ATTENTION BASED IMAGE RETRIEVAL

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÜLŞAH TÜMÜKLÜ ÖZYER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

SEPTEMBER 2012

Approval of the thesis:

ATTENTION BASED IMAGE RETRIEVAL

submitted by **GÜLŞAH TÜMÜKLÜ ÖZYER** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Prof. Dr. Fatoş T. Yarman Vural _____
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Assist. Prof. Dr. Sinan Kalkan
Computer Engineering Department, METU _____

Prof. Dr. Fatoş T. Yarman Vural
Computer Engineering Department, METU _____

Assoc. Prof. Dr. İlkay Ulusoy
Electrical and Electronics Eng. Dept., METU _____

Assist. Prof. Dr. Pınar Duygulu Şahin
Computer Engineering Department, Bilkent University _____

Assist. Prof. Dr. Ahmet Oğuz Akyüz
Computer Engineering Department, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: GÜLŞAH TÜMÜKLÜ ÖZYER

Signature :

ABSTRACT

ATTENTION BASED IMAGE RETRIEVAL

Özyer, Gülşah Tümüklü
Ph.D., Department of Computer Engineering
Supervisor : Prof. Dr. Fatoş T. Yarman Vural

September 2012, 128 pages

This thesis proposes a content-based image retrieval (CBIR) system based on the human visual attention, called Attention-based Image Retrieval (ABIR). The proposed ABIR system handles CBIR problem from the perspective of human perception. An efficient visual attention model specific to CBIR problem derived from the computational visual attention model of Itti and Koch is suggested. ABIR system defines the CBIR system as an attention task, where query and images in the database are considered together to extract region of interests.

The ABIR system consists of saliency map computing, region extraction, feature extraction and similarity matching steps using the saliency information. Bottom-up Normalization Algorithm, Top-down Normalization Algorithm and Top-down Feature Map Weighting Algorithm are proposed to compute the saliency maps. Bottom-up normalization and top-down normalization algorithms attack the normalization process of Itti-Koch model to compute saliency of images. Bottom-up normalization algorithm computes the normalization parameters from the all images in the dataset. On the other hand, top-down normalization algorithm normalizes the images in the

dataset by using query image. Top-down feature map weighting algorithm combines the feature maps of an image in the dataset by using the query image. The features of salient regions are computed by using proposed the saliency-based feature integration algorithm and saliency-based feature selection algorithm. A saliency-based similarity matching algorithm ranks the images with respect to the query image. The proposed ABIR system is tested on STIM and SIVAL object datasets and high resolution airport images. The retrieval results are superior compared to the selected state of the art CBIR systems.

Keywords: content-based image retrieval, visual attention.

ÖZ

DİKKAT TABANLI GÖRÜNTÜ ERİŞİMİ

Özyer, Gülşah Tümüklü
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Fatoş T. Yarman Vural

Eylül 2012, 128 sayfa

Bu tez çalışmasında, dikkat tabanlı görüntü erişim (DTGE) sistemi olarak adlandırılan görsel dikkate dayalı bir içerik tabalı görüntü erişim (İTGE) sistemi önerildi. Önerilen DTGE sistemi, içerik tabanlı görüntü erişimini, insan algılamasını göz önüne alarak tanımlıyor. Itti ve Koch tarafından geliştirilen görsel dikkat modeline dayalı, etkili ve İTGE problemine özgün bir sistem önerildi. Önerilen DTGE sistemi, sorgu görüntüsünü ve veri setindeki görüntüleri birlikte kullanarak İTGE yaklaşımını bir dikkat görevi olarak ele alır.

DTGE sistemi belirginlik bilgisini kullanan belirginlik haritası hesaplama, alan çıkarma, öznitelik çıkarma ve benzerlik eşleştirmesi adımlarından oluşmaktadır. Belirginlik haritası hesaplamak için aşağıdan yukarıya düzgeleme, yukarıdan aşağıya düzgeleme ve yukarıdan aşağıya öznitelik haritası ağırlıklandırma algoritmaları önerildi. Bu algoritmalar Itti-Koch modelinin düzgeleme ve öznitelik haritası birleştirme adımlarını İTGE problemine uyarlar. Belirgin bölgelerin öznitelikleri, belirginlik tabanlı öznitelik birleştirme ve belirginlik tabanlı öznitelik seçme algoritmaları ile çıkarıldı. İki görüntü arasındaki benzerlik yine görüntülerde ki bölgelerin belirginlik bilgisine day-

alı bir eşleştirme algoritması ile hesaplandı. Önerilen sistem, STIM ve SIVAL nesne veri setlerinde ve yüksek çözünürlüklü havalimanı görüntülerinde test edildi. Elde edilen deney sonuçlarına göre, önerilen sistem en son gelişmeleri yansıtan İTGE sistemlerinden daha iyi erişim perfomansı göstermiştir.

Anahtar Kelimeler: içerik tabanlı görüntü erişimi, görsel dikkat.

To my husband

ACKNOWLEDGMENTS

I would like to express my deepest grateful to my thesis advisor, Prof. Dr. Fatoş T. Yarman Vural, for her guidance, support and encouragement during my thesis at METU. Her knowledge and experiences always influence on me to change my thoughts and point of view to the problems and life.

Thanks to my thesis committee Assist. Prof. Dr. Sinan Kalkan and Assoc. Prof. Dr. İlkay Ulusoy for all the meetings they guide and support my progress in the thesis and also thanks to Assist. Prof. Dr. Pınar Duygulu Şahin and Assist. Prof. Dr. Ahmet Oğuz Akyüz for their advices and comments. I would like to thank Prof. Dr. James Z. Wang to inspire me about my thesis subject during my visit in Penn State University, USA.

I would like to thank my labmates, Sarper Alkan, Orhan Fırat, Ruşen Aktaş, Buğra Özkan, Ekin Gedik , Özge Öztimur Karadağ, Cüneyt Mertayak , Emre Akbaş, Mete Özay, Çağlar Şenaras, İsmet Yalabık, Ulaş Yılmaz, Gülcan Can, Fatih Titrek and all the others in the Image Processing and Pattern Recognition Laboratory for their friendship and helpfulness. My warmest thanks to Aykut Erdem and Erkut Erdem for their friendship and encouragement. I would like to thank my friends Levent Bayindir and Alev Mutlu for sharing hard times during our thesis. I want to give my special thanks to Ruken Çakıcı. She always with me in difficult times and encourages me to complete my thesis.

My parents Gülten (annem) and İsmail (babam), as well as my brother Uğur deserve spacial thanks for their kindness, patient, love and encouragement through the years.

Finally, I would like to give my very special thanks to my husband Barış for his love, support and being always with me in hard times. In fact, this thesis cannot be completed without him.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ vi

ACKNOWLEDGMENTS ix

TABLE OF CONTENTS x

LIST OF TABLES xiii

LIST OF FIGURES xv

CHAPTERS

1 INTRODUCTION 1

1.1 Motivation 3

1.2 The Major Contribution of This Thesis 5

1.3 Outline of the Thesis 6

2 BACKGROUND FOR VISUAL ATTENTION AND CONTENT-BASED IMAGE RETRIEVAL SYSTEMS 8

2.1 Introduction 8

2.2 Human Visual Attention 10

2.2.1 Theories for Human Visual Attention in Cognitive Psychology 12

2.2.2 Visual Attention Models in Neuroscience 15

2.3 Biologically Inspired Computational Visual Attention Models 16

2.3.1 The Itti-Koch Model 18

2.3.2 Otner Computational Models 19

2.4 Content-Based Image Retrieval 20

2.4.1 Saliency-based Image Retrieval 24

2.4.2 Localized Content Based Image Retrieval 27

2.5 Summary 28
3 AN ATTENTION BASED IMAGE RETRIEVAL SYSTEM 29
3.1 Content-Based Image Retrieval Problem Definition by a Human Centric Approaches 30
3.2 The Itti-Koch Model 32
3.2.1 Feature Map Extraction 33
3.2.2 Normalization and Conspicuity Map Computing . 35
3.2.3 Saliency Map Computing 36
3.2.4 Winner-take-all network and Inhibition of Return . 37
3.3 System Overview for the Suggested Attention Based Image Retrieval (ABIR) 38
3.4 Saliency Map Extraction for ABIR System 39
3.4.1 A Holistic Approach for Bottom-Up Normalization 43
3.4.2 An Object-Based Top-Down Normalization Method 47
3.4.3 Top-Down Feature Map Weighting Algorithm . . . 52
3.5 Saliency Region Extraction 55
3.6 Feature Extraction from the Salient Regions 56
3.6.1 Saliency-based Feature Integration Algorithm . . . 58
3.6.2 Saliency-based Feature Selection Algorithm 60
3.7 Similarity Matching 61
3.8 Summary 64
4 EXPERIMENTS 65
4.1 Datasets 67
4.1.1 The STIM Dataset 67
4.1.2 The SIVAL Dataset 68
4.1.3 The Airport Dataset 69
4.2 Experiments with the STIM dataset 71
4.2.1 Saliency Map Extraction 75
4.2.2 Analyzing the Number of Salient Regions 85
4.2.3 Time Complexity Analysis of Saliency Map Computation and Region Extraction 91

4.2.4 Analysing Feature Extraction from Salient Regions 92
4.2.5 Similarity Matching of Salient Regions 97
4.2.6 Comparison of Segmented Regions with Salient Regions 101
4.3 Experiments with Sival Dataset 103
4.4 Itti-Koch and SIFT comparison 107
4.5 Experiments with Satellite Images 109
4.6 Summary 113
5 CONCLUSION AND FUTURE DIRECTIONS 116
5.1 Future Directions 118
REFERENCES 121
CURRICULUM VITAE 128

LIST OF TABLES

TABLES

Table 2.1 Learning-based methods for CBIR systems (The table is adapted from [19]). 22

Table 4.1 The average precision results of Experiment Set_{SM} for the first 5 retrieved images by testing Saliency Map extraction algorithms. 78

Table 4.2 Prescision (Pre) and Recall (Rec) results of ABIR system for Itti-Koch Model. 86

Table 4.3 Prescision (Pre) and Recall (Rec) value for Bottom-Up Normalization Algorithm. 87

Table 4.4 Prescision (Pre) and Recall (Rec) value for Top-Down Normalization Algorithm. 88

Table 4.5 Prescision (Pre) and Recall (Rec) value for Top-Down Feature Map Weighting Algorithm. 89

Table 4.6 Average Saliency Map computation and region extraction times of an image for algorithms in seconds for different number of salient regions (DR) 92

Table 4.7 Feature Extraction precision results for the first 5 retrieved images in the database. 93

Table 4.8 Prescision (Pre) and Recall (Rec) value of the ABIR system with the Feature Integration Method 98

Table 4.9 Prescision (Pre) and Recall (Rec) value of the ABIR system with Feature Selection Method 99

Table 4.10 Precision results of Experiment Set_{match} for various criteria where $p = AP$, $p = RFxIPF$, $p = AVxIPF$, $p = AV$. 100

Table 4.11 Precision results for the CBIR system with Normalized cut segmentation and ABIR 103
Table 4.12 Average AUC values over 30 independent runs for all categories . . 106
Table 4.13 Comparison of SIFT and Itti-Koch 108
Table 4.14 SIFT, Reduced SIFT and SIFTmap results for STIM dataset 110
Table 4.15 SIFT, Reduced SIFT and SIFTmap results for SIVAL dataset 110
Table 4.16 Number of patches extracted from high resolution Airport images and the total area percentage of patches to the original high resolution images 112
Table 4.17 Results for Airplane Localization 115

LIST OF FIGURES

FIGURES

Figure 2.1 A general overview of the Visual Attention Methodologies by different disciplines. 9

Figure 2.2 (a) Attentional capture (b) Cueing Experiment. (Figures are taken from [5]). 11

Figure 2.3 (a) Feature Search (b) Conjunction Search. (Figures are taken from [5]) 12

Figure 2.4 Feature Integration Theory. (The figure is taken from [3]) 14

Figure 2.5 Guided Search Theory (The figure is taken from [4]). 15

Figure 2.6 A general scheme for extraction of Visual Attention points by different disciplines (The figure is adapted from [5]). 17

Figure 2.7 The structure of VOCUS (The figure is taken from [6]). 20

Figure 2.8 Visual attention model of Stentiford (The figure is taken from [7]). . 26

Figure 3.1 Object-based Approach of CBIR. 31

Figure 3.2 Holistic Approach of CBIR. 32

Figure 3.3 Itti-Koch Model (The figure is taken from [8]). 33

Figure 3.4 The normalization operator N(.). Note that the normalization operation accentuate the highest peak and suppresses similar neighboring peaks.(The figure is taken from [8]). 36

Figure 3.5 Itti-Koch Model Saliency Map (a) Input image (b)) Itti-Koch Saliency Map of input image. 37

Figure 3.6 The block diagram representation of Attention-based image retrieval 38

Figure 3.7 General structure of the proposed Computational Model of Visual Attention and ABIR. Red Arrows show steps which are for Bottom-Up Normalization Algorithm. Green Arrows for Top-Down Normalization Algorithm and Blue ones for Top-Down Feature Map Weighting Algorithm. 40

Figure 3.8 Top-down normalization with artificial images (a)original images (b) images normalized by Itti-Koch normalization operator N (c) images normalized by using Top-down normalization operator N_{TD} 51

Figure 3.9 Salient region extraction (The Figure is taken from [9]). 57

Figure 3.10 An image and its salient regions. (a) original image (b) saliency map (c) salient regions extracted from saliency map (d) the winner map for the first fixation point (e) region which is extracted by using winner map is shown. 58

Figure 4.1 Sample images from 'Mountain' category of Corel Images (Wang Database [10]). 66

Figure 4.2 Examples of tree types of images in the STIM dataset [11]. (a) Warning triangles, (b) road sings, (c) soda cans. 68

Figure 4.3 Sample Images from SIVAL dataset [12]. 70

Figure 4.4 (a) A sample airport image (b) corresponding mask image 71

Figure 4.5 Setup for experiments in STIM dataset 72

Figure 4.6 The salient regions extracted from saliency map of the image repeat after 5 salient regions are extracted 74

Figure 4.7 (a) original images, (b) Saliency Map from Itti.Koch Model (c) Saliency map from Bottom-up Normalization. 76

Figure 4.8 Saliency maps of Dataset Images according to the Query Image . . 77

Figure 4.9 Recall-Precision Curve for Experiment Set_{SM}. Results for category Autobahn. 79

Figure 4.10 : Recall-Precision Curve for Experiment Set_{SM}. Results for category Coke can. 80

Figure 4.11 : Recall-Precision Curve for Experiment Set_{SM}.Results for category Triangle. 81

Figure 4.12 The retrieval results of Experiment Set_{SM} for a query image from the Autobahn category (a) Itti-Koch (b) Bottom-Up Normalization (c) Top-down Normalization (d) Top-down Feature Weighting. 82

Figure 4.13 The retrieval results of Experiment Set_{SM} for a query image from the Coke Can category (a) Itti-Koch (b) Bottom-Up Normalization (c) Top-down Normalization (d) Top-down Feature Weighting. 83

Figure 4.14 The retrieval results of Experiment Set_{SM} for a query image from the Triangle category (a) Itti-Koch (b) Bottom-Up Normalization (c) Top-down Normalization (d) Top-down Feature Weighting. 84

Figure 4.15 Salient regions ordered between the most salient and least salient regions 91

Figure 4.16 Precision-recall curve of feature extraction methods employed to ABIR system for Autobahn category 94

Figure 4.17 Precision-recall curve of feature extraction methods employed to ABIR system for Coke can category 95

Figure 4.18 Precision-recall curve of feature extraction methods employed to ABIR system for Triangle category 96

Figure 4.19 Trace of Significance matrix 101

Figure 4.20 (a) Original images, (b) Saliency Map from the Itti-Koch Model (c) SIFT map. 109

Figure 4.21 Setup for composing an image database from single High resolution sattellite Image. 111

Figure 4.22 Shows the downsampled images with extracted "salient areas". From each region, image patches are extracted from original size of image. 114

Figure 4.23 An image patch and the corresponding query image 115

CHAPTER 1

INTRODUCTION

The amount of digital image has increased with the development of inexpensive hardware and software technologies for acquiring, storing and publishing images over the last decades. One of the most important challenges is to get a desired image in a huge image repository. Content-based image retrieval systems are developed to solve this problem. The aim of Content-based Image Retrieval (CBIR) is to search and browse images in a given database. The input of a CBIR system is mainly a query image and the output is a set of images that is most relevant to the query image in the database.

Retrieving a desired image from a database can be formalized as a one-class object recognition problem. In order to accomplish this task, initially, the desired object (query image or region of interest) is represented by an object recognition system. Then, the images in the database are tested on represented system to find if they are in the class of desired image or not. On the other hand, automatic image annotation systems present an alternative to find a desired set of images with their corresponding context or label. The aim of automatic image annotation systems is to assign label(s) to each image in the database. The image retrieval problem is formulated as a text retrieval problem in this case, where similar visual words and textual words are matched. However, while the object-recognition and automatic image annotation are challenging problems which are hard to solve, content-based image retrieval is a practical way of ranking images in a database.

The design of the content-based image retrieval system is achieved by the mathematical tools developed in the Computer vision literature. The methods developed for object recognition, object detection, object localization, automatic image annota-

tion, scene classification can be adopted to the content-based image retrieval systems depending on the application area of the image database.

A typical CBIR system searches and browse the images in a database by matching low-level visual features. The most relevant images to the query image are retrieved according to a similarity measure calculated between the features of the query and the target images. The features are selected heuristically depending on the application domain.

Visual features used in CBIR systems are categorized into two main types: global (holistic) features and local features. Global features are extracted from the whole image, whereas the local features are extracted from image regions. Global features are successful if the background of an image is also important for classification. However, in most cases, the background is irrelevant and the aim is to find objects in the foreground. For this reason, the local features are either used alone or in conjunction with the global features in the computer vision task.

Local descriptors such as SIFT [13], SURF [14], PCA-SIFT [15] are features that are extracted without the need to segment the image. They search for invariant points in the images at different scales. Local descriptors are successfully applied to various computer vision problems, such as object detection, recognition and image retrieval, when there is an invariant set of points for underlying object sets [16], [17]. Low-level visual features are commonly used in CBIR to retrieve images which contain high-level information. Using low-level features lead to the well-known semantic gap problem [18]. Relevance feedback, together with a variety of learning methods is used to partially improve the performance of CBIR systems. The Object-Based Image Retrieval (OBIR) approaches have been proposed recently to close the semantic gap between the low level and object-level features [18], [19]. OBIR systems usually rely on image segmentation to extract object in images. It is well known that this is an ill-posed problem in computer vision literature. Segmentation algorithms are expected to partition the images into segments or regions corresponding to the objects. The goal of segmentation is to separate objects from background of the given image according to the application. Segmentation is a difficult task, specifically when the objects lie on the cluttered background. Segmentation is an intermediate step between

the low-level features and high-level object concepts. However, it is difficult to find a generic solution to the image segmentation problem which is valid for a wide range of computer vision applications.

1.1 Motivation

An approach to the computer vision problems is to simulate the abilities of human vision for perceiving and understanding an image. In this thesis, we propose a CBIR system based on human visual attention behavior.

Human visual attention enables humans to perceive and understand the outside world. Visual attention is one of the fastest growing research areas in cognitive psychology [20], [21]. The main source of information for the human brain is the sensory inputs and it is claimed that 80 percent of this input is the visual data coming from the eyes [22]. Attention is the cognitive process of selectively focusing on only one part of the environment while ignoring the other parts. If we are able to attend to everything around us, we would be constantly distracted and unable to carry out any meaningfully action at all. In this sense, it is a useful adaptation that human process only a small portion of our surroundings at a given moment, and that only a limited range of objects can be attended to and acted upon at any one time. As James [20] pointed out, the concentration and focus are the essence of attention.

Human visual system has the ability to select relevant visual information based on saliency in the image using selective visual attention [5]. There are major mechanisms for human visual attention, namely bottom-up attention and top-down attention. Bottom-up attention denotes the process of attending by means of measuring the saliency of the items in a context. To give an example, suppose that there are considerably many red apples in a scene and only one green apple among the red ones. We attend the green apple in the scene due to the difference in color even if we do not have any purpose to find anything. Bottom-up attention which is defined also as stimulus driven process is a rapid and involuntary process of human vision system [23]. The second way of attention is the top-down attention, known as goal-driven attention. It assumes that prior knowledge of the content is known [24]. For exam-

ple, we attend to the faces in a crowded place if we are looking for somebody even in cluttered background. The available experimental studies suggest that top-down visual attention and bottom-up visual attention are used together in our visual system to direct our focus of attention in the environment.

Human visual system cannot process all image plane in parallel mostly. Therefore, the visual attention provides an important efficiency to compact the outside stimuli. Visual attention is responsible for deciding which part of visual information is to be fully processed. Therefore, computational visual attention systems, which aim to simulate human visual attention, are developed to select "relevant" or "salient" regions to reduce the irrelevant regions in the images. Salient region is the part of the images which is visually different than its surrounding regions. Hence, salient regions are extracted by using bottom-up models. Top-down information is needed to extract "relevant" regions by using Top-down visual attention models. However, how to model top-down information is a challenging problem, whereas top-down information is influenced by a set of factors such as task, memory and past experience. Therefore, bottom-up visual attention models are more common in the literature [5]. One of the currently best-known bottom-up visual attention system is suggested by Itti and Koch [11]. We call the system as Itti-Koch model in this thesis. The Itti-Koch model is a cognitive model which is designed for extracting salient points in still images.

