Høgskolen i Gjøviks rapportserie, 2008 nr. 2

Perceptual Image Difference Metrics –

Saliency Maps & Eye Tracking

Puneet Sharma

Gjøvik 2008

ISSN: 0806-3176

Perceptual Image Difference Metrics– Saliency Maps & Eye Tracking

Puneet Sharma

Masters Exchange Student

(Dr. B. R. Ambedkar National Institute of Technology-Jalandhar,India)

Norwegian Color Research Laboratory

Gjøvik University College, Norway

June 23, 2008

Abstract

Under natural viewing conditions humans tend to fixate on specific parts of the image that interests them naturally. Saliency map is the map of regions which are more prominent than other regions in terms of low level image properties such as intensity, color and orientation. With some modifications it can be used to simulate the natural human fixation also known as the gaze map. There are numerous applications in the field of engineering, marketing and art that can benefit from understanding of human visual fixation such as image quality evaluation, label design etc. The objective of this research is to understand the factors that influence the saliency map and gaze map and to modify the saliency map in order to make it similar to the gaze map. Eye movements of 20 test subjects were captured using eye tracking equipment available in the lab. The gaze maps obtained were averaged and superimposed over the corresponding original images. Saliency map toolbox [Walther(2006)] was modified by addition of face detection [Sauquet *et al.*(2005)Sauquet, Rodriguez & Marcel]. The gaze maps were analyzed and compared with modified saliency maps.

Preface

This master thesis has been carried out with the Norwegian Color Research Laboratory, Gjøvik University College, Norway and Dr. B. R. Ambedkar National Institute of Technology-Jalandhar,India. This project was funded by the Research Council of Norway.

Words elude me in expressing profound gratitude to my supervisors Dr. Faouzi Alaya Cheikh and Dr. Arun Khosla for their constant encouragement, constructive suggestions, thought provoking discussions and giving me full opportunity to work.

I also take the privilege to record my deepest appreciation and heartiest thanks to Dr. Jon Yngve Hardeberg for providing me all the resources and extending full support for working on this project.

I would like to thank all the members of the Norwegian Color Research Laboratory who gave valuable contribution to improve this project. I am beholden to my colleagues and friends who have always been ready to help me. I would also like to thank all those people who participated in the eye tracking experiment.

Puneet Sharma

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

The visual process starts when reflected light from an object in the outside world falls into the eye. Our brain's visual system processes the multitude of reflected frequencies into different shades and hues. The light that our eyes can see is also part of the electromagnetic spectrum (as shown in figure 1.1). The visible part of the electromagnetic spectrum is wavelength ranging between approximately 400 nm and 700 nm which is clearly a very small portion of the whole electromagnetic spectrum.

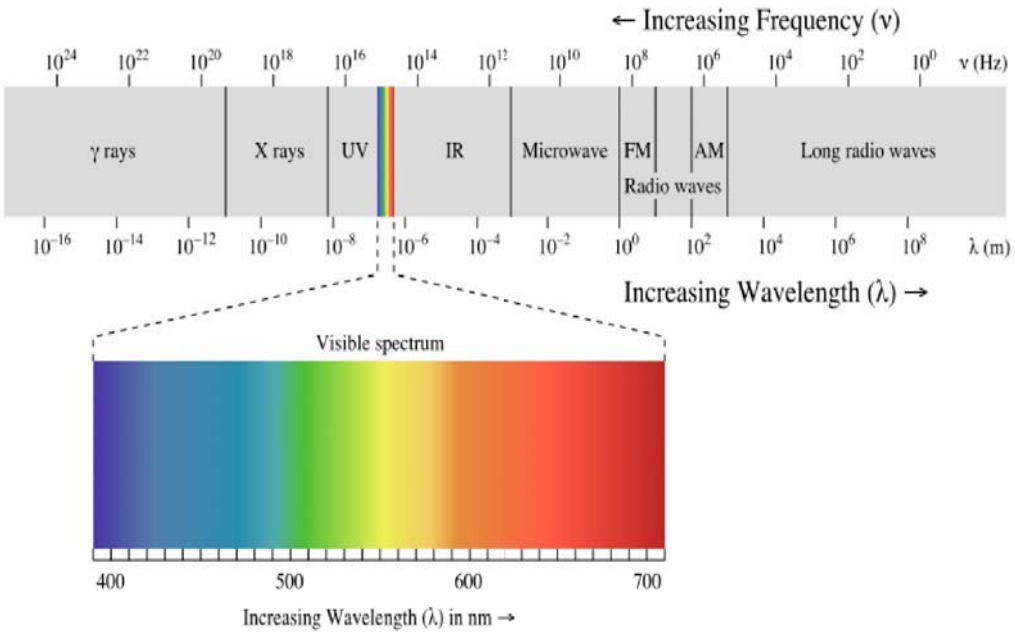


Figure 1.1: Electromagnetic spectrum with visible range.

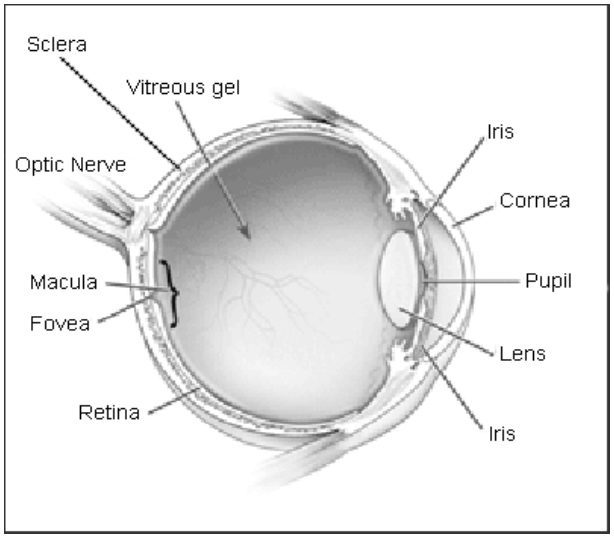


Figure 1.2: Different parts of eye.

Figure 1.2 shows different parts of eye.

- Cornea : Cornea is the clear bulging surface in front of the eye. It is the main refractive surface of the eye.

- Sclera : The sclera is the white of the eye and forms the outer coating of the eyeball.

- Retina : The retina is the light-sensitive inner lining of the back of the eye. Rays of light enter the eye and are focused on the retina by the cornea and lens.

- Lens : The lens is a transparent body behind the iris, the colored part of the eye. The lens bends light rays so that they form a clear image at the back of the eye on the retina. As the lens is elastic, it can change shape, getting fatter to focus close objects and thinner for distant objects.

- Iris :The iris is the colored circle surrounding the pupil. It changes the size of the pupil and allows different amounts of light to enter the eye.
- Macula : The macula is the small area at the center of the retina responsible for what we see straight in front of us, at the center of our field of vision. The macula is very important as it gives us the vision needed for detailed activities such as reading and writing, and the ability to appreciate color.
- Fovea : Fovea is the location on the retina of central gaze. Fovea is the location of highest visual acuity and best color vision.
- Vitreous gel : The vitreous gel is the clear, jelly-like substance that fills the inside of the eye from the lens to the retina.
- Optic Nerve: The nerve in the eye that serves to transmit information from the receptors in your eye to the brain.

Image forming light enters the cornea and is refracted by the cornea and the lens to be focused on the retina.Cornea has a greater refracting power than lens but it is constant. Whereas the refracting power of lens can be changed when the eye needs to focus at different distance. This process is called accommodation and occurs because of alteration in the lens shape.The diameter of incoming beam of light is controlled by iris. The eye rotates in the socket by the action of six extra-ocular muscles [Atchison & Smith(2000)].
When processing images for a human observer, it is important to consider how images are converted into information by the viewer. Understanding visual perception helps during image and video algorithm development. Image data represents physical quantities such as chromacity and luminance. Chromaticity is the color quality of light defined by its wavelength. Luminance is the amount of light. To the viewer, these physical qualities may be

perceived by such attributes as color and brightness. For the viewer, image perception begins at the eyeball. Light waves from the image are bent by cornea, the eye's fixed outer lens. Inside the eye, ciliary muscles flex, altering the shape of the lens to focus the image on the retina. The retina being the membrane lining the inner eyeball. Photons from the image flood the photo receptors inside the retina. The two basic types of photo receptors are rods and cones. Human retina has about 125 million rods and 7 million cones [Crane(1996)]. The rods are very sensitive to light intensity and enable us to see in dim light. They do not contribute to the color vision. The cones detect color and fine details. The cones are concentrated in the center of the retina in an area called the fovea. Cones form the basis of color perception and work at photopic conditions. Cones are of three types: L-cones, M-cones, and S-cones which are sensitive to long, medium, and short wavelengths, respectively. They are mainly located in the central part of the retina, called fovea, which is 2 degrees in diameter. Both psychological and physiological experiments give evidences to the theory of early transformation in the human visual system (HVS) of the L, M, and S signals issued from cones absorption. This transformation provides an opponent-color space in which the signals are less correlated. The principal components of opponent colors space are black-white (B-W), red-green (R-G), and blue-yellow (B-Y) [Meur *et al.*(2006)Meur, Callet, Barba & Thoreau]. There is a variety of opponent color spaces which differ in the way they combine the different cone responses.

There are at least three reasons why gaze control is an important topic in scene perception [Henderson(2003)]. First, vision is an active process in which the viewer seeks out task-relevant visual information. Secondly attention plays an important role in visual and cognitive processing. Thirdly eye

movements provide unobstructive, sensitive, real-time behavavioral index of ongoing visual and cognitive processing [Henderson(2003)].

Visual attention is the ability of a vision system, biological or artificial, to rapidly detect potentially relevant parts of a visual scene, on which higher level vision tasks, such as object recognition, can focus. It is generally agreed nowadays that under normal circumstances human eye movements are tightly coupled to visual attention [Jost *et al.*(2005)Jost, Ouerhani, Wartburg, Muri & Hugli]. This can be partially explained by the anatomical structure of the human retina, which is composed of a high resolution central part, the fovea, and a low resolution peripheral one. Visual attention guides eye movements in order to place the fovea on the interesting parts of the scene.

Understanding visual attention is a challenge due to the variability of human visual perception. Human eyes when looking on an image tend to fixate on some important parts of image because the complexity of visual world exceeds the processing capacity of the human brain [Fecteau & Munoz(2006)]. Attention implements an information-processing bottleneck that allows only a small part of incoming sensory information to reach short term memory thus understanding a complex scene is a series of computationally less demanding, local visual analysis problems [Itti & Koch(2001), Itti(2000)]. These regions are valuable for understanding the dynamics of human visual system and the development of applications in various fields. The selection of these regions depends on stimulus and goal driven objectives. The subjects selectively direct attention to objects in a scene using both bottom-up, image-based cues and top-down, task-dependent cues [Itti & Koch(2001)].

Human visual perception is task specific to objects in a scene for e.g. in a busy restaurant when we are looking for an empty table then we ignore all the other details like people, decorations, background etc. However if we are free to look we will pay attention to all the details like lights, decorations etc. of the same restaurant. So human vision is the perception of the scene depending on stimulus (saliency) and task assigned visual attention is the ability of a vision system, biological or artificial, to rapidly detect potentially relevant parts of a visual scene, on which higher level vision tasks, such as object recognition, can focus. It is generally accepted nowadays that under normal circumstances human eye movements are tightly coupled to visual attention [Jost *et al.*(2005)Jost, Ouerhani, Wartburg, Muri & Hugli]. This can be partially explained by the anatomical structure of the human retina, which is composed of a high resolution central part, the fovea, and a low resolution peripheral one. Visual attention guides eye movements in order to place the fovea on the interesting parts of the scene [Itti & Koch(2001)].

Saliency is the quality of an object or item to stand out from rest of the objects or items [Fecteau & Munoz(2006)]. There are many different physical qualities that can make an object more salient than other objects in the scene such as its color, orientation, size, shape, movement or unique onset .The analysis of the saliency maps shows the role played by low level image features. Intensity, color and orientation maps contribute to the saliency of the regions in an image. As saliency map emphasizes the object which is differentiated from the other objects in terms of its higher intensity, chromacity and orientation. The order in which the scene should be inspected is determined by saliency map [Fecteau & Munoz(2006)].

There are two types of visual attention

1. Top down
    - Slower process
    - Stimulus based (Observers knowledge is needed)
    - Intention plays a significant role
    - Cognitive or semantic understanding of the scene
      [Rajashekara *et al.*(2007)Rajashekara, van der Lindea, Bovika & Cormack]
2. Bottom up
    - Fast process
    - How people pay attention naturally
    - Strongly influenced by low-level image features
      [Rajashekara *et al.*(2007)Rajashekara, van der Lindea, Bovika & Cormack]

The model used for this project work will be bottom up and based on the model developed by Dirk Walther [Walther(2006)] .

1.2 A Model of Perceptual Processing [Ware(2004)]

[1] The model of human visual information processing can be divided into three stages as shown in figure 1.3

- Stage 1: Parallel Processing to Extract Low-Level Properties of Visual Scene
- Stage 2: Pattern Perception
- Stage 3: Sequential Goal-Directed Processing

[1]This section is reproduced with permission from Elsevier

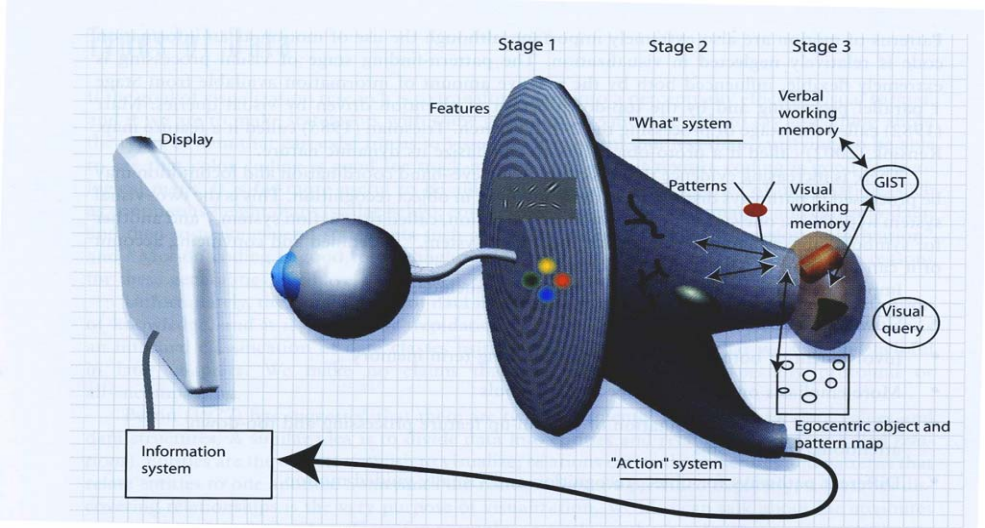


Figure 1.3: Three-stage model of human visual information processing [Ware(2004)].

1.2.1 Stage 1: Parallel Processing to Extract Low-Level Properties of Visual Scene

Visual information is first processed by large arrays of neurons in the eye and in the primary visual cortex at the back of the brain. Individual neurons are selectively tuned to certain kinds of information, such as the orientation of edges or the color of a patch of light. In stage 1 processing, billions of neurons work in parallel, extracting features from every part of the visual field simultaneously. It is also very rapid.
Information characteristics of Stage 1 processing include:

- Rapid parallel processing
- Extraction of features, orientation, color, texture, and movement patterns
- Transitory nature of information, which is briefly held in an iconic store

- Bottom-up, data-driven model of processing

1.2.2 Stage 2: Pattern Perception

At the second stage, rapid active processes divide the visual field into regions and simple patterns, such as continuous contours, regions of same color, and regions of the same texture. Patterns of motion are also extremely important, although the use of motion as an information code is relatively neglected in visualization.
Important characteristics of Stage 2 processing include :

- Slow serial processing
- Involvement of both working memory and long term memory
- More emphasis on arbitrary aspects of symbols
- In a state of flux, a combination of bottom-up feature processing and top-down attentional mechanisms
- Different pathways for object recognition and visually guided motion

1.2.3 Stage 3: Sequential Goal-Directed Processing

At the highest level of perception are the objects held in visual working memory by demands of active attention. In order to use an external visualization, we construct a sequence of visual queries that are answered through visual search strategies At this level only a few objects can be held at a time; they are constructed from available patterns providing answers to the visual queries. For example, if we use a road map to look for a route, the visual query will trigger a search for connected red contours(representing major highways) between two visual symbols (representing cities).

During human scene perception, high quality visual information is acquired only from a limited spatial region surrounding the center of gaze [Henderson(2003)]. Visual quality falls off rapidly and continuously from the center of gaze into a low resolution center surround. We move our eyes three times a second to reorient the fovea through the scene. Pattern information is only acquired during fixations.

Chapter 2

Literature Review

2.1 Saliency Maps

The complexity of visual information exceeds the processing capability of the human brain which forces us to select one or more object(s) in the scene for more detailed analysis at the expense of other items. Saliency is the quality of an object or item to stand out from rest of the objects or items [Fecteau & Munoz(2006)]. There are many different physical qualities that can make an object more salient than objects in the display such as its color, orientation, size, shape, movement or unique onset. Saliency at a given location is determined primarily by how different this location is from its surroundings in color, orientation, motion, depth etc. The saliency map was designed as input to the control mechanism for covert selective attention. Koch and Ullman [Koch & Ullman(1985)] state that the most salient location (in the sense defined above) in a visual scene would be a good candidate for attentional selection. Once a topographic map of saliency is established, this location is obtained by computing the position of the maximum in this map by a Winner-Take-All (WTA) mechanism. After the selection is made, suppression of activity at the selected location (which may correspond to the psychophysically observed "inhibition of return" mechanism) leads to selection of the next location at the location of the second-highest value in the saliency map and a succession of these events generates a sequential scan of the visual scene.

The authors of [Zhaoping(2005)] described the role of primary visual cor-

tex (V1) in creating a saliency map using autonomous intra-cortical mechanisms.Saliency of a visual location is described as the locations ability to attract attention without top-down factors. The firing rate of active V1 cell increases with respect to attention of the location in the image. The model produced usual contextual influences observed physiologically. V1 is necessary for normal visual awareness, and recent studies indicate that V1 activity is tightly correlated with awareness under various conditions [Tong(2003)].

As shown in figure 2.1 the procedure involved for saliency map [Walther(2006)] is described as follows :

- Input image is sub sampled and 6 different scales of the image are obtained.
- Intensity map is calculated from the R,G,B components of the image.
- Red-Green and Blue-Yellow maps of the image pyramid are calculated.
- Local orientation maps are obtained to the intensity pyramid levels for the angles of 0, 45, 90 and 135 degrees.
- Each iteration step consists of self excitation and neighbor induced inhibition, implemented by convolution with a "difference of Gaussians" filter followed by rectification.
- Feature maps are summed over the center surround combinations using across scale addition and the sums are normalized again. Conspicuity maps corresponding to color, intensity and orientation are obtained.
- All Conspicuity maps are combined into one saliency map.

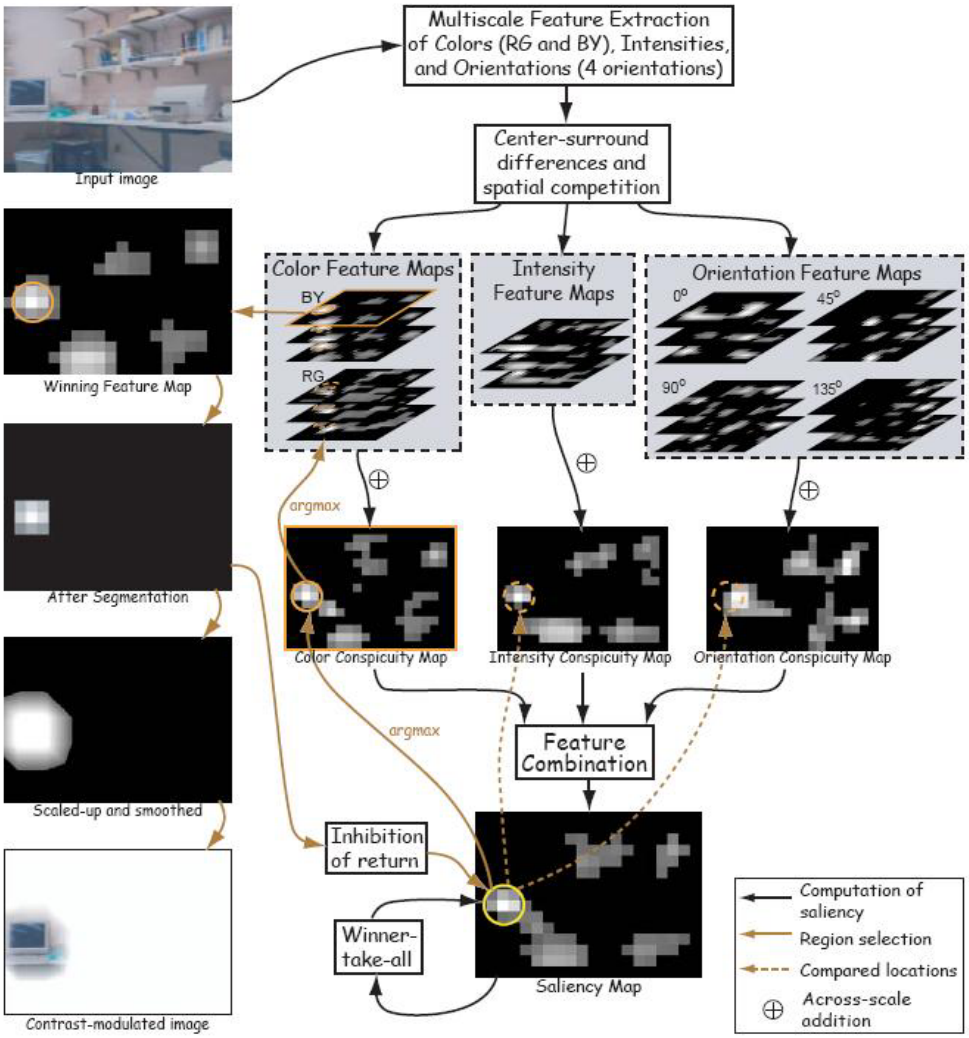


Figure 2.1: Saliency Map Algorithm (Dirk Walther).

The locations in saliency map [Walther(2006)] compete for the highest saliency value by means of WTA network of integrate and fire neurons. Winning location of this process is attended to and the saliency map is inhibited within a given radius of the location. Continuing WTA competition produces the second most salient location which is attended to and then inhibited.

