UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK UND
MATHEMATIK

Behavior Adaptive and Real-Time Model of Integrated Bottom-Up and Top-Down Visual Attention

Zur Erlangung des akademischen Grades

DOKTORINGENIEUR (Dr.-Ing.)

der Fakultät für Elektrotechnik, Informatik und Mathematik
der Universität Paderborn
vorgelegte Dissertation
von

MS. -CS. Muhammad Zaheer Aziz

Paderborn

Referentin: Prof. Dr.-Ing. Bärbel Mertsching
Korreferent: Prof. Dr.-Ing. Reinhold Häb-Umbach

Tag der mündlichen Prüfung: 03.09.2009
Paderborn, den 08.09.2009
Diss. EIM-E/259

Dedication

Dedicated to my wife Rabia and daughters Maryam and Fatima who went through all the hardships during the time of this research very patiently.

Declaration

I hereby declare that I have completed the work on this PhD dissertation with my own efforts and no part of this work or documentation has been copied from any other source. It is also assured that this work is not submitted to any other institution for award of any degree or certificate.

Paderborn, September 8, 2009

Muhammad Zaheer Aziz

Kurzfassung

Visuelle Aufmerksamkeit ist ein wichtiger Bestandteil des natürlichen Sehens, der dazu beiträgt, die Datenmenge, die das menschliche Gehirn erreicht, wesentlich zu optimieren. Der Aufmerksamkeitsmechanismus beinhaltet einen Filterungsprozess der visuellen Informationen, um nur relevante und wichtige Anteile der gesehenen Szene für eine Analyse auf höheren Verarbeitungsebenen weiter zu leiten. Computer-gestützte Aufmerksamkeitsmodelle versuchen diese Filterung für Verfahren des künstlichen Sehens zu realisieren.

In dieser Dissertation wird ein gebietsbasierter Ansatz zu Modellierung visueller Aufmerksamkeit vorgestellt, der eine Alternative zu existierenden Modellen darstellt. Das vorgeschlagene Modell integriert bottom-up- und top-down Pfade der Aufmerksamkeit in einer einzelnen Architektur und nutzt beide Pfade unter Berücksichtigung verschiedener visueller Verhalten. Eine derartige Integration ist bisher von anderen Modellen noch nicht berücksichtigt worden.

Um auf mobilen Seh-Systeme Ergebnisse in Echtzeit erzielen zu können, wurden schnellere Algorithmen zur Merkmalsextraktion und Salienzberechnung entwickelt. Diese Algorithmen berechnen den Kontrast in fünf Merkmalskanälen sowohl im Kontext lokaler Nachbarschaft, als auch im globalen Kontext des gesamten Bildes.

Die Neuerung hinsichtlich der top-down Aufmerksamkeit ist die Erzeugung von Salienz-Karten feiner Granularität, mit der die visuelle Suche eines gegebenen Objektes durchgeführt wird. Diese Karten besitzen eine hohe Salienz für jene Gebiete, die eine höhere Ähnlichkeit zu den Merkmalen des gesuchten Objekts aufweisen. Jüngste Untersuchungen im Bereich biologischer Sehsysteme unterstützen die Annahme einer Top-Down Verarbeitung im Aufmerksamkeits-Kanal des menschlichen Hirns. Andere existierende Ansätze nutzen zu diesem Zweck

bislang ausschließlich Bottom-Up Karten, wodurch die vorliegende Arbeit einen signifikanten Beitrag zur aktuellen Forschung auf diesem Gebiet leistet.

Das vorgeschlagene Modell lieferte sinnvolle Ergebnisse und erzielte eine gute Leistung im Vergleich zu anderen verfügbaren Aufmerksamkeitsmodellen. Diese Arbeit zeigt neue Richtungen für die Untersuchungen in diesem Bereich auf, die zum Erreichen des ultimative Ziels biologisch plausibler, künstlicher Sehsysteme führen können.

Abstract

Visual attention is an important component of natural vision that helps it to optimize the amount of data that reaches the brain for detailed processing. The Attention mechanism applies a filtration process in the visual input that selects only relevant and important portions from the viewed scene for high level analysis. Computational models of attention attempt to perform this filtration for the machine vision systems.

The work presented in this dissertation proposes a region-based approach for modeling visual attention as an alternative to the other existing paradigms. The proposed model integrates bottom-up and top-down pathways of attention into a single architecture and makes combined use of these pathways under different visual behaviors. This was not done by any computational model of attention before.

In order to obtain real-time results on mobile vision systems new faster algorithms were developed for feature extraction and saliency computation. These algorithms compute contrast in five feature channels in context of local neighborhood as well as the global context of the whole view. The innovation in terms of top-down attention is the creation of fine-grain saliency maps for visual search of a given object. In the proposed maps high saliency is given to regions that have more feature similarity with the search targets. Latest research on biological vision suggests that such fine-grain processing takes place in the top-down channel of attention in the brain. Other existing models have used the bottom-up maps for this purpose, hence the proposed approach makes a significant contribution to the state-of-the-art.

The proposed model produced valid results and has shown good performance in comparison to other available attention models hence this research has openned

new directions for investigations in this field that can lead to the ultimate target of biologically plausible machine vision.

Acknowledgements

First of all, praise and thanks be to God who enabled me to reach this level of academic achievement and after all it is the visual attention created by Him that we are trying hard to understand and model. Secondly, I am grateful to my parents whose moral support and prayers made it come so far.

I am extremely thankful to my supervisor Professor Bärbel Mertsching whose guidance and support in all helpful aspects during the time of my PhD kept things advancing.

My colleague Dirk Fischer deserves a special thanks and appreciation for his continuous technical support and keeping the machines up and running without a break.

Contents

1 Introduction

Visual attention is defined as the cognitive process of selectively concentrating on significant aspects of the environment while filtering out unwanted information. In context of computer vision the study and modeling of visual attention serves at least two purposes. Firstly, building vision systems after the role model of natural vision leads to efficiency, robustness, flexibility, and adaptivity in machine vision. Secondly, performance of a computational model gives a feedback for the theory and concepts developed for the natural vision that can help in progress towards better understanding of nature. Amount of work on the computational models of visual attention increased significantly during the last decade. These efforts have not only established a firm working relation between the fields of machine vision, neurobiology, and psychology but stimulated new directions of research in these areas as well.

This chapter introduces the related areas of research in sections 1.1 and 1.2 and then establishes the necessity of the investigations done under this project in section 1.3. As the problem for which this work is carried out is a multidisciplinary topic and involves integration of many fields of sciences, there is a need to explicitly formulate the problem and divide it into practicable steps. Section 1.4 presents this analysis of the requirements and section 1.5 highlights the contributions of the presented work in advancement in the state of the art of the attention modeling. Section 1.6 portrays an outline of this dissertation.

1.1 Natural Visual Attention

Visual attention enjoys a key position in the vision process in humans and animals with developed eyes. The amount of information flowing toward our brains through eyes is quite large as compared to the capacity that can be processed at a time. The count of units that can be sensed by a normal eye is depicted by

the number of receptors in the retina. Anatomical studies, such as [Ost35] show existence of 6.4×10^6 cones and 1.1×10^8 to 1.25×10^8 rods in a human retina. On the other hand the number of axions in the optic nerve has been reported to be between 8×10^5 [Pol41] and 1.2×10^6 [QAG82] [BRD$^+$84]. This convergence in numbers show that the information reaching the brain gets significantly filtered at the early stages of vision.

A special mechanism of filtering the input is used by the human vision system (HVS) to deal with this problem. Detailed tasks of recognition, classification, or learning are applied only to a small portion of the viewed scene called focus of attention (FOA). Within a fraction of a second, the FOA is shifted to another location by a swift saccade of the eye. Apart from this overt attention by eye movement an internal form of attention also exists that is called covert attention [CM01]. Overt attention does not involve movement of eyes (or head) rather the selection is performed on different locations in the given scene. During an overt or covert gaze towards some FOA, the rest of the scene remains excluded from detailed processing unless some stimulus motivates the eye to shift attention to a new location. Some identified visual features that stimulate the visual attention include color contrast [LPA95], orientation [Ner04], motion (or a sudden change) [Ita01], eccentricity [BFR84], and symmetry [OH03]. A detailed review of the features involved in the human attention mechanism can be seen in [WH04].

According to history of the work on attention collected in [TIR05], the first recorded evidence of investigations to understand and describe the phenomenon of attention dates back to 1649 [Des49], the earliest reported psychological experiment on visual attention was in 1871 [Jav71], and the first information processing model for attention was proposed in 1953 [Pou53]. Being a complex process there is little concrete knowledge available about details of the involved processes that take effect in natural visual attention. Theories and models are formulated based upon psychophysical experiments and neurobiological study of natural vision systems. These theories differ and sometimes even oppose each other in many aspects hence it is difficult to identify hard facts with the current state of the knowledge on natural attention. On the other hand, those concepts on which most of the theories agree on can be considered as trustable. The most commonly known components of visual attention include bottom-up feature anal-

ysis [TG80] [WH04], fusion of the contributing feature channels [Ner04] [Koc99] for pop-out selection, inhibition of return (IOR) [PKK97] [SK00] [CT03], and top-down influences of a given task [SMC+03] [SHS01]. These conceptual models provide cues for construction of computational models that can be implemented on machine vision systems and their results can then be compared with those of the natural attention in order to either evaluate the computational model or to verify the theory of the natural model.

1.2 Visual Attention Modeling and Its Applications

Visual input is a major sensing modality to explore the environment with mobile robots. With the availability of improved quality camera devices, it is possible to obtain full color and high resolution images for machine vision. On the other hand, processing time can exponentially increase when working on images of big size. This raises the need of fast algorithms that are able to handle information of complex scenes.

The problem of abundance of input data is even more crucial in artificial vision systems because they do not possess a massively parallel computing power as that of the biological systems. More intelligent and precise machine vision can be performed on selected objects if the strategy of the natural eyes is followed. Models of artificial visual attention are constructed keeping human or biological vision as a role model. Their main objective is to locate those areas in a scene that have certain significance in some respect hence they apply a filtration process in order to select salient and important locations from the visual input so that detailed processing could be restricted only to these locations.

Most of the feature-based models of visual attention have their foundations in the feature integration theory [TG80] according to which features are automatically registered in parallel across the visual field of the eye in an early stage before the objects are identified. This theory proposes that separable feature dimensions for color, orientation, spatial frequency, brightness, and direction of motion are coded and combined to formulate a single object in the focus of attention (FOA). Hence the typical procedure of finding the focus of attention by a majority of attention models starts with computing of feature maps that highlights the salient

areas from the visual input in terms of the related feature. Number and nature of the feature channels are chosen in conformity with the knowledge of natural vision. Then these channels are combined to obtain a master conspicuity map in which peaks of saliency compete to win the focus of attention. The attended peak(s) are suppressed in the succeeding attention attempts using the inhibition of return mechanism in order to allow other salient locations to get attention.

Efforts have been made to apply visual attention to achieve lower computational cost not only in areas of computer vision but in other fields like robotics and computer graphics as well. Although the exact elucidation of the natural process is yet a far target the enhancement in efficiency has been reported for many vision related tasks such as visual search [IK00], image compression [BS03], video compression [Itt04], scene rendering in 3D graphics [CCM03], object tracking [OH03], automatic image cropping [Ste07], perceptual grouping [AM07a], and gesture recognition [HRB+03] even with the current state of the art of attention modeling.

1.3 Need of Region-Based Approach

Some of the existing models of visual attention apply linear center-surround operations between fine and coarse scales of the input while others utilize frequency domain filters. Such approaches yield saliency regions amid cloudy clusters that are, though, sufficiently good for the purpose of visual attention but the actual shapes of the attended objects get totally lost at end of the attention process. A redundant procedure of feature extraction becomes necessary for shape analysis and recognition of each FOA. Additionally, the feature computation methods of the existing approaches are computationally heavy causing a further delay in the overall output. Furthermore, due to the use of coarse scales of the input, many small regions worth attending get faded away before reaching the final stages hence remain unattended.

Early recognition or shape analysis of the candidates for attention is also necessary for computing the contrast based upon identity of objects. For example, an red colored ball surrounded by (red) apples will be a salient focus of attention for the human attention. Hence an early pixel clustering needs to be investigated

in attention modeling so that groundwork for such complex aspects of attention gets established. Clustering of pixels, that belong together due to being components of a single object, into regions reduces the number of units that have to be processed by the attention procedures. As processes of attention are extensively iterative, such reduction in the number of items can significantly reduce the processing time. Hence the region-based approach can also be helpful in context of improving the overall performance.

There is yet another aspect that demands determination and preservation of object shapes within the early stages of attention. The recent trends in modeling of visual attention are emphasizing on the combination of top-down and bottom-up contexts [NI06b] [FBR05], which in turn need shape and feature signatures associated with the objects for a rudimentary recognition in order to inhibit already known objects. Catering for these requirements right from the early stages of processing can lead to significant advantages.

The above mentioned issues may be of trivial importance when the consideration is limited only to modeling of attention but the work presented in this dissertation is an effort to integrate the attention procedure into a comprehensive biologically inspired vision system. Hence not only the efficiency of each internal process of attention has to be optimized but they need to be made compatible with the rest of the vision system as well in order to bring visual attention into practical use on mobile and static robot platforms.

1.4 Formulation of the Problem

The work under discussion is concerned with investigations on an alternative architecture for attention modeling that involves early clustering of visual input, improvement in inner processes for accelerating the attention procedure, extending the scope of attention modeling by integrating influence of vision behaviors (such as exploration and search etc.) in internal steps of the model. Therefore the issues addressed in this work span from fundamental image processing to a complete architecture of the attention model. The following subsections analyze the whole problem into its constituent steps and describe requirements of each step in context of reaching a suitable solution.

1.4.1 Segmentation

Segmentation is one of the most important steps in the point of view of machine vision systems. It clusters neighboring pixels into groups or regions using some homogeneity criterion such as similarity of color. With emergence of new and more complicated applications, requirements of good results from the segmentation process have increased. Color segmentation has to follow a different approach as compared to gray-scale segmentation. In gray-scale, only intensity information is available for which the computer can discriminate 256 levels. On the other hand colored images contain more complex structure comprising of at least three components, leading to millions of colors per pixel. These components have different meanings and roles in different color models, called color-spaces, such as RGB, YUV, HIS, $I_1I_2I_3$, and CIEL*a*b* etc. [CJSW01]. Each space has advantages and limitations in terms of segmentation hence a selection of the best color space is one of the major issues in color image segmentation [TT96]. Light effects such as shades and shadows cause a major problem in achieving optimal regions because one region may either get over-split into many or different regions may merge together.

The ability of the natural vision system to perform selection and inhibition on objects [PKK97] [SF03] strongly indicates that the process of clustering, in which individual points are grouped to formulate complete objects, is performed somewhere in the processing pipeline. Existing models of attention that do not agree on explicit formulation of objects, for example [IKN98] and [TCW$^+$95], model the clustering implicitly while constructing pyramid of low resolution copies of the input. Only those regions survive in the layers near the summit of the pyramid that occupy sufficiently large area in the original input. Hence a point in the higher level of the pyramid represents a cluster of points in the lowest level that were grouped based upon only their spatial connectivity. Hence it can be safely argued that the process of clustering is an essential part of the attention mechanism. The main question remains that clustering is performed at which level of processing and how. In the work presented here, we investigate the possibility of shifting the clustering to the earliest stage of attention processing using color segmentation. Hence the segmentation will be a primary step in our model.

The type of result required from segmentation differs from one application to another. There are some applications such as shape-from-shading where different color shades on an object are used for 3D reconstruction of a scene [TT92]. In this case, retaining and distinguishing each different shade of color is important. On the other hand for object recognition applications it is needed that effects of illumination may be neglected in order to acquire more accurate object shapes. Generally a segmented area or region has common features between its pixels such as color value, intensity, and luminance etc. In some cases a combination of two or more features is required to correctly segment a specific region. As the segmentation in this work is meant for artificial visual attention, high tolerance to light effects is required so that uniformly colored parts of objects do not get split due to mere variation of illumination. Hence over-segmentation has to be avoided along with minimization of under-segmentation in order to obtain good input to support further steps of the attention process. HIS (Hue-Intensity-Saturation) space, sometimes also called HSV (Hue-Saturation-Value) or HSL (Hue-Saturation-Luminance) represent the psychophysical perception in humans [CL94]. It separates illumination from the representation of the basic color and has capability of dealing with highlights, shades, and shadows [TT96]. Keeping in view these advantages, the work presented here uses HIS color space and proposes an improved algorithm to obtain better results, especially in context of the proposed region-based attention model (see section 3.1 for details).

1.4.2 Preattentive Feature Computation

Other models have been able to incorporate limited number of feature channels mainly due to heavy computational cost of feature extraction functions. We intend to increase the number of channels through acceleration in the feature extraction procedures. As a high amount of accuracy in feature magnitudes is not required at the pre-attentive stage because the main issue is to evaluate the saliency rather than object recognition, feature extraction methods can be optimized on time instead on accuracy for the purpose of visual attention. On the other hand we are modeling top-down saliency in parallel with the usually computed saliency along the bottom-up pathway hence we have separated the feature extraction step from the saliency detection process. Due to this separa-

tion the feature magnitudes become available for saliency processing in both of the pathways.

1.4.3 Saliency Computation

All models of visual attention consider color as an important feature channel for determining saliency. A majority of existing models have their foundations of color saliency computation in the concept of opponent colors described in psychology [HJ57]. During our search in the literature we found valuable information in the work on color theory about the attributes of colors that contribute in making an object visually prominent or receding. Artists practice these aspects for creating effects of contrast, visual advancement, and activeness in their illustrations. Accordig to [For06], Johannes Itten was one of the first who formally described methods for color combinations offering contrast. He has defined different situations in which the human vision finds contrast in colored scenes. According to his research, the contrast can occur due to presence of objects posing high difference of intensities, saturation, and/or hue. Other reported causes include presence of opponent colors and co-occurrence of warm and cool colors [Itt61]. Another relatively modern source of theoretical concepts on colors is available in [Mah96]. We combine the concepts from [Itt61] and [Mah96] of these resources to formulate a set of computationally feasible premises for design of the proposed methodology, see section 3.4 of chapter 3 for details.

An important requirement while computing saliency is the consideration of the global context apart from the immediate neighborhood of individual objects. Existing region-based methods for color saliency computation such as [BMB01] ignore the global context of contrast due to which such models tend to produce false results in certain cases. For example, figure 1.1 (d) shows the output of color contrast by [BMB01] using an image where a red square is surrounded by green squares on a black background as shown in figure 1.1 (a). Each square has a high contrast with the background in local context but the red one supersedes the others when global context is considered. It can be seen that the said model fails to arbitrate the red square as salient because its inner methodology takes only the local contrast into account. Some of the existing models of attention have considered shape-based features such as orientation, eccentricity, and symmetry.

A shortcoming in these approaches is ignoring of the aspect of rarity in terms of these features. For example [BMB01] applies a bias towards high magnitudes of eccentricity and symmetry for computing prominence in these two feature channels. Figure 1.1 (e) demonstrates construction of eccentricity map by this model where neglecting the rarity criteria results into giving high saliency to items appearing abundantly in the input given in figure 1.1 (b). Feature channel of size is not explicitly included by any other model so far. For example the inability of the well known model proposed in [IKN98] to assess saliency with respect to size using the input given in figure 1.1 (c) can be seen in figure 1.1 (f).

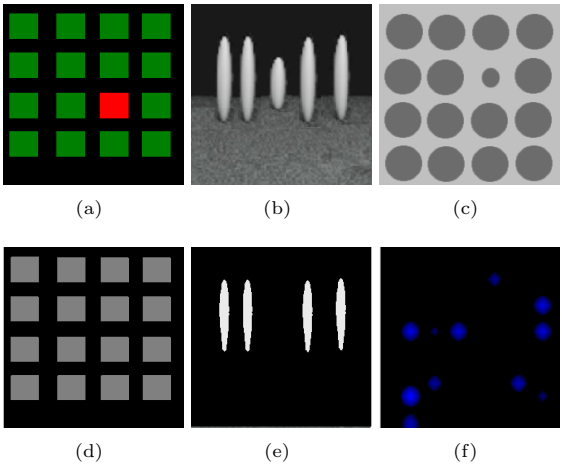


Figure 1.1: Samples of drawbacks of not using global contrast, ignoring rarity criteria, and using less feature channels. (a) Input with local color contrast. (b) Input with contrast of eccentricity. (c) Input having contrast of size. (d) Color saliency map by model of [BMB01] showing no saliency of the red box. (e) Eccentricity map by [BMB01] showing no saliency of the object having the rare eccentricity. (f) Saliency map by the model of [IKN98] showing no saliency to the actually salient object.

1.4.4 Inhibition of Return

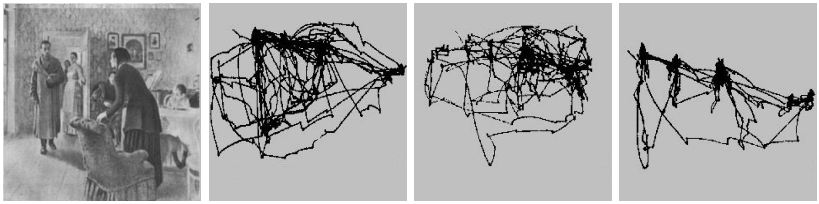
After attending a particular object or location in a given scene, the attention mechanism should perform an inhibition process on the attended locations in order to avoid frequent revisit of the same location. This process allows the system to fixate on all of the important locations present in the scene. The most commonly known inhibition is the spatial inhibition [PKK97] [CT03] in which a specific area around the attended location gets inhibited for a certain time [SK00]. There are also indications in literature on feature based inhibition as well [PKK97] [LPA95] [GGS05]. We keep all of these factors in the requirements list for the proposed method of inhibition of return.

1.4.5 Behavior Influence

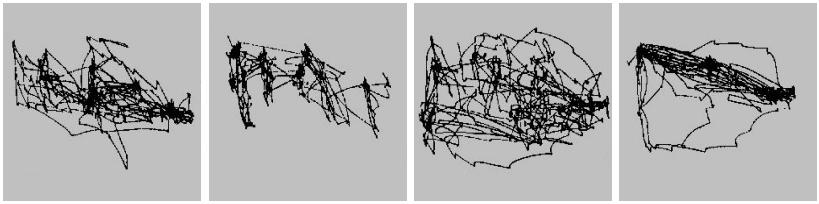
Fixation points entirely depend on the active visual behavior or the task at a hand, e.g the foci of attention while searching for a predefined object will be entirely different from those under free viewing. Figure 1.2 shows the results of experiments reported by [Yar67] in which the influence of task on the scan path of human attention is clearly visible. The experiments reported in [DLD04] also propose that adapting the human vision to a particular task selectively reduces sensory gain to a narrow range of the stimulus domain. Similarly the work presented in [NCMS04] reconfirms the findings of [Yar67] using a new set of experiments.

In order to make our model a more adequate model of human vision, we need to integrate the influence of behavior at each individual step of the model where the active task could play a part. It is not the peak selection only that has to be performed dependant on the active behavior but other internal steps of the attention have to be customized based upon the behavior also in order to make it flexibly adaptable according to the active vision behavior.

The existing models of artificial visual attention have implemented mainly three types of visual behaviors, namely *explore*, *search*, and *detect changes* (see section 2.8 of chapter 2 for references). In the proposed model we introduce *examine*



(a) Input (b) Free viewing (c) Estimate material circumstances of the family (d) Give the ages of the people



(e) Surmise what the family had been doing before arrival of the unexpected visitor (f) Remember the clothes worn by the people (g) Remember positions of people and objects in the room (h) Estimate how long the visitor had been away from the family

Figure 1.2: Results of psychophysical experiments reported by [Yar67] in which variation in attended locations and scanpaths on the same scene depending upon different visual behaviors is observable.

or *track* besides the commonly known behaviors. A brief description of the functionality of these behaviors is given in the following paragraphs.

Under the *explore* behavior the system performs attention under no influence from any task. It may also be called *free viewing*. Interesting locations emerge automatically due to their feature contrast compared to the background or neighboring objects. The feature maps are, therefore, task independent and the top-down pathway remains inactive during this behavior. On the other hand, the system tries to locate occurrences of a pre-defined object when working under *search* behavior. The definition of the search target is given to the system from an external source as a set of top-down conditions. The bottom-up pathway becomes primarily inactive under a *search* task. The behavior of *detect changes*

requires the system to respond on visual changes occurring in a given scene with respect to time.

In the work under discussion a new behavior has been introduced in which the system examines a set of similar looking objects to extract high level information, for example detecting the circle after observing the similar looking stars in the European Union flag, therefore we call it *examine* behavior. Top-down conditions do not come from an external source in this situation; rather they are generated from features of the previously attended object that are used as a basis in the forthcoming attempts of attention. These top-down conditions are used as a seed unit for extracting a bigger pattern formulated by these units. In case of dynamic scenarios with moving objects this type of task turns into *track* behavior as the system has to locate the identical features within the near vicinity of the previously focused object. Top-down and bottom-up pathways may work simultaneously as a highly bottom-up salient object will be allowed to distract the *examine* or *track* operations, although such distraction has to be minimized to make this behavior successful.

1.4.6 Dynamic Scenes and Overt Attention

Dynamic scenes consist of moving objects and shifting view due to movement of visual sensors with the mobile vision system. The sensors also have to rotate in order to bring the salient objects into center of view (overt attention). Such dynamic situation introduces new challenges to attention models especially in maintaining a correspondence between the subsequent frames of input. The locations already attended by the system get displaced due to motion of either the objects or the sensor leading to difficulty in applying location based inhibition of return. A dynamic memory based mechanism has to be designed in order to deal with this issue so that the inhibition in the succeeding frames of input is applied on the translated positions of locations attended in the previous frame. Moreover, the world locations of objects have to be involved in computations rather than locations in image frame so that objects could be localized even after movement of sensors.

1.4.7 Fine-Grain Top-Down Visual Search

Most of the models of visual attention have incorporated visual search by using the bottom-up feature saliency. Their methodology keeps the main focus on contrast detection during bottom-up map construction and no facility exists for highlighting a particular feature value under search. Models dedicated to visual search under top-down influence, such as [FBR05] and [NI05], also utilize the process of feature map construction given in [IKN98]. They apply adjustments to the weights of the whole feature channels rather than considering a particular magnitude (or a range of values) for locating the search target. Such adjustments in the map weights under top-down influence are somewhat helpful in allowing a quick popout of the target, due to excitation on the channel that makes the search target prominent among its environment, but the results can lack robustness and efficiency in situations where many other locations exist that possess a bottom-up saliency in the same feature channel.

The top-down pathway of attention has become an important topic of discussion in the recent research on visual attention as it is useful in solutions for many task driven attention behaviors such as visual search [NI05], tracking [BMB01], examining [AM07a], and loop closing in visual SLAM [FBR05]. We argue that the method of applying the top-down influence by adjusting weights of the bottom-up features channels, as most of the existing models do it, is not only inefficient but does not match with the natural process as well. Consider as example a search task in which an object with a special color is to be searched in a scene. According to the existing techniques, a very high weight will be assigned to the color channel if color is found to be the most prominent feature that would distinguish the object from the scene. As a result, the attention system will fixate on locations that are color-wise salient while suppressing other features. On the other hand, the human vision would excite the particular color associated with the target while suppressing other colors rather than exciting the whole feature channel of color in order to quickly locate a target. This concept is supported by the recent experiments that reveal fine grain nature of top-down selection [NI06b]. Other literature that establish ground for fine-grain top-down attention is discussed in section 2.2.4 of chapter 2. The proposed model has constructed the top-down

pathway using independent fine grain feature maps apart from the bottom-up saliency maps.

1.5 Contributions of this Work

A complete model for artificial visual attention has been designed and implemented that uses early clustering of input pixels through a color segmentation process. Such early clustering brings several advantages to the attention model. The clustered pixel groups act as substitutes for individual pixels not only at the original high-resolution scale of the input image but at lower resolution scales as well. At the high-resolution level, a cluster behaves as a representative of all the pixels included in it. For the low resolution requirement, the whole blob can be treated as representative unit of a bigger area of the input as done by a pixel in low resolution edition of the input. A major advantage of the segmented regions over the down-scaled pixels is that the regions represent clusters of homogeneously colored pixels that facilitate in finding contrast with respect to their neighbors. On the other hand, the down-scaled pixels group heterogeneously colored pixels the suppress the sharpness of contrast in the neighborhood. Hence, the segmented regions preserve more visual information especially for the saliency computation. Early segmentation has allowed obtaining the functionality of the pixel-based attention models with less computational complexity due to reduction of number of units being processed and simplification in feature extraction procedures on already grouped pixels.

Apart from the proposal of a new type of attention model, the work under discussion includes design of efficient and robust algorithms to obtain feature magnitudes and then saliency with respect to these features using clustered regions rather than individual pixels or frequency domain filters. The proposed feature extraction algorithms have not only enhanced the working efficiency of the proposed model but have made it feasible to include higher number of feature channels into the model as compared to other models. Another contribution is the implementation of the concepts from color theory while constructing a saliency map for the feature channel of color contrast. None of the existing models have investigated this option in their processing.

The region based approach has made it possible to retain the original shape of attended regions such that the regions could be sent directly to the pattern analysis or machine vision routines, hence a harmony can be established between the modules of visual attention and object recognition so that the attention phenomenon could become an integral and beneficial component of artificial vision systems. The retention of shape-based features also helps in tracking the attended objects in a sequence of frames in which the object alters their positions, using the feature signatures associated with the regions.

In terms of theoretical advancements, the influence of attention behavior is explicitly included into the internal steps of the proposed model. This was not done by any other existing model. Due to inclusion of the behavior influence in internal procedures of the model it was possible to integrate the bottom-up and top-down attention pathways into a single architecture. Such integrated model did not exist earlier. Another innovation in the proposed model is the implementation of fine-grain aspect of top-down attention. The existing models of attention use the bottom-up maps of attention for the purpose of top-down pathway which is a contradiction to the recent discoveries in the research on natural (human) attention that suggest fine-grain nature of this pathway. Until the time of writing this dissertation no other attention model published so far had proposed construction of fine grain feature maps for top-down pathway. Hence this aspect is also a new contribution into the research on visual attention. Another contribution is the design of a memory based inhibition mechanism in which world locations and features of the attended items are stored with a decaying inhibition factor age of the memory item increases. Such mechanism has facilitated working of the attention mechanism in dynamic scenarios where the vision system moves and its camera rotates to look around in the environment.

1.6 Thesis Outline

The work done on the project under discussion has multiple facets. The main focus is on the modeling of visual attention but some topics of low level image processing and pattern analysis also come under consideration because the

new model has a significantly different nature as compared to the existing approaches. These subtopics include color image segmentation, extraction of feature magnitudes for color contrast, eccentricity, orientation, symmetry, and size, construction of saliency maps for each of these feature channels, fusion of these maps, pop-out identification, inhibition of return, and dealing with attention in dynamic scenes. Effort has been made to organize this multi-disciplinary theme such that the concepts and existing literature on these issues, proposed methodology for each of the individual stages, and results are presented in a stepwise manner in order to make the presentation clearly comprehensible.

A thorough review of the literature from different areas of knowledge involved in the work on this project is provided in Chapter 2 where the demarcation of the concerned areas is done through different sections and subsections each dedicated to an individual issue. The methods for the basic image processing stage of the proposed attention model are presented in chapter 3 in which the newly designed algorithms for the foundation steps of segmentation, feature extraction, and saliency map construction have been described. Architecture and high-level processes of the proposed region-based visual attention model is explained in chapter 4 and results of experiments carried out using the developed model under different visual behaviors are presented in chapter 5. Chapter 6 evaluates the output of the proposed model and compares its results with some of the available models while chapter 7 summarizes the achievements and indicates the issues that requires further work in this direction of research.

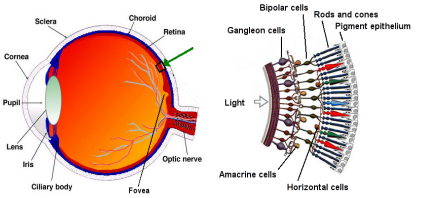
2 Related Literature

This chapter reviews the state-of-the-art related to the proposed work in one context or the other. At first a brief overview of the human vision system is provided and then the theories on natural visual attention are considered that provide a basis for the models of artificial attention. Then the conceptual models of natural attention are touched that provide a guideline for designing the computational procedures of the visual attention models. As the main focus of the dissertation is design of an artificial attention system hence literature on existing computational models is referred in detail in context of their overall architecture, processes of saliency computation, popout detection, and inhibition of return. Due to involvement of pixel clustering in the design of the proposed model, a review of literature on color segmentation is also provided. The proposed model also integrates the top-down pathway of attention into the bottom-up mechanism; hence publications in this regard are also referred in context of natural vision as well as modeling perspective.

Presentation of the literature is made by arranging methods and models under related categories. One model may be viewed under more than one perspectives especially when considering its constituting components in detail because one component of a model may fall under one category while another component may belong to a different one. Hence, repetition of references to models and methods may be found in the following sections but it was necessary do to this in order to review all aspects of the existing methods.

2.1 Human Vision System

Figure 2.1 shows the structure of eye along with a section of retina showing different types of cells in it. A circular field of approximately 6 mm around the fovea is considered the central retina while beyond this is peripheral retina stretching to the ora serrata, 21 mm from the center of the optic disc. The total retina is a circular disc of approximately 42 mm diameter [Pol41] [Kol91]. The optic nerve contains the ganglion cell axons running to the brain. The ganglion cells (the output neurons of the retina) lie innermost in the retina closest to the lens and front of the eye, and the photosensors (the rods and cones) lie outermost in the retina against the pigment epithelium and choroid. Light must travel through the thickness of the retina before striking and activating the rods and cones. The retinal message concerning the photic input and some preliminary organization of the visual image into several forms of sensation are transmitted to the brain from the spiking discharge pattern of the ganglion cells.

