks are plotted on a graph to determine the relationship between in-focus and out-of-focus coefficients. The ICT frequency coefficient amplitude pattern of an image is shown in Figure 6.2.

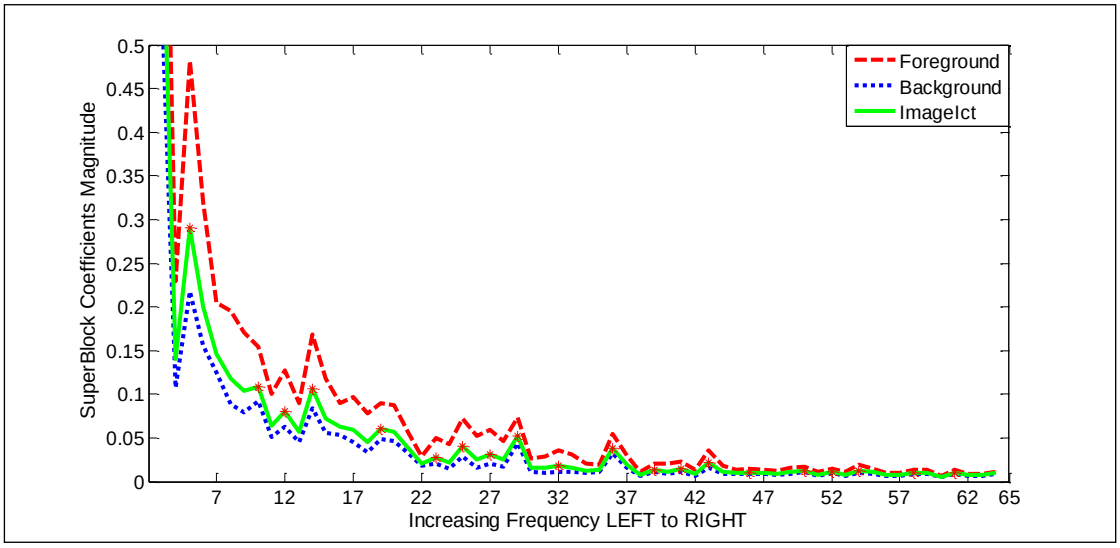
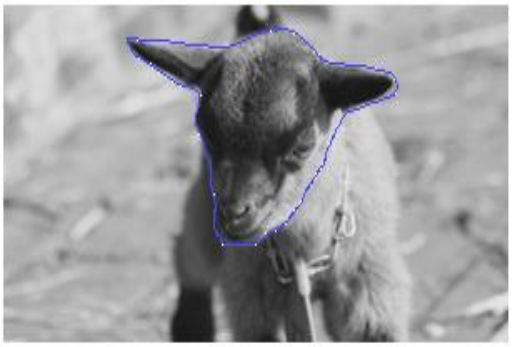


Figure 6.2: Superblock coefficients magnitude vs. Zig-zag scanned frequencies

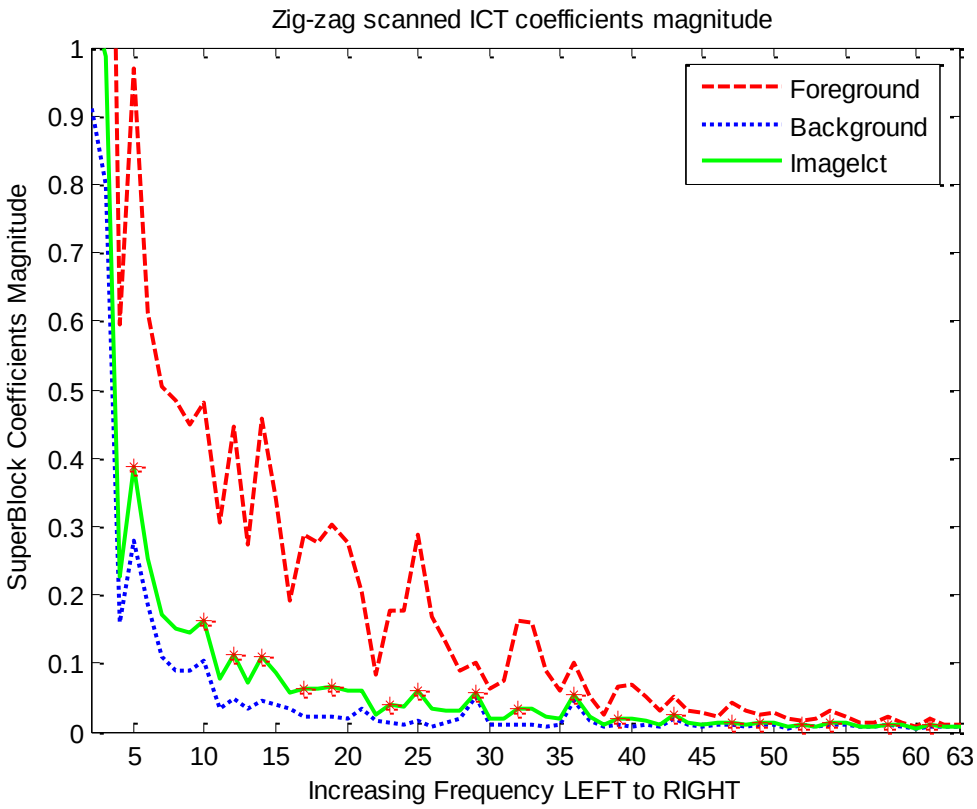
The graph reveals that the peaks of spatial frequency amplitude waveform of the foreground (in-focus), background (out-of-focus) and the entire image almost coincide. The maximum magnitude difference occurs between in-focus and out-of-focus waveform at these peak frequencies. These amplitude differences tends to be significant within a band of frequencies which excludes both very low (DC component and the first few very low frequencies are not shown here due to higher amplitudes) and high frequencies. It also reveals that the peak magnitude frequencies in the in-focus regions are absent in out-of-focus regions. Further, these peak frequencies have

a significant presence in the overall image ICT. Therefore, the peak frequency coefficients in the whole image can be used to identify in-focus regions. Experiments with a number of images revealed that the DC frequency coefficient represents gradual changes and high frequencies (35-63) do not show a significant difference between in-focus and out-of-focus areas as shown in the Figure 6.2, Figure 6.3 and Figure 6.4.

(a) Sheep-RGB picture



(b) Sheep-Y picture

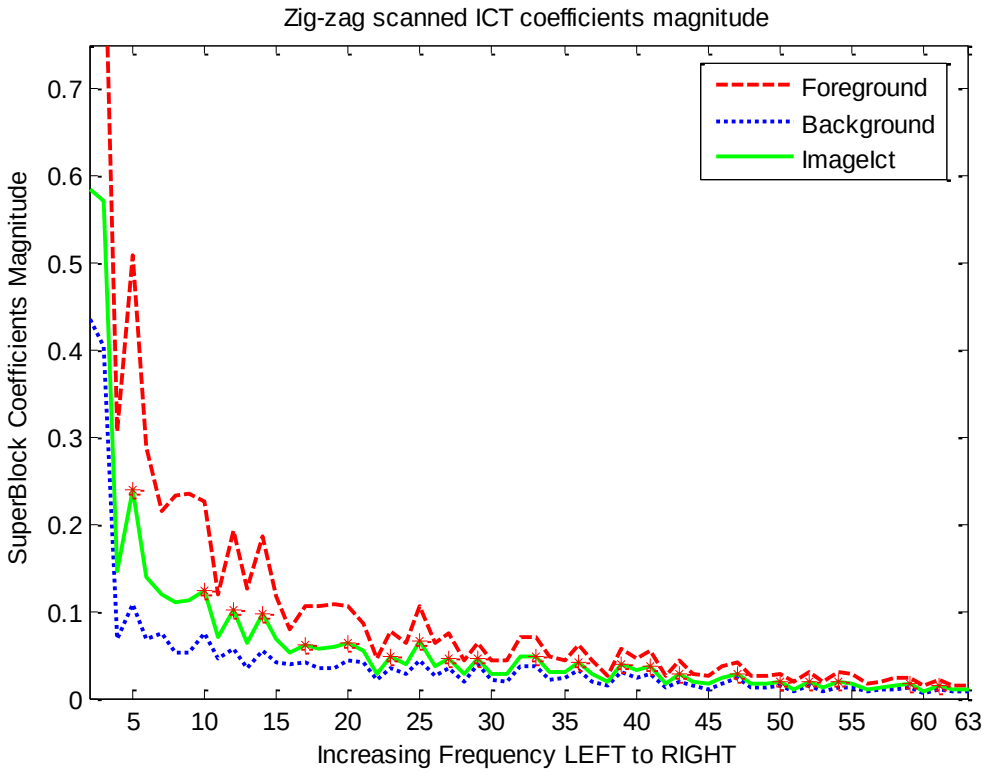


(C) Sheep - Zigzag scanned ICT coefficients

Figure 6.3: Sheep (a) RGB picture (b) Y-picture (c) Zig-zag scanned ICT coefficients

(a) Face-RGB picture

(b) Face-Y picture



(c) Face-Zigzag scanned ICT coefficients

Figure 6.4: Face (a) RGB picture (b) Y-picture (c) Zig-zag scanned ICT coefficients

Therefore, these zigzag scanned frequencies coefficients are band-pass filtered to remove the high and very low frequency ICT coefficients. All frequency coefficient positions that correspond to the peaks within the band of frequencies of the image superblock coefficients are identified and stored. These are the salient spatial frequencies present in the in-focus areas of the image.

In the previous model, the final focus map was generated by plotting the sum of peak frequencies in each DCT block for the entire image. However, in this work an initial focus map is generated using the ICT. Then the initial focus map is filtered with a

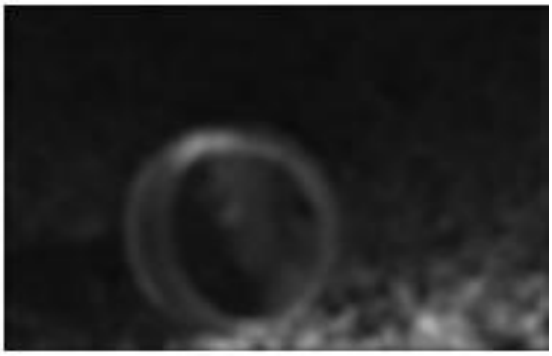


Figure 6.5: Focus map of the image shown in Figure 6.1

smoothing Gaussian kernel (size $n \times n$) and contrast stretched by multiplying each pixel in the focus map with a factor k. The smoothing filter is used to generate connected regions from the sparse focus map. The contrast is stretched to improve the overall intensity of the focus map. The resolution of the focus map is $1/8^{th}$ of the resolution of the image in each dimension. This is because the focus map is generated by using the sum of salient frequencies of each 8x8 block. Therefore, the focus map is up-sampled to the original image resolution using bi-cubic interpolation. The focus map corresponding to Figure 6.1 is shown in Figure 6.5.

6.4.2 Centre Sensitivity Map

Similar to Judd *et al.* [12] the centre sensitivity is explicitly taken into account but with one major difference. It was hypothesised that human eye saccades are affected by the height and width of the image. The eye saccades were oriented vertically in the centre if the height of the image is far greater than width. Similarly they are oriented horizontally if the width is greater than the height. As an example it can be seen in the Figure 6.6, the image with the ship and corresponding eye tracking map shows that most of the fixations are clustered around the centre and oriented towards the horizon. This phenomenon is different in the other image. As the height of the image is higher than the width the fixation are initially clustered at the centre and then oriented vertically as saccades transit towards the periphery. This phenomenon is computationally modelled by taking an anisotropic 2D Gaussian with standard deviations as a function of percentage of image resolution.

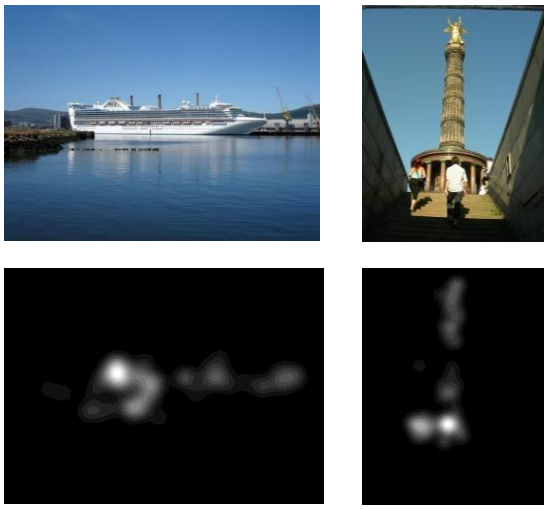


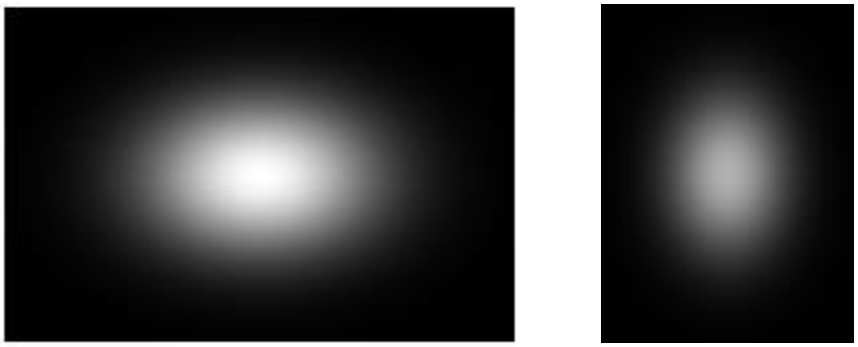
Figure 6.6: Images with Corresponding eye tracking maps

This percentage is empirically derived using hill climbing approach which is a heuristic technique of multi variable optimisation. The centre map is obtained using a 2D Gaussian with standard deviations σ_x and σ_y as shown below.

$$\sigma_x = \left(c * w\right)/6 \tag{6-2}$$

$$\sigma_y = \left(c * h\right)/6 \tag{6-3}$$

Where c is a fraction of height (h) and width (w) of the original image. The value of $c = 1$ means that the Gaussian distribution approximately fills the entire centre map horizontally and vertically (corresponding to 99.7% of area under curve for a distance of 3 standard deviations either side from the centre). A sample centre map is shown in Figure 6.7. The focus and the centre maps are normalised to the same dynamic range and are combined using the following equation. A linear summation approach was utilised to combine the maps as it has some psychophysical support and simple to apply [177]. Where, α is a weighting parameter. Therefore, the combined map consists of weighted additions of pixel values of focus and centre maps. The resolution of the generated saliency map is 1/8th of the resolution of the image in each dimension.



(a) Horizontal Centre map **(b) Vertical Centre map**

Figure 6.7: Centre map demonstration. (a) Horizontal centre map. (b) Vertical centre map

$$UniqueMap = \alpha \cdot FocusMap + (1-\alpha) \cdot CentreMap \quad (6\text{-}4)$$

This is because the focus map is generated by using the sum of salient frequencies of each 8x8 block. Moreover, the resolution of the centre map is chosen as the size equal to that of the focus map. Therefore, the obtained map is up-sampled to the original image resolution using bi-cubic interpolation.

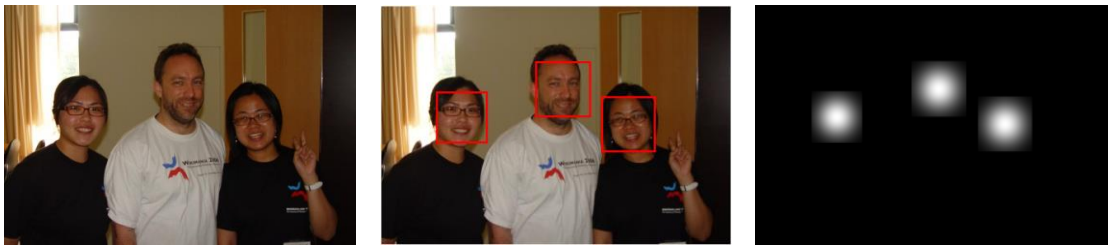
6.4.3 Face Map

The face detection algorithm from the authors of [90] was used in this work for detecting human faces. Square shaped bounding boxes are drawn around the detected faces to obtain the size and position of the face. To generate a face map, initially a binary map of zeros of original image resolution is constructed. For each detected face, 2D Gaussian blobs were generated with standard deviation values σ_x and σ_y as:

$$\sigma_x = \sigma_y = (S/6)*\beta \quad (6\text{-}5)$$

Where S denotes the length of the bounding box side and β is a parameter determining the size of the Gaussian blob or the fixation cluster in relation to the bounding box size. The value of $\beta = 1$ means that the Gaussian distribution approximately fills the entire bounding box, corresponding to 99.7% of area under curve for a distance of 3 standard deviations either side from the centre. When $\beta < 1$,

the fixation cluster moves towards the centre of the face and $\beta > 1$ expands the cluster towards the periphery of the box. However, the Gaussian distribution is confined to the bounding box dimensions (clipped at the boundary) because most of the actual fixations tend to cluster within the face region. A sample image with faces, bounding boxes and the corresponding face maps ($\beta = 2.3$) are shown in Figure 6.8. The optimal selection of β is discussed in the later sections. The generated face map is combined with the unique map (focus-centre) map as described in the next section.



(a) **(b)** **(c)**

Figure 6.8: Face map generation. (a) Image with 3 faces. (b) Face localisation using bounding boxes. (c) Face map using Gaussian blobs

6.4.4 Visual Saliency Map

The saliency map is obtained by combining focus and centre maps that are normalised to the same dynamic range. These normalised maps are combined using the following equation.

$$VisualSaliencyMap = UniqueMap + FaceMap \quad (6\text{-}6)$$

During the process of combining a number of weightings (including overlaying) were considered as part of the parameter tuning process and it has been empirically found that overlaying or simple addition of face map to the focus-centre map gives better accuracy as the importance of face map or top-down component is preserved.

6.4.5 Parameter Tuning

There are a number of model parameter values that need to be tuned. Due to the high number of parameters and their value ranges, the parameters of the model are tuned using hill climbing method [178]. These parameters are

i) Gaussian smoothing kernel size n ($n \times n$ kernel) of the focus map.

ii) Contrast multiplier k of the focus map.

iii) Fraction c of the centre map, used to calculate the standard deviations of the 2D Gaussian centre sensitivity distribution.

iv) Weighting parameter α, used to combine the centre and focus maps.

The objective of parameter tuning is to choose the model parameters in order to maximise the correlation between the eye fixation ground truth of images and the saliency maps obtained using the saliency model. Table 6.1 shows the parameters and the value ranges involved in the tuning process. An exhaustive search for the selection of the optimal set of model parameters will require large number of permutations to be tested on a dataset. This requires a prohibitive amount of processing resources. Therefore, a hill climbing algorithm [178] was used to tune the parameters of the model. The hill climbing approach belongs to the family of heuristic methods of local search for computational optimisation. It has been utilised in explanation based learning systems, utility analysis models and robotics [179], [180] for calculating optimal solutions. Hill climbing is an iterative process in which it starts with a random or arbitrary set of parameters as the solution to the problem. These parameters are purely based on

Table 6.1: Model Parameter Values

Parameter	Value range
n	3 to 24
k	0.1 to 4 in steps of 0.1
c	0.05 to 1.5 in steps of 0.05
α	0.05 to 1 in steps of 0.05

random guess. Then by incrementally changing each variable in the given set of parameters the process attempts to find a better solution. If a change in one particular variable produces a better result, then the process iterates by making an incremental change to the same parameter until no further improvements in solution are found. The same procedure is repeated with other sets of parameters. Generally the hill climbing is stopped based on three conditions.

i) No further improvements can be seen in the solution.

ii) The goal state is achieved.

iii) A fixed number of iterations have been performed.

The goal state or the fitness function of the hill climbing algorithm is evaluated using the Correlation Coefficient (CC) [181] metric. During the process different parameters are tested to optimise the prediction accuracy between the eye fixation ground truth and the proposed attention model under test for the Judd *et al.* [12] image database. In the current development the hill climbing process is terminated when no further improvements are observed. The tuned parameter values were found to be, $n = 12$, $k = 3$, $c = 0.95$ and $\alpha = 0.65$.

6.4.6 The Complete Visual Attention Model

The complete visual saliency detection model can be summarised as:

1. An ICT is performed across all 8x8 blocks of the image using the value channel of HSV colour space.
2. The image superblock is calculated and the frequency coefficients are zig-zag scanned.
3. The zig-zag scanned ICT coefficients are band pass filtered.
4. The peaks of the image ICT are obtained and the magnitude of the frequency coefficients corresponding to these peaks are summed and plotted as a salient frequency map.
5. The salient frequency map is Gaussian smoothed, contrast stretched and up-sampled using bi-cubic interpolation to original image resolution to generate the final focus map.
6. The centre maps are generated by placing anisotropic 2D Gaussian at the centre of the image with standard deviation as a function of the image resolution.
7. The focus map and the centre map are combined as per equation (6-4).
8. Face detection is performed on the original image and any detected faces are marked using Gaussian blobs.
9. The face map is overlaid on the combined focus and centre map to generate the final saliency map.

6.5 Experimental Results

The performance results of the proposed visual attention model are given in the following sections. The model has been initially tuned or trained using the Judd *et al.* [12] public dataset which contains 1003 images.

6.5.1 Quantitative analysis

A quantitative evaluation of the proposed model was carried out against ten popular visual saliency models. Three quantitative metrics, namely, Correlation Coefficient (CC) [181], Normalised Scanpath Saliency (NSS) [146] and Area Under receiver operating Curve (AUC) [12, 182] are used to evaluate the performance. In order to comprehensively evaluate the performance of the model, the Judd's database is manually split into two sub datasets comprising of images with and without human faces (264 images with human faces and 739 images without human faces).

The sub category which includes human faces is further divided into images with clear frontal faces (182) and images with non-frontal faces (82). Category of images with non-frontal faces include, faces that are angled sideways, unclear faces and very small faces relative to image size. The proposed model uses frontal face detection to detect human faces [90]. Therefore, the performance of the model was evaluated for the complete dataset and also for these sub categories of images. The results are shown in the Table 6.2. It can be seen that the proposed model performs better than the state-of-the art in terms of prediction accuracy for all the three metrics across all the three sub datasets. As shown in the Table 6.2 the model could only accurately detect the faces in clear frontal face category of images (182).

In the other category, the face detection did not perform accurately (i.e. non/partial detections and false detections). In the images with frontal human faces the proposed model achieves significant improvement in prediction accuracy. This can be attributed to the fact that humans have a predominant generic top down influence in looking at human faces. Visual saliency is a combination of both bottom-up and top-down influences. Therefore to achieve good prediction accuracy the saliency models should incorporate both generic bottom-up features in the images and generic top-down features such as human faces In the non-frontal face category, the proposed model does not perform due to challenges in face detection. However, the proposed model still performs better than other state-of-the-art attention models.

Table 6.2: Quantitative Comparison of Saliency Models on Subsets of Judd dataset

Year	Models	Images without faces (739)			Images with faces: Successful face detection (182)			Images with non-frontal faces: failed/partially failed face detection (82)		
		CC	NSS	AUC	CC	NSS	AUC	CC	NSS	AUC
1998	NVT [6]	0.24	1.10	0.76	0.21	1.02	0.77	0.19	0.88	0.69
2006	GBVS [7]	0.30	1.36	0.81	0.27	1.32	0.81	0.28	1.32	0.81
2007	SR [8]	0.18	0.84	0.69	0.18	0.89	0.72	0.19	0.87	0.70
2008	SUN [70]	0.16	0.76	0.68	0.14	0.69	0.67	0.16	0.76	0.68
2010	PQFT [10]	0.13	0.63	0.57	0.09	0.47	0.54	0.12	0.59	0.56
2010	CAS [9]	0.23	1.06	0.74	0.21	1.02	0.76	0.23	1.10	0.76
2012	SS [11]	0.23	1.07	0.74	0.22	1.08	0.77	0.26	1.22	0.77
2012	RCSS [122]	0.24	1.08	0.75	0.20	0.96	0.74	0.22	1.00	0.75
2013	SDSP [13]	0.22	0.99	0.72	0.19	0.95	0.73	0.20	0.93	0.72
2013	WBSD [127]	0.19	0.90	0.71	0.17	0.85	0.71	0.19	0.89	0.71
2017	**Proposed**	**0.32**	**1.43**	**0.82**	**0.33**	**1.64**	**0.83**	**0.30**	**1.38**	**0.82**

In the case of images without human faces, the model generates the saliency map only using the in-focus detection and centre sensitivity maps. The results show that the model performs better than the state-of-the-art saliency models. Similarly on the entire dataset the proposed model outperforms the state-of-the-art in predicting human fixations which can be seen in the Table 6.3 (section 6.5.2).