Our motivation, in this thesis, is to simulate human visual attention and employ it in image retrieval problem. In a classic content-based image retrieval system, the goal is to search and browse an image database to show the images which are similar to the query image.

The system that we propose adopts the Itti-Koch model to excavate the important regions for searching and retrieving similar images similar to the human visual perception. Using visual attention in the context of a CBIR problem provides a number of advantages such as extracting salient points from the images. Salient points can be considered as a pre-processing step which is responsible for deciding on the "critical" visual information to be further processed. In the context of CBIR, this mechanism can be used to extract region of interests in an image to select the relevant parts in

the query image. The recent research in cognitive science [10] reveals that this is the mechanism people actually use, when looking for images. Therefore, regions which are extracted by using visual attention represents the location of objects in the images. Cognitive studies show that humans attend objects before recognizing them [25] by using visual attention. In classic CBIR system, the goal is to find similar objects considering the query image. Hence, Itti-Koch model provides an important tool to find the possible locations of objects in the images.

1.2 The Major Contribution of This Thesis

In this thesis, we propose a new content-based image retrieval system based on visual attention, called Attention-Based Image Retrieval(ABIR). The proposed ABIR system handles CBIR problem from the perspective of human perception: we suggest an efficient visual attention model specific to CBIR problem derived from the Itti-Koch model. Although, the visual attention model suggested by Itti and Koch or other computational models are used in CBIR problems in literature [7], [26]. Our approach is the first study which suggest a visual attention system specific to CBIR problem. In this thesis, we define the CBIR problem as an attention task, where query and images in the dataset considered together to extract the region of interests.

Itti-Koch model computes saliency of a single image by using a bottom-up approach. In our model, we compute the saliency of images according to the query image and all the images in the dataset. For this purpose, we propose three new algorithms: Bottom-up Normalization Algorithm, Top-Down Normalization Algorithm and Top-Down Feature Map Weighting Algorithm.

Normalization process in visual attention is a crucial task to inhibit unimportant regions in human visual system. The normalization algorithm of the Itti-Koch model is biologically inspired, it coarsely replicates cortical lateral inhibition mechanisms, in which neighboring similar features inhibit each other via specific, anatomically-defined connections. Let us explain this with a ball example. Consider a stimulus image which contains only one green ball and a number of red balls. Red and green are both visually salient features, however we can inhibit the red balls and attend the

green ball. This task is achieved in our brain by a set of normalization processes. In this study, we employ the Itti-Koch model for both bottom-up and top-down normalization. In the bottom-up normalization algorithm, we consider all images in a dataset as a single image and normalize the entire dataset under a unique scheme. The parameters of the normalization operator is computed from all images in the dataset as described in Chapter 3. Additionally, we use the query image to compute the parameters of normalization for top-down visual attention.

We propose a top-down visual attention system specific to CBIR problem in Top-Down Feature Map Weighting algorithm. The saliency of images in the database are computed by using the dominant features of query image as described in Chapter 3.

We propose two algorithms for feature space design named as named as Saliency-based Feature Integration and Saliency-based Feature Selection algorithms by considering the saliency of the images. First algorithm proposes feature weighting algorithm to integrate different features according to the saliency of the region. Second algorithm proposes a way of selecting representative feature among a set of features by using saliency concept. Finally, we propose a significant criteria based on saliency of regions to match regions of query and regions of dataset images to compute similarity between them by using Integrated Region Matching [27]. The criteria are computed by using saliency of regions.

1.3 Outline of the Thesis

This thesis deals with developing a content-based retrieval system, which uses visual attention and introduces Attention-Based Image Retrieval(ABIR) System. The thesis is organized as follows: In order to establish the necessary background, Chapter 2 gives a literature survey for visual attention and content based image retrieval. The human visual attention system and related theories are introduced briefly. The most popular computational model of visual attention models studied in the literature during the last decades have been discussed to make clear of our motivation and proposed approach. Moreover, CBIR system are summarized in the context of attention based methods .

In Chapter 3, we introduce a new image retrieval system that we called " Attention-Based Image Retrieval (ABIR). The details of the our system is explained by introducing the new proposed algorithms for modeling visual attention of CBIR problem.

In Chapter 4, the experimental results on different databases are given. We compare our ABIR system with localized CBIR systems; ACCIO!, SBN, Simplicity and GMIL, in the literature. We discussed the experimental results, emphasizing the superiorities and weaknesses of the suggested ABIR system.

Chapter 5 concludes the thesis and gives the directions for future work.

CHAPTER 2

BACKGROUND FOR VISUAL ATTENTION AND CONTENT-BASED IMAGE RETRIEVAL SYSTEMS

2.1 Introduction

Visual attention is a highly interdisciplinary research topic at the intersection of cognitive Neuroscience, Psychology and Computer Vision. Figure-2.1 summarizes visual attention studies in the literature. Neuroscientists use brain imaging techniques such as PET and fMRI to understand the mechanism of visual attention in humans. The goal of psychological visual attention studies is to explain human perception and cognition based on experiments on human subjects. However, biologically inspired models try to develop computational visual attention models which are inspired from studies of Neuroscience and Psychology for solving computational vision problems such as object recognition, robot navigation or content-based image retrieval [5]. Furthermore, non-biological models aspire to reduce to irrelevant regions or detect interesting regions in the images based on the application domain by using visual attention indirectly. Non-biological models are computer vision methods or algorithms rather than a model which simulates human visual attention. For this reason, the non-biological models with application of CBIR systems are explained.

A literature survey on content-based image retrieval based on the biologically inspired methods is presented. A general description of human visual system and cognitive theories are given in the following section. The next section includes the details of the computational models of visual attention and briefly explains the Itti-Koch model. The last section gives the literature about content-based image retrieval and focuses

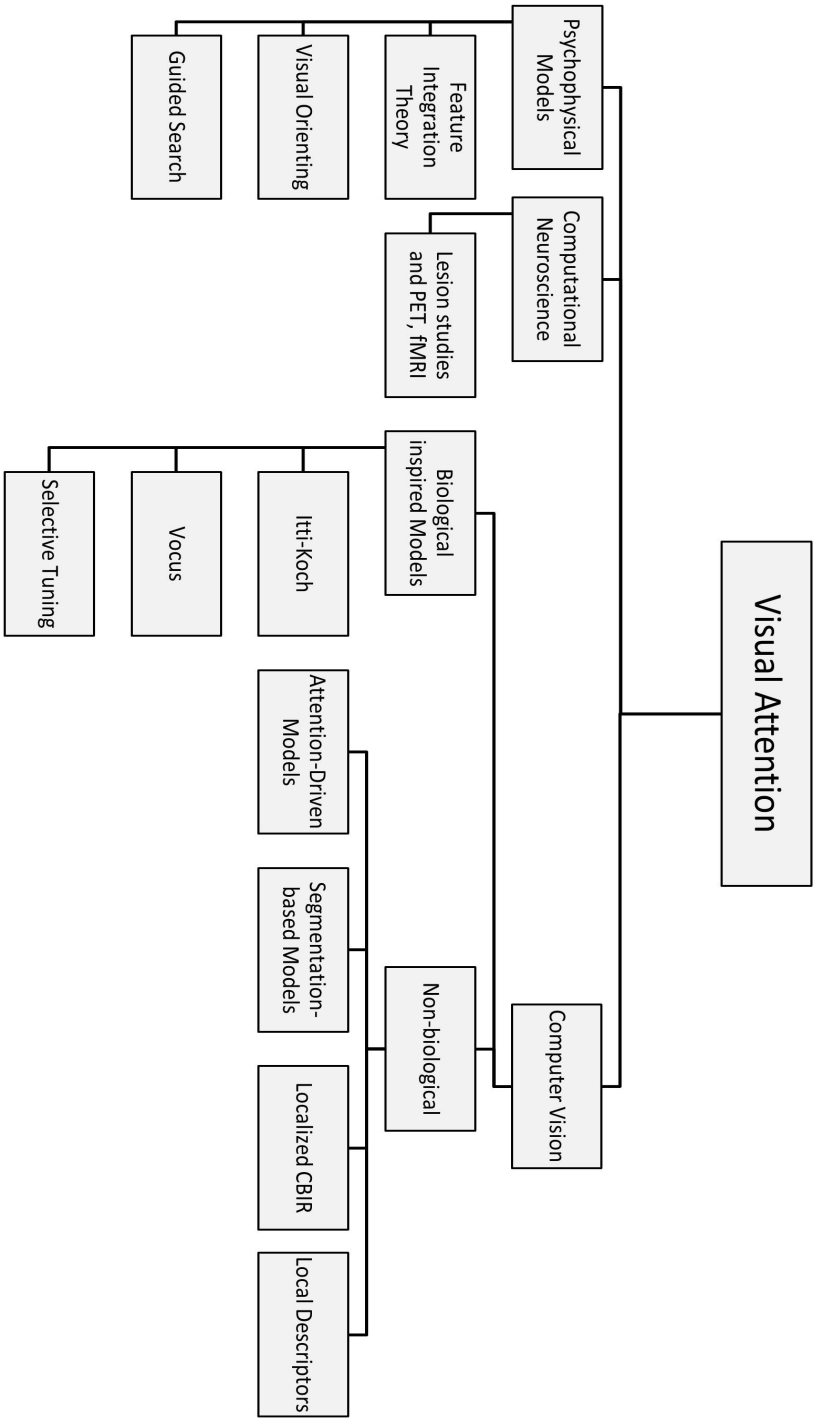


Figure 2.1: A general overview of the Visual Attention Methodologies by different disciplines.

on the attention-based methods.

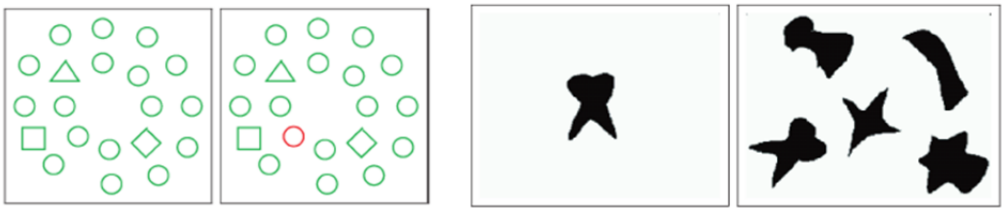
2.2 Human Visual Attention

How we see is one of the most anticipated questions since the early ages of science and many questions come along with these questions: "How do we see and feel?", "How do we see and perceive", "What do we see?" and so on. Different disciplines, such as Philosophy, Psychology and Neuroscience try to find an answer to these questions. In spite of a great effort, there exists no satisfactory answers to most of the questions. However, the existing inspires the computer vision research.

One of the known facts about the human visual system is that we select and focus on a region of interest in the scene and then process it in an efficient way. This cognitive process is called *visual attention*, and widely used in our brain for selectively concentrating on one aspect of the environment while ignoring other things. In this sense, it is a useful adaptation that we are aware of only a small portion of our surroundings at a given moment, and that only a limited range of objects can be attended to and acted upon at any one time. As James [20] pointed out, the concentration and focus are the essence of attention. He says: "Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Localization, concentration, of consciousness is of its essence. It implies withdrawal from some things in order to deal effectively with others." [20] pp. 403-4.

Eye movements are usually associated with visual attention for directing the focus to a region of interest which is called *overt attention*. However, human are able to attend to peripheral regions of interest without moving eyes which is called *covert attention*. *Overt attention* and *covert attention* usually works together: the region of interest is selected by covert attention which is followed by overt attention to fixate the region and perceive it at a higher resolution [28].

Visual attention is processed by the brain in two ways; namely Bottom-up and Top-down. Bottom-up attention is fast and involuntary process that is triggered by a stimulus presented such as a fast movement, bright color, or shiny surface [29]. Features of a scene that influence the direction of our bottom-up visual attention are the first to be considered by the brain and include color, movement, and orientation, among

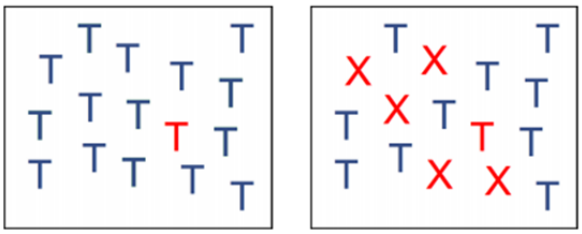


(a) Attentional capture (b) Cueing experiment

Figure 2.2: (a) Attentional capture (b) Cueing Experiment. (Figures are taken from [5]).

others [11]. For instance a flashing light immediately draws attention. On the other hand, top-down visual attention requires a priori knowledge coming from processing what we have learned and can recall. Top-down visual attention is initiated by memories and past experience [29]. For example, when we type a document, looking for a specific letter on a keyboard requires priori information about that letter and "all" the other letters. Both bottom-up and top-down visual attention decide on where we focus of interest on the scene. However, how they interact with each other is still unclear. Moreover, it has been argued that if a bottom up factor is too prominent it will suppress any top down effect on the perception. This effect is called *attentional capture* [5]. For example, the images in Figure-2.2(a) are presented to the human subjects [1]. The task is to find the diamond in both images. However, the red circle in the right image slows down the search about 65ms. This experiments shows that, bottom-up factors, which is color in this experiment, captures the focus of attention independent from the task.

Visual search experiments on human subjects are used in Psychology to understand the mechanism of visual attention [5]. The general structure of visual search experiments is as follows: given a target and a test image, the goal is to find the target in the test image. The time needed to find the target in given images are measured by the Reaction Time (RT). The RT is computed as a function of set size which is the number of elements (distractors) in the test image. One such example is the cueing experiments shown in Figure-2.2(b) where the subject is asked to find an object in the test image. Resulting measurements show that if the localization of the object in the target image is close to the localization of the object in the test image, experiment



(a) Feature Search (b) Conjunction Search

Figure 2.3: (a) Feature Search (b) Conjunction Search. (Figures are taken from [5])

is more likely to be successful i.e the target object is easier to find regardless of the orientation. This observation is called as Posner Cueing paradigm [30].

Another experiment conducted on visual search on human subjects is called feature search and conjunction search. The target in the feature search is distinguishable from the test elements (distractors) with unique features such as color, size, orientation or shape. Figure-2.3(a) gives an example of feature search that the target is represented as red T and all other distractors are represented as blue T. The target is quickly detected independent from the number of distractors due to difference in color feature [2]. The experiments for conjunction search aims to find a target which can be represented with more than two features according to the distractors, as shown in Figure-2.3(b). The target is red T that is distinguishable from the distractors by two features which are color (red) and shape (T). The reaction time is longer than the feature search experiment, because we are looking for conjunction of two different features. The reaction times increases steeply by increasing the number of distractors [2].

2.2.1 Theories for Human Visual Attention in Cognitive Psychology

There are many theories and hypotheses for describing human visual attention in cognitive psychology literature [5], [31]. Here, we will mention the theories which inspire computational models of visual attention. These are feature integration theory, guided search theory and visual orienting theory.

Feature Integration Theory (FIT) proposed by Treisman [3] suggests that the low

level features such as edge orientation, color difference, corners etc. are extracted in a parallel fashion at the low-level vision show in Figure-2.4. The theory assumes that " Features are registered early, automatically and in parallel across the visual field, while objects are identified separately and only at a later stage, which requires focused attention" [3] pp. 98. The first stage of the feature integration theory is the pre-attentive stage where the object features are analyzed automatically or unconsciously in the different part of brain area. The next stage is the focused attention stage. During this stage, all feature maps (Color maps, orientation maps) are collected in a master map of location which specifies where the attended regions are but no information about the region. The other regions are attended by scanning the master map of location serially. According to the theory, attention is required in order to perceive the whole object from the individuals features of the object. The object features are stored in object files which keeps the temporary object representation. In order to recognize the object, the object file associate with recognition network shown in Figure-2.4.

The Guided Search Theory proposed by Wolfe [4], [32], [33] takes into account both the top-down and bottom-up visual attention while FIT only considers bottom-up visual attention. The guided search theory is used to rank the regions in the stimulus image using feature maps and top-down commands as seen from the Figure-2.5. This theory is the most effective work for computational visual attention systems in the literature [5], [8], [21], [23].

According to the Guided Search Theory, the stimulus is filtered through the feature channel such as color, orientation. The theory expressed that the attention is focused onto the desired object in the scene in terms of adjusting the feature map weights. For this purpose, the bottom-up visual information and top-down information are combined to generate activation map that determines the attentional priority during the visual search. For instance, let us consider that the subject is searching black vertical lines in the scene. All black and vertical objects in the scene will firstly gain priority because of the top-down activation. In other words, top-down activation guide the feature map to activate the desired features. On the other hand, bottom-up attention activation indicates how the objects are different from the neighboring objects in the stimulus image.

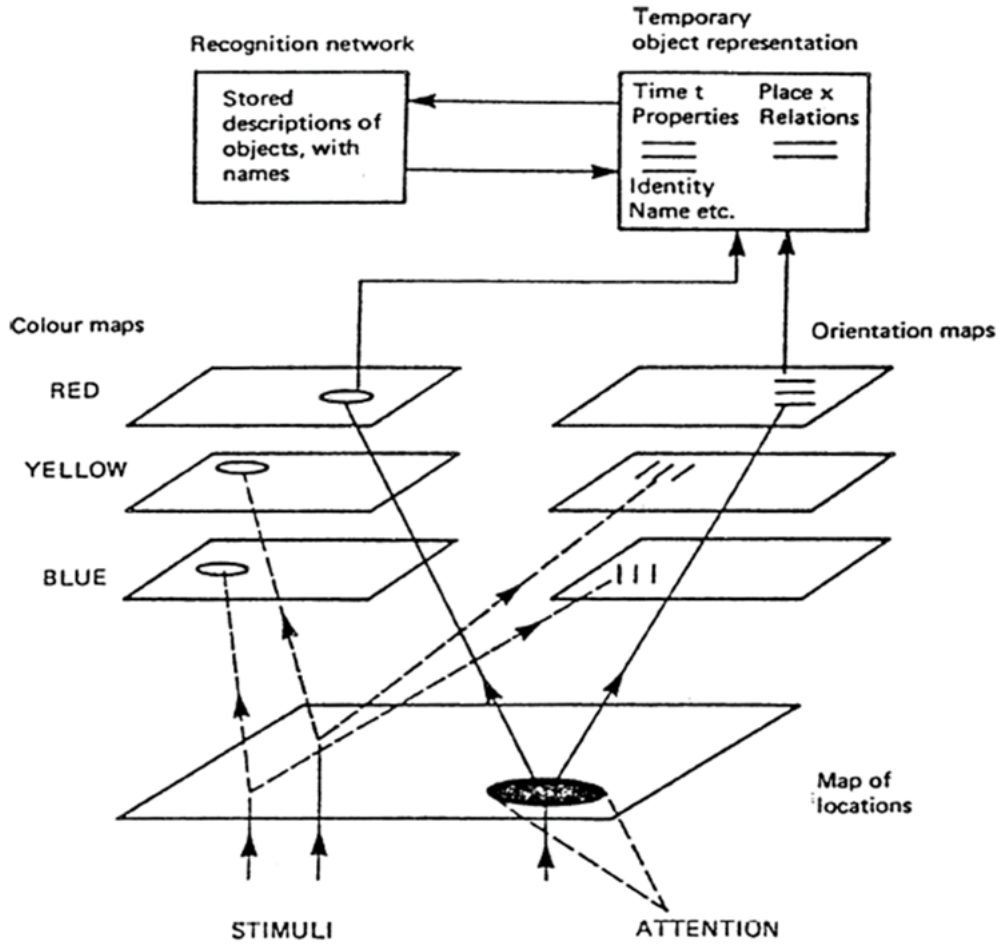


Figure 2.4: Feature Integration Theory. (The figure is taken from [3])

The previous two theories do not give an explanation how more than one objects can be attended in a scene which is defined as attentional shift [34]. The Visual Orienting Theory proposed by Posner [30] suggests the Inhibition of Return(IOR) approach that refers to an orientation mechanism of human visual attention system. The orientation term here is used to define attentional shift. IOR is the inability to return to a previously attended location unless a certain time period has passed. In order to explain more clearly, the reaction time (RT) needed to detect objects appearing in previously attended locations are longer when compared to locations which is not previously attended. The increase in RT is explained by suppressing the previously attended location in the scene. This mechanism allows humans to direct the attention in other regions in the scene.