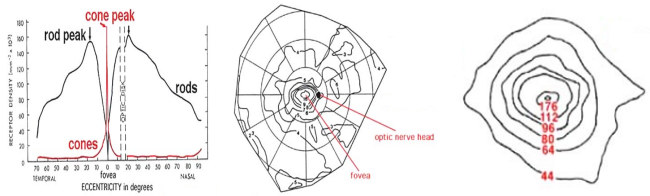


(a) (b)

Figure 2.1: (a) A sketch of the human eye with its major parts labelled. (b) Details of a retina section pointed by the arrow shown in subfigure.

Two basic types of photoreceptors exist in the vertebrate retina, namely rods and cones. The rods are sensitive to blue-green light and are used for vision under dark-dim conditions at night. There are three types of cones that are the basis of color perception depending upon their sensitivity to a particular range of wavelength of light. L-cones (red) are known to be maximally sensitive to wavelengths peaking at 564nm, M-cones (green) at 533nm and S-cones (blue) at 437nm respectively [Gou84].

Photoreceptors are organized in a mosaic that is a hexagonal packing of cones. Outside the fovea, the rods break up the close hexagonal packing of the cones but still allow an organized architecture with cones rather evenly spaced surrounded by rings of rods. The cone density is highest in the foveal pit and falls rapidly outside the fovea to a fairly even density into the peripheral retina as shown in figure 2.2. There is a peak of the rod photoreceptors in a ring around the fovea at about 4.5 mm or 18 degrees from the foveal pit. The region where the optic nerve begins (blind spot) has no photoreceptors.



(a) (b) (c)

Figure 2.2: (a) Graphs showing rod and cone densities along horizontal meridian of human eye [Ost35]. (b) Cone densities in differenet periphery areas of the human retina (in thousands) [CSP+87]. (c) Cone densities in fovea area of the human retina (in thousands) [CSP+87].

Ganglion cells are the major information processing units in the vertebrate retina. Ganglion cells collect information about the viewed scene from bipolar cells and amacrine cells and their final output goes to the brain visual centers through the optic nerve. Each ganglion cell has a fixel receptive field in which a high response is generated when the signal is incident on its center while a weaker response near the boundaries [Har68]. Three types of responses to light through optic nerve fibres attached to the gangleon cells has been reported. 'ON' type fibers respond when light turns from off to on and sustain an elevated discharge rate while the light signal remain on. 'ON-OFF' fibers respond when the light signal turns either from on to off or from off to on. 'OFF' fibers remain quiet until the stimulus light is turned off and remain active as long as the signal remains off in the receptive field. Using these three types of threads the gangleon can perform center-surround processing [MG75]. The ON-Center cells respond

when there exists a high light intensity at center of receptive field while a low light intensity in its surround. The OFF-Center cells respond when exactly the opposite conditions exists. For center-surround in terms of colors there exist tonic ganglion cells with green ON-center and red OFF-surround responses. Similarly Blue OFF-center and Yellow ON-surround type cells are also present. Including the opposite of these combinations a total of 12 spectral categories of center-surround combinations have been reported. The net impact of the center-surround receptive-field structure is that ganglion cells prefer small spots to large spots to drive visual attention. Different gangleon cells become selectively tuned to detect particular features of the visual scene, including color, size, and direction and speed of motion [LMMP59]. Interpretation of these signals is done by the brain in the context of events detected by other ganglion cells. Ganglion cell axons are directed to specific visual centers depending on the visual features they encode.

The ganglion cell have a characteristic of 'spatial tuning' of receptive fields is reflected. Each vertebrate ganglion cell is tuned to respond best for objects of a different size. Among the population of ganglion cells, a wide range of sizes is covered, perhaps corresponding to the wide range of object sizes in the visual image. This tuning reflects in part the variable dendritic span in ganglion cells. Dendritic span is one of the factors allowing ganglion cells to collect visual signals over a broad reach of visual space. Receptive field centers and dendritic fields can be similar in size [YM92] [YM94].

The part of brain responsible for visual activities is called visual cortex and is divided into further portions depending upon the neuron types and funtionality. Figure 2.3 displays these areas of the human visual cortex named as V1, V2, V3, V4, and IT. The optic nerves coming in from the eyes brings signals from the gangleon cells of the two retinas to the V1 area.

V1 transmits information to two primary pathways, called the dorsal (also called the "where") stream and the ventral (also named as "what") stream. The dichotomy of the dorsal/ventral pathways was first defined in [UM82]. The dorsal stream begins with V1, goes through visual area V2, then to the dorsomedial area and visual area V5 (sometime called MT) and to the posterior parietal

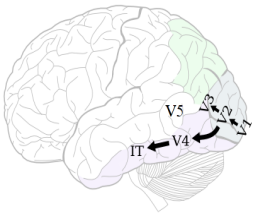


Figure 2.3: Human brain with its different parts of visual cortex labelled.

cortex. The dorsal stream is associated with motion, representation of object locations, and control of the eyes and arms, especially when visual information is used to guide saccades of visual attention or reaching objects by hands. On the other hand, the ventral stream begins with V1, goes through visual area V2, then through visual area V4, and to the inferior temporal cortex. The ventral stream is associated with form recognition and object representation. It is also associated with storage of long-term memory [GM92].

2.2 Models of Natural Visual Attention

This section reviews the concepts and theories that have influenced the computation models of visual attention existing today. There are some theories, like the feature integration theory and guided search theory, that have a large following in the community of computational modeling. The proposed model also combines concepts from these two theories.

2.2.1 Feature Integration Theory

The feature integration theory was proposed by Treisman and Gelade [TG80] in 1980. It has been one of the most influential psychological models of human visual attention. It suggests that the human vision system can detect and identify separable features in parallel across a display and this early, parallel process of feature registration mediates between texture segregation and figure-ground

grouping. They further conclude from their experiments that locating any individual feature or performing their conjunctions requires attention to be diverted to each relevant location.

According to this theory several primary visual features are processed and represented with separate feature maps in an early step of visual processing. These maps are later integrated in a saliency map that can be accessed in order to direct attention to the most conspicuous areas. Efforts can be found in experimental psychology to identify the features that stimulate the visual attention mechanism. Some of identified features in this regard include color contrast [LPA95], orientation [Ner04], motion (or a sudden change) [Ita01], eccentricity [BFR84], and symmetry [OH03]. A detailed review of the features involved in the human attention can be seen in [WH04]. Models regarding combination of the feature channels in the pre-attention phase are proposed in [Koc99] and [Ner04]. The operation is modeled as square of sum in [Koc99] while [Ner04] proposes that the features are combined in the visual cortex using a multiplication-style operation.

A search task is categorized into two kinds according to this theory, namely, feature search and conjunction search. Feature search can be performed fast and pre-attentively for targets defined by primitive features. Conjunction search is the serial search for targets defined by a conjunction of primitive features. It is much slower and requires conscious attention. Color, orientation, and intensity are proposed as primitive features for which feature search can be performed.

2.2.2 Spotlight Model

The attention spotlight is one of the metaphors used by researcher of this field becuase the attended location is considered to be under a spotlight while the rest of the scene being in darkness [FDJ99]. The issues on which debate has been carried out include the size, shape, and type of movement of the spotlight. Another question is whether the spotlight is splittable or not. The concept of an attention focus having similarity to a spotlight was given by LaBerge [LaB83] during an effort to find the relationship between the size of attentional focus and

the time taken to process the contents of the focused items. A formal spotlight model was first proposed by Eriksen and Yeh [EY85] in 1985. As the spotlight wa criticized due to its fixed size therefore the model was modified into a zoom-lense model [EJ86]. These models assume that while looking at a view the visual attention works like an internal spotlight that moves across the scene by which certain parts of the scene are illuminated. This hypothesis is taken from the overt attention performed by the eye in which the eye focuses on only selected locations of the whole view. The remaining portions of the view remain in dark or ignored area. There has been a debate in this paradigm about the movement of the spotlight that whether it moves continuously or jumps from one location to the next [LCWB97]. A later experimental work belonging to this category suggested that the beam of attention could be splitted leading to tracking of multiple objects at a time [KH95]. This theory of attention does not have much of following in computational modeling presumably because they do not propose any methodology about the internal working of attention.

2.2.3 Guided Search

The guided search model was introduced by Wolfe and colleagues in 1989 [WCF89] and then its revised version was presented in [Wol94]. This model is related to the process of visual search in which the main objective is reduction in response time to identify presence of a search target in a given scene. The model suggests that a number of features like color and orientation are computed and stored in maps in each of which presence of these features is encoded. The top-down influence of the target features controls the construction of these maps to apply bias to a particular category of feature values. A weighted summation method is used to integrate these maps together into a combined activation map. Peaks in this activation map are visited serially as targets of attention. Locations once visited are marked in an inhibition map that is used to avoid rapid revisiting of the same locations. This model, especially its part of map construction, has been an inspiration for many computational models of attention as its processes are clearly explained making them feasible for conversion into algorithms.

2.2.4 Fine-Grain Top-Down Attention

A majority of the existing attention models have demonstrated visual search as a primary area of application for their models. Most of these models have utilized manipulation on bottom-up saliency maps in order to let the search target pop-out quickly. Although such processing demonstrates visual search as an attentional behavior but in practical sense these models lag behind the performance of natural visual search significantly. This suggests that the top-down tasks of attention have a different nature and require a separate mechanism for computing saliency. There is another school of thought about the top-down influences that leads to fine-grain nature of this attentional pathway in natural vision, which appeals better than the strategy mostly followed by the contemporary attention models.

The models of human vision such as [LD04] suggest target related feature processing in the V4 area of brain. Similarly the models on feature and conjunction search, for example [LHG97], also presume excitation and inhibitions on particular feature magnitudes rather than whole channels. Results of psychophysical experiments reported by [Ham05], [Dec05], and [NI06b] also support the concept of search on particular feature values rather than excitation on a whole feature channel. The work of [Ham05] has shown that a population of neurons encoding the target color and/or orientation gets a gain while others get suppressed. According to [Dec05], each feature channel can adopt many values that are evaluated by a specialized layer of neurons in the human brain.

Recent psychologial models of attention such as [HT06] and [Knu07] agree on the concept that top-down modulations of neural responsiveness are precise for the features upon which attention is to be diverted. Apart from the excitation of the neurons concerned with the stimulus, it has been reported that neurons tuned for non-target stimulus parameters exhibit a decrease in sensivity [RD03]. The experiments reported by [NI06b] explicitly declare fine-grain nature of top-down attention. These findings suggest that the top-down saliency mechanism constructs task dependant maps to allow quick pop-out of the target rather than using the bottom-up saliency maps. The model proposed here follows this newly discovered strategy in its top-down pathway.

2.2.5 Inhibition and Facilitation of Return

After focusing one one salient object/location, the next important component of attention that gets activated is inhibition of return (IOR). This process enables the vision system to fixate on a variety of locations (or objects) in the scene, otherwise the gaze would stay fixed to one salient location. It is worth mentioning here that there has been a contunuing debate on early selection and late selection. In the early selection models, such as [Bro], [Tre60], and [TG67], attention is diverted to a location without forming a semantic meaning to the contents. Hence attention and IOR work only on location basis. On the other hand, according to late selection [DD63] the contents of sensory data are analyzed semantically before attending, therefore objects may be identified and used as units to perform attention processes. In the work presented in this dissertation the object based attention is taken into consideration.

It has been established by experiments in psychophysics that inhibition takes place in terms of both location and object features [GE94] [WLW98]. Evidence is provided for inhibition in the immediate vicinity of the attended location and a U-shaped function has been reported which strongly suppresses the immediate surroundings of the attended location and gradually fades to no suppression after a limited diameter [CT03]. The work of [LPA95] discovers the idea of feature based inhibition in which inhibition on color of the recently attended object has been reported in human vision. It was further confirmed by experiments reported in [PKK97] that inhibition takes place in terms of object identity apart from the spatial inhibition of return. The psychological model of attention proposed in [Knu07] defines an explicit role of a working memory while processing for bottom-up as well as top-down visual attention.

Under some visual behaviors, such as search and track, bias has to be given to certain features and/or locations so that the next fixations are driven towards similar looking features or nearby positions. This component of the attention mechanism is called facilitation of return (FOR) [OMY05] [CC06].

2.3 Computational Models of Attention

This section reviews the existing computational models of attention in a categorized arrangement. Categorization is made based upon the basic principle used for computing the focus of attention. Existing models could be classified into four types of approaches, namely, connectionist, saliency based, rarity based, and object based. The proposed model cobines attributes of the last three types of approaches in its current status. This review will also provide a basis to understand the innovations suggested in the proposed attention model.

2.3.1 Connectionist Models

A model of attention-based object recognition was proposed in [OAE93] in which a heirarchical system of connected layers for selection of attention window was introduced. It uses dynamic routing circuits and a pyramid with varying resolutions of the input. The attended region is mapped to the centers of the higher layers in the pyramid in order to sustain the spatial relations. The information flow between the pyramid layers is guided by control neurons. IOR is implemented by inhibiting control neurons connected to the attended pattern routing.

The prominent model belonging to this category was introduced by Tsotsos and colleagues, which was called Selective Tuning Model [TCW$^+$95]. It consists of a layered network with an input pattern on the lowest layer. The top layer calculates a global winner for focusing attention. On the lower layers signals converge layer-by-layer to select regions of interest in a feed-forward fashion. The final winner activates stimuli below it as a trace-back mechanism. This feed-forward and trace-back system is named attentional beam that links the layers with its sharp tip in the top layer and wider base in the bottom layer. The already attended area is completely inhibited and internal representations are computed again to find the next FOA.

The model presented in [PHH97] is known as Postma's SCAN that consists of more than one hierarchical layers named gating lattices, each of which contains

many overlapping sublattices. Only the winning sublattice can route its pattern to the higher layer of hierarchy after a WTA process. This selected pattern is sent to a classification network. Another model in this category [HBSH07] names its attention system SAIM. It consists of two networks, called content and selection, interacting with each other. Each unit in the contents network represents a correspondence between the input and the FOA; the selection network determines which correspondences are instantiated. The system acts to map retinal input into the FOA, based on competition between units in the selection network.

2.3.2 Saliency-Based Models

The saliency-based models have their foundations in the feature integration theory. A prominent model of this category was presented in [IKN98] and then refined in [IK00]. It builds saliency maps for three features, namely, color channels, intensity, and orientations. Each feature is computed by a set of linear center-surround operations between fine and coarse scales analogous to visual receptive fields. These feature maps are combined into three conspicuity maps for intensity, color, and orientation through across-scale addition. At any given time, the maximum in the resultant of saliency maps defines the most significant image location to which the focus of attention should be directed. This is done by a 2D layer of leaky integrate-and-fire neurons. This layer feeds into a biologically plausible winner-take-all (WTA) neural network. Shift of attention to the winner location causes a global inhibition of all WTA neurons and transient activation of local inhibition.

Many flavors of the above model can be found with different variations in methodology. The model described in [PSL02] uses the opponent color theory for constructing the feature map of color contrast using a computation scheme very similar to [IKN98]. It introduces new feature maps for edges and symmetry. It computes two color maps and the center-surround is implemented as the difference between fine and coarse scales of a Gaussian pyramid images. A total of 24 maps are computed and combined into four conspicuity maps. Unsupervised learning is used to determine the relative importance on different bases to gen-

erate a suitable salient region. The IOR process is implemented by masking the currently attended focus of attention for the next attention cycle.

The model presented in [MCBT06] implements the opponent color theory by computing the color distance in Krauskop's color space. Contrast sensitivity functions are applied on the three color components in the frequency domain. The saliency of an achromatic structure is enhanced if this structure is surrounded by a high contrast in chromatic channels.

2.3.3 Rarity-Based Models

Models of this category concentrate on finding locations in the visual input that contain rarity with respect to a considered feature. The method of [Ste01] for color saliency picks a selected set of neighborhood pixels around a target pixel and compares it with a similar pattern of neighborhood at several test locations. The exclusiveness is computed by subtraction of color components of every corresponding pixel in the neighborhood patterns around the target and each test location. A large value of this exclusiveness adds a score of saliency to the target. The sum of these scores after checking a number of test locations decides the final saliency value for the target. Another work presented in [Ahu96] generates the color contrast map according to rareness criteria on feature maps of intensity contrast, saturation contrast, and hue contrast. Intensity and saturation is convolved with a Laplacian of Gaussian kernel at each point. The circular nature of hue is normalized before applying the convolution. The orientation map is constructed using a rareness criteria using a Gabor kernel of four different angles on intensity, saturation, and hue and then picking the maximum as the resultant. The model of [AL06] also takes rarity in terms of visual features into account to identify salient regions in the scene.

2.3.4 Object-Based Models

In this category of attention modeling the computation of saliency is done on basis of higher level units instead of individual pixels. Objects are formulated by

combination of features and clustering of points that belong together due to similarity of some attribute. The model presented in [SF03] computes object-based saliency depending on groupings. A grouping is considered to be a hierarchical structure of "objects and space"; hence it may be a point, an object, a region, or a structure of other groupings. The primary features are extracted exactly as done in [IKN98], but it constructs the intensity, color (red, green, blue, and yellow), and orientation pyramids after applying a Gaussian filter and then a Gabor steerable filter on the five feature channels of intensity, red, green, blue, and yellow. The shift of attention is carried out by using an algorithmic approach with a coarse to fine strategy.

Some models have partially region-based components in their strategy. The attention model of [BMB01] uses a region-based approach for construction of maps for color and eccentricity. The conspicuity of a region in terms of color is calculated as the mean gradient along its boundary to the neighbor regions. The color gradient between the two regions is defined as the Euclidian distance between mean values of the color components in MTM color space. The eccentricity map is constructed using moments of segmented regions. The model proposed by [LLY+05] also utilizes a region-based method for the feature of color contrast and texture contrast. They include skin color and face existence as cognitive features for attention. A three step approach is used for color contrast in which the image is first clustered using a k-means algorithm. The biggest cluster having a large enough size is considered as background and then color difference of each cluster is computed in contrast to the background. The resultant map is scaled and truncated to remain within prescribed limits.

Formation of objects from raw pixel data is a signicantly complex task. The model proposed in [SF03] remained till a theoretical proposition without going into details of implementing formation of the so-called groupings. Other models in this category used the early clustering approach but suffered from the computational complexity resulting in fairly long response time. The approach proposed in this dissertation is an effort to make advancements and innovations in the methodology of the early clustering paradigm in order to make object-based approach usable in real-time attention systems.

2.4 Feature Extraction and Saliency Map Construction

This section discusses the methods of feature map construction used by different existing models of visual attention. A quick review of some literature other than attention modeling is also provided in order to have a glimpse of the contemporary trends for extraction of the concerned features. We may roughly categorize the feature computation methods in the attention models into three classes. First is the group that processes individual pixels at single or multiple scales of the input and then applies some sort of clustering for formation of objects under attention. We may name them as pixel-based methods. The second category of methods carries out its processing in the frequency domain by mostly applying Gabor filters. The third group performs a clustering first (such as region segmentation) and then computes features using these clusters.

2.4.1 Pixel-Based Approaches

Color contrast computation using pixel-based category of algorithms includes methods by [IKN98], [SF03], [PSL02], [Ahu96], and [Ste01]. For the feature map of color contrast, the model of [IKN98], and most of the other existing models, use the concept of opponent colors that was first introduced by Hering in 1872 [Her64] and further established by efforts such as [De 60] and [EZW97]. The attention models compute chromatic opponent colors of red-green and blue-yellow along with the achromatic opponent pair of white-black. Six maps are constructed for intensity feature by computing the absolute difference between intensities of the considered pixel and its surround at six different scales. For chromatic colors, each of red, green, and blue channels are normalized by the intensity channel and then double-opponency is determined by center-surround differences across scales. Six maps each are created for red/green and blue/yellow. A single conspicuity map for color is created after running an iterative lateral inhibition scheme on each feature map.

The model of [SF03] uses a similar basic concept for the color contrast computation. The nature and number of maps are also same as that of [IKN98] but a different calculation method is applied. The model presented in [PSL02] also uses the opponent color theory for constructing the feature map of color contrast

using a similar computation scheme as that of [IKN98]. They have introduced a feature map for edges that highlights strong boundaries in the input.

Another recent work presented in [Ats07] generates the color contrast map according to rareness criteria on feature maps of intensity contrast, saturation contrast, and hue contrast. They convolve intensity and saturation with Laplacian of Gaussian kernel at each point. The circular nature of hue is normalized before applying the convolution. The method of [Ste01] for color saliency picks a selected set of neighborhood pixels around a target pixel and compares it with a similar pattern of neighborhood at several test locations. The exclusiveness is computed by subtraction of color components of every corresponding pixel in the neighborhood patterns around the target and each test location. A large value of this exclusiveness adds a score of saliency to the target. Sum of these scores after checking a number of test locations decides upon the final saliency value for the target. A similar technique has been applied to detect facial symmetry in images containing already separated faces [Ste05].

Algorithms using pixel-based approaches for computing symmetry in visual input can be found in [PSL02], [OM02], [KG98], and [FS06]. The attention model in [PSL02] computes a symmetry map using a noise tolerant generalized symmetry transform algorithm on edge information of the input. The method of symmetry detection proposed by [OM02] uses a network of globally coupled maps associated with each pixel for finding reflection symmetry around it. It has shown success in images containing mainly one foreground object. The method of [KG98] considers a circular Gaussian window for local symmetry and determines the 2D symmetry as a resultant of 1D symmetry functions along line segments parallel to the examined axis. Although this method has a similarity with the basic idea of the proposed method (see section 3.5.1 in chapter 3) but the calculations involved in this method are mathematically complex making them computationally expensive. The work in [FS06] proposes a set of invariants based on complex moments to determine N-fold rotation symmetry.

The pixel-based approach applied for orientation map given in [SF03] computes the angle differences between centers and surround for the four angles of 0, 45, 90, and 135 degrees and then obtain the saliency value as product of Gaussian

distance between the considered pixels and the trigonometric sine of the angle difference between the center and its surround.

The selective tuning model mainly works with motion features [RT06] but other features such as color and orientation have also been utilized in some of their implementations such as [Zah04]. Depth as a feature can also be seen as mentioned in [BT05]. Although it is hard to find clear details of the methods for extraction of feature maps from literature on this school of attention modeling but they process individual pixels at different scales of the input hence their approaches belong to the pixel-based category.

2.4.2 Frequency Domain Methods

Frequency domain methods for color contrast computation have been applied by some attention models, for example [MCBT06]. It implements the opponent color theory by computing the color distance in Krauskop's color space. Contrast sensitivity functions are applied on the three color components in the frequency domain. The saliency of an achromatic structure is enhanced if this structure is surrounded by a high contrast in chromatic channels.

Orientation maps are constructed using frequency domain techniques by most of the attention models. The model of [IKN98] computes the local orientation in different scales of the image through creation of oriented Gabor pyramids from the intensity channel. They encode the magnitude of difference in orientation between a point and its surround for four angles of 0, 45 , 90, and 135 degrees using absolute center-surround differences between these channels. [Ats07] constructs the orientation map using a rareness criteria by applying a Gabor kernel of four different angles on intensity, saturation, and hue and then picking the maximum as the resultant.

Frequency domain operations for computing symmetry are also available, for example, [BMB01] constructs symmetry map for monocular images by applying Gabor filters in twelve orientations on edge image of the input. Their depth map is also computed in frequency domain through disparity in results of Gabor filtering. One of the recent methods presented in [KS06] utilizes angular corre-

lation computed by a pseudo polar Fourier transform to determine the center of symmetry.

2.4.3 Region-Based Techniques

The attention model of [BMB01] uses a region-based approach for construction of maps for color and eccentricity. The conspicuity of a region in terms of color is calculated as the mean gradient along its boundary to the neighbor regions. The color gradient between the two regions is defined as the Euclidian distance between mean values of the color components in MTM color space. The eccentricity map is constructed using moments of segmented regions. The model proposed by [LLY+05] also utilizes a region-based method for the feature of color contrast and texture contrast. They include skin color and face existence as cognitive features for attention. A three step approach is used for color contrast in which the image is first clustered using a k-means algorithm. The biggest cluster having a large enough size is considered as background and then color difference of each cluster is computed in contrast to the background. The resulting map is scaled and truncated to remain within prescribed limits.

Saliency with respect to size is also an important factor that contributes in natural visual attention, especially when other factors do not formulate a decisive focus of attention [SF03], but it is rarely implemented in existing attention models. The approaches used for this purpose are usually region-based. The method in [RC04] uses the Coherence Theory of Rensink [Ren00] to separate small foreground regions from the large background. The consideration of size in [Wol00] is also limited to discriminating between large and small objects in a given scene. [SF03] has mentioned the importance of a size contrast map but no proposal has been made to model it.

2.5 Feature Map Combination and Popout

In this section we review the methods adopted by different attention models to combine the feature maps to determine a focus of attention and then applying inhibition on the attended locations in order to shift the attention focus to a

different location. The model presented in [IKN98] and [IK00] first normalize the feature maps of color contrast C, intensity I, and orientation O using a normalization function N and then apply a simple weighted sum to obtain the input S for the resultant saliency map as follows:

$$S = [N(I) + N(C) + N(O)]/3$$

The saliency map is implemented as a 2D layer of leaky integrate-and-fire neurons that takes S as input and feeds into a Winner Take All (WTA) neural network. The WTA network ensures only one occurrence of most active location at a time. In this model the inhibition of return is implemented by spatially suppressing an area in the saliency map around the current focus of attention while feature-based inhibition is not considered. Another recent effort [NI05] by the same group includes the task driven top-down influence during the bottom-up saliency map construction. The elementary units of computation are pixels or small image neighborhoods arranged in a hierarchical structure.

The model proposed in [PSL02] uses a weighted sum of feature maps to obtain a combined saliency map. They use Independent Component Analysis algorithm for unsupervised learning to determine relative importance of features and to reduce redundancy. An adaptive mask is used to suppress the recently attended object for performing the inhibition of return. The model of [HRB+03] also computes a weighted sum of individual feature maps for obtaining an integrated attention map but introduce a manipulator map which is multiplied to the sum. The output conspicuity map C^m is obtained by applying a threshold function θ on the weighted sum of the feature maps M_i and multiplying it to the manipulator map M_m. Hence

$$C^m_{(x,y)} = \sum_{i=1}^{N} \theta(w_i \times M_{i(x,y)}) \times \prod_{m=1}^{l} M_{m(x,y)}$$

The maximum in C^m is taken as the point of attention. No inhibition function was used as it was not needed in their application.

The model presented in [BMB01] includes the aspect of tracking multiple objects while focusing attention in a dynamic scene. They first determine the features

that lead to activation of the neural fields that are in turn responsible for determination of pop-out. Then they adapt the weights of these feature maps so that a pop-out emerging due to a specific feature receives the main support from that particular feature map. A separate map is used for IOR where the visited location is marked as highly active. This activity inhibits the master map of attention to avoid immediate revisiting of the attended location. The activity of the inhibition map decays slowly in order to allow revisiting of the location after some time.

The method proposed in [SF03] implements a hierarchical selectivity process using a winner-take-all neural network. They apply a top-down influence to increase or decrease the baseline of neural activity of the most prominent feature channel. As their model deals with so called 'groupings' of pixels, the IOR process works on siblings of the current focus of attention in the hierarchy of groupings and sub-groupings. Another recent model in [MCBT06] uses the direct sum of the feature channels to compute a two-dimensional saliency map but they introduce an anisotropic Gaussian as the weighting function centered at the middle of the image.

2.6 Models for Inhibition of Return

Inhibition of return is an important aspect of attention that has to be modeled in artificial vision systems in order to avoid continuous attention to only one location or object. The commonly used approach by the existing models is to make a 2D inhibition map that contains suppression factors for one or more spots that were recently attended. The models of [IKN98], [FBR05], and [DBZ07] are examples where such an approach is utilized. Although this type of map can also serve in case of dynamic scenes but they are not able to handle the situations where inhibited objects change their locations or when the vision system itself is in motion. The model of [BMB01] relates the inhibitions to features of activity clusters (named as object files) hence inhibition can track an object while the later changes its location. As the information contained by object files are related to activity clusters rather than objects themselves hence the scope of dynamic

inhibition becomes very limited. The model of [Ats07] utilizes a queue of inhibited points to maintain inhibition in dynamic scenes. The information stored in the queue is pixel oriented data rather than object/region features hence it may be considered conceptually similar to the approach proposed in this dissertation but structures of the two approaches are totally different.

2.7 Modeling of Top-Down Attention

Early computational models of visual attention such as [IKN98] and [IK00] have proposed a comprehensive mechanism for determining bottom-up saliency using some feature channels and they use the same bottom-up saliency maps for search task as well. They apply high weight to the feature channel that facilitates highlighting the search target. Even the recent developments by the same group in this context, such as [NI05] and [NI06a], apply a similar strategy. The model of [FBR05] determines weights for the feature maps that would highlight the target in a learning stage and applies them in the searching stage. Although [MGS$^+$04] has separate components for bottom-up and top-down pathways in the model the same saliency maps are used to deal with the top-down pathway. The model presented in [HW06] also applies attentional bias towards the target by learning weights for the conspicuity maps that would make the required object prominent. Such approaches are likely to show inefficiency when distractors are also salient in the same feature channel.

The work presented in [TTW01] has provided a search mechanism to detect the target by looking for its constituent parts. This approach can be considered similar to fine-grain search but the methodology is inclined towards pure machine vision rather than following a biologically inspired approach. Using gist of the whole view to apply a top-down influence to restrict search locations as proposed by [PI07] is also a useful concept that can accelerate biologically plausible visual search. This concept deals with signature of the whole image rather than individual items. These signatures are used for estimating the identity of scene in order to bias the saliency on the locations that have high probability of existence of the target, for example probability of finding people in a beach scene is higher along the coast line.

2.8 Visual Behaviors in Attention Models

Some of the existing models have considered the affect of active visual behavior on the output of visual attention. A brief introduction of some of the commonly known visual behaviors was already provided in chapter 1 section 1.4.5. Here we recapitulate the behaviors implemented by the existing models. The well-known model discussed in [IKN98] and [IK00] mainly deals with *search* behavior but uses bottom-up procedure for this purpose. The selective tuning model [RT06] remains in a behavior resembling *explore* as it does not apply top-down conditions to excite the target of search and lets the salient items pop-out during a process of bottom-up saliency and inhibitions. The models of [PSL02] and [MCBT06] are restricted only to *explore* while the model given in [SF03] discusses both *explore* and *search* behavior by integrating bottom-up and top-down biasing in the process of hierarchical selectivity. The model of [BMB01] considers three behaviors of *explore*, *search*, and *detect changes* while [FBR05] implements *explore* and *search* for dynamic scenes.

2.9 Segmentation

We include a review of color segmentation approaches here because an innovative method for clustering is also developed during the course of this project. This review will help in developing an awareness about the existing techniques and judging the novelness of the proposed segmentation approach presented in section 3.1 of chapter 3.

Work on color segmentation has a history of almost three decades. The techniques proposed for this purpose may be categorized in four groups namely pixel-based, edge-based, region-based, and model-based [IPV00]. Pixel-based techniques group pixels into regions only on the basis of their color features without using the spatial context. Edge-based methods find region boundaries by locating discontinuities of segments in the image. Region-based algorithms take into account both color features and spatial constraints for construction of regions. In model-based schemes image regions are modeled as random fields and the segmentation problem is posed as a statistical optimization problem.

Early publications on color segmentation such as [OKS80] discussed color features that could be used for pixel comparison in segmentation. Histogram based techniques such as [FH04], divide the image into regions by applying thresholds on peaks of color histograms. The method of applying a quantization first and then segmenting spatially, as in [DM01], has been a common practice. In [CMKG03], a similar approach has been proposed involving color clustering and then merging clusters based on color similarity and spatial adjacency. The technique given in [LT04] constructs coarse regions first using a threshold on color distance in RGB space and then detailed segmentation using an irregular pyramid structure. Edge based techniques such as [HB90] and [Sin99] have a common problem that they fail to take into account the correlation among the color channels and miss certain crucial information revealed by color [IPV00]. Region-based techniques work best on images with an obvious homogeneity criterion and tend to be less sensitive to noise [CJSW01]. There are two typical approaches available under this category. One is to use region split-and-merge as in [OPR78] and other is region-growing, for example [TB97]. The later usually applies a merging step to combine segments having further similarity of color features.

Under the model-based category, the Markov Random Field model has been applied as in [DC04] and [Muk02] achieving a good quality of color and texture based segmentation. In [SSA04] a color-based segmentation approach is given for extracting regions of human skin from scenes. They use a second order Markov model on a HSV histogram. Model-based techniques forgo computation time for the quality hence are usable only in those cases where computational complexity is not an issue.

Region-based techniques are considered better when processing speed is a major issue as they provide acceptable quality within a reasonable computation time [CJSW01] [IPV00]. Region-growing methodology has a natural computational advantage over its split-and-merge counterpart in the same category. In terms of color quantization for segmentation, HSI space is considered most appropriate [LB01] and it has provision of overcoming the illumination effects such as shades and shadows.

Graph-based techniques, such as [FH04], can also produce good results. These techniques arrange regions as vertices of a graph where an edge between two vertices reflects the difference of color attributes between them. The idea of integrating the checks for boundary crossing and region expandability, as performed by the proposed approach, can also be found in [TA97] and [FYEA01]. In [TA97] a nonlinear transform is used for finding the attraction force on pixels that is exerted on them by the neighboring regions. This extracts structures from the given image at multiple scales and detects regions and edges in the transformed domain. This technique is able to handle only grey scale images and make use of computationally heavy processes. In [FYEA01] color edges in YUV space are obtained to get the major geometric structures in an image and the centroids between these adjacent edge regions are taken as the initial seeds for seeded region growing.