6.5.2 Measure of Dispersion

The research community has used popular metrics like CC, NSS and AUC to evaluate the performance of the saliency models. The performance is quantified by taking the mean of the metric scores of all the images within the dataset. Although they have used sophisticated metrics to quantify the performance, all of these metrics are used only to measure the central tendency or mean across the dataset. However, measuring the central tendency in itself is not enough to describe the performance. It only indicates the global performance of the saliency model across a dataset, but fails to determine how effective and consistent the model is across each image (different visual stimuli) within the dataset. In order to determine the consistency, the dispersion measure was computed across the dataset.

In statistics the commonly used techniques of dispersion measure are range, interquartile range and Standard Deviation (SD). Range is a simple measure of dispersion; however, it is very sensitive to outliers and does not use all the observations in the dataset. Interquartile range has an advantage of not being affected by the extreme values; however, the main disadvantage regarding this measure is not being amenable to mathematical manipulations. On the contrary SD is a widely used technique of dispersion measure in statistics. It considers all the values of each image in determining the spread. So far, the dispersion measure for saliency consistency has not been used by any author to evaluate the performance of a visual attention model. The higher the SD, the lower is the consistency. In applications like video compression, a higher mean and a lower SD are extremely important for a model to be used across each and every video frame. Therefore, depending on the SD and mean, the models were categorised as follows.

i) **High mean and low SD**: Indicates that the chosen model features and combination strategies are very good for computing saliency. Moreover the model consistently detects salient regions with high prediction accuracy across different visual stimuli in the dataset.

ii) **High mean and high SD**: Indicates that the model is highly inconsistent. Such a type of model is actually tuned to detect salient regions only in certain types of images. These models achieve high prediction accuracy in few images and a very low accuracy in rest of the images. Therefore, the features employed in developing the model are not really effective.

iii) **Low mean and high SD**: Indicates that the model is highly inconsistent with low prediction accuracy across all the images.

The performance of the saliency model is evaluated on the entire Judd's dataset using three metrics in terms of mean and SD. These results are shown in Table 6.3. It can be inferred from the table that the proposed model achieves higher CC and low SD compared to the state-of-the-art models. The RCSS model achieves low SD similar to the proposed model. However, the prediction accuracy of RCSS is very low compared to the proposed model. The other metrics (NSS and AUC) also indicate a high accuracy with low SD for the proposed model.

Table 6.3: Comparison of Dispersion measure of Saliency Models on Judd dataset

Judd's dataset (1003)						
Models	**CC**	**SD**	**NSS**	**SD**	**AUC**	**SD**
NVT	0.23	0.13	1.09	0.66	0.76	0.11
GBVS	0.29	0.11	1.35	0.63	0.81	0.08
SR	0.18	0.15	0.85	0.78	0.69	0.14
SUN	0.15	0.13	0.75	0.65	0.67	0.13
PQFT	0.12	0.12	0.59	0.67	0.56	0.08
CAS	0.22	0.14	1.05	0.73	0.74	0.12
SS	0.23	0.15	1.08	0.78	0.74	0.12
RCSS	0.23	0.10	1.05	0.53	0.75	0.08
SDSP	0.21	0.11	0.97	0.58	0.72	0.10
WBSD	0.18	0.12	0.88	0.64	0.71	0.12
Focus+Center	0.30	0.11	1.36	0.57	0.81	0.09
Proposed	**0.32**	0.10	**1.46**	0.52	**0.82**	0.07

6.5.3 Database Independence

The performance of the model is tested using a larger non-training publicly available DUT-OMRON [141] image dataset. The dataset consists of 5168 images manually selected from more than 140,000 images. According to the authors these images have salient regions with relatively complex background. The ground truth eye tracking fixations are collected using 5 participants with normal or corrected to normal vision. The results of applying our model on DUT-OMRON are presented in Table 6.4. It is clear from the table that the performance of the proposed model is not database dependent and once trained it is capable of predicting fixations irrespective of the type of visual stimuli.

Table 6.4: Quantitative Comparison of Saliency Models on DUT-OMRON Dataset

Models	CC	NSS	AUC
NVT	0.36	1.40	0.81
GBVS	0.40	1.45	0.83
SR	0.26	1.10	0.73
SUN	0.21	0.86	0.70
PQFT	0.18	0.76	0.62
CAS	0.32	1.37	0.78
SS	0.32	1.37	0.78
RCSS	0.33	1.33	0.80
SDSP	0.32	1.24	0.76
WBSD	0.30	1.23	0.77
Proposed	**0.41**	**1.50**	**0.84**

6.5.4 Qualitative Analysis

The saliency maps for four sample images from Judd's database for our model and ten state-of-the-art saliency models are given in Figure 6.9 (a) and Figure 6.9 (b). The Ground Truth Fixation Maps (GTFM) is also shown. In the first image (flower), the ground truth indicates the users gazed mostly at the centre of image. The saliency map of the proposed model exhibits higher intensity at the centre when compared to the other models. It also has lower false detections when compared to other models which detect unimportant areas as salient. In the second image, both GBVS and the proposed model gave closer performance to the ground truth. In the last two images the ground truth maps indicate high user sensitivity towards face regions. The proposed model also shows higher intensity at the face regions for both images, compared to other models. Most other saliency models fail to detect these regions as they ignore top-down features.

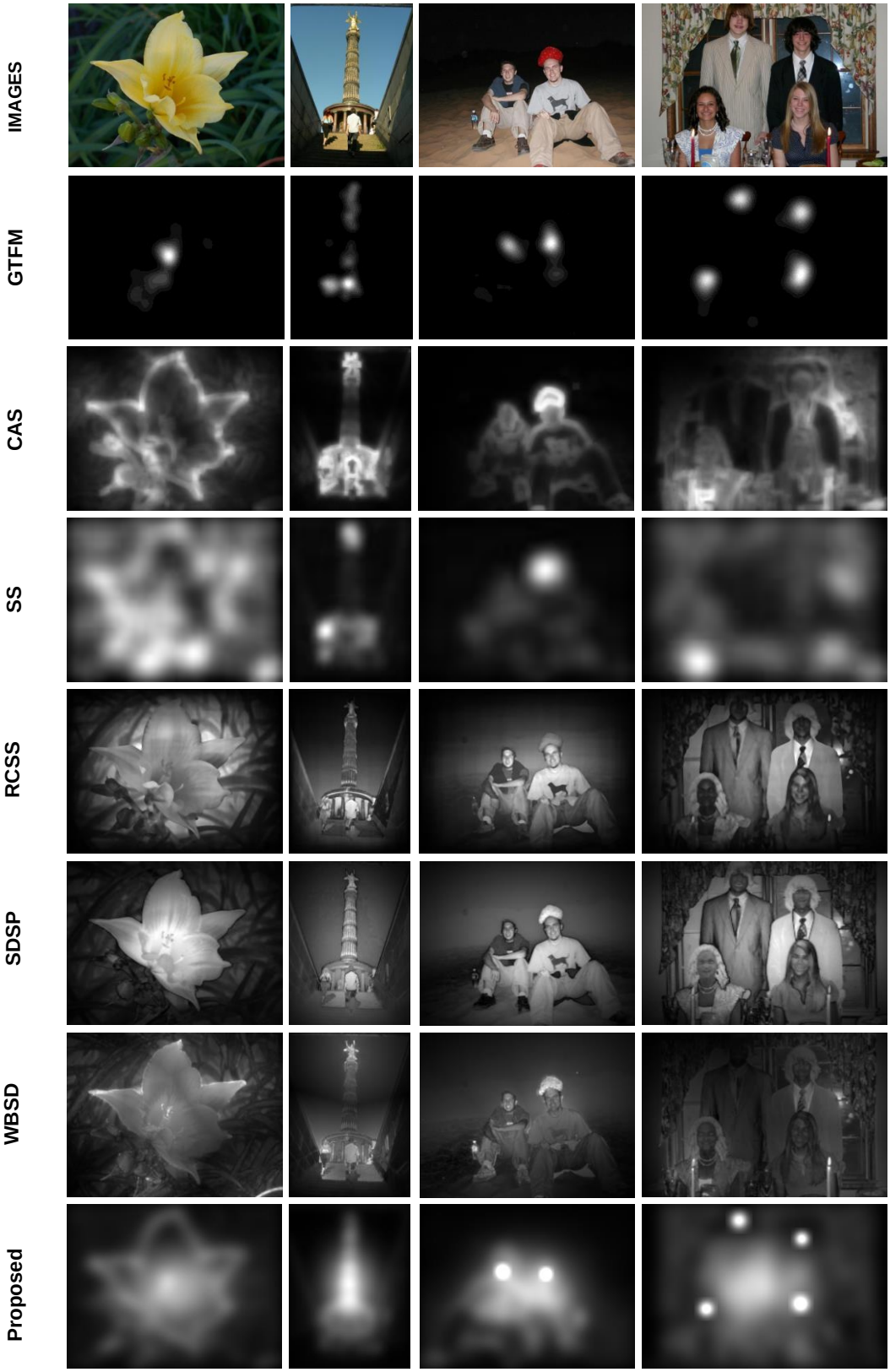


Figure 6.9a: Qualitative comparison of the state-of-the-art visual saliency models for four sample images from Judd's dataset

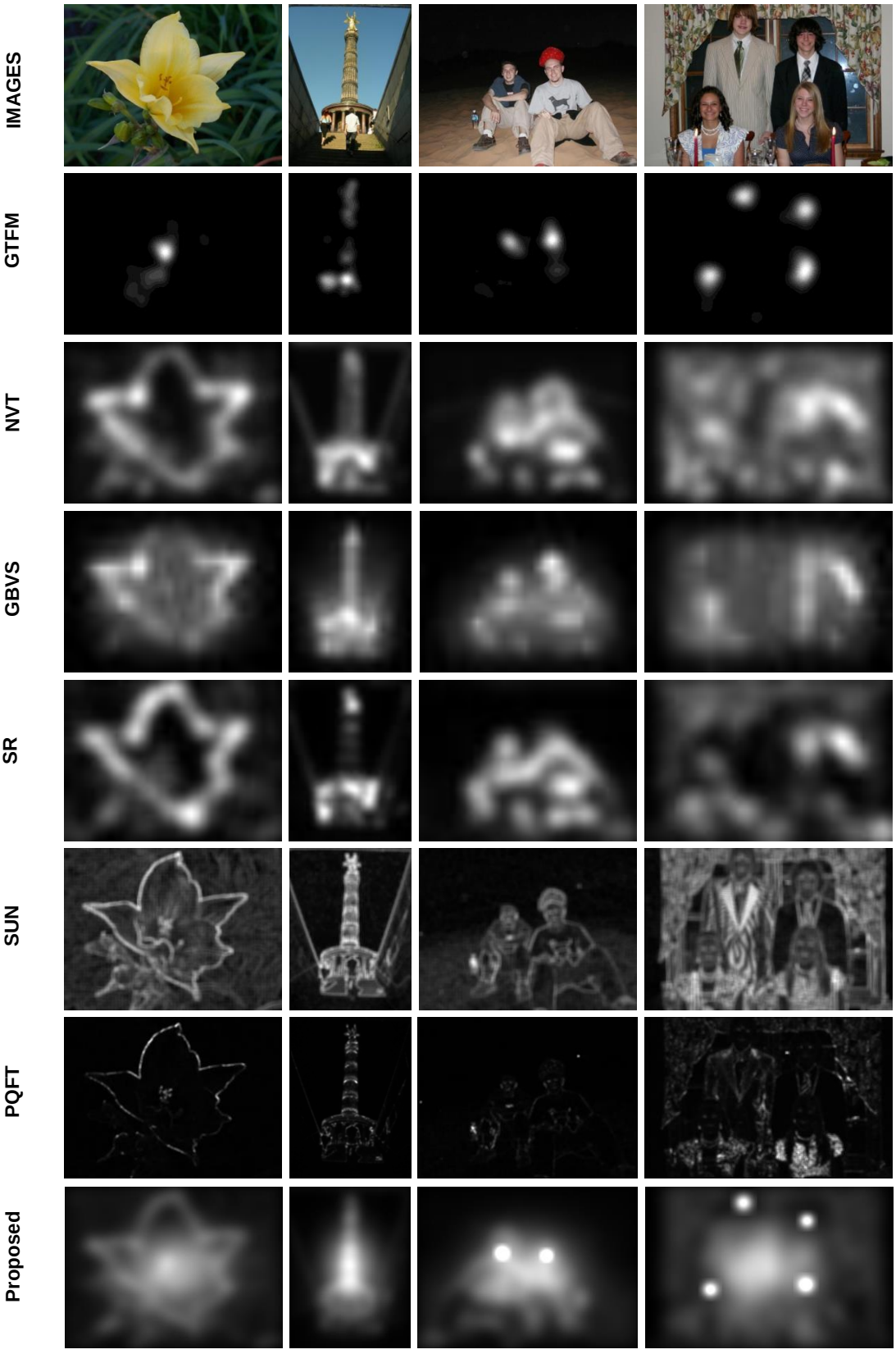


Figure 6.9b: Qualitative comparison of the state-of-the-art visual saliency models for four sample images from Judd's dataset

6.5.5 Computational Complexity

In a time constrained scenario, a saliency model should be fast enough to meet the real time performance requirements. As mentioned in section 6.4, the proposed model uses the Integer cosine transform (ICT) instead of the traditional Discrete Cosine Transform (DCT). The model complexity using DCT and ICT and the complexities of all the individual components of the proposed model are provided in Table 6.5 and 6.6 respectively. It can be inferred from Table 6.5 that the proposed model using ICT saves 30% of complexity with a slight gain in prediction accuracy compared to the model using DCT. In the Table 6.6 it can be observed that almost 50% of the time is occupied by face detection. However, as shown in Table 6.2 considering human faces for saliency detection contributed significantly to prediction accuracy in the images with faces and therefore makes it an important choice for saliency detection.

Table 6.5: Prediction accuracy and complexity comparison of the proposed model using DCT and ICT (MATLAB)

Saliency Model	Prediction Accuracy			Complexity (secs)
	CC	NSS	AUC	
Model using DCT	0.3202	1.4607	0.8222	0.98
Model using ICT	0.3208	1.4637	0.8224	0.68

Table 6.6: Complexity comparison of the individual components of the proposed model (MATLAB)

Saliency Model Individual Components	Complexity (secs)
In-Focus detection	0.33
Centre detection	0.01
Face detection	0.34

The complexity of the proposed model is compared with the state-of-the-art saliency models in Table 6.7. An unoptimised MATLAB code (without MEX code) was used for all of these saliency models in order to ensure a fair comparison. The average time required to compute a saliency map is calculated over 100 images with resolution

1024x768 from Judd's database. It is evident that, SR, SDSP and SS are the fastest among the models. However, they are limited in terms of prediction accuracy. Compared to our model, GBVS is the next best performing model in terms of accuracy. However, it is still significantly more complex than our model. The proposed model achieves 42% of complexity reduction and better prediction accuracy compared to GBVS and therefore outperforms state-of-the-art saliency models.

Table 6.7: Complexity comparison of state-of-the-art saliency models

Models	NVT	GBVS	SR	SUN	PQFT	CAS	SS	RCSS	SDSP	WBSD	**Ours**
Complexity (secs) (MATLAB)	0.76	1.18	0.01	9.36	0.83	45.96	0.03	5.01	0.06	24.28	**0.68**

6.6 Discussion

A low complexity visual attention model is proposed by improving the DCT based in-focus visual attention model proposed in chapter 5. The DCT operation is computationally intensive as it involves many floating point multiplications. Therefore, to speed up the entire process an Integer Cosine Transform (ICT) is used instead of the traditional DCT. The integer based transform involves integer arithmetic (additions and possibly multiplications), and thus its implementation is greatly simplified compared to the DCT. Moreover, in the earlier model, a DCT based focus map was calculated using Y-channel of YC_bC_r colour space. However, in the improved model the focus map is computed in HSV colour space. HSV is a perception oriented non-linear colour space. The colour information is represented by hue and saturation and the colour's brightness (the amount of light) is indicated by the value channel. HSV colour space is more intuitive to human vision for its good ability of representing the colours of human perception. The Human visual system (HVS) is more sensitive to lightness than saturation and hue. Therefore, the value channel of HSV colour space is used for calculating the saliency map. Moreover, as the saliency is calculated only in 2-D value space, it reduces the computational complexity and memory utilisation of the model. The results demonstrate that by using ICT and HSV colour space, significant computational complexity savings and a slight improvement in the prediction accuracy has been achieved. The in-focus detection consumes 0.33 seconds (refer to section 6.5.5) for computing the focus map for images with resolution 1024x768 on Intel core I7-2600K CPU.

Centre sensitivity is incorporated into the model to optimise the accuracy. The centre sensitivity map has been obtained by using an anisotropic 2D Gaussian as a function of image resolution. It has been hypothesised that the viewer's gaze is oriented vertically when the height of the image is greater than the width of the image and vice versa. This is the main reason behind using anisotropic Gaussian distribution for modelling image centred viewer's gaze. The time complexity of the model in calculating the centre map has been evaluated over 100 images with resolution 1024x768. The model needs an average time of 0.01 seconds for generating the centre map. The performance of the model was analysed qualitatively and quantitatively on Judd's image dataset available in the research community. The results have shown that the model has achieved a prediction accuracy of CC=0.30, NSS= 1.36 and AUC=0.81 on Judd's dataset (refer to section 6.5.2). The proposed model has achieved a 1% prediction accuracy (with respect to CC metric) improvement with 62% of computational complexity savings when compared to the GBVS model. GBVS is the best model among the chosen benchmark state-of-the-art saliency models [70], [10], [127], [6], [7], [8], [11], [122], [13], [9] according to qualitative and quantitative analysis. When compared to the earlier saliency model which detects in-focus regions using DCT coefficients, this model has achieved a 12% improvement in prediction accuracy (with respect to CC metric) with 57% of complexity reduction.

The prediction accuracy of the model is further improved by including human face map. The main reason behind choosing face sensitivity for modelling visual attention was that in free viewing conditions human faces attract viewers more when compared to other top-down features and it is also a common top-down bias for majority of viewers. Further, there is also evidence in the literature that viewers tend to look at human faces independent of the task at hand. The Viola Jones face detection algorithm was used to detect human faces due to its wide usage and effectiveness at detecting human faces. Face maps were mapped by using square shaped bounding boxes around the human faces. This process involves generating a binary map of zeros of original image resolution and placing Gaussian blobs in the regions where human faces are located. The Gaussian distribution has been confined to the bounding box to avoid the false detections (considering non salient regions as visually salient). The parameters of the face map have been tuned using the hill climbing approach. The computational complexity of the model in calculating the face map has been evaluated over 100 images with resolution 1024x768. The model on an average consumes 0.34 seconds for generating the face saliency map which includes time needed for face detection and generating the corresponding face map. The model achieved a prediction accuracy of CC=0.41, NSS=1.50 and AUC=0.84 on the DUT-OMRON

dataset and an accuracy of CC=0.32, NSS= 1.46 and AUC=0.82 on the Judd's dataset (refer to section 6.5.3 and 6.5.2). When compared to the GBVS model which is the best among the benchmark state-of-the-art visual attention models, the proposed model has achieved 3% and 1% of prediction accuracy (with respect to CC metric) improvement on Judd's and DUT-OMRON datasets respectively. Further, it has achieved 42% of computational complexity savings when compared to the GBVS model (the GBVS model has achieved the best prediction accuracy among the chosen benchmark saliency models). Note that the computational complexity of the entire model in calculating the saliency map has been evaluated over 100 images with resolution 1024x768 on Intel core I7-2600K CPU. Further, this model has achieved 2% improvement in the prediction accuracy (with respect to CC metric) when compared to the earlier visual attention model which detects in-focus regions using ICT coefficients and centre sensitivity on the Judd's image dataset. The advantages and disadvantages of this model are summarised below.

Advantages

- This model achieves better prediction accuracy with significant reduction in computational complexity when compared to the DCT based in-focus visual attention model.

- In the scenario when an image is completely out-of-focus or in-focus the model is still able to derive the salient regions by detecting image centre as salient. When there are no features of interest in the images then there is a high probability that a viewer's gaze will be oriented towards image centre because of photographer bias and viewing strategy.

- The proposed model has achieved a prediction accuracy of CC=0.33, NSS=1.64 and AUC=0.83 for images with in-focus regions and frontal human faces. The state-of-the-art has achieved a prediction accuracy of CC=0.27, NSS=1.32 and AUC=0.81 (refer to section 6.5.1). This clearly indicates 6% of improvement in prediction accuracy (with respect to CC metric) over the state-of-the-art attention models for images with frontal human faces. This indicates that the model is good for detecting salient regions with frontal human faces. Further, the model requires an average time of 0.68 seconds for computing the saliency map of images with resolution 1024x768. This is 42% of complexity reduction when compared to the GBVS model which achieves the best

prediction accuracy among the chosen state-of-the-art attention models for this research.

- The proposed model has achieved a prediction accuracy of CC=0.32, NSS=1.43 and AUC=0.82 in images where top-down features such as human faces are absent (refer to section 6.5.1). In this case the GBVS model (the best among the state-of-the-art models) has achieved a prediction accuracy of CC=0.30, NSS=1.36 and AUC=0.81. This indicates 2% of improvement in prediction accuracy (with respect to CC metric) over the state-of-the-art for images without human faces. (Note that no improvement can be seen in computational complexity as the face detection algorithm operates even on images without human faces). The proposed model is able to outperform the state-of-the-art based on in-focus regions and centre sensitivity in the absence of human faces.

Disadvantages

- The model might produce inaccurate results with peripheral salient regions. The main reason behind this is the model giving more priority to the image centre than periphery. Due to this, less importance is given to peripheral salient regions in the images. However, the probability of images with peripheral salient regions is very low because of photographer bias and viewing strategy.

- In an image with crowd scene there will be only few visually salient human faces in which viewers are interested. In contrast to this the proposed model detects all the faces present in the image as visually salient. These false detections in the face saliency map will result in the performance (prediction accuracy) drop of the attention model.

- The model uses frontal face detection algorithm and hence detects only frontal human faces present in the images. The model produces inaccurate results if there are non-frontal faces present in the images.

- Although the model is better than the state-of-the-art, its prediction accuracy and complexity has to be further improved to use it in real time applications like image/video compression.

6.7 Conclusion

In this chapter, a low-complexity visual saliency detection model for detecting salient regions in images is proposed. The salient regions are detected using three main aspects that attract the attention of the Human Visual System. They are, (a) in-focus areas – mapped using ICT based salient frequency detection, (b) image centre sensitivity – mapped using a 2D Gaussian distribution and (c) human faces – mapped using face detection and Gaussian blobs. The model parameters are tuned using a hill climbing algorithm. The performance of the saliency model in predicting human eye fixations is evaluated against ten state-of-the-art visual saliency detection models using two publicly available datasets. The results demonstrate that the proposed model shows higher prediction accuracy in saliency detection at significantly lower computational complexity compared to other state-of-the-art saliency models.

7 Evaluation of the Effectiveness of Video Quality Metrics in Quality Assessment of Pre-processed Video

7.1 Introduction

In the recent years a growing interest has been witnessed in pre-processing based perceptual video quality optimisation algorithms [183], [184], [185], [186], [187]. As subjective video quality evaluation is a complex and time consuming activity, objective video quality metrics which can be used to detect perceptual quality variations when videos are pre-processed will make the development of algorithms easier. These objective video quality metrics are essential for speeding up video quality tests. Moreover, objective metrics are widely employed during the development of perceptual video quality optimisation algorithms as they can be easily implemented in software or hardware to generate results automatically without viewer intervention. This chapter investigates nineteen state-of-the-art objective video quality metrics to determine the effectiveness of the metrics in detecting the perceptual variations induced by Gaussian pre-processing filter. These metrics include Full Reference (FR) [188], [189], [190], [191], [192], [193], No Reference (NR) [194], [195], [196], [197], [198] and Reduced Reference (RR) [199] video quality metrics. The results show that either of these metrics effectively detects the quality variations when videos are pre-processed. However, No Reference metrics show better performance when compared to both FR and RR metrics. In particular, the Naturalness Image Quality Evaluator (NIQE) [198] is notably better at detecting perceptual quality variations. Moreover, the traditional methods of evaluating video quality metrics such as the Spearman Rank Order Correlation Coefficient (SROCC) [200] and Pearson Linear Correlation Coefficient (LCC) [200] are shown to be ineffective in determining the effectiveness of the quality metric in detecting variations in the perceived quality, particularly when the metrics are to be employed during the development of pre-processing based perceptual video quality optimisation algorithms.

This chapter is self-contained with its own background, literature survey, methodology, experimental procedure and the results. In the section 7.2 the main hypothesis is explained. The subjective testing methods and the objective video quality metrics such as Full Reference (FR), Reduced Reference (RR) and No Reference (NR) metrics available in the literature are briefly described in the section 7.3. Moreover the benefits and drawbacks of each of these approaches are briefly outlined. In section 7.4, the perceptual video quality optimisation algorithms and the testing methods used

during the development of these algorithms are provided. The test video sequences, the procedure followed for pre-processing the selected video sequences and the video quality evaluation methods used in this work are explained in detail in the section 7.5. The results obtained with all the chosen quality metrics for this study are analysed in the section 7.6. Finally, in the section 7.7 this study is critically discussed and the main conclusions drawn from this study are summarised in section 7.8.