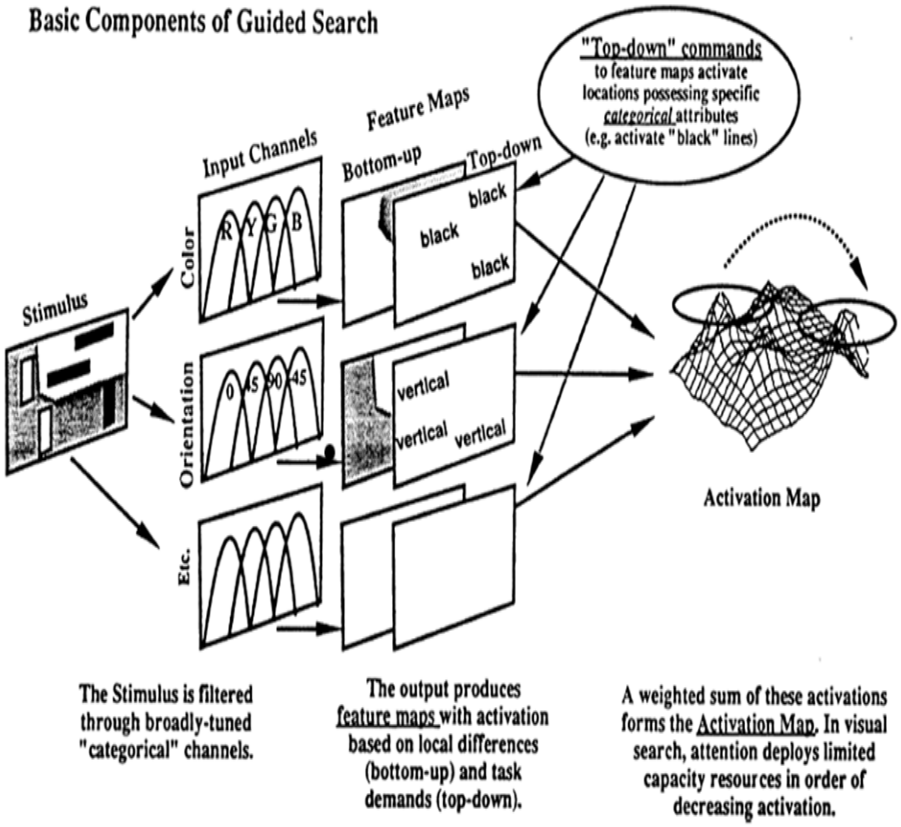


Figure 2.5: Guided Search Theory (The figure is taken from [4]).

2.2.2 Visual Attention Models in Neuroscience

Neuroscience models try to understand visual attention through lesion studies or brain imaging techniques, such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET). Brain studies reveal that multiple brain areas guide the attention during the visual processing. Important areas of brain associated with visual attention are Posterior Parietal cortex (PP), the Superior Colliculus (SC),the lateral intraparietal area (LIP), the Frontal Eye Field (FEF) and the pulvinar [35]. One of the findings on brain studies shows that attending to a salient location is carried out by interaction of brain areas PP, SC and pulvinar [36]. Brain studies throw light on human perception; however, the best know finding about human brain is that there is no single area which is responsible for visual attention. The function and interaction

of brain areas are still an open question.

2.3 Biologically Inspired Computational Visual Attention Models

Computational visual attention models are mostly developed to solve a computer vision problems rather than trying to understand human vision system. However, the computational attention systems used in computer vision and robotic applications are mostly inspired from psychological and neurobiological outcomes. Computational visual attention is an important tool for computer vision applications to select relevant and salient regions. Moreover, this mechanism is important for object recognition and localized content-based image retrieval problems to discard background information from the scenes.

Most biologically inspired computational visual attention systems have the similar structure as shown in Figure-2.6. This structure is also very similar to psychological theories such as the Feature Integration theory [3] and Guided Search model [32]. The main idea is to extract feature maps using several features in parallel and then integrate these maps to find a map which is usually called saliency map. There are mainly four steps in computational visual attention models: feature extraction, saliency map computing, finding the most salient location in the saliency map and finally finding focus of attention points in a given image scene.

In the first step, features are extracted in different scales as shown in Figure-2.6. Feature maps are computed by considering the intensity, color and orientation features [8]. Psychological and biological works [33], [37] proposed that these are basic features of human visual system which are also easy to compute. The feature maps collect the local within-map contrast. This is usually computed by center-surround mechanisms, also called center-surround differences [38], [39]. Center-surround operation compares the average value of a center region to the average value of a surrounding region, inspired from the ganglion cells in the visual receptive fields of the human visual system [40].

The saliency map is computed by taking summation of all feature maps as second step. Before the feature maps are summed up, they are usually normalized. Nor-

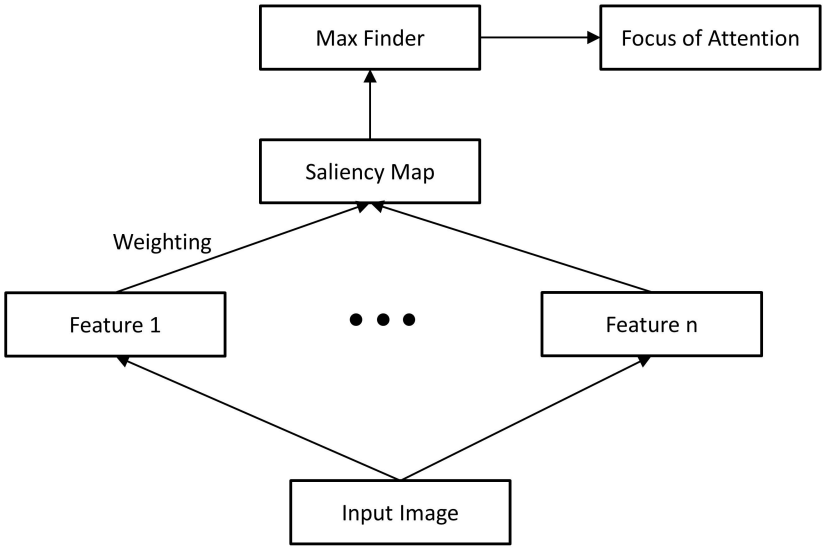


Figure 2.6: A general scheme for extraction of Visual Attention points by different disciplines (The figure is adapted from [5]).

malization is used for giving more emphasize to the feature maps which have small number of strong peaks, while suppressing feature maps which contain numerous comparable peaks [11].

The computed saliency map might already be regarded as an output of any computer vision system since it shows the saliency for each region of a scene. Hence, the next step in a computational model is to find locations of salient regions by using the saliency map of the input image. The salient image regions are local maxima in the saliency map. They might be determined by a winner-take-all (WTA) network which was introduced by Koch and Ullman [41]. It shows how the selection of a maximum is implementable by neural networks that means by single units which are only locally connected. This approach is strongly biologically motivated and shows how such a mechanism might work in the human brain. A simpler and relatively more technically motivated alternative to the WTA with the same result is to straightforwardly determine the pixel with the largest intensity value in the saliency map of the input image. This method requires fewer operations to compute the most salient region, but note that the WTA might be a good solution if implemented on a parallel architecture like a GPU.

Recall that the focus of attention (FOA) is usually not on a single point but defined by a region which is called MSR (Most Salient Region). Let us now discuss about the identification of this region. The simplest approach is to determine a fixed-sized circular region around the most salient point [11]. More sophisticated approaches integrate image segmentation on feature maps [9] or saliency maps [42] to determine a irregularly shaped attention region.

The most common method for obtaining a set of image regions which mimics the human eye fixation search trajectory is the Inhibition of Return (IOR). It refers to the observation that in human vision, the speed and accuracy with which a target is detected is impaired after the target was attended. It was first described in [43] and prevents that the FOA stays at the most salient region. In computational systems, IOR is implemented by inhibiting (reseting) the surrounding region in the saliency map. The surrounding region can be a fixed region around the FOA (spatial inhibition) or the MSR (feature-based inhibition), or a combination as in [44]. However, human subjects can re-attend the inhibited regions after a time period [45]. A possible implementation of IOR in computational models inhibits each MSR for a short time, dependent on a probabilistic value. An alternative way to IOR to obtain a set of image regions, which is simple to implement and obtains good results, is to determine all peaks in the saliency map, sort them by their saliency values, and direct the FOA attention subsequently to each salient region [46].

In the next sub-sections, we briefly introduce the Itti-Koch model and other computational visual attention models.

2.3.1 The Itti-Koch Model

The Itti-Koch model is based on feature integration theory proposed by Theirsman [3] that explains human visual search strategies. It is a biologically-inspired model based on a study of actual neurological processes in primates. In this model, an input image is decomposed into a set of multi-scale feature maps which extract local spatial discontinuities in the modalities of color, intensity and orientation. Each feature map is endowed with non-linear spatially competitive dynamics, so that the response of a neuron at a given location in a map is modulated by the activity in neighboring

neurons. Such contextual modulation also inspired from recent neurobiological findings [3], has proven remarkably efficient at extracting salient targets from cluttered backgrounds. All feature maps are then combined into a unique scalar saliency map, which encodes for the saliency of a location in the scene irrespectively of the particular feature which detected this location as conspicuous. The most salient region is at the maximum point of saliency map and focus of attention is directed to this point. It is used to find the location of focus of attention. The other regions are examined by inhibition of return of former attended region.

2.3.2 Otner Computational Models

The first computational visual attention model is proposed by Koch and Ullman [41] which Itti-Koch model [8], [11], [47], [48] is derived from. The "Saliency Map" concept is introduced firstly in this model. Also winner-take-all network (WTA) and inhibition of return processes are implemented with this model. However, Itti-Koch proposes a more efficiend method for computing saliency map by using multi-scale analysis of input image.

Because of the top-down mechanism of visual attention in humans is not resolved completely, computational models for top-down visual attention is less studied in computer vision literature [5]. In [49], top-down cues are added to Itti-Koch model. The VOCUS (Visual Object detection with a Computational attention System) [6] is an another top-down visual attention model which is also based on Itti-Koch model. In this model a top-down saliency map is computed by using target information and bottom-up saliency map and top-down saliency map are combined into one saliency map. In Figure-2.7 shows the basic structure of VOCUS.

Another well known computational model is Selective Tuning Model of visual attention [50]. The Selective Tuning model analyzes four kinds of visual features to identify the focus of attention in an input image: luminance, orientation, color, and motion. The model performs a pyramid style processing of information where the stimuli of interest are located at the top of pyramid and control an inhibitory beam. This inhibitory beam can inhibit or pass a zone for further processing. The top-down influence is modeled through manipulating the inhibitory beam.

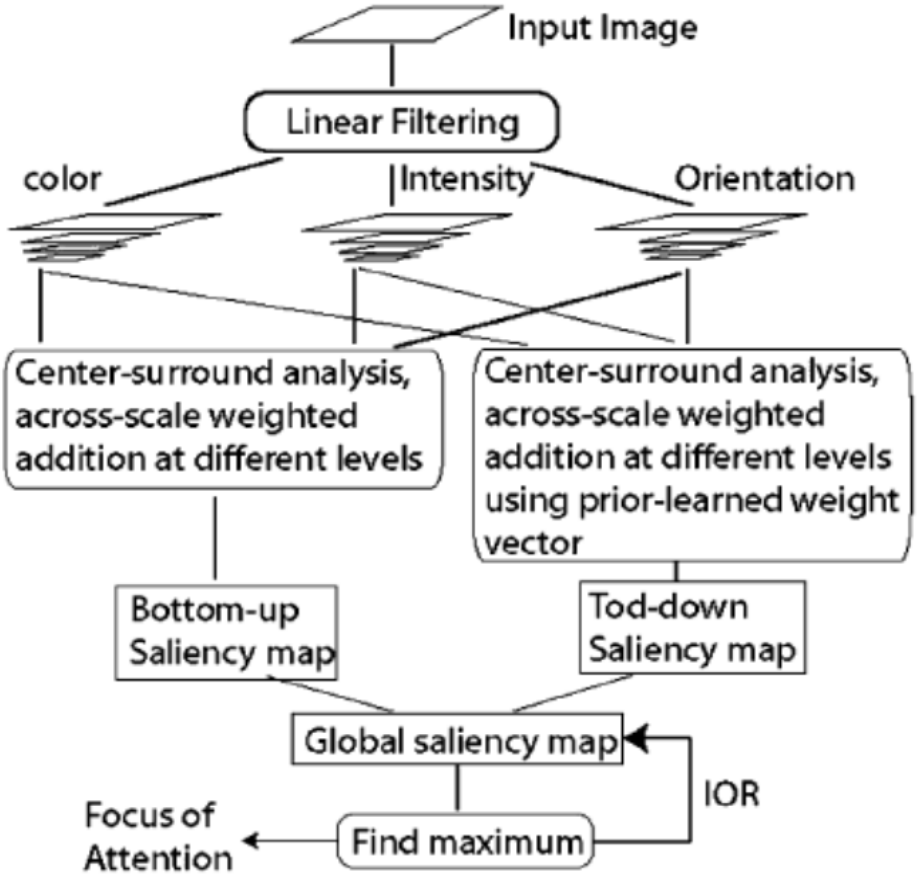


Figure 2.7: The structure of VOCUS (The figure is taken from [6]).

There are other computational models of visual attention [16], [51], [52] but the models discussed in this section are important and used mostly in computer vision applications. A more detailed literature about computational models of visual attention can be found [5].

2.4 Content-Based Image Retrieval

Almost 10-15 years ago, people had used photo albums to store the photos taken in special moments of their life and annotated them manually not to forget where and when they took them. However, the digital equipment now becomes a part of our life with the rapid development of technology. We take photos of every moment even necessary or not, and share them on social networks. Therefore, we face with a big

digital photo thrash. It is sometimes impossible to find a picture that we desired in a tremendous number of pictures. Image retrieval systems are developed to search and browse image databases for different applications such as digital photo album, surveillance systems and web search . The first retrieval systems used manual annotation to perform image retrieval. They needed database management systems [19]. Manual annotation is the best solution for content-based image retrieval problem but it has two disadvantages. The first is that manual annotation is too time consuming and labor intensive. The second is annotated words are too far from being objective [53]. For that reason, content-based image retrieval systems are developed to close this gap [54], [55].

The term content-based image retrieval first introduced in [56]. A typical content-based image retrieval system has three major aspects: feature extraction and similarity matching, high dimensional indexing, and system design [18]. Feature extraction is important for accuracy performance. The contents of images are represented by features, so extracting discriminative features are the main problem in content-based image retrieval systems. High dimensional indexing is important for speed performance of the system, when the growth of digital databases is considered. The last step of content-based image retrieval is system design which is also shows the usage performance for real life applications.

The features used in CBIR are categorized into two classes: global features and local features [53]. The basic global features, which have been studied at the time of the beginning of the image retrieval systems, are extracted from the whole image such as color, texture, and shape [57], [58], [59]. As they are simple and try to capture all information in the image, global features are widely studied and used in CBIR systems [18], [19], [60]. Color histogram, color layout, edge histogram, Gabor, wavelet, etc. are mostly used global features used in CBIR systems [18], [19], [51], [61] . The combination of various features is also utilized in CBIR systems. In [62], the combination of color and wavelet based texture features are used efficiently. A set of color and texture descriptors of Mpeg-7 standard is used in [63] and well suited to natural images.

Global features give successful results when the user considers the entire image such

Table 2.1: Learning-based methods for CBIR systems (The table is adapted from [19]).

Method	**Purpose**	**Techniques**	**User Involvement**
Clustering	Faster Retrieval, Efficient Storage	K-means, Hierarchical, N-cut	minimal
Classification	Pre-Processing,Fast/Accurate Retrieval, Automatic Organization	SVM, Statistical Models, k-NN, Bayesian Classifiers	Provide Training Data (not interactive)
Relevance Feedback	Capture user and query specific semantics, refine rank accordingly	Feature re-weighting, region weihting, active learning, boosting	Significant, interactive

as natural scenes. However, users usually look for an object or for several objects in the query image. In this case, local features are used to eliminate background and irrelevant regions from images.

Local features are extracted from the sub images or regions. Sub images could be computed using sliding windows, or simply dividing the image to equal parts. More commonly, the regions are extracted by using segmentation methods. In segmentation-based systems, shape information is an important feature because of its efficient and robust representation power. Using shape descriptors, reliable segmentation is a critical issue. In the literature, segmentation algorithms based on Normalized Cuts criterion [64] have an important place as being successful in [63], [65], [66].

An important problem in CBIR systems is semantic gap. Similarity between two images is mostly high-level whereas low-level features represent the images in systems. Moreover, the similarity between same images can be different for different users. Learning and user feedback are adapted to CBIR systems to handle these problems. Learning-based methods, which used in CBIR systems, are summarized in Table-2.1 [19].

Using clustering techniques are mostly required when handling large, unstructured image databases such as web images. Clustering achieves faster retrieval and visualizes data in a meaningful way but low-level features are used in clustering. In [62] clustering is applied to the results of the system and retrieval accuracy is increased in this way. Drawbacks of clustering in retrieval systems are using low-level features and poor user adaptability. K-means clustering and Normalized cut are widely used methods in CBIR systems [19].

SVM, k-NN, Bayesian Classifiers are mostly used classification methods in CBIR systems[19]. The drawback of using classification methods in CBIR systems is providing training data. Relevance feedback is mostly used method in learning based CBIR systems. It modifies the queries according to the users ratings. A detailed analysis of relevance feedback systems is given in [67].

Automatic annotation is an important tool for CBIR systems as it integrates semantic information to the images. Automatic annotation is the process of producing words for images automatically. The relation between words and images depends on the purpose of the application or content of the image database. There are two different approaches for automatic annotation, Joint Word-Picture Modeling Approach and Supervised Categorization Approach [67]. In the first approach, a translation is developed from a set of image segments to a set of words [63]. In the later approach, firstly supervised image categorization is done and then related words are assigned to the image [10]. The drawback of annotations systems is lack of insufficient and consistent labels for training sets.

Region-based methods are used widely in the context of image retrieval. These methods mostly rely on image segmentation, which is an ill posed problem in computer vision. In work [68] the user indicates a region in query image and this region is used for retrieval work. It computes only region-to-region similarity. The drawback of region-based systems is how the user decides criterions when indicating regions. In Simplicity [60], the region sets of images are used to matching images by similarity metric called Integrated Region Matching (IRM) [27].

The local descriptors are very successful in content-based image retrieval applications [17] and [69]. They do not need segmentation, and are sensitive to occlusion and image transforms. SIFT [13], [70], SURF [14], PCA-SIFT [15], Kadir-Brady [71], Harris and GLOH [69] are most used local descriptors and evaluations of them are represented in [16] and [17].

The CBIR systems which aim to find the salient or relevant regions are explained in the next section under the Saliency-based Image Retrieval Systems. After that, localized CBIR systems are summarized. The goal of localized CBIR systems is to find object of interest in database images based on query image.

2.4.1 Saliency-based Image Retrieval

The image retrieval systems which are based on region extraction and region matching are widely known as object-based image retrieval and examples to region-based systems given in previous section. Saliency-based image retrieval can be considered as a sub-class of object-based image retrieval which are based on salient features or salient regions such as interest point detectors or systems use visual attention.

In literature, the best known interest point detector and feature is Scale Invariant Feature Transform (SIFT) [13, 70]. The SIFT features are computed in four steps: scale-space extreme detection, key-point localization, orientation assignment and extracting the key point features. For the first step, the image is filtered with Gaussian mask at different scales, and then the difference of Gaussian images is computed. The local extremes of Difference of Gaussian images are identified as key-points. Scale-space extreme detection finds too many key-point candidates, which are not reliable. The next step eliminates the unreliable key-points such as edge points and low-contrast key-points. The Taylor series of Difference of Gaussian images are used to discard low-contrast key-points. The Hessian matrix method is used to eliminate key-points on the edges. Alternatively, Harris corner detection is used to eliminate edge response points [72].

The simplest use of SIFT features for content-based image retrieval is matching the SIFT features of query image with those of dataset images [73]. In this case, the SIFT features of background could suppress the object of interest. Moreover, computation time for feature matching is not convenient for most applications.

Even though local descriptors are very successful in cluttered backgrounds and under different image transforms, they do not differentiate interest of object from background in common backgrounds. Localized content-based image retrieval systems [12] try to solve this problem by using Multiple Instance Learning [55], [74], [75] or semi-supervised learning [76], which need positive and negative query images or training data.

Another way to extract region of interest from images is using Visual Attention models which is based on Human Visual System and without doing any segmentation. The

advantage of using visual attention in Content-based image retrieval systems does not need any segmentation process. The salient regions are computed from the saliency map easily by using winner-take-all network and inhibition of return mechanism. In CBIR systems, mostly bottom-up approaches are used, no top-down information is used for retrieval performance.

In work [77], regions extracted by using Itti-Koch model and SIFT [9] features are used for image retrieval. The attention driven study utilize Itti-Koch model with Expectation Maximization segmentation for extracting regions [11]. Another study [78] proposed a computational model of visual attention by integrated regions extracted by employing classical image segmentation and regions extracted by using Itti-Koch model.

Stentiford [7] proposes a computational model of visual attention which is used in CBIR applications. It functions by suppressing areas of the image with patterns that are repeated elsewhere. As a result, flat surfaces and textures are suppressed while unique objects are given prominence. Figure-2.8 explains the algorithm of Stentiford model. Regions are marked as high interest if they possess features not frequently present elsewhere in the image. The output of the Stentiford Model is a visual attention map that is similar in function to the saliency map generated by Itti-Koch. The visual attention map generated by Stentiford tends to identify larger and smoother salient regions of an image, as opposed to the more focused peaks in Itti-Koch's saliency map.

In [79] and [26] region of interest are extracted by using two bottom-up visual attention models: Itti-Koch [11] and Stentiford [7]. The seed points of region growing are selected from Itti-Koch saliency map and area of regions are considered by using Stentiford' s saliency map. The resulted regions are clustered for content-based image retrieval. The system biologically inspired but not worked with cluttered background.

A similarity measure which is computed by using Stentiford visual attention model is used for content-based image retrieval in study [80]. This similarity measure can determine the property of region is interested or not.