The concept of categorizing the whole color spectrum into a few classes, as also done in the proposed segmentation method (see section 3.1 of chapter 3), has been used in [BKV05] for developing a query-by-color method that takes into account the human cognition capabilities. Their concept is based upon the psychological findings that humans can perceive colors in so called focal color categories labeled after the colors that humans consider while thinking and speaking, namely, red, green, blue, yellow, orange, brown, pink, purple, black, white, and gray. They also segment the hue-intensity plane into eight regions representing the chromatic colors from the said list. Color categorization in normalized RGB space was done in [LB01] for segmenting objects in a RoboCup soccer field. They propose to put the possible variations of the colors into a lookup table and decide the class of input pixels by comparison to these values. The scope of this approach is limited to distinguishing regions with one of the three colors: blue, yellow, and orange.

2.10 Analysis

Existing literature on artificial visual attention focuses on mimicking the natural attention mechanism by modeling the theories from research in human vision, neurobiology, psychology, and other related disciplines. The procedures proposed by these models are mostly computationally expensive as the correlated processes

carried out in the brain are significantly complex. Such efforts do provide a good representation of the natural organisms but the requirement of computational resources needed for these modeled procedures may turn into an overhead for the vision system instead of enhancement in the overall efficiency through utilization of attention. Hence, there is a need to investigate ways to reduce the complexity of the computational model of attention and enable it to integrate with other vision algorithms so that the output of attention may become an effective contributor for improving performance of the vision system. In this chapter, literature on the theoretical and computational models of attention has been reviewed in order to have a picture of the state-of-the-art in this field. This review will also help in critical evaluation of the innovations proposed in the new model designed in the work under discussion.

This review has provided sufficient experience towards development of an attentive vision system able to operate in real time, which is the main objective of the work under discussion. The target robotic vision system will need to process more feature channels as input will be dynamic leading to variations in acquired features due to changing illumination conditions, motion blurr, and occlusions of objects. Having more channels in hand will reduce dependancy on certain features. A mobile system will need to deliver results of visual processing at a rate of multiple frames per second, therefore the feature processing and attention mechanism needs to be optimized for computation time. In order to make the attention process reliable, the future strategy of model design would be to incorporate all attributes of the existing paradigms discussed in section 2.3 because the required attention system should be able to perform feature processing as in saliency based models, determine rarity in terms of individual features as well as conjunctions to formulate objects (as done in rarity based and object based groups respectively), and construct a hierarchical connectivity as done by the connectionist models. After having a look into the fundamental procedures in sections 2.4 and 2.5 ranging from feature computations, through saliency detection up to inhibition of return, the option of early clustering seems to be a suitable direction that should be investigated to advance towards the ultimate goal of this research. The literature on theory of top-down attention reviewed in

2.2.4 clearly indicates that this pathway should be modeled based upon fine-grain feature maps in order to make it efficient and effective.

The aspect that the existing attention models generally lack is the integration of the involved attention pathways and adaptation of a unified attention system to a certain visual behavior. The biological attention system, on the other hand, has an integrated mechanism that activates the top-down and bottom-up pathways in required configurations to adapt the same system to the active visual behavior. Secondly, natural visual attention is able to perform inhibition of return in three dimensional world despite its perception of a two dimensional projection on the retina. This ability to handle objects in three dimensions is evident from its ability to fixate upon and track objects in space that undergo overlaps while moving. Also, the inhibition or facilitation of return remains effective even the previously focused object changes its position in the recent view frame. This capability of IOR can not be managed with the two dimensional inhibition maps as used by the contemporary models of attention. The pixel based processing, as done by most of the existing models, requires longer computation time and does not support shape-based feature extraction at the pre-attention stage. Clustering of pixels prior to attention processing not only accelerates the computation due to significant reduction of data to be processed but preserves shape features as well in order to enable involvement of more feature channels. In context of top-down attention, the trend of models is shifting towards fine-grain search but contruction of purely fine-grain saliency maps has not been considered by the models until the time of writing of this documentation. The work presented here makes advancements in the state-of-art by proposing solutions for the said issues along with the innovations in the design of attention system in order to make it fast and robust for use in mobile vision systems.

3 Proposed Region-Based Saliency Maps

Focus of this research is to investigate the early clustering approach for the purpose of visual attention hence the basic unit to be processed by the attention procedures will be regions constructed through a segmentation process. In order to optimize the clustering step according to needs of the attention model a new segmentation algorithm has been developed that produces regions in accordance with human perception and provides the output in form of a data structure that facilitates further procedures of the model. This chapter explains the procedure designed to obtain a list of regions from the raw input, extract the feature magnitudes, and compute the saliency of each region in terms of five different features, namely, size, color, eccentricity, orientation, and symmetry.

3.1 Transformation into Region List

An obvious target of a model of artificial attention is to produce results comparable to human (or natural) behavior. In order to remain close to the natural domain the basic step of region construction has to be done in accordance with the human perception as well. Segments that represent regions as perceived by human vision will be useful for a robust and faster artificial attention system. The first objective for such a segmentation is to construct regions that largely correspond to the shapes of actual objects in the image. This can be achieved with optimal tolerance to illumination effects so that neither too many regions are produced for a single object having variations of a uniform color nor distinct regions get merged into one. Secondly, in many situations objects with similar colors overlap each other and create a challenge of discriminating them without going into over-segmentation. The third objective is to complete the segmentation step in a minimum possible time so that enough time is left for the other procedures of visual attention and recognition etc. For this reason we

avoid using the existing model-based statistical techniques that produce quite good results, especially on textures, but require significantly long time to complete their processing. Another requirement is to be able to process a variety of input images without needing to tune the parameters in the meantime because a mobile vision system is expected to wander in unknown scenarios undergoing illumination variations where parameter adjustment will not be possible for each situation. The method provided here has emerged as evolutionary development from its experimental versions presented in [ASMM05], [ASM05a], and [AM06].

3.1.1 Selection of Color Space

Segmentation of color images has to follow a more complex approach as compared to gray-scale segmentation. In gray-scale, only intensity information is available for which the computer can discriminate 256 levels. On the other hand colored images contain compound structure comprising of at least three components, leading to millions of colors per pixel. These components have different meanings and roles in different models, called color-spaces, such as RGB, CYMK, YUV, HSI, and CIELAB etc. Each space has advantages and limitations in terms of segmentation and selection of the best color space is one of the major difficulties in color image segmentation and the decision is mostly made depending upon the requirements of target application. A survey of color spaces and their role in region segmentation can be seen in [CJSW01].

The RGB color space is used for display devices where colors are produced by emission of light whereas CYMK is utilized in printing systems where colors are produced by inks that absorb certain wavelengths while reflecting some other. These spaces are suitable for synthesis but not appropriate for analysis of colors as some shade of the same color may have a fairly distant representation making it difficult to consider them as similar. YUV space is meant for television devices and is hardly used for image processing purposes. HSI (Hue-Saturation-Intensity) space, sometimes also called HSV (Hue-Saturation-Value) or HSL (Hue-Saturation-Luminance), has a good representation of the colors of human perception [CL94] and has good capability of dealing with highlights, shades, and shadows [TT92]. The CIELAB space is a peceptually uniform representation of color but its values do not define absolute colors unless a reference

white point is specified whose definition is assumed to follow a standard and is not explicitly stated [LJBV05].

The main objective of the segmentation step for the purpose of visual attention is to obtain regions from colored images such that the regions may have minimum influence of shades and shadows so that a uniformly colored surface does not get split into many segments. Similarly colored regions in a neighborhood can lead to incorrect results of color contrast and loss of actual object shape would cause errors in computation of shape-based feature maps. As features of the HSI space suite the requirements of our application area hence it is selected for the proposed segmentation algorithm.

3.1.2 Technique Foundation and Innovations

As the target application of our segmentation routine is a biologically inspired attention system, the factor of human perception will have a prominent influence on each step of the process. We would like to build homogeneous segments in a given scene that are potentially distinct regions for humans. Surveys on the color segmentation techniques, for example [SK94] and [CJSW01] show that the area-based segmentation gives a good optimization of segmentation quality and processing time. In the area-based category, the region growing procedure has the advantage of computational simplicity over the split-and-merge strategy. The same surveys establish the advantage of using hue for discounting illumination effects like shading, shadowing, and highlights according to human perception over its other counterparts. As our problem domain has a severe restriction of computation time, we infer to select a region growing method using HIS (hue, intensity, saturation) color space in the proposed segmentation technique. However we suggest introduction of some innovative enhancements in order to improve the quality of output without causing much escalation in the computation time.

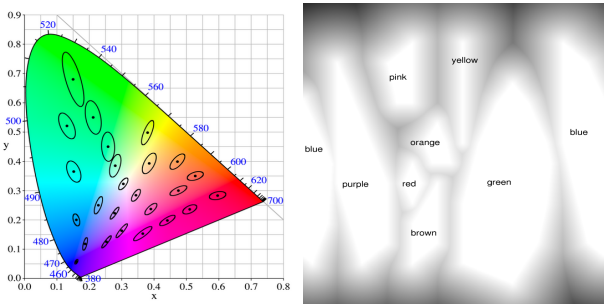
A typical region growing procedure compares the color of the region seed with the expected members of the region. This homogeneity criterion allows high degree of tolerance to color variations when a large threshold is applied on the allowed color difference. Although this is effective for obtaining fine segmentation quality, especially in real-life images, but this way under-segmentation is likely to occur. For example objects separated by a small fluctuation of color, for instance an

object overlapping another similarly colored object, cannot be distinguished. On the other hand, reducing the said threshold in order to deal with such situations leads to over-segmentation of smooth sections of the given input. We propose to include a continuity criterion between adjacently neighboring members of the considered region along with testing the homogeneity condition between the seed and the neighborhood. The other proposition is to have different sets of thresholds to deal with different ranges of colors. We discuss these factors in detail in the following subsections.

3.1.3 Color Dependant Thresholds

Segmentation algorithms depend on the values of thresholds or parameters that are used for their optimum functioning. Normally, these parameters have to be tuned, manually or automatically, according to the nature of the given input. The general practice is to have a single set of thresholds that deals with all types of objects existing in the given input but we argue that it is advantageous to adapt the thresholds according to nature of the seed color.

The human vision has diverse abilities of percieving different color categories. This phenomenon was reported in psychological experiments by MacAdams [Mac42] [Mac49] on the chromaticity diagram that represents hue and saturation attributes of colors. Ellipses were drawn on this diagram in order to represent the range of variations that are accepted as a single color by human observers, as shown in figure 3.1(a). It is easily noticeable that the ellipses in the area for green (tip in the top region) are larger in size as compared to those in other areas such as red (corner region at right) and blue (corner region at left-bottom). Hence the human vision tolerates greater variations in color components on a green object as compared those on a red or blue object. A similar discrimination in the nature of color perception have been reported in [BKV05]. Hue-Intensity is plotted and then segmented into regions representing the eight focal colors of human perception as shown in figure 3.1(b). It is again apparent that green covers a larger area than all other colors. The disparity of size for blue in the two experiments is due to the difference of color space and the level of saturation used by the two analysts. The conclusion that can clearly be derived is the fact

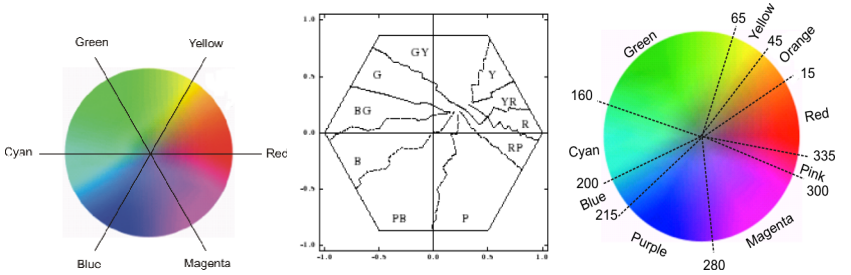


(a) (b)

Figure 3.1: (a) Chromaticity diagram with MacAdam's ellipses. The horizontal and vertical axes represent the x and y components of the CIE XYZ color space respectively. Wavelengths, in nanometers, of the saturated colors are specified on the boundary of the horseshoe (b) A visualization of the results of experiments as reported in [BKV05] on categorization of HSI color space into 9 named colors percieved by humans. The figure is included here with permission of the respective authors.

that each color range from the whole spectrum needs a different set of thresholds in order to optimize the separation of regions.

The next question is to limit the number of color categories for each of which an independent set of thresholds is to be formulated. We prefer performing categorization based upon the hue component of the seed color because it is computationally feasible. A usual division of the hue cycle is done by making six chunks of 60 degrees each and naming them according to the primary colors: red, yellow, green, cyan, blue, and magenta respectively, starting the red at zero degree, as shown in figure 3.2(a). Another division of the hue circle into ten named color categories has been done in [HE96] where the groups of hue angles are made under the names red, red-yellow, yellow, green-yellow, green, blue-green, blue, purple-blue, purple, and red-purple after experiments on human subjects as shown in figure 3.2(b). Their division also shows bigger pie slices for green (and its derivatives) and purple with smaller slices of different sizes for other colors. Combining these divisions and naming conventions with the



(a) (b) (c)

Figure 3.2: (a) The hue circle with angles of basic colors. (b) Hue cycle divided as done in [HE96] (c) Hue-saturation circle divided into nine slices of chromatic colors

focal color categories of [BKV05], we decide to categorize the hue angles of the chromatic colors into nine groups under the names red, orange, yellow, green, cyan, blue, purple, magenta, and pink, each having a different size of span in the hue circle. Figure 3.2(c) shows the ranges of hue angles during which the color remains under the same name for a human observer, represented by pie slices of different sizes for each group. It may be noted that we do not intend to segment objects possessing only these colors; rather we want to pick a set of thresholds that suits the nature of the related color. If, for example, the color of a seed pixel lies at an angle at boundary of two categories then the thresholds may allow construction of a region that has a mixture of the two colors residing at both sides of this boundary angle.

Keeping in view the the sensitivity of human vision to different saturations of the named colors as visible in McAdam's ellipses and the sensitivity to different intensities as reported by [BKV05] we summarize our conclusion in table 3.1, where high sensitivity means that a small variation in the concerned color channel will make a different color for the eye and low sensitivity means that it will seem consistent for human perception even with a high variation in that channel. The main purpose of this analysis is to get a good estimate about the tolerance to variations in intensity and saturation as a nature of each color in order select appropriate general-purpose thresholds that would work on a vast variety of

scenes. For the colors having a high sensitivity to a color component we will have to assign a small threshold to that particular component because regions having these colors should be split even by a small fluctuation in that component. On the other hand, a high threshold value is to be given to a component of a particular color that has low sensitivity in that component. For example, the threshold for tolerating intensity variations in green has to be kept high because a green region will still be counted as green even with lots of variation in its intensity (green has low sensitivity in intensity channel). Otherwise if a green region is treated with a low threshold in intensity it would lead to over segmentation splitting straightforward single regions into many. On the other hand a red region should be allowed to split even with a small variation of intensity because human vision considers red regions as separate even when small fluctuation of intensity exists between them (red is highly sensitive in intensity channel).

Table 3.1: Summary of sensitivity of human vision to intensity and saturation variations in named chromatic colors extracted from MaAdam's ellipses and the analysis done in [BKV05].

Color	*Intensity*	*Saturation*
Red	high	high
Orange	high	high
Yellow	medium	low
Green	low	low
Cyan	medium	medium
Blue	high	low
Purple	high	low
Magenta	high	low
Pink	high	medium

There are six thresholds for each color considered here that need to be adjusted depending on the nature of seed color. Three of them control the allowed amount of color difference between the seed and the other members of a region. The rest of the three are related to the allowed color fluctuation at the border of a region. Γ^h denotes the maximum hue difference that can occur between the seed and other pixels of a region while τ^h is the maximum amount of hue difference that can be tolerated between two adjacent neighbors of a region before declaring

that they belong to different regions separated by an edge. Similarly, Γ^i is the maximum intensity difference that a region pixel can have from the seed while the intensity difference above τ^i between two adjacent pixels of a region will mean that we have reached the edge of that region. Likewise Γ^s and τ^s are thresholds for allowed saturation differences from the seed and at the edge respectively. Γ^h gets a high value for colors that have a larger span in the hue cycle and a smaller value for the colors with the shorter intervals. Giving half of the value of Γ^h to τ^h has shown success in determining hue boundaries. The values of Γ^s and Γ^i, are set to high, medium, or low amounts for colors having low, medium, or high sensitivity to saturation and intensity, respectively. It is evident from experimentation that it is sufficient to give τ^s and τ^i values equal to one third of Γ^s and Γ^i respectively.

3.1.4 Seed classification and two phase operation

The proposed segmentation algorithm works in two phases of operation. In the first phase, chromatic pixels are picked as seeds for region growing that have high amount of saturation and intensity. This condition is limited only to high intensity for achromatic seeds. It facilitates to begin the region growing from prominent portions of objects and helps to position the seeds at central locations so that areas of their corresponding regions are evenly spread around them. Another advantage of this step is that it allows those regions to grow first that have a potential probability of attracting visual attention otherwise there are chances that such regions get merged into other unattractive segments and get neglected in the attention process. In the second phase, the restrictions on saturation and intensity of the seeds are lowered in order to allow the left over areas to get segmented. Values of all concerned thresholds are also relaxed in such a way that the remaining pixels get a higher chance of joining some segmented region. Seeds with black color are not allowed to grow in the first phase due to convergence of all hues and unpredictable behavior of saturation at extremely low intensity. Similarly gray seeds are avoided in the first phase because each color turns into gray at low saturations. Hence they can swallow major parts of neighboring regions with chromatic colors due to overlap of hues at this saturation level. These two colors get their opportunity in the second phase.

The color of seed pixel is classified at two stages. In the first stage, the purpose is to select an appropriate procedure according to the nature of the seed color. At this point the seed is evaluated to see if it is white, black, gray, or with a chromatic color because the process of region growing will be different for each of these seeds. The second stage categorization is needed only for the seeds with chromatic colors where class of the seed's hue has to be determined in order to activate a suitable set of thresholds as discussed in the previous subsection. Hence we define two classification functions $C_1(P)$ and $C_2(P)$ for the two stages where P is the color of a given pixel having the components of hue, saturation, and intensity. $C_1(P)$ classifies the given color as black if the intensity of the given color is below the low intensity for black i_b, as white if the given intensity is above the high intensity for white i_w, as gray when given saturation is below the low saturation s_g where every color turns to gray, and as a chromatic color otherwise. $C_2(P)$ assigns a category to the given hue angle from the previously defined nine classes of chromatic colors according to the value of hue in the given color P.

Let $\mathbf{S}$ be the set of categorized seed pixels in the two phases of operation. A pixel $P(x, y)$ of the input image $\mathbf{I}$ will be selected as a seed $S_p^c(x, y) \in \mathbf{S}$ in the first phase ($p = 1$) if $c = C_1(P)$ is either white or chromatic. A further condition for a chromatic seed in the first phase is that its saturation should be above the high threshold for seed saturation $seed_s$ and its intensity should be higher than the high threshold for seed intensity $seed_i$. In the second phase ($p = 2$), $P(x, y)$ belonging to all four color categories determined by $C_1(P)$ are entitled to be region seeds. The values of $seed_i$ and $seed_s$ are also decreased so that almost all $P(x, y)$ get the opportunity to begin a region.

Another difference between the two phases is that the values of thresholds related to the allowed hue, saturation, and intensity differences set after the color categorization done by $C_2(P)$ are increased in the second phase. In the first phase, the minimum allowed region size is kept higher so that small regions get deleted in order to allow their pixels to join some other region if possible. The value of the allowed size is reduced in the second phase so that even small regions may survive, ultimately minimizing the unprocessed spots in the final output.

3.1.5 Integrated Edge and Region Homogeneity Check

The proposed algorithm integrates the testing of the color homogeneity between the seed and rest of the region pixels with the check to determine if the region border has been reached. This strategy allows the regions to grow with suitable tolerance to illumination and other colors variations and, at the same time, distinguishes the fine edges between them. In order to achieve this we establish four different sets of checks K^c one of which will be activated according to the category of region seed determined by the classification function $C_1(P)$.

For growing a region around a black seed all we need to examine is that the neighborhood should be black. Being a condition to test only one color, there is no need to perform a boundary check. Hence, given the color of the neighbor pixel as N^t and a function $INT()$ that extracts the intensity component of the given color, the set of checks K^c for black seeds is defined as:

$$\begin{aligned} K^c &= \{INT(N^t) \leq i_b\}, \\ & \quad c = black \end{aligned} \tag{3.1}$$

Around a gray seed, we need to construct a region that has a gray shade closely matching the shade of the seed. Therefore, firstly the neighbor pixel being tested should be gray, i. e., its saturation has to be below the saturation for gray s_g. The second test is on the intensity difference between the seed color S and the neighbor color N^t, which should be under a threshold value Γ_g^i, the allowed intensity difference between the seed and other pixels of a gray region. Thirdly, let N^{t-1} be the color of a pixel adjacent to N^t that was made member of the same region before N^t then N^{t-1} and N^t should not form an edge for this region. Hence the intensity difference between N^{t-1} and N^t should be below the allowed gradient at edge τ_g^i. So, having a function $SAT()$ to extract the saturation component from a given color, the set of checks K^c for gray seeds can be defined as:

$$\begin{aligned} K^c &= \{SAT(N^t) \leq s_g,\, INT(N^t) - INT(S) < \Gamma_g^i, \\ & \quad INT(N^t) - INT(N^{t-1}) < \tau_g^i\}, \\ & \quad c = gray \end{aligned} \tag{3.2}$$

For white seeds, the situation is similar to that of black seeds except for the difference in condition in the intensity check that the neighborhood should have an intensity above a high threshold value i_w:

$$\begin{aligned} K^c &= \{INT(N^t) \geq i_w\}, \\ & \quad c = white \end{aligned} \tag{3.3}$$

The set K^c for a seed having a chromatic color consists of nine checks. The first three are meant for stopping the growth of a region if the neighboring pixel N^t is black, gray or white. These checks are important as the gray areas can have arbitrary values in their hue channel while the white and the black regions may contain subjective values for both hue and saturation, hence gray pixels can easily get swallowed when a neighboring chromatic region is growing. The next three checks are to allow small differences between hue, saturation, and intensity components of the seed color S and those of the neighbor N^t. Then the last three checks inspect if two adjacent neighbor pixels in the given region form an edge in terms of hue, saturation, or intensity. Hence, with a function $HUE()$ that extracts the hue component of the given color, this set can be defined as :

$$\begin{aligned} K^c &= \{INT(N^t) < i_w,\, INT(N^t) > i_b,\, SAT(N^t) > s_g, \\ & \quad HUE(N^t) - HUE(S) \leq \Gamma^h,\, INT(N^t) - INT(S) \leq \Gamma^i, \\ & \quad SAT(N^t) - SAT(S) \leq \Gamma^s,\, HUE(N^t) - HUE(N^{t-1}) < \tau^h, \\ & \quad INT(N^t) - INT(N^{t-1}) < \tau^i,\, SAT(N^t) - SAT(N^{t-1}) < \tau^s\}, \\ & \quad c = chromatic \end{aligned} \tag{3.4}$$

Before starting the region growing process, the first stage seed categorization $C_1(I)$ is performed that decides the basic class of the seed color. A proper set of checks K^c is activated based upon this classification and in case of $c = chromatic$ the second stage categorization $C_2(I)$ is executed that decides the chromatic color class to which the given hue angle belongs. An appropriate set of values are loaded into the thresholds Γ^h, Γ^i, Γ^s, τ^h, τ^i, and τ^s based upon the found category.

3.1.6 Region Construction

Region construction is carried out using a usual region growing procedure in which all pixels found eligible to join the region in an 8-connected neighborhood are labeled with the region identity and pushed in a stack. Later the stack is popped and the same procedure is repeated for the popped pixel considered as part of the region. This process continues until the stack gets empty. In order to determine the eligibility of the neighbor N^t to become a member of the current region, a set of attributes $A^c(N^t)$ is constructed for N^t according to the category c of the seed $S_p^c(x,y)$. Having a relation $\otimes$ ($\otimes \in \{>, <, =, \leq, \geq\}$) that can exist between comparable color components of two pixels, depending upon the seed category c an appropriate condition $A^c(N^t)$ with a matching value of c is activated from those given below. $A^c(N^t)$ can have one of the following structures according to the value of c.

$$A^c(N^t) = \{INT(N^t) \otimes i_b\}, \text{ for } c = black \tag{3.5}$$

$$\begin{aligned} A^c(N^t) = \{&SAT(N^t) \otimes s_g,\ INT(N^t) - INT(S) \otimes g_i, \\ &INT(N^t) - INT(N^{t-1}) \otimes g_i\},\ \text{ for } c = gray \end{aligned} \tag{3.6}$$

$$A^c(N^t) = \{INT(N^t) \otimes i_w\}, \text{ for } c = white \tag{3.7}$$

$$\begin{aligned} A^c(N^t) = \{&INT(N^t) \otimes i_w,\ INT(N^t) \otimes i_b, SAT(N^t) \otimes s_g, \\ &HUE(N^t) - HUE(S) \otimes \Gamma^h,\ INT(N^t) - INT(S) \otimes \Gamma^i, \\ &SAT(N^t) - SAT(S) \otimes \Gamma^s,\ HUE(N^t) - HUE(N^{t-1}) \otimes \tau^h, \\ &INT(N^t) - INT(N^{t-1}) \otimes \tau^i,\ SAT(N^t) - SAT(N^{t-1}) \otimes \tau^s\}, \\ &\text{for } c = chromatic \\ &\text{where } \otimes \in \{>, <, =, \leq, \geq\} \end{aligned} \tag{3.8}$$

Region constructed around a seed $S_p^c(x,y)$ will be a set of pixels $I(x,y)$ defined as follows:

$$\begin{aligned} R_i = \{&I(x,y) | (x,y) \in N^t(R_i) \text{ and } A^c(N^t) = \mathbf{K}^c, \forall (x,y) \in \mathbf{I}\} \\ &Ri = S_p^c(x,y) \text{ at } t = 0 \end{aligned} \tag{3.9}$$

where $N^t(R_i)$ is the neighborhood pixels of R_i at time t. The final output of the segmentation procedure is a list of regions $\Re$ consisting of n regions each represented as R_i. Each R_i is coupled with data regarding location, bounding rectangle, and magnitudes of each feature ϕ_i^f $(f \in \Phi)$. As five channels of color, orientation, eccentricity, symmetry and size are considered in the current status of our model hence we have $\Phi = \{c, o, e, s, z\}$.

3.1.7 Segmentation Results

As the segmentation process was not the main topic of this research hence results of this module are provided here separately from the output of visual attention presented in chapter 5. The proposed approach of color segmentation was tested using many artificial and real life images. The results are very encouraging and the segmentation output was found suitable according to the requirements of the region-based attention model. We have also compared our results with some existing segmentation methods that use computationally heavy statistical methods and produced fairly good results for general-purpose segmentation. Figure 3.3 presents results of the proposed and two existing segmentation methods. A qualitative comparison can be done by observing these results. The graph-based method [FH04] has performed very well with the chromatic colors but has flaws in the achromatic areas. For example, it splits the uniform black background of the image in the second column into many regions while it merges the white border line of the road into the gray road in the traffic scene. On the other hand the scale space method [DN04] handles these situations in a better way but it is over segmenting in chromatic regions. Both of the competitive methods are unable to separate the yellow colored melon overlapping the similarly colored banana in the fruit image. The proposed method has shown a good balance in separating distinct regions while tolerating illumination effects on uniformly colored objects. The ability of the proposed method in performing well both for chromatic and achromatic color can be very useful in situations when the mobile vision system goes through low light areas where color distinction based upon hue becomes difficult.

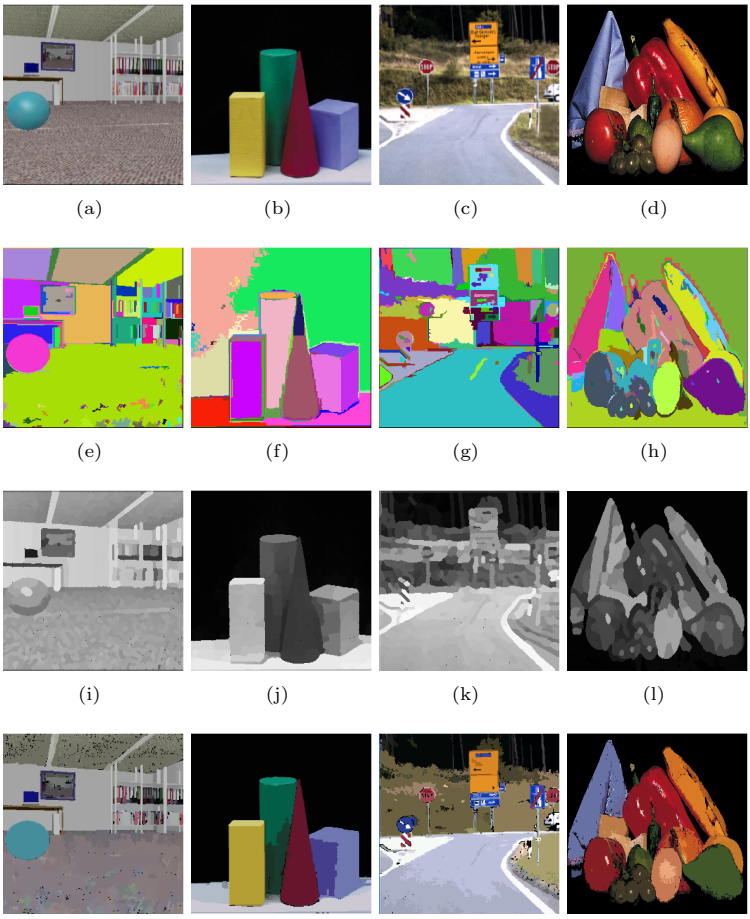


Figure 3.3: Results of segmentation by the proposed segmentation routine and two other segmentation methods. Top row: input images. Second row: segmentation results of a graph-based method [FH04]. Third row: results of a scale space method [DN04]. Bottom row: results of the proposed method.

3.2 Filtration of Useless Regions

Before processing the regions with the saliency computation methods, a filtration process removes those regions from the region list $\Re$ that are useless in the later procedures. The main criterion for this filtration is size of the region relative to the whole image. Regions having very small size are assumed to be result of noise or segmentation error. Similarly large regions covering a significant portion of the image are considered as background. We introduce a factor of region perimeter f_i^p for this purpose that will gain a high value (equal to 1) for regions with moderate size while a small value for very small and very large regions. Scaling the saliency values using this factor will eliminate noise and background from the resulting feature maps. The perimeter of the bounding rectangle of a given region is used instead of the area in order to deal with large porous regions. Such regions would have a large bounding rectangle but a small area covered under their pixels.

Let P_R^i be the perimeter of the bounding rectangle of R_i, P_I be the perimeter of the input image, k_{min}^p be the minimum percentage of the image perimeter below which a region should be neglected, and k_{max}^p be the maximum percentage of the image above which a region should be regarded as background then f_i^t, the unclipped value of perimeter factor, will be computed as

$$f_i^t = k_{scale}^p \frac{(P_R^i - k_{min}^p P_I)(k_{max}^p P_I - P_R^i)}{(P_I/2 + k_{min}^p P_I - k_{max}^p P_I)^2} \tag{3.10}$$

where k_{scale}^p is a scaling constant to bring the highest value obtained from the rest of the expression equal to 1. The resulting values are clipped between 0 and 1 to obtain the final value of f_i^p as follows

$$f_i^p = \begin{cases} 1 & \text{for } f_i^t > 1 \\ 0 & \text{for } f_i^t < 0 \\ f_i^t & \text{otherwise} \end{cases} \tag{3.11}$$

Using the constant definitions as given in Table-3.2 the curve of f_i^p for P_R^i ranging between 1% to 100% of P_I is shown in figure 3.4.

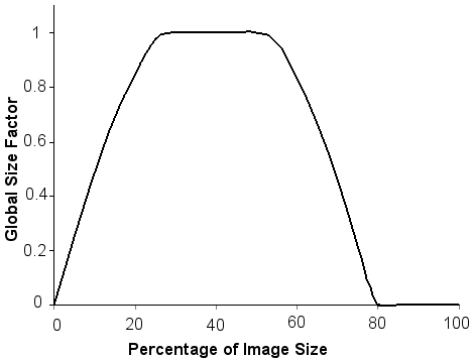


Figure 3.4: Values of perimeter factor f_i^p for regions with perimeters covering different percentages of the image.

3.3 Size Map Construction

Although prominence due to size may be suppressed in presence of high contrast in other visual attributes such as color, saliency in terms of area can play a useful role in situations where a target of attention does not surface due to other features. The size based saliency mainly contributes in suppressing large sized background regions and unnoticeable small sized regions. The contrast of size with respect to the neighborhood needs to be computed to find objects having exclusive size. Figure 3.5 provides examples of two scenarios where the only obvious feature to determine saliency is the size of objects and such uniquely sized objects are the obvious attractors of attention.

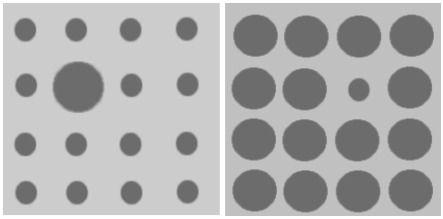


Figure 3.5: Examples of obvious size saliency due to uniqueness in region area.