Moran Cerf states that a combined model of face detection and low-

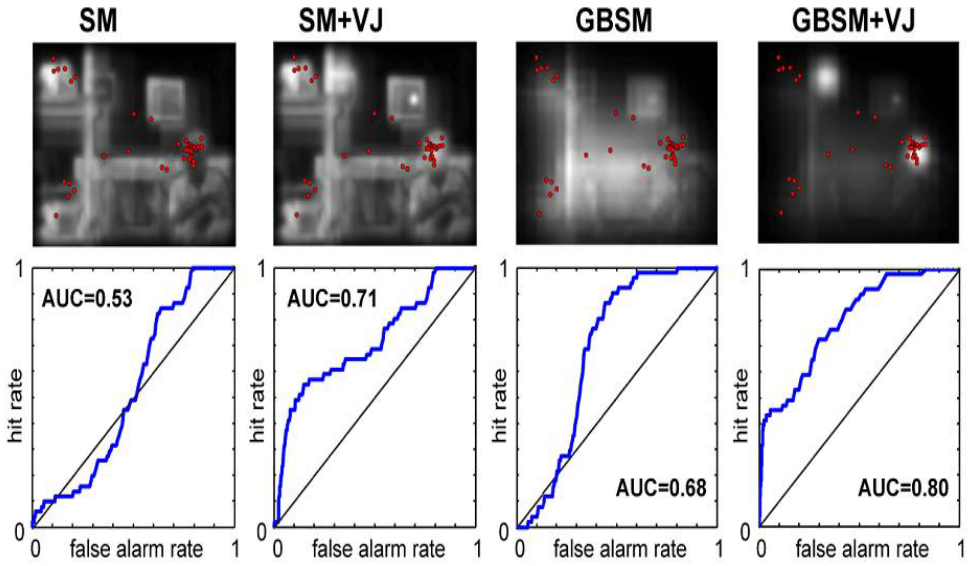


Figure 2.2: Comparison of Results.

level saliency outperforms a low-level model in predicting locations humans fixate on [Cerf *et al.*(2007)Cerf, Harel, Einhauser & Koch]. Viola & Jones [Viola & Jones(2001)] feature-based template matching algorithm combined with bottom-up saliency map model of Itti [Itti(2000)] was used. Seven subjects viewed a set of 250 images in a three phase experiment. Overall in both the experimental conditions i.e."free viewing" and "search" faces were powerful attractors of attention, accounting for a strong majority of early fixations when present. The new saliency model combined "bottom-up" feature channels of color, orientation and intensity with a clear special face detection channel based on Viola & Jones(VJ) algorithm [Viola & Jones(2001)]. Figure 2.2 shows the comparison of the area-under-the-curve (AUC) for an image. Top panel: image with the 49 fixations of the 7 subjects (red). First central fixations for each subject were excluded. From left to right, saliency map model of Itti et al [Itti(2000)].Saliency map(SM) with the VJ face detection

map (SM+VJ), the graph-based saliency map (GBSM), and the graph-based saliency map with face detection channel (GBSM+VJ). Red dots correspond to fixations. Lower panels depict ROC curves corresponding to each map. Here, GBSM+VJ predicts fixations best, as quantified by the highest AUC. The combination was linear in nature with uniform weight distribution for maximum simplicity. In attempting to predict the fixations of human subjects, additional face channel improved the performance of both a standard and a more recent graph-based saliency model in images with faces. In the few images without faces, false positives represented in the face-detection channel did not significantly alter the performance of the saliency maps - although in a preliminary follow-up on a larger image pool they boost mean performance. These findings pointed towards a specialized "face channel" in our vision system, which is subject to current debate in the attention literature. Inspired by biological understanding of human attention allocation to meaningful objects - faces - a new model for computing an improved saliency map which is more consistent with gaze deployment in natural images containing faces than previously studied models was developed. It suggested that faces always attract attention and gaze, relatively independent of the task. They should therefore be considered as part of the bottom-up saliency pathway [Cerf *et al.*(2007)Cerf, Harel, Einhauser & Koch].

Humans have a collection of passive mechanisms reducing the amount of incoming visual information. For instance, the signal stemming from the photo receptors is assumed to be compressed by a factor of about 130:1, before it is transmitted to the visual cortex [Meur *et al.*(2006)Meur, Callet, Barba & Thoreau]. Nevertheless, the visual system is still faced with too much information. To deal with the still overwhelming amount of input, an active selection, involving eye movement, is required to allocate processing resources to some parts of

our visual field. Oculomotor mechanisms involve different types of eye movements. A saccade is a rapid eye movement allowing jump from one location to another. The purpose of this type of eye movement, occurring up to three times per second, is to direct a small part of our visual field into the fovea in order to achieve a closer inspection. This last step corresponds to a fixation. Saccades are therefore a major instrument of the selective visual attention. This active selection is assumed to be controlled by two major mechanisms called bottom-up and top-down controls. The former, the bottom-up attentional selection, is linked to involuntary attention. This mechanism is fast, involuntary, and stimulus-driven. Observers unconsciously tend to select central locations of the image in order to catch the potentially most important visual information [Meur *et al.*(2006)Meur, Callet, Barba & Thoreau]. Human vision relies extensively on the ability to make saccadic eye movements to orient the high-acuity fovea region of the eye over the over the targets of interest in the visual scene [Rao *et al.*(2002)Rao, Zelinsky, Hayhoe & Ballard]. The authors of [Meur *et al.*(2006)Meur, Callet, Barba & Thoreau] used a coherent computational approach for the modeling of the bottom-up visual saliency. Contrast sensivity functions, perceptual decomposition, visual masking and center surround interactions were used in the model. Ten natural color images with various contents were selected. The quality of these pictures was degraded using different techniques (spatial filtering, JPEG, JPEG2000 coding etc.) Forty-six pictures were finally obtained. Every image was seen in random order by up to 40 observers for 15 seconds each in a task-free viewing mode. Qualitative or subjective evaluation showed that similarity between the predictions and the experimental results were good.The architecture of the proposed model was similar in spirit to the Koch and Ullman architecture. The fundamental difference was the normalization

of all the early visual features. The visibility threshold was modified by the context, and was incorporated by the modeling of visual masking. Linear correlation coefficient and the Kullback-Leibler divergence, were used to conduct the qualitative comparison. These coefficients were 0.71 and 0.46, respectively. The proposed model outperformed the model of Itti [Itti(2000)] in all the tested configurations [Meur *et al.*(2006)Meur, Callet, Barba & Thoreau]. A new bottom-up visual saliency model known as Graph-Based Visual Saliency(GBVS) was proposed in [Harel *et al.*(2006)Harel, Koch & Perona]. It consisted of two steps: first forming activation maps on certain feature channels, and then normalizing them in a way which highlighted the conspicuity and admitted combination with other maps. This model powerfully predicted human fixations on 749 variations of 108 natural images, achieving 98% of the ROC area of a human-based control, whereas the classical algorithms of Itti & Koch [Itti(2000)] achieved only 84%.

The authors of [Tatler *et al.*(2005)Tatler, Baddeley & Gilchrist] recorded eye movements of the subjects when they viewed natural images. Signal detection and information theoretic techniques were then used to compare fixated regions to those that were not. The results sugessted that the balance between the top-down and bottom-up control of saccade target selection changes over time. Bottom-up component is more influential early in viewing but becomes less so as viewing progresses. It was concluded that fixated locations tended to be more distinctive in higher spatial content [Tatler *et al.*(2005)Tatler, Baddeley & Gilchrist].

The authors of [Rajashekar *et al.*(2008)Rajashekar, van der Linde, Bovik & Cormack] used four low-level local image features: luminance, contrast and bandpass outputs of both luminance and contrasts to analyze that image patches around human fixations had, on average, higher values of each of these fea-

tures than image patches selected at random. Foveated framework was used for analysis of the fixations. Circular patches of diameters 32, 64, 96, 160, 192 pixels centered at each fixation were extracted. Each image was displayed for 5 sec. in a fixed order for all observers. The image patches around observers' fixation points were then analyzed to determine if the statistics of the four image features: luminance, contrast, luminance-bandpass, and contrast-bandpass were statistically different from image patches that were picked randomly. Contrast-bandpass showed the greatest difference between the human and random fixations, followed by luminance bandpass, RMS contrast and luminance. Using these measurements GAFFE(Gaze-Attentive Fixation Finding Engine) is proposed by the authors
[Rajashekar *et al.*(2008)Rajashekar, van der Linde, Bovik & Cormack].

2.2 Analysis of Saliency maps using simple images

This section describes the results obtained from the Saliency Map toolbox made by Dirk Walther [Walther(2006)], on simple images. Saliency map is obtained by linear combination of Intensity map, Color map and the Orientation map. Color map is obtained by the combination of Blue-Yellow and Red-Green which are obtained by across the scale addition of the image for 6 levels of scales. Orientation map is obtained for 4 angles of orientation i.e. 0, 45, 90 and 135 degrees, again for the same number of scales. Figure 2.3 shows the saliency map regions obtained for the figure containing three rectangles of black, red and yellow colors. Black rectangle is labeled as the first most salient region followed by red rectangle as the second most salient region and yellow rectangle as the least salient region in the image. Figure 2.4 shows the color map and intensity map of the same test image. Black rectangle because of its higher intensity map stands out as most salient region as the combined

maps of Blue-Yellow and Red-Green generate lower values (as compared to intensity map). Figure 2.5 shows the orientation map for the same test image. Higher intensity content also contributes to higher orientation map because the orientation map is obtained from Gabor filtering of intensity map. Thus the Black rectangle here also has highest orientation map. After the first salient region is selected the inhibition of return prevents the selection of the next salient region around that area. So the next most intense region in the Intensity map (i.e. figure 2.4) is the red rectangle. Orientation map (i.e. Figure 2.5) also shows higher orientation values for the red rectangle as compared to the yellow rectangle. So we get the red rectangle as second most salient region. When the saliency map is activated for the third salient location both the black and the red rectangle regions have been inhibited so it goes for the least salient region from the color map as shown on Figure 2.4.

Figure 2.6 shows the saliency maps for a modified image in which the black rectangle is replaced by cyan rectangle to study the effect of color map.The saliency toolbox selects the red rectangle as the most salient region based on intensity followed by the cyan rectangle as the second most salient region and the yellow rectangle as the least salient region. As shown in figure 2.7 the intensity map shows higher values for the red rectangle so it is selected as the most salient region based on higher values of intensity map and orientation map (shown in figure 2.8). Once this region is inhibited the next salient region is selected as cyan rectangle because of its higher value in combined Blue-Yellow and Red-Green color maps(as shown in figure 2.7).The least salient region comes out to be yellow rectangle because of its low intensity and color maps.

Figure 2.9 shows the saliency maps for a different image to study the effect of orientation maps. The intersection of the lines i.e. center is shown as the most salient region because of its higher orientation values as seen in the orientation maps (figure 2.10). The next most salient region is the region that stands out of the rest of the lines because of its high orientation values. This region is the extended line region as shown in figure 2.9. Color maps play no role in this image.

Figure 2.11 shows the saliency maps for regions on the basis of intensity map only. Because it is black and white so chromatic components play no role in the saliency map for this image. The most salient region is the triangle, polygon followed by circle and then rectangle. Figure 2.12 shows the intensity map and orientation maps for this image. The center of the triangle has higher intensity values and combined with the orientation map it results in the region of highest saliency. Because it generates higher values of orientation for the angles of 0,45 and 135 degrees. The polygon is chosen as the next highest salient region again on the basis of its higher orientation values in the map. After this region is inhibited the circle shows higher values of orientation map as compared to the rectangle and hence is chosen as the next salient region. Rectangle having the least values of the orientation is chosen as the least salient region. This analysis proves the basis of the saliency map which are intensity, chromatic and orientation maps. This analysis shows that the regions that are differentiated in terms of intensity, color and orientation combine to generate higher values for the saliency map.

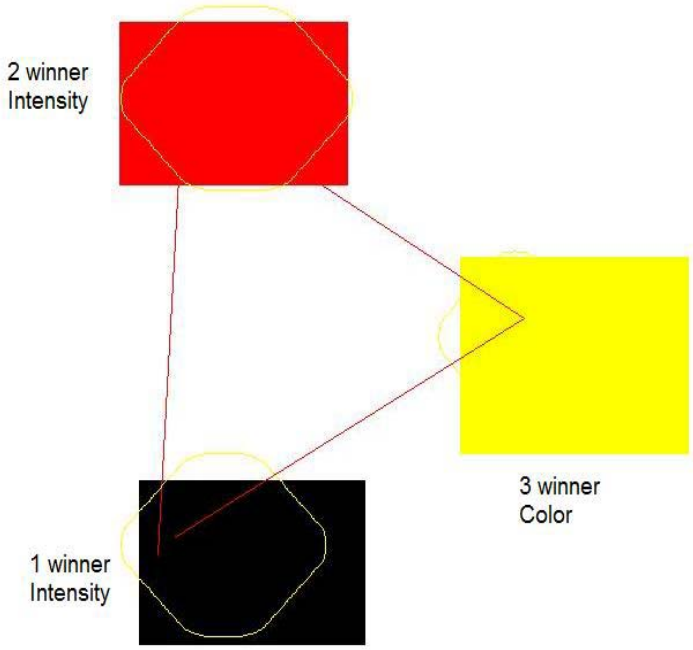


Figure 2.3: Saliency Maps.

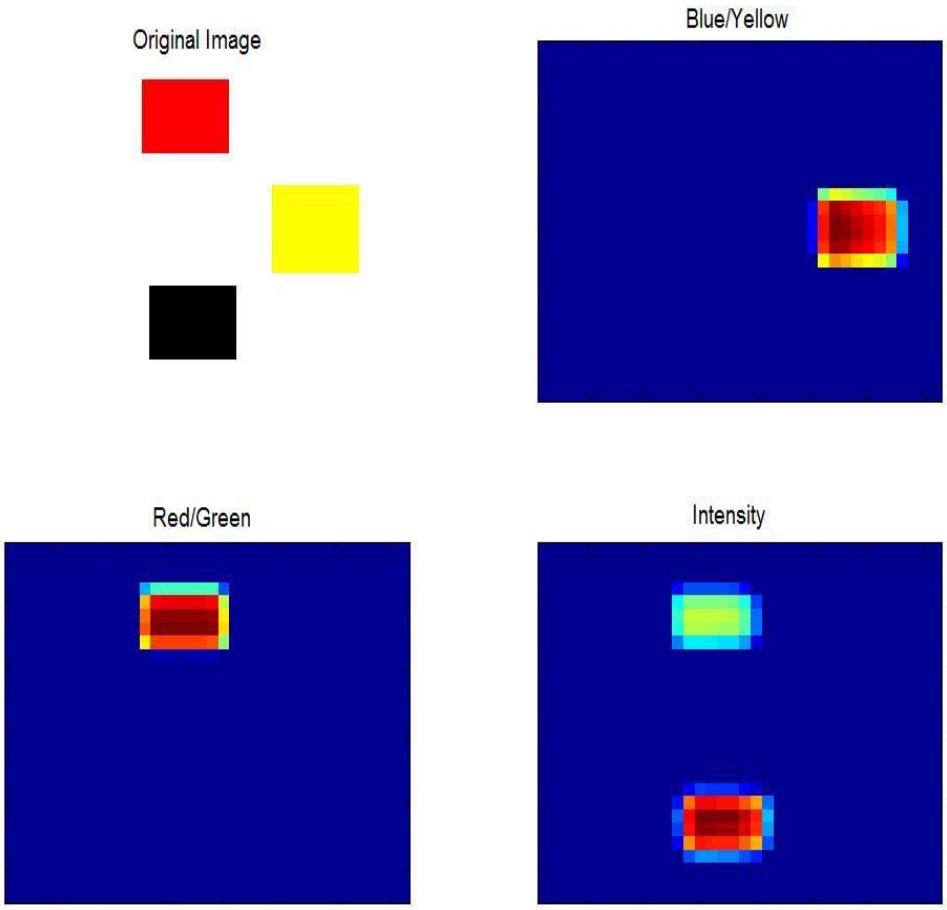


Figure 2.4: Color and Intensity Maps.

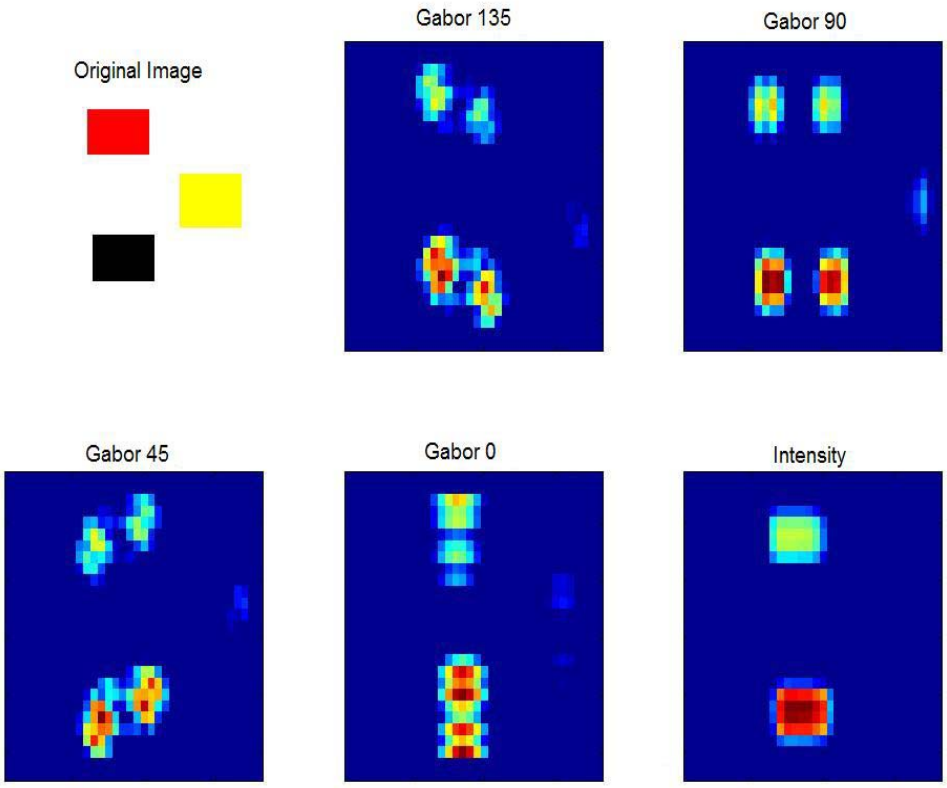


Figure 2.5: Orientation Maps.

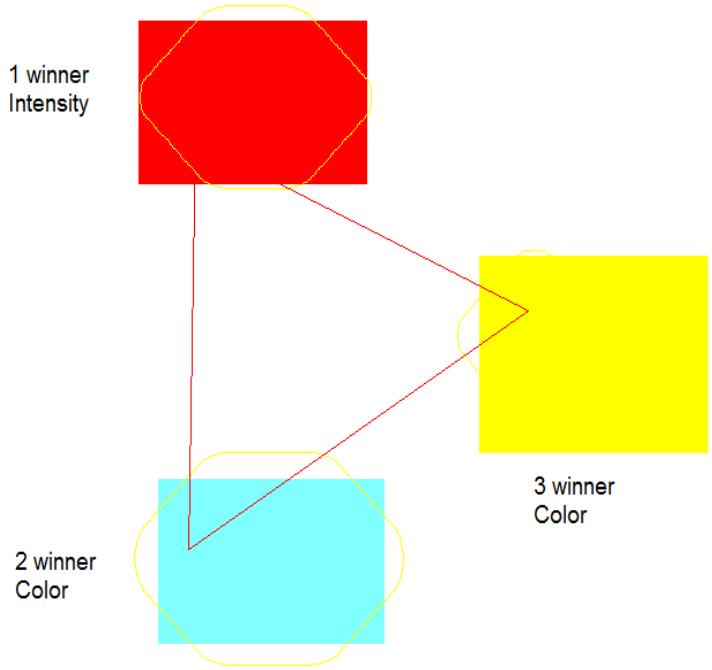


Figure 2.6: Saliency Maps.

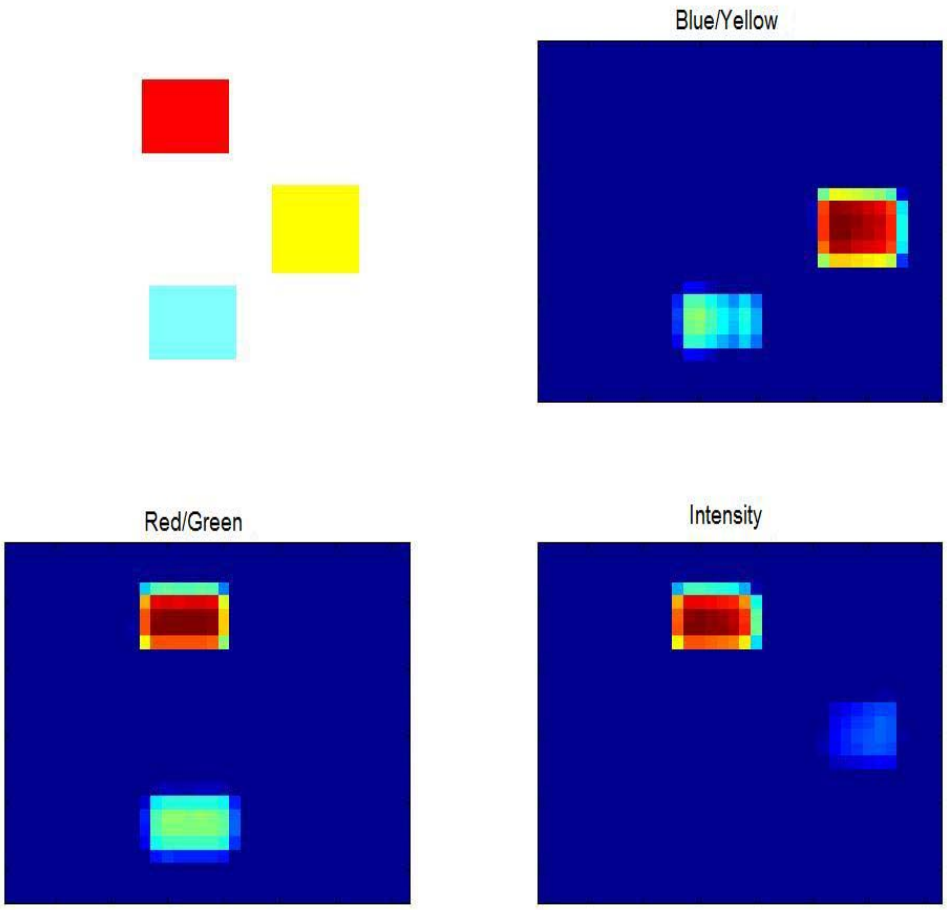


Figure 2.7: Color and Intensity Maps.

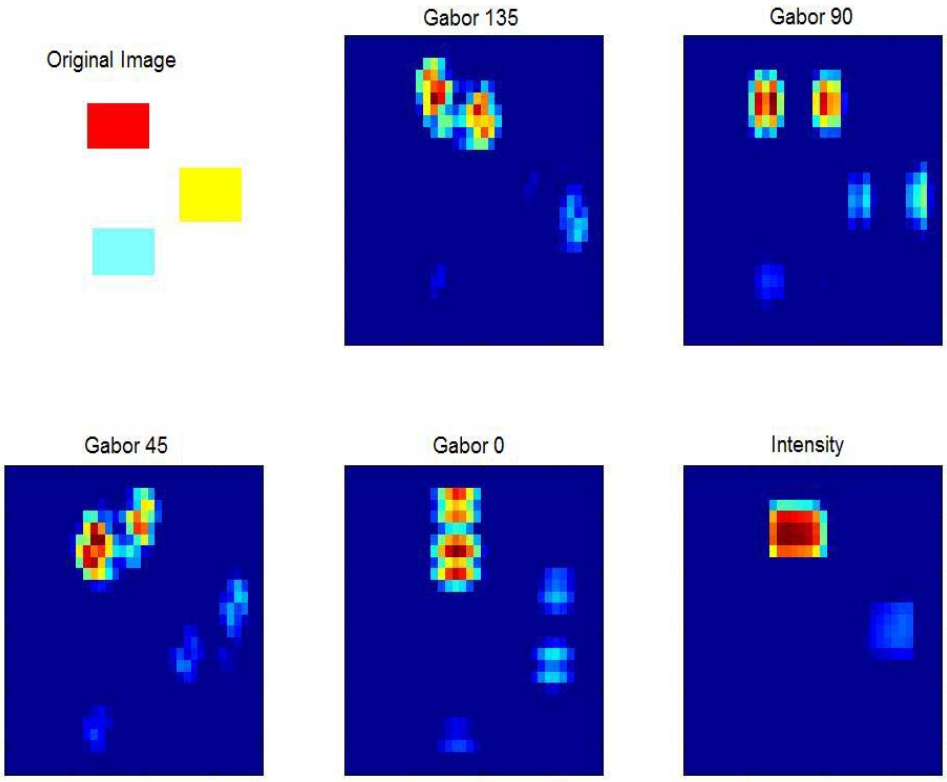


Figure 2.8: Orientation Maps

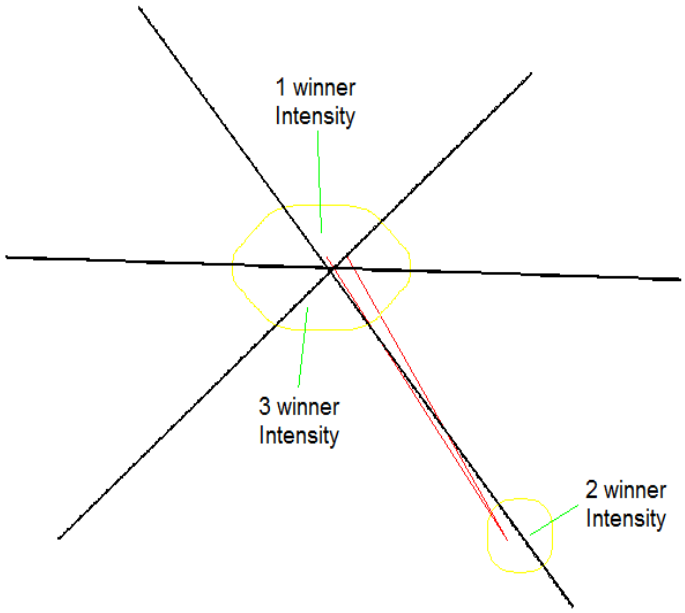


Figure 2.9: Saliency Map.