A novel region-based image retrieval algorithm using selective visual attention model

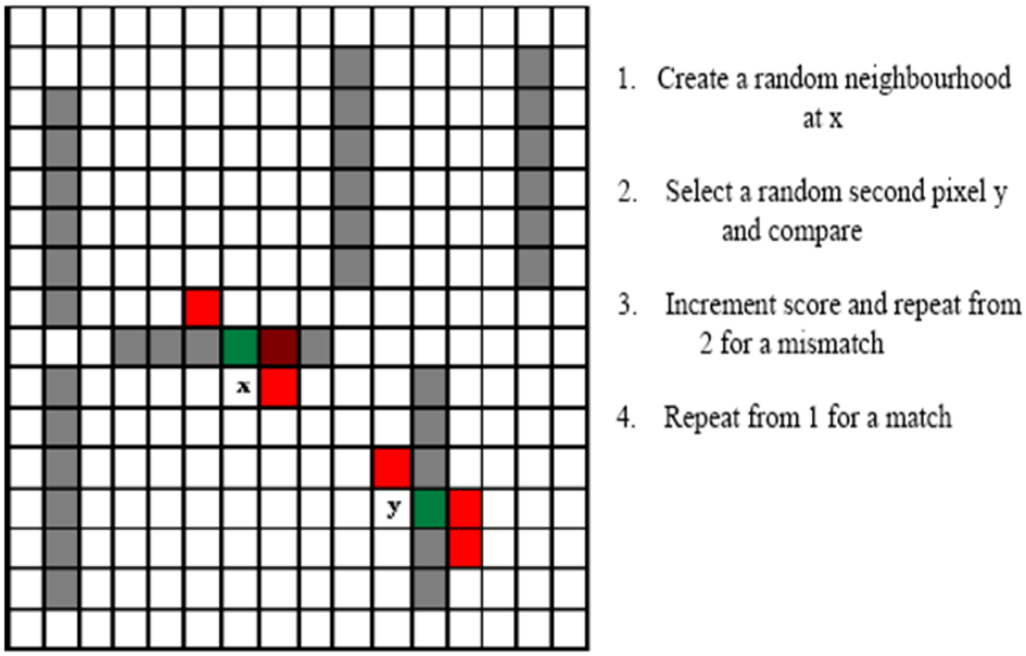


Figure 2.8: Visual attention model of Stentiford (The figure is taken from [7]).

is proposed in [81]. The algorithm computes three different saliencies for each pixel, each region and each scene. Finally, combines three saliency values and performs content-based image retrieval.

Bottom-up approaches give accurate results if the object of interest is conditioned with visual pop-out. Visual pop-out means that object has significantly different from its background. If the background is cluttered and object of interest has not dominant features, it is impossible to locate the object. In this case, top-down information is needed for correct localization of objects.

In literature, the top-down visual attention models are not used in content-based image retrieval problem. The aim of content-based image retrieval systems is to search and browse dataset images which are similar to query image, and mostly it is not known what the query image contains. So, it is hard to add top-down information to retrieval system. In some cases, the query image only contains object of interest. In this case, top-down information can be extracted from query image.

2.4.2 Localized Content Based Image Retrieval

Localized CBIR is an retrieval problem that defines as a CBIR task where the user is only interested in a part of the image. It assumes that the rest of the image is irrelevant. Localized CBIR must rely on positive and negative labeled images to learn which part of the image is interested. The query set which contains a set positive and negative images is provided by the user or obtained using relevance feedback.

In localized CBIR, the query set is used to determine the part(s) of the image that are relevant to the user's interests, as well to compute an appropriate weights for features. Most CBIR systems either subdivide the image into predefined blocks, or more commonly partition the image into different meaningful regions by applying a segmentation algorithm. In both cases, each region of the image is represented as a feature vector extracted from that region. Other CBIR systems extract salient points, which represents high variability in the features of the local neighborhood. One feature vector is created for each salient point for salient point-based methods.

The difference between the localized and region-based CBIR systems is how the image is processed by the system. Single feature vector CBIR systems computes one feature vector to represent the entire image. In contrast, multiple feature vector CBIR systems computes a set of feature vectors with one feature vector for each region to represent the image. Region are obtained from pre-defined blocks or segmentation methods.

Multiple instance learning algorithm (MIL) is used to develop object-based image retrieval systems.

The standard supervised learning algorithms uses training sets where all samples in the training set are definitely labeled. However, multiple instance learning algorithms uses labeled data in the bag level where each bag consists of multiple instances. Multiple instance learning method defines a bag as positive if it includes at least one positive instance, and negative if it includes all negative instances. The goal of a multiple instance learning method is to generate a classifier that will classify unseen bags correctly. A single query image in an object-based image retrieval system represents a positive bag. A query set with positive and negative instances represents

region of interest which the user search for in an image dataset. Semi-supervised learning methods are developed to solve unlabeled data problem for MIL methods [82]. However, most of semi-supervised learning algorithms are only developed for single-instance training sets and cannot be directly applied to multiple instance learning problem. Therefore, multiple-instance semi-supervised learning (MISSL) method has been developed by Rahmani et al.[83]. They first transform the MISSL setting into a bag-level graph, which is an input of the single-instance semi-supervised learning training set, and then directly solve the single-instance semi-supervised learning problem. They developed ACCIO! system by using MISSL method. The user choose a query set directly or by using relevance feedback in ACCIO' system. Then, system segments the images and converts the segmented image into the multiple instance using the bag generator. Finally, all images in the dataset are ranked by a ranking algorithm combines the hypotheses from the learning algorithm.

A graph-based multiple instance learning (GMIL) algorithms is proposed by [84] based on MISSL method. Another method called SBN (Single Blob with Neighbors) [85], multiple instance learning is applied to the problem of learning how to classify natural scenes. A important part of the system is the bag generator. The system which takes an image and generates set of instances, where each instance is a possible description of what the image is about. Diverse Density (DD) framework is used for bag generator. The main idea behind the DD is to find areas in feature space that are close to at least on instance from every positive bag and far from every negative instance.

2.5 Summary

In this chapter, we have given a background for content-based image retrieval problem by considering human vision attention. For this purpose, firstly, we discussed the fundamental concepts of human visual attention. Then, we explained the computational models of visual attention in literature briefly. Finally, we discussed the saliency and localized-based image retrieval systems after giving a general literature about CBIR. We also summarized the well-known localized content-based image retrieval methods in the literature, namely, ACCIO! [74], GMIL [84] and SBN [85].

CHAPTER 3

AN ATTENTION BASED IMAGE RETRIEVAL SYSTEM

In this chapter, we describe a content-based image retrieval system based on visual attention. The proposed system, called "Attention-Based Image Retrieval (ABIR)", is designed for general purpose image repositories. The system is a query by example CBIR system which has two input categories: query and dataset images. The output is a dataset of images ranked according to the similarity to the query image.

Visual attention is the ability of human visual system, which guides the selection of salient regions from visual data. This selection enables human to perceive the outside world by ordering regions according to the saliency and by ignoring the irrelevant data. Our perspective, here, is to design the content-based image retrieval problem by incorporating the very crucial cognitive process of visual attention. Our major assumption is that salient regions compose the object(s) or area(s) in the images that we are looking for. Visual attention is used for two major reasons in our proposed approach: Firstly, visual attention models enables us to focus on the salient regions in the images which we are looking for. Secondly, this approach avoids expressive and error-proof segmentation process. Specifically, we use visual attention to extract the salient regions from the images. However, the saliency in human visual system is a highly complicated process which is affected by bottom-up or top-down factors as described in the Introduction section. We design a new ABIR system based on the saliency of the images, represented by both top-down and bottom-up visual attention activities. Bottom-up visual attention is modeled by visual features. On the other hand, top-down factors is modeled by considering the images in the entire database. Therefore, the saliency of the query image affects the images in the database.

The visual attention model suggested in this study is based on Itti-Koch model. Itti-Koch model is a bottom-up approach which computes the saliency points of an image by emphasizing visually salient regions. The Itti-Koch model deals with only one image and the output of the model is the saliency map of the image under consideration and indicates the salient points. However, we compute the saliency map of an image by considering all the images in the database. The salient regions are extracted from saliency map. Therefore, the suggested CBIR model does not require segmentation.

In the following sections, two content-based image retrieval approaches are presented: Object-Based Approach and Holistic Approach. In the next section, Itti-Koch model is presented in detail. After that, the system overview is described in a block diagram.

3.1 Content-Based Image Retrieval Problem Definition by a Human Centric Approaches

In most of the classical CBIR systems, a predefined set of features are extracted from the query image and the images in a given database. Then, an algorithm compares the feature set of the query image and that of the database images to select the "similar" images in the database. In this study, this type of approaches are called object-based CBIR, where the query image represents a group of objects. In this approach, the properties of the objects in the database images are rather ignored. The design of the CBIR system is based on the properties of the query objects. In the object-based approach, there is one reference point for the CBIR system; that is the query object(s). No matter what is the content of the images in the database, the CBIR searches the similar objects in the database. If we make an analogy to our human perception, this approach ignores our priori knowledge about various objects gathered from our past experience. Let us give an example: suppose that we are shown an object, such as green apple as in Figure-3.1. In the next step, we are shown images sequentially and asked to select the similar objects. If the information about the objects in the database is not available to us, our choice of "similar" objects would be quite naive and perhaps would results in unacceptable objects.

In this study, we propose a new approach to CBIR system, which employs the avail-

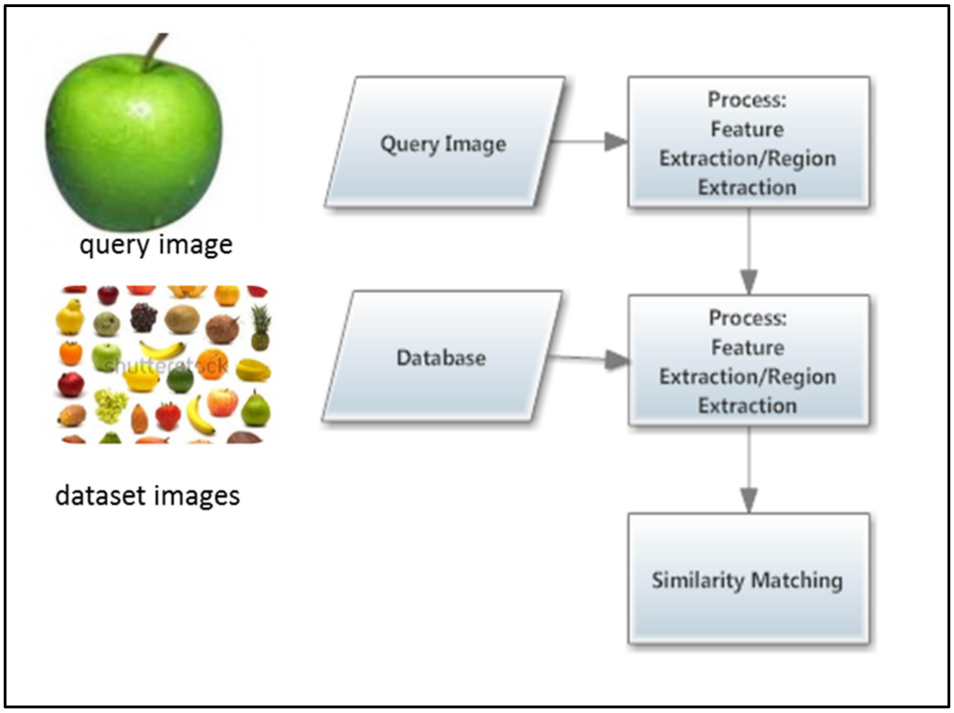


Figure 3.1: Object-based Approach of CBIR.

able information in both query image and database images. This new paradigm, called holistic approach, roughly simulates the past experience or memory of human perception. In order to achieve this task, we first analyze various properties of the images in the database. Then, we form the feature maps to capture the similarities and distinctions in the objects of the database. In object-based approach the focus of attention is on the query object, whereas in holistic approach the images in the database resembles our memories which represents the information of multiple objects in the entire database. Therefore, the proposed CBIR system processes the images in the database at the first step. Then, the query image is given as an input at the second step. For example, suppose that we have an image database which consists of fruit images as shown in Figure-3.2. The goal is to find green apples in the dataset. In the object-based approach, we do not consider any information about similarities and the distinctions about fruits. On the other hand, in the proposed holistic approach, we extract the feature maps of the images from the database. Then, by analyzing features of database images, we can localize the query image in the database. The proposed holistic CBIR system employs the visual attention system suggested by Itti and Koch.

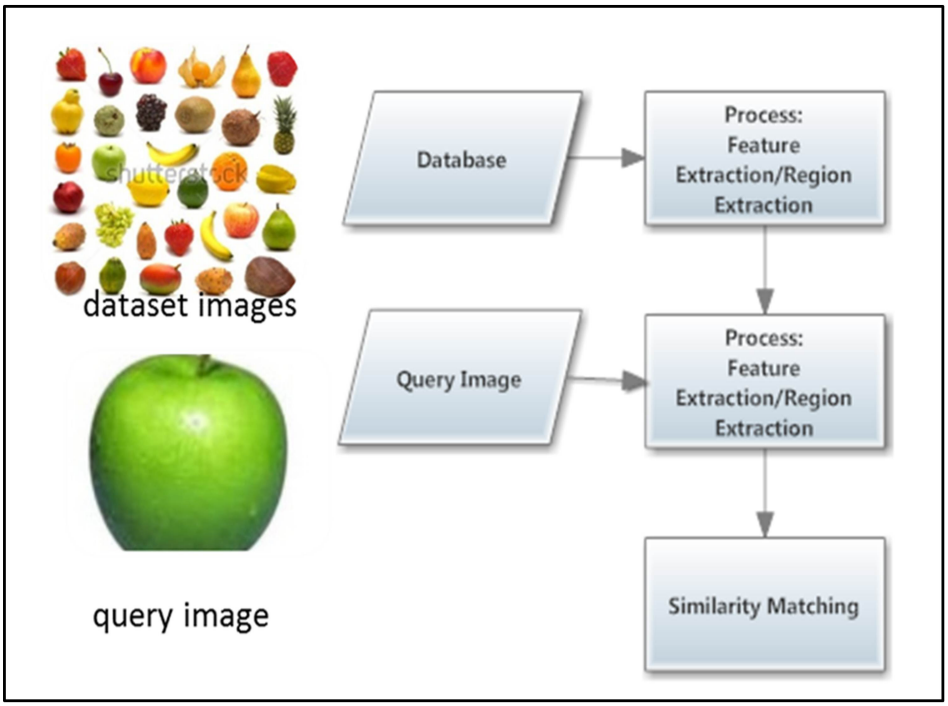


Figure 3.2: Holistic Approach of CBIR.

In the next section, we first provide a brief overview of this model. Then, we propose a revised version of Itti-Koch model to create an appropriate selection tool for CBIR problem.

3.2 The Itti-Koch Model

The Itti-Koch model is based on Treisman's Feature Integration Theory [3]. In this biologically-inspired model, an input image is decomposed into a set of multi-scale "feature maps" which extract local spatial discontinuities in the modalities of color, intensity and orientation. Each feature map is endowed with non-linear spatially competitive dynamics, so that the response of a neuron at a given location in a map is modulated by the activity in neighboring neurons. Such contextual modulation, also inspired from recent neurobiological findings [86], has proven remarkably efficient at extracting salient targets from cluttered backgrounds. All feature maps are then combined into a unique scalar "saliency map" which encodes for the salience of a lo-

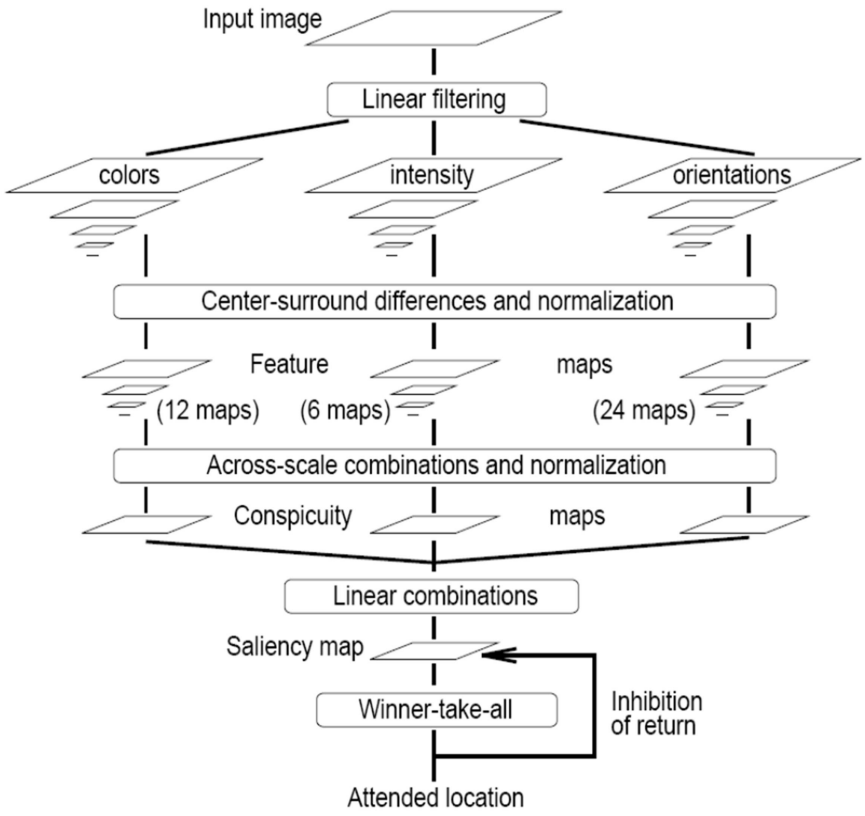


Figure 3.3: Itti-Koch Model (The figure is taken from [8]).

cation in the scene, irrespectively of the particular feature, which detects this location as conspicuous. A winner-take-all neural network, then, detects the point of highest salience in the map at any given time, and draws the focus of attention towards this location. In order to allow the focus of attention to shift to the next most salient target, the currently attended target is transiently inhibited in the saliency map. The model is depicted in Figure-3.3. The details of Itti-Koch model are explained below.

3.2.1 Feature Map Extraction

The visual input is, first, decomposed into different scales. Nine spatial scales are created using dyadic Gaussian Pyramids. Each feature is computed by a set of linear center-surround operation, inspired from the center-surround neuron activities. This operation compares the center region to surrounding region which detects spa-

tial discontinuities. The model implements the center-surround operation by taking the difference between fine and coarse scales. Center pixels are at scale $c \in \{2, 3, 4\}$ and surrounding pixels are at scale $s = c + \delta$, with $\delta \in \{3, 4\}$. This operation is called "across-scale difference" and denoted with $\ominus$, obtained by interpolation to the finer scale at point by point subtraction.

The Itti-Koch model uses three biologically inspired features: intensity, color and orientation. Intensity image is obtained as $I = (r+g+b)/3$, where r, g, and b represent red, green and blue channels of the image. I is used to create a Gaussian pyramid $I(\sigma)$, where $\sigma \in [0 \dots 8]$ is the scale. Intensity feature maps are created from the Gaussian pyramid images $I(\sigma)$, where $\sigma \in [0 \dots 8]$ by using center-surround differences $\ominus$ between fine c and course scales (s) ($c \in \{2, 3, 4\}$ and $s = c + \delta, \delta \in \{3, 4\}$) :

$$I(c, s) = |I(c) \ominus I(s)| . \tag{3.1}$$

The second feature is obtained from broadly-tuned color channels: $R = r-(g+b)/2$ for red, $G = g-(r+b)/2$ for green, $B = b-(r+g)/2$ for blue and $Y = (r+g)/2-|r - g| /2-b$ for yellow (negative values are set to zero). Four Gaussian pyramids $R(\sigma)$, $G(\sigma)$, $B(\sigma)$ and $Y(\sigma)$ are created from these color channels. Two color feature maps are created with these four Gaussian pyramids by using center-surround differences:

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| , \tag{3.2}$$

$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (B(s) - Y(s))| . \tag{3.3}$$

Orientation feature maps are obtained by using Gabor pyramids $O(\sigma, \theta)$, where $\sigma \in [0 \dots 8]$ represents the scale and $\theta \in \{0^o, 45^o, 90^o, 135^o\}$ is the orientation. The orientation feature map is computed as:

$$O(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)| . \tag{3.4}$$

Totally 42 feature maps are calculated: 6 for intensity, 12 for color and 24 for orien-

tation.

3.2.2 Normalization and Conspicuity Map Computing

Conspicuity maps are obtained by combining feature maps. A normalization operation is applied to feature maps to obtain comparable modalities. Itti-Koch model proposed a map normalization operator ($N(.)$) which globally promotes maps in which a small number of strong peaks of activity is present, while globally suppressing maps which contain multiple comparable peak responses.$N(.)$ consists of;

1. normalizing the values in the map to a fixed range $[0 \dots M]$, in order to eliminate modality-dependent amplitude differences,
2. finding the location of the map's global maximum $M$and computing the average m of all its other local maxima,
3. globally multiplying the map by $(M - \overline{m})^2$.

This normalization operator is also biologically inspired, it coarsely replicates cortical lateral inhibition mechanisms, in which neighboring similar features inhibit each other via specific, anatomically-defined connections. Figure-3.4, shows how normalization operator computes the normalized maps.