The exclusiveness of size with respect to the neighborhood and the global context is determined using a voting style mechanism. A given region R_j with a size similar to R_i will not contribute to the size saliency of R_i, one with significantly different size will give a partial supporting vote, and if such R_j surrounds R_i then the contribution is a full support because a situation similar to accent color is developed. Hence V_{ij}^s, the vote of R_j to the size saliency of R_i may be defined as follows:

$$V_{ij}^s = \begin{cases} 1 & \text{when } R_j \odot R_i \text{ and } \alpha(R_i)/\alpha(R_j) \leq k_1^a \\ 0.5 & \text{when } \alpha(R_i)/\alpha(R_j) \leq k_1^a \\ 0.5 & \text{when } \alpha(R_i)/\alpha(R_j) \geq k_2^a \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

where k_1^a and k_2^a are threshold constants, $R_j \odot R_i$ means R_j surrounds R_i, and $\alpha(R_i)$ extracts area covered by the given region. For $R_j \in \eta_i$, the contributions will have a higher weight as compared to the non-neighbor regions. The contributions from $R_j \in \eta_i$ is accumulated into X_1^i and those from the global context are summed up into Y_1^i. Now

$$X_1^i = \sum_{j=1}^{j=p_i} V_{ij}^s, \ \forall R_j \in \eta_i \tag{3.13}$$

$$Y_1^i = \sum_{l=1}^{l=n} V_{il}^s/2, \ \forall R_l \in \Re, l \neq i \tag{3.14}$$

Now, having the maximum amount of saliency value that can be assigned to a region due to a single feature as S^{max}, p_i as the count of neighbors in the neighborhood list η_i of R_i and n as the count of regions in $\Re$, the final value of area saliency S_a^i for a region R_i will be computed as:

$$S_a^i = f_i^p S^{max} \frac{X_1^i + Y_1^i}{p_i + (n-1)} \tag{3.15}$$

3.4 Color Saliency

3.4.1 Color Contrast in Color Theory

Apart from psychology, valuable information can also be found in the literature on color theory about the attributes of colors that contribute in making an object visually prominent or receding. In terms of color saliency, other methods of artificial visual attention have concentrated only on those attributes of colors that were reported in psychology and many important aspects described for this purpose in the color theory have been neglected. Artists practice these aspects for creating effects of contrast, visual advancement, and activeness in their illustrations. Johannes Itten was one of the first experts of color theory who described methods for color combinations offering contrast. He has defined different situations in which the human vision finds contrast in a colored scene. According to his research, the contrast can occur due to presence of objects posing high difference of intensities, saturation, and/or hue. Other reported causes include presence of opponent colors and co-occurrence of warm and cool colors [Itt61].

Another relatively modern source of theoretical concepts on colors is available in [Mah96]. We combine these concepts with those of Itten's and formulate a set of points that are feasible for computation. Another important issue is to decide that which color will receive the benefit of saliency in presence of a contrast. The summarized points with the mention of the saliency winning color in each situation are listed below:

1. *Contrast of Saturation*: A contrast is produced by low and highly saturated colors. The value of contrast is directly proportional to the magnitude of the saturation difference. Highly saturated colors tend to attract attention in such situations unless a low saturated region is surrounded by highly saturated one.
2. *Contrast of Intensity*: A contrast will be visible when dark and bright colors co-exist. The greater is the difference in intensity the higher is the effect of contrast. Bright colors catch the eye in this situation unless the dark one is totally surrounded by the bright one.

3. *Contrast of Hue*: The difference of hue angles on the color wheel contributes to creation of contrast. High difference will obviously cause a more effective contrast. Due to the circular nature of hue, the highest difference between two hue values can be 180°.

4. *Contrast of Opponents*: The colors that reside on the opposite sides of the hue circle produce a high amount of contrast. This naturally means that the difference of the hue angles should be close to 180. The colors residing in the first half of the hue circle, known as the active color range, will dominate on the rest of the passive ones.

5. *Contrast of Warm and Cool*: The warm colors namely red, yellow, and orange are visually advanced. These colors are present in the first 45° of the hue circle. Warm and cold colors create a contrast in which warm colors remain dominant.

6. *Accent Colors*: The color of the object covering a large area of the scene will become the ground color (trivial for attention). Colors covering a small relative area, but offering a contrast, are called accent colors. Accent colors get the benefit of contrast in terms of attracting visual attention.

7. *Dominance of Warm Colors*: The warm colors dominate their surrounding whether or not a contrast in the environment exists.

8. *Dominance of Brightness and Saturation*: Highly bright and saturated colors are considered as active regardless of their hue values. Such colors have more chances of attracting attention.

The effect of contrast is controlled by the saturation value of both of the involved colors in the situations mentioned in points 2 to 5. Highly saturated colors will offer stronger contrast. The attention models of [IKN98], [SF03] and [PSL02] mainly concentrate on points 2 and 4 from the above list by computing feature maps for achromatic opponent colors of black and white while working on chromatic opponents of red-green and blue-yellow. The rest of the existing models compute only the relative difference between the color of current pixel (or region) and their neighborhood. We build our model for determining color saliency of regions by covering all of the facts gathered above.

3.4.2 Color Map Construction

We divide the procedure of color saliency computation into seven steps each of which adds to the saliency magnitude of a region R_i in a voting style mechanism. The sixth point from the above list is concerned with the decision of dominance in context of relative size of differently colored objects, hence it is used within other steps rather than implementing it as a separate one. The first version of this technique was presented in [AM07b].

The first five steps of the algorithm use one or more of the factors of saturation f_{ij}^s, intensity f_{ij}^b and area f_{ij}^a in their calculations. The subscript of each factor denotes that it is effective between regions R_i and R_j whereas R_i will actually use the factor. The first part of the factor of saturation f_{ij}^s is modeled as the mean of the saturation components of R_i and R_j so that higher effect takes place when both regions possess high amount of saturation and vice versa. The second part depends upon only the saturation of R_i and holds a minimum value equal to κ_{min} in order to let the regions with near zero saturation survive. The rest of the second part comes from the saturation of the region color scaled by $\kappa/255$ where $\kappa = 1 - \kappa_{min}$. The factor for intensity f_{ij}^b is computed in a similar fashion using intensity component of the region color instead of saturation. For the area factor f_{ij}^a we use the concept of accent color according to which a small region surrounded by a large one will receive the benefit of color contrast in terms of becoming visually attractive. Hence a region R_i should get the full support of saliency only if it is sufficiently smaller than the neighbor region R_j and is surrounded by the later. The value of each factor is kept to lie between 0 and 1 so that it could play its role as a multiplicative factor in further computations. Having the maximum level of saturation and intensity at 255, $\xi(R_i)$ as the saturation component of the color of R_i, $\Im(R_i)$ as the intensity component, and $\alpha(R_i)$ as the area covered by the pixels of R_i, these three factors are defined as:

$$f_{ij}^s = \frac{(\xi(R_i) + \xi(R_j))/2}{255} \left(\kappa_{min} + \frac{\kappa \xi(R_i)}{255} \right) \tag{3.16}$$

$$f_{ij}^b = \frac{(\Im(R_i) + \Im(R_j))/2}{255} \left(\kappa_{min} + \frac{\kappa \Im(R_i)}{255} \right) \tag{3.17}$$

$$f_{ij}^{a} = \begin{cases} 1 & \text{for } \alpha(R_i)/\alpha(R_j) \leq k_1^a \text{ and } R_j \odot R_i \\ 0 & \text{for } \alpha(R_i)/\alpha(R_j) \geq k_2^a \text{ and } R_i \odot R_j \\ 0.5 & \text{otherwise} \end{cases} \tag{3.18}$$

where k_1^a and k_2^a are constants used as thresholds. The $\odot$ operator indicates that the region mentioned at its left surrounds the one at its right. Two contexts are considered for comparison of color values in the first five steps. First is the local neighborhood in which regions having a common boundary with R_i (members of η_i counted as p_i) are considered. The second is the global context in which all regions in $\Re$ except R_i are used. For each vote V_ς^i for the region R_i in the step $\varsigma(\varsigma \in \{2..6\})$, the part of vote due to the local context is stored into X_ς^i and the part coming in from the global context is stored into Y_ς^i.

The first step collects votes for each R_i from other regions that possess opponent hues. Two hues are said to be opponent if they lie at the opposite sides on the color wheel. In other words the hue difference should be close to 180 degrees. It is a generalization of the red-green and blue-yellow opponents as used by existing methods. Due to the circular nature of the hue we calculate the hue difference between two regions R_i and R_j as:

$$\Delta_{ij}^{h} = \begin{cases} \Delta_{ij}^{\mu} & \text{for } \Delta_{ij}^{\mu} \leq 180 \\ 360 - \Delta_{ij}^{\mu} & \text{otherwise} \end{cases} \tag{3.19}$$

where $\Delta_{ij}^{\mu} = |\mu(R_i) - \mu(R_j)|$, $\mu(R_i)$ being the hue angle of the color of R_i. A region R_j contributes a unit value scaled by the three factors of area, saturation, and intensity to the vote of this step if R_j possesses a color with an opponent hue to the color of R_i. Let Δ_o^h be the minimum hue difference for two colors to be opponents then X_ς^i and Y_ς^i for $\varsigma = 2$ are computed as

$$X_2^i = \sum_{j=1}^{j=p_i} f_{ij}^a f_{ij}^s f_{ij}^b \ \forall R_j \in \eta_i \text{ when } \Delta_{ij}^h \geq \Delta_o^h \tag{3.20}$$

$$Y_2^i = \sum_{l=1}^{l=n} f_{il}^a f_{il}^s f_{il}^b \ \forall R_l \in \Re \text{ when } \Delta_{il}^h \geq \Delta_o^h, l \neq i \tag{3.21}$$

For exact opponents the value of Δ_o^h should be 180 but in order to give a relaxation of 10 degrees at both sides we set $\Delta_o^h = 170$. The second step collects votes from regions that are at large hue distances from R_i. The computation of X_3^i and Y_3^i is performed as follows:

$$X_3^i = \sum_{j=1}^{j=p_i} f_{ij}^a f_{ij}^s f_{ij}^b \Delta_{ij}^h / 180 \ \forall R_j \in \eta_i \tag{3.22}$$

$$Y_3^i = \sum_{l=1}^{l=n} f_{il}^a f_{il}^s f_{il}^b \Delta_{il}^h / 180 \ \forall R_l \in \Re, l \neq i \tag{3.23}$$

The division by 180 is performed to normalize the hue difference to be between 0 and 1. Neighborhood with a high hue difference will contribute more weight to this vote.

In the third step we extend the contrast of warm and cool to contrast of active and passive colors. A color is considered as active if its hue is in the first half of the color wheel. Hence, when R_i has an active color then a region R_j with a passive color will contribute to the saliency of R_i. Higher difference in the hue will obviously make this contrast more prominent. Hence we can model this step as

$$X_4^i = \sum_{j=1}^{j=p_i} f_{ij}^a f_{ij}^s f_{ij}^b \Delta_{ij}^h \forall R_j \in \eta_i \tag{3.24}$$
$$\text{when } \mu(R_i) < 180 \text{ and } \mu(R_j) \geq 180$$

$$Y_4^i = \sum_{l=1}^{l=n} f_{il}^a f_{il}^s f_{il}^b \Delta_{il}^h \forall R_l \in \Re, l \neq i \tag{3.25}$$
$$\text{when } \mu(R_i) < 180 \text{ and } \mu(R_l) \geq 180$$

The fourth step conducts voting for contrast of saturation. Regions possessing highly different saturations in local and global context will add to the saliency of R_i as

$$X_5^i = \sum_{j=1}^{j=p_i} f_{ij}^a f_{ij}^s f_{ij}^b \Delta_{ij}^s / 255 \ \forall R_j \in \eta_i \tag{3.26}$$

$$Y_5^i = \sum_{l=1}^{l=n} f_{il}^a f_{ij}^s f_{ij}^b \Delta_{il}^s / 255 \ \forall R_l \in \Re, l \neq i \tag{3.27}$$

where Δ_{ij}^s is the difference of saturation between regions R_i and R_j. The value of Δ_{ij}^s is divided by 255 to bring the effect of saturation difference between 0 and 1 to keep the contribution from each region within a unit amount.

The fifth step collects votes for R_i from regions having highly different intensity (contrast of intensity). The computations have a similar format as the fourth step, hence

$$X_6^i = \sum_{j=1}^{j=p_i} f_{ij}^a f_{ij}^s f_{ij}^b \Delta_{ij}^b / 255 \ \forall R_j \in \eta_i \tag{3.28}$$

$$Y_6^i = \sum_{l=1}^{l=n} f_{il}^a f_{ij}^s f_{ij}^b \Delta_{il}^b / 255 \ \forall R_l \in \Re, l \neq i \tag{3.29}$$

where Δ_{ij}^b is the difference of intensity (brightness) between regions R_i and R_j.

Keeping in view that p_i regions have given votes in the local context and $n-1$ regions have contributed in the global context of the first five steps, the resultant magnitude of each vote V_ς^i to R_i in these steps will be

$$V_\varsigma^i = \frac{X_\varsigma^i + Y_\varsigma^i}{p_i + (n-1)} \ \forall \varsigma \in \{2..6\} \tag{3.30}$$

The warm colors consisting of the red, orange, and yellow ranges are given an extra vote to strengthen their color saliency in the sixth step. These ranges reside in the first 45 degrees of the color wheel. Hence

$$V_7^i = \begin{cases} \Im(R_i)\xi(R_i)/(255)^2 & \text{for } 0 \leq \mu(R_i) < 45 \\ 0 & \text{otherwise} \end{cases} \tag{3.31}$$

Finally, the seventh step supports saliency of highly bright and saturated colors. These color components of R_i are scaled to a unit value to determine the weight of the seventh vote

$$V_8^i = \Im(R_i)\xi(R_i)/(255)^2 \tag{3.32}$$

Let δ be the part of full weight vote that a step can contribute to the color saliency of R_i, then having restricted the maximum value of feature saliency to S^{max}, the value of δ turns out to be $S^{max}/7$. Hence the resultant color saliency S_c^i of R_i is computed as

$$S_c^i = f_i^p \sum_{\varsigma=2}^{8} \delta V_\varsigma^i \tag{3.33}$$

3.5 Shape Based Saliency

3.5.1 Symmetry Magnitude

Computing local as well as global symmetry is a computationally expensive process. An efficient solution is needed in order to bring this feature into the attention pipeline while keeping the overall processing time within the rate of multiple frames per second. The proposed algorithm accelerates the computation by its innovative simplified design and determines the value of symmetry for a given region with a reasonable accuracy. A further speed up is obtained for the purpose of attention by limiting the amount of precision to the level required by this application. In this subsection we present the novel method of determining symmetry magnitude of each region and the procedure to construct the saliency map using these values will be discussed later.

Let $\Psi(L, P_s)$ be a scanning function that counts the symmetric points around a given point P_s along a line L by investigating pairs of points each denoted as $\{a^t, b^t\}$ on it (see figure 3.6). L is one of the line segments perpendicular to the axis A_θ^s around which symmetry is being evaluated. L intersects A_θ^s at P_s and the points a^t and b^t are equidistant on opposite sides of P_s. The scan starts with a^t and b^t at unit distance from P_s and continues with an increment of one in the distance. The scan stops when anyone of a^t and b^t touches the bounding rectangle of the region being scanned. The count of symmetric points is incremented when both a^t and b^t belong to the same region.

The set of lines and points that provide input to $\Psi(L, P_s)$ is generated by first taking a desired number of reference lines A_θ^s by calculating their end points

incident on the bounding rectangle of R_i. Here θ reflects the angle of the line and in a typical case $\theta \in T$ where $T = \{0, 15, 30, 45, 60, 75, 90, 105, 120, 135, 150, 165\}$ to examine symmetry of a given pattern around twelve axes. Figure 3.6 shows four axes around which symmetry is computed for the attention purpose. Each point on every A^s_θ is parsed as P_s and the end points of the line L on the bounding rectangle of R_i are generated for each P_s.

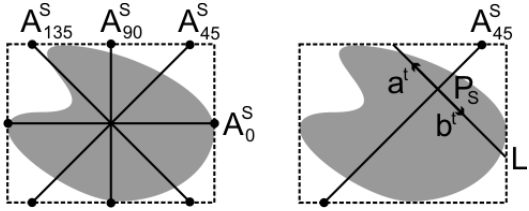


Figure 3.6: (Left) Four of the scan axes around which symmetry is computed for purpose of attention. (Right) Scan method to find symmetric points on the region along a line perpendicular to the axis of symmetry.

Having the count of symmetric points along one scan line as $\Psi(L, P_s)$, the measure of symmetry around an axis A^s_θ is computed as

$$M^\theta_s = \sum_{s=1}^{l} \Psi(L, P_s)/\alpha(R_i) \tag{3.34}$$

where l is the length of the investigated axis A^s_θ. Division by the area of given region is performed to normalize the result between 0 and 1. The total measure of symmetry for R_i with respect to all considered axes will be

$$M^i_s = \sum^{\forall \theta \in T} M^\theta_s / n_T \tag{3.35}$$

where n_T is the count of members in T

The method described so far can be useful for computing local symmetry and can be generalized to evaluate global symmetry if the complete image is taken as input and similarity criterion is set to color similarity of a^t and b^t. Although gain in computation speed is achieved due to the simple structure of the proposed algorithm but a further decrease in time consumption can be achieved for the

purpose of attention as only an approximation of symmetry values is needed. This can be done by reducing the iterations by taking a small subset of T for the computations. We use a set $T_4^i \subset T$ by picking only four angles for symmetry axes. The members of T_4^i include angles starting from the inclination angle of the major axis (orientation) of R_i and three other angles each at an increment of 45, 90, and 135 degrees respectively from the first. Hence, for example, a region inclined at approximately 30 degrees will have $T_4^i = \{30, 75, 120, 165\}$. This estimation of symmetry against four selected angles is sufficient to construct a feature map of symmetry in the attention model. As the orientation of R_i comes as a byproduct during construction of the orientation feature map (discussed later), hence it poses no overhead on the overall performance of the attention model.

3.5.2 Region Angle and Eccentricity Magnitudes

We use a traditional technique involving moments for finding the orientation and eccentricity of regions. Later the feature values are used to determine saliency with respect to these features. Three types of discrete two dimensional moments $m_{1,1}^i$, $m_{2,0}^i$, and $m_{0,2}^i$ are computed for each R_i as follows

$$m_{1,1}^i = \sum (x - \overline{x})(y - \overline{y}) \ \forall (x, y) \in R_i \tag{3.36}$$

$$m_{2,0}^i = \sum (x - \overline{x})^2 \ \forall (x, y) \in R_i \tag{3.37}$$

$$m_{0,2}^i = \sum (y - \overline{y})^2 \ \forall (x, y) \in R_i \tag{3.38}$$

where $(\overline{x}, \overline{y})$ is the center of R_i. Now the orientation ϕ^i and eccentricity ϵ^i are computed as

$$\phi^i = \frac{1}{2} tan^{-1} \left(\frac{2m_{1,1}^i}{m_{2,0}^i - m_{0,2}^i} \right) \tag{3.39}$$

$$\epsilon^i = \frac{(m^i_{2,0} - m^i_{0,2})^2 + 4m^i_{1,1}}{(m^i_{2,0} + m^i_{0,2})^2} \tag{3.40}$$

3.5.3 Shape Based Feature Maps

After having computed the values of feature magnitudes for the rest of the three features, the remaining task in building the saliency maps is to determine exclusiveness with respect to these features in local as well as global context. As the prominence of a region mainly arises from the rarity of its feature values [Ner04], hence the saliency of an object with respect to a particular feature decays strongly when another object with a similar value of that feature exists is the near surrounding. A weaker decline of saliency occurs in case of similarity existing outside a certain radius r^{fovea}.

We adopt an incremental approach for saliency with respect to orientation S^i_o as we model no bias to any particular angle, hence S^i_o initializes with a zero for every region and increments when other regions differ in orientation. On the other hand, an adversative approach is adopted for the other two features because high amount of symmetry and eccentricity are attractors of attention [WH04] [BMB01]. Hence, a head start is given to regions having high amount of symmetry or eccentricity and their saliency is decremented when similarly featured regions are found in the neighborhood. S^i_e and S^i_s, the saliencies of R_i with respect to eccentricity and symmetry, are initialized with the feature magnitudes ϵ^i and ϕ^i respectively (both range between 0 and 1). A multiplicative decay is given to these saliency values when a neighbor has similarity of attributes. We define v^o_{ij}, v^e_{ij}, and v^s_{ij}, the contributions of region R_j to the saliency of R_i in terms of orientation, eccentricity, and symmetry respectively as:

$$v^o_{ij} = \begin{cases} \Delta^\phi_{ij}/90 & \forall \Delta^\phi_{ij} > k_\phi,\ \|R_i R_j\| < r^{fovea} \\ \beta_o \Delta^\phi_{ij}/90 & \forall \Delta^\phi_{ij} > k_\phi,\ \|R_i R_j\| \geq r^{fovea} \\ 0 & \text{otherwise} \end{cases} \tag{3.41}$$

$$v^e_{ij} = \begin{cases} \beta^1_e & \text{for } \Delta^e_{ij} < k_e \ \&\ \|R_i Rj\| < r^{fovea} \\ \beta^2_e & \text{for } \Delta^e_{ij} < k_e \ \&\ \|R_i Rj\| \geq r^{fovea} \\ 1 & \text{otherwise} \end{cases} \tag{3.42}$$

$$v_{ij}^{s} = \begin{cases} \beta_s & \text{for } \Delta_{ij}^{s} < k_s \\ 1 & \text{otherwise} \end{cases} \tag{3.43}$$

where Δ_{ij}^{ϕ} is absolute difference of orientation between R_i and R_j and it is divided by 90 to bring the outcome within a unit amount. This is done so because equation (3.39) always gives ϕ_i between 0°and 180°which means that the maximum difference of orientation between two regions can be 90°(the smaller angle between two straight lines). $\|R_i R_j\|$ is the distance between the centers of R_i and R_j. Δ_{ij}^{e} and Δ_{ij}^{s} are absolute differences between eccentricity and symmetry of the two considered regions while k_ϕ, k_e, and k_s are thresholds and β_o, β_s, β_e^1 and β_e^2 are constants having values less than 1. See Table 3.2 for values of these constants used in our experiments. Now S_o^i, S_e^i, and S_s^i are obtained as:

$$S_o^i = \sum_{j=1}^{j=n} v_{ij}^{o}, \;\; S_e^i = \epsilon^i \prod_{j=1}^{j=n} v_{ij}^{e}, \;\; S_s^i = M_s^i \prod_{j=1}^{j=n} v_{ij}^{s} \tag{3.44}$$

3.6 Top-Down Saliency Maps

As mentioned in section 2.2.4, the findings from experiments on human vision by [LD04], [LHG97], [Ham05], [Dec05], and [NI06b] suggest that the top-down saliency mechanism constructs task dependant maps to allow quick pop-out of the target rather than using the bottom-up saliency maps, hence we propose to model the top-down pathway independent of the bottom-up process. The top-down pathway uses the magnitudes of the same features ϕ_i^f associated with each region R_i regarding color components (hue, intensity, and saturation), size, angle of orientation, eccentricity, and symmetry which were computed during the bottom-up processes ($f \in \Phi = \{c, o, e, s, z\}$).

For the construction of fine-grain saliency maps for each feature channel f considered in the model, the search target is defined as a set of top-down feature values F_{td} in which the individual features are referred as F_{td}^f. For constructing the saliency map with respect to color ($f = \{c\}$), we define D^h as the difference of hue that can be tolerated in order to consider two colors as similar, D^s

as the tolerable saturation difference, D^I as the allowed intensity difference for equivalent colors, and ϕ_i^c as the magnitude of the color feature for R_i. Now, the top-down color saliency γ_i^c of each region R_i is determined as follows:

$$\gamma_i^c = \begin{cases} \frac{a(D^h - \Delta_i^h)}{D^h} + \frac{b(D^s - \Delta_i^s)}{D^s} + \frac{c(D^I - \Delta_i^I)}{D^I} & \text{for } \Delta_i^h < D^h \text{ \& chromatic } \phi_i^c, F_{td}^c \\ \frac{(a+b+c)(D^I - \Delta_i^I)}{D^I} & \text{for } \Delta_i^I < D^I \text{ \& achromatic } \phi_i^c, F_{td}^c \\ 0 & \text{otherwise} \end{cases} \tag{3.45}$$

where a, b, and c are weighting constants to adjust the contribution of each color component into this process. Δ_i^h, Δ_i^s, and Δ_i^I are magnitudes of the difference between ϕ_i^c and F_{td}^c in terms of hue, saturation, and intensity respectively. We take $a = 100$, $b = 55$, and $c = 100$ because the saliency values of a region lie between the range of 0 and 255 in our model. The value of b is kept smaller in order to keep more emphasis on the hue and intensity components. Hence a perfect match would result in a saliency value equal to 255.

The color map had specific requirements being a composite quantity whereas the other feature channels consist of single-valued quantities; hence they can be processed using a simpler procedure. Having Θ^f as the normalized ratio of the feature magnitudes ϕ_i^f and F_{td}^f (for $f \neq \{c\}$) defined as

$$\Theta^f = \begin{cases} \phi_i^f / F_{td}^f & \text{for } \phi_i^f < F_{td}^f \\ F_{td}^f / \phi_i^f & \text{otherwise} \end{cases} \tag{3.46}$$

which always keeps $1 \geq \Theta^f \geq 0$, the top-down saliency γ_i^f of a region R_i with respect to a feature f ($f \in \Phi, f \neq \{c\}$) will be computed as

$$\gamma_i^f = \begin{cases} k\Theta^f & \text{for } \Theta^f > D^\Theta \\ 0 & \text{otherwise} \end{cases} \tag{3.47}$$

where k is a scaling constant and D^Θ is the ratio above which the two involved quantities may be considered equivalent. Values of constants introduced in these equations that were used in our experiments are provided in table 3.2.

3.7 Chapter Summary

The prominent contributions of the work presented in this chapter are the innovations in the computation schemes of feature maps for color contrast and symmetry, proposal for determining size contrast as a formal feature map, and inclusion of eccentricity map together with orientation map as feature channels in the process of visual attention. Hence, efficient algorithms for construction of five feature maps are proposed that are able to be integrated into a region-based model of visual attention and, in turn, into other intelligent vision systems. The color contrast map is generated based upon the extended findings from the color theory, the symmetry map is constructed using a novel scanning based method, and a new algorithm is proposed to compute a size contrast map as a formal feature channel. Eccentricity and orientation are computed using the moments of obtained regions and then saliency is evaluated keeping rarity criteria into consideration. Evolutionary steps of the feature extraction and saliency detection algorithms presented in this chapter can be seen in [ASM05b], [AMSS06], [AM07b], [AM07c], and [AM08a].

The efficient design of the proposed algorithms allows incorporating five feature channels while maintaining a processing rate of multiple frames per second. A salient advantage over the existing techniques is the reusability of the salient regions in the high level machine vision procedures due to preservation of their shapes and precise information about locations. Results of implementation of the methods presented in this chapter are given in chapter 5. The values that were used in our experiments for the constants introduced in the proposed methodology are listed in table 3.2.

Table 3.2: Values of constants used in experiments

S.No	Constant	Value
1	k^p_{scale}	0.6
2	k^p_{min}	0.005 (0.5%)
3	k^p_{max}	0.79 (79%)
4	κ_{min}	0.21
5	κ	0.79
6	k^a_1	0.5
7	k^a_2	2.0
8	S^{max}	255
9	k_ϕ	25
10	k_e	0.21
11	k_s	0.17
12	β_o	0.75
13	β_s	0.7
14	β^1_e	0.7
15	β^2_e	0.9
16	r^{fovea}	51
17	k	255
18	D^Θ	0.91

4 Proposed Region-Based Attention Model

This chapter presents the proposed attention model in perspective of its overall architecture and high level procedures. The proposed model is designed to be behavior adaptive such that the same architecture could operate diversely under different visual behaviors. This allows to incorporate integration of bottom-up and top-down pathways of attention into a single architecture. The structural design of the proposed model is described in section 4.1. The low level image processing procedures including feature computation and feature saliency detection are already described in chapter 3 hence the discussion proceeds here with the high level procedures involved for attention processing. As most of these procedures are influenced by the visual behaviors of attention, hence a description of the behaviors implemented so far in the system is provided in section 4.2 before explaining the model components. Description of behavior dependant processing for feature map fusion for bottom-up and top-down pathways is provided in sections 4.3 and 4.4 respectively followed by the procedure for popout detection in section 4.5. As we apply a mechanism based upon saccadic memory for inhibition of return (IOR) therefore the working of the said memory is explained in section 4.6 before presenting the IOR module in section 4.7.

4.1 Model Architecture

The proposed model separates the steps of feature magnitude computation and saliency evaluation as shown in figure 4.1. The primary feature extraction function F produces a set of regions $\Re$ as explained in section 3.1 of chapter 3. Computation of the bottom-up saliency using the rarity criteria is performed by the group of processes S, which were discussed in detail in sections 3.4 to 3.5 in chapter 3. Output of S is combined by the procedure W that applies weighted

fusion of these maps according to the active visual behavior to formulate a resultant bottom-up map. Details of this procedure are given in section 4.3. The function G considers the given top-down conditions to produce fine grain saliency maps as already explained in chapter 3 section 3.6. Behavior dependant combination of these maps is performed by the function C that results into a resultant top-down map as explained ahead in section 4.4.

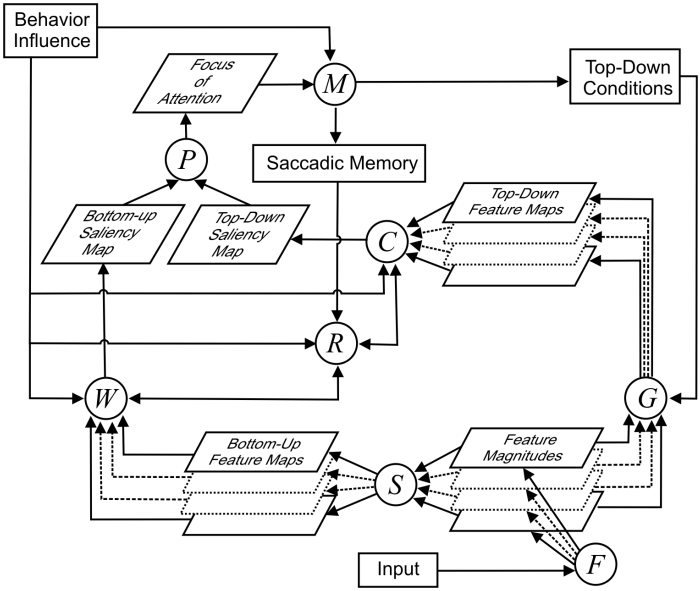


Figure 4.1: Architecture of the proposed region-based attention model.

The function P combines the resultant saliency maps into a master conspicuity map and applies a peak selection mechanism to choose one pop-out at a time. Details of this process are coming in section 4.5. The focus of attention at a particular time t is stored in the inhibition memory using which the process of IOR, denoted as R in the architecture diagram, suppresses the already attended location(s) at time $t + 1$ in order to avoid frequent revisiting of the same location(s). The operational visual behavior strongly affects this process hence the proposed design of this module explicitly incorporates this aspect (see section 4.7

for details). The memory management function M decides whether to place the recent focus of attention in inhibition segment or excitation segment of the saccadic memory according to the active behavior. Functioning of this component is described in section 4.6. One of the main focuses of this work is to establish the capability of demonstrating various visual behaviors in the attention system such that the same model may operate adaptively under different visual behaviors without requiring to make any architectural changes therefore the behavior influence is embedded in the major four internal functions for bottom-up map combination (W), top-down map summation (C), inhibition of return (R), and memory management (M).

4.2 Visual Behaviors

Locations and order of eye fixations are largely dependant on the active visual behavior or the task given to the vision system. This concept was introduced in chapter 1 section 1.4.5 with the support of psychophysical experiments reported by [Yar67]. We implement the influence of active visual behavior using a set of behavior-dependant weights for each individual behavior in which every set member is associated to a specific feature channel $f \in \Phi$ where $\Phi = \{c, o, e, s, z\}$ is the set of visual features considered in the model introduced in chapter 3 section 3.1.6. We may denote these sets of weights as $W^b_{td}(f)$ and $W^b_{bu}(f)$ where b represents the behavior while td and bu indicate the relation of the weights to the top-down and bottom-up pathways respectively. Having *explore*, *search*, *examine*, and *track* as the so far considered behaviors for our model, b can be a member of the set of behaviors Ξ where $\Xi = \{e, s, x, t\}$. Functionality of these behaviors in natural vision has been discussed in section 1.4.5 of chapter 1.

Until sufficient knowledge about details on the context of behavior-based processing in the natural attention becomes available, we use a set of quantized values for populating behavior-dependant weights $W^b_{td}(f)$ and $W^b_{bu}(f)$. We use four categories, namely *inactive*, *low*, *medium*, and *high* for this purpose. For an inactive channel the weight has to be set such that the concerned channel becomes totally non-contributing hence the *inactive* has a value of zero. We assign *low*= 1 meaning that no extra emphasis will be given on the related channel

when computing the resultant saliency. The value *medium* is taken equal to the sum of all *low* weights at that particular time in order to make influence of the involved channel higher than the others. If some channel has to be given higher weight in presence of a channel with a *medium* importance then the category *high* will be used in which the weight will be again sum of all weights counting the involved channel as a *medium* one. Hence, for our model with five channels in hand, we can have *medium*= 5 and *high*= 13 (when we have one other channel with *medium* weight while the rest having *low*). At time of initialization all weights are set to *low* and then updates are made only in the set of weights concerned with the active behavior as described in the following subsections.