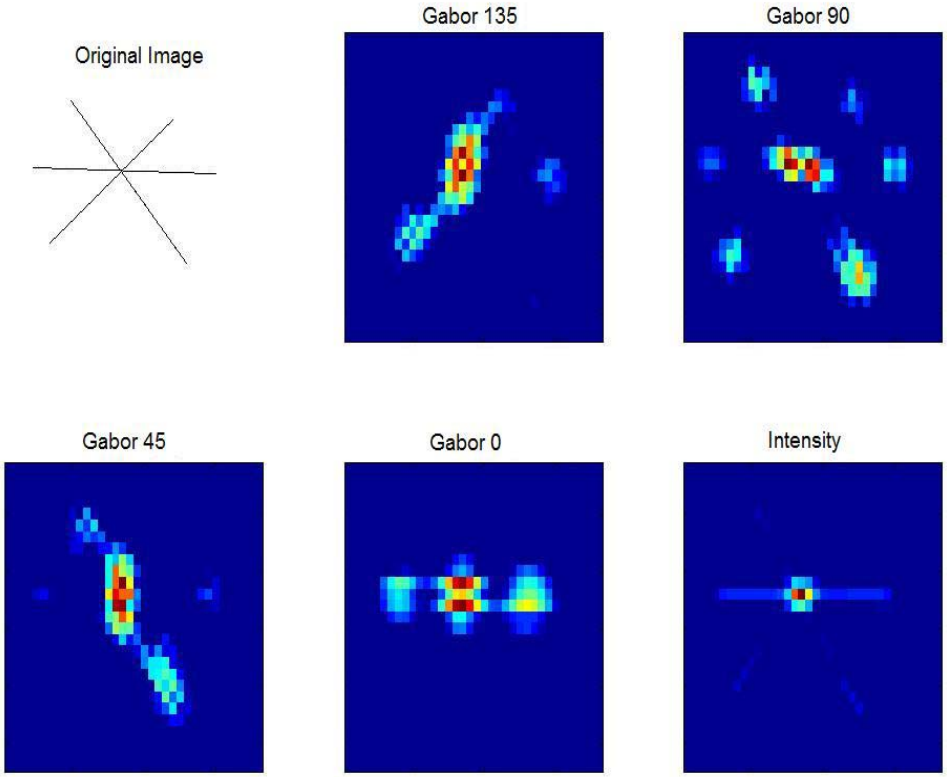


Figure 2.10: Orientation and Intensity Maps.

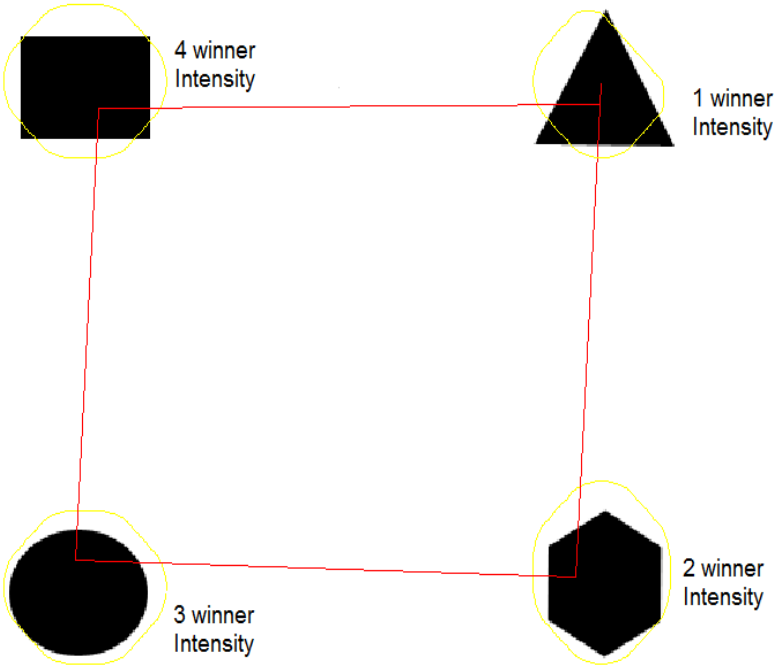


Figure 2.11: Saliency Map.

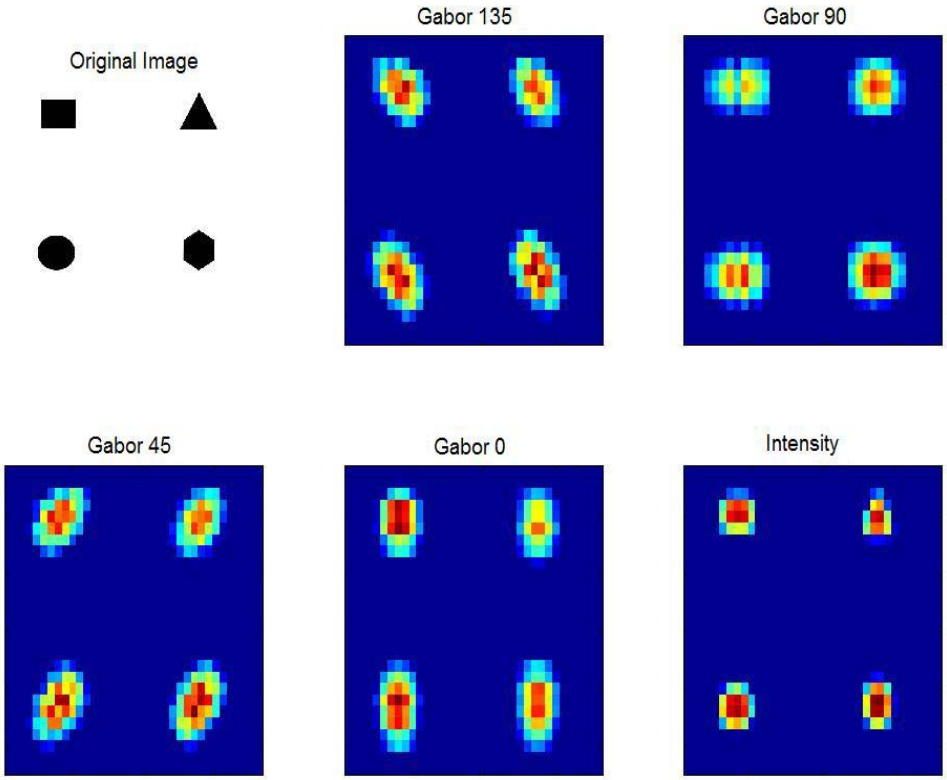


Figure 2.12: Orientation and Intensity Maps.

Chapter 3

Experiment Setup & Gaze Maps

3.1 Experimental Setup

Subjects were shown a set of 190 images and data was recorded as they viewed each image. Each image was shown for 2 sec only, similar to the experiment performed in [Cerf *et al.*(2007)Cerf, Harel, Einhauser & Koch]. The experiment was divided into two phases in order to provide adequate rest to the test subjects. As shown in figure 3.2 the experiment was performed in 6 steps. First of all the dominant eye of the test subject was found by using Porta test [Roth *et al.*(2002)Roth, Lora & Heilman].For the Porta test subjects were asked to extend one arm and align pointer finger with extended arm at the corner of the room with both eyes open. Subjects then closed one eye or the other alternately and reported with the eye closure the largest alignment change. Dominant eye was regarded as the eye when closed caused more change [Roth *et al.*(2002)Roth, Lora & Heilman]. After finding the dominant eye of the test subject the eye tracker was calibrated to it. Performing the calibration on the dominant eye yields accurate results which is an important issue in this project [Roth *et al.*(2002)Roth, Lora & Heilman, Pedersen(2007), Oishia *et al.*(2005)Oishia, Tobimatsub, Arakawaa, Taniwakia & ichi Kira]. The calibration time differed from one user to another. After calibration experiment phase 1 was started. In this phase each subject was shown a set of 90 images for a total period 180 seconds. Each image was shown for a period of 2 seconds only. After this phase subject was given a break to rest his or her eyes. The eye tracker is recalibrated before entering the experiment phase

2. In experiment phase 2 each subject was shown a set of 100 images for a total period of 200 seconds. Each image was shown for 2 seconds only. After completing this step subject was given a questionnaire to fill. The distance between the eye of the subject and the monitor was kept at 70 cm giving a viewing angle of 30x23 degrees. The viewing angle (as shown in figure 3.1) is calculated by measuring the width of the monitor and distance between the viewing location and the monitor.

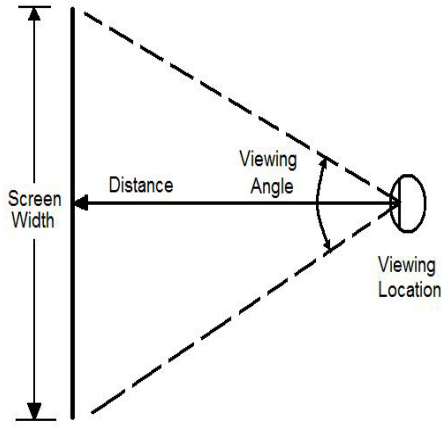


Figure 3.1: Viewing Angle.

The participants were seated so that they had a downward gaze angle. Several studies done by Von Noorden showed that participants had head movements even at small movements of the eyes [Burian & Noorden(1985), Pedersen(2007)].According to Von Noorden a downward gaze angle results in smaller head movements , this will decrease errors in the data set [Burian & Noorden(1985), Pedersen(2007)]. The chair used had 4 legs, armrests and backrest, this type of chair was chosen to minimize observer movement. The intensity of light at the front of display, back of the display was 159 lux and 146 lux respectively as measured from i1Display device. The subjects were not given any instructions to judge the images or look for something in particular. The

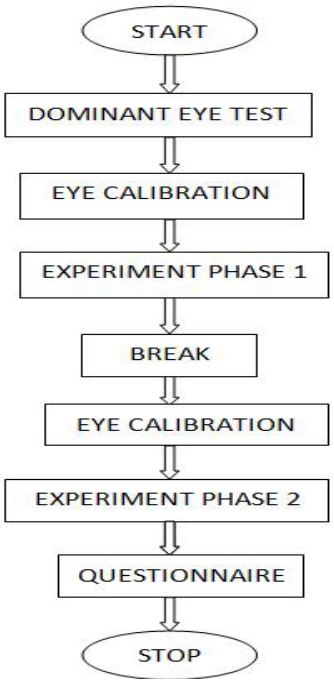


Figure 3.2: Experiment Workflow.

trails were performed under free view conditions on the dominant eye of the user using eye tracker equipment available in the lab. The eye tracker shown in figure 3.3 was used to measure the fixation of the user's eye at important parts of the image. Software allows online gaze position computation, real-time visualization, online fixation analysis, and digital output for control purposes. Results can be analyzed with a 3rd party software. However Matlab was used for analyzing the results due to the complexity of the experiment and for better analysis of the results. The eye tracker has the following features [Pedersen(2007), smi(2005)] :

- Sampling Rate : 50/60 Hz
- Tracking Resolution, Pupil/CR : 0.1 deg.
- Eye tracking equipment is a contact free gaze measurement device.

- Gaze Position Accuracy : 0.5 - 1 deg.
- Operating Distance Subject-Camera : 0.4 - 1.0
- Head Tracking Area : 40 x 40 cm at 80 cm distance.

Figure 3.4 shows the eye tracker with both the display pc (on the left) and the pc connected to the eye tracker (on the right). Figure 3.5 shows the setup of the experiment with the subject seated on the chair. Subject looking on the display pc (on the left) and the eye movements being recorded by the pc connected to the eye tracker (on the right).

Figure 3.3: Eye Tracker.

The lab has two types of eye tracker equipment :

- Head Mounted Eye Tracking Device (HED)
- Remote Eye Tracking Device (RED)

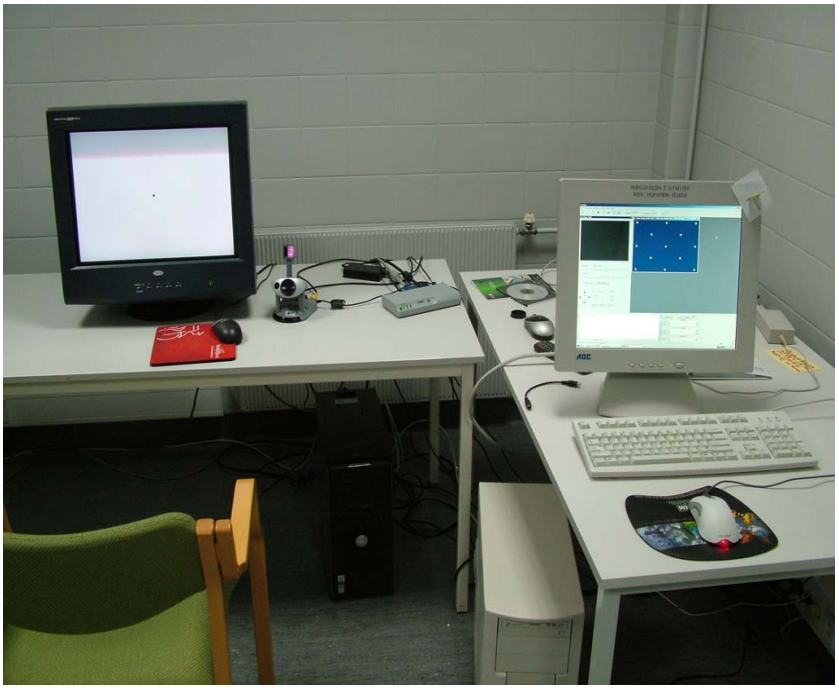


Figure 3.4: Experiment Setup 1.

Remote Eye Tracking Device was used for the experiment. Observers felt worse with Head Mounted Eye Tracking Device(HED) than Remote Eye Tracking Device(RED) because of the size and weight of HED. The difference of precision between HED and RED is about 10-16 pixels for printed images [Kominkova *et al.*(2008)Kominkova, Pedersen, Hardeberg & Kaplanova] These reasons favored the use of RED for this experiment. While viewing a stationary object, the eyes alternate between fixations and saccades (very fast eye movements). Each saccade leads to a new fixation. Typically there are about three saccades per second, but since saccades are so fast, they occupy only about 10 percent of the total viewing time. Vision is suppressed during the saccades and almost all the visual information is collected during the fixations [Schubiger *et al.*(1991)Schubiger, Moser, Hacisalihzade & Muller]. Appendix B shows the thumbnail view of images from the database that are

Figure 3.5: Experiment Setup 2.

used in the experiment. It can be observed that none of the two images have the similar background but same people are appearing in different images. Some random images were added to break the sequence feeling during viewing of similar images and to maintain the level of interest among the subjects.

3.2 Results for Gaze Map

Fixation is the pause over the informative region of interest and saccade is the rapid movement in between fixations.The process of identification of fixations and saccades is complex in nature. Fixation detection for the SMI eye tracker is based on dispersion based algorithm which provides robust and accurate information [Salvucci & Goldberg(2000)]. The gaze maps obtained for 20 users were summed and normalized to get 190 gaze maps cor-

responding to 190 images. There were 190 images with 212 faces in them, 17 grayscale images and 39 images with no faces. Subjects fixated on 194 of the total 212 faces in the database which validate one of the observations made by [Cerf *et al.*(2007)Cerf, Harel, Einhauser & Koch]. Subjects fixated on the faces in all the grayscale images also. Objects including faces at center of the image were fixated more as compared to faces or objects at other parts of the image. But in a few images faces at the other parts of images were fixated more than faces at the center. This was possibly because subjects fixated more on the changes i.e. new faces or new objects appearing in the images. The area of the face regions in the image was also a factor as larger face (as a result of person standing near to the camera) was fixated more as compared to smaller face. After the faces the most prominent regions for subjects were objects like toy car, toy banana, mobile phones, magic cube, books etc. Numbers, alphabets on posters were also strong in attracting attention of subjects. Figure 3.7 shows the gaze map obtained after gaussian blurring and normalization. The range of values varies from 0 to 1 where 1 represents the red and 0 represents the blue. Such a representation is known as heat map. Figure 3.8 shows the gaze map in the form of heat map with the original image superimposed in the background. The regions fixated by the subjects are face, book (at the center of the image), scanner and the pc monitor. The regions that are fixated appear red or yellow depending on the fixation or gaze time spent by the subject on those regions within the 2 second time of each image view. The maps are filtered using Gaussian filter to enhance the representation of the map Figure 3.8 shows the averaged gaze map. We can see that there is large variability in the viewing patterns of the observers. But some regions are fixated more than others. In this picture the face region and the book at center of the book shelf are the two most

fixated regions which can be judged by the redness in these regions of the heatmap. Scanner and the PC monitor can also be classified as other less fixated regions. However there are also other regions near the top of book shelf and corner of the book shelf that have been fixated which may or may not contain any information. This may be due to the randomness in the viewing pattern of eyes.

Figure 3.9 shows the gaze map of another image. Clearly the most fixated regions are the two faces and magic cube followed by a fixation at the center of the image. This gaze map again establishes the importance of the faces (which can be judged by the amount of fixation on the faces) in the images.

Figure 3.10 shows the gaze map of another image. There are lots of fixations in this image. The face in the poster of Einstein is the strongest attractor followed by the text at the bottom of the poster. The top left edge of the poster is also very strongly fixated. The less fixated regions include a face that is in dark environment. The selection of top left corner of poster and the corner of book shelf (in figure 3.8) may be due to the importance of edges or orientation in the gaze map. Since studies of gaze control have sugessted that empty and uninformative regions are often not fixated [Henderson(2003)]. Hence also establishing the the fact that contrast and edge information are more discriminatory than luminance or chromacity [Tatler *et al.*(2005)Tatler, Baddeley & Gilchrist].

Figure 3.11 shows another interesting gaze map with a number of fixations. The face,center edge of window and the left arm of the person are the most fixated regions. Balloon , light (at the top) with decorations and the edge of the cupboard in the background are the other fixated regions. The fixation on the balloon is obvious because of presence of face like pattern on it. The selection of regions on cupboard and window may be due to the natural

visual saliency and edge information as discussed previously.

Figure 3.6: Original.

3.3 Results from Fixation time/count

Event Detector software (which is part of SMI vision package [smi(2005)]) was used to generate a file which contained the fixation time, fixation count, blink time and blink count for each image. These parameters were added for all the images and classified for each user as shown in table 6.2(Appendix

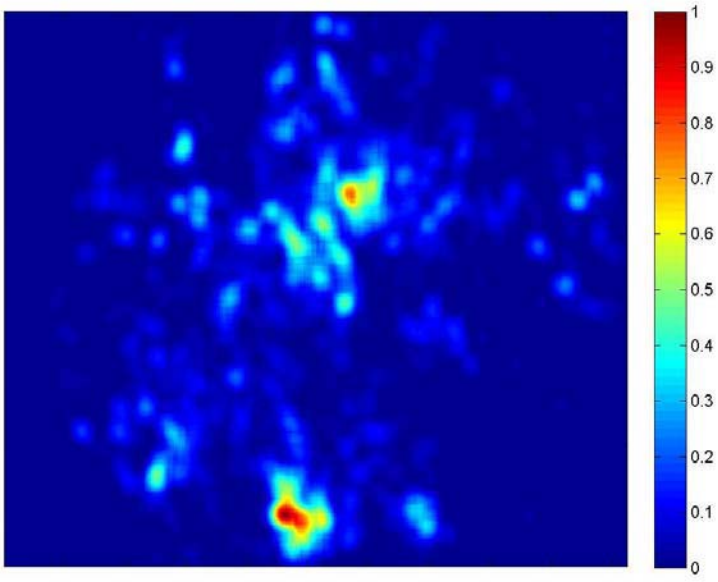


Figure 3.7: Gaze map.

A). The results obtained were averaged and are shown below :

$$Total\ Time\ for\ experiment = 380sec$$

$$Average\ blink\ time = 38.25sec$$

$$Average\ blink\ count = 157.45$$

$$Average\ fixation\ time = 136.365sec$$

$$Average\ fixation\ count = 701.8$$

Apart from this fixations were also counted by thresholding the fixation data by 50% of the maximum value. This was done to get the extent to which faces are fixated after thresholding. The average of the results shown

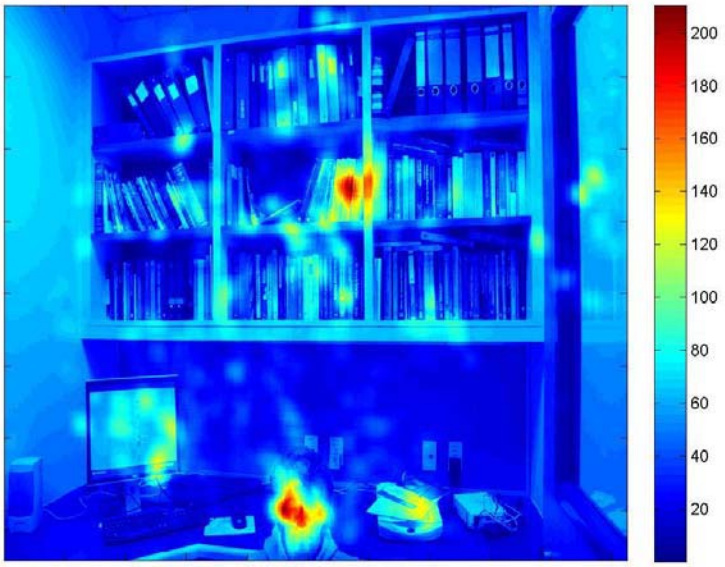


Figure 3.8: Gaze map superimposed over the original image.

in table 6.2(Appendix A) showed that:

$$Average\ number\ of\ fixations\ (with\ threshold) = 540.25$$

$$Average\ number\ of\ fixations\ on\ face\ (with\ threshold) = 151.5$$

So 28% of the total average fixations were on face inspite of repetetion of the same faces in the images. Which emphsizes the importance of face in the gaze maps.