Conspicuity maps $\overline{I}$, $\overline{C}$, $\overline{O}$ are obtained through across-scale addition $\bigoplus$ for each feature as follows:

$$\overline{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s)), \tag{3.5}$$

$$\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \left[\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))\right], \tag{3.6}$$

$$\overline{O} = \sum_{\theta \in \{0^0, 45^0, 90^0, 135^0\}} \mathcal{N}\Big(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta))\Big). \tag{3.7}$$

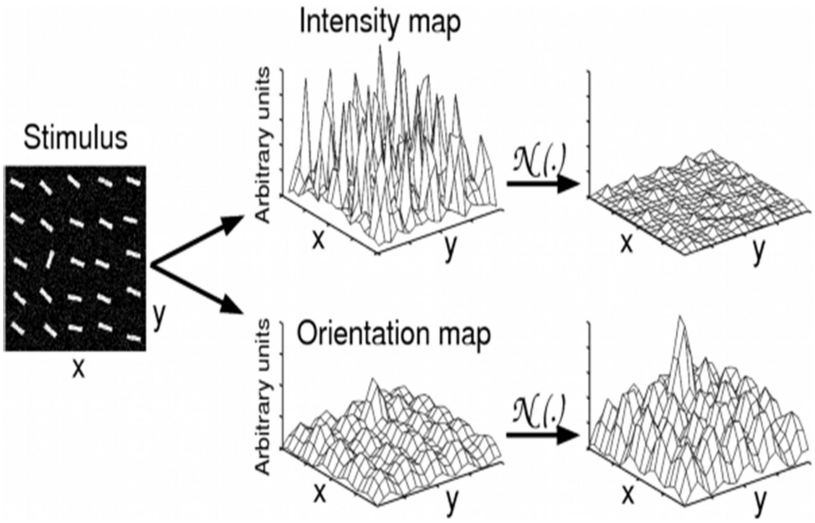


Figure 3.4: The normalization operator N(.). Note that the normalization operation accentuate the highest peak and suppresses similar neighboring peaks.(The figure is taken from [8]).

3.2.3 Saliency Map Computing

Finally, three conspicuity maps are normalized and summed into the saliency map:

$$S = \frac{1}{3}(\mathcal{N}(\overline{I}) + \mathcal{N}(\overline{C}) + \mathcal{N}(\overline{O})). \tag{3.8}$$

The input image and corresponding saliency map of Itti-Koch model is shown in Figure-3.5. As seen from the figure, the brighter regions are more salient than other regions. However, the more salient regions are not always point of interest. For example, the sharp contrast between the wall and grasses is also detected as salient as seen from the figure. Because of the region is visually different from its surroundings, it is expected that Itti-Koch model detects it. Besides this, the model indicates the other two objects in the image as seen from the saliency map of the image.

Note that the above saliency map represents a bottom up visual attention, where the

(a) (b)

Figure 3.5: Itti-Koch Model Saliency Map (a) Input image (b)) Itti-Koch Saliency Map of input image.

salient regions indicate "important regions within a single image". As it is mentioned before the saliency map, extracted by Itti and Koch does not employ any information about our past knowledge, nor does it seek particular object within an image or set of images.

3.2.4 Winner-take-all network and Inhibition of Return

A winner-take-all (WTA) network [8] is used to detect the location of highest saliency regions in terms of elementary features such as color, orientation and intensity. These features are represented in various type of feature maps that enhances the local conspicuity. The output of these maps is combined with the saliency map. The WTA network localizes the most active unit in the saliency map. In other words, the WTA network equivalent to a maximum finding operator which operates the units in the saliency map. A more detailed explanation of WTA can be found in [87] The other regions are examined by inhibition of return of former attended region.

In this thesis, we extend the bottom up visual attention model of Itti-Koch in such a way that we can model also top down visual attention with in the same framework. With the suggested generation of Itti-Koch model, we suggest a new approach to CBIR problem. The details of the suggested CBIR, called Attention Based Image Retrieval is introduced in the next section.

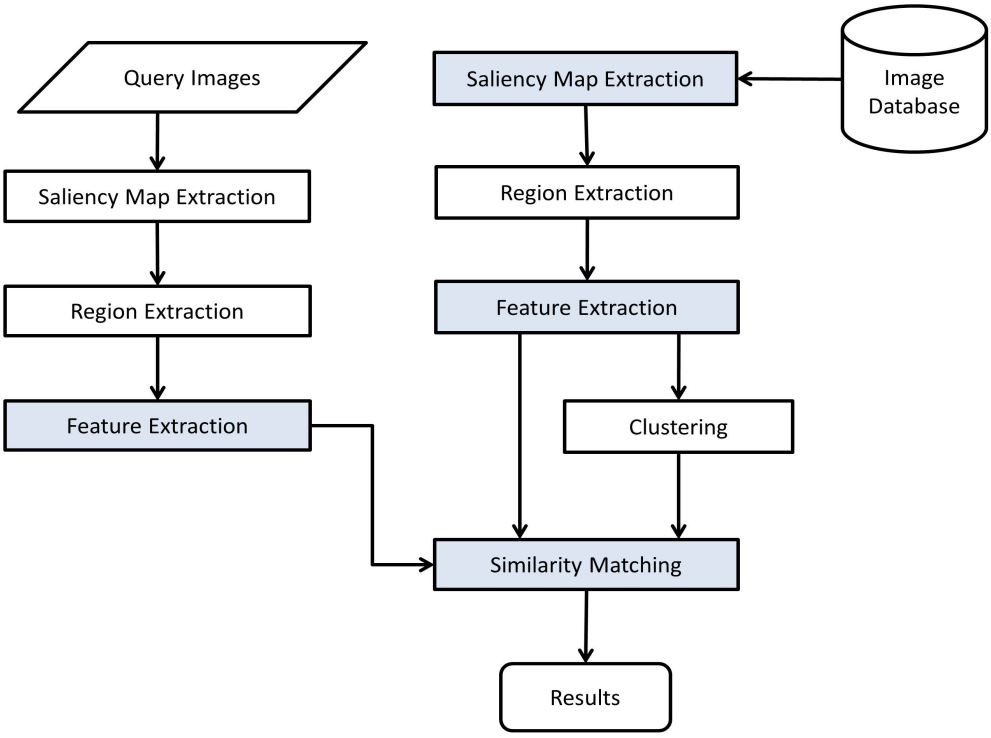


Figure 3.6: The block diagram representation of Attention-based image retrieval

3.3 System Overview for the Suggested Attention Based Image Retrieval (ABIR)

The block diagram representation of the suggested ABIR system is shown in Figure-3.6. The first step of the suggested attention-based image retrieval system is saliency map extraction. In this study, three algorithms are proposed to extract saliency maps of query and dataset images: Bottom-Up Normalization, Top-Down Normalization and Top-Down Feature Map Weighting algorithms.

The next step is region extraction. Region extraction is done on the saliency maps of images by using region growing and adaptive thresholding methods as suggested in [9]. Saliency-based feature integration algorithm and Saliency-based feature selection algorithm are proposed for feature extraction in the third step. The final step is the similarity matching between regions of query image and regions of the dataset images. The details of the components for the suggested ABIR system are explained in next sub sections.

3.4 Saliency Map Extraction for ABIR System

The first step of the suggested ABIR system is to compute saliency map of query and dataset images. The saliency maps will be used to extract regions from images in the next step. In this section, we propose a set of new saliency map extraction algorithms which are specific to the content-based image retrieval problem. The proposed algorithms extracts top-down and bottom-up saliency maps based on the algorithm of Itti-Koch Visual Attention model. The output of the suggested model is a saliency map, which indicates the degree of salient points both in a single image and the images in the dataset.

The Itti-Koch model computes the saliency map of images in a database one by one which corresponds to bottom up visual attention in human visual system. However, it is well known that bottom-up visual attention is associated with top-down visual attention to perceive the objects in a scene. The perception of objects are realized in two steps: object localization and object recognition. Object localization or detection is the "where" question and object recognition is the "what" question in the human brain. In fact that means that we can attend the objects before recognizing them [9]. It is believed that bottom-up visual attention has a role to find the locations of objects in human brain [9]. The role of bottom-up attention is to emphasis salient regions according to the neighboring regions in a scene, whereas the objects have salient regions mostly.

As mentioned previously, the goal of CBIR systems is to retrieve images in an image database, which are similar to a query image. In most applications, a part of the query image rather than the whole image is the focus of interest. In this case, the goal of CBIR is to retrieve the images which contain the focus of interest in the query image. In most cases, focus of interest is an object. However, where is the object(s) or focus of interest in the image and what is the object or focus of interest are unknown to the CBIR system. Hence, bottom-up visual attention, suggested by Itti-Koch model, provides an important tool to determine the possible location(s) of object or focus of interest in images. Since, the similarities between the query and database images is used to rank and retrieve the images in the database, one can skip the object recognition task and directly compare saliency maps.

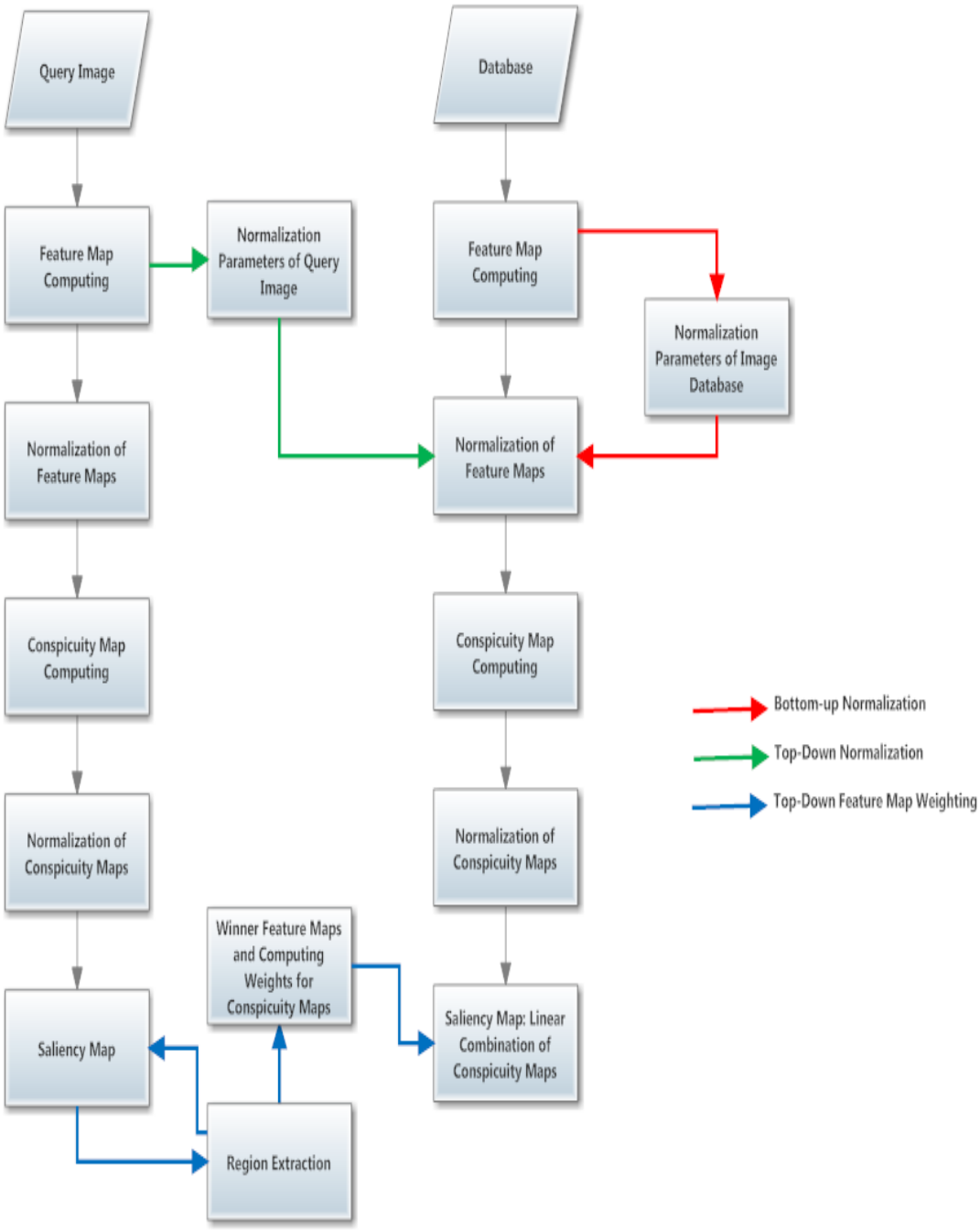


Figure 3.7: General structure of the proposed Computational Model of Visual Attention and ABIR. Red Arrows show steps which are for Bottom-Up Normalization Algorithm. Green Arrows for Top-Down Normalization Algorithm and Blue ones for Top-Down Feature Map Weighting Algorithm.

Although, the Itti-Koch model extracts salient regions from the images, it is proposed for a single image. However, there are a set of query and database images in a CBIR system. We design a visual attention model based on the Itti-Koch model to compute saliency of images by considering the all images in an system. We assume that the salient regions in the images can compose the object of interest in the query and database images.

Recall that the Itti-Koch model has three main steps: feature map computation, normalization-conspicuity map computation and saliency map computation. Let us know, reevaluate these steps and revise them for our Attention-Based Image Retrieval System. One of the most crucial step of Itti-Koch model is the normalization and conspicuity computation. In this study, this step is reformulated for ABIR by three main approaches:

1. Bottom-Up Normalization (Algorithm 1)
2. Top-Down Normalization (Algorithm 2)
3. Top-Down Feature Map Weighting (Algorithm 3)

The first and the second methods are adapted from the normalization step of the Itti-Koch model. Recall that the normalization operator of the Itti-Koch model has two parameters: the global maximum of maps (M) and the average of the local maxima (m) of the maps. Bottom-up normalization and top-down normalization methods attack the global maximum parameter (M) in different ways according to the query and dataset images. Bottom-up normalization algorithm normalizes the feature maps of images in the database by computing the global maxima from the all images, whereas the Itti-Koch computes the global maxima from a single image. Top-down normalization algorithm normalizes the feature maps of images in the database by using global maxima of query image. Firstly, the saliency map of query image is computed, then saliency map of images in the database are computed. Recall that, Itti-Koch compute the conspicuity maps of an image by employing across-scale addition to the feature maps. All feature maps have equal weights which is 1 by default. Top-down feature map weighting algorithm adjust the weights of feature maps of images in the database by using salient regions of query image.

The overview of our proposed saliency map extraction approach is depicted in Figure-3.7. While the block diagram on the left hand side is used to employ the query image, the right hand side employs the database images to generate the saliency maps. Each side has similar stages as in Itti-Koch model: feature map computing, normalization, conspicuity map computing and saliency map computing. Red connections indicates the Bottom-up Normalization Algorithm. As seen from the figure, red connections are related with only database. Green connections shows the specific stages of Top-down Normalization algorithm. Green connections add information to the database at the normalization step as seen from the figure. Blue connections shows how the query image is used to compute weights of database images for top-down feature map weighting algorithm. The details of the proposed algorithms are explained in the following sub-sections.

We give the definitions of parameters which are used in this section for the sake of clarity:

- **Itti-Koch Saliency Map**: The output of the classical Itti-Koch model.
- **Bottom-up Saliency Map**: The output of Bottom-up Normalization algorithm, suggested in this study.
- **Top-down Saliency Map**: The output of Top-down Normalization algorithm, suggested in this study.
- **Top-down Feature Weighting Saliency Map**: The output of Top-Down Feature Map Weighting algorithm, suggested in this study.
- **Feature Maps**: Color, Intensity or Orientation feature maps are computed at the first step of Itti-Koch model.
- **Winner Map**: Feature Map which has the highest saliency value at the location of most salient point in the saliency map.
- **Fixation points**: The pixel location of the most salient point in the saliency map.

3.4.1 A Holistic Approach for Bottom-Up Normalization

The eye fixation experiments on human subjects with stimulus images shows that the output of the Itti-Koch model is consistent with the output of eye fixation experiments on human subjects [47]. The Itti-Koch model computes 42 feature maps by applying center surround difference operator to the features extracted from the different scales of the image. The normalization operator makes the model more biologically plausible by enabling to combine these 42 feature maps into one saliency map [47], [48]. The normalization operator puts forward the maps with few strong peaks and suppresses maps with many comparable peaks. The key point in normalization process in Itti-Koch model is the computation of global maximum and local maximum parameters of feature maps as described in the previous sections. These parameters are computed from the image under the consideration. Therefore, Itti-Koch is designed to compute the saliency map of a single image which emphasis salient regions. However, we should deal with all the images in our content-based image retrieval problem. We proposed a new bottom-up normalization algorithm to compute saliency maps of dataset image by considering a holistic approach for CBIR. In the holistic approach, images in the database can be treated as a single image. Hence, the saliency map of an image depends on all images in the database.

The first step of the Bottom-up Normalization algorithm is to compute feature maps of all images in the database. The global maximum (M) of each feature map which is used in normalization is computed by considering all feature maps of all the images in the database. The second step is the normalization of the feature maps to compute conspicuity maps of images by using the global parameters. The next step is to compute the saliency map of the images in database by linear combination of normalized conspicuity map. The output of Bottom-up normalization is called "bottom-up saliency map".

The mathematical representation of the Bottom-up normalization algorithm is given as follows:

Let us define the images in the database such as $D = \{D_1, D_2, \ldots, D_N\}$ where N is the number of images in the database. For each image D_i a set of feature maps FM_i

which includes intensity feature maps I_i, color feature maps C_i and orientation feature maps O_i at different scales are computed by using equations (3.1), (3.2), (3.3), (3.4)

$$\begin{aligned} FM_i &= \{I_i, C_i, O_i\}, \\ I_i &= \{I_i(c,s)\}, \\ C_i &= \{RG_i(c,s), BY_i(c,s)\}, \\ O_i &= \{O_i(c,s,\theta)\}, \end{aligned} \tag{3.9}$$

where $c \in \{2,3,4\}$, $s = c + \delta$, $\delta \in \{3,4\}$ and $\theta = \{0, 45, 90, 135\}$.

Let's define the bottom-up normalization operator represented as $N_{BU}(P_1, P_2)$. The first parameter of the operator $P_1 \in FM_i$ is feature map to be normalized. The second parameter P_2 is the global maxima, computed from the feature maps of all images in the database by the following steps:

$$\begin{aligned} &\textit{compute global maxima } \textit{for each } FM_i \textit{ as} \\ &I_{max}(c,s) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\max\{I_i(c,s)\}) && i = 1,2,\ldots,N \\ &RG_{max}(c,s) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\max\{RG_i(c,s)\}) && i = 1,2,\ldots,N \\ &BY_{max}(c,s) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\max\{BY_i(c,s)\}) && i = 1,2,\ldots,N \\ &O_{max}(c,s,\theta) = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\max\{O_i(c,s,\theta)\}) && i = 1,2,\ldots,N \end{aligned} \tag{3.10}$$

The normalization operator $N_{BU}(P_1, P_2)$ is employed to the feature maps by the following steps:

1. *Compute the local maxima $\overline{LM}$ of the feature map P_1,*
2. *Compute the global maxima P_2 by using equation (3.10),*
3. *Multiply the feature map P_1 by $(P_2 - \overline{LM})^2$.*

The normalization operator N_{BU} is employed to compute conspicuity maps for each image in the following calculations,

$$\overline{I_i} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N_{BU}(I_i(c,s),\ I_{max}(c,s)) \tag{3.11}$$

$$\overline{C_i} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [N_{BU}(RG_i(c,s)\ , RG_{max}(c,s)) + N_{BU}(BY_i(c,s)\ , BY_{max}(c,s))] \quad (3.12)$$

$$\overline{O_i} = \sum_{\theta \in \{0^0, 45^0, 90^0, 135^0\}} N(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N_{BU}(O_i(c,s,\theta)\ , O_{max}(c,s,\theta))) \quad (3.13)$$

Finally, saliency map for bottom-up normalization $SMBU_i$ of each D_i is computed by linear combination of conspicuity maps in the following,

$$SMBU_i = \frac{1}{3}(N(\overline{I}_i) + N(\overline{C}_i) + N(\overline{O}_i)) \quad (3.14)$$

Bottom-up normalization algorithm is given in Algorithm 1 for all D_i. The computations are simplified for the sake of clarity. As seen from the algorithm, all saliency maps of images in the dataset are computed by parallel programming, processing images in high performance computing environment. Note that, the saliency map of an image depends not only on a single image but also all the images due to the normalization operation performed over all the images by finding the global maxima and minima of the saliency maps. However, the saliency map of images does not take into account on the saliency maps of the images in the dataset. Moreover, the saliency of images is not affected by any top-down factor. The saliency of the query image only depends on images in the database by treating them as if there is a single image which is obtained by concatenating all the images under one scene image. Henceforth, we choose the term "bottom-up" to describe the proposed algorithm.

Algorithm 1 Bottom-Up Normalization

1: *% processing dataset images*
2: **for** i=1:N **do**
3: $FM_i(:) = ComputeFeatureMap(D_i)$
4: **end for**
5: *% Find Global Max for each feature maps*
6: **for** i=1:numOfFeatureMaps **do**
7: **for** j=1:N **do**
8: $maxFM(j)) = max(FM_j(i))$
9: **end for**
10: $globalMax(i) = mean(maxFM)$
11: **end for**
12: *% Normalize Feature Maps*
13: **for** i=1:numOfImages **do**
14: **for** j=1:numOfFeatureMaps **do**
15: $FM_i(j) = N_BU(FM_i(j), globalMax(j))$
16: **end for**
17: **end for**
18: *% Compute Saliency map of dataset images*
19: **for** i=1:numOfImages **do**
20: $SMBU_i = ComputeSMBU(FM_i)$
21: **end for**

{ *numOfFeatureMaps are 42*
N(featureMaps) => takes the all feature maps and return the normalized feature maps for an image
ComputeFeatureMap(image) => takes an image and returns corresponding feature maps
ComputeSaliencyMap(featureMaps) => takes the feature maps of an image and returns saliency map }

3.4.2 An Object-Based Top-Down Normalization Method

In the previous section, we adopted the bottom-up normalization approach of Itti-Koch model to CBIR problem by computing the global maximum of each feature map over the entire set of images. This holistic approach enables us to calibrate the feature map of the query image relative to all the images in the dataset.