4.2.1 Search Behavior

For the visual behavior of *search* the top-down channel plays the main role hence the manipulations of behavior dependant weights will be focused on $W^s_{td}(f)$. The bottom-up feature maps have trivial importance in the *search* task hence they are blocked by setting $W^s_{bu}(f) = \mathit{inactive}\ \forall f \in \Phi$. In this top-down pathway the color channel is given more emphasis over the other four channels because color is the most stable feature while searching an object whereas other features like size, eccentricity, symmetry, and orientation can significantly vary when a vision system goes around in the environment or the objects themselves rotate and translate in three dimensions. Therefore color will have a *medium* weight while other channels will be given *low* weights under *search* behavior. Hence

$$W^s_{td}(f) = \begin{cases} \mathit{medium} & \text{for } f = c \\ \mathit{low} & \text{otherwise} \end{cases} \tag{4.1}$$

4.2.2 Examine and Track Behaviors

For an *examine* or *track* task, the subsequent fixation is supposed to be on the nearest and very similar occurrence of the previously attended object (or, when tracking, the target is the object itself that has gone through a minor translation during the time between successive frames). Hence the top-down size channel is expected to be the most stable after color. In rest of the top-down channels

there are chances to obtain fluctuations because of the alterations in shades and shadows due to change in object location can vary the eccentricity, symmetry, and orientation of the region(s) associated with the object. Therefore for the *examine* and *track* behaviors we keep *high* weight for the color channel while *medium* for size in the top-down pathway. Other channels remain at *low* in these two behaviors. The bottom-up channels are kept active but without any discrimination of any channel, hence $W^x_{bu}(f) = low$ and $W^t_{bu}(f) = low\ \forall f \in \Phi$. The top-down weights for these behaviors may be summarized as

$$\text{for } b = x \text{ and } b = t,\ W^b_{td}(f) = \begin{cases} high & \text{for } f = c \\ medium & \text{for } f = s \\ low & \text{otherwise} \end{cases} \tag{4.2}$$

4.2.3 Explore Behavior

When the system works under *explore*, the bottom-up channels become the major players and all top-down channels are hindered from participating in further steps, hence $W^e_{td}(f) = inactive\ \forall f \in \Phi$. Under this behavior the bottom-up feature channels have to be assigned optimal weights that facilitate automatic pop-out of the visually salient object in the given scene. The mechanism of assigning the optimal bottom-up weights to $W^e_{bu}(f)$ is explained below.

Earlier to the process of finding the weights for the involved feature channels, weight of each feature channel f, $W^e_{bu}(f)$, is first initialized (at $t = 0$) such that the color map gets the highest weight because it plays a major role in attention and the size map gets the lowest weight because it is effective only when other channels do not contain significant bottom-up saliency. The orientation channel is reported as one of the confirmed feature channels and the other shape-based features like eccentricity and symmetry are listed as probable channels (a level below the confirmed channels) in human vision [WH04], hence we assign them medium weights. Thus at time of initialization:

$$W^e_{bu}(f) = \begin{cases} high & \text{for } f = c \\ medium & \text{for } f = s, e, o \\ low & \text{for } f = s \end{cases} \tag{4.3}$$

Later, the weights are adjusted such that the feature map offering the sharpest peak of saliency contributes more in the accumulated saliency map. It is done by finding the distance Δ_f between the maximum and the average saliency value in each map. The feature map with the highest Δ_f is considered as most active and its weight is increased by a multiplicative factor δ (we take $\delta = 2$).

4.3 Bottom-Up Map Fusion

The function W of the model, sketched in figure 4.1, performs a combination of the bottom-up feature maps under the influence of the active visual behavior. Figure 4.2 presents its working in an architectural perspective. This function takes the raw feature maps for the very first frame of visual input while for the rest of the frames (or successive attention attempts) it uses the maps that have gone through the inhibition process R (explained in section 4.7). The resultant bottom-up saliency map is obtained by summation of all feature maps of this pathway after applying the weights related to the active behavior of attention obtained from the procedure explained in the previous section for this purpose.

Computation of the total bottom-up saliency $\beta_i(t)$ of a region R_i at time t can now be modelled as:

$$\beta_i(t) = \frac{\sum_{f \in \Phi} \left(W^a_{bu}(f) S^i_f(t) \right)}{\sum_{f \in \Phi} W^a_{bu}(f)} \tag{4.4}$$

where $W^a_{bu}(f)$ are weights of the feature maps for the active visual behavior. $S^i_f(t)$ represents the bottom-up saliency at time t in feature channel f for a region R_i.

4.4 Top-Down Map Fusion

The top-down feature maps are combined by the function C of the model (see figure 4.1) in accordance with the influence of the active visual behavior. Working of this module is presented pictorially in figure 4.3. Similar to the bottom-up

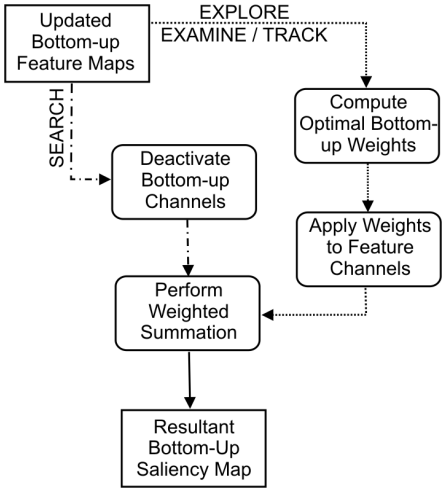


Figure 4.2: The module W for behavior dependant combination of bottom-up feature maps.

function W described above, this function also takes the raw feature maps for the first frame of visual input and then the maps processed by the inhibition function R for the rest of the input.

The resultant top-down saliency map is constructed through summation after applying the weights related to the top-down pathway. The resultant top-down saliency $\gamma_i(t)$ of a region R_i at time t is computed as

$$\gamma_i(t) = \frac{\sum\limits_{f \in \Phi} \left(W^a_{td}(f) \gamma^f_i(t) \right)}{\sum\limits_{f \in \Phi} \left(W^a_{td}(f) \right)} \tag{4.5}$$

where $W^a_{td}(f)$ are weights of the active top-down visual behavior obtained through the process discussed in section 4.2.1 and $\gamma^f_i(t)$ are the top-down feature maps at a given time t.

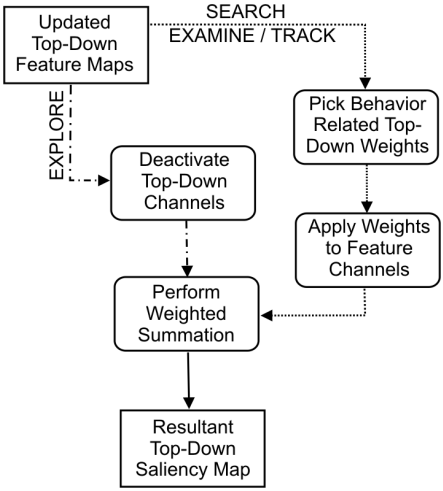


Figure 4.3: The module C for behavior dependant combination of top-down feature maps.

4.5 Pop-out Selection

The function P in the proposed model combines the bottom-up and top-down saliency maps $\beta_i(t)$ and $\gamma_i(t)$ for the current frame to produce the final conspicuity map and selects one region on which attention should be focused for this moment of time. As the influence of behavior is already applied during the steps that sum up the feature maps therefore here a simple combination is required. Each region possesses a quantity for its saliency in terms of bottom-up as well as top-down aspect to compete for attention in the final step. We propose to model the step of finding the combination of saliency in bottom-up and top-down channels for given region R_i as a function that picks the maximum saliency value from the two pathways as its final conspicuity value. Therefore the final conspicuity value $\alpha_i(t)$ of a region R_i at a given instance t will be computed as

$$\alpha_i(t) = max(\beta_i(t), \gamma_i(t)) \tag{4.6}$$

Having processed all the important steps at the lower stages of the model, the popout selection at time t remains as simply picking an R_i that possesses the maximum amount of $\alpha_i(t)$. An important component of attention that contributes in computation of the final saliency of regions, namely inhibition of return, is yet to be discussed in the next sections but we have elaborated pop-out selection after the methods for saliency map fusion to establish a sequence of presentation. Pop-out for the first cycle of attention will be selected before the influence of IOR but for the later cycles this step will receive input processed by the IOR routine.

4.6 The Saccadic Memory

In order to deal with dynamic visual input the process of inhibition of return is applied using a saccadic memory consisting of two main segments, namely inhibition and excitation. The inhibition memory M^I is designed to remember the locations and features of the last m foci of attention because a series of recently attended locations have to be inhibited. Hence:

$$M^I = \{M_k^I\},\ k \in \{1, 2, ..m\} \tag{4.7}$$

where k denotes the age of the memory item M_k^I and m represents the maximum number of items that this memory can store. According to [CCSP03] the average period for which a location remains inhibited is about 1500 ms (ranging from 50 to 3000 ms). Keeping in view the average processing rate of our model equal to 10 frames per second (a single frame is processed in 100 ms), which is similar to the frame rate of human eye according to the, now abandoned, classical theory of persistence in vision, the number of items m that should be remembered in the inhibition memory turns out to be approximately 15. Hence, we keep $m = 15$. For the most recent item M_k^I stored in M^I, the value of k is set to 1. k increases with the age of M_k^I in the memory. When a fresh item arrives, it replaces the stored item with $k = m$ and its k is reset to 1. All other items stored in the memory get an increment in their values of k, i.e. their age.

Under certain visual behaviors a process of facilitation of return (FOR) is also performed in which bias is given to certain features and/or locations while focusing attention in the successive frame of visual input [OMY05] [CC06]. In order to handle the behavior dependant facilitation of return (FOR), we use the excitation memory M^E designed to remember the location and features of the last attended region. This location and features will be preferred while determining the next FOA under *examine* and *track* behaviors. The size of this memory is set to a single item because it possible to excite only one set of features and/or location while deciding a focus of attention.

Figure 4.4 shows the working of the saccadic memory unit under different visual behaviors. Under every type of behavior the last few attended locations have to be inhibited hence they are stored in the inhibition memory in all cases. In case of *explore*, we are implementing the feature based inhibition also (see section 4.7 for details) hence features of the recently attended regions are also stored under this behavior. While working under *examine* and *track* behaviors, facilitation of return is needed on the features and location of the recently attended object hence these informations are saved in relevant portion of the saccadic memory. The *search* behavior requires only inhibition of the recently attended locations hence memory management for this process is simpler than others.

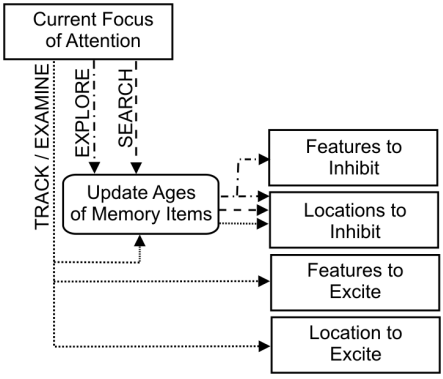


Figure 4.4: The module M for behavior dependant management of storing recently attended items.

4.7 Inhibition and Facilitation of Return

After having attended a region at time t, the mechanism of inhibition of return (IOR) prevents the system from keeping continuous attention on the same point in order to allow exploration of other relatively less salient locations in the scene also. For this purpose the focus of attention at time t is inhibited while finding pop-out at time $t+1$. In the proposed model, we consider three types of inhibition mechanisms namely spatial, feature based, and feature-map based. Most of the existing models of attention implement either the spatial inhibition in which a specific area around the point of attention is inhibited or the feature-map based inhibition in which the weight of the wanted feature channel is adjusted to obtain required results. In some recent studies in psychophysics the feature based inhibition has also been reported [WLW98]. For example, according to [LPA95], the color of the focus of attention is ignored in the successive attempts of attention. The proposed approach models this feature-based inhibition also.

Figure 4.5 draws the mechanism of the inhibition and facilitation (or excitation) processes in the proposed model under different active visual behaviors. The inhibition process is common in all behaviors but the excitation process denotes the facilitation of return under certain visual behaviors (*examine* and *track* from the currently implemented behaviors).

For applying the IOR and FOR, the locations and features of the salient regions attended in the last few saccades are stored in the saccadic memory and salient locations from the current frame are compared with them. Under the *explore* behavior the current salient regions are inhibited in two ways. Firstly, when their locations match with one of the locations stored in the inhibition segment of the saccadic memory M^I and, secondly, if they have features similar to one of the stored regions M^I_k. Stronger inhibition is applied if the current region matches a recently attended region (a memory item with young age) and weaker inhibition takes place in case of a match with an older item in memory. In case of *search* behavior only location based inhibition occurs using the saccadic memory of many previous FOAs. The top-down conditions to excite the searched features related to the search target come from a source external to the attention mechanism and are used in fine-grain feature map construction rather than in

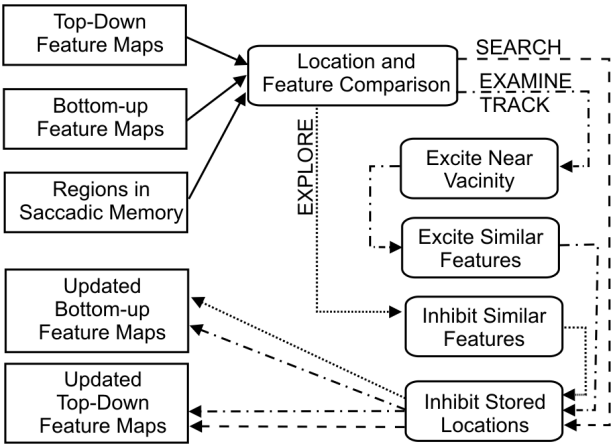


Figure 4.5: The module R for behavior dependant inhibition and facilitation of return.

applying facilitation of return, hence features of the previously attended objects do not play any part in *search* behavior. While having the *track* or *examine* behavior active, the module R applies inhibition as well as facilitation of return. For examining a bigger pattern made up of small individual patterns features of the last attended region are to be excited in the next coming frame and the locations nearby the last FOA also need to be excited in order to maintain good sequence of fixations. The inhibition process applies spatial inhibition as done in other behaviors. Under *track* behavior the inhibition/facilitation process will be identical to *examine* because occurrence of a moving object in the successive frame will be close to its position in the last frame and the features will of course be the same. As both bottom-up and top-down channels work together in the last two behaviors, maps of the both pathways are updated by the module R whereas in case of the first two behaviors maps of only the concerned pathway is updated. We describe details of the internal functionality of different modes of the module R in the following subsections.

4.7.1 Feature-based and Spatial Inhibition

In the functions W and C, shown in the architecture diagram in figure 4.1, the inhibition process influences the raw (un-inhibited) saliency maps obtained from the current frame (let us call this instant t') using the foci of attention stored in the saccadic memory up to the time $t-1$ before production of the current conspicuity maps at time t. In other words, the bottom-up feature maps $\beta_i^f(t')$ and the top-down feature maps $\gamma_i^f(t')$ are updated by R to obtain $\beta_i^f(t)$ and $\gamma_i^f(t)$, in W and C respectively, before summing them into master conspicuity maps for bottom-up and topdown pathways $\beta_i(t)$ and $\gamma_i(t)$ respectively. Hence at time t, the updated saliency of a region in bottom-up and top-down context can be written as

$$\beta_i^f(t) = \beta_i^f(t')\Im^s(R_i, M_k^I, k)\Im^f(R_i, M_k^I, k) \;\forall\; k \in \{1, 2, ..m\} \tag{4.8}$$

$$\gamma_i^f(t) = \gamma_i^f(t')\Im^s(R_i, M_k^I, k) \;\forall\; k \in \{1, 2, ..m\} \tag{4.9}$$

The inhibition function $\Im^s(R_i, M_k^I, k)$ performs the spatial inhibition whereas feature-based inhibition is applied by $\Im^f(R_i, M_k^I, k)$ around all m locations stored in M^I. These two functions are designed in such a way that a decreasing suppression will be applied with the increasing value of k, i. e. less suppression when the age of the memory item becomes higher. The spatial inhibition factor $\Im^s(R_i, M_k^I, k)$ is modeled as:

$$\Im^s(R_i, M_k^I, k) = \frac{\delta^s k D^s\left(L^c(R_i), L^c(M_k^I)\right)}{m r^{inh}} \tag{4.10}$$

where δ^s is the spatial inhibition factor such that $0 < \delta^s < 1$, $L^c(.)$ extracts the midpoint of provided region, $D^s(L^c(R_i), L^c(M_k^I))$ is the spatial distance between centers of the considered region R_i and the region in the memory location M_k^I, and r^{inh} is the radius within which inhibition takes effect. The value of $D^s(L^c(R_i), L^c(M_k^I))$ is clipped to r^{inh} when it rises above r^{inh} in order to maintain $\Im^s(R_i, M_k^I, k) \leq 1$. $\Im^s(R_i, M_k^I, k)$ has the lowest magnitude when a region is at a small distance from the previous focus of attention and it gradually grows to 1 while approaching the radius r^{inh}. In other words, the decay is strongest near the center and it weakens to no decay as the boundary of inhibition circle

is approached. The spatial distance $D^s(P_1, P_2)$ is calculated using a simplified approximation in order to keep it computationally inexpensive. Hence for any two points $P_1(x_1, y_1)$ and $P_2(y_1, y_2)$

$$D^s(P_1, P_2) = |x_1 - x_2| + |y_1 - y_2| \tag{4.11}$$

Inhibition is performed on world coordinates in order to tackle attention in dynamic scenarios, therefore a location in space is inhibited rather than a location in the view frame. This involves head angles of the vision system along with the position within the view frame, in which the top corner of the view frame is considered as origin. The x-coordinates grow towards right whereas y-coordinates grow downward following the convention of computer graphics systems. Hence, taking P_1 as center of the considered region R_i and P_2 as the inhibited location stored in memory location M_k^I, the coordinates x_1, x_2, y_1, and y_2 for equation (4.11) are computed as follows:

$$x_1 = \delta_x^\theta \Theta_x^v(R_i) + L_x^c(R_i) \tag{4.12}$$

$$x_2 = \delta_x^\theta \Theta_x^v(M_k^I) + L_x^c(M_k^I) \tag{4.13}$$

$$y_1 = \delta_y^\theta \Theta_y^v(R_i) + L_y^c(R_i) \tag{4.14}$$

$$y_2 = \delta_y^\theta \Theta_y^v(M_k^I) + L_y^c(M_k^I) \tag{4.15}$$

where δ_x^θ is the number of pixels in the view frame that are covered after rotating the camera head through one degree in horizontal direction, $\Theta_x^v(R_i)$ is the head angle of the vision system while looking at the region R_i and $L_x^c(R_i)$ gives the x-coordinate of the center of R_i inside the view frame. Figure 4.6 shows a situation where the horizontal world coordinates for a region are computed using the horizontal camera angle of the robot and the x-coordinate of the attended region in the view frame. $\Theta_x^v(M_k^I)$ and $L_x^c(M_k^I)$, respectively, are the horizontal head angle and the x-coordinate of the inhibited region stored at k^{th} location in the inhibition memory. δ_y^θ, $\Theta_y^v(R_i)$, $L_y^c(R_i)$, $\Theta_y^v(M_k^I)$, and $L_y^c(M_k^I)$ are similar quantities for the vertical direction. This arrangement can yield sufficient accuracy either for stationary vision systems able to rotate their camera head or indoor robots that move for short distances with a moderate speed. This simple method is not expected to be accurate for systems that undergo complex and fast move-

ments such as sharp turns or driving on uneven roads. Such situations demand an integrated localization and mapping system able to tackle object positions in three dimensional space.

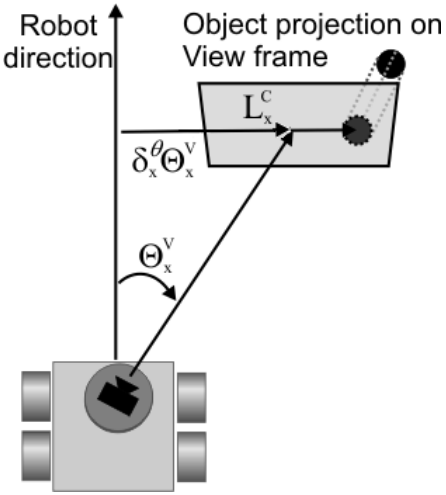


Figure 4.6: Computation of horizontal world coordinates with respect to the robot for a region using the horizontal camera angle of the robot and the x-coordinate of the attended region in the view frame.

The second inhibition factor $\Im^f(R_i, M_k^I, k)$ in the equation (4.8) inhibits in context of feature similarity with respect to each feature channel f. Regions having feature similarity with the inhibited regions get a suppression in saliency while the system looks for next interesting locations. This inhibition step is modeled as:

$$\Im^f(R_i, M_k^I, k) = \begin{cases} \frac{\delta^f k}{m} & \text{for } D^f(R_i, M_k^I) < \tau^f \\ 1 & \text{otherwise} \end{cases} \tag{4.16}$$

where δ^f is the inhibition constant such that $0 \leq \delta^f \leq 1$ and $D^f(R_i, M_k^I)$ determines the difference between feature values of the given region R_i and the inhibited one stored in M_k^I. Two regions should have a feature difference below the threshold τ^f in order to be considered similar.

It may be noted that the inhibition of return has to decay with time in order to allow attention to the same or similar objects after a certain period of time. This decay is automatically managed in the proposed mechanism by use of the age in memory. The strenght of inhibition decays as the age of inhibited item increases in the memory leading to no decay when the item gets old enough to be replaced by a new entry.

4.7.2 Feature Map Based Inhibition

The feature-map based inhibition is modeled for preventing a feature map from gaining extraordinary weight so that other features do not get excluded from the competition. When a weight $W_{td}^b(f)$ becomes equal to $max(W_{td}^b(f) \forall f \in \Phi)$ then it is set back to its original value that was assigned to it during the initialization step. This mechanism keeps the weights of feature maps in a cycle because the map weights keep rising when the concerned feature map contains a sharp peak until this peak gets attended or gets inhibited due to attention to some neighboring region. The system iterates back to the feature fusion step mentioned in equation 4.4 after applying these inhibition procedures to reach the next focus of attention.

4.7.3 Facilitation of Return

Here we describe the process for excitation or facilitation of return. Excitation has to be performed in context of location as well as in terms of features. Functionality of facilitation of return will be similar to the inhibition of return except that excitation will be applied on saliency instead of suppression using the last FOA stored in the excitation segment M^E of the saccadic memory. As in the current status of the attention model such excitation is included only in the module C, which is responsible for the top-down pathway, this process will affect only the top-down feature maps $\gamma_i^f(t)$ as follows:

$$\gamma_i^f(t) = \gamma_i^f(t')\Omega^s(R_i, M^E)\Omega^f(R_i, M^E) \tag{4.17}$$

$\Omega^s(R_i, M^E)$ and $\Omega^f(R_i, M^E)$ are functions for spatial and feature similarity based excitation respectively. The previous FOA stored in the excitation segment M^E of the saccadic memory is used for comparison with the considered region R_i. These two functions are defined as

$$\Omega^s(R_i, M^E) = \frac{\delta^e D^s\left(L^c(R_i), L^c(M^E)\right)}{r^{inh}} \tag{4.18}$$

$$\Omega^f(R_i, M^E) = \begin{cases} \delta^e & \text{for } D^f(R_i, M^E) < \tau^f \\ 1 & \text{otherwise} \end{cases} \tag{4.19}$$

where δ^e is the excitation factor such that $\delta^e > 1$, $D^s\left(L^c(R_i), L^c(M_E)\right)$ computes the distance between center of the considered region $L^c(R_i)$ and center of the region stored in the excitation memory $L^c(M^E)$, and $D^f(R_i, M^E)$ computes the difference between R_i and M^E in terms of the considered feature f. Due to this design of these two functions, a region having feature similarities with the previous FOA and closest distance from it will get the highest amount of excitation.

4.8 Chapter Summary

This chapter has presented the core of the research work being discussed in this dissertation. The overall architecture of the proposed attention model has been sketched with description of its constituting building blocks. The main advancement as compared to the existing attention models is the integration of bottom-up and top-down pathways into a single architecture and inclusion of explicit influence of visual behaviors in internal steps of the model. The behavior dependant functionality of different modules have been explained in detail. Further advancements in the state-of-the-art through this model is inclusion of facilitation of return (FOR) besides the normally implemented inhibition of return (IOR) and consideration of attention in three dimensional space apart from the commonly used two dimensional image planes. Intermediate milestones of the steps presented in this chapter can be seen in [AM07d], [AM07c], [AM07e], [AM07b], [AM07a], and [AM08c].

5 Experiments and Results

The proposed attention model is implemented as a complete software using object oriented programming techniques with C++. This chapter presents the output generated by the model on various test cases selected to judge its performance. The experiments with the model were carried out for three attentional behaviors, namely *explore*, *search*, and *examine*, for which details are described in sections 5.2, 5.3 and 5.4 respectively. Being conceptually similar, results on *examine* behavior also cover the testing of *track*. Results were obtained using three different experimentation platforms, described in section 5.1, for testing the model's performance on static snapshot images, controlled virtual environments in a simulation framework, and real-life scenarios using a camera head mounted on a mobile robot with pan and tilt capability.

5.1 Experimentation Platforms

Experiments were performed to test the capabilities of the proposed model for different visual behaviors using three experimentation platforms. The first platform is an evaluation framework for single images. Images can be loaded as input for the model and a behavior can be activated through the options provided in a graphical user interface. Using this software the saliency maps can be seen at each attempt of attention and the fixated regions get marked by prominently colored rectangles. The sequence numbers of fixations are also displayed within the rectangles to keep record of the previously selected regions for the given input. Figure 5.1 shows a screenshot of the graphical user interface of this system in which saliency maps and fixated locations for a sample image are also visible. Computation time taken to process the given input is also displayed in a prescribed part of the window.

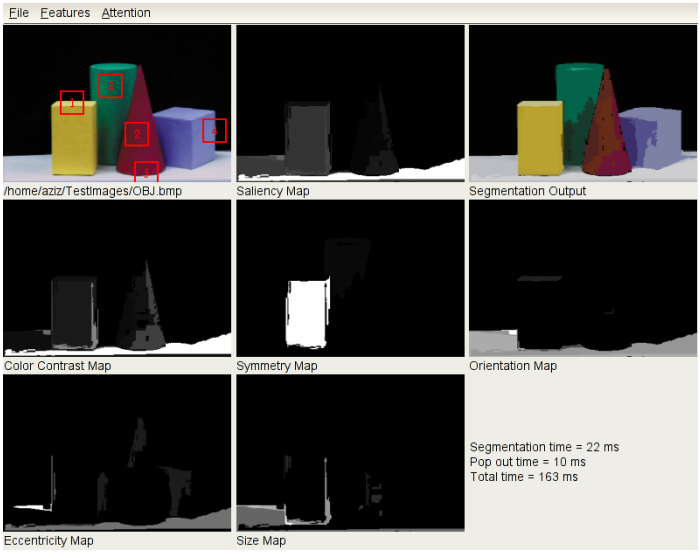


Figure 5.1: Graphical user interface of the implementation of the proposed attention model. Processing on a sample input is shown with the intermediate results.

The second platform is a robot simulation framework, SIMORE, developed in our group [KHSM08]. It has the capability of assimilating 3D models of environments, simulated robots, individual sensors and actors created with virtual reality modeling software into an integrated test scenario [Sim09]. The robots can be manipulated at the level of their individual components, such as rotatable wheels and sensor head, using different means including control by manual input devices and a graphical user interface. The interface functions of SIMORE can be called from another application (such as the attention model) in order to manuever the modeled robot autonomously in the virtual reality scene based upon the data retrieved from the simulated sensors.

In addition to the 3D graphics engine the simulator has a physics engine to guarantee a correct physical behavior of the simulated objects. The input from the sensors, for example continuous image stream from the simulated camera head for the requirement of the work under discussion, is provided to the control-

ling program as it would be done by a real sensor wandering inside the given environment [Hil08]. This platform helps in testing the algorithms in a three dimensional world with the ability of maneuvering the sensor head as well as the whole robot to experiment with active vision. The test scenes can be created with scalable complexity and they are utterly reproducible as illumination conditions remain stable and the arrangement of objects remains intact for an arbitrarily long period of time. Moreover, experiments can be conducted uninterruptedly without disturbances from hardware failures and emptying of batteries. Hence the core functionality of the algorithms can be verified and validated through this system. Figure 5.2 presents a sample virtual environment with a simulated robot maneuvering inside it. The visual input seen through the cameras are also shown.

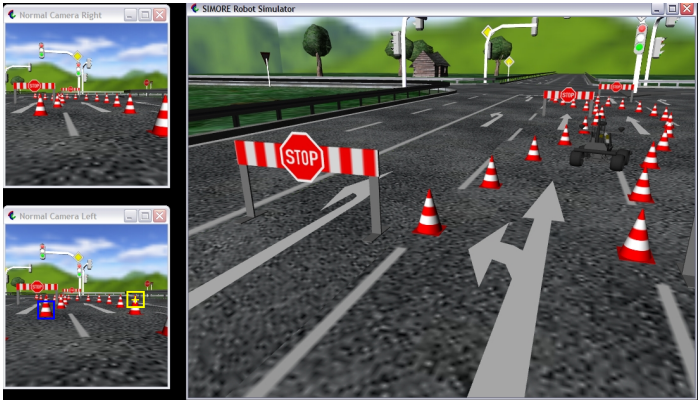


Figure 5.2: User interface of the simulation framework SIMORE. Global view with the robot controlled through the attention model is shown in the right window. The views through the simulated robot's left and right cameras are visible in the smaller windows at left side. In the left camera view the current focus of attention is marked by a yellow rectangle while a previously attended (now inhibited) region is marked by a blue one.

The third experimentation platform consists of actual mobile robots available in our laboratory. One of them is a teleoperated rover robot (TSR) built in

our laboratory [BDST04]. The TSR is equipped with a stereo camera head and a wide range of sensors like electronic compass, GPS device, infrared sensor, and motion detector. It can be remotely controlled by input devices such as joystick, force-feedback steering wheel, and head mounted display with tracking device for head motion, or through computer programs running either on the on-board computer or on a remote machine connected to the robot via wireless network. Another rover robot is based upon the commercial Pioneer 3AT system. Sensors for visual input, laser range detection, ultra sound range detection, heat detection, and inertial measurement unit (IMU) have been installed on it to enable its autonomous wandering. Yet another system is a commercial flying robot (quadrocopter) which can be controlled through wireless network and can be used to obtain visual data from top of scenes. Work is underway to extend the capabilities of this platform and install new sensors on it. Figure 5.3 shows pictures of these three robotic systems.

(a) (b) (c)

Figure 5.3: Robotic platforms currently available in our laboratory for experimentation (a) The teleoperated robot system TSR with its stereo camera head shown at top corner of the image (b) Pioneer based system GETBOT (c) Flying robot (quadrocopter).

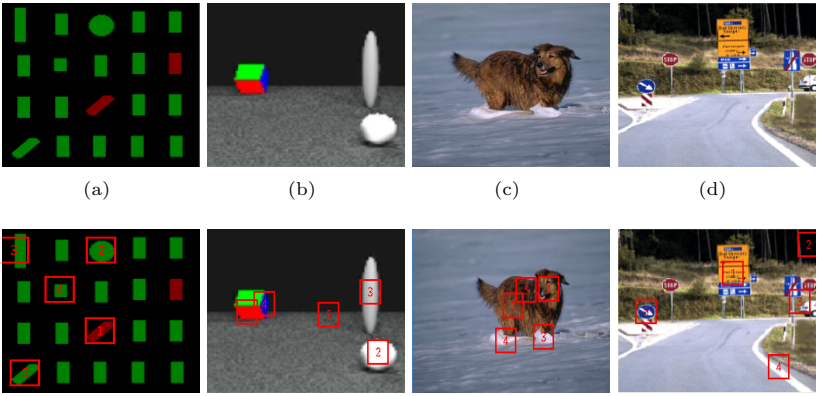
5.2 Exploration Results

Under the *explore* visual behavior the vision system performs free viewing with no top-down task given to it. The requirement is to identify those locations in the given scene that would be salient for the natural/human vision. The

developed attention model was tested with static images to verify its ability to locate salient regions and then its capability to work in dynamic environments was tested using the robot simulation framework SIMORE. The experiments on overt attention were also carried out using the camera head of the actual robot. We arrange these results in form of subsections with each subsection dedicated to results from an experimental platform.

5.2.1 Static Scene Exploration

A variety of visual scenes including snapshots taken from camera of robot head, synthetic environment of the simulation framework, and other images collected from image databases on internet were used to experiment with the proposed attention model. Figure 5.4 presents results of attention on images having conjunction of different features in synthetic and natural scenes: The image in subfigure (a) contains one object possessing saliency due to conjunction of multiple features while other objects have saliency due to contrast of only one feature. The synthetic scene in subfigure (b) offers rendered 3D objects having saliency with respect to different features with a very simple background. The picture in part (c) is a real life scene offering one foreground object (a dog) composed of different regions (head, body, tail, colliding water, etc) with complex combinations of feature saliencies. The input image in part (d) is a traffic scene in which the traffic signs offer a high bottom-up saliency in presence of a fairly distracting background. It may be observed in the output of the proposed model (subfigures (e) to (h)) that the fixated locations are mostly over regions where some visually salient object exists. In subfigure (f) the fixation in the middle of the image (boundary of the upper and lower background) may seem to be an error but this occurred because of the strong feature-based and spatial inhibition of return that motivated the system to explore new regions of interest even with lower bottom-up saliency. Similarly the fixation on the dark region between the trees in subfigure (h) may appear to be an error but that region gained saliency because of its strong contrast of brightness with its surrounding. A formal evaluation of the results to judge robustness and efficiency of the proposed approach is provided in the next chapter.



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

(e) (f) (g) (h)

Figure 5.4: (a) - (d): Input images containing regions with conjunction of different features. (e) to (h) are results of attention on the corresponding input with the first five fixations marked by the proposed attention model.

Figure 5.5 presents the results of step by step inhibition of return by showing the saliency maps during the first three selections in a given input image. Subfigure (a) shows the input sample, the original status of saliency at time t can be seen in its subfigure (b), and the first four fixated locations (time t to $t+3$) are shown in subfigure (c). Subfigure (d) shows inhibition in the spatial domain only after the first cycle of inhibition $(t+1)$ while subfigure (e) contains the saliency map after the first inhibition on regions possessing features similar to the last attended region. The combined effect of both inhibition functions in the first cycle is shown in subfigure (f). Subfigures (g) to (i) demonstrate the results after the second inhibition cycle $(t+2)$ and subfigures (j) to (l) contain the output after the third inhibition $(t+3)$. In each case the brightness of the region with the highest saliency is raised to white in order to indicate the next pop-out region. The brightness of the rest of the regions is also scaled up accordingly.