7.2 Hypothesis

In the literature there is very limited evidence to determine the suitability of mathematical error based measurements such as PSNR or other FR, RR or NR perceptual quality metrics for the measurement of perceptual video quality variations induced by video pre-processing. In this work, it is hypothesised that image/video quality metrics which includes error-based and/or FR, RR and NR metrics tend to produce inaccurate measurements when significant pixel variations are induced by pre-processing them using filters. This is mainly because the pixel variations that are induced by filtering the frequency components are typically interpreted as distortion by these metrics. Although some kind of filtering operations achieve subjective quality gain, the objective video quality metrics tend to detect this as perceptual loss. Therefore, the objective of this work is to evaluate the effectiveness of state-of-the-art image/video quality metrics in measuring the quality of pre-processed and coded video. During the investigation, a number of videos will be pre-processed at different filter intensities and coded at various bitrates. Later the quality of the coded video sequences is evaluated using subjective quality testing procedure. The performance of the video quality metric is evaluated by determining their effectiveness in detecting the perceptual gain/loss that is observed during subjective video quality evaluations.

7.3 Video Quality Evaluation

Video quality evaluation is the process of determining the quality of the videos using either subjective or objective video quality measurement strategies. These two kinds of approaches are described in this section. The subjective testing methodology used in the current work and the different categories of objective quality metrics chosen for this study is discussed. Moreover, the advantages and disadvantages of each of the approaches are also outlined.

7.3.1 Subjective Video Quality Assessment

During subjective video quality assessment human subjects are used to judge the quality of the videos. In the literature several subjective quality assessment

strategies are proposed for multimedia applications [201]. Some of the popular methods include Single Stimulus Continuous Quality Evaluation (SSCQE), Double Stimulus Continuous Quality Scale (DSCQS), Double Stimulus Impairment Scale (DSIS) and Pair Comparison (PC). In the current study, the SSCQE method is used to evaluate the quality of the videos as it has the ability to obtain repeatable results during video quality evaluation [202], [203]. In this method test video sequences are presented one at a time and then rated using a rating scale. After presenting the test sequence, a voting time of less than or equal to 10 s is given to rate the video quality. The timing of the stimulus presentation is shown in the Figure 7.1.

Figure 7.1: Test sequence presentation in the SSCQE method (source [201])

Figure 7.2: Eleven point quality rating scale (source [201])

Generally a 5-point rating scale is used however, in the current work an eleven point scale is used as a higher discriminative power is needed during the quality evaluation. The scale is shown in the Figure 7.2. The number 0 indicates worse quality or the reproduction of the video is in no way similar to the original video shown. The number 10 indicates best quality and the reproduced video sequence has the quality similar to the original video and in no way can be improved further. The quality ratings after watching the test sequences by the subjects are written down on a response sheet. These numerical responses from the viewers are averaged to draw conclusions regarding the quality of the video. The number of subjects used for the test generally range from 4 to 40. The general recommendation is to use at least 15 viewers for the test and the actual number depends on the required validity. Prior to the actual subjective quality test a small group with 4 to 8 video quality experts can be used for obtaining indicative results. These subjective quality methods generally provide a reliable video quality assessment; however, they are very expensive in terms of amount of time and complexity.

7.3.2 Objective Video Quality Assessment

The video quality evaluation using a mathematical algorithm is called objective video quality assessment. Extensive research performed in the area of video quality assessment has produced three categories of algorithms namely FR, RR and NR algorithms. The Full Reference (FR) algorithms access the quality of the degraded image/video sequence by making a comparison with reference image/video. In the case of Reduced Reference (RR) metrics specific features are extracted from the reference and the image/video under test. These specific features that are extracted from the reference Image/video are sent to the receiving system via a communication channel to evaluate the quality. These features include blurriness, blockiness, spatial and temporal information. No Reference (NR) metrics quantify the image quality without the need of pristine images. It means these video quality metrics do not require original image/video as the reference for comparison with the image/video under test. In contrast to both FR and NR video quality metrics, NR metrics determine the visual quality based on the local statistics of the video. An objective video quality evaluation system is shown in Figure 7.3. The input video sequence is initially encoded and sent over the network or stored. The received video is then decoded and displayed at the end user system. It can be seen in the figure that the Full Reference quality metric needs both the original and the decoded video or video under test to evaluate the video quality whereas the No Reference (NR) video quality metric needs only the video under test to quantify the perceived quality. The current investigation considers FR, RR and

NR metrics in evaluating their performance. The different types of FR, RR and NR metrics available in the literature are briefly discussed in the following sections.

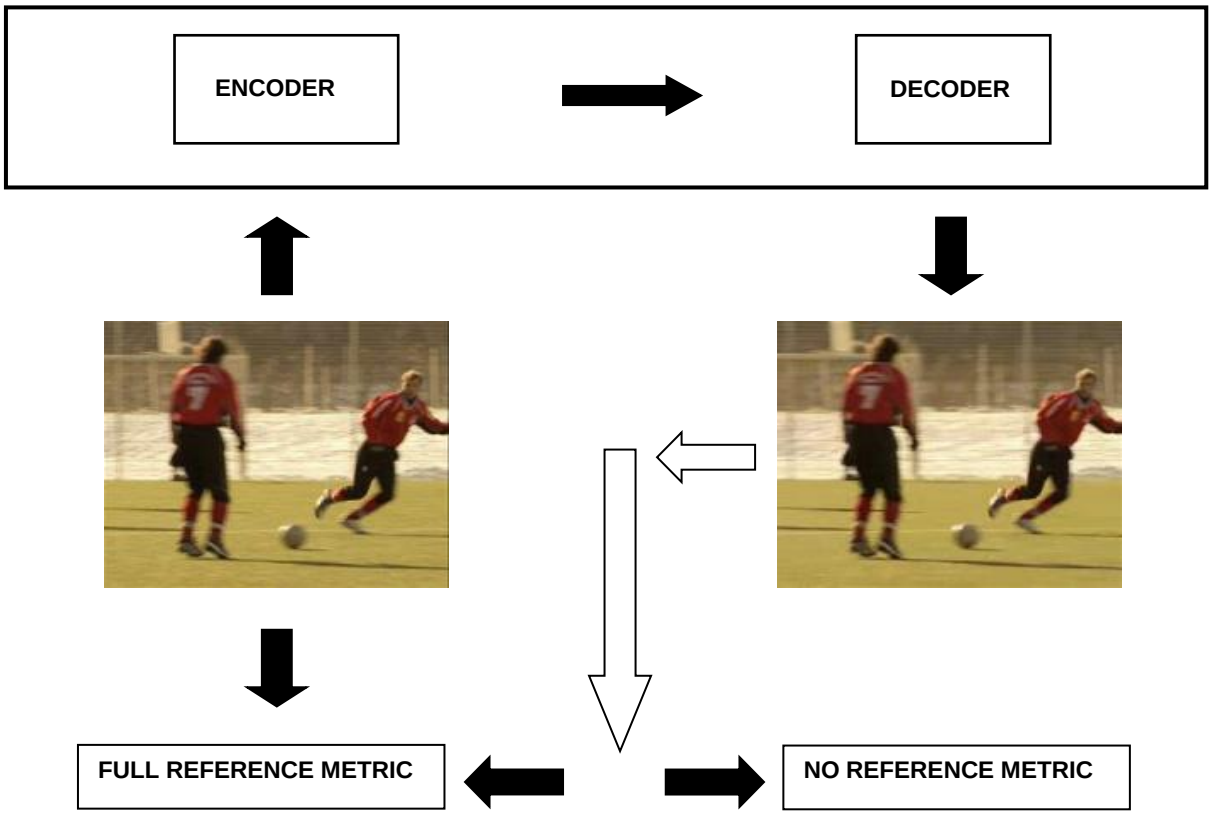


Figure 7.3: Objective Video Quality Evaluation System

7.3.2.1 Full Reference (FR) quality metrics

In traditional error based metrics such as Mean Squared Error (MSE), PSNR, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Signal to Noise Ratio (SNR) the fidelity is computed as the sum of mathematical operations on individual pixels (i.e. pixel error measurements). Similar to these metrics, Zhou and Bovik [204] proposed Universal Quality Index (UQI) which is also mathematically defined metric to measure the perceived quality. In this metric, the distance between the test and the original signal is measured as a function of luminance, loss of correlation and contrast distortion. According to the authors the metric has better prediction ability at detecting blurring distortion compared to MSE. However, the effectiveness of the metric is studied only with respect to images but not with temporally correlated images such as videos. The authors of [188] further extended UQI and proposed Structural Similarity Index (SSIM) FR metric based on the hypothesis that changes in structural information can be well approximated to perceived distortion in the videos. To compute the visual distortion they compare pixel

intensity patterns of normalised luminance and contrast for both the original and the video under test. The SSIM is further improved in [189], named MS-SSIM and it has the ability to incorporate details of the images at different resolutions.

The Video Quality Metric (VQM) [190], also known as NTIA-VQM, considers distortions that occur during the video coding process and the transmission phase in computing the perceived visual quality. The authors of [205] for assessing the video quality consider the statistical information shared between the source and the test image as the fidelity measure. This concept is further extended to Visual Information Fidelity (VIF) measure in [191] .

The video quality metrics based on the characteristics of Human Visual System (HVS) are proposed in [192, 193]. In [192] a Noise Quality Measure (NQM) is proposed. It is a two-step process in which the source and the modeled restored image (Original image processed using restoration algorithm) are initially processed by using a contrast pyramid. Later the NQM is computed as the SNR of the restored degraded image and model restored image. The authors of [193, 206] quantified fidelity based on near and suprathreshold distortions. It is also a two stage process similar to NQM where a contrast threshold is defined for detecting perceived distortions. The perceived contrast of the distortions is computed and the extent to which it degrades the visual quality is modeled as Visual Signal to Noise Ratio (VSNR).

As the approach has access to the reference, Full Reference (FR) video quality metrics usually tend to obtain better accuracy at predicting the perceived quality. These metrics are used in designing image/video quality optimisation algorithms. However, the main drawback is the computational complexity and the requirement of reference image/video under test.

7.3.2.2 Reduced Reference (RR) quality metrics

The authors of [199] proposed a RR algorithm known as Reduced Reference Entropy Differencing (RRED) metric in which the scaled local entropy differences between the source and test image in the wavelet domain are computed to determine the perceived video quality. Their work achieves a Spearman correlation value of 94.53%, when images are Gaussian blurred.

Reduced complexity compared to Full Reference metrics is the main advantage of these metrics. In contrast to these FR video quality metrics, in RR metrics the specific information related to the reference is transmitted over a communication medium that is needed for the computation of perceived video quality. The need of a communication channel for the delivery of the features is the main drawback of RR metrics.

7.3.2.3 No Reference (NR) quality metrics

In addition to the FR and RR metrics, several NR metrics are also proposed in the literature. The statistics of locally normalised coefficients in the spatial domain are used to determine the overall perceived quality in Blind/Reference less Image Spatial Quality Evaluator (BRISQUE) [194]. A two stage NR metric that determines quality based on Natural Scene Statistics (NSS) is proposed in Blind Image Quality Index (BIQI) [195]. The compression induced effect on the nonlinear dependencies of natural scenes is quantified as a measure of perceived quality in [207].

In the literature, several blur based NR metrics are also proposed. In the work by Ferzli and Karam [196] initially the Just Noticeable Blur or the blur intensity that can be masked around the edges is determined. The perceived image quality is then determined by using JNB as a function of contrast. The metric is further extended in [197] by combining JNB with Cumulative Probability of Blur Detection (CPBD). The authors of [208] estimated the blur in the DCT domain. The perceived quality is approximated to changes in the blur levels caused by the variations in edge strength. A low complexity video quality evaluation method for JPEG compressed images is proposed in [209]. The blocky artifact influences and the observed blur are both used in evaluating the image quality. In [210] a computationally efficient metric is developed in which the image quality is computed by estimating the sharpness based on localised frequency content analysis. The software implementation of blur based metrics are available to download from [211].

An opinion and distortion unaware Naturalness Image Quality Evaluator (NIQE) NR metric is proposed by the authors of [198]. This metric uses a 'quality aware' collection of statistical features, using a database of natural undistorted images, based on a space domain Natural Scene Statistics (NSS) model. A similar set of features based on the NSS model, are extracted from the image under test. The fidelity between quality aware features and multivariate Gaussian fit of NSS features extracted from the image under test is quantified as perceived quality. The NIQE implementation along with some other FR, RR and NR metrics are available to download from [212].

The prediction ability of these metrics tends to be low when compared to FR and RR metrics as the reference image/video is not available. However, their main advantage is the ability to assess the quality without the reference. All the FR, RR and NR metrics discussed above will be evaluated using the experimental procedure described in section 7.6 to determine their effectiveness in detecting quality variations.

7.3.2.4 Performance Evaluation of Video Quality Metrics

According to Video Quality Experts Group (VQEG) [200] the performance of

any objective video quality metric can be evaluated by analysing three important attributes such as prediction accuracy, monotonicity and consistency of the developed assessment model. The VQEG stated that these attributes can be computed by using mathematical measures such as Spearman Rank Order Correlation Coefficient (SROCC), Linear Correlation Coefficient (LCC), Outlier ratio, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Among these, the LCC and SROCC are widely used methods to quantify prediction accuracy and monotonicity respectively. The Pearson Correlation measures the strength of linear association between two variables. The Pearson Correlation (r_p) for N data pairs (x_i, y_i) can be defined as:

$$r_p = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}_m\right)\left(y_i - \overline{y_m}\right)}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x}_m)^2}\sqrt{\sum_{i=1}^{N}\left(y_i - \overline{y_m}\right)^2}} \tag{7-1}$$

(r_p)- is the Pearson correlation and (x_m, y_m) indicate Means.

Accordingly the Spearman Correlation measures the strength of association between two ranked variables. The Spearman Correlation (r_s) for N ranked data pairs (x_r, y_r) can be defined as:

$$r_s = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}_r\right)\left(y_i - \overline{y_r}\right)}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x}_r)^2}\sqrt{\sum_{i=1}^{N}\left(y_i - \overline{y_r}\right)^2}} \tag{7-2}$$

(r_s)- is the Spearman correlation and (x_r, y_r) indicate Mid-ranks,

As the Spearman correlation is computed on ranks, it depicts a monotonic relationship between the variables. In contrast, the Pearson correlation is measured on true variables and it depicts only linear relationship between the variables. The published correlation values with respect to subjective Mean Opinion Scores (MOS) for some of the discussed metrics, when images are Gaussian blurred are shown in Table 7.1. The SROCC and LCC values shown in the table are extracted from the references given against the corresponding metrics. These values are not directly comparable between the metrics as they are evaluated on different datasets. However, they give a performance estimation of a particular metric in predicting the video quality when images are Gaussian blurred. It is clear from the table that the majority of the metrics

Table 7.1: Published correlation values of objective video quality metrics

Type	Metric	SROCC	LCC
Full Reference	PSNR [195]	0.761	0.782
	UQI [193]	0.938	0.945
	NQM [193]	0.874	0.903
	VIF[193]	0.973	0.975
	VSNR [193]	0.941	0.934
	SSIM [194]	0.9321	0.9395
	MS-SSIM [194]	0.9607	0.9762
No Reference	BRISQUE [194]	0.9435	0.9498
	BIQI [213]	0.8463	0.8293
	NIQE [198]	0.9341	0.9525
	JNBM [196]	0.932	0.936
	CPBD [197]	0.9437	0.9107

achieve very high correlation values when the measured quality is compared with the actual score. However, it is not clear whether these metrics can detect quality variations accurately despite exhibiting high scores. For a metric to be employed during the development of perceptual quality optimisation algorithms its ability to measure quality variations is a key factor in determining its suitability. Therefore, in this investigation, a Gaussian low pass filter is used to simulate the quality variations and a HEVC CODEC [214] to compress the videos in determining the ability of the discussed metrics in detecting quality variations.

7.4 Video Quality Measurement for Perceptual Quality Optimisation Algorithms

The main aim of perceptual video quality optimisation algorithms is to enhance the perceived subjective video quality of the compressed video. The video coding tools that are employed in the popular video coding standards such as H.264 [80] and HEVC [81] generally aim to minimise the overall pixel errors between original and encoded video frames to optimise the perceptual quality. In contrast to these video coding tools, in pre-processing based algorithms visually insignificant frequency components are reduced to achieve improvements in the perceived quality. These are the components to which the HVS is less sensitive and this process discards them during visual processing. In the literature, a number of pre-processing based perceptual video quality optimisation algorithms that employ low-pass pre-processing filter were proposed [183-

187, 215-223].

In [183], the authors present an adaptive edge-preserving smoothing and detail enhancement pre-processing filter for perceptual quality optimisation. The authors presented results in the form of subjective MOS scores as well as PSNR. However, the results clearly showed that the PSNR does not always correlate well with the subjective quality variations. In [184] a pre-processing filter is used to remove spurious noise and insignificant features present in the video frames. The results indicated PSNR improvements, however, without any subjective verification.

In the research by Mancuso and Borneo [185], the filtering intensity is dynamically adjusted according to the amount of noise present in the video sequence to generate perceptually optimised videos. The PSNR is used as the quality metric to show that their nonlinear filters achieve higher quality videos. Similarly, in [186] and [187], the quality improvement is presented in the form of achieved gains in PSNR. Further, the authors have shown the screenshots of video frames to highlight the artefact reduction. However, no subjective evaluation was carried out.

The authors of [215] have used variable Gaussian pre-processing filters controlled by a visual quality map which indicates the distance to the Region of Interest (ROI). They have used PSNR as the objective video quality metric to show that a variable number of Gaussian filters improve the perceptual quality. However, actual subjective quality testing results were not presented. In [216], the authors interpret that the filtered surgical video with quality improvements in regions of interest is visually equivalent to non-filtered surgical video for a telesurgery application. De-Frutos-Lopez *et al.* [217] proposed texture and motion adaptive filtering in which the bilateral filter parameters are dynamically estimated based on the motion and texture present in the video. PSNR and visual comparisons of the video frames are used to demonstrate the improvements in the performance of the algorithm.

In [218], a low complexity version of a traditional bilateral filter is used to achieve faster pre-processing of videos. The authors show quality improvements using the RMSE metric and they have also provided the visual evidence of frame 31 from Foreman video sequence. However, more substantial evidence would be to use subjective quality testing procedures to complement the achieved results. In the work of Liang-Jin and Ortega [219] PSNR is used to validate the proposed rate control scheme. The rate control algorithm determines the pre-filtering strength which is then coupled with block classification to achieve improvements in PSNR.

Young and Evans [220] proposed an image pre-processing algorithm based on attribute morphology. The authors choose the image area and the power criterion based on subjective evaluations. However, they do not correlate well with the results

obtained from chosen objective (RMSE vs. compression ratio) measure. Further, to improve the compression efficiency the authors have investigated Multidimensional attribute (MA) morphology filters and proposed a sliding window AM filter in [221]. PSNR values for different attributes are presented but without bit rate savings to ascertain the compression efficiency of the proposed pre-processing algorithm.

The authors of [222] selectively attenuate the insignificant high frequency content while preserving the significant frequency content which is of interest to the human eye. PSNR is used as the measurement metric to show improvements in visual quality. However, the bit rate savings and subjective testing results are not shown. The authors of [223] claimed that by filtering noise from the videos a better compression can be achieved. However, no rate distortion curves of the pre-processed video with respect to the original videos are shown. Moreover, neither a subjective nor an objective video quality metric is used to validate the achieved visual quality improvements.

In spite of the availability of many objective video quality metrics and subjective video quality testing methods, most perceptual video quality research typically employed PSNR [183-187, 215, 217, 219, 221, 222] and RMSE [218, 220] to evaluate perceptual quality. This is mainly because the subjective video quality testing methods are expensive in terms of amount of time and the number of viewers needed to carry out the testing process. Therefore, these objective video quality metrics can be used to generate large number of test result data points that are necessary to develop robust algorithms (e.g. hundreds/thousands of bit rate – quality data points corresponding to different parameters necessary for generating mathematical models/algorithms). It is not practically possible to use subjective quality tests for a very large number of test cases. Therefore, the use of objective video quality testing for perceptual quality optimisation algorithms during the development stages is perceived to be justified, given the practical limitations of subjective testing procedures. Further, it is also evident that, when videos are pre-processed there is no reliable and widely accepted method of measuring perceived quality to assess the performance of perceptual quality optimisation algorithms. Therefore in this work, the state-of-the-art video quality metrics are evaluated to determine their effectiveness in measuring the quality of pre-processed video.

7.5 Experimental Procedure

The main aspects of the experimental procedure used in this investigation are,

I) Selection of test video sequences,

II) Pre-processing of test video sequences

III) Video quality testing procedure.

These are explained in the following sections.

A. Selection of Test Video Sequences

Experiments were carried out using five different widely used CIF resolution YUV 4:2:0 format video sequences chosen from International Telecommunication Union (ITU) test video materials. The sequences are *Coastguard, Mother and Daughter, Soccer, Hallmonitor and Crew* [224]. Sample video frames from these sequences are shown in the Figure 7.4. They are chosen to represent a different level of motion and detail, a range of video content and camera movements. These sequences are described below.

(a) Soccer: This is a scene with players on the soccer field. The players continuously run on the field kicking the ball. There are high levels of detail and movement in the sequence with continuous change in the background. The camera pans to track the players.

(b) Mother & daughter: This is a scene of a woman and a child sitting in a room. The woman talks to the camera while stroking the child's hair. The sequence has moderate amount of detail with some head and hand movements. The camera is static and captures the frontal view of both mother and daughter. It has low to moderate amount

(a) Soccer

(b) Mother & daughter

(c) Crew

(d) Hall monitor

(e) Coastguard

Figure 7.4: Sample frames from video sequences in CIF format – (a) Soccer (b) Mother & daughter (c) Crew (d) Hall monitor (d) Coastguard

of movement.

(c) Crew: This is a scene with a team of crew members walking with some of them smiling and waving their hands. There are many head, leg and hand movements with high amount of detail. There is a moderate change in the background. The camera moves to capture the view of the crew.

(d) Hall monitor: This is a scene with two people walking opposite towards each other in the hall way. The surveillance camera is located at the end of the hall monitoring the people entering and leaving the hall. The video clip has less amount of detail. The camera is stationary and the background does not change.

(e) Coastguard: This is a scene with a cruise moving at high speed in the water. The video sequence has moderate to high amount of detail with its background changing continuously. The camera pans to capture the cruise in motion.

B. Pre-processing stage

The state-of-the-art perceptual quality algorithms induce spatially adaptive filtering variations to achieve perceptual quality optimisation. The majority of them have used filters (such as Gaussian and Bilateral) to obtain these filtering variations. In the current work Gaussian pre-processing filter has been used to induce the pre-processing based quality variations. Note that the objective of the work is to test the performance of the metrics in detecting quality variations but not Gaussian pre-processed video. Here a Gaussian filter is only used to simulate the pre-processing used in many perceptual quality optimisation algorithms. Gaussian filters have become a common choice for image pre-processing because of the applications in areas such as human vision models [225], edge detection techniques [226] and scale space filtering [227]. Gaussian filters are applied during the pre-processing stage of video compression. In a perceptual quality optimisation scenario, Gaussian filtering is employed to improve the overall bitrate vs. visual distortion efficiency. The filter eliminates some unwanted high frequency components which are of lesser interest to the HVS. This process may also result in some visual distortions in the uncompressed filtered frames. However, the reduction in high frequencies may also result in better compression performance (i.e. improved bitrate vs. visual quality). Therefore, as a result, the overall quality for an equivalent bitrate may improve.

During the pre-processing phase, three different kernel sizes namely 3x3,2x2 1x1, and ten different standard deviations ranging from σ = 0.1, 0.2, 0.3,….,1.0 are chosen. Each video sequence among the chosen five video sequences are pre-

processed using the chosen kernel sizes and standard deviations. This results in ten different pre-processed versions of each original video sequence at each kernel size. These sequences are encoded and are then subjectively evaluated using the methodology described in the next section.

C. Video quality testing

In the current work, the subjective quality assessment methods recommended by ITU [228-230] are adopted for video quality evaluation. The single stimulus Absolute Category Rating (ACR) method is chosen for the experiment. A 40 inch monitor with 1920x1080 resolution was used to display the test videos. All videos were displayed at their native CIF resolution. To avoid distractions, the videos are played on a homogeneous screen and they are centered as they occupy a very small area of the screen. The videos are played at a rate of 30 frames per second with duration of 10 seconds per each sequence. They are presented one at a time with an interval of 7 seconds duration for voting time. In order to ensure an unbiased assessment of the video quality, all the video sequences are presented in a random order. Precautions are taken to avoid random votes from incoherent voters. An extended 11-point quality rating scale (from Bad to Excellent) is used during the experiment to identify subtle differences in perceptual video quality.