Considering the fact that in a CBIR system, the goal is to retrieve the most "similar" objects to the query object. This approach allows us to retrieve the "most similar" objects comparing the "most salient" regions.

In this section, we suggest a top-down normalization method by computing the saliency map of all the images based on the query image. We call this approach as object-based approach. In this approach, the feature maps of the query image are used to normalize the feature maps of images in the database. First, the saliency map of the query image is computed. The global maxima parameters of each feature map of the query image is computed over stored the database images. Then, these parameters are used to normalize the feature map of images in the database. We introduce top-down normalization operator N_{TD} to normalize feature maps. The detail of the top-down normalization algorithm is given in Algorithm 2. The output saliency map of this algorithm is called "top-down normalization saliency map".

The mathematical representation of top-down normalization is given in below:

- Firstly, feature map of a query image D_q is computed by using equations (3.1), (3.2), (3.3), (3.4) by applying top-down normalization operator given as below,

$$\begin{aligned} FM_q &= \{I_q, C_q, O_q\}, \\ I_q &= \{I_q(c,s)\}, \\ C_q &= \{RG_q(c,s), BY_q(c,s)\}, \\ O_q &= \{O_q(c,s,\theta)\}. \end{aligned} \tag{3.15}$$

  In the above formulation, $c \in \{2,3,4\}$, $s = c + \delta$, $\delta \in \{3,4\}$, $\theta = \{0, 45, 90, 135\}$, O_q is the orientation feature map, I_q is the intensity feature map and C_q is the color feature map of the query image.

- The next step is to normalize the feature maps of each image in the database by N_{TD} to compute the conspicuity maps. Top-down normalization operator $N_{TD}(P_d, P_q)$ is defined with two arguments. The first argument $P_d \in FM_i$ is the feature map of database image D_i as described in bottom-up normalization algorithm with equation (3.9). The second argument P_q is the global maxima of feature maps which are computed by using feature maps of query image defined as below,

$$N_{TD}(.,.) = \begin{cases} mI_q(c,s) = \max(I_q(c,s)) \\ mRG_q(c,s) = \max(RG_q(c,s)) \\ mBY_q(c,s) = \max(BY_q(c,s)) \\ mO_q(c,s,\theta) = \max(O_q(c,s,\theta)) \end{cases} \tag{3.16}$$

  where each of the equation represent the maxima of corresponding feature maps according to the query image. The normalization operator $N_{TD}(P_d, P_q)$ is employed to the feature maps as the following steps:

  1. *Compute the local maxima* $\overline{LM}$ *of feature map* P_d,
  2. *Compute the global maxima defined as* P_2 *from the all feature maps in the database*,
  3. *Compute* $N_{TD}(P_d, P_q) = P_d(P_q - \overline{LM})^2$.

- The next step is to compute the conspicuity maps by using N_{TD} operator as described in below,

$$\overline{I_i} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N_{TD}(I_i(c,s),\ mI_q(c,s)), \tag{3.17}$$

$$\overline{C_i} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [N_{TD}(RG_i(c,s)\ , mRG_q(c,s)) + N_{TD}(BY_i(c,s)\ , mBY_q(c,s))], \tag{3.18}$$

$$\overline{O_i} = \sum_{\theta \in \{0^0, 45^0, 90^0, 135^0\}} N(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N_{TD}(O_i(c,s,\theta)\ , mO_q(c,s,\theta))). \tag{3.19}$$

- Finally, the saliency map $SMTD_i$ of image D_i is computed as follow,

$$SMTD_i = \frac{1}{3}(N(\bar{I}_i) + N(\overline{C}_i) + N(\overline{O}_i)). \tag{3.20}$$

The details of top-down normalization algorithm for each image D_i is given in Algorithm 2. The computations are simplified for the sake of clarity of the algorithm.

We use the term 'top-down' to describe this Algorithm 2, since the extra information computed from normalization parameters of query image, which is not available in the image under the consideration, is added to the system for computing saliency map. However, this information, called top-down information (the normalization parameters) reflects the whole query image, not the region of interest in query image. For instance, if the red color is salient in query image, the saliency of red color in database is increased by top-down normalization.

Let us give an example with artificial images as shown in Figure-3.8(a). Suppose that, in a hypothetical example, there are two database images and one query image represented in one dimensional. Assume that this dimension corresponds to the red feature map. The second Figure-3.8(b) shows that images normalized by using the normalization operator $N(.)$ of Itti-Koch. The third figure Figure-3.8(c) shows the images normalized by the suggested top-down normalization operator N_{TD}. Note that, there is a peak in the feature map of query image, which means there is salient region in the red feature map. As it is seen from the Figure-3.8 that feature map of query image increases the saliency of red feature maps in the database images.

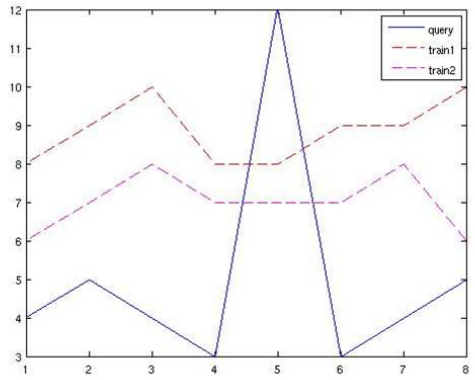
Algorithm 2 Top-down Normalization

1: *% process the Query Image*
2: $FM_Query(1,:) = ComputeFeatureMap(QueryImage)$
3: *% Find Global Max from Query Maps*
4: **for** i=1:numOfFeatures **do**
5: $maxFM(i,:) = max(FM_Query(1,i))$
6: $globalMax(i) = max(maxFM)$
7: **end for**
8: *% processing dataset images*
9: **for** i=1:numOfImages **do**
10: $FM(i,:) = ComputeFeatureMap(image(i))$
11: **end for**
12: *% Normalize Feature Maps*
13: **for** i=1:numOfImages **do**
14: **for** j=1:numOfFeaturesMaps **do**
15: $FM(i,j) = N_TD(FM(i,j), globalMax(j))$
16: **end for**
17: **end for**
18: *% Compute Saliency map of dataset images*
19: **for** i=1:numOfImages **do**
20: $SMTD(i,:) = ComputeSM(FM(i,:))$
21: **end for**

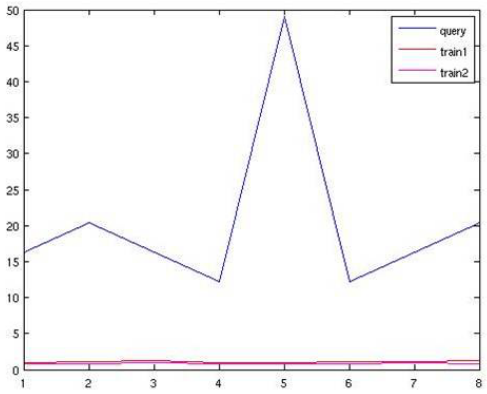
{*Normalization(featureMaps) => takes the all feature maps and return the normalized feature maps for an image*

ComputeFeatureMap(image) => takes an image and returns corresponding feature maps

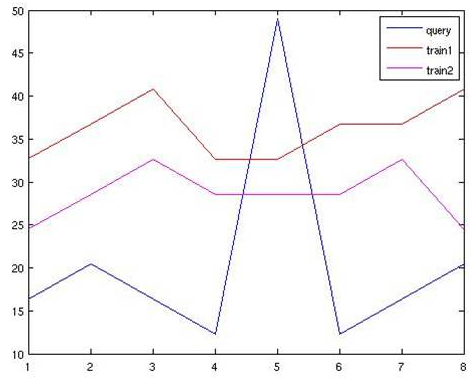
ComputeSaliencyMap(featureMaps) => takes the feature maps of an image and returns saliency map }



(a)



(b)



(c)

Figure 3.8: Top-down normalization with artificial images (a)original images (b) images normalized by Itti-Koch normalization operator N (c) images normalized by using Top-down normalization operator N_{TD}

3.4.3 Top-Down Feature Map Weighting Algorithm

In the previous section, we proposed two different algorithms, bottom-up normalization and top-down normalization, to extract saliency map of images in the database by taking into account of normalization of feature maps, based on Itti-Koch model. In this section, we propose a new algorithm called Top-Down Feature Map Weighting which presents a CBIR method to use top-down information for content-based image retrieval problem. This algorithm realizes the object-based approach for CBIR problem.

As opposed to Top-Down normalization algorithm, this algorithm uses the point-of-interests of the query image, which are the fixation points extracted from saliency map of query image, to compute the saliency map of images in the database. There are two main steps of the algorithm, namely, processing the query image and extracting the saliency maps of images in the database:

Step 1 Query Image processing: Firstly, the saliency map of the query image is computed by using Itti-Koch model. After that, fixation points are computed from the saliency map by using winner-take-all-network and applying inhibition of return as described in [8]. The number of fixation points can be changed according to the image, but we take it as a constant value for the sake of comprehensibility of the algorithm. The set of fixation points of the query image F_q is represented as below,

$$F_q = \{f_1, f_2 \ldots, f_t\}, \tag{3.21}$$

where t is the number of fixation points of the query image and f_j is the pixel coordinate in two dimension of the $j-th$ fixation point.

The next step is to find the winner map for the each fixation point. The winner map is the feature map which has the highest value for each fixation point among all feature maps that compose the saliency map of the current query image. The fixation point f_j is back tracked through the conspicuity map $CM(f_j)$ and feature map FM_j to find the winner map [88]. The winner conspicuity map cm_j of the fixation point f_j is computed as,

$$cm_j = \underset{CM \in \overline{I}_q, \overline{C}_q, \overline{O}_q}{\arg\max} \; CM(f_j). \tag{3.22}$$

Then, the cm_j is back tracked to the its feature maps to find winner feature map represented by the following equations,

$$FM_j = \begin{cases} I_q & \text{if } cm_j = \overline{I}_q \\ C_q & \text{if } cm_j = \overline{C}_q, \\ O_q & \text{if } cm_j = \overline{O}_q \end{cases} \tag{3.23}$$

where FM_j is the set of feature maps of the query image. The scale and orientation of the feature map is computed as

$$(FM_j, c_j, s_j, \theta_j) = \arg\max FM_j(cx, sx, \theta x)(f_j), \tag{3.24}$$

where $cx \in \{2, 3, 4\}$, $sx \in \{cx + 3, cx + 4\}$, $\theta x \in \{0^o, 45^o, 90^o, 135^o\}$ for $FM_j = O_q$, c_j is the coarse scale, s_j is the fine scale and θ is the orientation of FM_j. The θ_j is computed if and only if FM_j is equal to O_q. Otherwise, θ_j is $NULL$.

Step 2 Saliency Map Extraction for the dataset images: The feature maps and conspicuity maps of image D_i are computed by using equation (3.9).

In the top-down and bottom-up normalization algorithms computed for the query image D_i, feature maps have evenly contributed to construct the conspicuity maps. In other words, the weights of all feature maps of D_i were equal to 1 represented as:

$$\forall w(k, c, s, \theta) = 1, \tag{3.25}$$

where w is the weight vector, $k \in \{I, C, O\}$, $c \in \{2, 3, 4\}$, $s \in \{c + 3, c + 4\}$, $\theta \in \{0^o, 45^o, 90^o, 135^o\}$ for $k = O$ otherwise $\theta = NULL$.

For the dataset images, the weights are updated for each fixation point f_j by using the winner feature map of the query image as follows,

$$w(FM_j, c_j, s_j, \theta_j) = 1 + FM_j(c_j, s_j, \theta_j)(f_j), \tag{3.26}$$

where j is ranging from 1 to t.

Then, the conspicuity maps of D_i are computed by the following equations,

$$\overline{I}_i = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} w(I, c, s) \cdot N(I_i(c, s)), \tag{3.27}$$

$$\overline{C}_i = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} w(RG, c, s) \cdot N(RG_i(c, s)) + w(BY, c, s) \cdot N(BY_i(c, s)), \tag{3.28}$$

$$\overline{O}_i = \sum_{\theta \in \{0^0, 45^0, 90^0, 135^0\}} N(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} w(O, c, s, \theta) \cdot N(O_i(c, s, \theta)). \tag{3.29}$$

Then, saliency map $SMTDA_i$ extracted from top-down feature map algorithm of D_i is obtained by,

$$SMTDA_i = \frac{1}{3}(N(\overline{I}) + N(\overline{C}) + N(\overline{O})). \tag{3.30}$$

The top-down feature map weighting algorithm is described in details in Algorithm 3 for images in dataset D_i and query image D_q with a simplied computation for the sake of clarity of the algorithm. The contribution of the Top-Down Feature Map Weighting algorithm is to add top-down information of the query image to the computation of saliency map of images in the database. To give an example, if the color feature is salient in the query image, the weight of the corresponding color feature map of database images is increased with the value of fixation point in the winning feature map of the query image. The scale of the salient point of the query image is emphasized in the database by adjusting weights according to the scale of the feature maps. Top-down Normalization algorithm emphasis the all features of query image in the database images by using normalization. However, Top-down feature map weighting algorithm emphasis the only salient feature maps of query image in the database images by adjusting weights of feature maps in the database images.

Algorithm 3 Top-down Feature Map Weighting Algorithm

D_q is query image
t = number of regions for Q
Q_{SM} is the saliency map of Q
SM is the matrix of all saliency maps of training images
N = number of images in the database
Find the winner maps index, M, from Q_{SM}, $M = [m_1, m_2, \ldots, m_t]$
Calculate the weight vector $W = [w_1, w_2, \ldots, w_t]$
for i=1:N **do**
    for j=1:t **do**
        $SM(i, M(j)) = W(j) * SaliencyMap(i, M(j))$
    end for
    update $SM(i, :)$
end for
Return SM

3.5 Saliency Region Extraction

After extracting top-down, bottom-up and weighted feature saliency maps algorithms, the next step is region extraction from the saliency maps. The saliency map is a topological map, which indicates visually salient points in an image. We attend to these salient locations to extract regions from the corresponding image. The size and shape of the region around the attended locations may be adjustable according to the image, object or the application domain.

Region extraction is done by using the saliency map and feature maps as described in [88],where the extracted region are called as proto-object. Let us define the saliency regions of image D_i with the set $SR = \{SR_{i,1}, SR_{i,2}, \cdots, SR_{i,t}\}$ where t is the number of regions in image D_i and $SR_{i,j}$ is the $j-th$ region of D_i. The steps of the method to extract the saliency region $SR_{i,j}$ from the image D_i are given as below:

Step 1 Computing Fixation Points: Firstly fixation points of D_i are computed by using a winner-take-all (WTA) network of integrate-and-fire neurons and inhibition of return from the saliency map of D_i. The set of t fixation points

for image D_i is $F_i = \{f_{i,1}, f_{i,1}, \cdots, f_{i,t}\}$. Feed-forward connections in Figure-3.9 shows the computation of saliency map and fixation points.

Step 2 Finding winner feature map for $f_{i,j}$: The next step is to find winner conspicuity map for fixation point $f_{i,j}$ by back tracking the location of fixation point from saliency map to the all conspicuity maps. Let us define the winner conspicuity map with $CM_{i,x}$ where $x \in \{\overline{I}_i, \overline{C}_i, \overline{O}_i\}$ as defined previously. Then, $CM_{i,x}$ is back tracked through feature maps to find the winner feature map $WFM_{i,j}$ where $WFM_{i,j} \in FM_i$ as described in equation (3.9). This step is shown with feedback connections in Figure-3.9.

Step 3 Segmentation of Winner Map $WFM_{i,j}$: The winner feature map $WFM_{i,j}$ is segmented around the most salient location by using region growing and adaptive thresholding methods. A binary map is obtained which is called $Mask_{i,j}$. Local connections in Figure-3.9 shows the segmented winner map. Scalar multiplication of $WFM_{i,j}$ and D_i gives the salient region or proto object $SR_{i,j}$ as seen from the Figure-3.9.

Step 4: After the region $SR_{i,j}$ is extracted, the step-2, step-3 are repeated for other fixation points in F_i. Further details and the discussions of the method are given in the work [88].

In Figure-3.10, an image and an extracted salient regions are shown. In the first row, the region is extracted from the saliency map by segmenting the saliency map around the fixation point. In the second row, the region is extracted from the winner feature map by the method explained in [88]. The method explained above finds the winner location in Intensity feature map by backtracking through the conspicuity and feature maps. As seen from the figure, winner map gives a better result in salient region extraction.

3.6 Feature Extraction from the Salient Regions

It is well known that the visual features are used to describe the content of the digital images. Therefore, feature extraction is a very crucial operation for computer

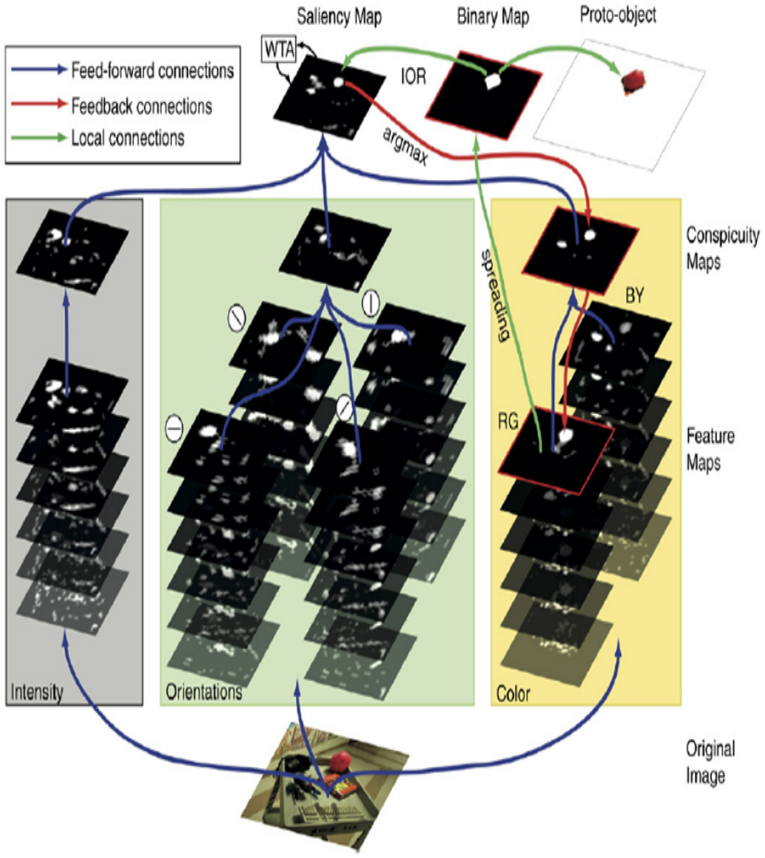
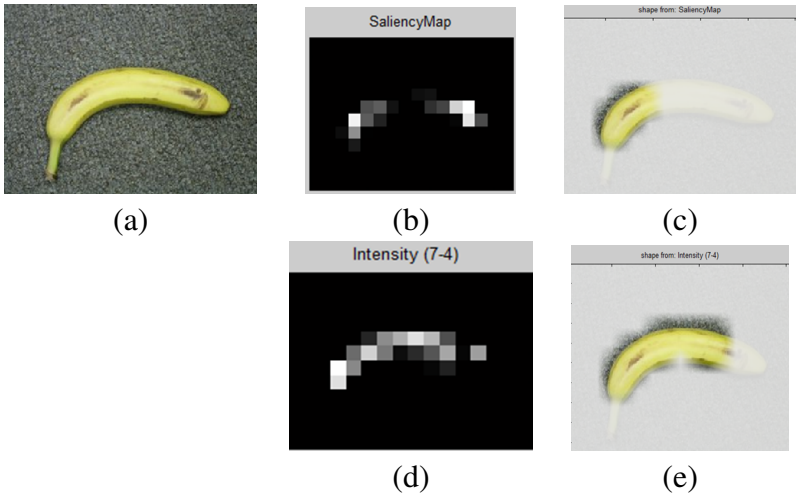


Figure 3.9: Salient region extraction (The Figure is taken from [9]).

vision applications. The visual features describe low-level characteristics such as the shape, the color, the texture or the intensity. There are many feature extraction and description methods to detect these low-level features in computer vision and image processing literature [19], [18], [51].

In this work, our focus of attention is the salient regions and visual attention, therefore simple and low-level features are selected to describe these regions. Color histogram, edge histogram, entropy and average intensity features are extracted from salient regions of both the query and the dataset images.

According to Feature Integration Theory [3], low-level visual features are computed in parallel fashion in human brain. However, in computer vision applications, the common way is to concatenate the features. Feature concatenation is a simple, yet a



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

Figure 3.10: An image and its salient regions. (a) original image (b) saliency map (c) salient regions extracted from saliency map (d) the winner map for the first fixation point (e) region which is extracted by using winner map is shown.

controversial way which requires normalization of the feature.

We propose two different algorithms for feature integration and feature selection: Saliency-based Feature Integration and Saliency-based Feature Selection. The details of the algorithms are explained in the following sections.