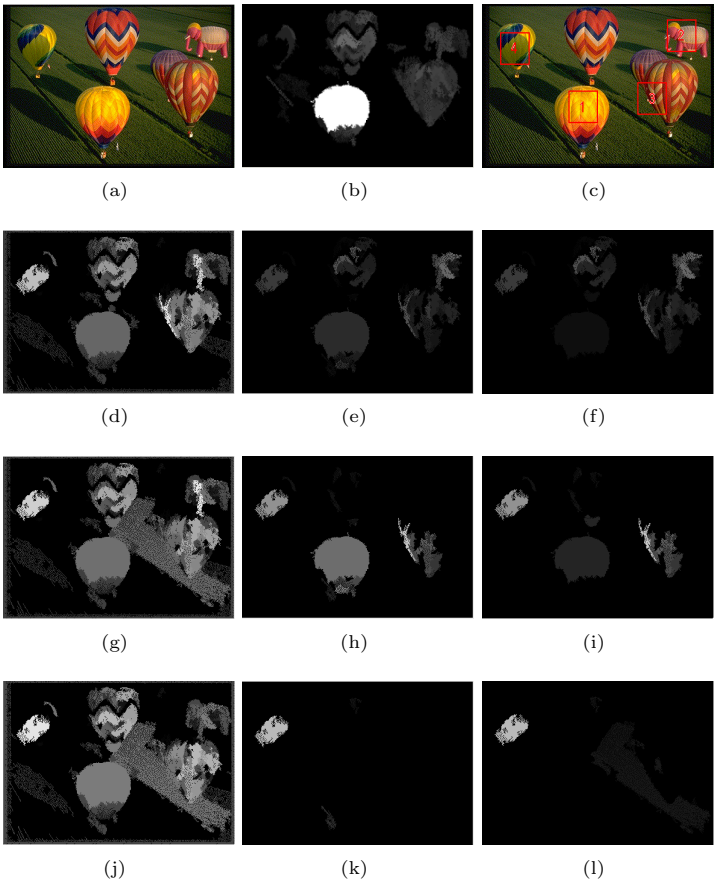


Figure 5.5: (a) Input image. (b) Saliency map at time t. (c) First four foci of attention representing time t to $t+3$. (d) Saliency map after only spatial inhibition (see equation 4.10) at time $t+1$. (e) Saliency map after only feature inhibition (see equation (4.16)) at time $t+1$. (f) Saliency map after combination of both inhibitions. (g) - (i) Spatially inhibited, feature-wise inhibited, and resultant saliency maps at time $t+2$. (j) - (l) Spatially inhibited, feature-wise inhibited, and resultant maps at time $t+3$.

5.2.2 Exploration in Simulated 3D Environments

Figure 5.6 demonstrates results of exploration performed by the proposed system in a dynamic scenario experimented in a virtual environment using the robot simulation framework SIMORE. Subfigure (a) shows the environment in which the simulated robot drives on the path marked by the red arrow while subfigure (b) shows the scene viewed through one camera of the stereo camera head of the robot. Subfigures (c) to (g) present the output of bottom-up attention for five selected frames each picked after equal intervals of time. The current focus of attention is marked by a yellow rectangle whereas blue ones mark the inhibited locations. It may be noted that the vision system was able to inhibit previously attended locations while being in motion in the 3D world.

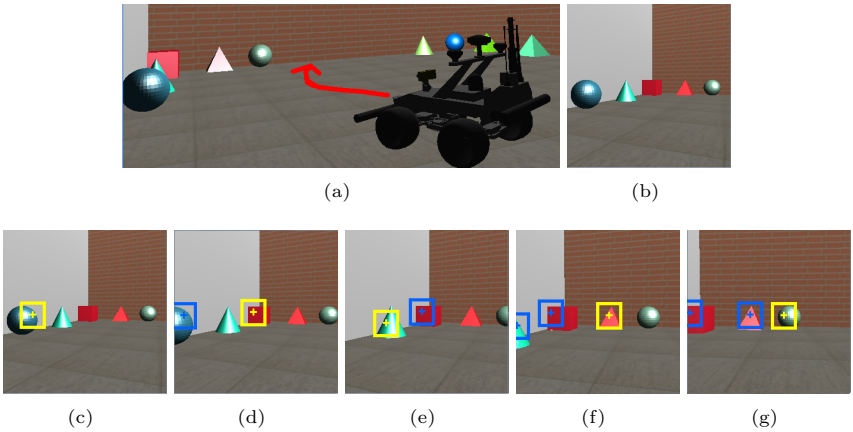
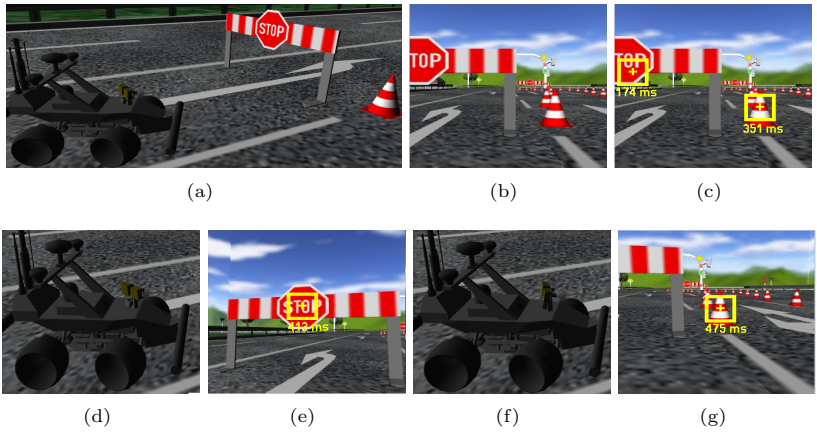


Figure 5.6: Results in dynamic scenario using a simulated mobile vision system. (a) Simulated robot moving in virtual environment. (b) Scene viewed through left camera of the robot. (c) to (g) Fixated locations are indicated by yellow marks and inhibited locations are indicated by blue marks while the robot moves along the path marked by the red arrow.

Figure 5.7 presents results of overt attention performed by the simulated robot in the environment shown in subfigure (a). The parts (b) and (c), respectively, show

the scene viewed through the camera head and the first two locations selected to be attended. Subfigures (d) and (e) demonstrate status of the camera head and the current view seen through it after automatically bringing the first FOA into center of frame. Similarly subfigures (f) and (g) demonstrate the situation for the second focus of attention.



(a) (b) (c)

(d) (e) (f) (g)

Figure 5.7: Results of overt attention performed by simulated robot. (a) The virtual environment including the simulated robot. (b) Scene viewed through the left camera of the robot. (c) Top two salient locations to be overtly attended. (d) Camera head rotated to bring the first FOA into center of view frame. (e) Scene viewed through the left camera after overt attention shift. (f) and (g) Status of camera and the camera view after overt attention to second FOA.

5.2.3 Exploration Using Robotic Camera

The developed model was integrated with the control system of the robot platform in order to perform overt attention using its camera head to identify visually salient items in the environment and bring them into center of view using the pan-tilt camera. In the exploration mode the system was required to rotate the camera in a scanning manner from one end (e. g. right) to the other (left) and

focus at items that offer high visual contrast. Figure 5.8 demonstrates a sample from the situations in which the robot identifies salient objects and brings them into the center of camera view one by one.

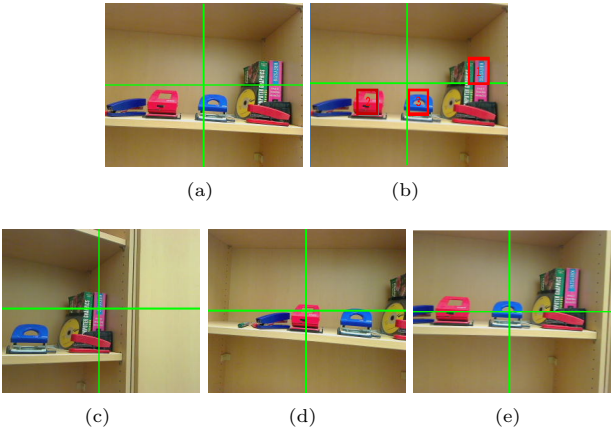


Figure 5.8: Results of overt attention performed by the robotic camera head under *explore* behavior. (a) Scene viewed through the camera of the robot with the cross hair indicating center of view frame. (b) Top three salient locations to be overtly attended. (c) Scene viewed through the camera after overt attention shift to first FOA. (d) and (e) Scenes viewed through the camera after overt attention shifts to second and third FOAs respectively.

5.3 Search Results

For experiments to test functioning of the proposed model under the visual behavior of *search*, description of the objects to be searched is given to the attention model in form of images containing the isolated targets over a blank background. Such descriptions of the target may be considered as the top-down conditions for the attention mechanism. In the current status, the system is able to work with single regions at a time rather than composite objects hence the system picks the largest foreground region from the given image of the search target. Similar

to the arrangement of the previous section, the results under this visual behavior are also presented in subsections each related to the used experimentation platform.

5.3.1 Search in Static Scenes

The first scenario of experiments was the search in static scenes in which the attention mechanism was allowed to mark as many occurrences of the target as possible. These experiments tested the ability of the system to select all relevant locations. Figure 5.9 reflects this scenario with the search field as a still scene having four occurrences of the target (a dull blue box with some texture) in the scene. Results of the first five fixations ($t = 1$ to $t = 5$) by the attention system are reported in figure 5.10. The current focus of attention is marked with a black rectangle while blue rectangles are drawn at the inhibited locations. It may be noted that the four target locations are marked in the first five fixations in which the extra fixation is due to a repeated saccade on an object instance that had such a high top-down saliency that, even after inhibition, it still remained higher than the fourth object, which has relatively less similarity with the target. This aspect can be noticed in the saliency maps provided in the second column of figure 5.10.

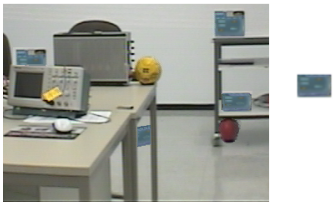


Figure 5.9: A sample from visual input used in experiments on visual search using top-down visual attention. Left image is the search field and the right one is the target to be searched.

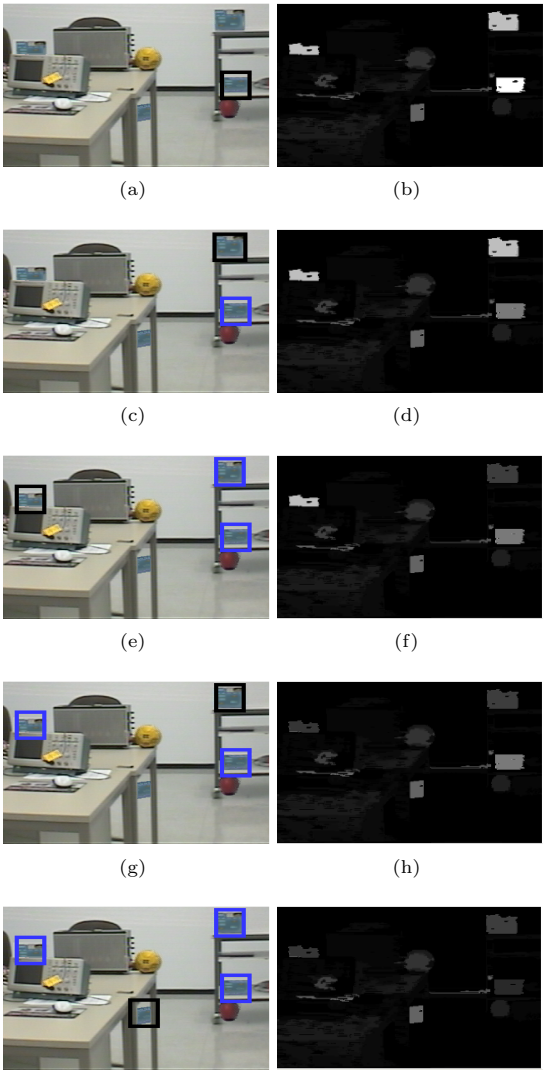


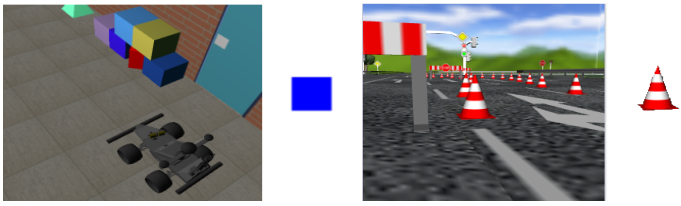
Figure 5.10: Results of covert attention under *search* behavior. Left column: Fixated locations marked by black rectangles and inhibited locations represented by blue rectangles in the scenario given in figure 5.9. Right column: Top-down saliency maps at time of each fixation.

5.3.2 Search in Dynamic Virtual Scenes

In the second scenario, a simulated vision system was set into motion that was required to mark the locations matching the search target using one fixation per frame. Figure 5.11(a) represents this scenario in the simulation framework SIMORE in which the target specified by the small image at the right side is to be searched in the 3D environment shown at left. On the other hand, subfigure (b) is a scenario in which the simulated robot performs overt attention on the searched locations by rotating its simulated camera head. Results of search without moving the camera head in the scenario given in figure 5.11(a) are provided in figure 5.12. After covertly fixating on the best matches, the system tries to pick target locations even when they have less similarity with the target, for example, the later fixations are done based only upon color similarity. Figure 5.13 presents results of overt attention performed on instances of the search target in the scenario shown in figure 5.11(b).

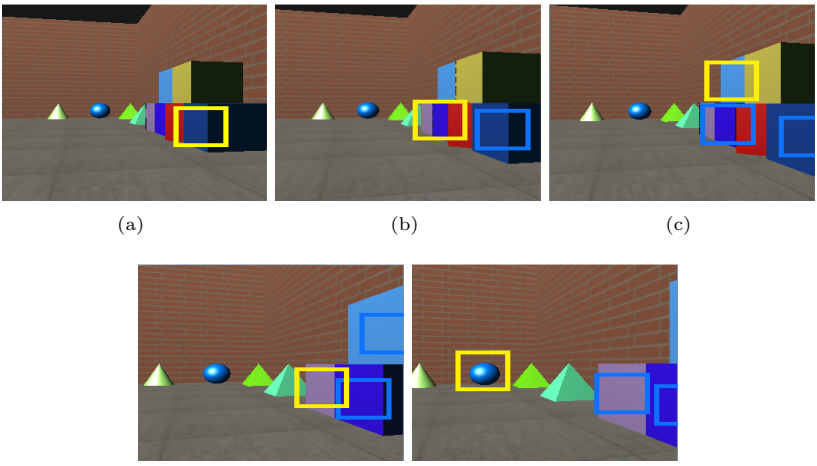
5.3.3 Search With Robot Camera Head

In the third scenario, involving real robot platform, the attention mechanism was required to perform overt attention to the best matching location by bringing the target into center of camera view with rotation of the robot camera head in two degrees of freedom. The search target was provided in form of a picture of the search object and the system was required to mark its presence in a given scene in the first glance of viewing. Then overt attention shift was made by pointing the camera head towards the found target that brought the target locations into center of view frame one after the other. These experiments tested the ability of the system to locate the (estimated) position of the search target in three dimensional space. The top-down conditions (or the search target) provided to the system is shown in figure 5.14(a). Figure 5.14(b) shows the left camera view in which a given search target is to be searched. The system computed the top-down saliency and marked two locations in the scene as visible in figure 5.14(c). Figures 5.14(d) and (e) show the camera view after the covert attention shift that brought the search target into center of camera view. The head angles at



(a) (b)

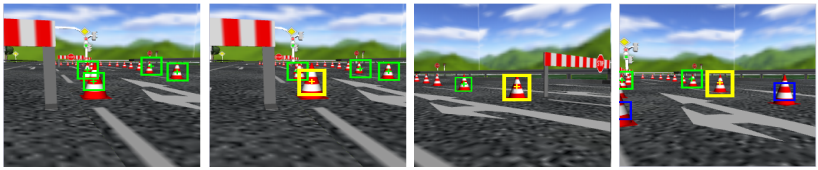
Figure 5.11: Scenarios to test attentive search in 3D environments of SIMORE. (a) A simple indoor scenario. Left image is the search field whereas the right one is the search target. (b) Another scenario in which search with overt attention shift was experimented.



(a) (b) (c)

(d) (e)

Figure 5.12: Fixated locations marked by yellow rectangles while searching for the target using simulated mobile robot scenario given in figure 5.11(a). Inhibited locations are marked with blue rectangles.



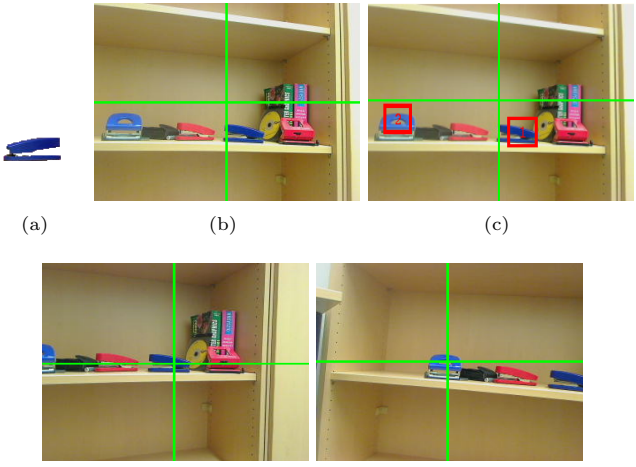
(a) (b) (c) (d)

Figure 5.13: Results of overt attention while searching in the scenario presented in figure 5.11(b). (a) Instances of the target found by the system marked by green rectangles. (b) - (d) First three instances of the search target brought into center of view by overt camera movement. Current FOA is marked by yellow rectangle, the inhibited regions are marked by blue rectangles, and the remaining target instances to be attended are marked by green rectangles.

this status provide sufficient information for other sensors in order to perform some further processing, for example marking the object location in the map of the environment. It is observable that the best matching target location is marked even when it has a different orientation as compared to the given picture in figure 5.14 (a). The second fixation is of course a less matching object marked after inhibition of the best matching case.

5.4 Perceptual Grouping

Grouping of visually similar regions, which are distributed in a given scene, into a perceptible pattern is performed by the attention system while working in *examine* mode. Experiments on this aspect were performed using images in which high level patterns composed of small objects were present. The left column of figure 5.15 demonstrates some of these test cases. The model was first executed in exploration mode until it fixated on one of the components of the macro-level pattern and then the behavior was switched to *examine* in order to highlight this pattern in the subsequent saccades. Exploration of reasons and mechanism of autonomous switching between behaviors is a topic for vast future research. In the current status of the model the switching between behaviors



(a) (b) (c)

(d) (e)

Figure 5.14: Results of attentive search using camera head of the mobile robot platform. (a) Picture of the search target provided to the system. (b) View through the camera of robot. (c) Regions with highest top-down saliency marked by the system. (d) and (e) Camera view after overt attention to the found targets.

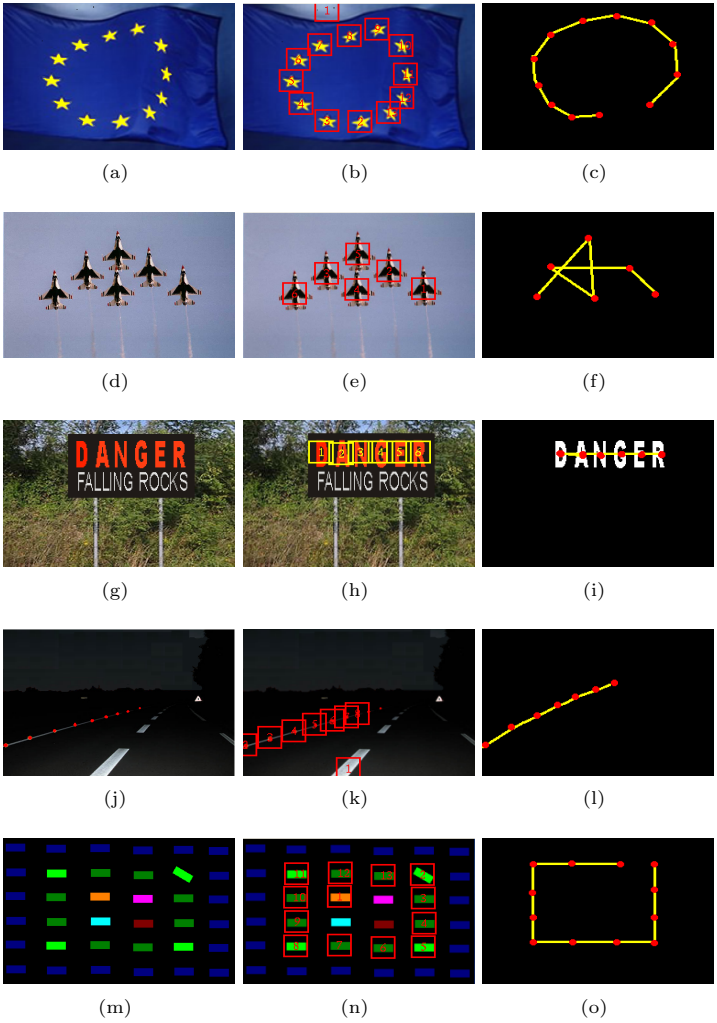
was done manually. Output of *examine* behavior is presented in the middle column of figure 5.15. The images of the European Union flag, the night drive scene, and the image with a hidden rectangle were among the samples in which a few saccades had to be made before entering the required pattern as visible in subfigure (b), (k), and (n) where the first focus of attention (marked as 1 in the small rectangle) is outside the main pattern. It is observable in the output that the model has successfully selected the components of the global pattern in a suitable sequence that follows a scan path reflecting the shape of the respective pattern. This can be seen in the scan paths followed for these images given in the rightmost column of figure 5.15. The circle of stars is picked with fair accuracy, the formation of airplanes is also picked but the scan path needs to be corrected to form a triangle, the letters on the sign board are attended in a suitable sequence that can facilitate reading of the message, and the road side

marked by cat eyes is also followed correctly. Due to the attentional nature of the method, the sequence of fixations is highly dependant on the visual saliency of the individual items in comparison to their neighborhood that may pull attention of the system before the other items. Hence the scan path may not always draw the concerned shape but this problem can be tackled by normalizing the curve using the points of fixation as guiding information.

Output of the system using the test image given in figure 5.15(m) very clearly demonstrates the ability of the attention model to maintain its focus on the pattern under examination even in presence of distractors that possess saliency higher than the components of the pattern. For example, the rectangles with orange and red colors have higher color saliency than the green ones but the system continues to examine the pattern due to excitation of the examined features.

5.5 Chapter Summary

This chapter has presented results obtained from the attention model after running it on three different experimentation platforms. The software to work with static images allows viewing the results along with the intermediate processing being carried out. The results for visual behaviors of *explore*, *search*, and *examine* has been obtained on this platform. Working of attention in *explore* and *search* behaviors has been experimented using the simulation framework SIMORE and the camera head mounted on mobile robot also. The output under all scenarios is promising and shows success of the proposed methodology in advancing the state of the art in this area of research. A formal evaluation of the results and comparison with other existing attention models is performed in the next chapter.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Figure 5.15: Output of the attention system working in *examine* mode. The leftmost column contains the input images, the middle one shows the fixated locations under *examine* behavior, and the rightmost column sketches the scan path followed by the fixations.

6 Evaluation

This chapter presents results of those experiments that were meant specifically for evaluating the performance of the proposed attention model. Currently a standardized set of benchmark images and ground truth data is not available for visual attention. Creation of such a resource is part of the extended work on the research presented in this dissertation. For evaluating validity of the results from the proposed attention system, it was tested using a set of self-created benchmark samples in order to verify the ability of the model to identify saliency with respect to different visual features. Output of the proposed model is evaluated using experiments related to bottom-up as well as top-down aspects of attention. The second context of evaluation was to test robustness of the model under different situations that could degrade its performance. For this purpose experiments were performed using transformed and noisy visual input. Output of the proposed model is compared with the attention models of [IKN98], [AL06], and [BMB01] in order to quantify the achievements gained through the new model. The software of the models of [IKN98] and [AL06] was obtained from their websites, [Itt09] and [Avr09] respectively, while the model of [BMB01] was developed by a former member of our research group, hence its code was available. The chapter is arranged in three main subsections related to evaluation in terms of validity of output, robustness, efficiency, and effectiveness of the model for use in attentive vision applications.

6.1 Validity of Results

The first step of validation of results is to check whether the model creates correct saliency maps for a given input and if the locations fixated by the system are acceptably correct. As the proposed model works for both bottom-up and

top-down pathways, evaluation is needed for both types of maps. For a better organization of presentation, these two facets of assessment are divided into separate subsections given below.

6.1.1 Validation of Bottom-up Attention

In order to verify the ability of the proposed model to determine saliency with respect to the individual features, benchmark images each containing salient objects in context of only one feature were used as input for the model. Figure 6.1 presents results of these evaluation experiments. The first row of figure 6.1 shows the benchmark images each consisting of objects having saliency with respect to only one feature, namely, color, eccentricity, orientation, symmetry, and size. The second row displays the corresponding saliency maps and the third row shows the foci of attention on which the proposed system fixated. It can be clearly seen that the outstanding object due to each feature was marked by the proposed system in the first attempt hence the system's response to individual features is valid.

Results of bottom-up attention produced by some other models of attention using the images shown in figure 6.1(a) to (e) are given in figure 6.2 in order to assess the comparative performance of the proposed model. The first row of figure 6.2 presents output of the previous model developed in our group [BMB01], the second row shows output of the extended-saliency method proposed by [AL06], while the third row contains results of the method proposed in [IKN98]. It is observable that the model of [BMB01] is not successful in identifying color saliency when the regions are separated by some other regions (black background between the colored boxes in figure 6.1(a)) because it computes contrast on the region edges only and ignores the global context. The contrasts due to orientation and size are also not considered in this model, hence, it could not identify the saliency correctly for the image given in figures 6.1(d) and (e). The model proposed in [AL06] does not perform well in the benchmark images because it concentrates mainly on global rarity while ignoring local feature-based saliency. The model of [IKN98] does not compute the feature channels of eccentricity,

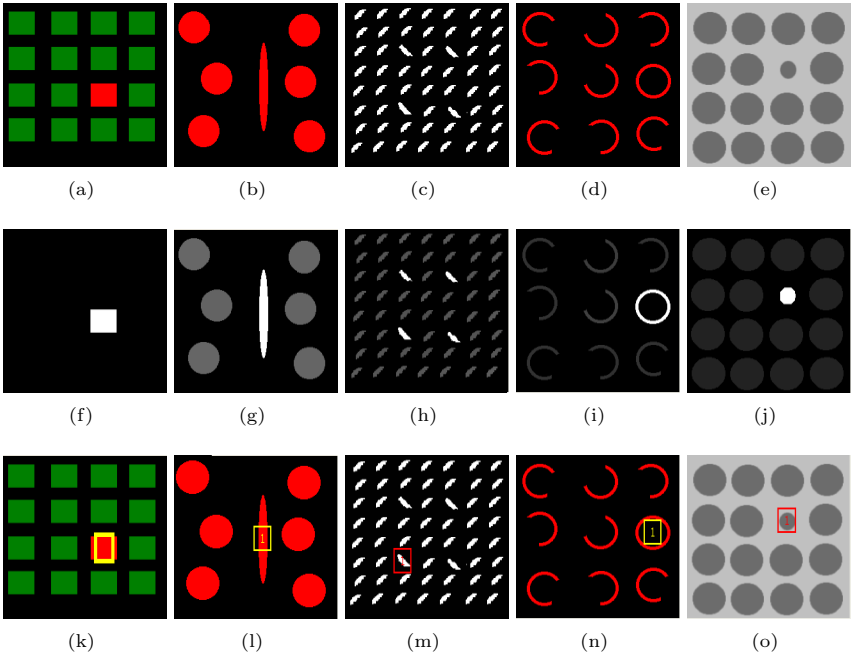


Figure 6.1: Evaluation of response to individual feature channels. Row 1: (a) to (e) are benchmark samples having salient objects due to only one feature, namely, color contrast, eccentricity, orientation, symmetry, and size respectively. Row 2: Corresponding saliency maps produced by the proposed model. Row 3: Fixated locations.

symmetry, and size hence it was unable to pick the correct objects from the input samples given in figures 6.1 (b), (d), and (e).

For a quantitative performance evaluation of the proposed model and comparing it with other existing models, we use the criteria of detection rate σ^d and error rate σ^e from the list of different evaluation metrics mentioned in [AM08b]. The readings to be noted for these metrics while running a model over an input are the number of salient locations marked by human subjects in that image N_s, the count of erroneous fixations falling outside the salient locations N_e, and the number of fixations taken by the model to cover all salient locations N_a. Having

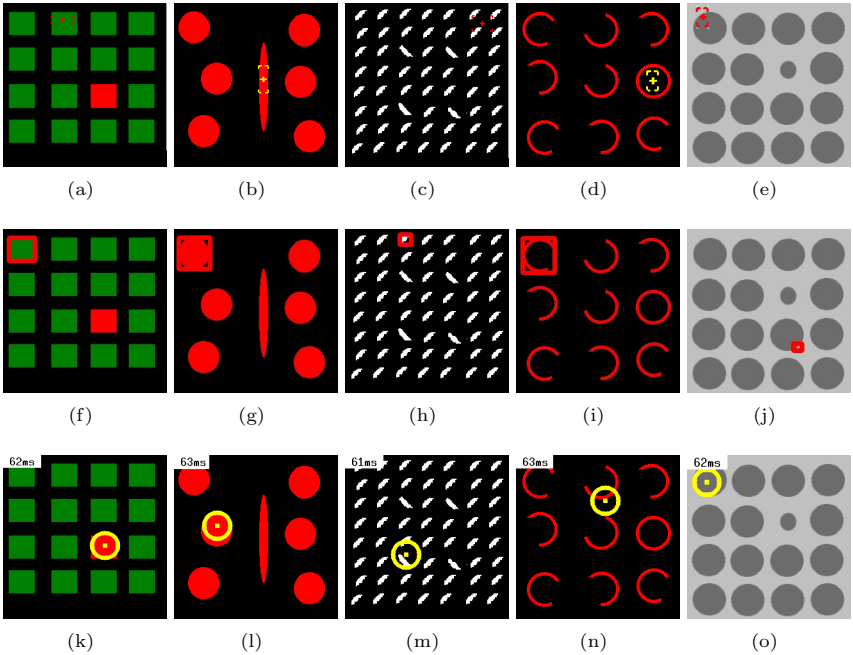


Figure 6.2: Results of three other existing attention models on static scenes given in figure 6.1 (a) to (e). Row 1: Locations fixated by the model of [BMB01]. Row 2: Locations fixated by the model of [AL06]. Row 3: Locations fixated by the model of [IKN98].

these values available the said two performance metrics σ^d and σ^e are computed as follows

$$\sigma^d = N_s/N_a \tag{6.1}$$

$$\sigma^e = N_e/N_a \tag{6.2}$$

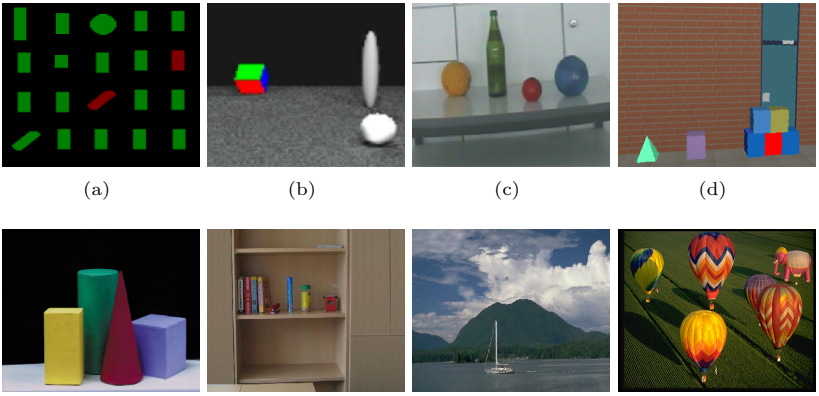
Some input images were selected to perform the experiments for this quantitative evaluation, which are shown in figure 6.3. The number of salient locations N_s as marked by human subjects are given in the second row of table 6.1.

The proposed model and the models given in [BMB01], [AL06], and [IKN98] were allowed to run until all manually marked salient locations were detected, hence N_a was obtained for each image by each model. In order to avoid running of a model for indefinite period of time, in case marking all salient location is beyond the model's capability, the maximum limit for N_a is kept as N_s^2 for $N_s > 3$ and it is set to 10 for $1 \leq N_s \leq 3$. The recorded readings of N_s, N_a, and N_e for the compared attention models using the input images shown in figure 6.3 are listed in table 6.1.

Comparison of the four models under discussion in terms of detection rate σ^d and error rate σ^e in graphical format is provided in figures 6.8 and 6.9. Results of experiments from which data for the said metrics was extracted are shown in figures 6.4 to 6.7. Graphs for the computed values σ^d and σ^e are presented in figures 6.8 and 6.9 respectively. It is noticeable in figure 6.8 that the average detection rate of the proposed model is higher than the other models while the average error rate is equivalent to the lowest rate from the other models as visible in figure 6.9. This analysis shows that the results of the proposed model are valid in context of the state of the art as it performs equally good or a little better in comparison to the existing models in terms of correctness of output.