3.4 Results from Data Collection

The gaze maps were collected using the SMI eye tracking device from 20 test subjects who participated in the experiment and data was collected regarding age, level of education, dominant eye and right/left handedness.

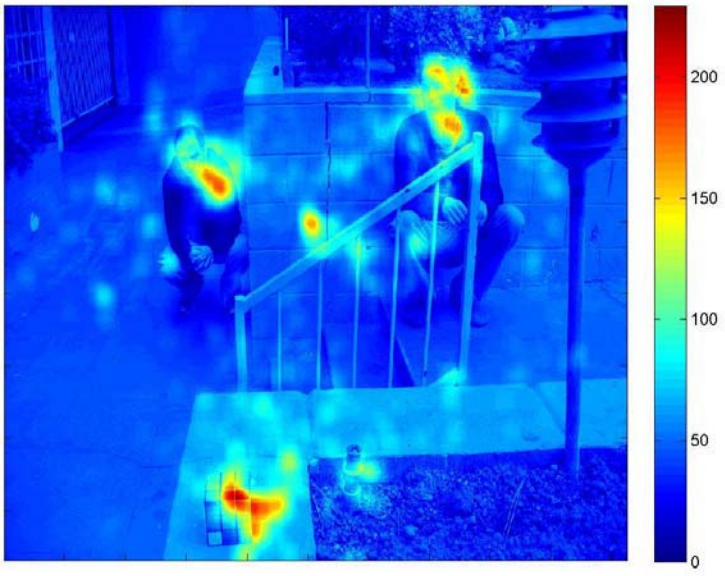


Figure 3.9: Gaze map superimposed over the original image.

The users were classified into age groups of 21-25, 26-30, 31-35, 36-40 and 41-45. As shown in figure 3.12(a) most of the subjects belong to the age group 21-25 and 26-30. 35% of the subjects belonged to age group of 21-25 and 30% belonged to age group of 26-30. 20% of the subjects belonged to age group of 36-40 and rest 15% belonged to age group of 41-45. No subject belonged to age group of 31-35. Figure 3.12(a) shows that 55% of the subjects had studied color science as one of the subjects and rest(45%) had not studied color science at all. Figure 3.12(c) shows that 30% of the subjects had participated in similar eye tracking experiments before. Figure 3.12(d) shows the classification of subjects on the basis of dominant eye and right/left handedness. The results show that 70% right eye dominant subjects were right handed and 20% of the left eye dominant subjects were left handed. Whereas only 5% of the left eye dominant subjects were right

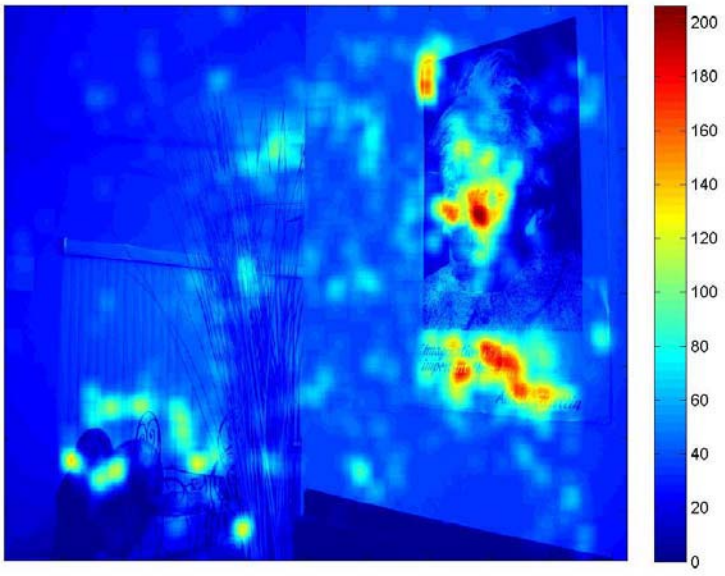


Figure 3.10: Gaze map superimposed over the original image.

handed and 5% of the right eye dominant subjects were left handed.

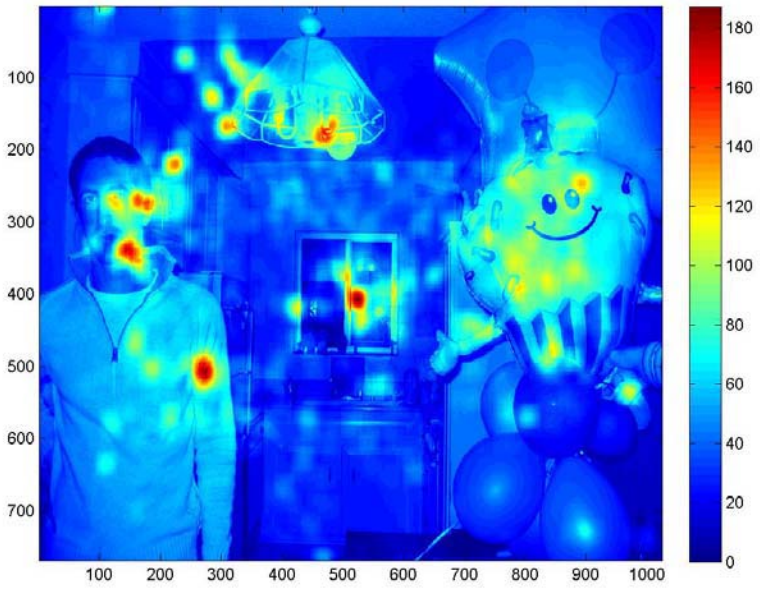
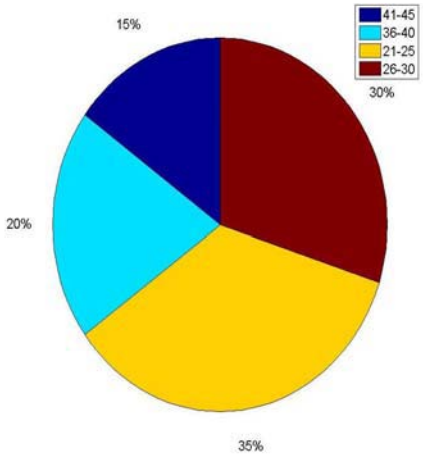
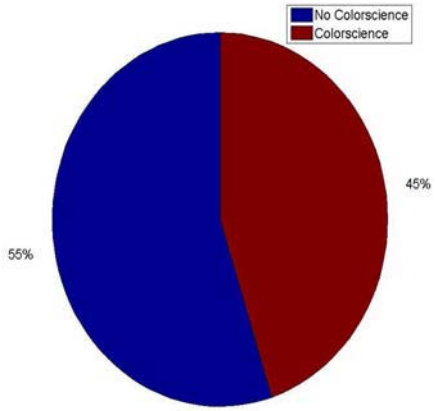


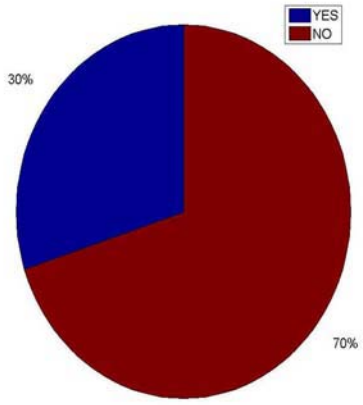
Figure 3.11: Gaze map superimposed over the original image.



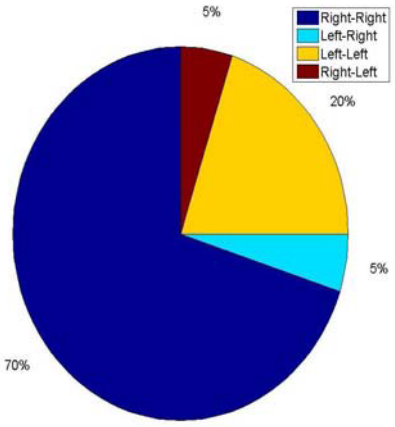
(a) Classification of subjects on the basis of age groups.



(b) Classification on the basis of knowledge of Color science.



(c) Percentage that have participated in eye tracking experiments before.



(d) Classification on the basis of Right/Left eye dominance to Right/Left handedness.

Figure 3.12: Pie Charts.

Chapter 4

Proposed Models

4.1 Modified Saliency Map

As discussed earlier the analysis of gaze map shows that the saliency map algorithm alone cannot closely simulate the human gaze. Therefore a top-down approach of face detection is added to the original algorithm. Moreover most researchers imply that eye movement targeting involves a combination of top-down and bottom-up factors [Tatler *et al.*(2005)Tatler, Baddeley & Gilchrist]. As bottom-up guidance can only account for up to 63% of fixation position data at the best, top-down modulation is likely to be involved. The addition of face detection as one of the channels for conspicuity map takes into account the natural bias of human vision for face patterns in the image. So the final saliency map is formed by the combination of 4 conspicuity maps (as shown in figure 4.1):

- Color conspicuity map
- Intensity conspicuity map
- Orientation conspicuity map
- Face conspicuity map

The saliency map is computed by Winner-Take-All (WTA) network after the feature combination of four conspicuity maps. The feature which generates the highest salient value is considered by the WTA and other values are neglected. Inhibition of return (IOR) helps in computing other

salient regions as it prevents the next salient region from coming into the vicinity of the previous salient region. In this manner we can get a number of salient regions. A set of 7 salient regions were computed for the database [Cerf *et al.*(2007)Cerf, Harel, Einhauser & Koch] of images used for this work. The average fixation duration scene viewing is 330 ms [Henderson(2003)]. Since each image was viewed for 2 seconds only, the number of fixations (for each image) can be predicted by :

$$nof = \frac{2000}{330} = 6.06 \tag{4.1}$$

where nof is number of fixations. As the average fixation duration varies around 330 ms and can be less than, Therefore the number of fixations expected can be more than calculated . So to be on the safer side we take $nof = 7$. The number of salient regions chosen is 7 in accordance with the average number of fixations.

The brain areas that provide guidance for the top-down attentional effects seem to be tightly linked to those areas responsible for the planning and execution of eye movements, which is in agreement with the frequent need to focus on salient regions of the visual environment for a more detailed analysis [Treue(2003)]. A set of 10 gaze maps along with their corresponding saliency maps) were chosen at random from the database. The weight parameters of color, intensity, orientation and face were changed for generating the saliency maps and the correlation was calculated with gaze maps. Various combinations of the weights were tried and the results obtained are shown in Appendix A, table 6.2. In figure 4.2 the trend of correlation values show that increase in the weight of face increases the correlation significantly. But the correlation values saturates after weight is increased beyond 4 so we selected the weight for face as 4. There was no significant change in correlation when weights of color, intensity and orientation were changed. So the weights of

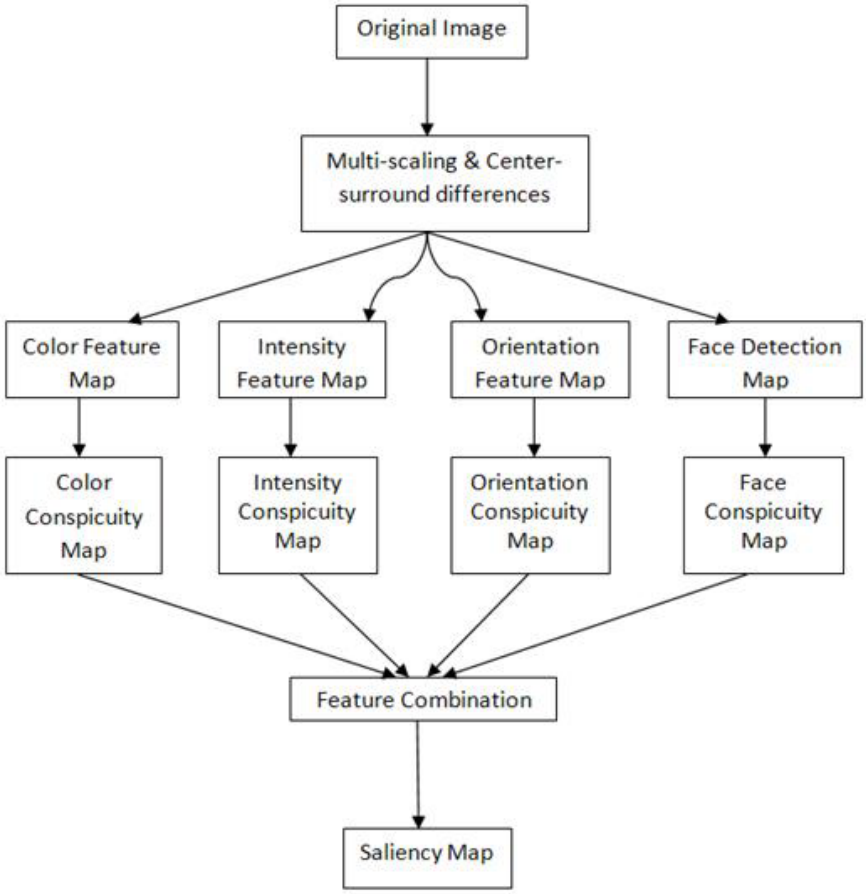


Figure 4.1: Modified Saliency Map.

color, intensity and orientation are taken as unity. The original saliency map is computed as mean of normalized color(C), orientation (O), and intensity(I) maps [Cerf *et al.*(2007)Cerf, Harel, Einhauser & Koch, Itti(2000)].

$$M = 1/3 * (N(I) + N(C) + N(O)) \tag{4.2}$$

where $N(I)$ is normalized intensity map, $N(C)$ is normalized color map, $N(O)$ is normalized orientation map

The modified saliency map is computed as follows :

$$M = 1/7 * (N(I) + N(C) + N(O) + 4 * N(T)) \tag{4.3}$$

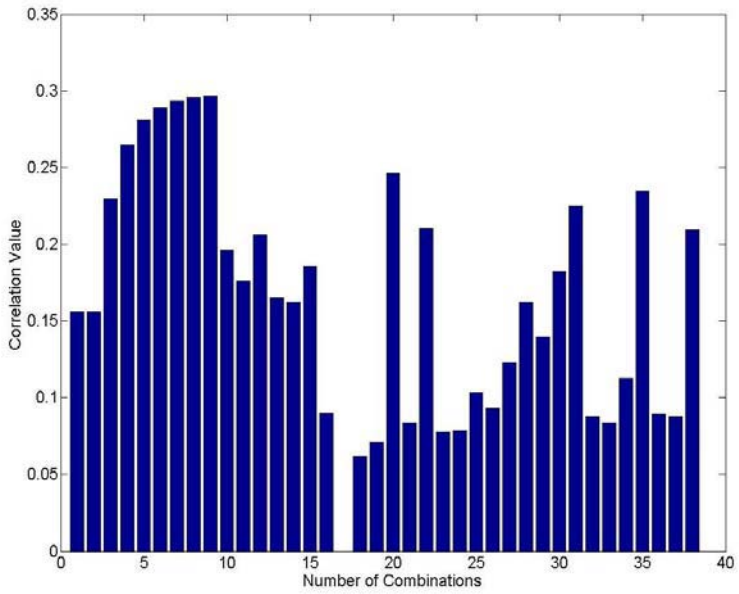


Figure 4.2: Correlation values for different weights of color, intensity, orientation and top down.

where $N(I)$ is normalized intensity map, $N(C)$ is normalized color map, $N(O)$ is normalized orientation map, $N(T)$ is normalized face map

4.2 Face Detection

[FaceonIt()] library was used for face detection procedure. The frontal face detection system [Sauquet *et al.*(2005)Sauquet, Rodriguez & Marcel] was developed at the IDIAP research institute and inspired by the work of Viola and Jones [Viola & Jones(2001)]. The feature selection and classifier training was done using a variant of Ada Boost algorithm [Viola & Jones(2001)]. Each weak classifier depends on a single Modified Census Transform (MCT) feature. They were combined into a cascade to make a strong classifier, so called MCT cascade, which is a weighted combination of weak classifiers. All features were computed on the face training set and non face training

subset. The features chosen were those which minimize the classification error. The final frontal face detector was 4 stage cascade of classifier which included a total of 262 features. Each stage was trained to detect 99.5% of the faces. First classifier in the cascade was constructed using two features and it rejected about 50% of the non faces. The second classifier had 10 features and rejected another 40% of the non faces. The third classifier had 50 features and the last one had 200. The overlapping detections were merged to obtain one final detection per face and only detections which had 60% of their surface as common were kept. Final face detection was the mean of these detections.Modified Census Transform (MCT) is a non parametric local transform that captures the local spatial image structures. It maps the neighborhood of a pixel and the pixel itself into a bit string, comparing the intensity values of the pixels with the intensity mean on the neighborhood of the pixel [Sauquet *et al.*(2005)Sauquet, Rodriguez & Marcel].
The Face detection algorithm [Sauquet *et al.*(2005)Sauquet, Rodriguez & Marcel] is able to detect 189 correct faces out of the total 212 faces in the 190 images. The number of false faces detected are 30, and 18 faces present in the images are not detected at all.

Chapter 5

Comparison of Results

5.1 Comparison of Gaze Maps & Modified Saliency Maps

In this section a detailed comparison of gaze maps and saliency maps is done. Figure 5.1(a) on the left side shows the gaze map superimposed over the original image and figure 5.1(b) shows the corresponding modified saliency map superimposed over the original image. This type of representation helps in the analysis and comparison of the two maps.

Figure 5.1(a) and figure 5.1(b) shows that face, toy truck and toy banana are the common regions on both maps. Where as the gaze map shows that object in the hand of the person in the picture and some parts of the table are fixated by the observers. These regions are not salient according to the modified saliency map algorithm.

Figure 5.1(c) and figure 5.1(d) shows that the two faces, toy dog (upper-left corner) and book (bottom-left corner) are the common regions in both the maps. Whereas the gaze map (figure 5.1(c)) shows a number of fixations on the wall regions, left-top bottom of the shelf and the edge of the shelf which are not salient regions according to the saliency map (figure 5.1(d)). Also the saliency map shows the white fruit basket as salient region which is not fixated in the gaze map. The differences between the gaze maps and saliency maps suggest for changes in the saliency map for further improvement in predicting the gaze maps.

Figure 5.2(a) and figure 5.2(b) shows a number of common regions between the two maps. Face, toy on the sofa, region in the painting on the wall,

laptop, toy car as the common fixated region. Whereas the gaze map (figure 5.2(a)) also shows fixations on the objects on the table like chess, books, toys, lamp top and the edge of the left arm of the person in picture which are not salient according to the saliency map (figure 5.2(b)).

The difference between the gaze and modified saliency map can be attributed to the presence of a large number of objects like mobile phones, magic cube, chess board and chess pieces which may or may not be salient but are strongly attracting early fixation due to our natural bias to familiar objects seen in daily life.

Now when some changes are made in the picture with same background the gaze map changes and so does the modified saliency map. Figure 5.2(c) and figure 5.2(d) show that two faces, toy banana, region in the painting on the wall are the common regions in both the maps. Whereas the gaze map (figure 5.2(c)) shows the toy car and the hand of the person as fixated regions which are not salient according to the modified saliency map (figure 5.2(d)).

Figure 5.3(a) and figure 5.3(b) show that face, white bag as common regions in both maps. Whereas the gaze map(figure 5.3(a)) shows fixations on objects like bottle, vegetable and in a region on the ground, which are not salient regions according to the modified saliency map (figure 5.3(b)). Also the saliency map shows the right arm of the person in the picture, the white shoe and a region on the background as salient regions which are not fixated in the gaze map.

Figure 5.3(c) and figure 5.3(d) show that two faces, toy banana, reflection in window at the center of the image as common regions in both the maps. Whereas the gaze map (figure 5.3(c)) shows fixations around the toy banana and the number (on the top of image) and address board (on the left side of image near the face), which are not salient regions according to the modified

saliency map (figure 5.3(d)).

The difference in the gaze and modified saliency map can be attributed due to the bias of the human vision towards words and alphabets while viewing images. That is the primary reason that there were more fixations around the text and number patterns for the given image as observed by gaze map (figure 5.3(c)).

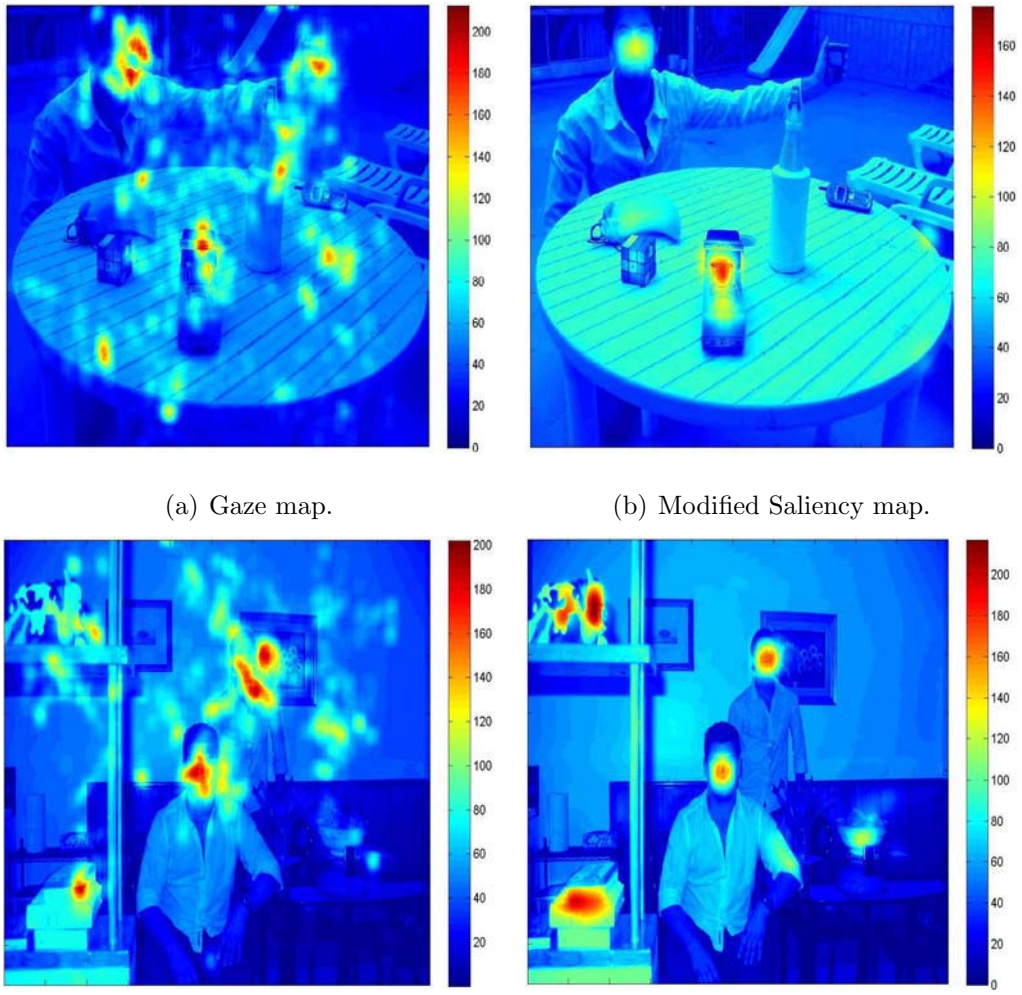
Figure 5.4(a) and figure 5.4(b) show that two faces, center of face (Einstein in the poster) are the common regions in both maps. The gaze map(figure 5.4(a)) shows fixations on top-right corner of the poster of Einstein, text written on the poster and curtain edge (near the left side of face) which are not salient according to the saliency map (figure 5.4(b))

The difference in the maps is again due to the bias of the human vision to text appearing on the posters and the top-right corner, curtain edge (near the left side of face) due to the presence of edge information [Henderson(2003), Tatler *et al.*(2005)Tatler, Baddeley & Gilchrist]. As investigators have found that high spatial frequency content and edge density are somewhat greater at fixation sites [Henderson(2003), Tatler *et al.*(2005)Tatler, Baddeley & Gilchrist].