The subjective perceptual quality results are used as a benchmark to evaluate the effectiveness of the metrics chosen for evaluation. These subjective evaluations were carried in two phases. During the first phase of the subjective evaluation, a limited number of test subjects were utilised for quality assessment (however with large number of quality variations) to identify and select suitable quality variations that achieve perceptual gain and loss over the original video. In the second phase, some of the selected quality variations are tested using more comprehensive subjective video quality evaluation.

The 30 variations of each video sequence that were generated during pre-processing phase were encoded at four different quantisation values (namely 16, 24, 32 and 40) using an HEVC encoder, resulting in four encoded rate-quality points per each version of the sequence. A higher QP difference of 8 was chosen to ensure the effectiveness of the metrics in detecting the variations in perceptual quality can be studied at both low and high bitrates. 15 non-expert viewers were used for the subjective video quality assessment. Based on the subjective results two test cases (kernel sizes) were identified that achieve perceptual gain and perceptual loss. These two selected test cases will undergo a comprehensive subjective video quality assessment in the next phase.

During the second phase, the pre-processed and coded versions of each video sequence (QP = 16, 24, 32 and 40) of the selected test cases were subjectively evaluated by 60 non-expert viewers using standard test methodology as described earlier. The higher number of subjects and an extended 11-point rating scale helps in identifying subtle video quality variations. Furthermore, during the objective video quality assessment, 19 objective video quality metrics were chosen for evaluating the perceptual quality of preprocessed videos.

7.6 Results and Discussion

The Mean Opinion Score (MOS) values from the actual subjective tests were plotted against the bit rate for all encoded versions of the Crew video sequence at different standard deviations as shown in Figure 7.5. It can be observed that the Gaussian filter with standard deviation σ = 0.3 produces higher perceptual quality vs. bitrate performance. All the other blur levels (standard deviations) produced a rate-perceptual quality loss. Figure 7.6 shows the PSNR vs. bitrate plots for the same Crew video sequence. It is evident that PSNR is steadily decreasing with an increase in standard deviation.

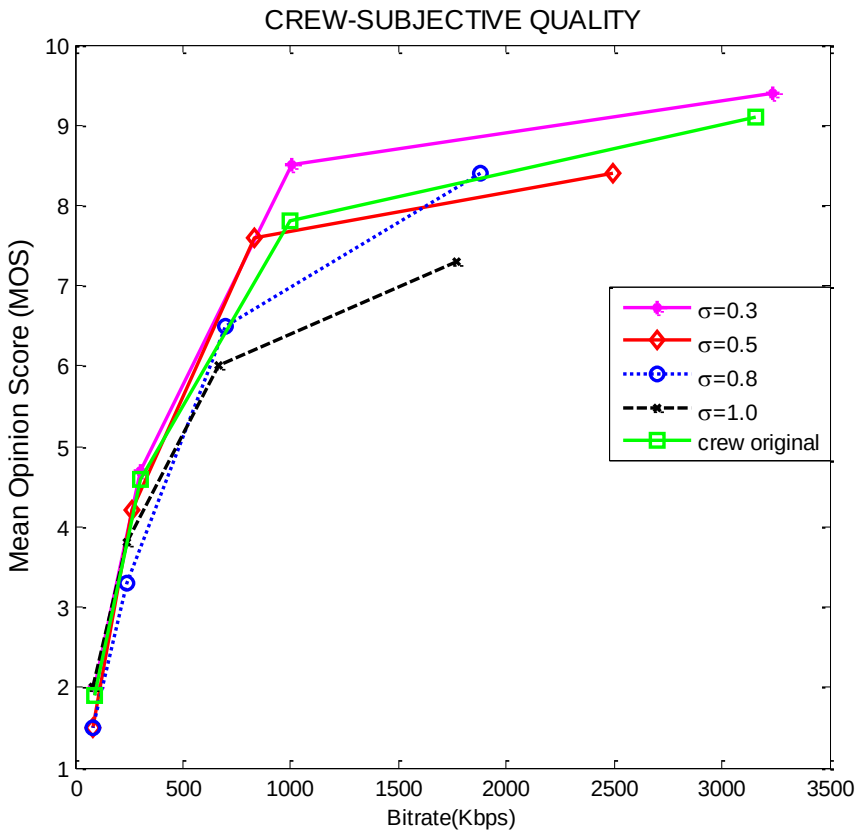


Figure 7.5: Crew video quality evaluation using ACR

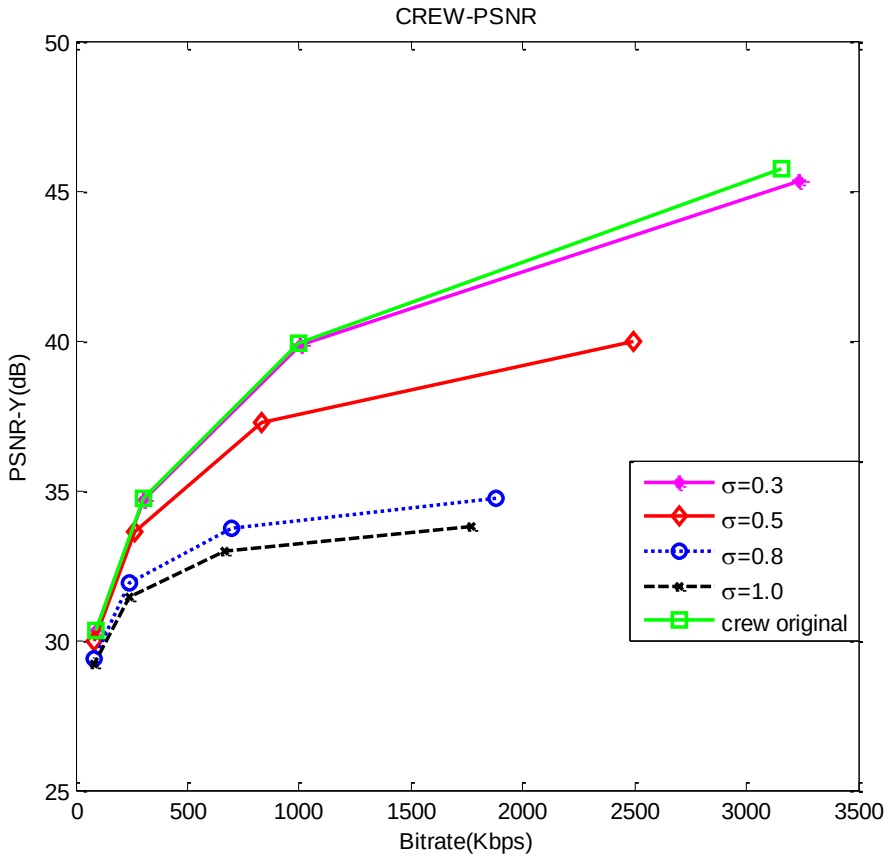


Figure 7.6: Crew video quality evaluation using PSNR

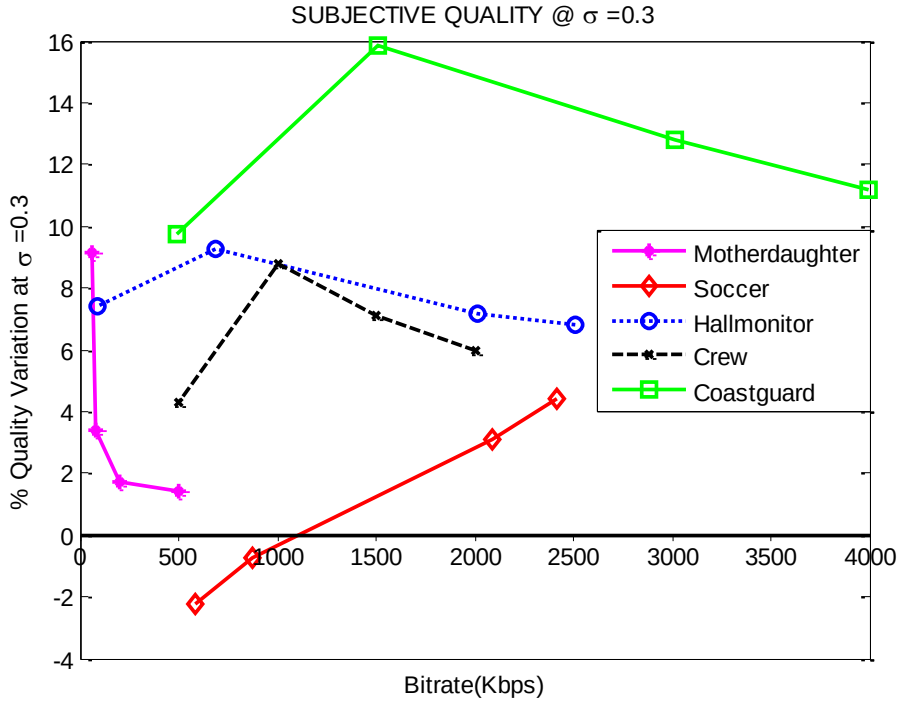


Figure 7.7: All sequences percentage subjective gain/loss at σ=0.3

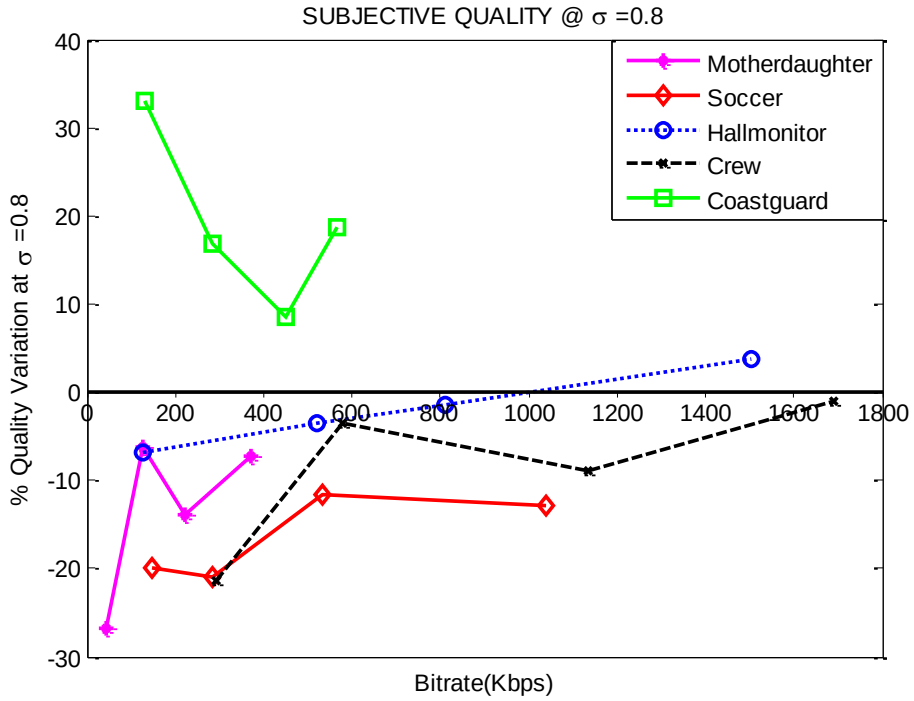


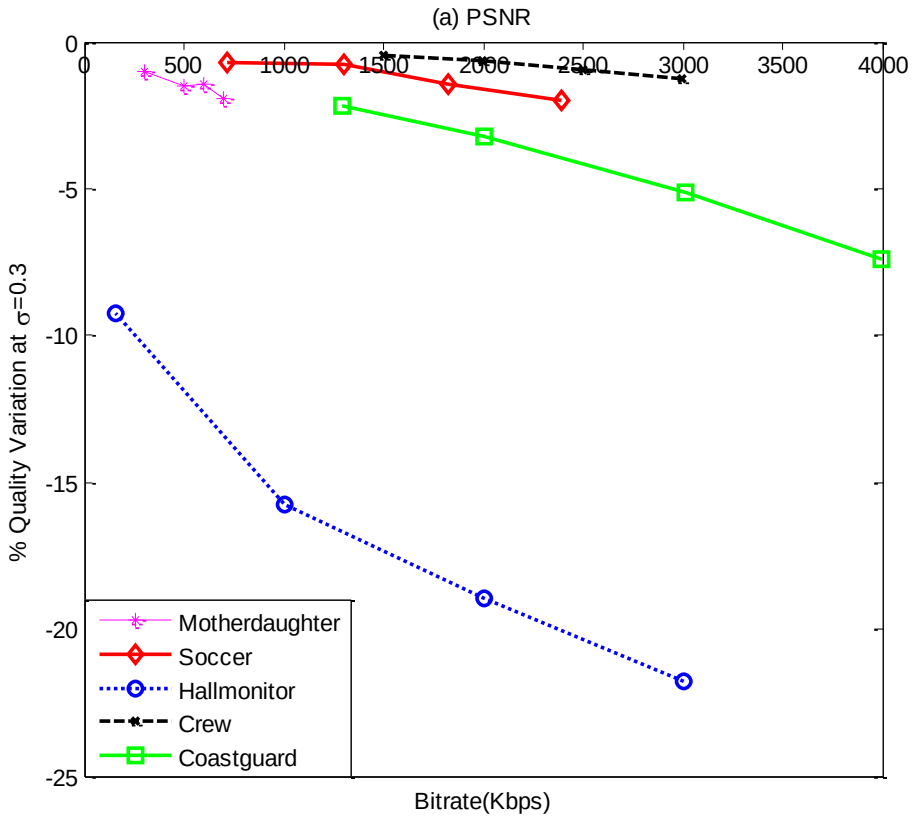
Figure 7.8: All sequences percentage subjective gain/loss at σ =0.8

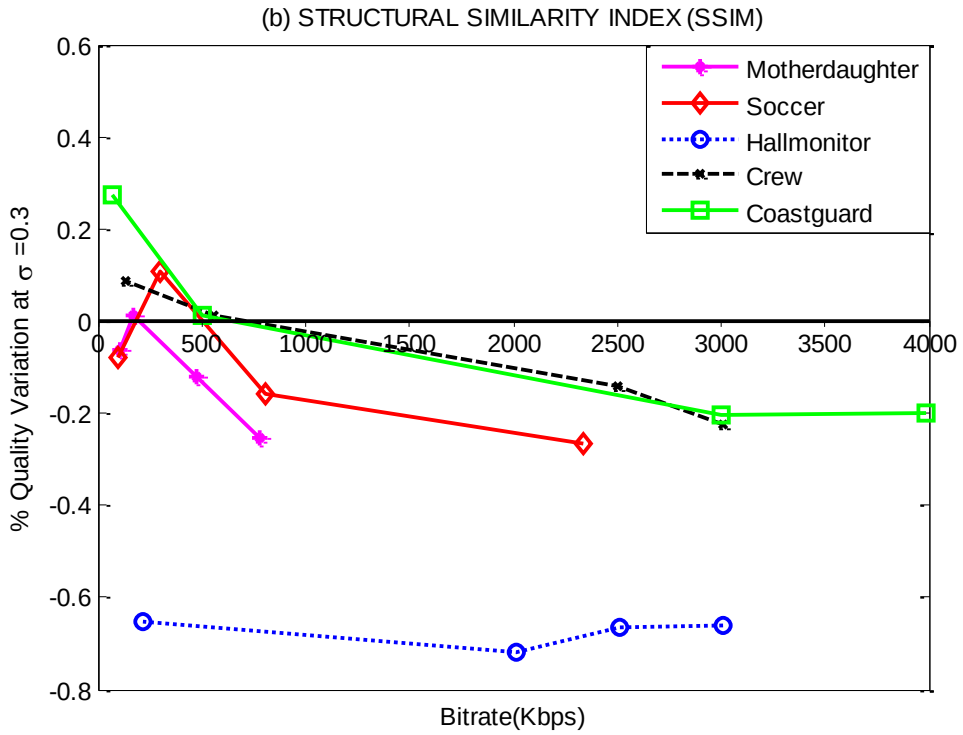
Therefore, PSNR does not show the gain in subjective quality that was observed at σ = 0.3. This can be attributed to the induced variation in pixel values by the Gaussian filter (i.e. higher filter strength leads to lower PSNR). In Figure 7.7 and Figure 7.8 the percentage subjective gain/loss is plotted against bitrate at σ = 0.3 and 0.8 respectively for all tested video sequences. The percentage of quality gain/loss of filtered sequence is computed with respect to the original video sequence. In Figure 7.7 the area above the horizontal bitrate axis indicates quality gain and that below is the percentage of quality loss.

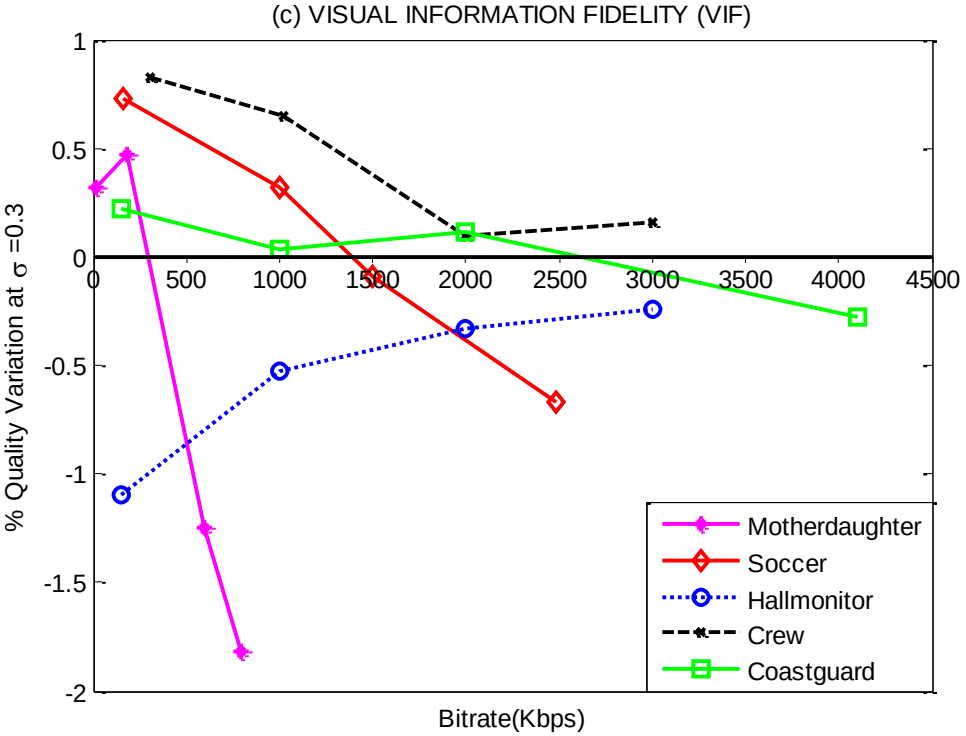
Almost all the sequences, as shown in Figure 7.7 achieve higher perceptual quality when pre-processed with standard deviation equal to 0.3 when compared with the original video sequence (with an exception for soccer sequence at lower bitrates). Whereas in Figure 7.8, the graph shows rate-perceptual loss for all sequences at σ = 0.8 except for coastguard sequence. These subjective results from all the video sequences at σ = 0.3 and 0.8 serve as the benchmark for comparison with the chosen metrics to determine the metric that best correlates with subjective perception. Note that the actual percentage values are not directly comparable between different metrics due to their unique non-linear algorithms.

The percentage PSNR variation at σ = 0.3 for all the sequences is plotted in Figure 7.9 (a). It is clear that in all the tested video sequences PSNR shows a rate

perceptual loss. A similar behaviour is also observed in SSIM as shown in the Figure 7.9 (b). Full reference metric VIF in the Figure 7.9 (c) shows partial detection performance for the video sequences Soccer, Coastguard and Crew at σ = 0.3. Moreover, the metric detected the perceptual gain that is observed at σ = 0.8 in coastguard sequence. The video quality variations detected by all the tested FR and RR metrics are shown in abbreviated format in Table 7.2. The performance of the metrics at two different standard deviations (namely σ = 0.3 and σ = 0.8) is shown in the table. The first row indicates whether subjective results showed a gain or a loss in quality for a specific video sequence. The X or √ for each metric indicates whether that particular metric was able to correctly detect the actual gain or loss. It is evident that all FR and RR metrics can detect a loss in video quality corresponding to σ = 0.8 (except for the coastguard sequence that shows a gain at σ = 0.8). However, these metrics fail to reliably detect quality gains corresponding to σ = 0.3.




(b) STRUCTURAL SIMILARITY INDEX (SSIM)
% Quality Variation at σ =0.3
Bitrate(Kbps)
Motherdaughter
Soccer
Hallmonitor
Crew
Coastguard
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
500
1000
1500
2000
2500
3000
3500
4000


(c) VISUAL INFORMATION FIDELITY (VIF)
% Quality Variation at σ =0.3
Bitrate(Kbps)
Motherdaughter
Soccer
Hallmonitor
Crew
Coastguard
1
0.5
0
-0.5
-1
-1.5
-2
0
500
1000
1500
2000
2500
3000
3500
4000
4500

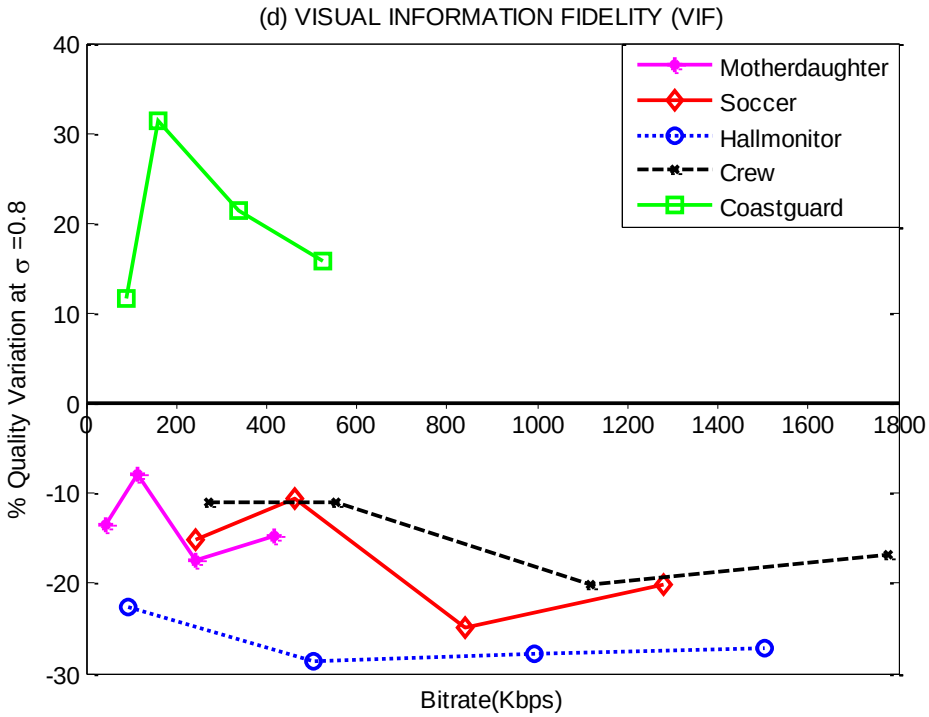


Figure 7.9: Full Reference metrics percentage gain/loss at σ = 0.3 and 0.8

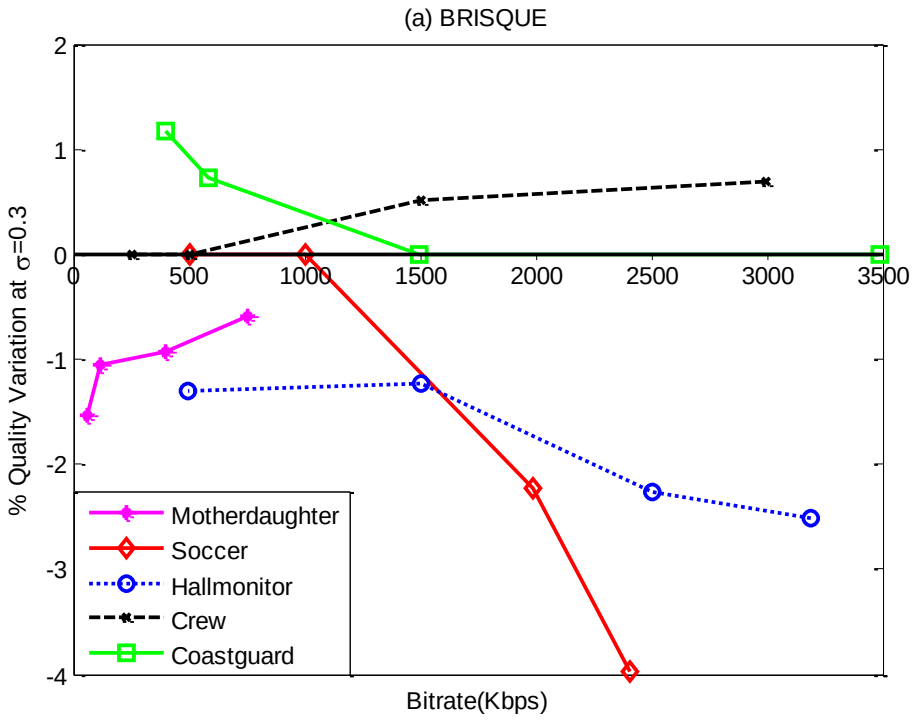
Table 7.2: Full and Reduced Reference metrics