6.1.2 Validation of Top-Down Attention

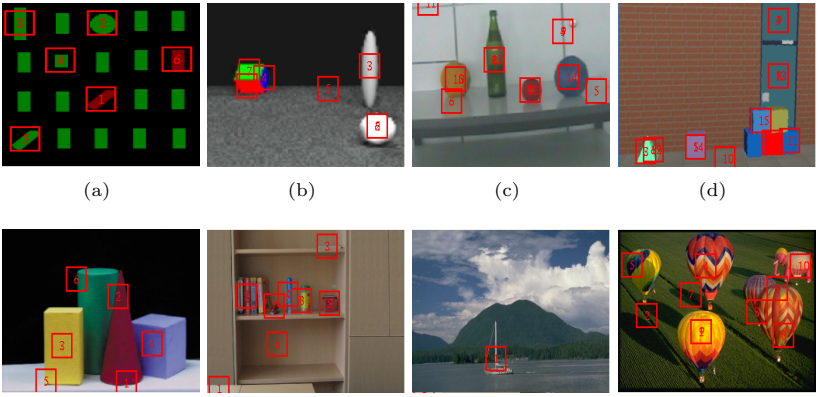
The main measure of success for visual search is the number of fixations before the system finds the best match to the target. The top-down pathway of the proposed model has shown 100 percent success rate with detection of the all n best matching targets in first n fixations in almost all experiments. Therefore, instead of discussing this pathway in terms of success and error rate we examine the validity of results in two other aspects. Firstly, we demonstrate construction of fine-grain saliency maps for a test image with different search targets. Under the fine-grain paradigm the maps should be different when the search target is different. Secondly, the natural system inhibits the already attended locations and the gaze is shifted to other locations having even less similarity with the searched object as fixation on the search target does not remain for a long time.



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

(e) (f) (g) (h)

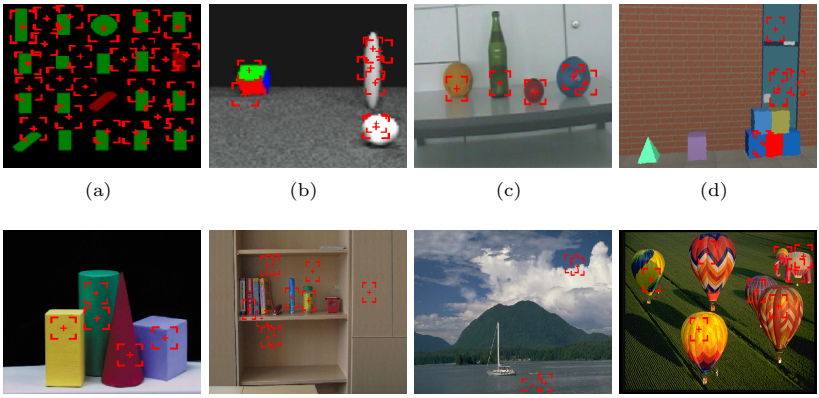
Figure 6.3: Input images used for performing quantitative evaluation of proposed and other attention models. Image codes: (top row) conj, o3d, ball, sim, (bottom row) obj, off, boat, bln



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

(e) (f) (g) (h)

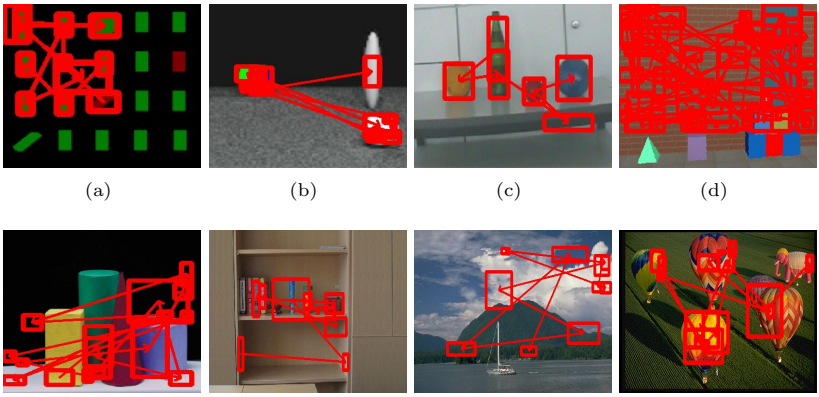
Figure 6.4: Fixations performed by the proposed model (N_a) to cover all salient locations (N_s) on images given in figure 6.3.



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

(e) (f) (g) (h)

Figure 6.5: Fixations performed by the model of [BMB01] (N_a) to cover all salient locations (N_s) on images given in figure 6.3.



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

(e) (f) (g) (h)

Figure 6.6: Fixations performed by the model of [AL06] (N_a) to cover all salient locations (N_s) on images given in figure 6.3.

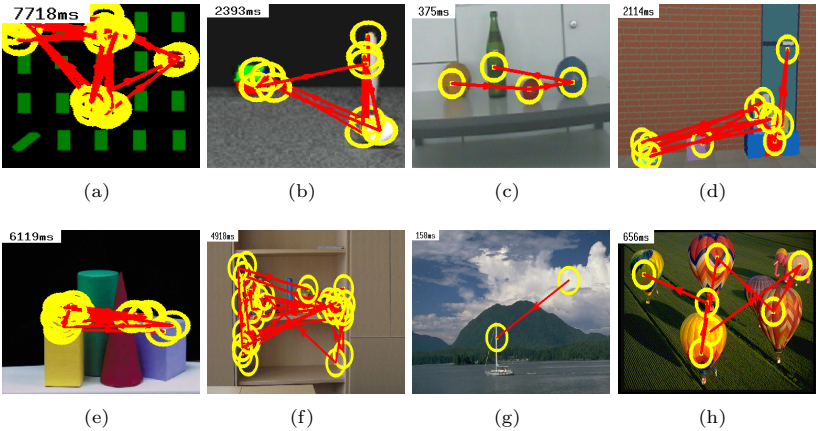


Figure 6.7: Fixations performed by the model of [IKN98] (N_a) to cover all salient locations (N_s) on images given in figure 6.3.

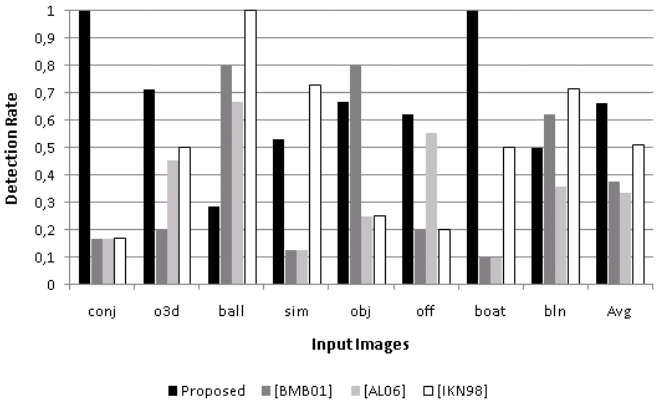


Figure 6.8: Comparison of the proposed model with the models by [BMB01] (Backer), [AL06] (Esal), and [IKN98](Itti) in terms of detection rate σ^d.

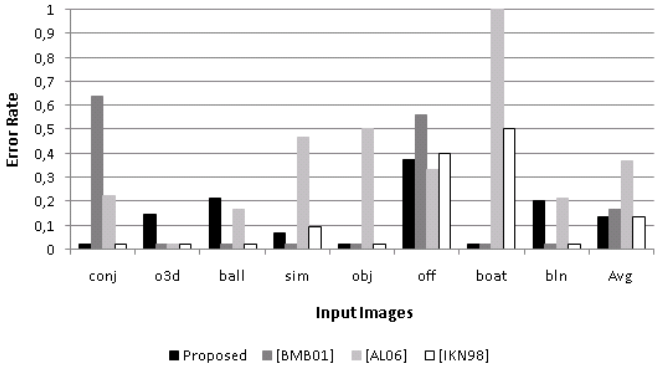


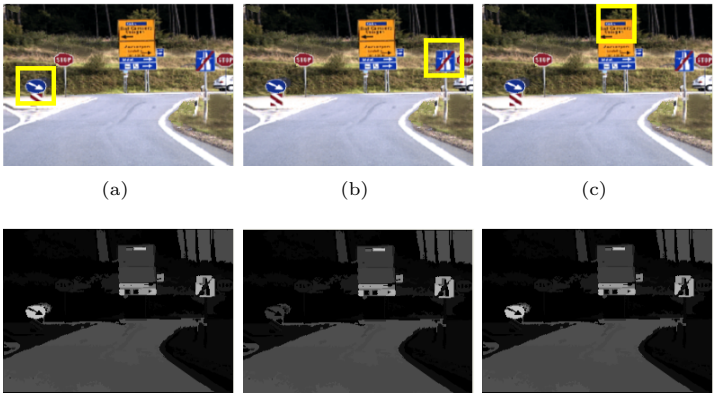
Figure 6.9: Comparison of proposed model with the existing models of [BMB01] (Backer), [AL06] (Esal), and [IKN98](Itti) in terms of error rate σ^e.

Hence for the second aspect we observe the role of inhibition of return using which the system locates less probable targets after detecting the best matches.

Figure 6.10 demonstrates one of the test cases in which two search targets and a search area are shown. Figure 6.11 presents the result of search on the target given in figure 6.10(b) which is fixated as the first FOA as visible in figure 6.11(a). The top-down saliency map causing this FOA is shown in figure 6.11(d) in which the region matching the search target has the highest saliency reflected by its bright color. In the subsequent saccades the system tries to ignore the already attended object by applying an inhibition in order to explore other possible occurrences of the search target in the scene. The suppression on saliency of the first FOA can be seen in figure 6.11(e). In this test case there are no more good matches to the target hence the subsequent FOAs, such as the one in 6.11(f), have decreasing similarity with the target. Figure 6.12 demonstrates results of search on the target given in figure 6.10(c) in the same search area. It can be seen that this time the top-down saliency maps are built totally different as compared to the previous case. These two advantages are achieved due to the fine-grain nature of the proposed methodology. The other approaches of attentional search have not applied this concept yet hence the search targets pop out after several iterations, especially in complex real-life situations.

(a) (b) (c)

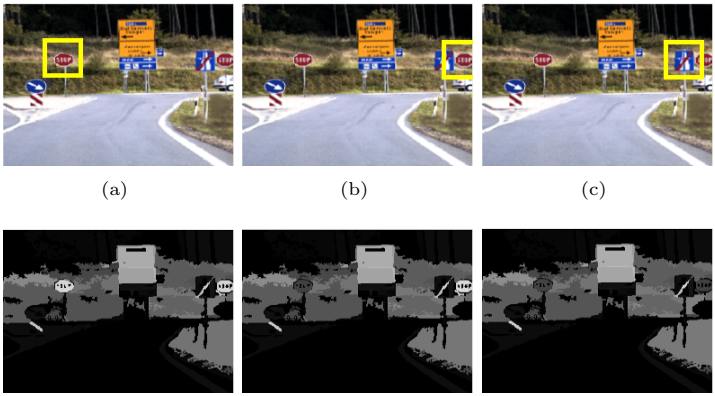
Figure 6.10: Input images for evaluation of top-down attention. (a) Image used as search area. (b) First search target. (c) Second search target.



(a) (b) (c)

(d) (e) (f)

Figure 6.11: Search results for finding the target given in figure 6.10(b) into the figure 6.10(a). (a) to (c) are foci of attention and (d) to (f) are saliency maps each corresponding to the FOA above it.



(a) (b) (c)

(d) (e) (f)

Figure 6.12: Search results for finding the target given in figure 6.10(c) in the figure 6.10(a). (a) to (c) are foci of attention and (d) to (f) are saliency maps each corresponding to the FOA above it.

6.2 Efficiency

We take the computation time taken by a model to process a single image as a metric for measuring the efficiency. The compared models were allowed to fixate for five FOAs and the CPU time taken by the models was noted on the same machine and operating system. For these experiments three models, namely the proposed one, model of [BMB01], and the model of [IKN98] were included. The CPU time of the model by [AL06] could not be noted because its software does not provide this information. The obtained readings are shown graphically in figure 6.13. It is observable that the proposed model is faster than the existing models in most of the cases.

The drop of time curve by Itti's model below the proposed model's curve at the right hand side of figure 6.13 is needed to be investigated. One possible explanation is that the pixel based models use downsized copies of the input hence their reaction time does not rise linearly or exponentially with increase of image resolution. To confirm this aspect, run time of the models using a

set of images containing identical contents with growing resolutions (64x64 to 512x512) was recorded. This data is shown graphically in figure 6.14. A curve for segmentation time alone is also plotted for the proposed method. It is observable that the response of other two models against rise of resolution is nonlinear while the time taken by the proposed method grows steadily with increasing size of images. It is worth mentioning that the model of [IKN98] takes this much time for two feature maps while the proposed method computes five maps including the heavy map of symmetry.

6.3 Effectiveness

In order to evaluate the effectiveness of a model in showing useful attention behavior, we choose to apply the metrics of success rate σ_ϕ^s in context of a given phenomenon ϕ and explorative capability ε^x from the list of evaluation metrics proposed in [AM08b]. The purpose of the measure σ_ϕ^s is to see how many out of the N_s salient locations does the system mark in its first N_s fixations. Hence the model is allowed to fixate only for N_s times and the number of salient objects found by the model is counted as N_f. Using the readings obtained for N_s and N_f, σ_ϕ^s is computed as

$$\sigma_\phi^s = N_f/N_s$$

The metric ε^x quantifies the capability of a model to explore new locations while attending to a scene rather than repeating fixations on already attended objects. For a total N_s salient locations in a scene, if a model fixates on N_D distinct objects out of its N_a fixations then the degree of exploration capability ε^x will be computed as

$$\varepsilon^x = 1 - (N_a - N_D)/N_a$$

It may be noted that a fixation will be counted in N_D even if it falls on different parts of a large object or region that may be considered as distinct from each other in terms of spatial distance or visual characteristics (such as shade of light or shape of corner).

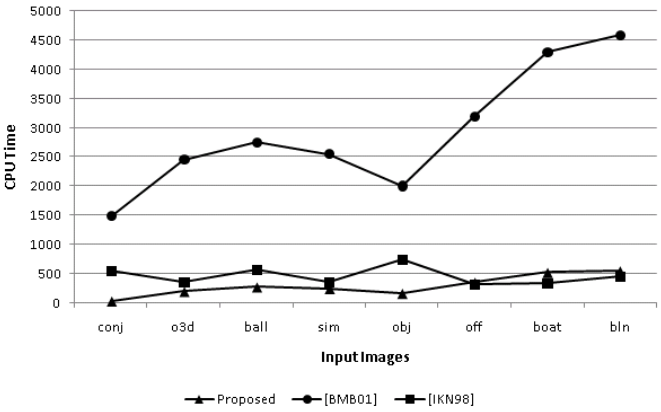


Figure 6.13: Comparison of computation time of the proposed method with [IKN98] and [BMB01] using images shown in figure 6.3.

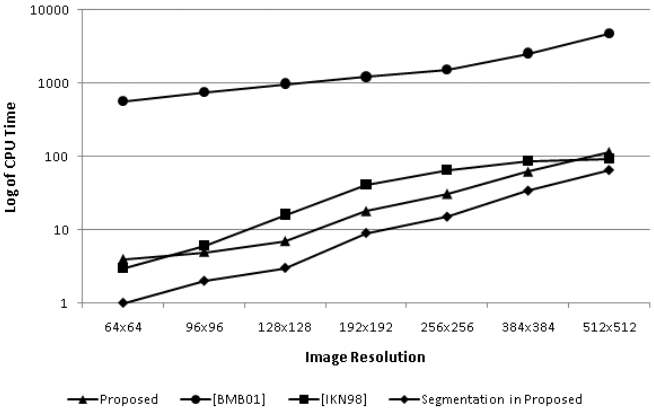
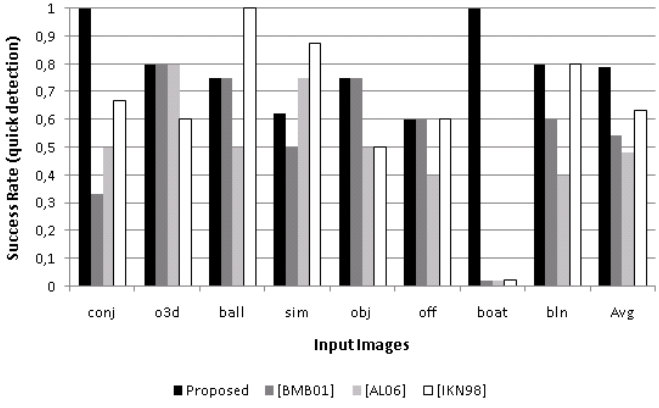


Figure 6.14: Comparison of computation time of the proposed method with [IKN98] and [BMB01] using images with different resolutions but having same contents.

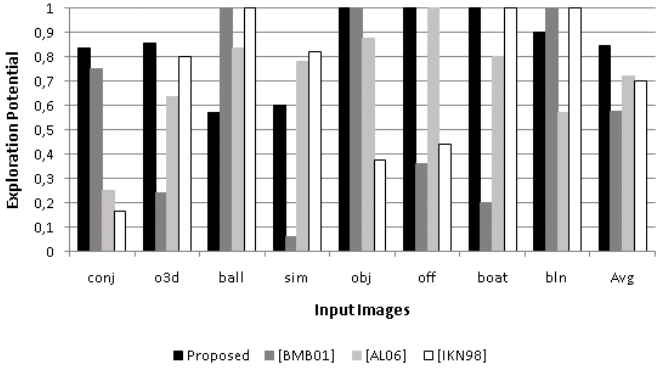
During the experiments that produced the output shown in figures 6.4 to 6.7 the readings of N_f and N_D required for σ^s_ϕ and ε^x were also recoded and are listed in table 6.1. Comparison of the proposed model with the models given in [BMB01], [AL06], and [IKN98]) in terms of these two metrics is presented in figure 6.15. It is observable in figure 6.15(a) that the rate of success in quickly identifying salient objects, i. e. finding the salient locations within the first n fixations when n main salient objects exist in the given input, shown by the proposed model is higher than the existing models. The explorative potential of the proposed model is also highest in the compared models as apparent in figure 6.15(b).

Table 6.1: Readings from evaluation experiments using the images given in figure 6.3. For each model the recorded values are actual salient regions N_s, fixations taken by the model to cover them N_a, error fixations on non-salient regions N_e, targets found in first N_s fixations N_f, and distinct regions N_D fixated out of N_a fixation.

	Input	conj	o3d	ball	sim	obj	off	boat	bln
	N_s	6	5	4	8	4	5	1	5
Proposed	N_a	6	7	14	15	6	8	1	10
	N_e	0	1	3	1	0	3	0	2
	N_f	6	4	3	5	3	3	1	4
	N_D	5	6	8	9	6	8	1	9
[BMB01]	N_a	36	25	5	64	5	25	10	8
	N_e	23	0	0	0	0	14	10	0
	N_f	2	4	3	4	3	3	0	3
	N_D	27	6	5	4	5	9	2	8
[AL06]	N_a	36	11	6	64	16	9	10	14
	N_e	8	0	1	30	8	3	10	3
	N_f	3	4	2	6	2	2	0	2
	N_D	9	7	5	50	14	9	8	8
[IKN98]	N_a	36	10	4	11	16	25	2	7
	N_e	0	0	0	1	0	10	1	0
	N_f	4	3	4	7	2	3	0	4
	N_D	6	8	4	9	6	11	2	7



(a)



(b)

Figure 6.15: Comparison of effectiveness of the proposed model with the existing models by [BMB01], [AL06], and [IKN98]. Comparison in terms of success rate σ_{ϕ}^{s} in context of quickly identifying salient objects is shown in (a) and exploration capability ε^{x} is compared in (b).

6.4 Robustness

The first criterion that can be considered in context of testing robustness of an attention model is its ability to perform equally well on transformed images as mentioned in [DL03]. For this purpose the input images given in figure 6.3 were rotated at 90°, 180°, and 270° and response of the models under comparison was noted to see how many out of N_s fixations on the rotated input matched with those on the untransformed image. Figure 6.16 shows samples from the results out of these experiments. Response of the proposed model and the models of [BMB01], [AL06], and [IKN98] is shown on one of the input images at different angles of rotation.

The robustness of a model against a transformation may be measured using the metric σ^r_T, which is defined as the ratio between the number of matching fixations N^T_m between the first N_s fixations on the untransformed input and the N_s fixations on the transformed image. Hence σ^r_T may be computed as

$$\sigma^r_T = N^T_m / N_s$$

where T may be replaced by a symbol for the related transformation, for example σ^r_{R90} may be used to represent robustness of the model against rotations along 90 degree rotations. Figure 6.17 shows comparison of the four considered models in terms of σ^r_{R90}, σ^r_{R180}, and σ^r_{R270}. It is evident from the bars of average (AVG) in figures 6.17(a), (b) and (c) that the proposed model performs fairly well as compared to the existing ones in all considered cases of rotations.

In addition to the robustness of performance on transformed images, we incorporate the robustness on distorted input as an extended criterion for evaluating attention models because in many situations the vision systems get fairly distorted input because of digitization and transmission errors. For this purpose a selected image was distorted for levels between 0 and 100 using a jpeg compression software, in which 0 denotes no distortion and 100 is the maximum level of distortion that the software could produce. Figure 6.18 shows response of the four models under discussion on the original version of the input and its distorted version at level 100. Results of these experiments using the complete

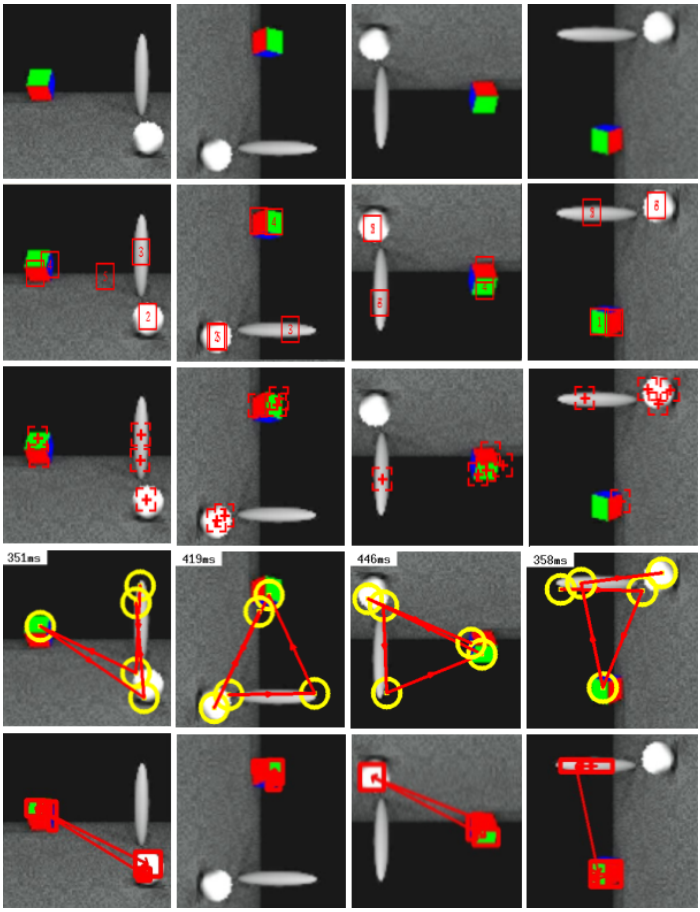
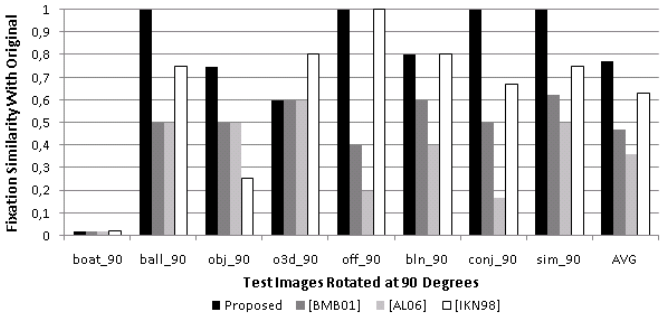
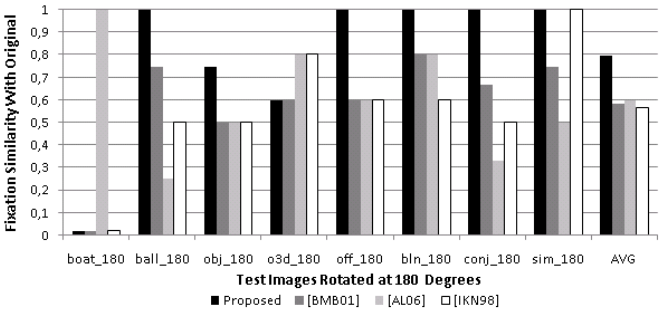


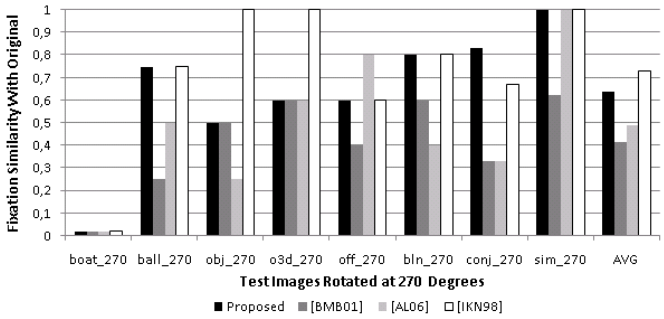
Figure 6.16: Top row: Untransformed input image and its rotated versions at 90°, 180°, and 270°. Second row: Foci of attention marked by the proposed system. Third to fifth rows: Fixations by models of [BMB01], [AL06] and [IKN98] respectively.



(a)



(b)



(c)

Figure 6.17: Comparison of robustness of the proposed model with the existing models by [BMB01], [AL06], and [IKN98] against rotation of input at 90, 180, and 270 degrees. Comparison in terms of σ^r_{R90} is given in (a), for σ^r_{R180} in (b), and for σ^r_{R270} in (c).

set of distorted images are plotted in figure 6.19. The average of the readings for all distortion levels (AVG) shows moderate robustness of the proposed model on distorted input while Itti's model shows the highest robustness. The reason for this is naturally the region-based nature of the proposed model in which segmentation errors on distorted input causes errors in the results of higher stages of the model also.

6.5 Chapter Summary

In this chapter an extensive evaluation of results produced by the proposed attention model has been discussed along with a comparison with other existing models for which software is pubically available. Quantitative metrics have been applied to judge the validity of results, efficiency of model performance, effectivity in attending required locations, and robustness against transformations and distortions in the input. The evaluation has shown that the proposed model has shown not only valid results but has better performance in many aspects as compared to the contemporary attention models.

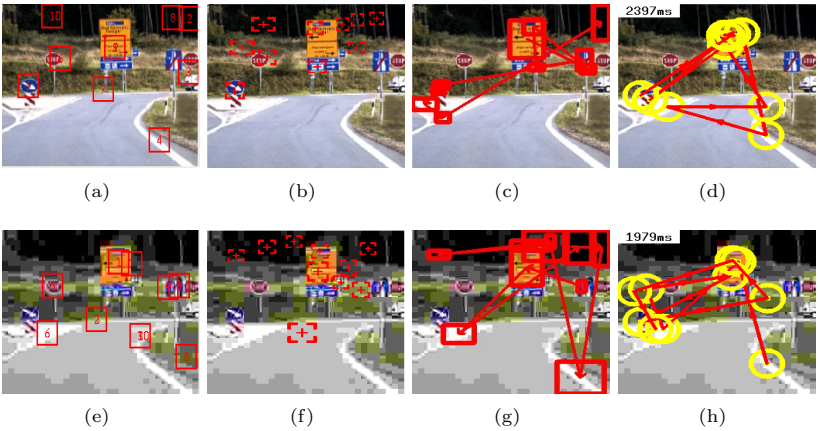


Figure 6.18: (a) to (d) Output of the proposed model, [BMB01], [AL06], and [IKN98] respectively on distortion level 0 of a sample input. (e) to (h) Output of the four models on distortion level 100.

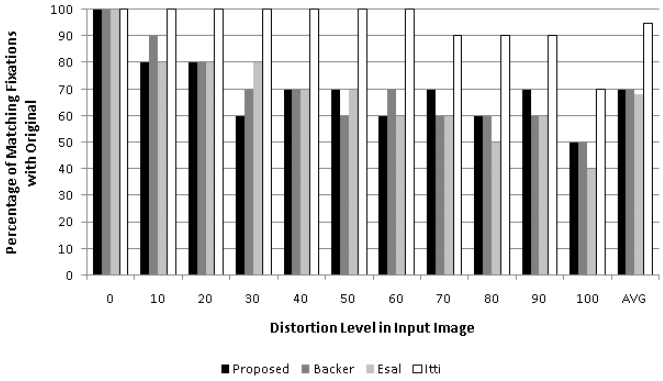


Figure 6.19: Comparison of the proposed model with models of [BMB01], [AL06], and [IKN98] in terms of response on distorted input.

7 Conclusion and Outlook

This chapter first summarizes the contributions of the new model discussed in this dissertation and then reviews the achievements made in the field of visual attention modeling. After that a critical discussion on the theoretical aspects of the proposed early clustering verses the commonly used late clustering is presented. After completion of the work presented here many directions have become visible that need to be investigated further for reaching an ultimate model of attention that could perform like the human vision system. The dissertation is concluded with indications of such directions.

7.1 Scientific Contributions

The main objective of this work was to develop a model of visual attention that could perform fast enough such that it could be used as a building block for a real-time robotic vision system. Another objective, which was rather contradicting with the first goal, was to increase the number of feature channels so that the scope of saliency detection may be expanded to more complex features apart from the basic ones. For this purpose an approach to apply an early clustering on the visual data was investigated. This clustering reduced the amount of data for the subsequent steps of the model while sufficiently preserving the visual information. Owing to the acceleration in the processing speed, five feature channels could be included into the attention process maintaining the ability to process multiple input frames per second. These many features have not been considered in any of the known attention models. The formal inclusion of the feature channel of size contrast for the first time is also an innovative contribution in this area.

The first milestone of this work was to develop a suitable method to convert the visual input into its constituent clusters while optimizing computation time and quality of segmentation. A new algorithm was developed for the segmentation part that kept human visual perception into consideration and, hence, the clusters obtained by this technique were mostly biologically plausible [AM06]. The other milestone of this work was developing a fast and robust algorithms for extracting features involved in visual attention from segmented regions. For the bottom-up pathway, new methods were developed to compute saliency maps using channels of color [AM07b] and other features including eccentricity, orientation, symmetry, and size [AM08a]. Methods for applying inhibition of return and determining pop-out in the region-based paradigm were also specially developed [AM07d]. As this model was meant for mobile vision systems, a solution for handling attention and IOR in dynamic scenarios was needed. A memory-based design for the inhibition mechanism was developed to tackle with this problem that also incorporated capability of handling attention in three dimensional space [AM07c]. Another major contribution of this work is the introduction of fine-grain saliency computation in the modeling of top-down attention pathway. Groundwork for using fine-grain saliency using color channel was established in [AM07e] and it was extended to work with other feature channels with design of methods for top-down map fusion and IOR on both top-down and bottom-up saliency maps [AM08c]. The proposed model integrates the bottom-up and top-down pathways in a single architecture which was also not done before by any other model. As a byproduct of this work an application of visual attention in perceptual grouping [AM07a] and a proposal for metrics and methods for quantitative evaluation of visual attention models [AM08b] also came into being.

7.2 Discussion

Early segmentation has shown some advantages in terms of increasing efficiency, handling the global contrast, helping in finding shape based features, and precise localization of the salient objects. On the other hand it lags behind the pixel based approaches in context of robustness because output of segmentation is sig-

nificantly affected when the visual input gets noisy, distorted, or transformed. The work presented in this dissertation has shown that the early clustering approach can yield substantial advantages over the pixel based approaches but the technique or format of clustering may be improved in order to get more biologically plausible clusters rather than simple segmentation. The use of downscaled copies of the visual input as used by pixel based approaches may also be considered as very crude form of clustering but it results in loss of visual data while converting a block containing pixels of different colors into a singled downscaled unit. Hence, development of a clustering approach that would lead to splitting of the given input into logical visual components with optimization between data reduction and data loss would be a significant step forward.

Region-based techniques have shown more explorative capability as visible in results given in figure 6.15 in chapter 6. The proposed model and the region-based model by [AL06] have a higher average of fixating on new locations as compared to the other two pixel-based techniques. The reason behind is that the units of processing are bigger clusters in the region-based approaches that allow computation of contrast at a broader (or global) level leading to a large jump while making a saccade. On the other hand, pixel based approaches may remain on a close vicinity of the last fixated point because the unit of processing is a single pixel or a small cluster in the down-sized version of the input.

In context of top-down attention, the proposed region-based methodology with the innovation of constructing the fine-grain saliency maps separate from the bottom-up maps is an efficient and robust alternative to the existing approaches. The early clustering or segmentation allowed associating more shape-based features to the salient locations and also using these features while performing visual search. The proposed method is also immune to bottom-up saliency of distractors in every feature channel and does not require any tuning of parameters or adjusting of weights. The memory based inhibition mechanism has also shown success in static as well as dynamic scenarios.

Evaluation of the proposed technique as presented in chapter 6 shows that the early clustering approach is a competitive alternative for the pixel-based approaches. Its performance is better than the existing techniques in many as-

pects. Although due to dependency on the output of segmentation results from the region based attention lag behind in robustness against distortions and noise in the input. Further improvement can be achieved if research is done in this direction.

7.3 Outlook

Investigations on the proposal of early clustering have shown a significant potential in this concept as evident from the promising results demonstrated by the proposed model. The main questionable item is the format of clustering because the region segmentation is a solution for computer processing of images and details of clustering in human vision are not clearly understood yet. A study leading to a computational model of clustering mechanism in human vision, that formulates objects from point and color information, will be a substantial contribution into this field. The research will involve understanding of the role of neuron hierarchy starting from receptors, through the Gangleon cells, up to the visual cortex. A mixture of pyramid techniques as used by [IK00] and [TCW+95] and segmentation may be required that would treat regions in multiple resolutions rather than a single high resolution layer. The part of multi-resolution nature of the visual input, as suggested by the literature such as [JSN92], [FFLB85], [EW02], and [RR03] etc., has to be kept in consideration also.