Figure 5.4(c) and figure 5.4(d) show the common regions for an image without faces in it. Lamp top and region in the poster of Einstein are common in both the maps. Again the gaze map (figure 5.4(c)) shows more fixation around the text of the poster.

5.1.1 Receiver Operating Characteristic(ROC)

An Receiver Operating Characteristic (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their performance [Fawcett(2004)]. ROC curves provide a visual tool for examining the trade off between the ability of a classifier to correctly identify positive cases and the number of

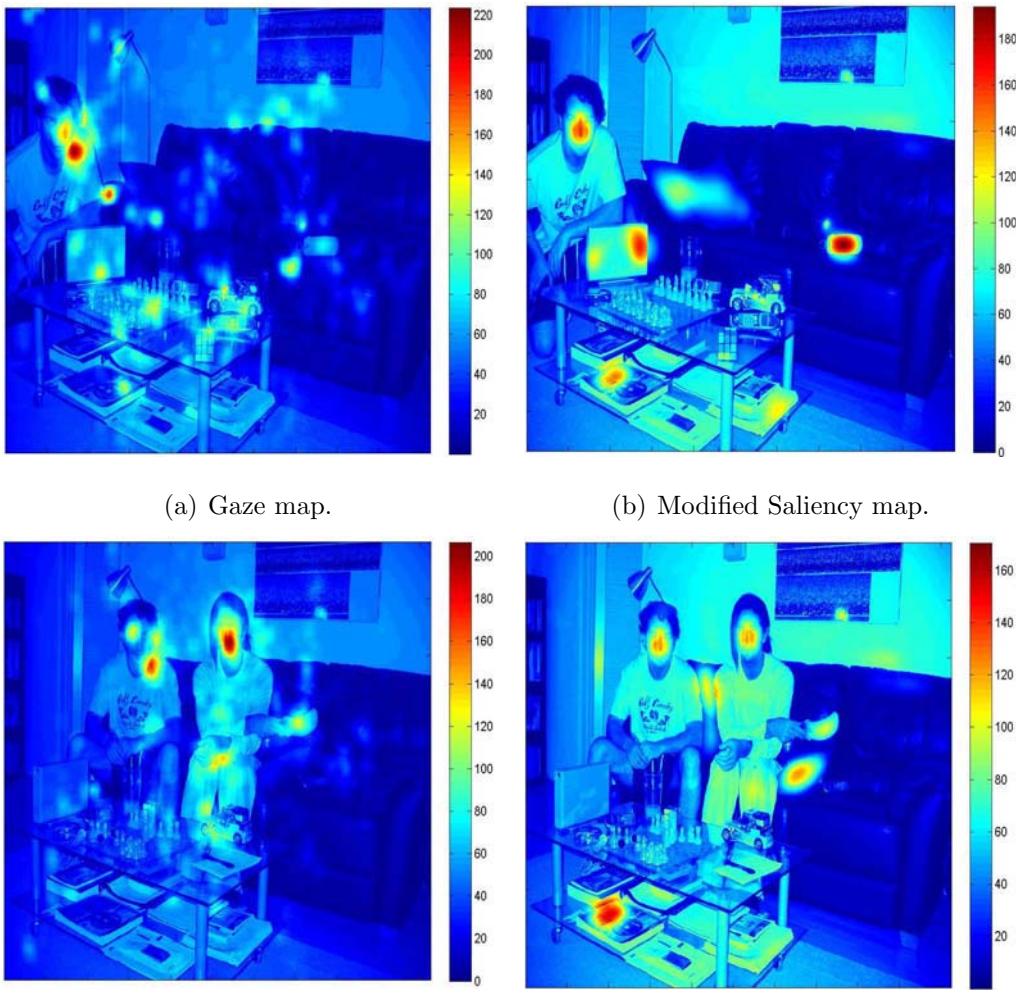


(a) Gaze map. (b) Modified Saliency map.

(c) Gaze map. (d) Modified Saliency map.

Figure 5.1: Comparison of gaze map with modified saliency map.

negative cases that are incorrectly classified. ROC graphs are commonly used in medical decision making, and in recent years have been used increasingly in machine learning and data mining research [Fawcett(2006)]. We begin by considering classification problems using only two classes. Formally, each instance I is mapped to one element of the set {p,n} of positive and negative class labels. A classification model (or classifier) is a mapping from instances to predicted classes [Fawcett(2006)]. To distinguish between the actual class



(a) Gaze map.

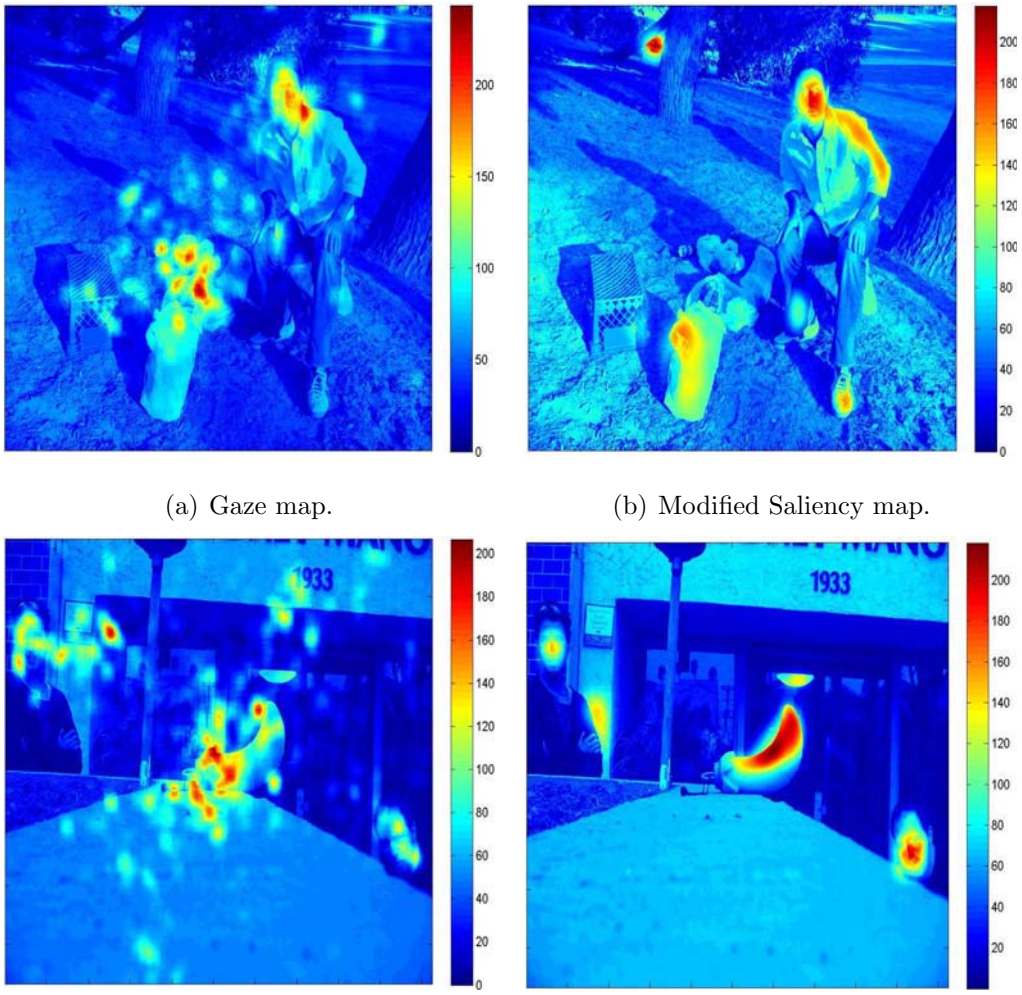
(b) Modified Saliency map.

(c) Gaze map.

(d) Modified Saliency map.

Figure 5.2: Comparison of gaze map with modified saliency map.

and the predicted class we use the labels {Y,N} for the class predictions produced by a model. Given a classifier and an instance, there are four possible outcomes. If the instance is positive and it is classified as positive, it is counted as a true positive; if it is classified as negative, it is counted as a false negative. If the instance is negative and it is classified as negative, it is counted as a true negative; if it is classified as positive, it is counted as a false positive. Given a classifier and a set of instances (the test set), a



(a) Gaze map.

(b) Modified Saliency map.

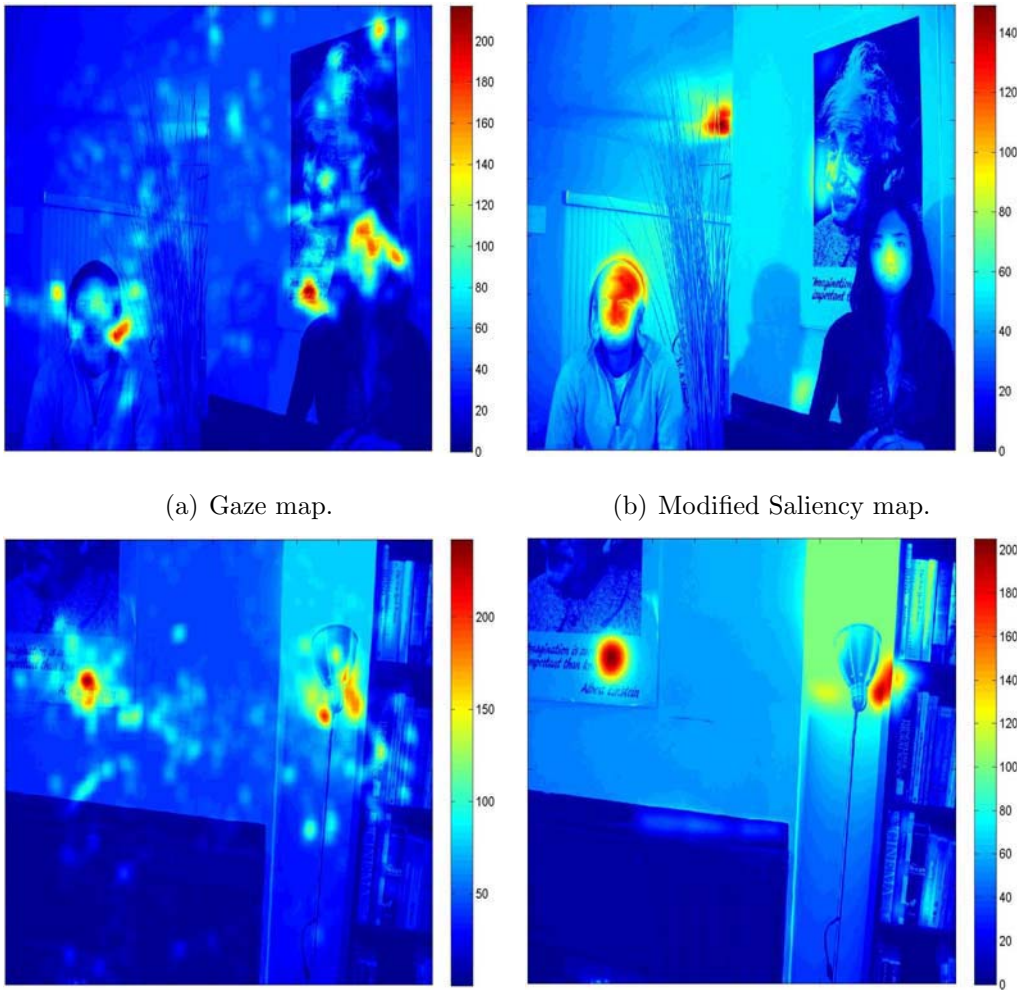
(c) Gaze map.

(d) Modified Saliency map.

Figure 5.3: Comparison of gaze map with modified saliency map.

two-by-two confusion matrix constructed representing the dispositions of the set of instances as shown in figure 5.5 [Fawcett(2004), Fawcett(2006)]. The true positive rate (also called hit rate and recall) of a classifier is estimated as :

$$\text{tp rate} = \text{positives correctly classified/total positives} \tag{5.1}$$



(a) Gaze map.

(b) Modified Saliency map.

(c) Gaze map.

(d) Modified Saliency map.

Figure 5.4: Comparison of gaze map with modified saliency map.

The false positive rate (also called false alarm rate) of the classifier is :

$$\text{fp rate} = \text{negatives incorrectly classified/total negatives} \tag{5.2}$$

ROC graphs are two-dimensional graphs in which True Positive (TP) rate is plotted on the Y axis and False Positive (FP) rate is plotted on the X axis. An ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positives). A point on ROC curve that is nearest to northwest

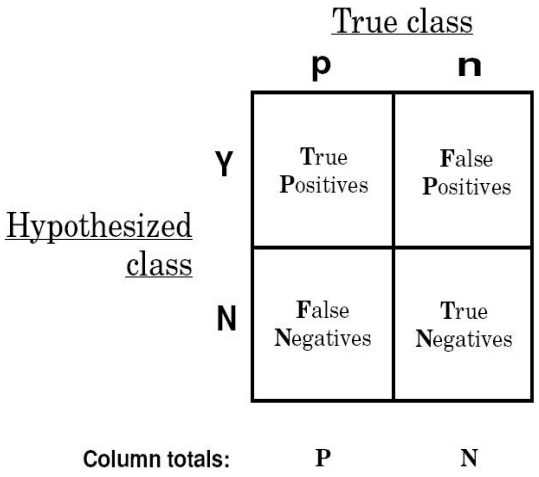


Figure 5.5: Confusion matrix.

point is considered a better classifier due to its high true positive rate and low false positive rate (as shown in figure 5.6 the performance of D is perfect. Whereas the point A is considered to be more conservative than B because it positive classification only with strong evidence that it makes few false positive errors but having a low true positive rate as well. The classifier like C is said to be liberal, because it makes positive classification with a weak evidence and classifies positives correctly but has high false positive rate as well. A random classifier will produce a ROC point that slides back and forth on the diagonal based on the frequency with which it guesses the positive class. Any classifier that appears in the lower right triangle performs worse than random guessing as shown in figure 5.6, E performs much worse than random guessing [Fawcett(2004)].

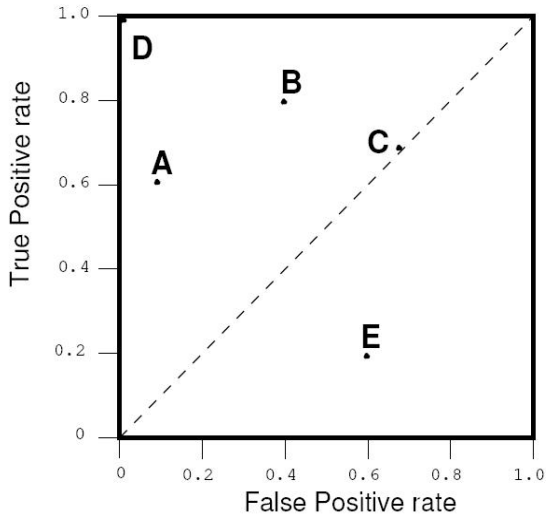


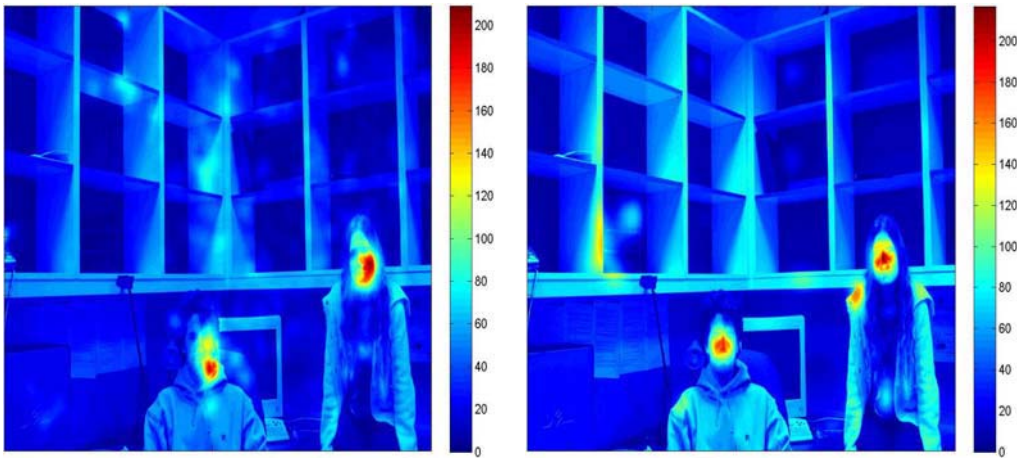
Figure 5.6: A basic ROC graph

5.1.2 Area under the ROC Curve(AUC)

An ROC curve is a two-dimensional depiction of classifier performance. To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0. However, because random guessing produces the diagonal line between (0, 0) and (1, 1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5. The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [Fawcett(2004), Fawcett(2006)].

Figure 5.7 shows the gaze map, modified saliency map, original image and the ROC graph. The ROC graph (5.7(d)) shows a good classifier result and peak performance at (0.1,0.68) which is above the conservative region, and

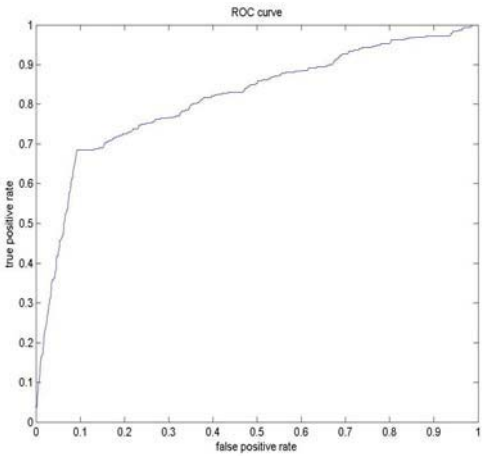
near to the ideal region. The AUC for the ROC graph comes out to be 0.8150 which means that its performance is much better than random guessing.



(a) Gaze map.

(b) Modified Saliency map.

(c) Original image.



(d) ROC Graph (Area under the ROC Curve (AUC) = 0.8150).

Figure 5.7: Gaze map, Modified saliency map, Original image, ROC graph.

The scatter plot (as shown in figure 5.8) shows the performance of AUC for Saliency map (SM) and Modified Saliency map (MSM). The AUC for SM is represented in green diamond shape and AUC for MSM is represented as blue circles. The X-axis represents the image number and the Y-axis

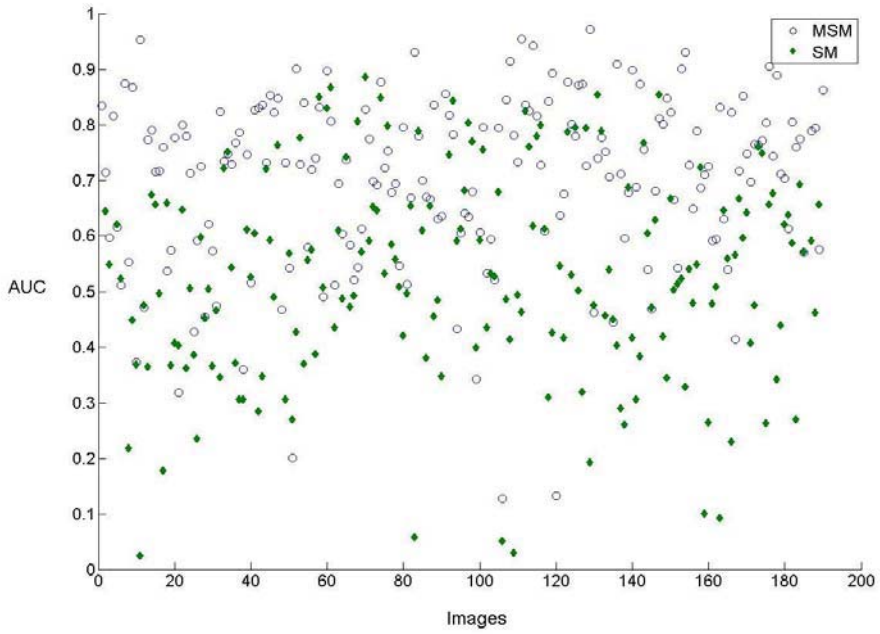


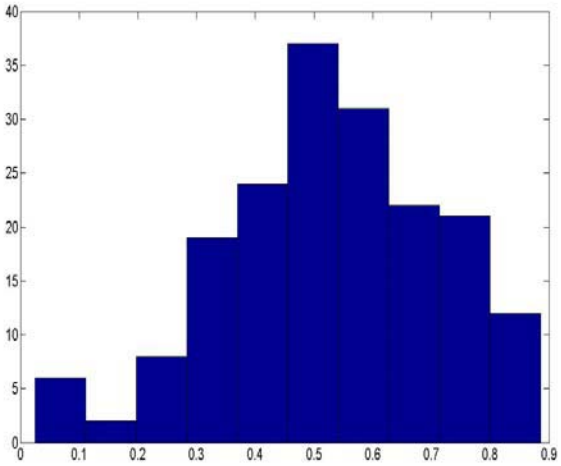
Figure 5.8: Scatter plot of AUC(MSM) and AUC(SM).

represents its corresponding AUC values for SM and MSM.
The comparison of histogram of AUC for SM and MSM (in figure 5.9) shows that MSM performs better than SM as for SM most of the values lie around 0.5 (figure 5.9(a)) whereas for MSM most of the values lie around 0.7-0.8 (figure 5.9(b)).

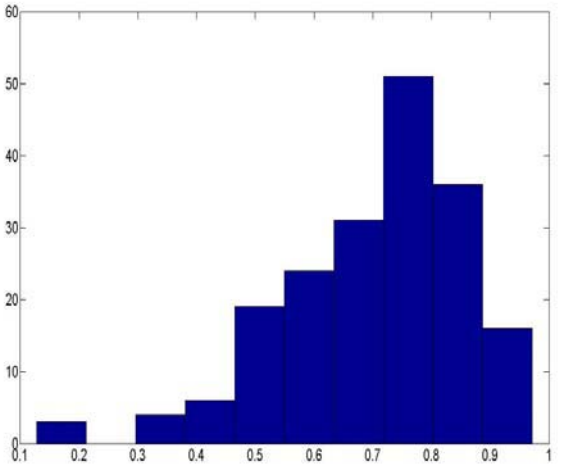
$$\text{Mean AUC for Saliency maps} = 0.5281$$

$$\text{Mean AUC for Modified Saliency maps} = 0.7035$$

Clearly these results indicate that the performance of Modified Saliency map is better than Saliency map. The addition of the top down (Face detection) is responsible for improving the performance of the saliency map. This establishes the importance of the role played by face detection in improving



(a) Histogram for Saliency map.



(b) Histogram for Modified Saliency map.

Figure 5.9: Comparison of histogram of AUC for Saliency map and Modified Saliency map.

the traditional saliency map for the prediction of gaze maps.

5.2 Comparison of Gaze Maps & Modified GAFFE Maps

GAFFE (Gaze Attentive Fixation Finding Engine) produces a true prediction map as combination of four low level image parameters as shown below:

$$M = 1/4 * (N(L) + N(C) + N(BL) + N(BC)) \tag{5.3}$$

where $N(L)$ is normalized luminance map, $N(C)$ is normalized contrast map, $N(BL)$ is normalized bandpass luminance map and $N(BC)$ is normalized bandpass contrast. The maximum values of the true prediction map are selected as the predicted fixation points by GAFFE. So GAFFE gives us the coordinates of fixation points and true prediction map. In order to obtain the saliency map for 7 fixations the 35 pixel radius was drawn around fixation points and the values of true prediction map around these regions were selected. The circular regions were selected with mid point centering around the fixation points. All values outside these circular regions were assigned as zero. In this manner saliency map consisting of 7 salient regions was formed for each image of the database.