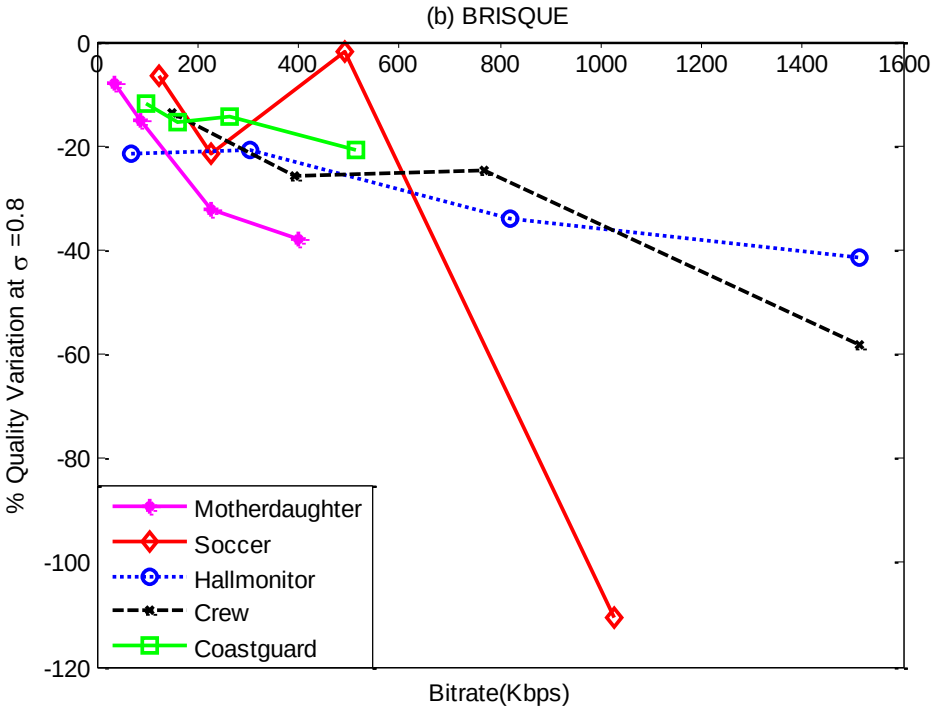
Quality Metric	Standard Deviation	Video Sequences				
		Mother and	Soccer	Hall monitor	Crew	Coastguard
Subjective Quality	0.3	gain	gain	gain	gain	gain
	0.8	loss	Loss	loss	loss	gain
PSNR	0.3	X	X	X	X	X
	0.8	✓	✓	✓	✓	X
UQI	0.3	X	X	X	✓	Partial
	0.8	✓	✓	✓	✓	X
SSIM	0.3	X	X	X	X	X
	0.8	✓	✓	✓	✓	X
MS-SSIM	0.3	X	X	X	X	X
	0.8	✓	✓	✓	✓	X
VQM	0.3	X	X	X	X	X
	0.8	✓	✓	✓	✓	X
IFC	0.3	X	Partial	X	✓	Partial
	0.8	✓	✓	✓	✓	✓
VIF	0.3	X	Partial	X	✓	Partial
	0.8	✓	✓	✓	✓	✓
NQM	0.3	X	X	X	X	X
	0.8	✓	✓	✓	✓	X
VSNR	0.3	X	✓	Partial	✓	Partial
	0.8	✓	✓	✓	✓	X
RRED	0.3	X	X	X	Partial	X
	0.8	✓	✓	✓	✓	X

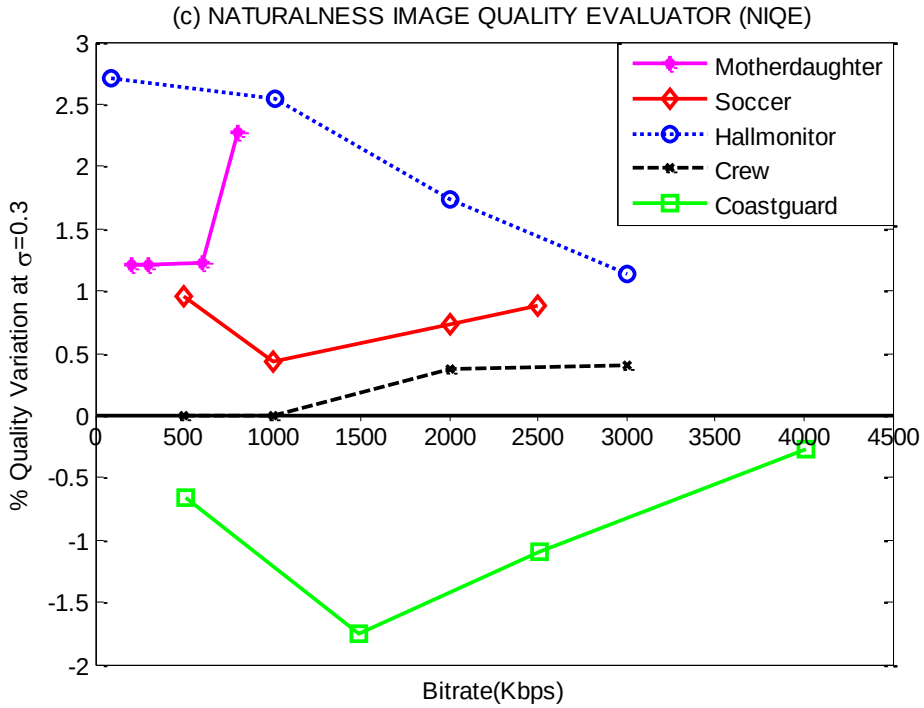
A few exceptions to this are UQI, Information Fidelity Criterion (IFC), VIF and VSNR which show partial detection performance at some bitrates and clear detection in one or more sequences. This behaviour could be attributed to the fundamental assumption behind FR and RR metrics, which is that the reference image is the ideal representation and any difference in image under test is considered as a deviation from ideal quality. Therefore, it can be understood that pixel differences induced by pre-processing algorithms are typically interpreted as perceptual loss in video quality by FR and RR quality metrics.

In contrast to the FR metrics some of the NR metrics have shown better detection ability. It can be seen in Figure 7.10 (a) that NR metric BRISQUE has detected perceptual gain in coastguard and crew video sequence at σ = 0.3, but failed to detect the gain at σ = 0.8 (Figure 7.10 (b)). The video quality variations detected by all the chosen NR metrics for the study are shown in abbreviated form in Table 7.3. No Reference metric BIQI showed partial gains for Soccer, Mother and Daughter, and a clear gain in Coastguard video sequence. However, BIQI failed to detect the perceptual gain in Hall monitor, crew and the perceptual loss in Hall monitor and mother daughter at σ = 0.8.

In the current study blur based metrics are also chosen to determine their effectiveness. The blur based metrics (such as JNBM, CPBD, HP and Marichal metric) despite quantifying the quality using sensitivity of HVS towards sharpness and blur in the image, show only partial detection ability in detecting quality variations.




(b) BRISQUE
% Quality Variation at σ =0.8
Bitrate(Kbps)
0
-20
-40
-60
-80
-100
-120
0
200
400
600
800
1000
1200
1400
1600
Motherdaughter
Soccer
Hallmonitor
Crew
Coastguard


(c) NATURALNESS IMAGE QUALITY EVALUATOR (NIQE)
% Quality Variation at σ=0.3
Bitrate(Kbps)
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Motherdaughter
Soccer
Hallmonitor
Crew
Coastguard

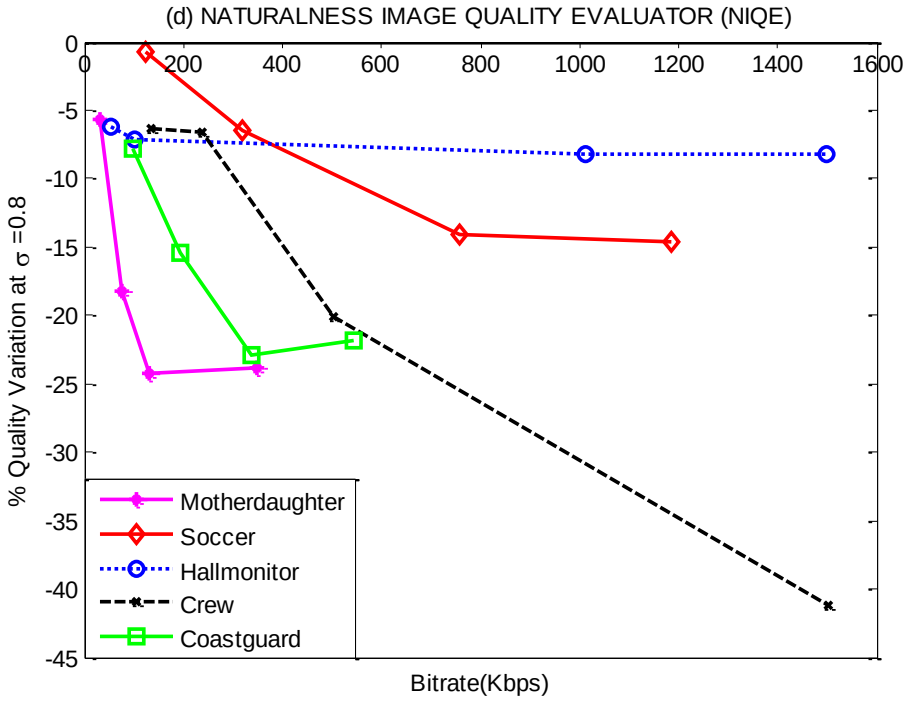


Figure 7.10: No Reference metrics percentage gain/loss at σ = 0.3 and 0.8

Table 7.3: No Reference (NR) metrics

Quality Metric	Standard Deviation	Video Sequences				
		Mother and Daughter	Soccer	Hall monitor	Crew	Coastguard
Subjective Quality (MOS)	0.3	gain	gain	gain	gain	gain
	0.8	loss	Loss	loss	loss	gain
NR JPEG	0.3	X	X	Partial	X	X
	0.8	✓	✓	✓	✓	X
BRISQUE	0.3	X	X	X	✓	✓
	0.8	✓	✓	✓	✓	X
BIQI	0.3	Partial	Partial	X	X	✓
	0.8	X	✓	X	✓	✓
JNBM	0.3	Partial	X	X	X	X
	0.8	✓	✓	✓	✓	X
CPBD	0.3	Partial	Partial	X	Partial	X
	0.8	✓	✓	✓	✓	X
Marichal	0.3	Partial	Partial	Partial	X	X
	0.8	X	X	X	X	✓
NR for JPEG 2000	0.3	X	X	X	X	X
	0.8	✓	✓	✓	✓	X
HP	0.3	X	X	Partial	X	X
	0.8	✓	✓	✓	✓	X
NIQE	**0.3**	✓	✓	✓	✓	**X**
	0.8	✓	✓	✓	✓	**X**

The NR metric NIQE detects the perceptual gain and loss that is shown by subjective results in four out of five tested video sequences except for coastguard video (depicted in Figure 7.10 (c) and Figure 7.10 (d)). Therefore, NIQE shows better detection ability compared to all other objective video quality measurement metrics.

The published correlation values of a number of metrics for Gaussian blurred images were presented earlier in Table 7.1. From the table it can be seen that NIQE has low monotonicity and prediction accuracy according to the SROCC and LCC scores. However, NIQE outperforms other metrics that have higher SROCC and LCC scores in detecting video quality variations. Although metrics such as VIF, SSIM and MS-SSIM show higher correlation scores, they perform poorly in detecting video quality variations. Therefore, SROCC and LCC are not fully reliable in indicating the ability of a quality metric to detect video quality variations.

7.7 Discussion

The objective of this work is to investigate the existing video quality metrics such as Full Reference (FR), Reduced Reference (RR) and No Reference (NR) quality metrics to determine their ability in detecting perceptual quality variations induced by the pre-processing filter. Nineteen state-of-the-art metrics have been investigated with five different video sequences pre-processed and coded at various filter intensities. The videos are pre-processed using a Gaussian low pass filter to simulate the pre-processing used in many perceptual quality algorithms and these filtered videos have been encoded using HEVC video CODEC.

The results clearly show that No Reference (NR) metrics are more effective when compared to Full Reference (FR) and Reduced Reference (RR) video quality metrics. Further, among the No Reference (NR) metrics investigated the Naturalness Image Quality Evaluator (NIQE) has been shown to be better at evaluating the quality of pre-processed videos.

In the literature the Spearman Rank Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (LCC) have been widely used for evaluating the performance of many objective video quality metrics. However, in this work it has been shown that they are ineffective at determining the effectiveness of the quality metric in detecting variations in the perceived quality, particularly when the metrics are used during the development of perceptual video quality optimisation algorithms. The investigation has identified that NIQE shows better performance at detecting perceptual quality variations when videos are pre-processed. NIQE uses a collection of features on a database of undistorted images based on the Natural Scene Statistics (NSS)

model as the reference for quantifying perceived quality. The NSS model of this metric has to be further improved for detecting quality variations more effectively.

The study can be further improved by evaluating the performance of the video quality metrics against spatially-adaptive Gaussian filtering techniques. The current work relied on spatially invariant Gaussian filtering techniques for assessing the performance of metrics. Moreover, the prediction performance is only compared on images with perceptual quality better than the original. The global performance of the video quality metrics has not been analysed.

7.8 Conclusion

In this work, 19 state-of-the-art objective video quality metrics were investigated in order to determine the ability to detect the perceptual variations when videos are pre-processed. This work highlighted the issues related to effective perceptual quality measurement in the development of perceptually optimised compression techniques. Here the main conclusions of our research work are listed.

1) The existing Full Reference (FR) and Reduced Reference (RR) perceptual quality metrics, although having high overall correlation with subjective quality, do not effectively identify the changes in perceptual quality when video frames are pre-processed. This may be because the pixel differences induced by filtering/pre-processing being interpreted as a quality loss by these metrics.

2) No reference (NR) metrics are far more effective compared to FR and RR metrics in detecting the variations. In particular, Naturalness Image Quality Evaluator (NIQE) is notably better at detecting the perceptual gain/loss shown by the subjective evaluations.

3) Moreover, the traditional techniques of evaluating video quality metrics such as Spearman Rank Order Correlation Coefficient (SROCC) and Pearson (Linear) Correlation Coefficient (LCC) are shown to be weak in determining the effectiveness of the quality metric in detecting quality variations, particularly when the quality metric is to be employed during the development of pre-processing based perceptual quality optimisation algorithms.

Furthermore, No Reference (NR) video quality measurement metrics such as NIQE may be further developed to detect small visual quality variations, by extending it with a computational model of visual attention (visual saliency model) [231]. So that it could be employed during the development and evaluation of perceptual video quality algorithms.

PART THREE: CONCLUSIONS

8 Conclusions and Future Directions

Visual attention is an essential component of human vision which selects relevant information from the scene and allows the Human Visual System (HVS) to allocate needed resources for further processing of the attended locations in the scene. This mechanism efficiently solves the trade-off between the visual information entering the human eye and the processing capacity of the human brain.

In computer vision, the visual attention paradigm is of high relevance because the computational complexity is a major issue. Computer vision systems often deal with high resolution images introducing significant amount of overhead during real time operation. To address this problem many computational models of attention have been developed in the literature. These models predict where humans look in the images and thereby reducing the amount of information that has to be processed. The existing visual attention models have either achieved better prediction accuracy with high complexity or low prediction accuracy with faster operation at detecting salient regions in the images. To attain better accuracy they have modelled more number of feature channels present in the images. This resulted in an increase of computational complexity. Moreover they have relied mostly on bottom-up features compared to top-down features as these features are easier to model. These scenarios made it difficult for the existing models to be efficiently employed in computer vision applications. Therefore, the objective of this research work is to develop a novel computational model of visual attention for detecting salient regions in images. The model developed should be computationally fast with better prediction accuracy compared to the state-of-the-art attention models. This is achieved by modelling in-focus regions in the images which are bottom-up in nature and top-down aspects such as image centre and human faces in this thesis. This has allowed an efficient extraction of salient regions in the images.

The research project has been achieved using a preliminary study work and a list of key objectives as described in chapter 1. Each of these objectives has been completed successfully. A brief summary of each of these is given below.

Objective 1: Study the state-of-the-art visual attention models available in the literature. Further, critically analyse them and empirically evaluate their performance.

During the preliminary study work for this research project a comprehensive literature review of the existing state-of-the-art visual attention models was carried out to obtain the relevant theoretical knowledge of what is already known in the field. The ideas behind these models were critically analysed by identifying their advantages, disadvantages and interesting aspects. This background knowledge and the critical review of the attention models have been presented in the chapter 2 and chapter 3 of this thesis. The image datasets, qualitative and quantitative assessment techniques used in the literature to validate the saliency models have been identified. The characteristics of the datasets and assessment techniques used in this work have been described in chapter 4. These attention models, datasets and assessment techniques were obtained using different data collection strategies mentioned in chapter 4. Ten state-of-the-art attention models have been evaluated using qualitative and quantitative research methods to analyse their performance in terms of prediction accuracy and computational complexity. It has been observed that GBVS has achieved highest prediction accuracy across two image datasets.

Objective 2: Development of a novel bottom-up computational model of attention based on in-focus regions.

During the first objective of this project a novel bottom-up visual attention model based on in-focus regions present in images has been developed. The in-focus regions have been detected using the peak frequencies present in the DCT domain. The performance results of the proposed model have been presented in chapter 5 of this thesis. Qualitatively it has been shown that the in-focus detection algorithm has very good ability to differentiate the in-focus and out-focus regions. The attention model performance has been evaluated on a dataset of 1003 images. The results indicated that the proposed model has achieved similar or better prediction accuracy (CC=0.18, NSS= 0.84 and AUC=0.71) compared to the state-of-the-art visual attention models. The model takes an average time of 0.80 seconds for generating a saliency map of an image with 1024x768 resolution. The WBSD [127] model achieves similar prediction accuracy as the proposed model; however, it requires 24.28 seconds with respect to the same image resolution and testing platform. This indicates that the proposed model has good ability to detect salient regions with a lower computational complexity.

Objective 3: Manage the computational complexity of bottom-up visual attention model developed in the second objective and improve the prediction accuracy using top-down components of human attention.

During the third objective of this project the computational complexity of the bottom-up visual attention model has been further reduced by using the Integer Cosine Transform (ICT) instead of the DCT. The prediction accuracy of the model has been enhanced by choosing HSV colour space instead of YC_bC_r colour space and integrating it with location based top-down component known as centre sensitivity. The parameters of the model developed have been tuned using a hill climbing approach. The results of this attention model have been presented in chapter 6 of this thesis. Its performance has been evaluated on the Judd's image dataset. The model has achieved 1% of prediction accuracy (with respect to CC metric) improvement with 62% of computational complexity savings when compared to the GBVS model. GBVS is the best model among the chosen benchmark state-of-the-art saliency models. Furthermore, the literature related to face sensitivity has been studied and it has been found that human faces highly attract the viewer's gaze. The Viola Jones face detection algorithm was used to detect human faces present in the images. The human face maps have been developed by initially drawing square shaped bounding boxes around the human faces and then placing Gaussian blobs in the regions where human faces are located. Similar to the model developed in second objective, the parameters of the model have been tuned using a hill climbing approach to improve the overall prediction accuracy. When compared to the GBVS model which is the best among the benchmark state-of-the-art visual attention models, the proposed model has achieved 3% and 1% of prediction accuracy (with respect to CC metric) improvement on Judd's and DUT-OMRON datasets respectively. Further 42% of computational complexity savings has been achieved when compared to the GBVS model. This work achieves the main objective of this research project, which is to develop novel computational model of visual attention with high prediction accuracy and low computational complexity.

In addition to achieving the main objective, an investigation of the existing video quality metrics for detecting quality variations in pre-processed video was carried out in chapter 7. The study has identified suitable metrics that can be utilised during the development of perceptual video quality optimisation algorithms.

The thesis fulfils the aims and objectives of this research project which is to propose a novel computational model of visual attention for detecting salient regions in images. The proposed model should have the ability to extract salient information with higher prediction accuracy and low computational complexity when compared to the state-of-the-art attention models. It proposed new solutions for predicting salient regions in the images. The main contributions made to the body of knowledge in computational modelling of visual attention can be summarised as:

- Development of a novel DCT based bottom-up in-focus visual attention model. To develop the model a self-dataset of images with different regions in-focus and out-of-focus has been created. During the development stage these images have been used for hypothesis generation and for testing the focus detection performance. The proposed model detected salient regions using the peak frequencies present in the DCT domain with a lower computational complexity. The main novelty of this model lies in detecting in-focus regions using the peak frequencies present in the DCT domain.

- Development of a low complexity visual attention model for detecting salient regions in the images. The computational complexity of the DCT based attention model has been further reduced by using the Integer Cosine Transform (ICT) instead of the DCT. The prediction accuracy is enhanced using the value channel of HSV colour space and location based top down component known as centre sensitivity. The novelty of this work compared to the earlier developed model is in a) Using the Integer transform instead of the traditional DCT for detecting peak frequencies. b) Modelling the centre map by placing the Gaussian blob at the image centre as a function of image resolution and c) Tuning the parameters using a hill climbing approach.

- Development of high prediction accuracy model by integrating learning based top down feature known as human face sensitivity into the earlier model with focus and centre sensitivity. Further, the parameters of the model have been tuned using a hill climbing approach to optimise the overall prediction accuracy of the model. The novelty of this work is in modelling a face map using 2D square shaped Gaussian distribution as it makes the centre of the face more salient compared to the periphery. Further, the parameters of the model have been tuned using the hill climbing approach to optimise the overall prediction accuracy.

- Investigation of the existing video quality metrics to determine their effectiveness in detecting perceptual quality variations when videos are pre-processed. This investigation has identified that No Reference (NR) metrics are far more effective than Full Reference (FR) and Reduced Reference (RR) metrics in detecting quality variations. Furthermore, the NR metric NIQE is the most effective one among the NR metrics that have been tested.

The proposed models are based on firm theoretical foundations and do not depend on empirically obtained thresholds. The hypothesis of detecting salient regions based on in-focus regions, image centre and human faces has been formulated using formal logic, inductive, deductive and analogical inferences. The formulated hypothesis has been corroborated by testing it with an improved methodology when compared to the earlier models. The model has been tested with large datasets which have very high number of ground truth images with complex and varied image statistics. Moreover, it has been analysed empirically using methodological triangulation in which blended qualitative and quantitative approaches using multiple mathematical metrics have been utilised. They have the ability to withstand the test of time and can be subjected to constant testing by other researchers, modification and even refutation as new ideas, datasets and metrics with different viewpoints emerge in the literature. Further, the proposed models have the predictive capabilities that can guide future investigation.

Novel contributions of this work may be used in applications such as object recognition, image segmentation, perceptual video coding and video quality assessment. Specifically during the development of perceptual video coding algorithms, a higher accuracy attention model can detect the salient regions in the video frames more effectively and thereby the non-salient/irrelevant information can be compressed more efficiently. Further as the proposed attention model is computationally fast at detecting the salient regions it can be used in videos with different formats (resolutions) and frame rates. Based on the major issues covered in this thesis, it can be concluded that this thesis work has given rise to a computational model of visual attention that can operate with low complexity and better prediction accuracy with applicability to computer vision tasks.

8.1 Future Directions

The future directions related to this research work and the general directions for the saliency research are presented in the following sections.

8.1.1 Future Directions Related To The Proposed Model

This section presents the future directions mainly aimed at addressing the disadvantages of the developed models and improving them to achieve better performance in terms of computational complexity and prediction accuracy. They are

1. The focus detection algorithm developed to detect in-focus regions initially generates a sparse focus map. Gaussian blurring is later used to connect these

regions. This process makes region edges denser making them more salient compared to the centre of the in-focus regions. The human fixations are mostly clustered near the centre of any salient region. The fixation density is reduced towards the edges. This indicates that the centre of the salient regions have to be given more priority during the development of a saliency map. In contrast to this the proposed focus detection algorithm gives relatively higher priority to the edges. This reduces the prediction accuracy of the model. Further, when the edges are blurred they fade into the out-of-focus area and tend to become visually salient. Therefore, better approaches have to be developed to connect these sparse salient regions giving higher priority to the centre of the in-focus regions to improve the prediction accuracy of the model.