Some investigations in the human visual attention, such as [MI99], are suggesting an active role of object recognition in attention. A further step in this direction will be applying the top-down influence of knowledge (or memory) on the clustering process that helps in filling up the missing pieces of information, for example vision in the blind spot of human eye. Involvement of knowledge in the attention processing may also lead to obtaining saliency in terms contrast of known and unknown objects and attention based upon nature of objects, e. g. a ball between oranges will be attended by a human observer even when both types of objects have similar visual attributes.

Computation of symmetry as a feature channel is a computationally heavy process, at least with the currently available state of the technology and paradigm of

algorithms. Hence it is suggested to replace the symmetry channel with a simpler channel such as closure to further improve computational efficiency. Symmetry is counted as a doubted channel in natural visual attention [WH04] hence its exclusion will not have a major affect on validity aspect of results. Computation of other important feature channels including texture, motion, and depth from stereo using region-based approaches are also further steps to be investigated.

As an ultimate target, the attention mechanism should be implemented in autonomous robotic machines that should be able to perform all required visual behaviors without human intervention. The system should be able to switch between behaviors autonomously depending upon the situation and requirements of the active task. In real-world scenarios, attention is performed in a three dimensional space. An improvement in the current state of the art will be to determine saliency in 3D rather than two dimensional maps. This aspect needs merging of map information and localization of objects in 3D environment because items will have to be overtly attended and inhibited in a three dimensional world.

Bibliography

[Ahu96] Ahuja, N.: A Transform for Multiscale Image Segmentation by Integrated Edge and Region Detection. In: *Transactions on Pattern Analysis and Machine Intelligence* 18 (1996), pp. 1211–1235

[AL06] Avraham, T.; Lindenbaum, M.: Esaliency - A Stochastic Attention Model Incorporating Similarity Information and Knowledge-Based Preferences. In: *WRUPKV-ECCV 06*. Graz, 2006

[AM06] Aziz, M. Z.; Mertsching, B.: Color Segmentation for a Region-Based Attention Model. In: *Workshop Farbbildverarbeitung (FWS06)*. Ilmenau - Germany, 2006, pp. 74–83

[AM07a] Aziz, M. Z.; Mertsching, B.: An Attentional Approach for Perceptual Grouping of Spatially Distributed Patterns. In: *DAGM 2007, LNCS 4713*. Heidelberg - Germany, 2007, pp. 345–354

[AM07b] Aziz, M. Z.; Mertsching, B.: Color Saliency and Inhibition in Region Based Visual Attention. In: *WAPCV 2007*. Hyderabad - India, 2007, pp. 95–108

[AM07c] Aziz, M. Z.; Mertsching, B.: Color Saliency and Inhibition using Static and Dynamic Scenes in Region Based Visual Attention. In: *Attention in Cognitive Systems, LNAI 4840* (2007), pp. 234–250

[AM07d] Aziz, M. Z.; Mertsching, B.: Pop-out and IOR in Static Scenes with Region Based Visual Attention. In: *WCAA-ICVS 2007*. Bielefeld - Germany, 2007

[AM07e] Aziz, M. Z.; Mertsching, B.: Region-Based Top-Down Visual Attention Through Fine Grain Color Map. In: *13. Workshop Farbbildverarbeitung (FWS07)*. 13. Workshop Farbbildverarbeitung (FWS07), 2007, pp. 83–92

[AM08a] Aziz, M. Z.; Mertsching, B.: Fast and Robust Generation of Feature Maps for Region-Based Visual Attention. In: *Transactions on Image Processing* 17 (2008), pp. 633–644

[AM08b] Aziz, M. Z.; Mertsching, B.: Towards Standardized Metrics and Methods for Evaluation of Visual Attention Models. In: *WAPCV 2008*. Santorini - Greece, 2008, pp. 180–193

[AM08c] Aziz, M. Z.; Mertsching, B.: Visual Search in Static and Dynamic Scenes using Fine-Grain Top-Down Visual Attention. In: *ICVS 08, LNCS 2008*. Santorini - Greece, 2008, pp. 3–12

[AMSS06] Aziz, M. Z.; Mertsching, B.; Shafik, M. S.; Stemmer, R.: Evaluation of Visual Attention Models for Robots. In: *ICVS 06*. New York - USA, 2006, pp. index–20

[ASM05a] Aziz, M. Z.; Shafik, M. S.; Mertsching, B.: Saliency Based Color Segmentation for Visual Attention of Mobile Robots. In: *GVIP 05*. Cairo, 2005

[ASM05b] Aziz, M. Z.; Stemmer, R.; Mertsching, B.: Region-based Depth Feature Map for Visual Attention in Autonomous Mobile Systems. In: *AMS 2005*. Stuttgart - Germany, 2005, pp. 89–95

[ASMM05] Aziz, M. Z.; Shafik, M. S.; Mertsching, B.; Munir, A.: Color Segmentation for Visual Attention of Mobile Robots. In: *ICET 05*. Islamabad, 2005, pp. 115–120

[Ats07] Atsumi, M.: Stochastic Attentional Selection and Shift on the Visual Attention Pyramid. In: *ICVS 2007*. Bielefeld - Germany, 2007

[Avr09] Avraham, T.: *http://www.cs.technion.ac.il/ tammya/*. Internet Resource, last accessed March 2009

[BDST04] Brockers, R.; Drüe, S.; Stemmer, R.; Thiem, J.: TSR - eine Forschungsplattform für TeleSensorische Robotikanwendungen. In: *Robotik 2004, VDI-Berichte*. München - Germany, 2004, pp. 863–870

[BFR84] Boch, R.; Fischer, B.; Ramspergern, E.: Express-Saccades of the Monkey: Reaction Times Versus Intensity, Size, Duration, and Eccentricity of Their Targets. In: *Experimental Brain Research* (1984), pp. 223–231

[BKV05] Broek, E. L. d.; Kisters, P. M. F.; Vuurpijl, L. G.: Content-Based Image Retrieval Benchmarking: Utilizing Color Categories and Color Distributions. In: *Journal of Imaging Science and Technology* 49 (2005), pp. 293–301

[BMB01] Backer, G.; Mertsching, B.; Bollmann, M.: Data- and Model-Driven Gaze Control for an Active-Vision System. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001), pp. 1415–1429

[BRD+84] Balaszi, A. G.; Rootman, J.; Drance, S. M.; Schuttzer, M.; Douglas, G. R.: The Effect of Age on the Nerve Fibre Population of the Human Optic Nerve. In: *American Journal of Ophthalmology* 97 (1984), pp. 760–766

[Bro] Broadbent, D.

[BS03] Bradley, A. P.; Stentiford, W. M.: Visual Attention for Region of Interest Coding in JPEG 2000. In: *Journal of Visual Communication & Image Representation* (2003), pp. 232–250

[BT05] Bruce, N.D.B.; Tsotsos, J. K.: An Attentional Framework for Stereo Vision. In: *Canadian Conference on Computer and Robot Vision*. Victoria, 2005, pp. 88–95

[CC06] Campana, G.; Casco, C.: Interaction Between Facilitation and Inhibition of Return Facilitates Visual Search. In: *ECVP Abstract Supplement, Perception* Bd. 35, 2006

[CCM03] Chalmers, A.; Cater, K.; Maflioli, D.: Visual Attention Models for Producing High Fidelity Graphics Efficiently. In: *19th Spring Conference on Computer Graphics*, 2003, pp. 39–46

[CCSP03] Castel, A. D.; Chasteen, A. L.; Scialfa, C. T.; Pratt, J.: Adult Age Differences in the Time Course of Inhibition of Return. In: *Psychological Sciences and Social Sciences* 58 (2003), pp. 256–259

[CJSW01] Cheng, H. D.; Jiang, X. H.; Sun, Y.; Wang, J.: Color image segmentation: Advances and Prospects. In: *Pattern Recognition* 34 (2001), pp. 2259–2281

[CL94] Carron, T.; Lambert, P.: Color Edge Detector Using Jointly Hue, Saturation, and Intensity. In: *IEEE Conference on Image Processing*. Austin, USA, 1994, pp. 977–1081

[CM01] Carrasco, M.; McElree, B.: Covert Attention Aaccelerates the Rate of Visual Information Processing. In: *Psychology* 98 (2001), pp. 5363–5367

[CMKG03] Chen, T. Q.; Murphey, Y. L.; Karlsen, R.; Gerhart, G.: Color Image Segmentation in Color and Spacial Domain. In: *Lecture Notes on Artificial Interlligence* 2718 (2003), pp. 72–82

[CSP+87] Curcio, C. A.; Sloan, K. R.; Packer, O.; Hendrickson, A. E.; Kalina, R. E.: Distribution of Cones in Human and Monkey Retina: Individual Variability and Radial Asymmetry. In: *Science* 236 (1987), pp. 579–582

[CT03] Cutzu, F.; Tsotsos, J. K.: The Selective Tuning Model of Attention: Psychophysical Evidence for a Suppressive Annulus Around an Attended Item. In: *Vision Research* (2003), pp. 205–219

[DBZ07] Dankers, A.; Barnes, N.; Zelinsky, A.: A Reactive Vision System: Active-Dynamic Saliency. In: *ICVS 07*. Bielefeld, Germany, 2007

[DC04] Deng, H.; Clausi, D. A.: Unsupervised Image Segmentation Using a Simple MRF Model with a New Implementation Scheme. In: *Pattern Recognition* 37 (2004), pp. 2323–2335

[DD63] Deutsch, J. A.; Deutsch, D.: Attention: Some Theoretical Considerations. In: *Psychological Review* 70 (1963), pp. 80–90

[De 60] De Valois, R. L.: Color Vision Mechanisms in the Monkey. In: *Journal of General Physiology.* 43 (1960), pp. 115–128

[Dec05] Deco, G.: The Computational Neuroscience of Visual Cognition: Attention, Memory and Reward. In: *WAPCV 2004, LNCS 3368*, 2005, pp. 100–117

[Des49] Descartes, R.: *Les Passions de l'âme.* Paris : Le Gras, 1649

[DL03] Draper, B. A.; Lionelle, A.: Evaluation of Selective Attention under Similarity Transforms. In: *WAPCV 03*, 2003

[DLD04] Dao, D. Y.; Lu, Z.-L.; Dosher, B. A.: Orientation Bandwidth of Selective Adaptation. In: *Journal of Vision* 4 (2004), pp. 11–21

[DM01] Deng, Y.; Manjunath, B.S.: Unsupervised Segmentation of Color-Texture Regions in Images and Video. In: *Transactions on Pattern Analysis and Machine Intelligence* 23 (2001), pp. 800–810

[DN04] Dow, M.; Nunnally, R.: An Edge Based Image Segmentation Method. In: *ISMRM 2004*, 2004, pp. poster

[EJ86] Eriksen, C.W.; James, J.D.S.: Visual Attention Within and Around the Field of Focal Attention: A Zoom Lens Model. In: *Perception and Psychophysics* 40 (1986), pp. 225–240

[EW02] Eglen, S. J.; Willshaw, D. J.: Influence of Cell Fate Mechanisms Upon Retinal Mosaic Formation: A Modelling Study. In: *Development* 129 (2002), pp. 5399–5408

[EY85] Eriksen, C.W.; Yeh, Y.: Allocation of attention in the visual field. In: *Journal of Experimental Psychology: Human Perception and Performance* 11 (1985), pp. 583–597

[EZW97] Engel, S.; Zhang, X.; Wandell, B.: Color Tuning in Human Visual Cortex Measured with Functional Magnetic Resonance Imaging. In: *Nature* 388 (1997), pp. 68–71

[FBR05] Frintrop, S.; Backer, G.; Rome, E.: Goal-Directed Search with a Top-Down Modulated Computational Attention System. In: *DAGM 2005, LNCS 3663*, 2005, pp. 117–124

[FDJ99] Fernandez-Duque, D.; JOHNSON, M. L.: Attention Metaphors: How Metaphors Guide the Cognitive Psychology of Attention. In: *Cognitive Science* 23 (1999), pp. 83–116

[FFLB85] Farber, D. B.; Flannery, J. G.; Lolley, R. N.; Bok, D.: Distribution Patterns of Photoreceptors, Protein, and Cyclic Nucleotides in the Human Retina. In: *Investigative Opththalmology and Visual Science* 26 (1985), pp. 1558–1568

[FH04] Felzenszwalb, P. F.; Huttenlocher, D. P.: Efficient Graph-Based Image Segmentation. In: *International Journal of Computer Vision* 59 (2004), pp. 167–181

[For06] Ford, J. L.: *www.worqx.com*. Internet Resource, last accessed August 2008, 2006

[FS06] Flusser, J.; Suk, T.: Rotation Moment Invariants for Recognition of Symmetric Objects. In: *Transactions on Image Processing* 15 (2006), pp. 3784–3790

[FYEA01] Fan, J.; Yau, D. K. Y.; Elmagarmid, A. K.; Aref, W. G.: Automatic Image Segmentation by Integrating Color-Edge Extraction and Seeded Region Growing. In: *Transactions on Pattern Analysis and Machine Intelligence* 10 (2001), pp. 1454–1466

[GE94] Gibson, B. S.; Egeth, H.: Inhibition of Resturn to Object-Based and Environment-Based Locations. In: *Perception and Psychophysics* 55 (1994), pp. 323–339

[GGS05] Goolsby, B. A.; Grabowecky, M.; Suzuki, S.: Adaptive Modulation of Color Salience Contingent Upon Global Form Coding and Task Relevance. In: *Vision Research* (2005), pp. 901–930

[GM92] Goodale, M. A.; Milner, A. D.: Separate Pathways for Perception and Action. In: *Trends in Neuroscience* 15 (1992), pp. 20–25

[Gou84] Gouras, P.: Color Vision. In: *Progress in Retinal Research* 3 (1984), pp. 227–261

[Ham05] Hamker, F. H.: Modeling Attention: From Computational Neuroscience to Computer Vision. In: *WAPCV 2004, LNCS 3368*, 2005, pp. 118–132

[Har68] Hartlin, H. K.: *Les Prix Nobel en 1967: Visual Receptors and Retinal Interaction*. Nobel foundation, 1968

[HB90] Haddon, J. F.; Boyce, J. F.: Image Segmentation by Unifying Region and Boundary Information. In: *Transactions on Pattern Analysis and Machine Intelligence* 12 (1990), pp. 929–948

[HBSH07] Heinke, D.; Backhaus, A.; Sun, Y.; Humphreys, G. W.: The Selective Attention for Identification Model(SAIM): Simulating Visual Search in Natural Color Images. In: *WAPCV 2007, LNAI 4840* (2007), pp. 141–154

[HE96] Healey, C. G.; Enns, J. T.: *A Perceptual Color Segmentation Algorithm, Technical Report TR-96-09*. Department of Computer Science, University of British Columbia, 1996

[Her64] Hering, E.: *Outlines of a Theory of the Light Sense*. Cambridge : Harvard University Press, 1964

[Hil08] Hilker, Christian: *Entwicklung einer dynamisch beeinflussbaren 3D-Engine zur Robotersimulation*, Studienarbeit, Universität Paderborn, Diplomarbeit, January 2008

[HJ57] Hurvich, L. M.; Jameson, D.: An Opponent-Process Rheory of Color Vision. In: *Psychological Review* 64 (1957), pp. 384–404

[HRB+03] Heidemann, G.; Rae, R.; Bekel, H.; Bax, I.; Ritter, H.: Integrating Context-Free and Context-Dependant Attentional Mechanisms for Gestural Object Reference. In: *ICVS 03, LNCS2626*, 2003, pp. 22–33

[HT06] H.Maunsell, J.; Treue, S.: Feature-Based Attention in Visual Cortex. In: *Trends in Neurosciences* 29 (2006), pp. 317–322

[HW06] Hawes, N.; Wyatt, J.: Towards Context-Sensitive Visual Attention. In: *Second International Cognitive Vision Workshop (ICVW06)*, 2006

[IK00] Itti, L.; Koch, C.: A Saliency Based Search Mechanism for Overt and Covert Shifts of Visual Attention. In: *Vision Research* (2000), pp. 1489–1506

[IKN98] Itti, L.; Koch, U.; Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. In: *Transactions on Pattern Analysis and Machine Intelligence* 20 (1998), pp. 1254–1259

[IPV00] Ikonomakis, N.; Plataniotis, K. N.; Venetsanopoulos, A. N.: Color Image Segmentation for Mutimedia Applications. In: *Journal of Intelligent and Robotic Systems* 28 (2000), pp. 5–20

[Ita01] Itakura, S.: Attention to Repeated Events in Human Infants (Homo Sapiens): Effects of Joint Visual Attention Versus Stimulus Change. In: *Animal Cognition* (2001), pp. 281–284

[Itt61] Itten, J.: *The Elements of Color*. New York, USA : John Wiley & Sons Inc., 1961

[Itt04] Itti, L.: Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention. In: *IEEE Transactions on Image Processing* 13 (2004), pp. 1304–1318

[Itt09] Itti, L.: *http://ilab.usc.edu/toolkit/downloads.shtml*. Internet Resource, last accessed March 2009

[Jav71] Javons, W. S.: The Power of Numerical Discrimination. In: *Nature* 3 (1871), pp. 218–282

[JSN92] Jonas, J. B.; Schneider, U.; Naumann, G.O.H.: Count and Density of Human Retinal Photoreceptors. In: *Graefe's Archive for Clinical and Experimental Ophthalmology* 230 (1992), pp. 505–510

[KG98] Kiryati, N.; Gofman, Y.: Detecting Symmetry in Grey Level Images: The Global Optimization Approach. In: *International Journal of Computer Vision* 29 (1998), pp. 29–45

[KH95] Kramer, A. F.; Hahn, S.: SPLITTING THE BEAM:. Distribution of Attention Over Noncontiguous Regions of the Visual Field. In: *Psychological Science* 6 (1995), pp. 381–386

[KHSM08] Kutter, Oliver; Hilker, Christian; Simon, Alexander; Mertsching, Bärbel: Modeling and Simulating Mobile Robots Environments. In: *3rd International Conference on Computer Graphics Theory and Applications (GRAPP 2008)*. Funchal - Portugal, January 2008

[Knu07] Knudsen, E. I.: Fundamental Components of Attention. In: *Annual Review of Neuroscience* 30 (2007), pp. 57–78

[Koc99] Koch, C.: *Biophysics of Computation*. New York : Oxford University Press, 1999

[Kol91] Kolb, H.: The Neural Organization of the Human Retina. In: *Principles and Practices of Clinical Electrophysiology of Vision* (1991), pp. 25–52

[KS06] Keller, Y.; Shkolnisky, Y.: A Signal Processing Approach to Symmetry Detection. In: *Transactions on Image Processing* 15 (2006), pp. 2198–2207

[LaB83] LaBerge, D.: Spatial extent of attention to letters and words. In: *Journal of Experimental Psychology: Human Perception and Performance* 9 (1983), pp. 371–379

[LB01] Leclercq, P.; Bräunl, T.: A Color Segmentation Algorithm for Real-Time Object Localization on Small Embedded Systems. In: *Robot Vision 2001, LNCS 1998*, 2001, pp. 69–76

[LCWB97] LaBerge, D.; Carlson, R.L.; Williams, J.K.; Bunney, B.G.: Shifting Attention in Visual Space: Tests of Moving-Spotlight Models Versus an Activity-Distribution Model. In: *Journal of experimental psychology. Human perception and performance* 23 (1997), pp. 1380–1392

[LD04] Lanyon, L.; Denham, S.: A Model of Object-Based Attention That Guides Active Visual Search to Behaviourally Relevant Locations. In: *WAPCV 2004, LNCS 3368*, 2004, pp. 42–56

[LHG97] Laar, P.V.D.; Heskes, T.; Gielen, S.: Task-Dependent Learning of Attention. In: *Neural Networks* 10 (1997), pp. 981–992

[LJBV05] López, F.; J.M.Valiente; Baldrich, R.; Vanrell, M.: Fast Surface Grading Using Color Statistics in the CIE Lab Space. In: *IbPRIA 2005, LNCS 3523* (2005), pp. 666–673

[LLY+05] Lu, Z.; Lin, W.; Yang, X.; Ong, E; Yao, S.: Modeling Visual Attention's Modulatory Aftereffects on Visual Sensitivity and Quality Evaluation. In: *Transactions on Image Processing* 14 (2005), pp. 1928–1942

[LMMP59] Lettvin, J. Y.; Maturana, H. R.; McCulloch, W. S.; Pitts, W. H.: What the Frog's Eye Tells the Frog's Brain. In: *Proc. Institute of Radio Engineers* 47 (1959), pp. 1940–1951

[LPA95] Law, M. B.; Pratt, J.; Abrams, R. A.: Color-Based Inhibition of Return. In: *Perception & Psychophysics* (1995), pp. 402–408

[LT04] Loo, P. K.; Tan, C. L.: Adaptive Region Growing Color Segmentaion for Text Using Irregular Pyramid. In: *DAS 2004, LNCS 3163*, 2004, pp. 264–275

[Mac42] MacAdam, D. L.: Visual Sensitivities to Color Differences in Daylight. In: *Journal of the Optical Society of America* 32 (1942), pp. 247–274

[Mac49] MacAdam, D. L.: Colour Discrimination and the Influence of Colour Contrast on Acuity. In: *Documenta Ophthalmologica* 3 (1949), pp. 214–237

[Mah96] Mahnke, F.: *Color, Environment, and Human Response*. Detroit : Van Nostrand Reinhold, 1996

[MCBT06] Meur, O. L.; Callet, P. L.; Barba, D.; Thoreau, D.: A Coherent Computational Approach to Model Bottom-Up Visual Attention. In: *Transactions on Pattern Analysis and Machine Intelligence* 28 (2006), pp. 802–817

[MG75] Monasterio, F.M. D.; Gouras, P.: Functional Properties of Ganglion Cells of the Rhesus Monkey Retina. In: *Journal of Physiology* 251 (1975), pp. 167–195

[MGS+04] Michalke, T.; Gepperth, A.; Schneider, M.; Fritsch, J.; Goerick, C.: Towards a Human-like Vision System for Resource-

Constrained Intelligent Cars. In: *ICVS 2007*. Bielefeld - Germany, 2004, pp. 264–275

[MI99] Murata, A.; Iwase, H.: Visual Attention Models - Object-Based Theory of Visual Attention. In: *SMC 99*. Tokyo, Japan, 1999, pp. 60–65

[Muk02] Mukherjee, J.: MRF Clustering for Segmentation of Color Images. In: *Pattern Recognition Letters* 23 (2002), pp. 917–929

[NCMS04] Nelson, J. D.; Cottrell, G. W.; Movellan, J. R.; Sereno, M. I.: Yarbus Lives: A Foveated Exploration of How Task Influences Saccadic Eye Movement. In: *Journal of Vision* 4 (2004), pp. 741–741

[Ner04] Neri, P.: Attentional Effects on Sensory Tuning for Single-Feature Detection and Double-Feature Conjunction. In: *Vision Research* (2004), pp. 3053–3064

[NI05] Navalpakkam, V.; Itti, L.: Modeling the Influence of Task on Attention. In: *Vision Research* (2005), pp. 205–231

[NI06a] Navalpakkam, V.; Itti, L.: Optimal Cue Selection Strategy. In: *NIPS 2006*. Cambridge, 2006, pp. 1–8

[NI06b] Navalpakkam, V.; Itti, L.: Top-Down Attention Selection is Fine-Grained. In: *Journal of Vision* 6 (2006), Oct, pp. 1180–1193

[OAE93] OIshausen, B. A.; Anderson, C.H.; Essenla, D. C. V.: A Neurobiological Model of Visual Attention and Invariant Pattern Recognition Based on Dynamic Routing of Information. In: *The Journal of Neuroscience* 13 (1993), pp. 4700–4719

[OH03] Ouerhani, N.; Hügli, H.: A Model of Dynamic Visual Attention for Object Tracking in Natural Image Sequences. In: *IWAN 2003, LNCS 2686*, 2003, pp. 702–709

[OKS80] Ohta, Y. I.; Kanade, T.; Sakai, T.: Color Information for Region Segmentation. In: *Computer Graphics and Image Processing* 13 (1980), pp. 222–241

[OM02] Oliveira, R.; Monteiro, L. H. A.: Symmetry Detection Using Global-Locally Coupled Maps. In: *ICANN 02, LNCS 2415*, 2002, pp. 75–80

[OMY05] Okubo, M.; Mugishima, Y.; Yosuke, G.: Facilitation of Return in Voluntary Orienting to Visual Attributes. In: *Japanese Psychological Research* 47 (2005), pp. 271–279

[OPR78] Ohlander, R.; Price, K.; Reddy, D. R.: Picture Segmentation Using Recursive Region Splitting Method. In: *Computer Graphics and Image Processing* 8 (1978), pp. 313–333

[Ost35] Osterberg, G.: Topography of the Layer of Rods and Cones in the Human Retina. In: *Acta Ophthalmologica* 6 (1935), pp. 11–102

[PHH97] Postma, E. O.; Herik, H. J. d.; Hudson, P.T. W.: SCAN: A Scalable Model of Attentional Selection. In: *Neural Networks* 10 (1997), pp. 993–1015

[PI07] Peters, R. J.; Itti, L.: Beyond Bottom-Up: Incorporating Task-Dependent Influences into a Computational Model of Spatial Attention. In: *CVPR 07*, 2007

[PKK97] Pratt, J.; Kingstone, A.; Khoe, W.: Inhibition of Return in Location- and Identity-Based Choice Decision Tasks. In: *Perception & Psychophysics* 59 (1997), pp. 964–971

[Pol41] Polyak, S.L.: *The Retina*. Chicago, USA : University of Chicago Press, 1941

[Pou53] Poulton, E. C.: Two Channel Listening. In: *Journal of Experimental Psychology* 46 (1953), pp. 91–96

[PSL02] Park, S. J.; Shin, J. K.; Lee, M.: Biologically Inspired Saliency Map Model for Bottom-up Visual Attention. In: *BMCV 02, LNCS 2525*, 2002, pp. 418–426

[QAG82] Quigley, H. A.; Addicks, E. M.; Green, W. R.: Optic Nerve Damage in Human Glaucoma: III Quantitative Correlation of Nerve Fibre Loss and Visual Defect in Glaucoma Ischemic Neuropathy and Toxic Neuropathy. In: *Archives of Ophthalmology* 100 (1982), pp. 135–146

[RC04] Ramstrom, O.; Christensen, H. I.: Object Based Visual Attention: Searching for Objects Defined by Size. In: *WAPCV 2004*. Prague, 2004

[RD03] Reynolds, J. H.; Desimone, R.: Interacting Roles of Attention and Visual Salience in V4. In: *Neuron* 37 (2003), pp. 853–863

[Ren00] Rensink, R.A.: The Dynamic Representation of Scenes. In: *Visual Cognition 7(1)* 7 (2000), pp. 17–42

[RR03] Raven, M. A.; Reese, B. E.: Mosaic Regularity of Horizontal Cells in the Mouse Retina Is Independent of Cone Photoreceptor Innervation. In: *Investigative Ophthalmology and Visual Science* 44 (2003), pp. 965–973

[RT06] Rothenstein, A.; Tsotsos, J.K.: Selective Tuning: Feature Binding Through Selective Attention. In: *LNCS 4132, ICANN 06*. Athens, Greece, 2006, pp. 548–557

[SF03] Sun, Y.; Fischer, R.: Object-Based Visual Attention for Computer Vision. In: *Artificial Intelligence* 146 (2003), pp. 77–123

[SHS01] Shinoda, H.; Hayhoe, M. M.; Shrivastava, A.: What Controls Attention in Natural Environments. In: *Vision Research* 41 (2001), pp. 3535–3545

[Sim09] Simon, Alexander: *Modellierung der physikalischen und dynamischen Eigenschaften mobiler Roboterplattformen in einer virtuellen Realität*, Diplomarbeit, Universität Paderborn, Diplomarbeit, January 2009

[Sin99] Sinclair, D.: *Voronoi Seeded Color Segmentation, Technical Report 3*. Cambridge : AT&T Laboratories, 1999

[SK94] Sharbek, W.; Koshan, A.: *Color Image Segmentation: A survey, Technical Report 94 - 32*. Berlin - Germany : Fachbereich 13 Informatik, Technical University Berlin, October 1994

[SK00] Snyder, J.; Kingstone, A.: Inhibition of Return and Visual Search: How Many Separate Loci are Inhibited? In: *Perception & Psychophysics* 62 (2000), pp. 452–458

[SMC+03] Shulman, G. L.; McAvoy, M. P.; Cowan, M. C.; Astafiev, S. V.; Tansy, A. P.; Avossa, G.; Corbetta, M.: Quantitative Analysis of Attention and Detection Signals During Visual Search. In: *Journal of Neurophysiology* 90 (2003), pp. 3384–3397

[SSA04] Sigal, L.; Sclaroff, S.; Athitsos, V.: Skin Color-Based Video Segmentation under Time-Varying Illumination. In: *Transactions on Pattern Analysis and Machine Intelligence* 26 (2004), pp. 862–877

[Ste01] Stentiford, F.: An Estimator for Visual Attention Through Competitive Novelty With Application to Image Compression. In: *Picture Coding Symposium*. Seoul - Korea, 2001, pp. 101–104

[Ste05] Stentiford, F.: Attention Based Facial Symmetry Detection. In: *ICAPR 05, LNCS 3687*. Bath - UK, 2005, pp. 112–119

[Ste07] Stentiford, F.: Attention Based Auto Image Cropping. In: *WCAA - ICVS 2007*. Bielefeld - Germany, 2007

[TA97] Tabb, M.; Ahuja, N.: Multiscale Image Segmentation by integrated Edge and Region Detection. In: *Transactions on Pattern Analysis and Machine Intelligence* 6 (1997), pp. 642–654

[TB97] Tremeau, A.; Borel, N.: A Region Growing and Merging Algorithm to Color Segmentation. In: *Pattern Recognition* 30 (1997), pp. 1191–1203

[TCW+95] Tsotsos, J. K.; Culhane, S. M.; Wai, W. Y. K.; Lai, Y.; Davis, N.; Nuflo, F.: Modeling Visual Attention Via Selective Tuning. In: *Artificial Intelligence* 78 (1995), pp. 507–545

[TG67] Treisman, A.; Geffen, G.: Selective Attention: Perception or Response? In: *Quarterly Journal of Experimental Psychology* 19 (1967), pp. 1–18

[TG80] Treisman, A. M.; Gelade, G.: A Feature-Integration Theory of Attention. In: *Cognitive Psychology* 12 (1980), pp. 97–136

[TIR05] Tsotsos, J. K.; Itti, L.; Rees, G.: *A Brief and Selective History of Attention. In: Neurobiology of Attention.* Elsevier, 2005

[Tre60] Treisman, A.: Contextual Cues in Selective Listening. In: *Quarterly Journal of Experimental Psychology* 12 (1960), pp. 242–248

[TT92] Tsang, P.W.M.; Tsang, W. H.: A Fast Linear Shape From Shading. In: *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 1992, pp. 734–736

[TT96] Tsang, P.W.M.; Tsang, W. H.: Edge Detection on Object Color. In: *IEEE International Conference on Image Processing*, 1996, pp. 1049–1052

[TTW01] Tagare, H. D.; Toyama, K.; Wang, J. G.: A Maximum-Likelihood Strategy for Directing Attention during Visual Search. In: *Transactions on Pattern Analysis and Machine Intelligence* 23 (2001), pp. 490–500

[UM82] Ungerleider, L.G.; Mishkin, M.: Two Cortical Visual Systems. In: *In: Analysis of Visual Behavior* (1982), pp. 549–586

[WCF89] Wolfe, J. M.; Cave, K.R.; Franzel, S.L.: Guided Search: An Alternative to the Feature Integration Model for Visual Search. In: *Journal for Experimental Pscychology: Human Perception and Performance* 15 (1989), pp. 419–433

[WH04] Wolfe, J. M.; Horowitz, T. S.: What Attributes Guide the Deployment of Visual Attention and How Do They Do It? In: *Nature Reviews, Neuroscience* 5 (2004), pp. 1–7

[WLW98] Weaver, B.; Lupianez, J.; Watson, F. L.: The Effects of Practice on Object-Based, Location-Based, and Static-Display Inhibition of Return. In: *Perception & Psychophysics* 60 (1998), pp. 993–1003

[Wol94] Wolfe, J. M.: Guided Search 2.0: A Revised Model of Visual Search. In: *Psychonomic Bulletin and Review* 1 (1994), pp. 202–238

[Wol00] Wolfe, J.: Visual attention. In: *De Valois KK, Academic Press*, 2000, pp. 335–386

[Yar67] Yarbus, A. L.: *Eye Movements and Vision.* New York : Plenum Press, 1967

[YM92] Yang, G.; Masland, R. H.: Direct Visualization of the Dendritic and Receptive Fields of Directionally Selective Retinal Ganglion Cells. In: *Science* 258 (1992), pp. 1949–1952

[YM94] Yang, G.; Masland, R. H.: Receptive Fields and Dendritic Structure of Directionally Selective Retinal Ganglion Cells. In: *Journal of Neuroscience* 14 (1994), pp. 5267–5280

[Zah04] Zaharescu, A.: *Towards a Biologically Plausible Active Visual Search Model.* Masters Thesis, York University, 2004

List of Abbreviations

S.No	Abbreviation	Explanation
1	FOA	Focus of Attention
2	FOR	Facilitation of Return
3	HIS	Hue, Intensity, Saturation
4	HVS	Human Vision System
5	IOR	Inhibition of Return
6	RGB	Red, Green, Blue
7	ROI	Region of Interest
8	WTA	Winner Take All

List of Tables

List of Figures