The GAFFE algorithm was modified by addition of face detection procedure. So the true prediction map was obtained as combination of five features as shown below :

$$M = 1/7 * (N(L) + N(C) + N(BL) + N(BC) + 3 * N(FD)) \tag{5.4}$$

where $N(L)$ is normalized luminance map, $N(C)$ is normalized contrast map, $N(BL)$ is normalized bandpass luminance map, $N(BC)$ is normalized bandpass contrast and $N(FD)$ is normalized face detection map. The face detection map was assigned different weights on hit and trail basis. Assigning the weight more than 3 leads to saturation of the algorithm around the face regions only. Thus the weight of face detection map was assigned a value of

3.

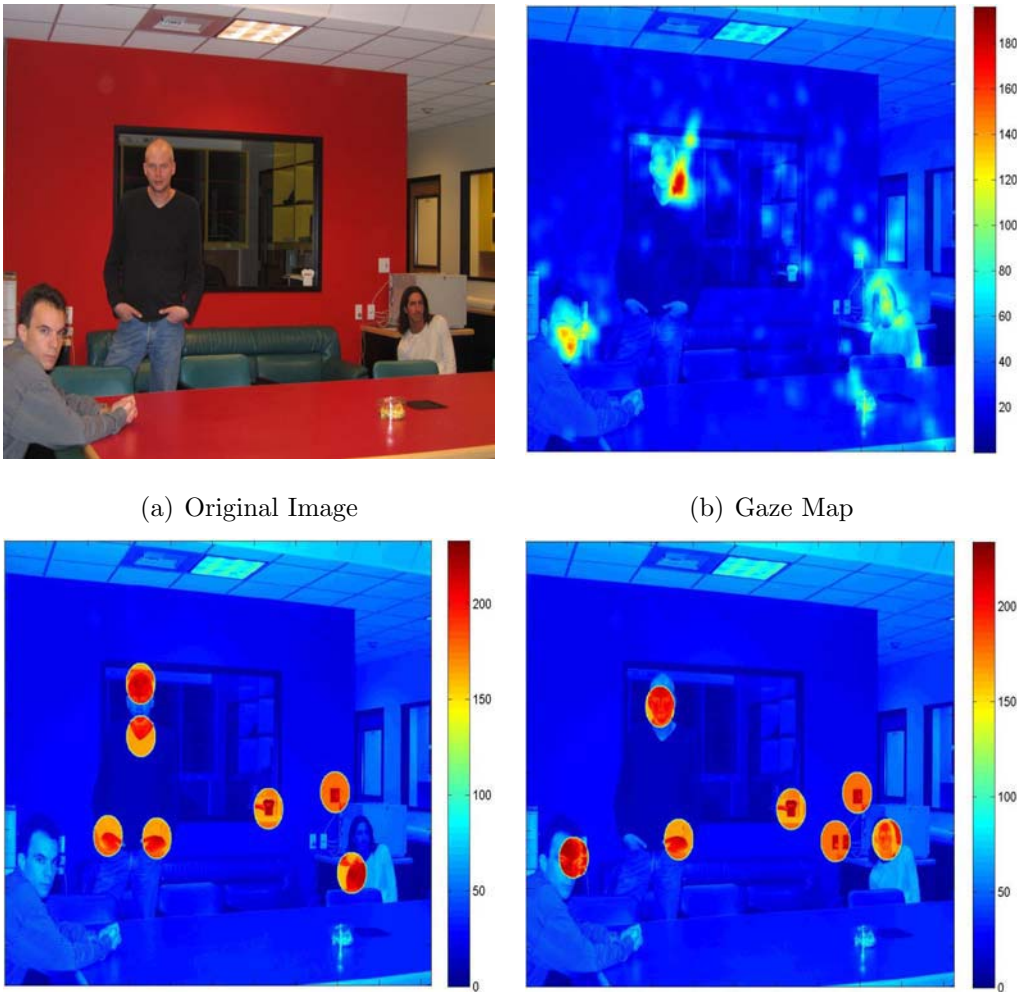
Gaze map (figure 5.10(b)) obtained for the image shown in (figure 5.10(a) depicts fixations around the three faces in the image followed by fixations on the edge of window at the center of the image, electrical outlet and ashtray on the table. Figure 5.10 clearly shows that Modified GAFFE Map(MGM)(figure 5.10(d)) gives better performance as compared to GAFFE Map(GM)(figure 5.10(c)). All three face regions are computed by the MGM, along with the the region around electrical outlet. As MGM is more closer to gaze map as compared to GM so MGM is better for comparison with Gaze map.

Figure 5.11(a) shows the fixations around the face regions, edge of the poster, text of the poster, curtains and some other regions in the Gaze map. Figure 5.11(b) shows the MGM. The common regions between the two maps are two faces, text on the poster and fixation at the edge of the wall near the center.

Figure 5.12(a) shows the gaze map with fixations around the face, pc monitor, lamp, wall and various other regions. Figure 5.12(b) shows the MGM. The common regions between the two maps are face, fixations around the pc monitor and objects on the tables.

Figure 5.13 shows the Gaze Map(figure 5.13(a)) for the image shown in figure 5.13(c). MGM (figure 5.13(b)) shows the predicted fixations around face, white shirt and on the magazine cover which are also common to the gaze map. The ROC Graph (in figure 5.13(d)) shows that the performance of the classifier is better than random as the Area under the ROC curve comes out to be 0.7962. The ROC graph shows that the performance of the classifier for a given image is better than conservative but lesser than ideal.

Figure 5.14 shows the comparison of histograms for GM and MGM. Histogram of MGM (in figure 5.14(b)) shows that on the average the value

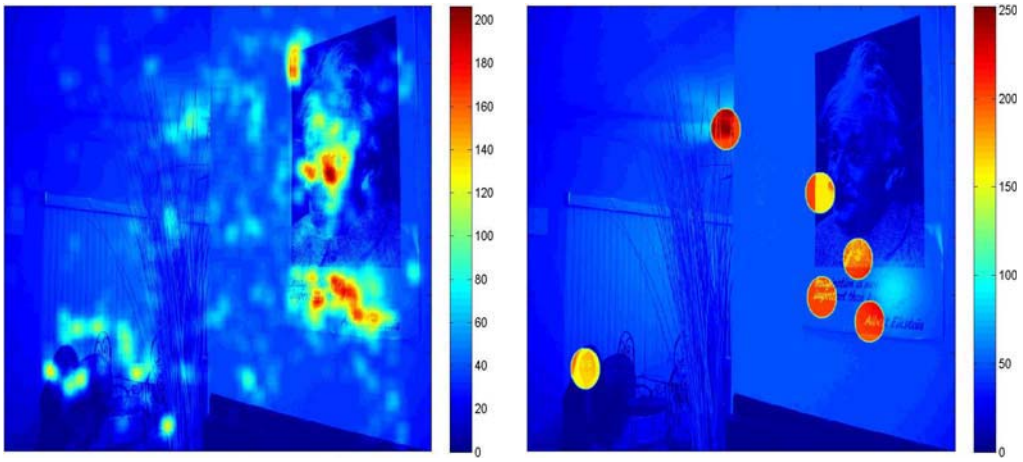


(a) Original Image (b) Gaze Map

(c) GM (d) MGM

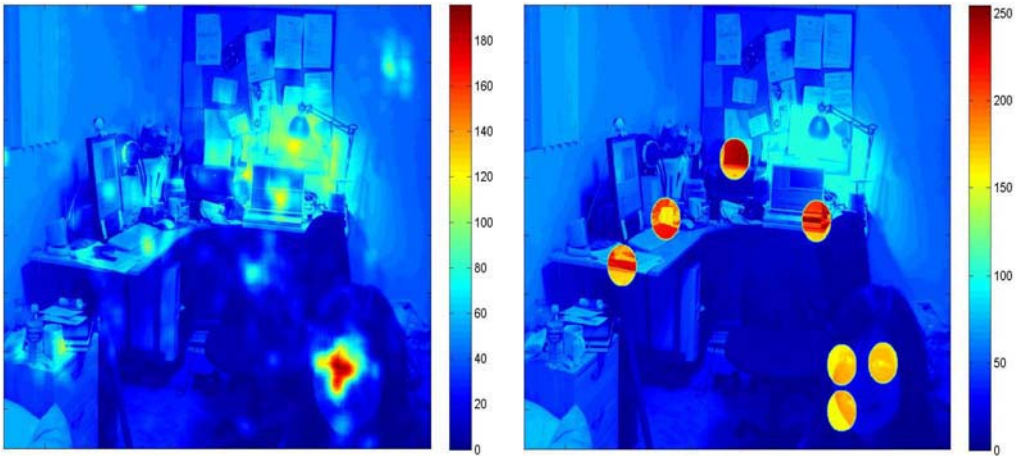
Figure 5.10: Comparison of GAFFE Map with Modified GAFFE Map and Gaze Map.

of AUC lies in the range between 0.7 to 0.8. Thus the performance of classifier is much better than random. But for the GM the average values lie around



(a) Gaze Map (b) MGM

Figure 5.11: Comparison of Gaze Map with Modified GAFFE Map.



(a) Gaze Map (b) MGM

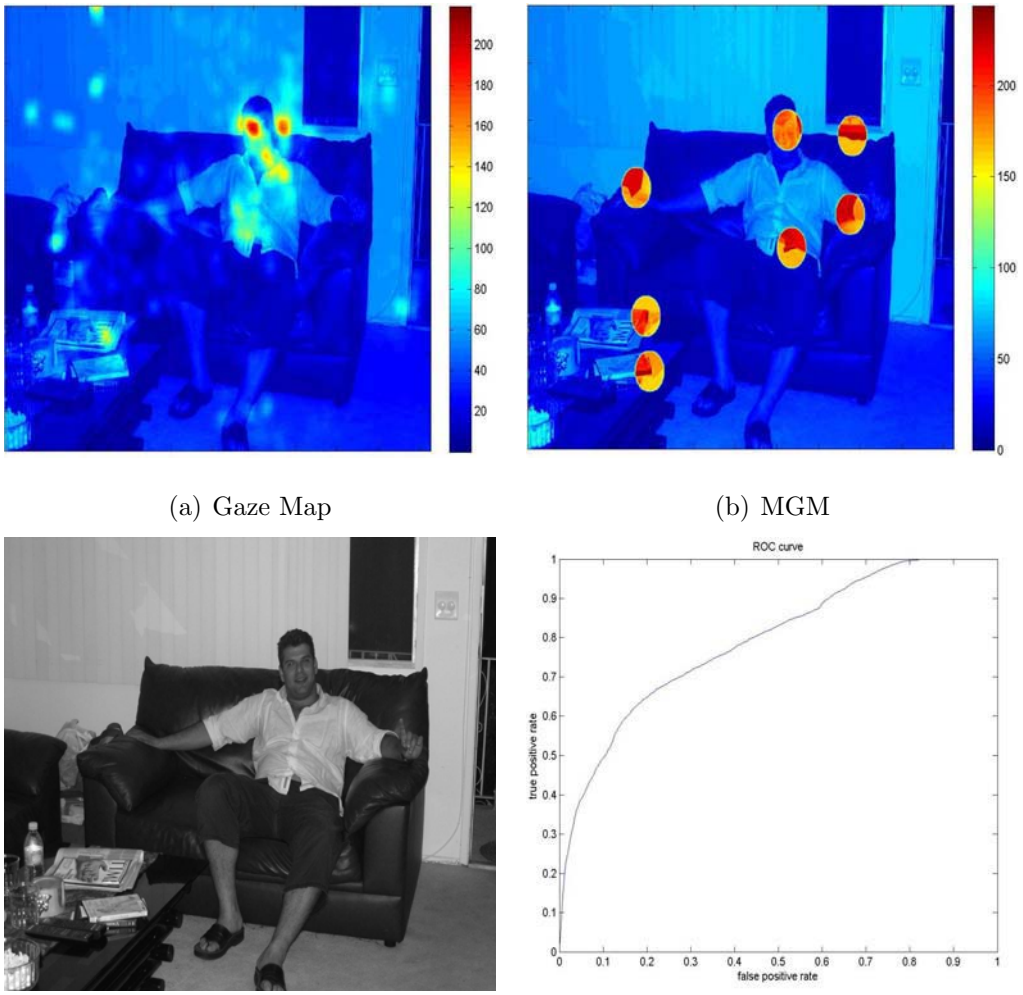
Figure 5.12: Comparison of Gaze Map with Modified GAFFE Map.

0.4 to 0.6.

$$\text{Mean AUC for GAFFE maps} = 0.4587$$

$$\text{Mean AUC for Modified GAFFE maps} = 0.7003$$

Since the Mean AUC for GAFFE is less than 0.5 this means that the perfor-



(a) Gaze Map

(b) MGM

(c) Original Image

(d) ROC Graph (Area under the ROC Curve (AUC = 0.7926))

Figure 5.13: Original Image, Gaze Map, Modified GAFFE Map and ROC Graph.

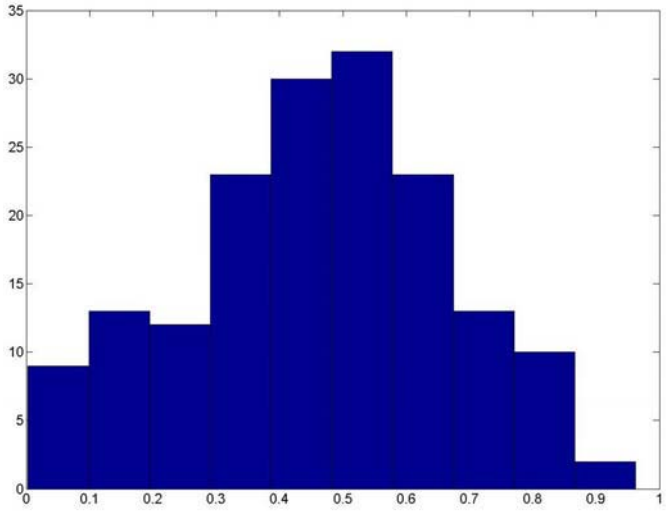
mance of the classifier is worse than random. It is clear from the mean values of GM and MGM that the addition of the face detection procedure improves the performance of this algorithm significantly. Figure 5.15 shows the Scatter plot for AUC of Saliency map and GM along y-axis and the image numbers on the x-axis. The blue circle represents the AUC for

Saliency map and red diamond represents the AUC for GM corresponding to each image. 23 images gave significantly higher values of AUC for GM as compared to AUC for Saliency map. Since the GM is formed by combination of four features i.e. luminance map, contrast map, bandpass luminance map and bandpass contrast map. It was observed that bandpass luminance and bandpass contrast contributed towards fixations around the face and other salient objects in the GM and it was due to the prediction of these regions that GM for these images were more correlated with gaze maps. It may be due to modeling of Gabor kernel for spatial frequencies of human patches that differed significantly from random patches [Rajashekar *et al.*(2008)Rajashekar, van der Linde, Bovik & Cormack].

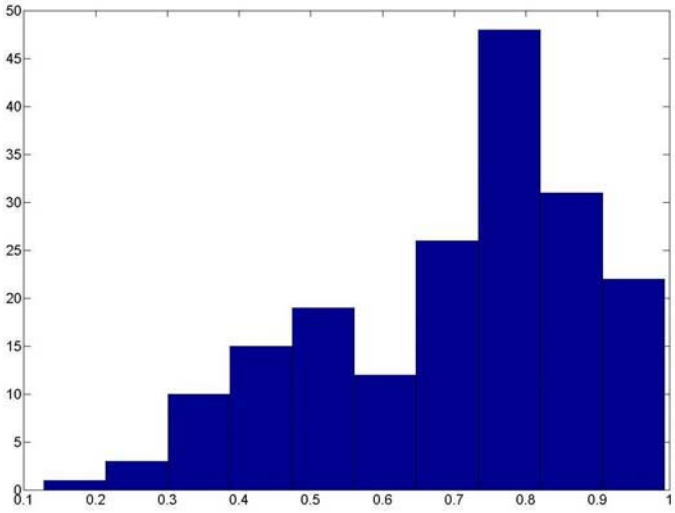
5.3 Comparison of Modified Saliency Maps & Modified GAFFE Maps

Figure 5.16 shows the scatter plot for AUC of MSM and MGM. It is clear from the figure that for 98 images the value for AUC of MGM is slightly greater than AUC of MSM. 12 images have significantly higher values AUC of MGM as compared to AUC of MSM. The higher values of AUC for MGM is due to the higher values of face region in the final map obtained after the combination of luminance, contrast, bandpass luminance, bandpass contrast and face maps. Since the gaze maps have higher fixation values around the face regions so the classifier gives slightly better results for some images in case of MGM. MGM is obtained after the operations on fixation coordinates of the images hence their direct comparison with MSM does not implies that MSM is better than MGM or vice versa. Moreover the MGM does not operates on the color component of the image whereas MSM does.

Figure 5.17 shows the similarities and differences in the MSM and MGM.



(a) Histogram for GM



(b) Histogram for MGM

Figure 5.14: Scatter plot for AUC of GM and MGM.

Figure 5.17(c) shows the most salient regions as face, objects on the top left corner of the image. Figure 5.17(b) shows gaze map with fixations around face region, objects on the top left corner of the image and some fixations on the door at the center of the image. Figure 5.17(d) shows the MGM with

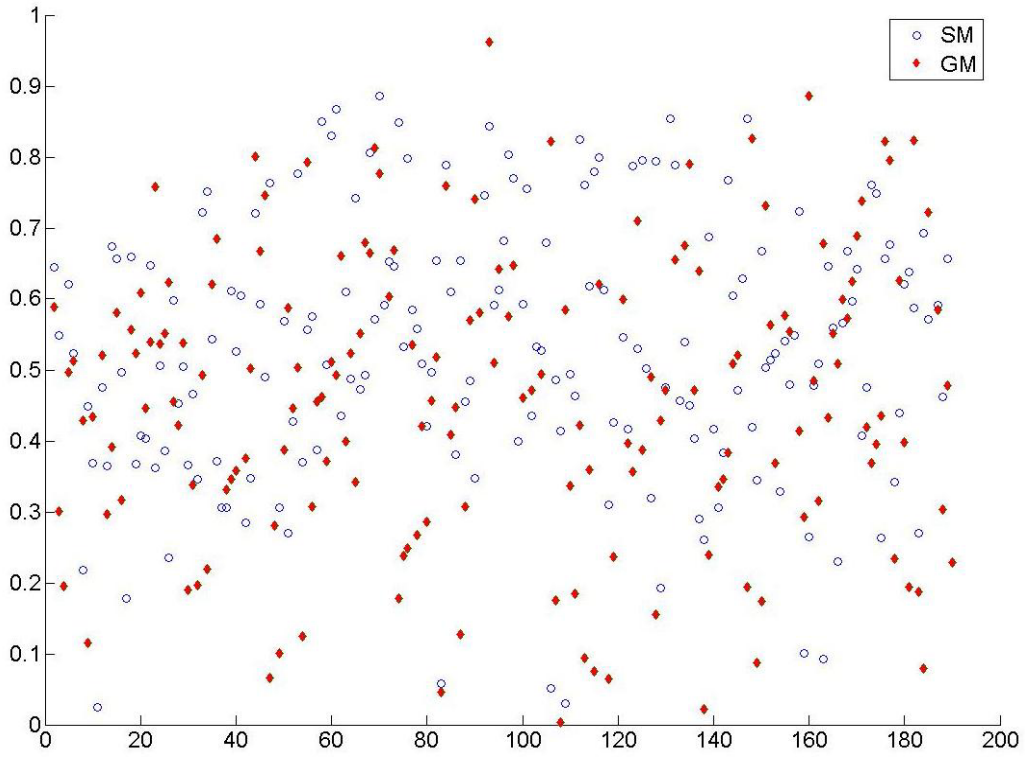


Figure 5.15: Scatter plot of AUC(SM) and AUC(GM).

salient regions as face, object on the top left corner, regions on the book cabinet and door at the center of the image. The ROC Graph values for the image shown in figure 5.17(a) is given below :

Area under the ROC Graph for MSM = 0.9042

Area under the ROC Graph for MGM = 0.9383

The ROC value for MGM is slightly higher than MSM because of larger salient region on the top left corner of the image by the MGM as compared to MSM. But they select the same two common salient regions, one is the face and other is the white object on the top left corner of the image. Gaze map (figure 5.17(b)) also shows fixations around these two regions. Thus

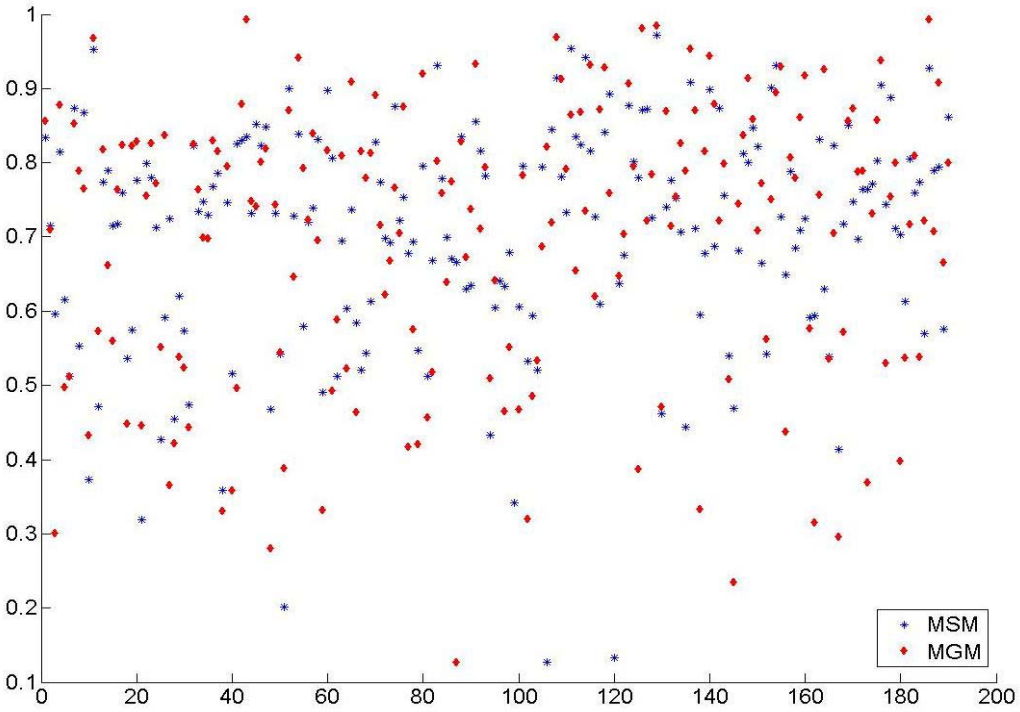
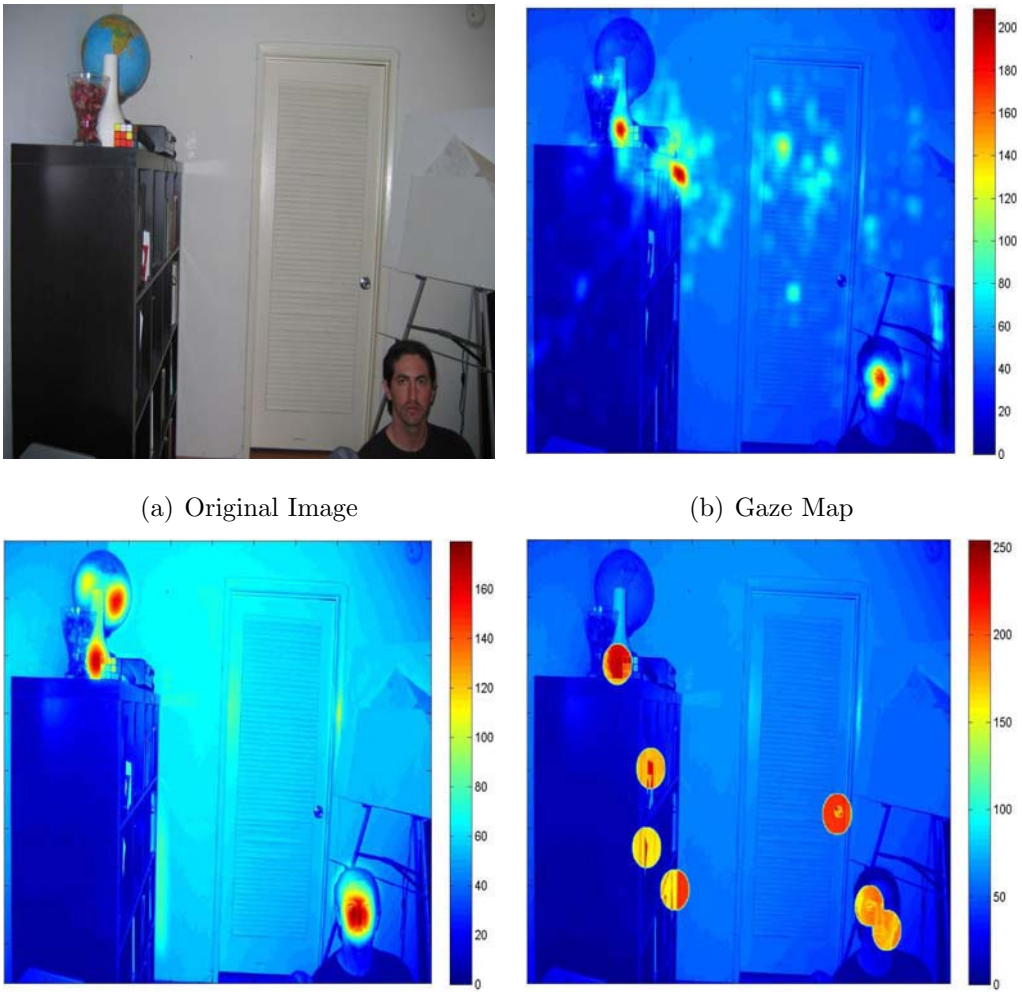


Figure 5.16: Scatter plot for AUC of MSM and MGM

accounting to near ideal values of AUC for both MSM and MGM.