2. The model considers the centre of the image as highly salient compared to the periphery of the image. The centre of the image is modelled as a function of image resolution. In a real-time scenario, although the majority of the fixations are clustered more at the centre, they are not consistent across stimuli with different image statistics. For example, when human faces are present in an image, the viewer gaze is mostly oriented towards the face and the image centre is highly ignored. Even in such a scenario the proposed models consider both human faces and centre as salient and will result in false detections. Therefore, to improve the prediction accuracy the centre bias has to be modelled dynamically depending on the image statistics.

3. The models developed in this thesis give more priority to the image centre in detecting the salient regions of an image. They have to be further improved to detect peripheral salient regions. Detecting these regions will help in improving the prediction accuracy of the model. The nature of these regions has to be investigated and the key characteristics of these regions have to be modelled as a visual saliency map.

4. The visual attention model which has been developed considers human faces as visually salient. Human faces are detected using a face detection algorithm. The faces are mapped using square shaped bounding boxes and 2D Gaussian blobs. However, the influence of other shapes (such as oval, ellipse and circular enclosing of human faces) on prediction accuracy has to be evaluated. Mapping using Gaussian blobs indicates that the centre of the face is given higher

priority when compared to the periphery. Although this is a better approximation based on Ground truth, it ignores the sensitivity of humans towards different parts of human faces. For example, viewers look at human eyes more often compared to nose and mouth. Furthermore, the impact of image resolution on human face fixations and sensitivity towards bigger and smaller human faces should be considered during the development of the face saliency map.

5. In an image with many human faces there will be only few faces which are visually salient. As the current research has shown that regions in-focus are more salient. A possible direction is to study the gaze sensitivity towards in-focus and out-of-focus human faces in the crowd scene and develop face map as a function of focus strength across different faces. This helps in developing a face map that has the ability to distinguish between salient and non-salient human faces.

6. The Viola Jones face detection used in the attention model which has been developed has the ability to detect only frontal faces. A non-frontal face detection algorithm can be incorporated to detect faces that are turned sideways. Further, human fixations patterns towards non-frontal faces have to be studied using an eye tracker and a novel attention map has to be developed. The developed non-frontal face map has to be integrated into the attention model to improve the overall prediction accuracy.

7. The proposed models use weighted addition for focus and centre. The face map is simply added to preserve its importance. Majority of the models have used linear summation as it has psychophysical support and simplicity in application. Better ways have to be investigated to combine the feature maps for increasing the prediction accuracy of the attention models.

8.1.2 General Directions For Visual Saliency Research

This section suggests some of the general directions needed for future saliency modelling. They are

1. Text can be as effective as human faces in attracting gaze. A text detection algorithm can be used to detect text present in images. Novel techniques have to be developed for generating text maps. For example the letter edges, text

size and font are some of the key aspects which are to be considered during the development of text saliency maps.

2. A general direction for the future is to model overt versus covert attention. All the models developed in the literature (including the proposed model) consider that viewers pay attention to whatever they look at in the images. However, in reality, although the human eyes fixate (overt attention) on some regions, they might not pay attention. This is due to the covert attention interested in some other regions in the images. For example, when a person is driving a car his overt attention may be on the road but his covert attention will continuously monitor traffic lights.

3. To develop real time visual attention models, the models should be made faster by reducing the computational complexity, improve the prediction accuracy, make them robust to noise, illumination changes and image transformations.

4. The proposed models can be further developed for detecting salient regions in videos. Videos consist of many in-focus, out-of-focus regions and temporal characteristics such as motion etc. that needs to be computationally modelled. The proposed models can be further developed by considering these temporal characteristics and evaluating their performance on video saliency datasets.

5. The influence of culture, gender, age on human fixations is still an open research question.

6. Connectionist approaches with deep learning architectures such as convolution neural networks, deep Boltzmann machine, ensemble learning, stacked auto encoders and deep belief networks should be investigated for developing efficient visual attention models.

References

[1] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. on Applied Perception,* vol. 7, pp. 1-39, 2010.

[2] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," *Proc. IEEE Int'l Conf. on Image Processing,* pp. 253-256, 2003.

[3] A. Torralba, "Modeling global scene factors in attention," *JOSA A,* vol. 20, pp. 1407-1418, 2003.

[4] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?," *Vision Research,* vol. 41, pp. 3559-3565, 2001.

[5] J. N. Bailenson and N. Yee, "Digital chameleons automatic assimilation of nonverbal gestures in immersive virtual environments," *Psychological Science,* vol. 16, pp. 814-819, 2005.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 20, pp. 1254-1259, 1998.

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

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

[9] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 1915-1926, 2012.

[10] G. Chenlei and Z. Liming, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," *IEEE Trans. on Image Processing,* vol. 19, pp. 185-198, 2010.

[11] H. Xiaodi, J. Harel, and C. Koch, "Image Signature: Highlighting Sparse Salient Regions," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 194-201, 2012.

[12] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *IEEE Int'l Conf. on Computer Vision,* pp. 2106-2113, 2009.

[13] Z. Lin, G. Zhongyi, and L. Hongyu, "SDSP: A novel saliency detection method by combining simple priors," *Proc. IEEE Int'l Conf. on Image Process. ,* pp. 171-175, 2013.

[14] J. Chilukamari, S. Kannangara, and G. Maxwell, "A DCT based in-focus visual saliency detection algorithm," *IEEE Int'l Conf. on Consumer Electronics Berlin,* pp. 1-5, 2013.

[15] J. Chilukamari, S. Kannangara, and G. Maxwell, "A low complexity visual saliency model based on in-focus regions and centre sensitivity," *IEEE Int'l Conf. on Consumer Electronics Berlin,* pp. 411-414, 2014.

[16] J. Chilukamari, S. Kannangara, and G. Maxwell, "Investigation of the effectiveness of video quality metrics in video pre-processing," *IEEE Int'l Conf. on Consumer Electronics Berlin,* pp. 1-5, 2013.

[17] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry II, V. Balasubramanian, and P. Sterling, "How Much the Eye Tells the Brain," *Current Biology,* vol. 16, pp. 1428-1434, 2006.

[18] L. Itti, "Models of bottom-up and top-down visual attention," Ph.D thesis, California Inst. of Technology, 2000.

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

[20] R. Jain, R. Kasturi, and B. G. Schunck, *Machine vision* vol. 5: McGraw-Hill New York, 1995.

[21] R. A. Rensink, J. K. O'Regan, and J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science,* vol. 8, pp. 368-373, 1997.

[22] Aristotle. On Sense and the Sensible. The Internet Classics Archive, 350 B.C.E, translated by J.I. Bearfe.

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

[24] Gescheider. G, *Psychophysics: the fundamentals*, 3rd ed. Lawrence Erlbaum Associates, 1997.

[25] G. P. R. Bruce. V, Georgeson M.A, *Visual perception*, 3rd ed. Psychology Press, 1996.

[26] Hugh J.Foley and Margaret W.Matlin. *Sensation and Perception*. Available: http://www.skidmore.edu/~hfoley/Perc3.htm

[27] R. Desimone, M. Wessinger, L. Thomas, and W. Schneider, "Attentional control of visual perception: cortical and subcortical mechanisms," *Cold Spring Harbor symposia on quantitative biology,* pp. 963-971, 1990.

[28] N. Ouerhani, "Visual attention: from bio-inspired modeling to real-time implementation," Ph. D. thesis,Université de Neuchâtel, 2004.

[29] D. R. Bull, "Chapter 2 - The Human Visual System," *Communicating Pictures,* pp. 17-61, 2014.

[30] O. J. Braddick, J. M. O'Brien, J. Wattam-Bell, J. Atkinson, T. Hartley, and R. Turner, "Brain areas sensitive to coherent visual motion," *Perception-London,* vol. 30, pp. 61-72, 2001.

[31] R. T. Born and D. C. Bradley, "Structure and function of visual area MT," *Annu. Rev. Neuroscience,* vol. 28, pp. 157-189, 2005.

[32] A. Yarbus, *Eye movements and vision*. New York: Plenum, 1967.

[33] M. Land and B. Tatler, *Looking and acting: vision and eye movements in natural behaviour*: Oxford University Press, 2009.

[34] R. J. Jacob, "Eye tracking in advanced interface design," *Virtual environments and advanced interface design,* pp. 258-288, 1995.

[35] L. W. Stark and S. R. Ellis, "Scanpaths revisited: cognitive models direct active looking," *Eye movements: cognition and visual perception,* pp. 193-226, 1981.

[36] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 35, pp. 185-207, 2013.

[37] M. I. Posner, "Orienting of attention," *Quarterly J. of experimental psychology,* vol. 32, pp. 3-25, 1980.

[38] B. Fischer and H. Weber, "Express saccades and visual attention," *Behavioral and Brain Sciences,* vol. 16, pp. 553-567, 1993.

[39] D. Brockmann and T. Geisel, "Are human scanpaths Levy flights?," *Int'l Conf. on Artificial Neural Networks,* pp. 263-268, 1999.

[40] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Ann. Rev. of Neuroscience,* vol. 18, pp. 193-222, 1995.

[41] H. E. Egeth and S. Yantis, "Visual attention: Control, representation, and time course," *Ann. Rev. of Psychology,* vol. 48, pp. 269-297, 1997.

[42] Sabra A.I, *The Optics of Ibn Al-Haytham*. The Warburg Institute, University of London.

[43] J. M. Wolfe, "Visual search," *Attention,* vol. 1, pp. 13-73, 1998.

[44] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Rev. Neuroscience,* vol. 3, pp. 215-229, 2002.

[45] J. Jonides, "Voluntary versus automatic control over the mind's eye's movement," *Attention and performance IX,* vol. 9, pp. 187-203, 1981.

[46] A. Borji, D. N. Sihite, and L. Itti, "Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study," *IEEE Trans. on Image Processing,* vol. 22, pp. 55-69, 2013.

[47] A. Wells and G. Matthews, *Attention and emotion: A clinical perspective*: Psychology Press, 1994.

[48] N. Fragopanagos and J. G. Taylor, "Modelling the interaction of attention and emotion," *Neurocomputing,* vol. 69, pp. 1977-1983, 2006.

[49] Neil D. B. Bruce, "Saliency, attention and visual search: An information theoretic approach," Ph. D. thesis, Graduate Program in Computer science and Engineering, York university, Toronto, Ontario, 2008.

[50] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Rev. Neuroscience,* vol. 3, pp. 201-215, 2002.

[51] J. Theeuwes, "Top-down search strategies cannot override attentional capture," *Psychonomic Bull. & Rev.,* vol. 11, pp. 65-70, 2004.

[52] A. B. Leber and H. E. Egeth, "It's under control: Top-down search strategies can override attentional capture," *Psychonomic Bull. & Rev.,* vol. 13, pp. 132-138, 2006.

[53] M. A. Schoenfeld and C. M. Stoppel, "Chapter 9 - Feature- and Object-Based Attention: Electrophysiological and Hemodynamic Correlates " *Cognitive Electrophysiology of Attention,* pp. 107-122, 2014.

[54] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology,* vol. 12, pp. 97-136, 1980.

[55] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic Bull. & Rev.,* vol. 1, pp. 202-238, 1994.

[56] S. Yantis, J. Schwarzbach, J. T. Serences, R. L. Carlson, M. A. Steinmetz, J. J. Pekar, and S. M. Courtney, "Transient neural activity in human parietal cortex during spatial attention shifts," *Nature Neuroscience,* vol. 5, pp. 995-1002, 2002.

[57] J. W. Bisley and M. E. Goldberg, "Neuronal activity in the lateral intraparietal area and spatial attention," *Science,* vol. 299, pp. 81-86, 2003.

[58] B. Giesbrecht, M. Woldorff, A. Song, and G. Mangun, "Neural mechanisms of top-down control during spatial and feature attention," *Neuroimage,* vol. 19, pp. 496-512, 2003.

[59] S. Shomstein and S. Yantis, "Control of attention shifts between vision and audition in human cortex," *J. of Neuroscience,* vol. 24, pp. 10702-10706, 2004.

[60] O. Ben-Shahar, B. J. Scholl, and S. W. Zucker, "Attention, segregation, and textons: Bridging the gap between object-based attention and texton-based segregation," *Vision Research,* vol. 47, pp. 845-860, 2007.

[61] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. of Vision,* vol. 8, p. 18, 2008.

[62] S. P. Vecera and M. J. Farah, "Does visual attention select objects or locations?," *J. of Experimental Psychology,* vol. 123, p. 146, 1994.

[63] S. Yantis and J. T. Serences, "Cortical mechanisms of space-based and object-based attentional control," *Current opinion in Neurobiology,* vol. 13, pp. 187-193, 2003.

[64] Z. W. Pylyshyn, *Seeing and visualizing: It's not what you think*: MIT Press, 2003.

[65] E. Awh and H. Pashler, "Evidence for split attentional foci," *J. of Experimental Psychology,* vol. 26, pp. 834-846, 2000.

[66] S. A. McMains and D. C. Somers, "Multiple spotlights of attentional selection in human visual cortex," *Neuron,* vol. 42, pp. 677-686, 2004.

[67] L. Nian, J. Han, D. Zhang, W. Shifeng, and T. Liu, "Predicting eye fixations using convolutional neural networks," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 362-370, 2015.

[68] J. Li, L. Y. Duan, X. Chen, T. Huang, and Y. Tian, "Finding the Secret of Image Saliency in the Frequency Domain," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 37, pp. 2428-2440, 2015.

[69] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual Saliency Based on Scale-Space Analysis in the Frequency Domain," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 35, pp. 996-1010, 2013.

[70] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. of Vision,* vol. 8, 2008.

[71] T. Judd, "Understanding and predicting where people look in images," Ph. D. thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2011.

[72] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Springer Matters of Intelligence,* pp. 115-141, 1987.

[73] S. Frintrop, "VOCUS: A visual attention system for object detection and goal-directed search," Ph. D. thesis, Rheinische Friedrich-Wilhelms-Universitat Bonn Institut Fur Informatik and Fraunhofer Institut Fur Autonome Intelligente Systeme, 2006.

[74] R. Milanese, "Detecting salient regions in an image: from biological evidence to computer implementation," Ph.D thesis, University of Geneva, Switzerland, 1993.

[75] N. Ouerhani, T. Jost, A. Bur, and H. Hugli, "Cue normalization schemes in saliency-based visual attention models," *Int'l Cognitive Vision Workshop,* 2006.

[76] S. Frintrop, E. Rome, A. Nüchter, and H. Surmann, "A bimodal laser-based attention system," *Computer Vision and Image Understanding,* vol. 100, pp. 124-151, 2005.

[77] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. of Electronic Imaging,* vol. 10, pp. 161-169, 2001.

[78] L. Itti and C. Koch, "Comparison of feature combination strategies for saliency-based visual attention systems," *Electronic Imaging'99,* pp. 473-482, 1999.

[79] ITU-T Recommendation H.263, "Video coding for low bit rate communication," 1998.

[80] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services," May 2003.

[81] ISO/IEC 23008-2 MPEG-H Part 2 and ITU-T Rec. H.265, "High Efficiency Video Coding," April 2013.

[82] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, and F. Pellandini, "Adaptive color image compression based on visual attention," *Proc. Int'l Conf. on Image Analysis and Processing,* pp. 416-421, 2001.

[83] L. Yu-Bei and Z. Xing-Ming, "Recent developments in perceptual video coding," *Int'l Conf. on Wavelet Analysis and Pattern Recognition,* pp. 259-264, 2013.

[84] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating higher level perceptual factors," *IEEE Int'l Conf. on Image Processing,* pp. 414-418, 1998.

[85] R. Barland and A. Saadane, "Blind Quality Metric using a Perceptual Importance Map for JPEG-20000 Compressed Images," *IEEE Int'l Conf. on Image Processing,* pp. 2941-2944, 2006.

[86] F. Miau, C. S. Papageorgiou, and L. Itti, "Neuromorphic algorithms for computer vision and attention," *Int'l Symp. on Optical Science and Technology,* pp. 12-23, 2001.

[87] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks,* vol. 19, pp. 1395-1407, 2006.

[88] A. A. Salah, E. Alpaydin, and L. Akarun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and

face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 420-425, 2002.

[89] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *Neural Information Processing Systems,* pp. 241-248, 2008.

[90] P. Viola and M. J. Jones, "Robust real-time face detection," *Int'l J. of Computer Vision,* vol. 57, pp. 137-154, 2004.

[91] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

[92] N. Ouerhani, N. Archip, H. Hügli, and P. Erard, "A color image segmentation method based on seeded region growing and visual attention," *Image Processing & Communications,* vol. 8, pp. 3-11, 2002.

[93] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," *Proc. of the ACM Int'l Conf. on Multimedia,* pp. 374-381, 2003.

[94] B. C. Ko and J.-Y. Nam, "Object-of-interest image segmentation based on human attention and semantic region clustering," *JOSA A,* vol. 23, pp. 2462-2470, 2006.

[95] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," *Int'l Conf. on computer vision systems,* pp. 66-75, 2008.

[96] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," *European Conf. on Computer Vision,* pp. 366-379, 2010.

[97] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *J. of Neuroscience,* vol. 13, pp. 4700-4719, 1993.

[98] M. C. Mozer, *Early parallel processing in reading: A connectionist approach*: Lawrence Erlbaum Associates, 1987.

[99] R. H. Phaf, A. Van der Heijden, and P. T. Hudson, "SLAM: A connectionist model for attention in visual selection tasks," *Cognitive Psychology,* vol. 22, pp. 273-341, 1990.

[100] G. Li and Y. Yu, "Visual Saliency Detection Based on Multiscale Deep CNN Features," *IEEE Trans. on Image Processing,* vol. 25, pp. 5012-5024, 2016.

[101] S. He, R. W. H. Lau, W. Liu, Z. Huang, and Q. Yang, "SuperCNN: A Superpixelwise Convolutional Neural Network for Salient Object Detection," *Int'l J. of Computer Vision,* vol. 115, pp. 330-344, 2015.

[102] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li, "Two-Stage Learning to Predict Human Eye Fixations via SDAEs," *IEEE Trans. on Cybernetics,* vol. 46, pp. 487-498, 2016.

[103] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 1265-1274, 2015.

[104] L. Wang, H. Lu, X. Ruan, and M. H. Yang, "Deep networks for saliency detection via local estimation and global search," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 3183-3192, 2015.

[105] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence,* vol. 146, pp. 77-123, 2003.

[106] Free image distribution [online]. Available: www.pixcove.com

[107] B. A. Draper and A. Lionelle, "Evaluation of selective attention under similarity transformations," *Computer Vision and Image Understanding,* vol. 100, pp. 152-171, 2005.

[108] F. H. Hamker, "The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision," *Computer Vision and Image Understanding,* vol. 100, pp. 64-106, 2005.

[109] F. Fraundorfer and H. Bischof, "Utilizing saliency operators for image matching," *Proc. of Int'l Workshop on Attention and Performance in Computer Vision,* pp. 17-24, 2003.

[110] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt visual attention for a humanoid robot," *Proc. Int'l Conf. on Intelligent Robots and Systems,* vol. 4, pp. 2332-2337, 2001.

[111] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *SPIE's 48th Ann. Meeting Optical Science and Technology,* pp. 64-78, 2004.

[112] T. Kadir and M. Brady, "Saliency, scale and image description," *Int'l J. of Computer Vision,* vol. 45, pp. 83-105, 2001.

[113] G. Backer, B. Mertsching, and M. Bollmann, "Data-and model-driven gaze control for an active-vision system," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 23, pp. 1415-1429, 2001.

[114] N. D. Bruce and J. K. Tsotsos, "An attentional framework for stereo vision," *Canadian Conf. on Computer and Robot Vision,* pp. 88-95, 2005.

[115] K. Lee, H. Buxton, and J. Feng, "Selective attention for cue-guided search using a spiking neural network," *Proc. of the Int'l Workshop on Attention and Performance in Computer Vision,* pp. 55-62, 2003.

[116] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modelling search for people in 900 scenes: A combined source model of eye guidance," *Visual Cognition,* vol. 17, pp. 945-978, 2009.

[117] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Trans. on Graphics,* vol. 26, p. 4, 2007.

[118] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in Cognitive Sciences,* vol. 9, pp. 188-194, 2005.

[119] J. M. Henderson and A. Hollingworth, "High-level scene perception," *Ann. Rev. of Psychology,* vol. 50, pp. 243-271, 1999.

[120] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition,* vol. 7, pp. 17-42, 2000.

[121] M. Sodhi, B. Reimer, J. Cohen, E. Vastenburg, R. Kaars, and S. Kirschenbaum, "On-road driver eye movement tracking using head-mounted devices," *Proc. of the symp. on Eye tracking research & applications,* pp. 61-68, 2002.

[122] T. N. Vikram, M. Tscherepanow, and B. Wrede, "A saliency map based on sampling an image into random rectangular regions of interest," *Pattern Recognition,* vol. 45, pp. 3114-3124, 2012.

[123] J. Schwiegerling, *Field Guide to Visual and Ophthalmic Optics*: SPIE Press, 2004.

[124] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

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

[126] P. Bian and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency," *Advances in Neuro-Information Processing,* pp. 251-258, 2009.

[127] N. Imamoglu, L. Weisi, and F. Yuming, "A Saliency Detection Model Using Low-Level Features Based on Wavelet Transform," *IEEE Trans. on Multimedia,* vol. 15, pp. 96-105, 2013.

[128] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 73-80, 2010.

[129] A. Borji, M. M. Cheng, H. Jiang, and J. Li, "Salient Object Detection: A Benchmark," *IEEE Trans. on Image Processing,* vol. 24, pp. 5706-5722, 2015.

[130] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 438-445, 2012.

[131] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient Object Detection: A Discriminative Regional Feature Integration Approach," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 2083-2090, 2013.

[132] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The Secrets of Salient Object Segmentation," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 280-287, 2014.

[133] E. Vig, M. Dorr, and D. Cox, "Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 2798-2805, 2014.

[134] C. Shen, M. Song, and Q. Zhao, "Learning high-level concepts by training a deep network on eye fixations," *NIPS Workshop,* 2012.

[135] Y. Lin, S. Kong, D. Wang, and Y. Zhuang, "Saliency detection within a deep convolutional architecture," *Proc. of the AAAI Workshops at the AAAI Conf. on Artificial Intelligence,* pp. 27-31, 2014.

[136] C. Xia, F. Qi, and G. Shi, "Bottom Up Visual Saliency Estimation With Deep Autoencoder-Based Sparse Reconstruction," *IEEE Trans. on Neural Networks and Learning Systems,* vol. 27, pp. 1227-1240, 2016.

[137] S. Winkler and R. Subramanian, "Overview of Eye tracking Datasets," *Int'l Workshop on Quality of Multimedia Experience,* pp. 212-217, 2013.

[138] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Neural Information Processing Systems,* pp. 155-162, 2005.

[139] G. Kootstra and L. R. Schomaker, "Prediction of human eye fixations using symmetry," *Ann. Conf. of the Cognitive Science Society,* pp. 56-61, 2009.

[140] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 802-817, 2006.

[141] Y. Chuan, Z. Lihe, L. Huchuan, R. Xiang, and Y. Ming-Hsuan, "Saliency Detection via Graph-Based Manifold Ranking," *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 3166-3173, 2013.

[142] N. Ouerhani, R. Von Wartburg, H. Hugli, and R. Muri, "Empirical validation of the saliency-based model of visual attention," *Electronic letters on computer vision and image analysis,* vol. 3, pp. 13-24, 2003.

[143] G. Kootstra, B. de Boer, and L. R. Schomaker, "Predicting eye fixations on complex visual stimuli using local symmetry," *Cognitive Computation,* vol. 3, pp. 223-240, 2011.

[144] T. Jost, N. Ouerhani, R. v. Wartburg, R. Müri, and H. Hügli, "Assessing the contribution of color in visual attention," *Computer Vision and Image Understanding,* vol. 100, pp. 107-123, 2005.

[145] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research,* vol. 45, pp. 2397-2416, 2005.

[146] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research,* vol. 42, pp. 107-123, 2002.

[147] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *J. of Vision,* vol. 9, p. 10, 2009.

[148] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," *IEEE Int'l Conf. on Computer Vision,* pp. 921-928, 2013.

[149] S. Y. Lee, S. S. Park, C. S. Kim, Y. Kumar, and S. W. Kim, "Low-power auto focus algorithm using modified DCT for the mobile phones," *Int'l Conf. on Consumer Electronics,* pp. 67-68, 2006.

[150] E. Barth, J. Drewes, and T. Martinetz, "Dynamic predictions of tracked gaze," *Int'l Symp. on Signal Processing and Its Applications,* pp. 245-248, 2003.