3.6.1 Saliency-based Feature Integration Algorithm

In the previous steps of our ABIR system, we extracted the saliency map and salient regions of the query and dataset images. At this stage, we extract features from each region for similarity matching. We propose an algorithm called Saliency-based feature integration to extract features from the regions, which are used in the final step of our ABIR system. Saliency-based Feature Integration Algorithm describes how these features are integrated by using the saliency of the region shown in Algorithm 4. In this approach, we suggest a new technique to integrate the features, extracted from each salient region $SR(i)$ by employing the saliency map.

Recall that, color, intensity and orientation feature maps are extracted for computing the saliency map in Itti-Koch model. In parallel to this approach, we extract three different features from salient regions, namely average color, intensity and entropy defined respectively as

Algorithm 4 Saliency-based Feature Integration

C = Color histogram feature vector of region
E = Entropy values of region
I = Average intensity value of region
$regionFeatures = [C \;\; E \;\; I]$
var = Visual attention values of salient region
$var = vav + 1$
$label$ = The region label ($Color, Orientation, Intentisty$)
if $label =' Color'$ **then**
    $regionFeatures = [vav * C \;\; E \;\; I]$
else if $label =' Orientation'$ **then**
    $regionFeatures = [C \;\; vav * E \;\; I]$
else
    $regionFeatures = [C \;\; E \;\; vav * I]$
end if
return $regionFeatures$

$$\begin{aligned} Ave_{Color} &= [mean(R), mean(G), mean(B)], \\ Ave_{Intensity} &= (R + G + B)/3, \\ Entropy &= E[-ln(P(x))]. \end{aligned} \tag{3.31}$$

where $P(x)$ is the color intensity histogram.

The feature integration method of a single salient region is explained in Algorithm 4. An image D_i is represented by its salient regions and fixation point with the sets respectively $SR_i = \{SR_{i,1}, SR_{i,2}, \ldots, SR_{i,t}\}$ and $F_i = \{F_{i,1}, F_{i,2}, \ldots, F_{i,t}\}$ where t is the number of regions in D_i.

The feature vector $FV_{i,j}$ of region $SR_{i,j}$ that refers to $j-th$ region of D_i is obtained by concatenation of average color, average intensity and entropy represented as follows,

$$FV_{i,j} = [Ave_{Intensity}(SR_{i,j}), Ave_{Color}(SR_{i,j}), Entropy(SR_{i,j})]. \tag{3.32}$$

Our goal is to adjust weights of features according to the saliency of regions. For that

purpose, we assign a label for each region. Labels are computed with the following formula,

$$label_{i,j} = \underset{CM \in \overline{I}_i, \overline{C}_i, \overline{O}_i}{\arg\max}\ CM(f_{i,j}) \tag{3.33}$$

The initial weights of $FV_{i,j}$ are defined as,

$$WR_{i,j} = \{wI_{i,j}, wC_{i,j}, wE_{i,j}\} \tag{3.34}$$

where $WR_{i,j}$ is the vector of weights for region $SR_{i,j}$ consisting of average intensity $wI_{i,j}$, average color $wC_{i,j}$ and $wE_{i,j}$ entropy weights. The weight vector is updated by using $Label_{i,j}$ defined as

$$\begin{aligned} wI_{i,j} &= S_i(f_{i,j}) + 1 \quad \text{if } label_{i,j} = \overline{I}_i, \\ wC_{i,j} &= S_i(f_{i,j}) + 1 \quad \text{if } label_{i,j} = \overline{C}_i, \\ wE_{i,j} &= S_i(f_{i,j}) + 1 \quad \text{if } label_{i,j} = \overline{O}_i, \end{aligned} \tag{3.35}$$

where S_i is the saliency map of D_i. Finally, the weight vector is normalized between 0 to 1. The weighted feature vector $FV_{i,j}$ is computed as,

$$\begin{aligned} FV_{i,j} = [&wI_{i,j} \cdot AverageIntensity(SR_{i,j}) \\ &wC_{i,j} \cdot AverageColor(SR_{i,j}) \\ &wE_{i,j} \cdot AverageEntropy(SR_{i,j})]. \end{aligned} \tag{3.36}$$

3.6.2 Saliency-based Feature Selection Algorithm

In the previous section, we propose a new method to integrate to features extracted from each saliency region. In this section, we propose an algorithm to select the best representative feature for the region of an image with respect to the saliency map of the corresponding region. This method is called Saliency-based Feature Selection, and it is given in Algorithm 5.

Algorithm 5 Saliency-based Feature Selection Algorithm

C = Color histogram feature vector of salient regions, 125 dimensional vector
E = Entropy values of region
I = Average intensity value of region
$regionFeatures = [C \;\; E \;\; I]$
vav = Visual attention values of region
$label$ = The region label ($Color, Orientation, Intentisty$)
if $label =' Color'$ **then**
    $regionFeatures = [C]$
else if $label =' Orientation'$ **then**
    $regionFeatures = [E]$
else
    $regionFeatures = [I]$
end if
return $regionFeatures$

As described in the previous section, the regions are extracted by using winner feature map. The weighted feature vector $FV_{i,j}$ of the region $SR_{i,j}$ of the D_i is computed similar to the saliency-based feature integration algorithm. The difference of saliency-based feature selection algorithm is to compute the weight vector equation (3.35) by,

$$\begin{aligned} wI_{i,j} = 1, wC_{i,j} = 0, wE_{i,j} = 0 \quad & \text{if } label_{i,j} = \overline{I}_i \\ wI_{i,j} = 0, wC_{i,j} = 1, wE_{i,j} = 0 \quad & \text{if } label_{i,j} = \overline{C}_i \\ wI_{i,j} = 0, wC_{i,j} = 0, wE_{i,j} = 1 \quad & \text{if } label_{i,j} = \overline{O}_i \end{aligned} \tag{3.37}$$

where all other equations (3.32), (3.33), (3.34), (3.36) are applied for saliency-based feature selection algorithm.

3.7 Similarity Matching

At this point, a similarity measure is needed to compare the salient regions between the query and dataset images. It is well-known that region-based image retrieval systems require an initial segmentation, which is an erroneous process [89]. In this

study, the region of interest is extracted by using visual attention model to avoid the classical image segmentation problems. This approach, also, enables us to reduce the irrelevant data in the background and unattended regions by using only the salient regions.

For this purpose, we employ the Integrated Region Matching, IRM proposed in [27]. IRM matches the images based on region sets according to a significance measure between the regions. In an attention-driven method, which region is attended first is an important information, since salient regions are expected to reflect a better semantic information. In this study, image similarity between two images is computed using IRM with a new significance measure which utilizes saliency measure of regions.

Mathematically speaking, let Image-I and Image-II be represented by region sets $R_I = \left\{r_1^I, r_2^I, \ldots, r_m^I\right\}$ and $R_{II} = \left\{r_1^{II}, r_2^{II}, \ldots, r_n^{II}\right\}$ where r_i^I or r_j^{II} is the descriptor of regions in Image-I and Image-II, respectively. Denote the distance between region r_i^I and r_j^{II} as $d\left(r_i^I, r_j^{II}\right)$, which is written as $d_{i,j}$ in short. This distance can be computed by using the classical Euclidean distance. However, a matching between r_i^I or r_j^{II} is assigned with a significance credit $s_{i,j} \geq 0$. The distance between two region sets can be defined by using $d_{i,j}$ and $s_{i,j}$ as,

$$d(R_I, R_{II}) = \sum_{i,j} s_{i,j} \cdot d_{i,j}. \tag{3.38}$$

Now, the problem is how to define significance matrix S; $S = \left\{s_{i,j}\right\}$, $1 \leq i \leq m$, $1 \leq j \leq n$. The fulfillment of significance should be enforced to ensure that all the regions play a role for image similarity. Assume that the significance of r_i^I in Image-I is p_i and r_j^{II} in Image-II is $p_j{}'$, we require that

$$\sum_{j=1}^{n} s_{i,j} = p_i \qquad i = 1, \ldots, m, \tag{3.39}$$

$$\sum_{j=1}^{m} s_{i,j} = p_j{}' \qquad j = 1, \ldots, n, \tag{3.40}$$

A greedy scheme is used to compute of the matrix S. The difference between IRM and proposed method is in the assignment of p_i values for each region. IRM method,

proposed in [27], chooses p_i as area percentage of regions (AP), assuming larger regions are more occupied by real objects. The p_i is defined as $p_i = (AP)_i$, where $(AP)_i$ is the area percentage of region r_i^I. In [90], p_i is defined as the product of region frequency (RF) and inverse picture frequency (IPF).

In order to compute (IPF), for each region feature vector x_i of the Image-I the closest group centroid from the list of k centroids are computed as follows,

$$\left\|x_i - \widehat{x_{co}}\right\| = \min_{1 \le c \le k} \left\|x_i - \widehat{x_c}\right\|, \tag{3.41}$$

where $\widehat{x_{co}}$ is the closest group centroid to x_i. N is the number of images in the database. If N_{co} is the number of pictures in the database that include regions in cluster c_o, then we define the Inverse Picture Frequency of the region r_i^I as follows,

$$(IPF)_i = \log\left(\frac{N}{N_{co}}\right) + 1. \tag{3.42}$$

On the other hand, Region Frequency (RF) of r_i^I is defined as:

$$(RF)_i = \log\left((AP)_i . N\right) + 1. \tag{3.43}$$

Based on the above definitions, the significance value of region r_i^I is defined as:

$$p_i = (RF)_i . (IPF)_i . \tag{3.44}$$

In this study, the saliency of regions is used to define the significance criterion. For this purpose, an attention value (AV) is assigned to each region by using saliency map of image. When extracting region of interest from saliency maps, the most salient location of feature map is used as key. The numeric values of these locations, (MS), are used to compute the Attention Value (AV) parameter. Therefore, the (AV) of r_i^I in Image-I is defined as,

$$(AV)_i = \log(MS)_i + 1, \tag{3.45}$$

$$p_i = (AV)_i . \tag{3.46}$$

In the above formulation, the attention value, (AV), measures the significance of regions according to the saliency map.

After computing p values for each image region, the significance matrix of image pairs S is computed as in [27] with a greedy method.

3.8 Summary

In this chapter, the content-based image retrieval problem is approached from a visual attention viewpoint of view. The general framework of the proposed Attention-Based Image Retrieval (ABIR) system and the related algorithms are explained, emphasizing the contributions presented in this thesis work. Firstly, we introduce Holistic and Object-based approaches for content-based image retrieval problem by considering human visual attention system based on Itti-Koch model.

Thereafter, three new algorithms employing Itti-Koch model are proposed to extract saliency map for content-based image retrieval problem: Bottom-up Normalization, Top-Down Normalization and Top-Down Feature Map Weighting Algorithm. However, these algorithms are also a new computational model of visual attention for CBIR problem. Itti-Koch model deals with only one single, whereas our algorithms are designed for CBIR problem by considering query and dataset images together. Bottom-up Normalization and Top-Down Normalization algorithms use the normalization concept of Itti-Koch model to introduce a model for CBIR. On the other hand, Top-Down Feature Map Weighting algorithm attacks the CBIR problem by adding top-down information which is extracted from the saliency of query image.

Afterwards, we propose Saliency-based Feature Integration and Saliency-based Feature Selection algorithms to extract features from region of images. Both algorithms are based on the saliency of regions. The algorithms propose a practical way for feature extraction, however, the application oriented problems are not considered such as size of feature vectors in the algorithms. Finally we propose saliency-based similarity matching criteria for region matching.

CHAPTER 4

EXPERIMENTS

The proposed Attention-Based Image Retrieval system is realized in *Matlab* environment. The *Saliency Toolbox* is utilized to implement Itti-Koch model. The algorithms are parallelized in NAR, the High Performance Computing Facility of the Department of Computer Engineering.

The major goal of this chapter is to show the pros and cons of the suggested a content-based image retrieval system for the object of interests in a given image database. This goal is achieved by applying this cognitive CBIR system to object datasets for saliency detection and comparison using the Itti-Koch model [11]. In the proposed approach, we assume that the images in the dataset consist of objects which can be represented by salient regions. Therefore, while we select the database to validate and test our proposed approach we focus the objects which have high salient regions compared to the rest of the regions.

There are various standard databases to evaluate CBIR systems for different applications in computer vision literature [19], [51], such as Corel Image Database. However, the image categories in Corel cannot be represented by their object contents, but rather general scenes. For example, some sample images from Mountain category is shown in Figure-4.1. Notice that the mountain object in these scenes is not separated from the background, but should be considered as a whole with its background. Therefore, these are no saliency regions which represent the mountain objects. However, the suggested ABIR system is applicable when the database of images consists of object which can be represented by a set of salient regions. For this reason, our system is tested on object-based datasets. Two appropriate databases for the suggested

Figure 4.1: Sample images from 'Mountain' category of Corel Images (Wang Database [10]).

systems are STIM and SIVAL. These image databases are chosen due to a particular requirement of this work: the object in the images should contain at least one salient region. The STIM database is used to analyze the proposed system in detail. Every step of the suggested ABIR system is tested separately and results are discussed in detail in this chapter.

The second dataset, SIVAL, is used to compare the proposed system with the selected state of the art content-based image retrieval systems. The compared systems are region-based and localized systems, where the aim is to retrieve images in reference to the object in the query image. These systems are ACCIO [74], GMIL [84], Simplicity [60] and SBN [85].

Finally, a comparison for the available local descriptors [16], [69] and visual attention is made and their performances are tested on CBIR systems. The local descriptors aim to find interested points in images. The salient regions carry interest points, since they are the visually different areas according to the neighbouring pixels. Many local descriptors are suggested in computer vision literature [16], [17], [69]. However, it is reported that SIFT-based methods perform better than the others [16] in region-based applications [16]. Therefore, we compare CBIR systems which employ the local SIFT descriptors [13], [70] and proposed ABIR system by a set of experiments in STIM and SIVAL datasets and the results are discussed.

Furthermore, we test our system on satellite images to localize airplanes in a complex airport region. Satellite images have different properties with respect to standard image databases such as high resolution and multispectral bands. Airplane detection

in satellite images have difficulties such as variations of airplane type, pose and size [91]. However, it is observed that airplanes are salient objects in an airport region. We perform visual attention on airport images to localize airplanes.

The following section gives a brief description of datasets which are used in experiments. Then, the results of the experiments are discussed. Finally, a brief summary is provided about experimental results.

4.1 Datasets

4.1.1 The STIM Dataset

Stim Datasets includes 8-bit color natural scenes with salient traffic signs, red soda can and vehicle emergency triangle. STIM database [11] consists of three categories. STIMcoke, STIMtriangle and STIMAutobahn. STIMcoke dataset consisted of 104 images in which a red aluminum can is the salient object. The second category, STIMtriangle, consisted of 64 images in which an emergency triangle was the target. The third category, STIMautoban consisted of 90 images obtained by a video camera on German roads, and contained one or more traffic signs. The number of traffic signs in a single image varies between 1 and 5 for this category. The object in all categories could be fully visible or partially occluded, shiny or dull, in the shadow or not, light or dark, large or small, and viewed frontally or at an angle, in scenes which also demonstrated high degree of variability. The pixel resolution of images in STIM datasets is 512x384 for STIMautobahn and 640x480 for STIMcoke and STIMtriangle categories. Figure-4.2 represents some sample images from the STIM dataset.

A subset of the STIMautobahn, STIMcoke, and STIMtriangle image databases [92] was selected randomly for this work. A dataset of 112 images was selected: 42 images from the STIMautobahn set (road signs, Figure-4.2(a)), 42 from the STIMcoke dataset (red soda cans in a variety of settings, Figure-4.2(b)), and the remaining 28 from STIMtriangle (road emergency triangles, Figure-4.2(c)). A dataset of 15 images are used as query images: 5 images from each set of images.

a b c

Figure 4.2: Examples of tree types of images in the STIM dataset [11]. (a) Warning triangles, (b) road sings, (c) soda cans.

4.1.2 The SIVAL Dataset

The SIVAL (Spatially Independent, Variable Area, and Lighting) [12] dataset is created to test the performance of ACCIO! which is the first Localized Content-based Image Retrieval system in literature. The dataset includes 25 different image categories with 60 images for each category. The images are 8-bit color. The pixel resolutions of all images are 1024x768 pixel. All categories share 10 different highly cluttered backgrounds. The object size and position of objects in images varies objects occupy 10-15 per cent of the most images. There are 25 image categories which are: AjaxOrange, Apple, Banana, BlueScrunge, CandleWithHolder, CardboardBox, CheckeredScarf, CokeCan, DataminingBook, DirtyRunningShoe, DirtyWorkGloves, FabricSoftenerBox, feltflowerrug, GlazedWoodPot, GoldMedal, GreenTeaBox, JuliesPot, LargeSpoon, RapBook, SmileyFacedDoll, SpriteCan, StripedNotebook, TranslucentBowl, Wd40Can and WoodRollingPin. A set of sample images from each category of SIVAL database is shown in Figure-4.3.

4.1.3 The Airport Dataset

An airport database is created by using satellite images captured by GeoEye-1. The database includes 5 multisprectral airport images at resolution of 50 cm. The original images supplied by GeoEye-1 are four band (Red, Green, Blue and Infared), pansharped and 16 bits. However, we use three bands (Red, Green and Blue) and quantize the images into 8 bits. The spatial resolutions of images are approximatly 8000x10000 pixels. The goal of experiments on satellite images is to localize the airplanes in the airport areas. For this purpose, first, the airport areas in the images are detected. Then, by using the mask, the search is applied to just the airport area. In Figure-4.4, a sample airport image and the corresponding mask image used for retrieving the airplanes shown.

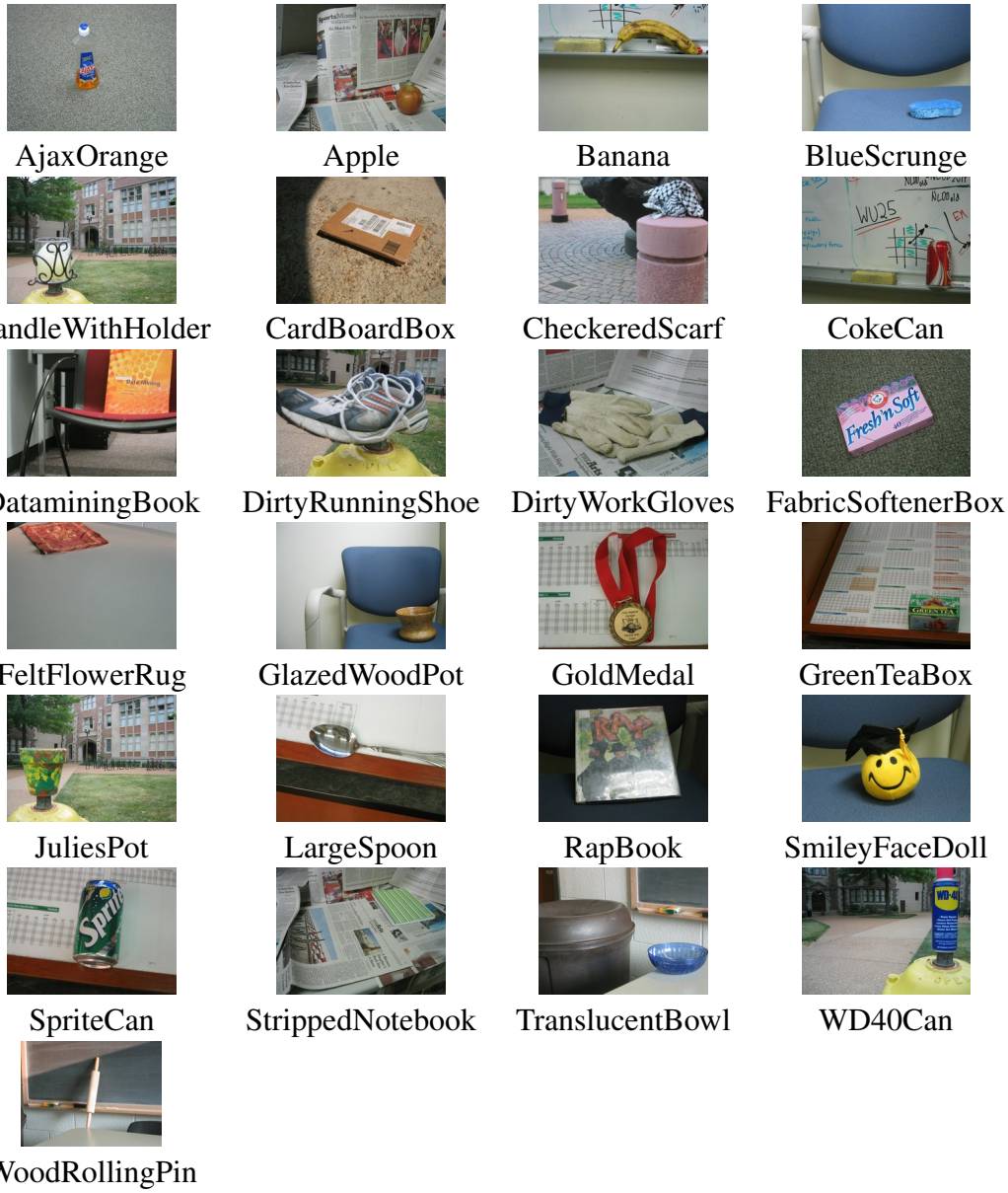


Figure 4.3: Sample Images from SIVAL dataset [12].

(a) (b)

Figure 4.4: (a) A sample airport image (b) corresponding mask image

4.2 Experiments with the STIM dataset

The STIM dataset with of 112 images may be considered rather small. It is employed to test and validate the proposed approaches and algorithms of our system in detail. Recall that, the proposed ABIR system has four main steps which are saliency map computation, region extraction, feature extraction from salient regions and similarity matching as shown in Figure-4.5. We analyze the all steps of the system separately with a set of experiments.