(a) Original Image (b) Gaze Map

(c) MSM (d) MGM

Figure 5.17: Original Image, Gaze Map, Modified Saliency Map and Modified GAFFE Map.

Chapter 6

Conclusions & Future Work

6.1 Conclusions

Understanding visual attention is a challenging task due to the complexity of human visual perception. The analysis of saliency maps shows the role played by low level image features. Intensity, color and orientation maps play a key role in the selection of the salient regions in an image. As saliency map selects the object which is differentiated from the other objects in terms of its higher intensity, color and orientation. By assuming that attention is drawn to the changes in the environment of the image. The order in which the scene should be inspected is determined by the saliency map. Inhibition of return suppresses the most salient regions in the image thus giving us options to look for a number of other salient regions. The conclusions that can be derived from the results are as follows :

- It is apparently clear from the comparison of MSM and gaze maps (in section 5.1.2) that the MSM gives us higher mean value of 0.7035 for AUC, than the SM which gives us the mean value of 0.5281 for AUC. Thus the MSM is able to predict the gaze map more closely as compared to SM. As MSM is formed after the addition of face detection (as one of the conspicuity maps) into the traditional SM. This results clearly suggest better prediction of gaze maps by MSM. This portrays the importance of top down bias of human vision for face images.

- Objects seen in daily life like mobile phones, computers, books, bal-

loons, pc monitor, scanner, water sprinkler,chess board, toy car, toy banana etc. may contribute to early visual fixation as analyzed from the results of gaze map (as shown in section 3.2). Under normal conditions also we tend to fixate more often on known objects which may or may not be salient. The concept of saliency must be further broadened to include other top down approaches (possibly familiar objects detection).

- The observations (in section 3.2, 5.1) emphasize the importance of text, alphabets on posters, paintings, sign boards, white boards etc. in getting early fixation of the subjects. Thus they portray the top down bias of human vision for text patterns under natural viewing conditions.

- The observations (in section 5.2) show that the performance of GAFFE is slightly lower than that of Saliency map. As the mean AUC for GAFFE is 0.4587 whereas the mean AUC for Saliency map is 0.5281. So Saliency map is better for prediction of gaze maps. The addition of face detection in GAFFE improved the performance of the algorithm significantly. This again emphasizes the role of top-down factors like face detection in visual saliency.

6.2 Future Work

Human visual perception is biased by familiar objects seen in daily life. Objects like books, balloons, clothes, boxes, toys etc. that we see in our daily life tend to attract early attention in the image. Under normal conditions also we tend to fixate on known objects rather than variations in the image. Therefore familiar object recognition procedure is needed in addition to saliency maps for simulating the human visual perception. The future work

involves training the model for detection of these objects in the saliency map. The performance of the model will be highly efficient if the model can identify such objects of interest in the image which the human eyes tend to fixate.

Bibliography

[smi(2005)] (2005). *iViewX system manual.* SMI (SensoMotoric Instruments), 1.05.09 ed.

[Atchison & Smith(2000)] Atchison, D. A. & Smith, G. (2000). *Optics of the Human Eye.* Elsevier Health Sciences, 1 ed.

[Burian & Noorden(1985)] Burian, H. M. & Noorden, G. K. V. (1985). *Binocular Vision and Ocular Motility : Theory and Management of Strabismus.* St. Louis : Mosby, 3 ed.

[Cerf *et al.*(2007)Cerf, Harel, Einhauser & Koch] Cerf, M., Harel, J., Einhauser, W. & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. In: *Advances in Neural Information Processing Systems (NIPS 2007)* (Platt, J., Koller, D., Singer, Y. & Roweis, S., eds.). Cambridge, MA.

[Crane(1996)] Crane, R. (1996). *A Simplified Approach to Image Processing (Hewlett-Packard Professional Books.* Prentice Hall PTR.

[FaceonIt()] FaceonIt (). Face on it library. URL `http://www.faceonit.ch/`.

[Fawcett(2004)] Fawcett, T. (2004). Roc graphs: Notes and practical considerations for researchers. HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto, CA 94304: Kluwer Academic Publishers, pp. 1–38.

[Fawcett(2006)] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters, Elsevier* (27), 861–874.

[Fecteau & Munoz(2006)] Fecteau, J. H. & Munoz, D. P. (2006). Salience, relevence, and firing: a priority map for target selection. *Trends in Cognitive Sciences, Elsevier* **10**, 382–390.

[Harel *et al.*(2006)Harel, Koch & Perona] Harel, J., Koch, C. & Perona, P. (2006). Graph-based visual saliency. In: *Advances in Neural Information Processing Systems (NIPS 2006).*

[Henderson(2003)] Henderson, J. M. (2003). Human gaze control during real-world scene perception. *TRENDS in Cognitive Sciences, Elsevier* **7**(11), 498–504.

[Itti(2000)] Itti, L. (2000). *Models of Bottom-Up and Top-Down Visual Attention.* Ph.D. thesis, California Institute of Technology, Pasadena, California.

[Itti & Koch(2001)] Itti, L. & Koch, C. (2001). Computational modelling of visual attention. *Neuroscience 2001* **2(3)**, 194–203.

[Jost *et al.*(2005)Jost, Ouerhani, Wartburg, Muri & Hugli] Jost, T., Ouerhani, N., Wartburg, R. V., Muri, R. & Hugli, H. (2005). Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding, Elsevier* **100**, 107–123.

[Koch & Ullman(1985)] Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology* **4**, 219–227.

[Kominkova *et al.*(2008)Kominkova, Pedersen, Hardeberg & Kaplanova] Kominkova, B., Pedersen, M., Hardeberg, J. Y. & Kaplanova, M. (2008). Comparison of eye tracking devices used

on printed images. In: *Human Vision and Electronic Imaging XIII.Proceedings of the SPIE*, vol. 6806.

[Meur *et al.*(2006)Meur, Callet, Barba & Thoreau] Meur, O. L., Callet, P. L., Barba, D. & Thoreau, D. (2006). A cohorent computational approach to model bottom-up visual attention. *IEEE Transcations on Pattern Analysis and Machine Intelligence* **28**, 802–817.

[Oishia *et al.*(2005)Oishia, Tobimatsub, Arakawaa, Taniwakia & ichi Kira] Oishia, A., Tobimatsub, S., Arakawaa, K., Taniwakia, T. & ichi Kira, J. (2005). Ocular dominancy in conjugate eye movements at reading distance. *Neuroscience Research, Elsevier* **52**(3), 263–268.

[Pedersen(2007)] Pedersen, M. (2007). *Importance of region-of-interest on image difference metrics*. Master's thesis, Department of Computer Science and Media Technology, Gjovik University College.

[Rajashekar *et al.*(2008)Rajashekar, van der Linde, Bovik & Cormack] Rajashekar, U., van der Linde, I., Bovik, A. C. & Cormack, L. K. (2008). Gaffe: A gaze-attentive fixation finding engine. *IEEE TRANSACTIONS ON IMAGE PROCESSING* **17**(4), 564–573.

[Rajashekara *et al.*(2007)Rajashekara, van der Lindea, Bovika & Cormack] Rajashekara, U., van der Lindea, I., Bovika, A. C. & Cormack, L. K. (2007). Foveated analysis of image features at fixations. *Vision Research, Elsevier* **47**, 3160–3172.

[Rao *et al.*(2002)Rao, Zelinsky, Hayhoe & Ballard] Rao, R. P., Zelinsky, G. J., Hayhoe, M. M. & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research, Elsevier* (42), 1447–1463.

[Roth *et al.*(2002)Roth, Lora & Heilman] Roth, H. L., Lora, A. N. & Heilman, K. L. (2002). Effects of monocular viewing and eye dominance on spatial attention. *Brain* **125(9)**, 2023–2035.

[Salvucci & Goldberg(2000)] Salvucci, D. D. & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In: *Tracking Research and Applications Symposium.*

[Sauquet *et al.*(2005)Sauquet, Rodriguez & Marcel] Sauquet, T., Rodriguez, Y. & Marcel, S. (2005). Multiview face detection. IDIAP-RR 49, IDIAP.

[Schubiger *et al.*(1991)Schubiger, Moser, Hacisalihzade & Muller] Schubiger, R., Moser, J., Hacisalihzade, S. & Muller, M. (1991). Machine vision based on human prception and eye movements. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 13.

[Tatler *et al.*(2005)Tatler, Baddeley & Gilchrist] Tatler, B. W., Baddeley, R. J. & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Reserach, Elsevier* (45), 643–659.

[Tong(2003)] Tong, F. (2003). Primary visual cortex and visual awareness. *Nuroscience* **4**, 219–229.

[Treue(2003)] Treue, S. (2003). Visual attention: the where, what, how and why of saliency. *Current Opinion in Neurobiology, Elsevier* (13), 428–432.

[Viola & Jones(2001)] Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In: *IEEE Com-*

puter Society Conference on Computer Vision and Pattern Recognition (CVPR'01).

[Walther(2006)] Walther, D. (2006). *Interactions of Visual attention and Object Recognition : Computational Modeling, Algorithms and Psychophysics*. Ph.D. thesis, California Institute of Technology, Pasadena, California.

[Ware(2004)] Ware, C. (2004). *Information Visualization - Perception for Design*. Morgan Kaufmann, Elsevier.

[Zhaoping(2005)] Zhaoping, L. (2005). *Neurobiology of Attention*. 93. Elsevier.

Appendix A

Subject ID	Total Fixations(with thresholding)	Fixations on Face(with thresholding)	Age	Sex	Dominant Eye(R/L	Right/Left Handed	Color Science Knowledge	Education	Participation in eye tracking before?	Seen the images before?	Fixation Count	Fixation Time (in second)	Blink Count	Blink Time (in second)
TS102	500	115	25	M	R	R	Y	PhD Stud.	N	N	851	140.5	18	5.1
TS103	512	122	29	M	R	R	Y	PhD Stud.	N	N	918	197.2	18	3.2
TS104	559	262	30	M	R	R	Y	PhD	Y	N	819	157.4	93	28.7
TS106	571	283	24	F	R	R	Y	BS Stud.	N	N	821	178.5	78	16.4
TS107	563	161	26	M	L	R	Y	MS Stud.	N	N	396	56.9	330	94.5
TS108	647	183	23	M	R	R	N	BS Stud.	N	N	427	57.9	241	77.8
TS109	467	101	23	M	R	R	N	MS Stud.	N	N	482	76.9	330	62.8
TS110	457	127	24	M	L	R	N	BS Stud.	Y	Y	871	126.0	35	8.5
TS111	535	152	24	M	R	L	Y	MS Stud.	N	N	842	162.1	172	29.2
TS112	450	138	27	M	L	R	N	MS Stud.	Y	N	598	183.5	130	27.8
TS113	633	124	45	M	R	R	Y	BS Stud.	N	N	918	197.2	18	3.2
TS114	499	114	27	M	R	R	Y	MS Stud.	N	N	576	134.7	167	59.6
TS116	555	146	36	M	L	L	Y	PhD	N	N	730	112.2	297	64.1
TS118	465	108	40	M	L	R	Y	PhD	Y	Y	480	113.3	199	55.1
TS119	624	169	26	M	R	R	Y	PhD Stud.	Y	N	860	191.3	64	16.3
TS120	516	71	24	M	R	R	N	BS Stud.	N	N	365	55.5	147	45.1
TS122	656	181	39	M	R	R	N	PhD Stud.	N	N	1001	190.4	18	2.9
TS123	516	159	45	M	R	R	N	PhD	Y	Y	810	186.9	54	13.0
TS124	619	160	39	M	R	R	N	PhD Stud.	N	N	803	117.7	292	58.4
TS125	461	154	30	M	R	R	N	PhD	N	N	468	91.2	448	93.3

Table 6.1: Fixation data for all the test subjects.

Correlation	Color	Intensity	Orientation	Face
0.1563	1	1	1	0
0.1560	1	1	1	1
0.2293	1	1	1	2
0.2647	1	1	1	3
0.2811	1	1	1	4
0.2893	1	1	1	5
0.2935	1	1	1	6
0.2957	1	1	1	7
0.2968	1	1	1	8
0.1963	2	1	1	2
0.1760	1	2	1	2
0.2059	1	1	2	2
0.1650	2	2	1	2
0.1623	1	2	2	2
0.1855	2	1	2	2
0.0899	0	0	1	0
N.A.	0	0	0	1
0.0617	1	0	0	0
0.0710	0	1	0	0
0.2462	0	0	1	1
0.0839	0	1	1	0
0.2103	1	0	0	1
0.0779	1	0	1	0
0.0787	1	1	0	0
0.1033	3	1	1	1
0.0930	1	3	1	1
0.1231	1	1	3	1
0.1623	3	1	1	2
0.1397	1	3	1	2
0.1822	1	1	3	2
0.2254	1	1	3	3
0.0879	4	1	1	1
0.0838	1	4	1	1
0.1126	1	1	4	1
0.2344	1	1	4	4
0.0892	1	4	4	1
0.0879	4	4	1	1
0.2091	4	1	1	4

Table 6.2: Correlation with different weights of color, intensity, orientation and face.

Appendix B

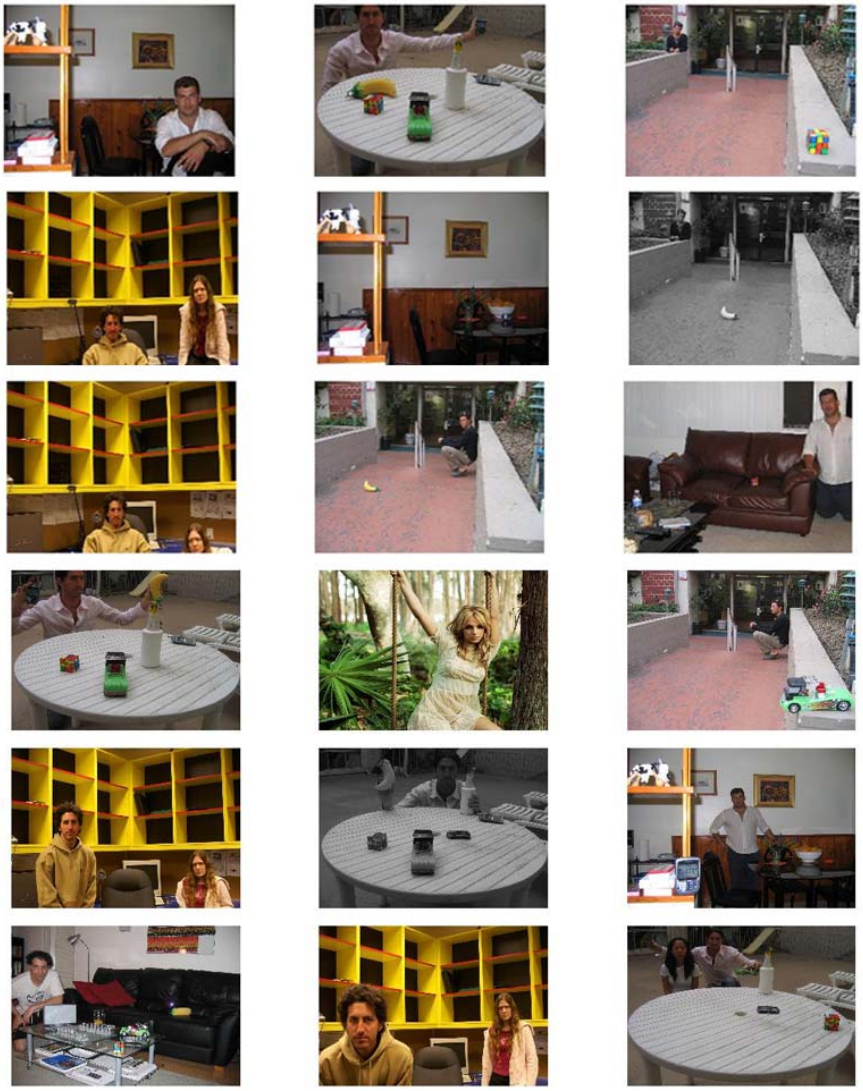


Figure 6.1: Thumbnail view of database of images.

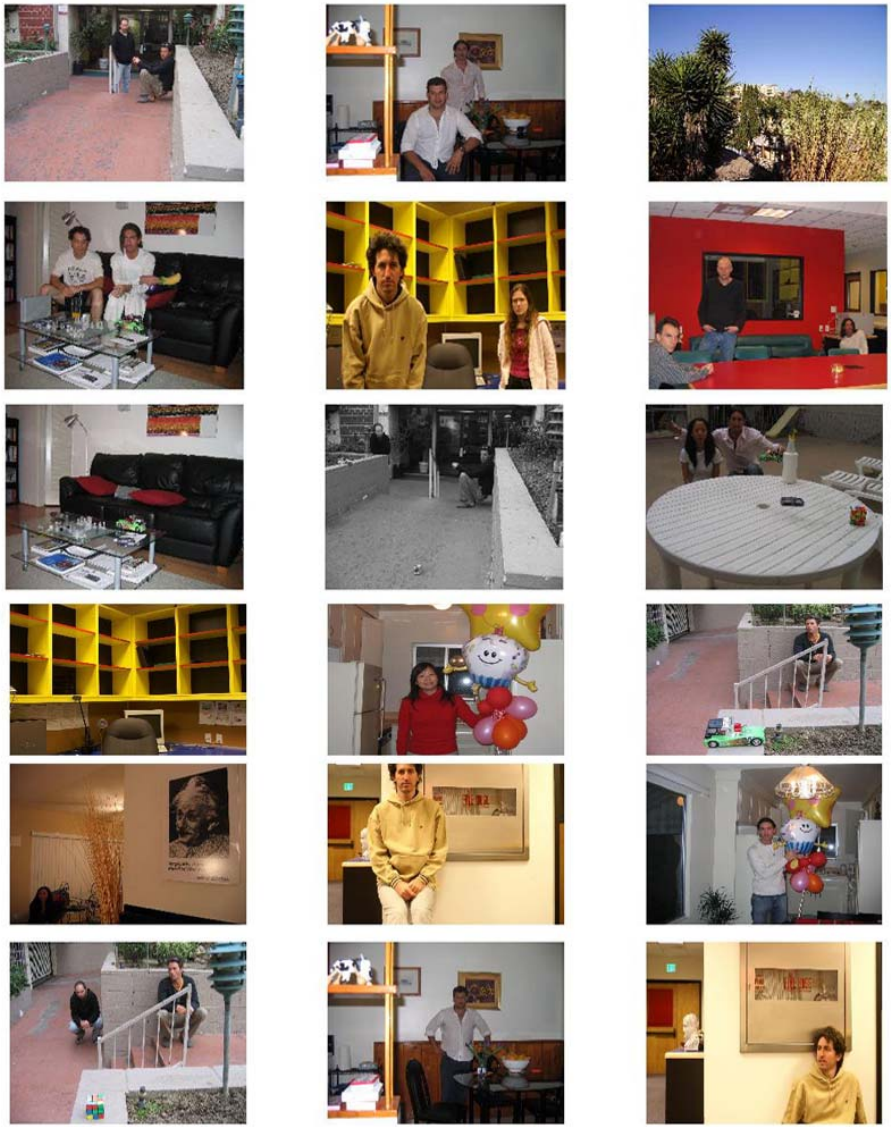


Figure 6.2: Thumbnail view of database of images.

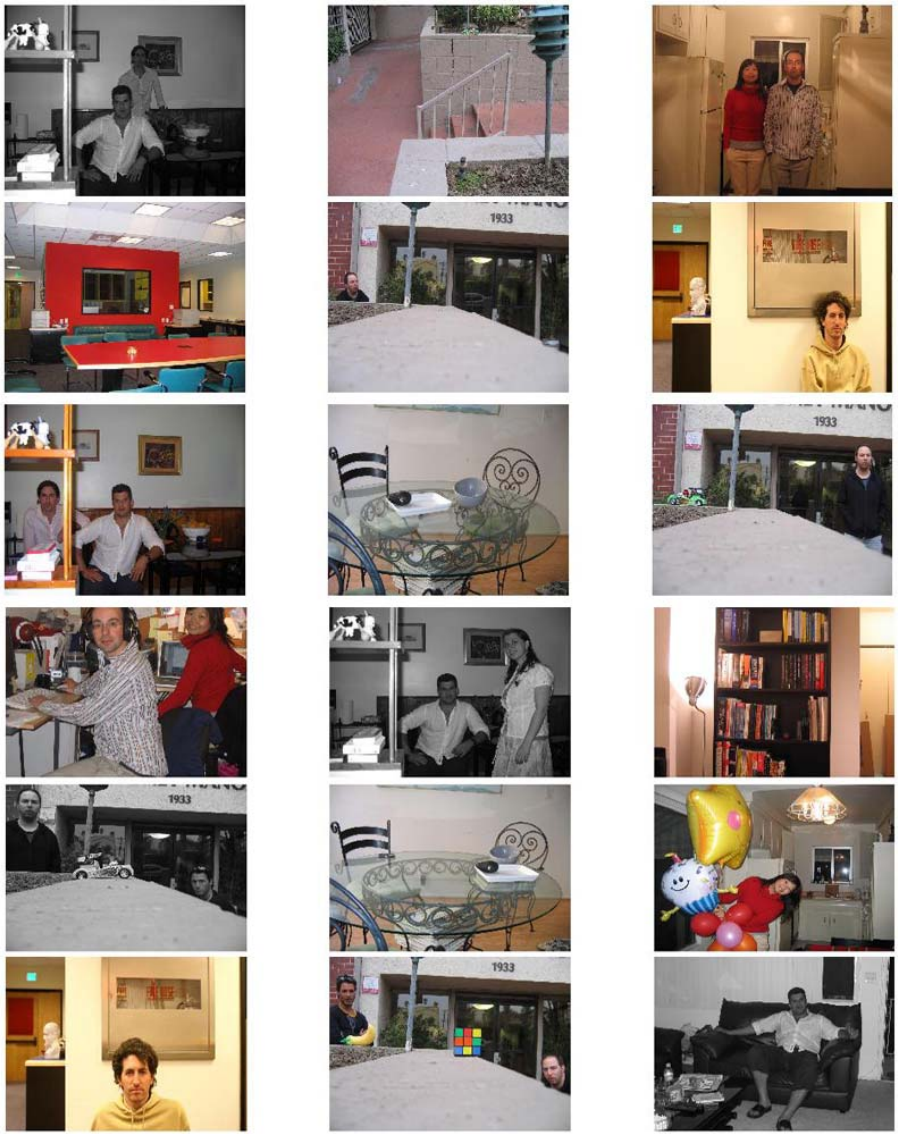


Figure 6.3: Thumbnail view of database of images.