[151] P. Ki Tae, P. Min Su, L. Jeong Ho, and M. Young Shik, "Detection of visual saliency in Discrete Cosine Transform domain," *IEEE Int'l Conf. on Consumer Electronics,* pp. 128-129, 2012.

[152] S. Bae and F. Durand, "Defocus magnification," *Computer Graphics Forum,* pp. 571-579, 2007.

[153] S. Zhuo and T. Sim, "Defocus map estimation from a single image," *Pattern Recognition,* vol. 44, pp. 1852-1858, 2011.

[154] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient Region Detection by UFO: Uniqueness, Focusness and Objectness," *IEEE Int'l Conf. on Computer Vision,* pp. 1976-1983, 2013.

[155] Y. Jiang and D. Xu, "A Visual Attention Model Based on DCT Domain," *TENCON Conf.,* pp. 1-5, 2005.

[156] B. Schauerte and R. Stiefelhagen, "Predicting human gaze using quaternion DCT image signature saliency and face detection," *IEEE Workshop on Applications of Computer Vision,* pp. 137-144, 2012.

[157] Y. Fang, W. Lin, Z. Chen, C. Tsai, and C. Lin, "Video saliency detection in the compressed domain," *Proc. ACM Int'l conf. on Multimedia,* pp. 697-700, 2012.

[158] M. Charfi, A. Nyeck, and A. Tosser, "Focusing criterion," *Electronics Letters,* vol. 27, pp. 1233-1235, 1991.

[159] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition," *J. of Vision,* vol. 8, 2008.

[160] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. of Vision,* vol. 7, 2007.

[161] D. J. Parkhurst and E. Niebur, "Scene content selected by active vision," *Spatial Vision,* vol. 16, pp. 125-154, 2003.

[162] F. Schumann, W. Einhäuser, J. Vockeroth, K. Bartl, E. Schneider, and P. König, "Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments," *J. of Vision,* vol. 8, 2008.

[163] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," *Vision Research,* vol. 46, pp. 4333-4345, 2006.

[164] J. H. Fuller, "Eye position and target amplitude effects on human visual saccadic latencies," *Experimental Brain Research,* vol. 109, pp. 457-466, 1996.

[165] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision Research,* vol. 45, pp. 643-659, 2005.

[166] F. Vitu, Z. Kapoula, D. Lancelin, and F. Lavigne, "Eye movements in reading isolated words: Evidence for strong biases towards the center of the screen," *Vision Research,* vol. 44, pp. 321-338, 2004.

[167] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *J. of Vision,* vol. 10, 2010.

[168] P. Tseng, R. Carmi, I. G. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. of Vision,* vol. 9, 2009.

[169] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. of Vision,* vol. 11, 2011.

[170] M. Bindemann, A. M. Burton, I. T. Hooge, R. Jenkins, and E. H. de Haan, "Faces retain attention," *Psychonomic Bull. & Rev.,* vol. 12, pp. 1048-1053, 2005.

[171] T. Ro, C. Russell, and N. Lavie, "Changing faces: A detection advantage in the flicker paradigm," *Psychological Science,* vol. 12, pp. 94-99, 2001.

[172] C. H. Cashon and L. B. Cohen, "The construction, deconstruction, and reconstruction of infant face perception," *The development of face processing in infancy and early childhood: Current perspectives,* pp. 55-68, 2003.

[173] C. K. Friesen and A. Kingstone, "The eyes have it! Reflexive orienting is triggered by nonpredictive gaze," *Psychonomic Bull. & Rev.,* vol. 5, pp. 490-495, 1998.

[174] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," *Proc. ECCV,* pp. 414-429, 2012.

[175] Graph Based Visual Saliency [online]. Available: http://www.klab.caltech.edu/~harel/share/gbvs.php

[176] H. Woong and K. Chong-Min, "A Multitransform Architecture for H.264/AVC High-Profile Coders," *IEEE Trans. on Multimedia,* vol. 12, pp. 157-167, 2010.

[177] H. Nothdurft, "Salience from feature contrast: additivity across dimensions," *Vision Research,* vol. 40, pp. 1183-1201, 2000.

[178] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River: NJ: Prentice Hall., 2003.

[179] W. C. Cohen, R. Greiner, and D. Schuurmans, *Probabilistic Hill-Climbing. Computational Learning Theory and Natural Learning Systems*, 1994.

[180] B. P. Gerkey, S. Thrun, and G. Gordon, *Multi-Robot Systems: From Swarms to Intelligent Automata*, 2005.

[181] U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis," *J. of Vision,* vol. 6, 2006.

[182] P. Verghese, "Visual search and attention: A signal detection theory approach," *Neuron,* vol. 31, pp. 523-535, 2001.

[183] K. Ji-Hye, L. Ji Won, P. Rae-Hong, and P. Min-Ho, "Adaptive edge-preserving smoothing and detail enhancement for video preprocessing of H.263," *Int'l Conf. on Consumer Electronics,* pp. 337-338, 2010.

[184] C. A. Segall, P. V. Karunaratne, and A. K. Katsaggelos, "Preprocessing of compressed digital video," *Photonics West 2001-Electronic Imaging,* pp. 163-174, 2000.

[185] M. Mancuso and A. Borneo, "Advanced pre/post-processing for DCT coded images," *IEEE Trans. on Consumer Electronics,* vol. 44, pp. 1039-1041, 1998.

[186] J. Xu, R. J. Sclabassi, L. Bing, and M. Sun, "Content-Based Video Preprocessing for Remote Monitoring of Neurosurgery," *Transdisciplinary Conf. on Distributed Diagnosis and Home Healthcare,* pp. 67-70, 2006.

[187] M. Li and Z. Xu, "An adaptive preprocessing algorithm for low bitrate video coding," *J. of Zhejiang University-Science A,* vol. 7, pp. 2057-2062, 2006.

[188] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing,* vol. 13, pp. 600-612, 2004.

[189] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *Asilomar Conf. on Signals, Systems and Computers,* pp. 1398-1402, 2003.

[190] Stephen Wolf and Margaret Pinson, " Video quality measurement techniques," NTIA Report 02-392, June 2002.

[191] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Processing,* vol. 15, pp. 430-444, 2006.

[192] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. on Image Processing,* vol. 9, pp. 636-650, 2000.

[193] D. M. Chandler and S. S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Trans. on Image Processing,* vol. 16, pp. 2284-2298, 2007.

[194] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/Referenceless Image Spatial Quality Evaluator," *Asilomar Conf. on Signals, Systems and Computers,* pp. 723-727, 2011.

[195] A. K. Moorthy and A. C. Bovik, "Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality," *IEEE Trans. on Image Processing,* vol. 20, pp. 3350-3364, 2011.

[196] R. Ferzli and L. J. Karam, "A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB)," *IEEE Trans. on Image Processing,* vol. 18, pp. 717-728, 2009.

[197] N. D. Narvekar and L. J. Karam, "A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD)," *IEEE Trans. on Image Processing,* vol. 20, pp. 2678-2683, 2011.

[198] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a Completely Blind Image Quality Analyzer," *IEEE Signal Processing Letters,* vol. 20, pp. 209-212, 2013.

[199] R. Soundararajan and A. C. Bovik, "RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment," *IEEE Trans. on Image Processing,* vol. 21, pp. 517-526, 2012.

[200] Video Quality Experts Group, ""Final Report from the VQEG on the validation of Objective Models of Video Quality Assessment, Pase II"," August 2003.

[201] ITU Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

[202] Q. Huynh-Thu, M. Ghanbari, D. Hands, and M. D. Brotherton, "Subjective video quality evaluation for multimedia applications," *Human Vision and Electronic Imaging XI,* 2006.

[203] H. T. Quan, M. N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of Rating Scales for Subjective Quality Assessment of High-Definition Video," *IEEE Trans. on Broadcasting,* vol. 57, pp. 1-14, 2011.

[204] W. Zhou and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters,* vol. 9, pp. 81-84, 2002.

[205] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. on Image Processing,* vol. 14, pp. 2117-2128, 2005.

[206] MeTriX MuX Visual Quality Assessment Package [Online]. Available: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[207] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. on Image Processing,* vol. 14, pp. 1918-1927, 2005.

[208] X. Marichal, M. Wei-Ying, and Z. HongJiang, "Blur determination in the compressed domain using DCT information," *Proc. Int'l Conf. on Image Processing,* pp. 386-390, 1999.

[209] W. Zhou, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," *Int'l Conf. on Image Processing,* pp. 477-480, 2002.

[210] D. Shaked and I. Tastl, "Sharpness measure: towards automatic image enhancement," *IEEE Int'l Conf. on Image Process.,* vol. 1, 2005.

[211] A. Murthy and L. Karam, "IVQUEST- Image and video quality evaluation software," [Online]. Available: http://ivulab.asu.edu/Quality/IVQUEST.

[212] LIVE Image Quality Assessment Database[Online]. Available: http://live.ece.utexas.edu/research/quality/

[213] A. K. Moorthy and A. C. Bovik, "A Two-Step Framework for Constructing Blind Image Quality Indices," *IEEE Signal Processing Letters,* vol. 17, pp. 513-516, 2010.

[214] HEVC software repository [Online]. Available: https://hevc.hhi.fraunhofer.de/trac/hevc/browser

[215] L. S. Karlsson and M. Sjostrom, "Improved ROI video coding using variable Gaussian pre-filters and variance in intensity," *IEEE Int'l Conf. on Image Processing,* pp. II-313-16, 2005.

[216] X. Jian, R. J. Sclabassi, L. Qiang, L. F. Chaparro, R. Marchessault, and S. Mingui, "Human Perception based Video Preprocessing for Telesurgery," *Ann. Int'l Conf. of the IEEE Engineering in Medicine and Biology Society,* pp. 3086-3089, 2007.

[217] M. de-Frutos-Lopez, H. Medina-Chanca, S. Sanz-Rodriguez, C. Pelaez-Moreno, and F. Diaz-de-Maria, "Perceptually-aware bilateral filtering for quality improvement in low bit rate video coding," *Picture Coding Symposium,* pp. 477-480, 2012.

[218] T. Q. Pham and L. J. van Vliet, "Separable bilateral filtering for fast video preprocessing," *IEEE Int'l Conf. on Multimedia and Expo,* 2005.

[219] L. Liang-Jin and A. Ortega, "Perceptually based video rate control using pre-filtering and predicted rate-distortion characteristics," *Proc. Int'l Conf. on Image Processing,* pp. 57-60, 1997.

[220] N. Young and A. N. Evans, "Psychovisually tuned attribute operators for pre-processing digital video," *Vision, Image and Signal Processing,* vol. 150, pp. 277-86, 2003.

[221] N. Young and A. N. Evans, "Digital video pre-processing with multi-dimensional attribute morphology," *Int'l Conf. on Visual Information Engineering,* pp. 89-92, 2003.

[222] D. A. Florencio, "Motion sensitive pre-processing for video," *Int'l Conf. on Image Processing,* pp. 399-402, 2001.

[223] S. Ke, "Video preprocessing techniques for real-time video compression-noise level estimation techniques," *Symp. on Circuits and Systems,* pp. 778-781, 1999.

[224] Xiph.org Test Media. [online]. Available: http://media.xiph.org/video/derf/

[225] D.Marr, *Vision*. New York: Freeman, 1982.

[226] J. Canny, "A Computational Approach to Edge Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. PAMI-8, pp. 679-698, 1986.

[227] A. P. Witkin, "Scale-space filtering: A new approach to multi-scale description," *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing,* pp. 150-153, 1984.

[228] ITU-R Recommendation BT.500, "Methodology for the subjective assessment of the quality of television pictures," Geneva, 2002.

[229] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Geneva, 2008.

[230] ITU-R Recommendation BT.710, "Subjective assessment methods for image quality in high-definition television," Geneva, 1998.

[231] H. Liu and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data," *IEEE Trans. on Circuits and Systems for Video Technology,* vol. 21, pp. 971-982, 2011.

[232] Visual Studio [online]. Available: https://www.visualstudio.com/en-gb

[233] QT project [online]. Available: http://qt-project.org/

[234] OpenCV [online]. Available: http://opencv.org/

[235] FFmpeg [online]. Available: https://www.ffmpeg.org/

Bibliography

[1] A. Yarbus, *Eye movements and vision*. New York: Plenum, 1967.

[2] D.Marr, *Vision*. New York: Freeman, 1982.

[3] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. New Jersey:Prentice Hall, 2002.

[4] Cohen.W, Greiner.R and Schuurmans.D, *Probabilistic Hill-Climbing. Computational Learning Theory and Natural Learning Systems,* vol. II, 1994.

[5] M. Land and B. Tatler, *Looking and acting: vision and eye movements in natural behaviour*: Oxford University Press, 2009.

[6] R. Jain, R. Kasturi, and B. G. Schunck, *Machine vision* vol. 5: McGraw-Hill New York, 1995.

[7] Gescheider. G, *Psychophysics: the fundamentals*, 3rd ed. Lawrence Erlbaum Associates, 1997.

[8] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, advantages, applications*. San Diego, CA: Academic Press, 1990.

[9] Y. Zhao, Complexity Management of Video Encoders, PhD Thesis, Robert Gordon University, March 2004.

[10] C. S. Kannangara, Complexity management of H. 264/AVC video compression, PhD Thesis, Robert Gordon University, 2006.

Appendix A: List of Publications

A.1 Conference publications

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

[2] J. Chilukamari, S. Kannangara, and G. Maxwell, "Investigation of the effectiveness of video quality metrics in video pre-processing," *IEEE International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, pp. 1-5, 2013. Available: https://www.youtube.com/watch?v=VJSyUSoa-bE

[3] J. Chilukamari, S. Kannangara, and G. Maxwell, "A low complexity visual saliency model based on in-focus regions and centre sensitivity," *IEEE International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, pp. 411-414, 2014. Available: https://www.youtube.com/watch?v=O4aulsbaj6g **[Best paper award]**

A.2 Journal Publications

[1] J. Chilukamari, S. Kannangara, Y.Zhao and G. Maxwell, "Frequency based in-focus visual saliency model for predicting fixations," *submitted to IEEE Transactions on Multimedia.*

Appendix B: Image/Video Saliency Detection

B.1 Introduction

An image/video saliency detection software application based on the attention model proposed in this thesis is implemented. The software detects salient regions of an image/video using focus, image centre and face detection. The application has been extended to include video encoding and decoding using H.264 video CODEC.

B.2 Project Objectives

The objectives of this project are:

1. Identify the algorithmic components of the proposed model required to be implemented.
2. Identify the libraries of programming functions that are needed for the real-time software implementation.
3. Investigate the required functionalities and design the high level architecture of the software. The proposed architecture should allow the provision for future modifications and extensions.
4. Implement focus, image centre and face detection components of the model and integrate them.
5. Profile the computation complexity of the implemented software and optimise the code.

B.3 Software Development Environment

The set of technologies, frameworks and libraries needed for the development of software prototype are briefly discussed in the following sub sections.

B.3.1 Visual Studio

Visual studio [232] is an Integrated Development Environment (IDE) developed by Microsoft for its .NET technologies. It uses software development platforms such as

Windows API and Windows Forms with the ability to produce both native and managed code. It supports various programming languages such as VC++, C# and F#.

B.3.2 Programming Language

The software prototype is implemented using C++ programming language. The main reasons for this choice are high performance, less run-time overhead and a high level of abstraction. It is an object-oriented programming language which allows the usage of plain C and assembler code. This helps in reducing the computational complexity by implementing instructions directly on the processor's register. Further, it is often used for real time computer vision applications because of computational efficiency when compared to other programming languages like MATLAB and Java.

B.3.2 Qt

Qt [233] is a cross-platform application framework for developing application software with graphical user interfaces (GUI). It is an open source project which is easy to use and very flexible. It can be used to run on various software and hardware platforms with little change in the underlying code. It helps to create good user interfaces with high performance in speed.

B.3.3 OpenCV

OpenCV [234] is an open source computer vision library. It has C++, C, Python and Java interfaces and supports Windows, Linux, Mac OS, iOS and Android. It was designed for computational efficiency aiming at real-time computer vision applications. The library is written in optimised C/C++ and it can take advantage of multi-core processing. OpenCV has a very big user community and it is easy to find documentation and information. In this project it is used for image processing and for face detection purposes.

B.3.4 FFmpeg

FFmpeg [235] is a cross-platform multimedia framework used to convert audio and video formats. It includes audio/video CODEC library which can be used to encode and decode the video frames. It can grab pictures from live audio/video source. In this project it is mainly used for decoding H.264 encoded videos.

B.4 Prototype Interface

The prototype's interface is shown in the Figure B.1. It has been developed using Qt project. The interface menu is shown on the left side of the application window. The menu gives the possibility to select the type of input. The input can be an image or a movie clip or it can be a live video feed from an external camera. The browse button is used to select the file location on the disk. The software is run by using the start/stop button. When the program is launched, if the camera source is not detected, then the camera radio button is disabled. During the run time, it is possible to dynamically select the core components of the attention model that have to be executed using the check boxes under the algorithms label. This process helps in studying the behaviour of individual components of the attention model. The output of the attention model is a grey scale image with salient regions shown with higher intensity. The input (left) and the output (right) images have been placed next to each other within the application window.

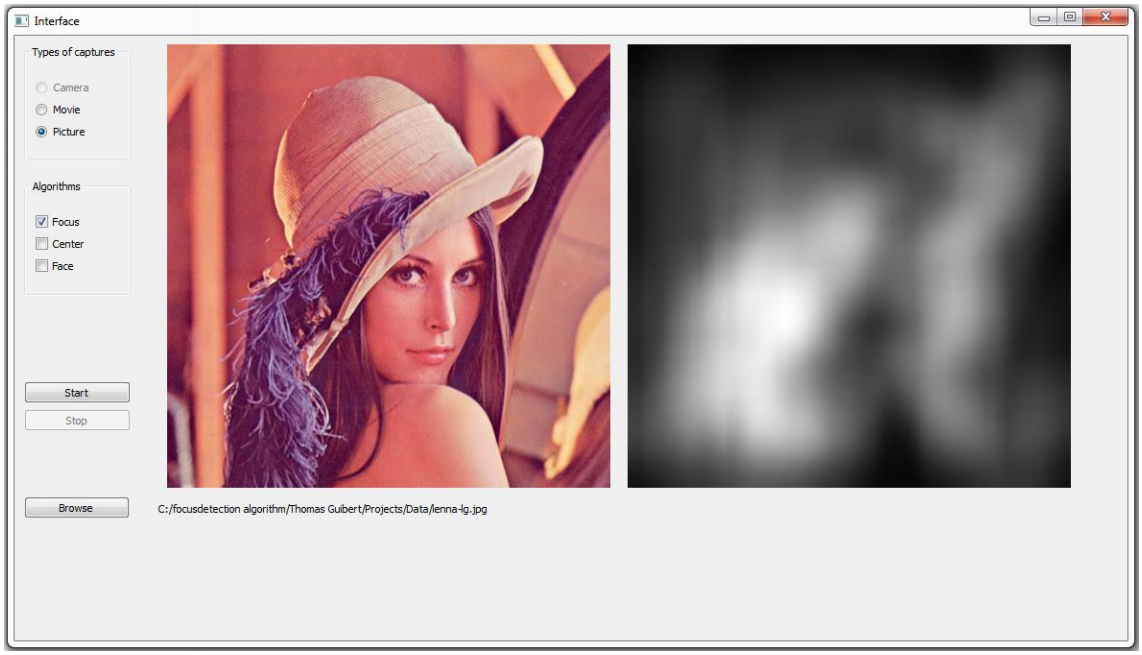


Figure B.1: Software prototype user interface

B.5 Software Architecture

B.5.1 Unified Modelling Language (UML) Diagram

This project architecture is composed of four namespaces. They are interface, algorithm, input and tools namespaces. These are briefly discussed below.

Interface: It handles the interface and all information that the user can see on the screen.

Algorithm: It is the main namespace of the prototype. It comprises of classes that process input pictures to generate focus map, image centre map and face map and combines them into a visual saliency map.

Input: It contains classes and member functions that handle the input from different sources such as external video camera, movie clips and images.

Tools: It contains classes that are useful for other namespaces.

B.5.2 Interface Namespace

This namespace is composed of four classes namely Image, Interface and QOpenCV2Widget as shown in the UML diagram in Figure B.2. The purpose of these classes along with their operation is discussed below.

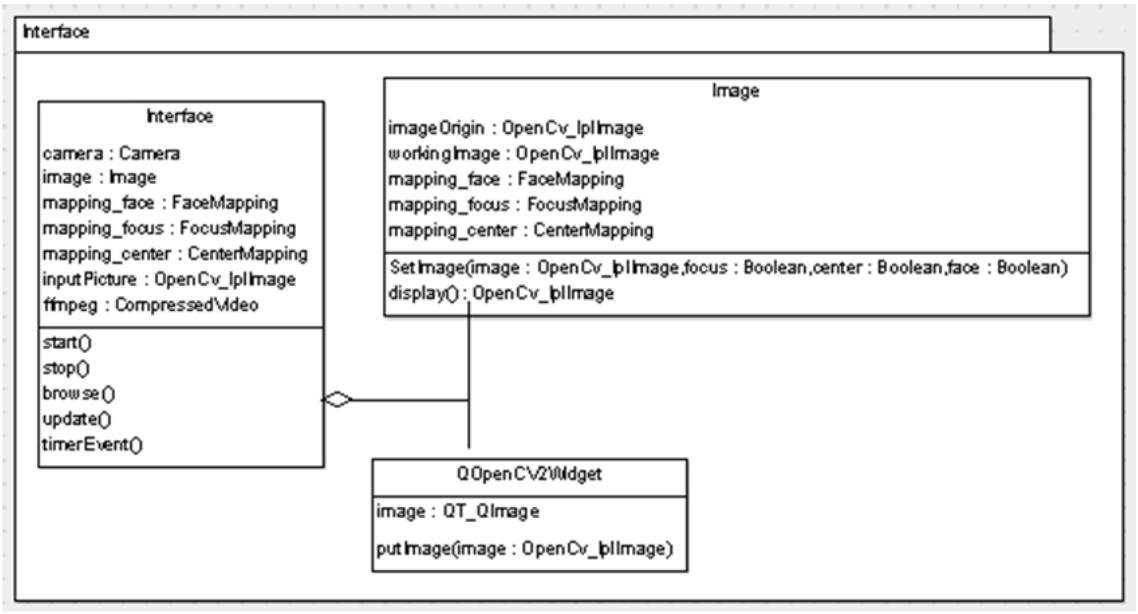


Figure B.2: Interface Namespace UML diagram

Image: It processes different types of algorithms such as focus detection, image centre and face detection on the given input picture to detect the saliency map.

Interface: It contains all the information that the user needs to use the application and run it.

QOpenCV2Widget: It converts OpenCV generated image (IplImage) into Qt image (QImage) and displays it on the application window.

When the user runs the application by choosing the necessary input type, the program initially creates two image data structures with image pointers *IplImage Saliency* and *IplImage origin*. *IplImage* is an OpenCV image data structure. The input picture selected by the user is assigned to the image pointer *IplImage origin*. The *IplImage saliency* is a grey scale image with same resolution as the original picture. *IplImage saliency* is the image pointer which points to the final visual saliency map. The program runs the attention model on *IplImage origin* and modifies the *IplImage saliency* pointer variable. Finally, the two image pointers are converted into two *QImage* type objects to display on the user interface. The program uses *QOpenCV2Widget* class to convert the image type. This class converts *IplImage*'s into RGB and assigns it to a *QImage* object. *QImage* class belongs to Qt interface which is used to display the images. Further, the program has a timerEvent, which is used to read a new frame from the input video clip or from the external camera at regular intervals. The application extracts a picture, runs the attention model on the picture and displays the saliency map on the user interface. This process iterates until end of the video file is reached.

B.5.3 Algorithm Namespace

The algorithm namespace consists of classes that implement the main components of the visual attention model. It contains *CenterMapping* class, *FaceMapping & Detection Mapping* class, *FocusMapping* class, *Process* and *Runnable* classes as shown in the Figure B.3.

Focus Mapping: This class implements DCT based in-focus visual saliency algorithm of the attention model.

Centre Mapping: This class implements the image centre algorithm of the attention model. It generates a Gaussian blob as a function of the input image resolution and places it in the centre of the grey scale map.

Face mapping and detection mapping: These classes detect the frontal human faces present in the images and generate the face saliency map. The prototypes uses OpenCV implementation of Viola Jones face detection algorithm to detect the human faces. Once the faces are detected the face mapping class generates the face saliency map.

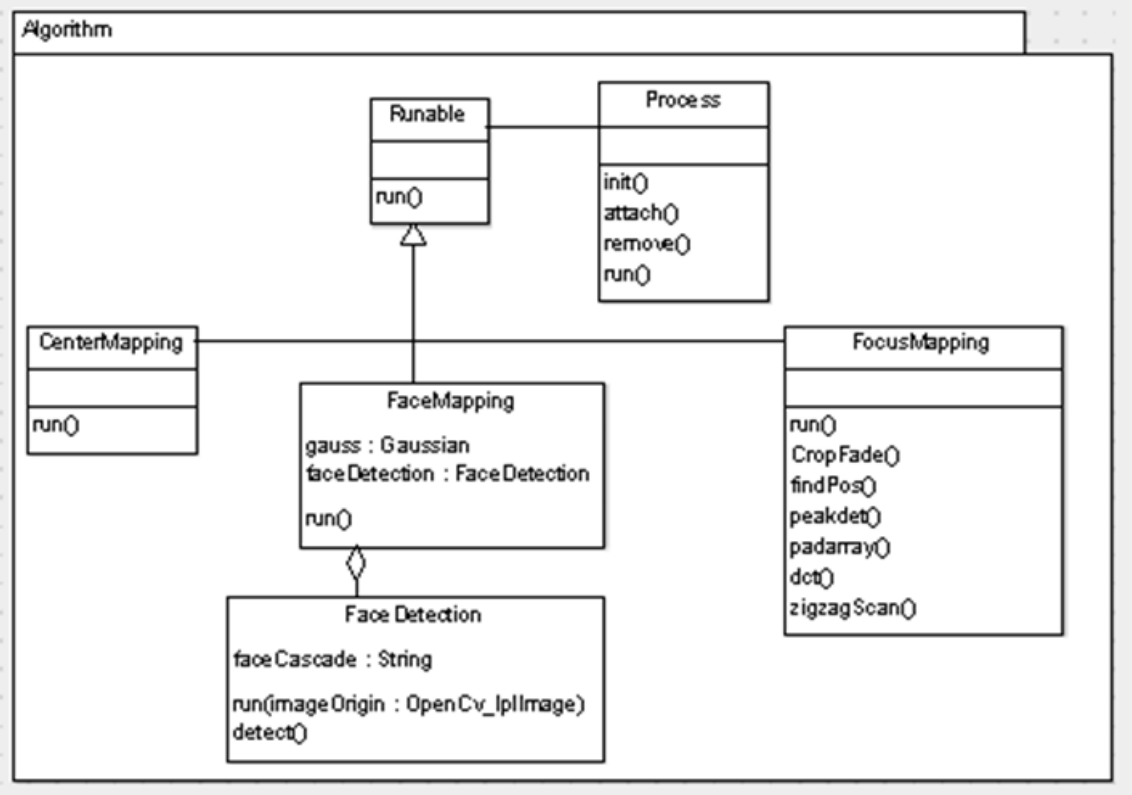


Figure B.3: Algorithm Namespace UML diagram

The final saliency map is generated by performing a weighted addition as described in chapter 6 of this thesis. Firstly the focus and the centre map is combined using weighted addition. Finally, the face map is overlaid on to the combined maps.

Process and Runnable: These classes are useful to extend the application in the future. To add another algorithm to this application, the developer just needs to add a new class to this namespace that extends *Runnable*. The *Process* class is called from the *Image* class to run all the algorithms implemented in the *Runnable* class.

B.5.4 Tools Namespace

The tools namespace is composed of three classes namely *Gauss* class, Tool class and *Test* class as shown in the Figure B.4. These are briefly discussed below.

Gauss: It is mainly used to create a Gaussian blob. *Gauss* class dynamically generates a Gaussian filter based on the information (kernel size, position) from the attention model.

Tool class: It encapsulates two member functions that are useful in other classes. For example, *roundf* function is used in all the mapping algorithm implementations to round the values.

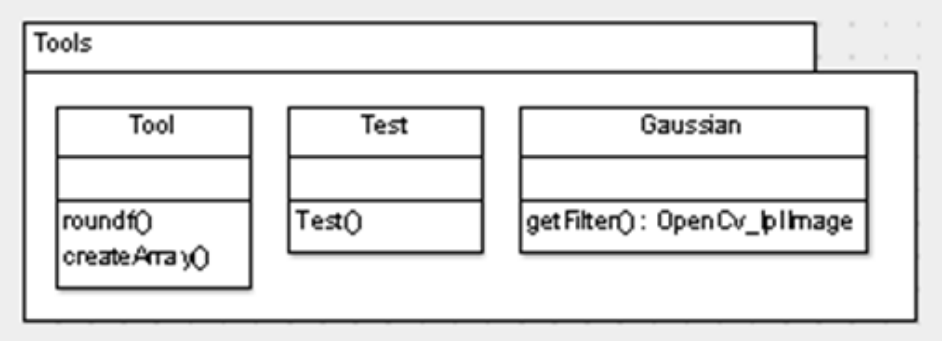


Figure B.4: Tools Namespace UML diagram

Test: *Test* class compares the output from the developed prototype to the output from the original MATLAB implementation.

B.5.6 Input Namespace

The input namespace is composed of five classes namely Input, Video, Picture, Compressed Video and Camera class as shown in the Figure B.5. The purpose of these classes is briefly discussed below.

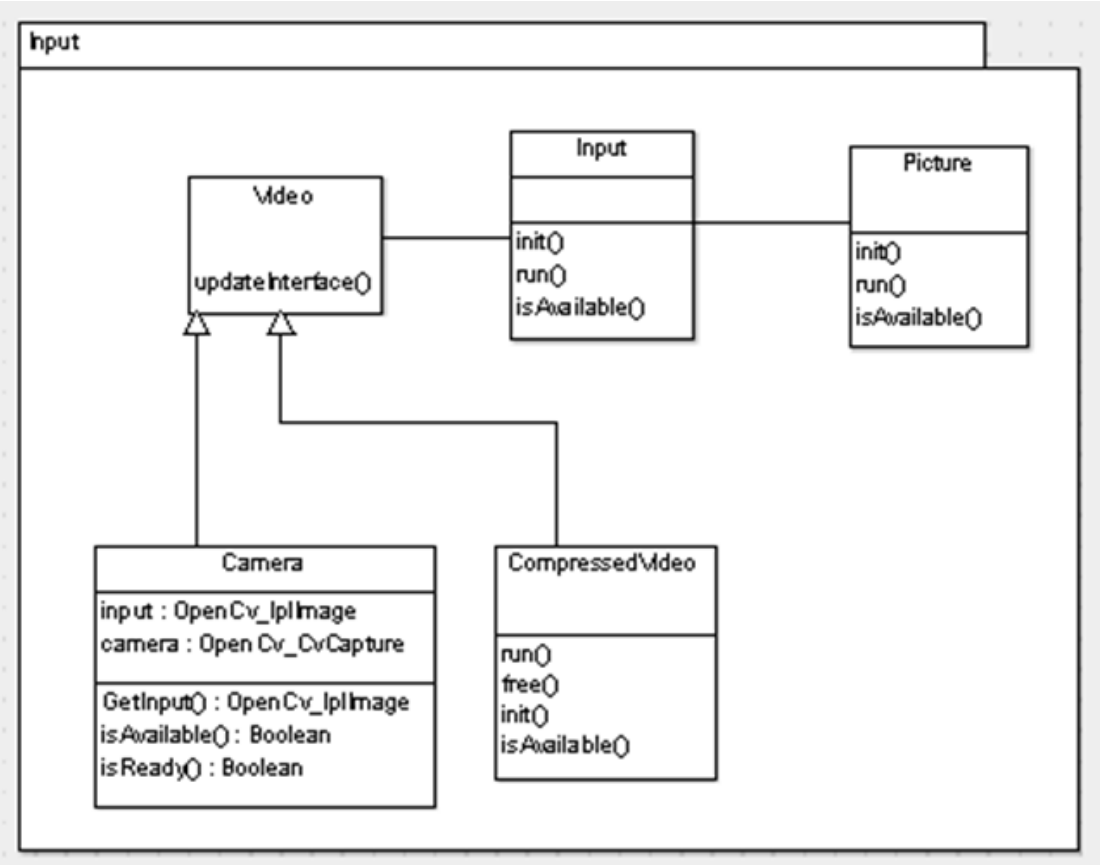


Figure B.5: Input Namespace UML diagram

Input: This class will act as an interface to take the picture, video from external source and movie clip as the input.

Picture and Camera: These classes handle the image and external camera objects and are given as the input to the Input class for further processing.

Video: It contains general function for video inputs from *Camera* and *Compressedvideo* classes.

CompressedVideo: The objective of this class is to extract image frames from compressed video file. A video sequence is a set of continues images (frames) captured at a particular frame rate. Each frame consists of a number of pixels (picture units) depending on the video format. A video file consists of both audio and video and is compressed in a particular format. It is necessary to separate these audio and video tracks before processing an image from the video. FFmpeg uses DeMux command to separate these tracks. The video track that is separated from the video file is then decompressed using a video codec.

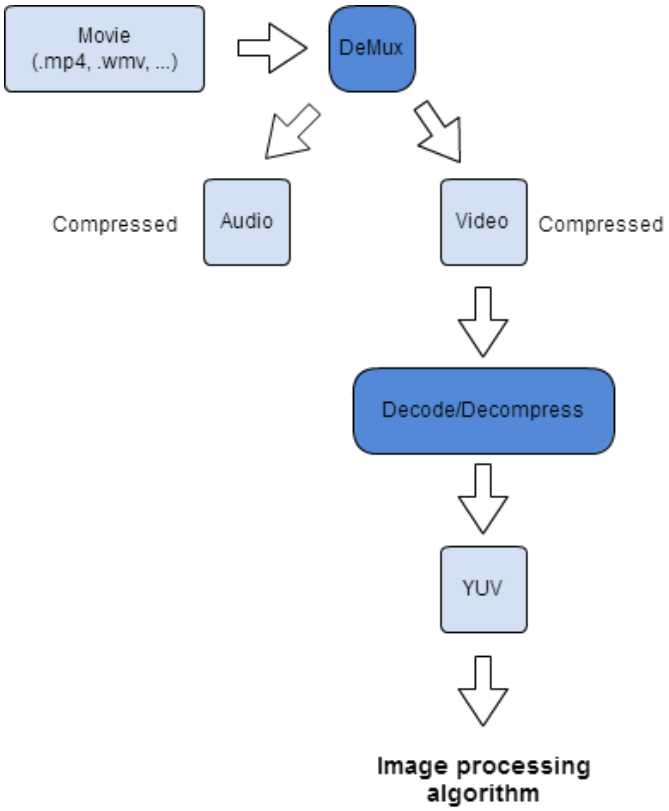


Figure B.6: FFmpeg video frame extraction

The video codec is selected from the FFmpeg library based on the compression format used to encode the video. The YUV frames are then extracted from the decompressed video for further processing. This entire process is illustrated in the Figure B.6.

B.5.7 Algorithm Behaviour

The Figure B.7 shows how the software works when the pictures are captured at regular intervals. First, the interface creates the instances of all the classes and checks for the availability of external camera before the first input picture is processed. The captured picture is displayed on the user interface using the *putImage* member function of QOpenCV2Widget's class. The OpenCV *IplImage* format is converted into *QImage* for displaying the image on the user interface. The displayed image is assigned to the Image object. The software runs the attention model on the Image object and generates a saliency map of the input Image. The Interface calls *Image*

class display function to get the output saliency map and displays it on the interface using *putImage* function.

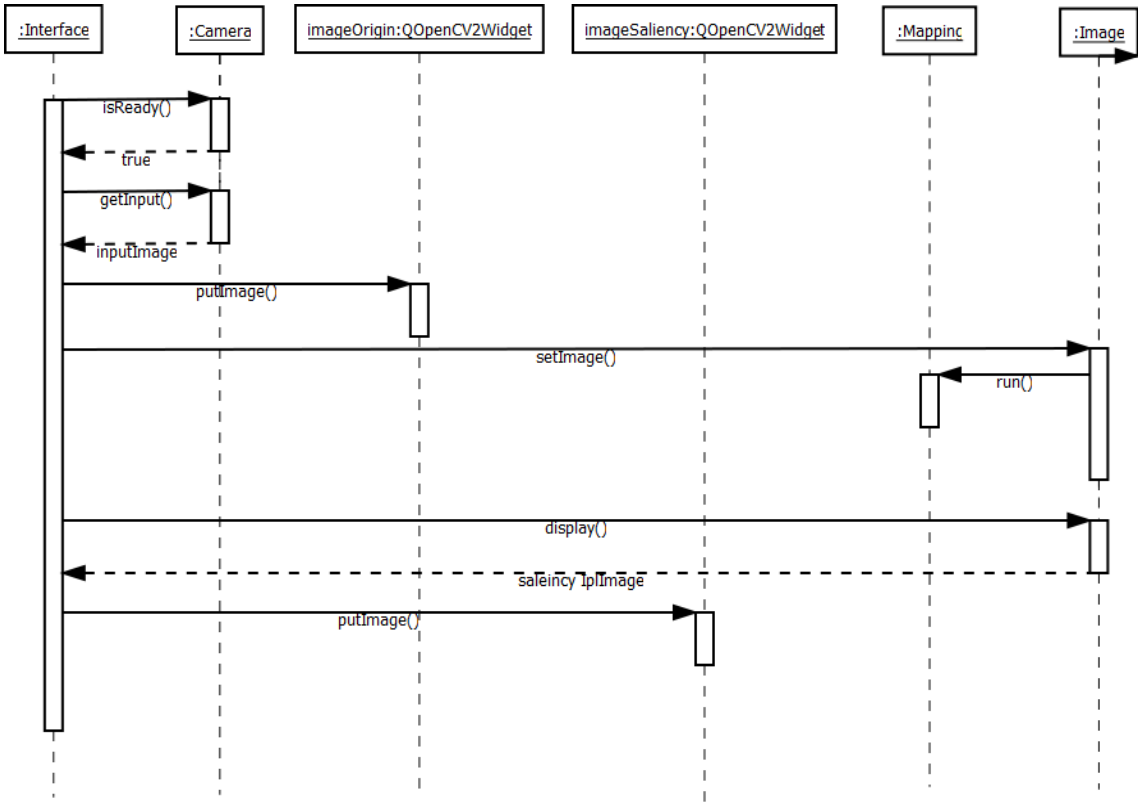


Figure B.7: Camera's pictures processing

B.6 Code Optimisation

The code optimisation process involves identifying the portion of the code or module of the program that is running slow. These computationally intensive modules are modified to reduce the time complexity.

B.6.1 Software Profiling

The software profiling is the investigation of the software behaviour during the execution. A profiler is used to do the software profiling. It is a performance analysis tool that measures the frequency and duration of the function calls, total time required by the software from the point of invocation to termination. In this project Visual studio's in-built profiler has been used to profile the code. The profiler has identified that some of the OpenCV and Qt project functions were called repeatedly and they causing delay

during the program execution. The different techniques that have been used to optimise the code are briefly discussed below.

B.6.2 Optimisation Methods

Array: The OpenCV implements a container for images called *CV::MAT* to access raw image data. This matrix has been often used in the program. *Mat* is a class with two data parts namely matrix header and pointer to the header containing pixel values. The size of the matrix header is constant and stores information such as matrix size, storage method, address etc. The *Mat* class is easy to use. However, the pointer to the header apart from storing pixel values stores other information such as picture type, pointer to the data, number of rows and columns etc. A better way is to use a 2D array to store the data. Thereby when a loop processes a picture, the program looks into a 2D array and doesn't need to use a pointer (using a pointer will make it to point to the matrix initially, then it points to the matrix data and finally to the pixel value). A 2D array is still an array of pointer. The 2D array can be further simplified by using a 1D array. The implementation will be less readable and requires more time and effort to code however it is a computationally faster approach. Indeed, to access a value from a 2D array the row and column number is enough as shown below.

$array[row][column]$.

However, in the case of a 1D array the code is less readable and the application needs to calculate position of the array as shown below.

$$array[row * numberOfRow + column].$$

Moreover, when we want to read an 8x8 array in a bigger array as shown in the Figure B.8 another variable is used to indicate the position of 8x8 array within the bigger array.

$$array[(Sum\ of\ no.rows\ in\ bigger\ and\ smaller\ \ array) * (no.of\ elements\ of\ \ bigger\ array) +\ \ bigger\ array\ column + smaller\ array\ column]$$

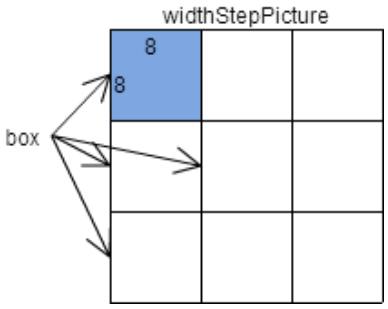


Figure B.8: Accessing 8x8 block in an image

This way of indexing the array is an optimised version used in the software. However, this version can be further improved by using a pointer on this array and incrementing it as *array Address* + 0, then *array Address* + 1, and then *array Address* + 2... Each time the program starts from base address to find the right position.

C function: During the run time the software initialises many arrays and also copies memory blocks. C functions such as memcopy and memset are used instead of *for* loops which consumes significant amount of time during the run time.

Increment/decrement operators: A simple operator creates a temporary object whereas addition operation creates additional memory spaces whose creation and destruction requires much time.

Bit shifting: Bit wise operations are faster than multiplication and division. The focus algorithm involves working frequently with 8x8 arrays of an image. Shifting operations are used instead of division and multiplication. For example shift number by 3 instead of multiplying it by 8. Depending on the processor type multiplication or division requires more number of processor cycles compared to shifting operation.

Datatypes: Memory and time can be saved by using a float variable instead of double. Floating point variables occupy only 32 bits whereas double variable needs 64bits of space for storing the data.

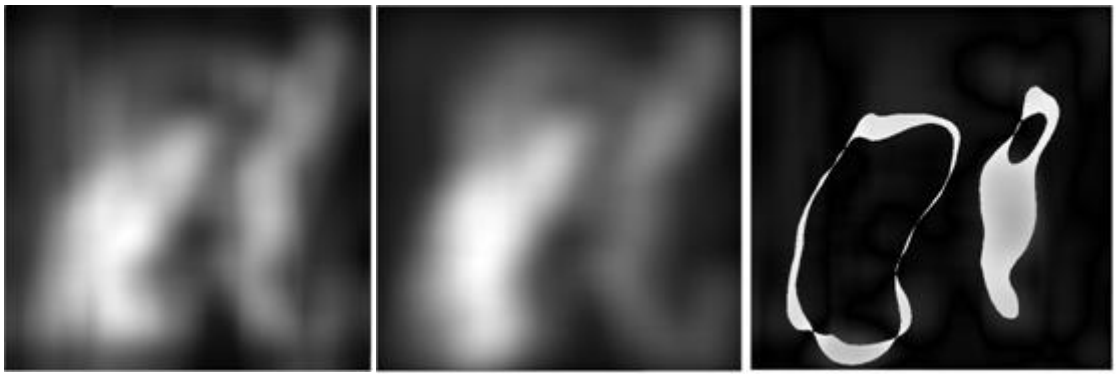
Loop unrolling: During the process of loop unrolling, the loop counter is updated less often and due to this the loop overhead is decreased. A better performance is achieved by using loop unrolling, however at the expense of increased code size (number of instructions between the branches increases). Compilers automatically unroll the loops with fixed number of iterations but when the number of iterations is unknown loops have to be unrolled manually.

B.7 Results and Discussion

During the software testing the output from the software prototype is compared with the output from the proof of concept developed in MATLAB. The testing process involved visual comparison of the outputs from focus detection algorithm, image centre algorithm, face map algorithm and the entire visual attention model's output. The absolute difference of outputs from MATLAB and C++ has been calculated to identify the minor differences. Further, the computational complexity of the individual components of the prototype is given.

B.7.1 Focus Detection

The focus maps from the C++ and MATLAB are shown in the Figure B.9. It can be seen that visually they both look similar however the absolute difference highlights minor differences. These minor differences could be due to the type of variables used. In the MATLAB code, variables are stored as double which needs 64 bits. In the C++ code, for optimisations reason, float variables which only require 32 bits to store the data are used. It can have a minor impact on the output focus map as values are rounded.



C++ **MATLAB** **Pixel Difference**

Figure B.9: Comparison between C++ and MATLAB focus detection implementation

B.7.2 Image Centre

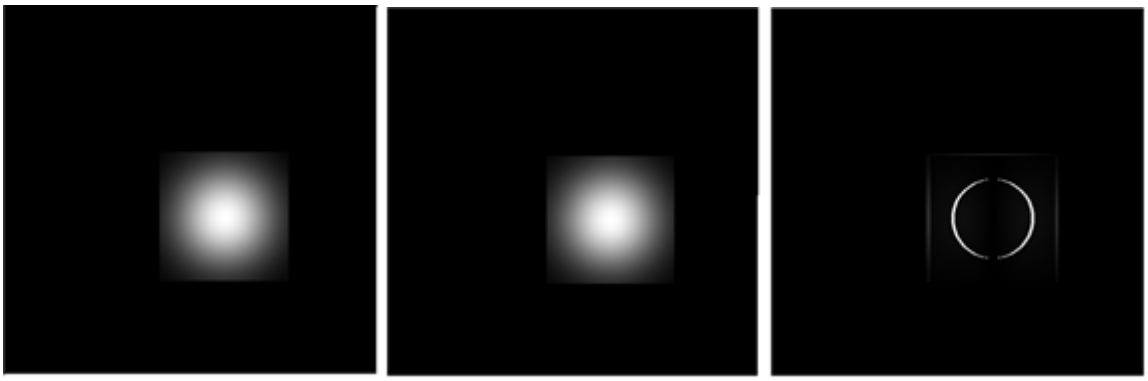
The centre map from the MATLAB and C++ is shown in the Figure B.10. The comparison shows that the output from the C++ and MATLAB are visually same. The absolute difference also shows no difference between the two images. Therefore, the C++ image centre implementation is exactly identical to the MATLAB version.

C++ **MATLAB** **Pixel Difference**

Figure B.10: Comparison between C++ and MATLAB image centre implementation

B.7.2 Face Map

The face map from the MATLAB and C++ is shown in the Figure B.11. The visual comparison between the C++ and the MATLAB shows that they are almost similar.



C++ **MATLAB** **Pixel Difference**

Figure B.11: Comparison between C++ and MATLAB face map implementation

However, the absolute difference indicates minor differences between the two outputs. This minor difference is from the OpenCV face detection algorithm. As already indicated it is due to the change in the type of variables used for optimisation reasons.

B.7.3 Visual Saliency Map

The visual saliency map from the MATLAB and C++ is shown in the Figure B.12. Visually the output from C++ and MATLAB are almost similar.

C++ **MATLAB** **Pixel Difference**

Figure B.12: Comparison between C++ and MATLAB visual saliency map implementation

The absolute difference also indicates this by showing a minor difference between the two outputs. These negligible differences clearly shows that the C++ implementation is very close to original MATLAB version and can be reliably used for visual saliency detection purposes.

B.7.4 Computational Complexity

The computational complexity of the software prototype is calculated over 100 images with resolution 1024x1024 on Intel core I7-2600K CPU operating at 3.40 GHz. The time complexity of the individual components of the model is given in the Table B.1. This performance can be further improved by using a low complexity version of *Haar* cascade in the OpenCV during face detection in the images. However, the speed improvement is achieved at the expense of loss in the accuracy of face detection.

Table B.1: Computational complexity of individual components of the attention model

Attention model Individual Components	Complexity (secs)
In-Focus detection	0.046
Centre detection	0.028
Face detection	0.131

B.7 Future Work

The future directions related to this development work are given below.

1. When the program is processing a movie, it is not possible to select a particular frame within the movie file. It could be interesting to add a scroll bar on to the application window to select the frame of the movie the viewer wants to study.

2. Recording the live feed that is obtained from the camera, provision to change the software internal variables such as Gaussian blur, algorithm variables through the user interface are some of the things that can be done to improve the software.

3. Code security is a major concern of the software prototype. The code security can be improved in three steps. In the first step, to protect confidentiality and integrity it is better to use security classes, crypto keys and encryption algorithms within the code. Buffer overflows which is kind of vulnerability in the computer software should be avoided. In the second step, the entire code has to be obfuscated. This process makes the code logic harder to understand. In the third step, as the .NET framework produces assemblies that are similar to assembly language code. Any programmer can use an assembly editor to reverse engineer and get the original source code. This problem has to be solved by converting the DLL's into native machine code (binary).

B.8 Conclusion

A software prototype has been developed to detect visually salient regions. It implements visual attention model proposed in this thesis. The user can detect salient regions in a picture, a video or a live feed from an external camera. The software provides front end flexibility of choosing the algorithms between focus, centre detection, face mapping algorithm and the overall integrated attention model. The results show that the prototype achieves similar results when compared to the proof of concept developed in MATLAB. Moreover, optimisation methods have been used to reduce the computational complexity of the software.