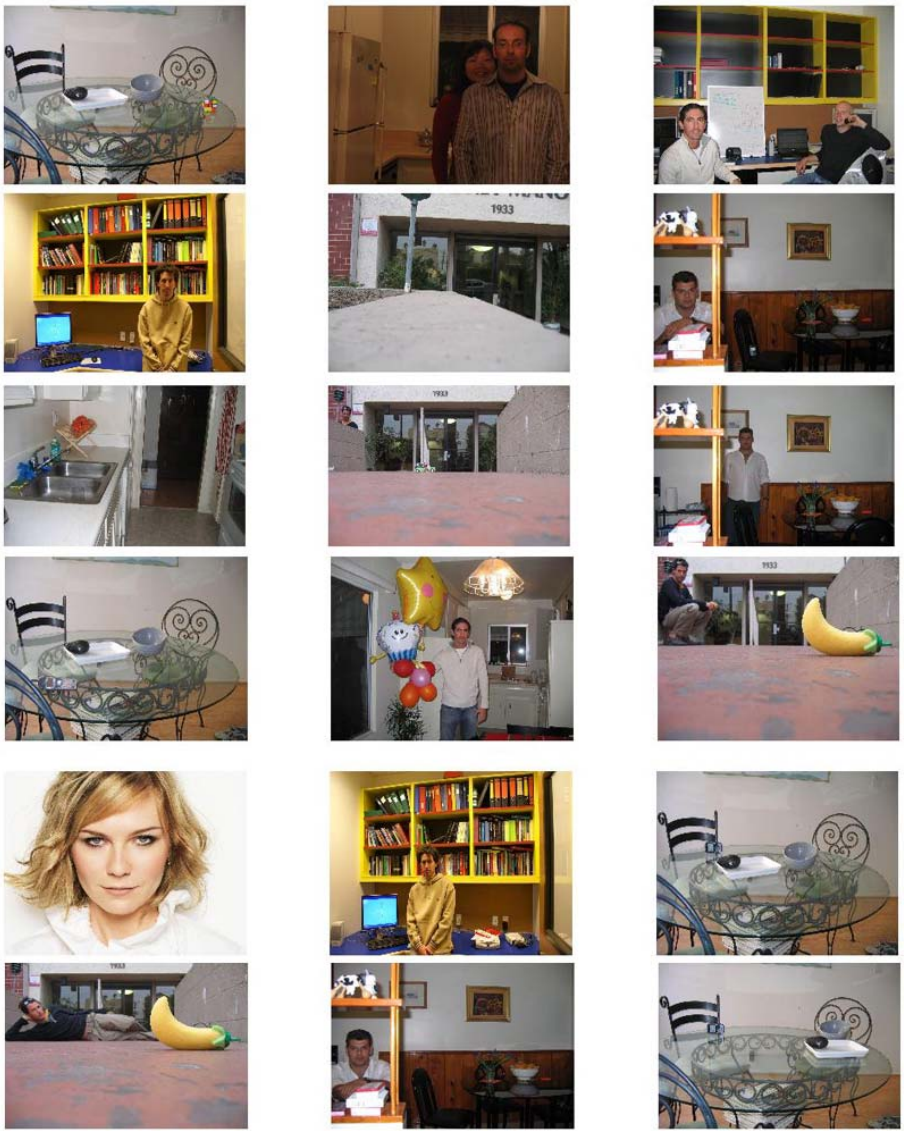


Figure 6.4: Thumbnail view of database of images.

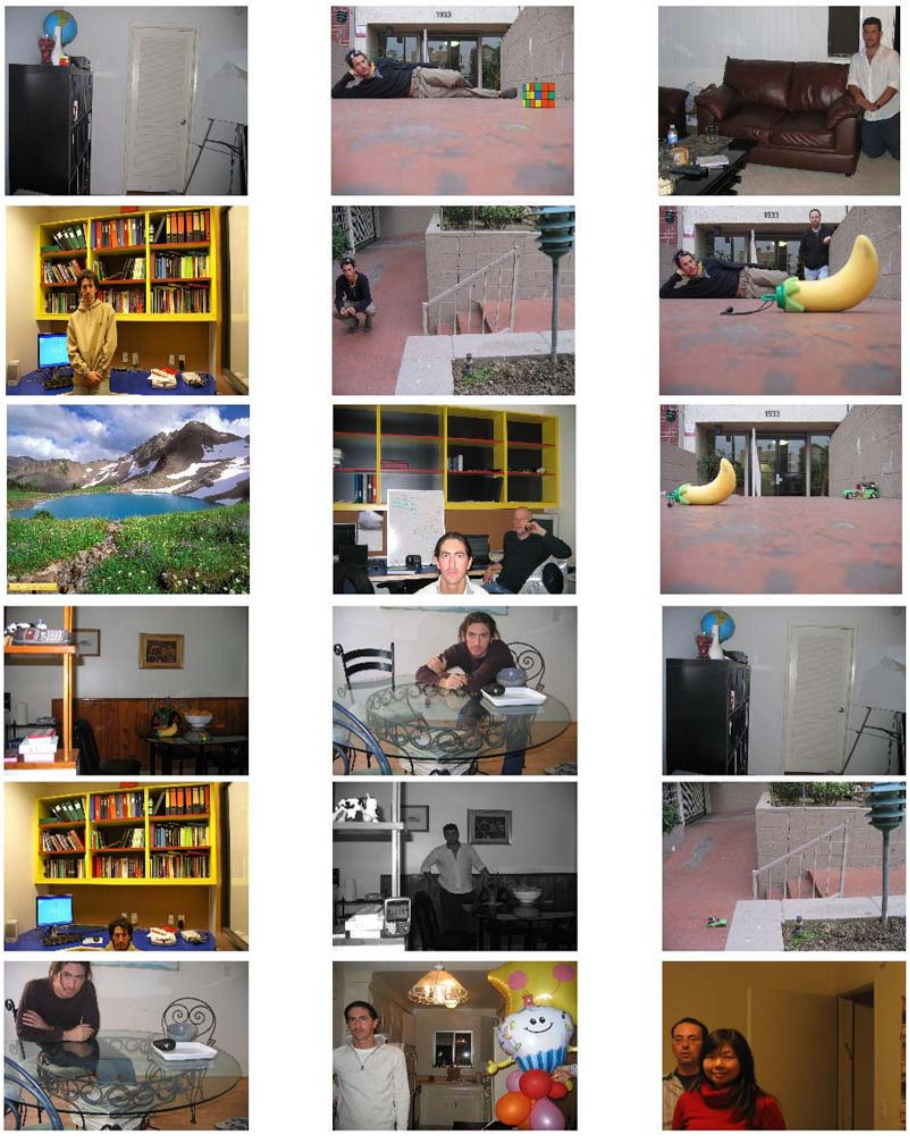


Figure 6.5: Thumbnail view of database of images.

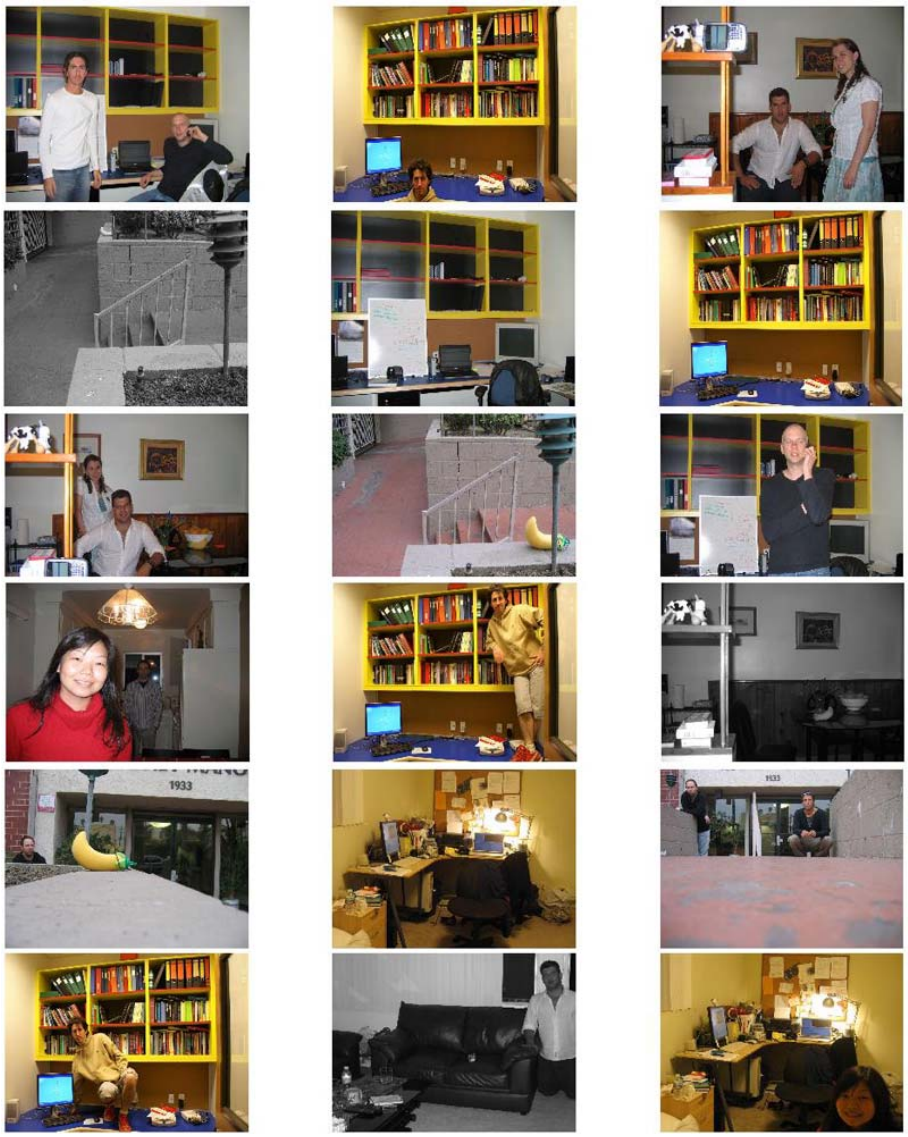


Figure 6.6: Thumbnail view of database of images.

Figure 6.7: Thumbnail view of database of images.

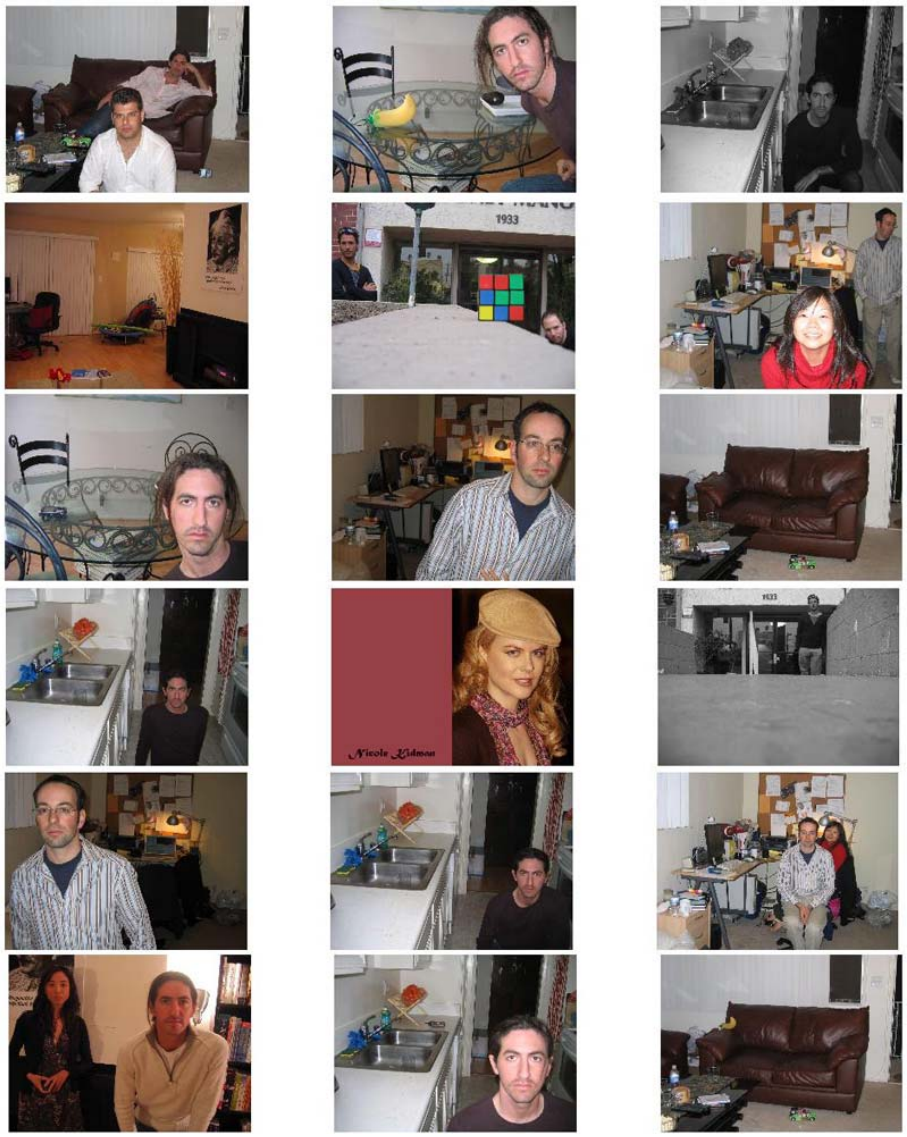


Figure 6.8: Thumbnail view of database of images.

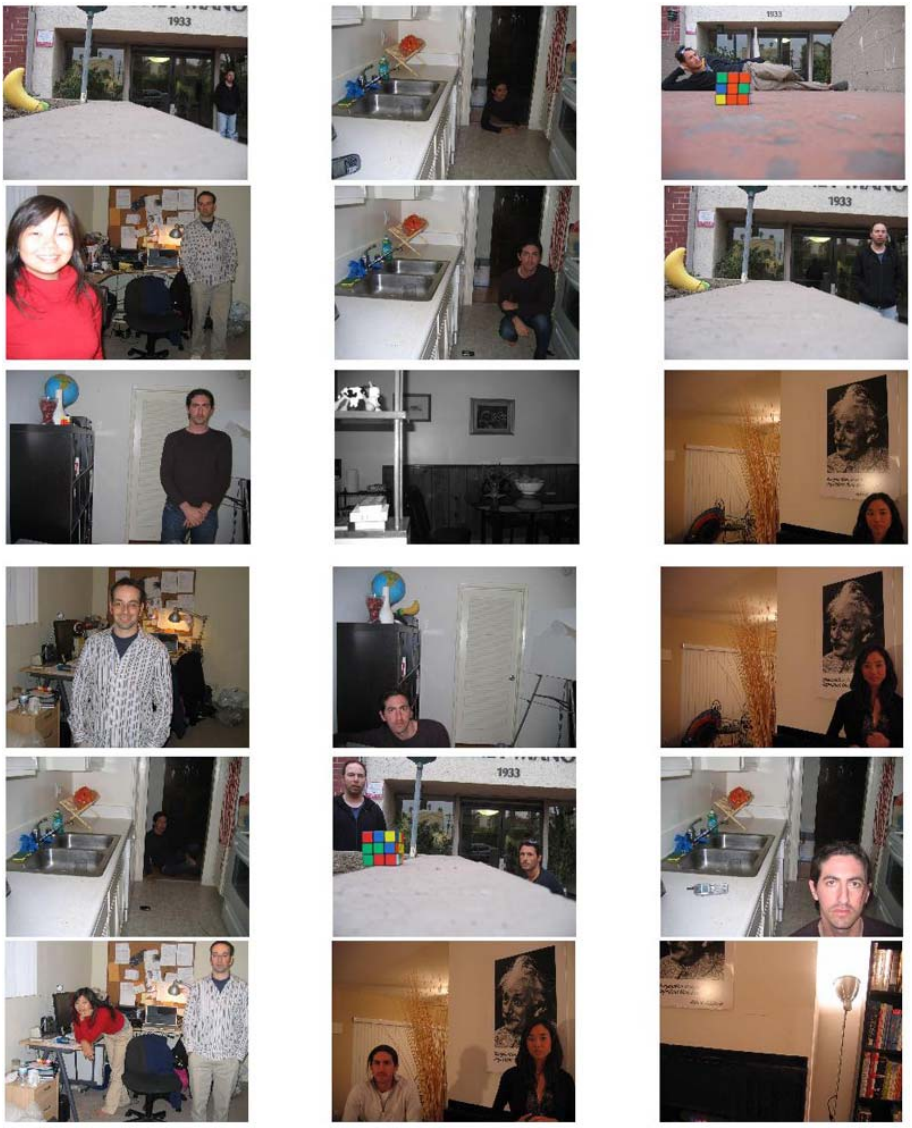


Figure 6.9: Thumbnail view of database of images.

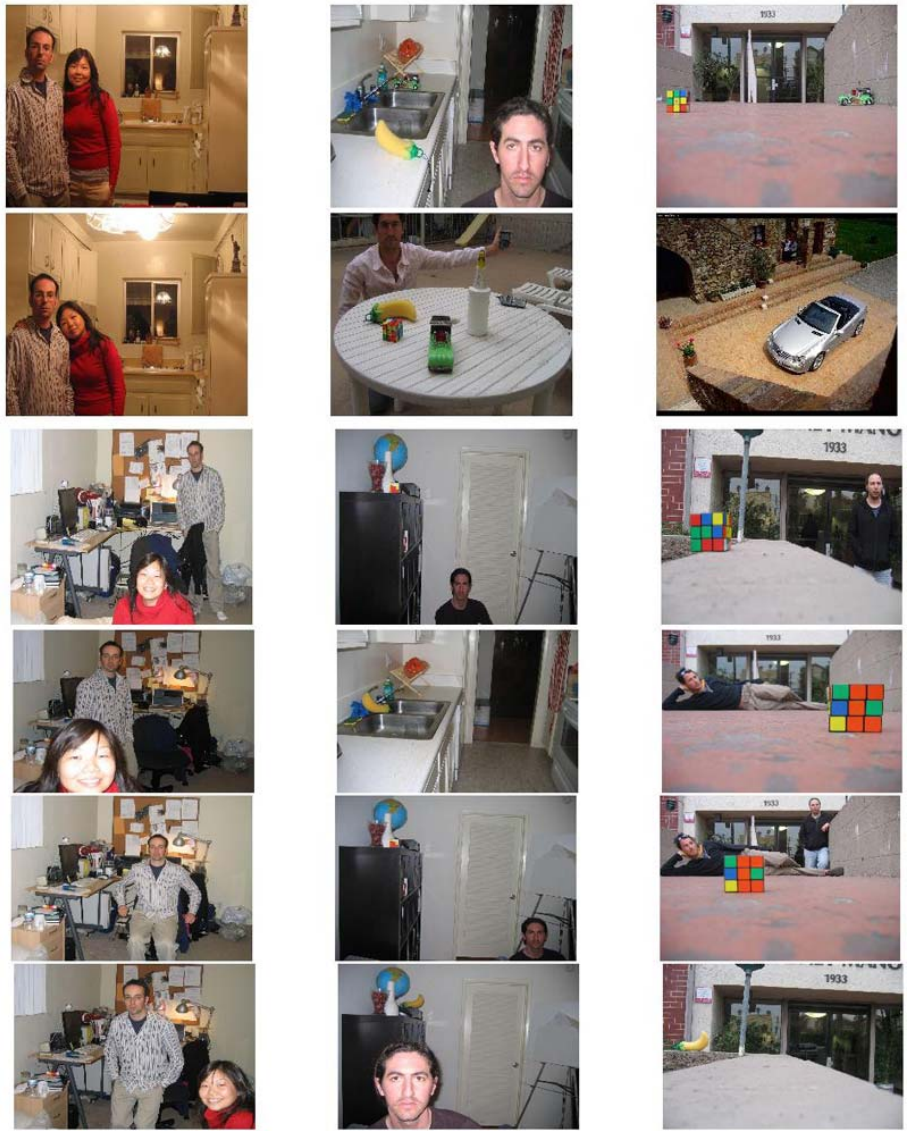


Figure 6.10: Thumbnail view of database of images.

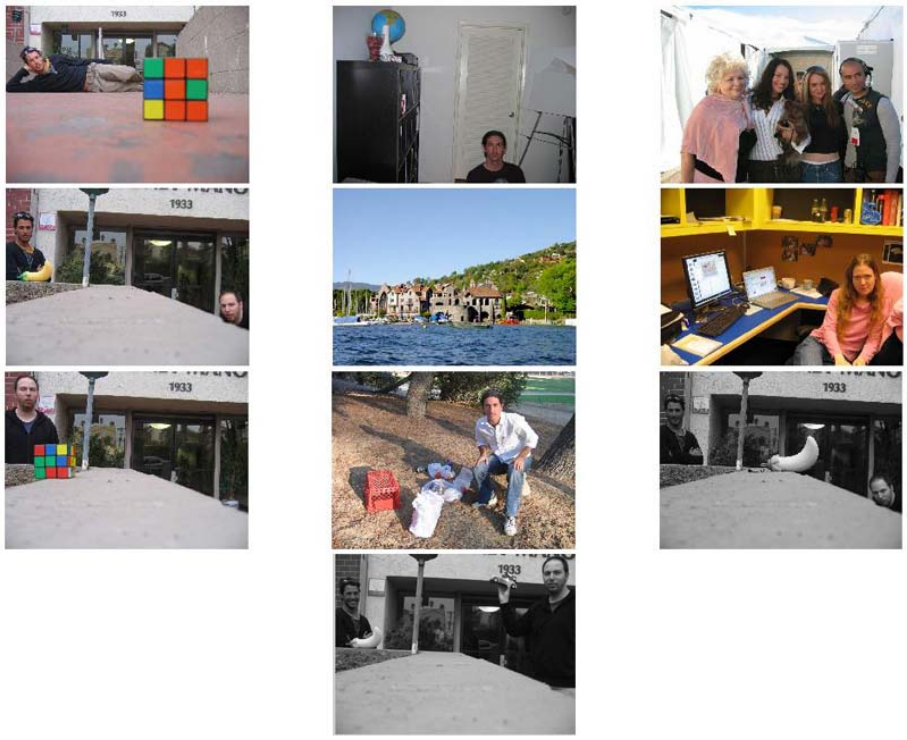


Figure 6.11: Thumbnail view of database of images.