In the first group of experiments, the effect of the proposed algorithms for saliency map extraction is discussed. We have three major visual attention algorithms based on Itti-Koch model which are specific to CBIR problem to compute saliency map of images: Bottom-up normalization, Top-down Normalization and Top-Down Feature Weighting. The performance evaluation of visual attention models is based on the purpose of implementation. The output of any visual attention model is a saliency map which indicates the salient or interested regions in an image. The eye fixation experiments on human subjects are used to indicate how much a visual attention model simulates human vision perception. Our motivation is to utilize the advantageous of computational visual attention, such as extracting salient regions by not employing segmentation, in CBIR problem. Therefore, we evaluate the performance of our proposed saliency map computing algorithms with the retrieval results of ABIR system.

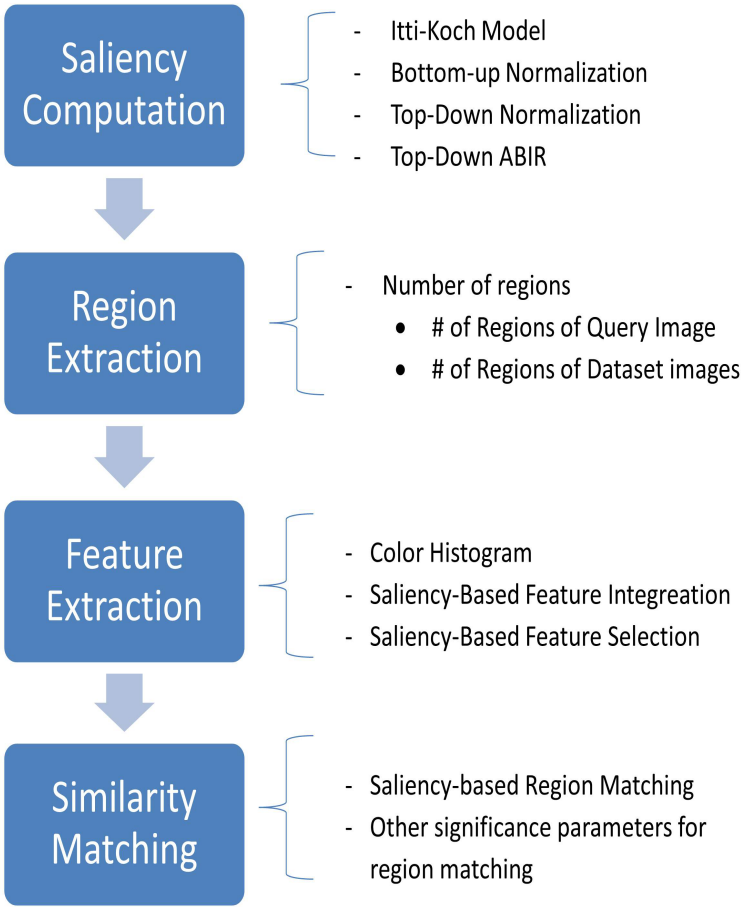


Figure 4.5: Setup for experiments in STIM dataset

In the next set of experiments, the effect of the number of regions extracted from saliency maps on the proposed ABIR system is tested and results are discussed. The number of regions which represents an image is fixed according to the database in most of the CBIR systems. In these experiments, we investigate the relation between the number of region and saliency of image.

The following set of experiments for features are employed in feature extraction step: Color Histogram, Saliency-based Feature Integration and Saliency-based Feature Selection. Finally, a set of experiments is employed to analyze the similarity matching step of ABIR system. We propose a similarity matching algorithm which utilizes saliency value of regions, based on Integrated Region Matching (IRM). In this set of experiments, we test the different significant values (p) with the proposed Similarity matching algorithm such as $p = AV, IPFxRForAP$.

We have four set of experiments as described above. In fact, we employ all steps of ABIR system in each set of experiments while analyzing effect of only one step with proposed algorithms. For this purpose, we define each experiment with steps as a set for the sake of clarity of experiments such as:

Experiment Set = {Saliency-Computation, Region-Extraction, Feature-Extraction, Similarity-Matching}.

where

Saliency-Computation $\in$ *{Itti-Koch model,Bottom-up Normalization, Top-down Normalization, Top-down Feature Weighting},*

Region-Extraction indicates the number of salient regions in query image (QR) and the number of regions in database images (DR),

Feature-Extraction $\in$ *{Color Histogram, Saliency-based Feature Integration, Saliency-based Feature Selection}*,

Similarity-Matching indicates the significance values, p $\in$ *{AV, AP, IPFxRF}.*

A sample experiment set, Experimemt-Set1, which analysis the experimental results of Bottom-up Normalization and Top-down Normalization is given in below:

Experiment-Set1:{

Saliency Computation*: 'Bottom-Up Normalization', 'Top-down Normalization'*

Region Extraction: QR = 5*, DR* = 5

Feature Extraction: 'Color Histogram'

Similarity Matching: p = AV}.

The bold step shows the algorithms which is under the consideration with this set of experiments. In this step, we employ the ABIR system two times; once for Bottom-up Normalization and once for Top-down Normalization. The other three steps uses

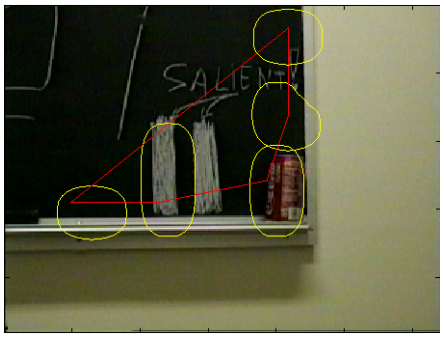


Figure 4.6: The salient regions extracted from saliency map of the image repeat after 5 salient regions are extracted

the same algorithms and parameters: number of regions are 5, Color Histogram for feature extraction and $p = AV$ for similarity matching.

Recall that, we use the work of Walter and Itti [88] to extract salient regions of images from the saliency maps. The fixation points are computed by using winner-take-all network and inhibition of return of previously attended regions as explained in Chapter-3. However, the work of Walter and Itti does not suggest any method to determine the number of fixations, in another words number of salient regions. The number of salient regions in images depends on the resolution of the images, and on the amount of visual information. In low-resolution images with few objects, three fixations may be sufficient to cover the relevant parts of the image. In high-resolution images with a lot of visual information, up to 30 fixations are required to sequentially attend to most or all object regions.

The salient regions extracted from the saliency map of images are repeated after a number of regions are extracted. The number of salient regions in an image can be determined by monitoring the saliency map to see after how many fixations the serial scanning of the saliency map starts to cycle for a few typical examples from the dataset. Cycling of fixation points usually occurs with 4-7 regions when the salient regions have covered approximately 30-50% of the image area for STIM dataset. For example, the salient regions are start to cycle in the image which is shown in Figure-4.6. We use the same number of fixation points for all images in an image set to ensure consistency throughout the respective experiment.

The results of experiments are presented with the precision and recall measures. Pre-

cision is the ratio of relevant images to total number of images retrieved in the query. Recall is ratio of images retrieved in a query to total number of images in the database. Mathematically,

$$precision = \frac{number\ of\ relevant\ images\ retrieved}{total\ number\ of\ images\ retrieved}, \tag{4.1}$$

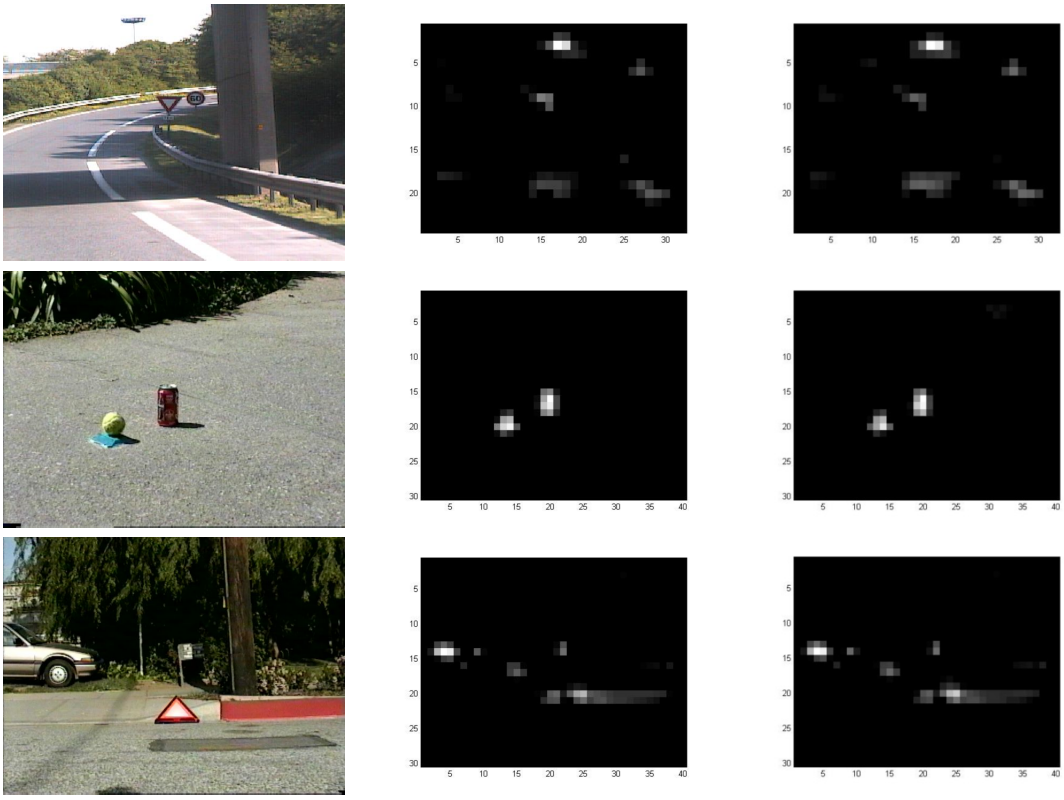
$$precision = \frac{number\ of\ relevant\ images\ retrieved}{total\ number\ of\ relevant\ images\ in\ dataset}. \tag{4.2}$$

In the following sub-sections, the experiment sets for saliency map extraction, effects of number of salient regions, feature extraction and similartiy matching are given in detail with the precision and recall results.

4.2.1 Saliency Map Extraction

An important step of the suggested ABIR system is the saliency map computing. The salient regions which represent the images are extracted from saliency map of images. The goal of extracting salient regions is to localize the objects of interest in images. Meanwhile, salient region extraction enable us to eliminate background clutter in the images. Therefore, the saliency map of an image affects the performance of the suggested ABIR systems significantly. Itti-Koch model considers only a single image to extract saliency map. On the other hand, we adopt this model for content-based image retrieval problem, where the saliency map of an image in the database varies according to the query image or other images in the database based on whole system.

Our first algorithm for saliency map extraction is called bottom-up normalization. The idea behind the bottom-up normalization is to consider all dataset images as one image during the normalization operation of the feature maps. Therefore, we define a global normalization parameter, N_{BU}, based on the normalization method described in Itti-Koch model. The normalization operator N_{BU} computes the global maxima of feature maps by using all feature maps of images in the database. Note that, in this approach, the query image is kept outside the database. Figure-4.7 shows saliency maps of sample dataset images. The second colon indicates saliency maps extracted by the original Itti-Koch model. The third colon is the saliency maps extracted by



(a) (b) (c)

Figure 4.7: (a) original images, (b) Saliency Map from Itti.Koch Model (c) Saliency map from Bottom-up Normalization.

using suggested bottom-up normalization. As it is seen from Figure-4.7, the saliency map extracted at the output of the bottom-up normalization algorithm is quite different than Itti-Koch saliency map. The saliency values of regions are decreased but new regions appeared in saliency map of bottom-up normalization created by the normalization operator N_{BU}. For example, the car in the second image is not in the region of interest, because the category of image is emergency triangles. The region corresponding to the car is salient in the saliency map. However, the car region is less salient in the bottom-up saliency map. Then approach creates an advantage in retrieval step to detect region of interest according to the object categories under consideration. The newly appeared regions does not cause any problem, because their saliences are very low compared to the most salient regions.

The second algorithm for saliency map extraction is called top-down normalization algorithm. In this case, the normalization parameters of feature maps of images in the database are computed by using global maxima of feature maps of the current query

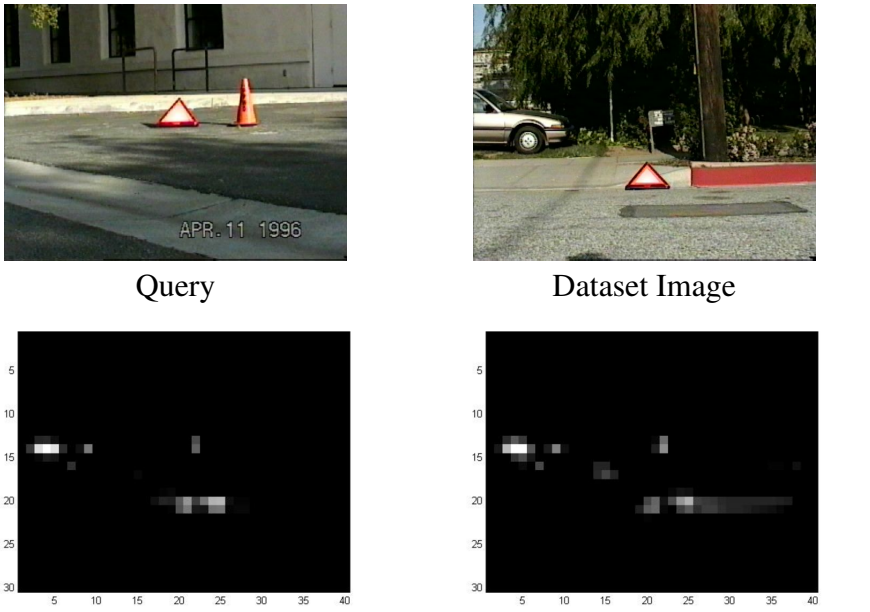


Figure 4.8: Saliency maps of Dataset Images according to the Query Image

image as described in Chapter 3. We define normalization operator N_{TD} to normalize feature maps of images in the database by using query image.

The third algorithm for saliency map extraction is called the top-down feature map weighting algorithm, where the feature maps which are used to compute the saliency map are weighted by using the query image. Note that, the saliency map of an image in the database is computing according to the query image for both top-down normalization and top-down feature map weighting algorithms. In Figure-4.8, the saliency maps which are computed by using top-down normalization and top-down feature map weighting algorithms are shown. The first column shows the query image, and the next image is the database image. At the second row of figure, saliency maps of the dataset image is shown. As it is seen from the Figure-4.8, the saliency values are decreased with respect to the maxima of feature maps of query image when using Top-Down Normalization. The topography of saliency map which is computed by using Top-Down Feature Weighting is similar to the saliency maps of Itti-Koch and Bottom-Up normalization.

How to evaluate the performance of a visual attention models is challenging problem in computer vision literature. A method is used human experts to test the output of

Table 4.1: The average precision results of Experiment Set_{SM} for the first 5 retrieved images by testing Saliency Map extraction algorithms.

	Autobahn	Coke-Can	Triangle	Average
Itti-Koch	0.88	0.64	0.56	0.69
Bottom-up Normalization	0.96	0.64	0.68	0.76
Top-Down Normalization	0.84	0.76	0.56	0.72
Top-Down Feature Map Weighting	0.96	0.72	0.60	0.76

models with eye fixation experiments. However, lack of this type of setup, we evaluate our proposed saliency map computing algorithms with the retrieval performance of the proposed ABIR system. We apply a set of experiments to indicate how the saliency maps extracted from Itti-Koch, Bottom-up Normalization, Top-down normalization or Top-down Feature Weighting algorithms effects the performance of the attention-based image retrieval system. The following experiment set is used :

Experiment Set_{SM}*: {*

> ***Saliency Computation: 'Itti-Koch', 'Bottom-Up Normalization'*** 'Top-down Normalization', 'Top-down Feature Weighting'
>
> *Region Extraction: QR=5, DR=5*
>
> *Feature Extraction: 'Color Histogram'*
>
> *Similarity Matching: 'p=AV'}.*

The color histogram extracted from salient regions is a 125-bin RGB color histogram. The results are interpreted by using average precision and recall-precision curve. Table-4.1 shows the average precision results for the first five retrieved images according to the categories for experiment Set_{SM}. Furthermore, precision-recall curves for each algorithm are given in Figure-4.9,Figure-4.10 and Figure-4.11

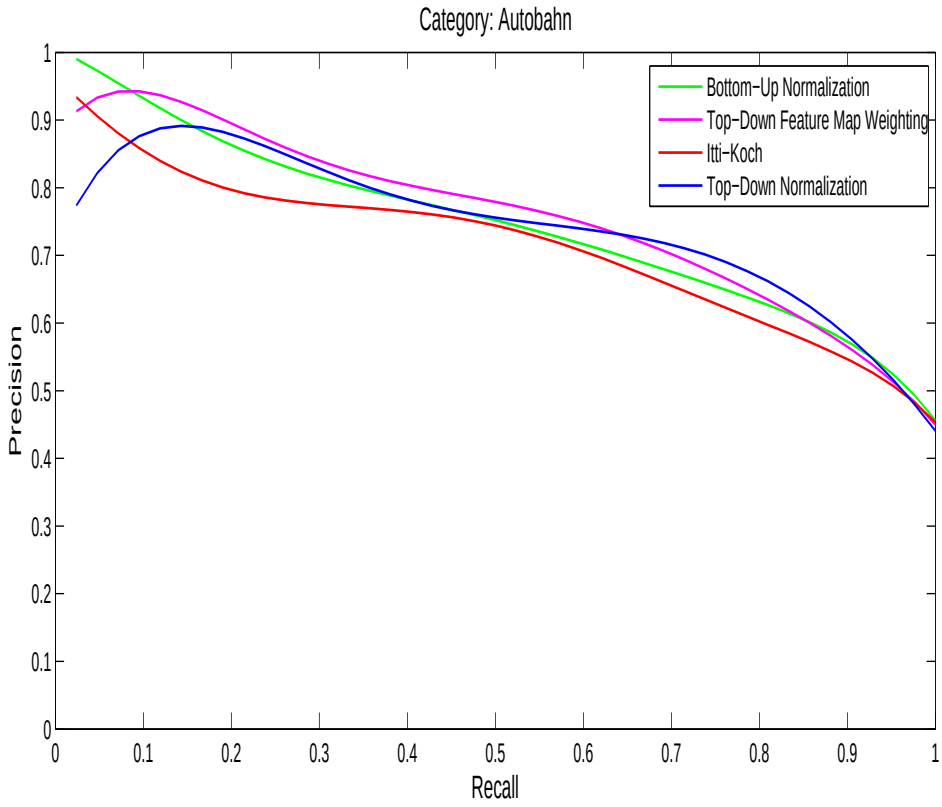


Figure 4.9: Recall-Precision Curve for Experiment Set_{SM}. Results for category Autobahn.

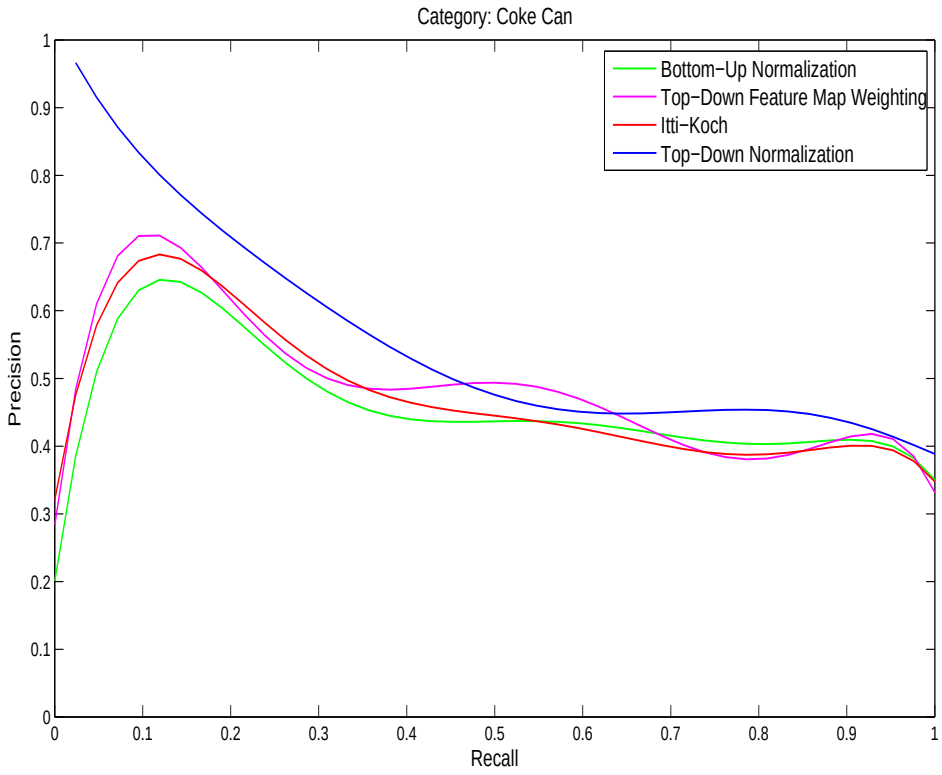


Figure 4.10: : Recall-Precision Curve for Experiment Set_{SM}. Results for category Coke can.

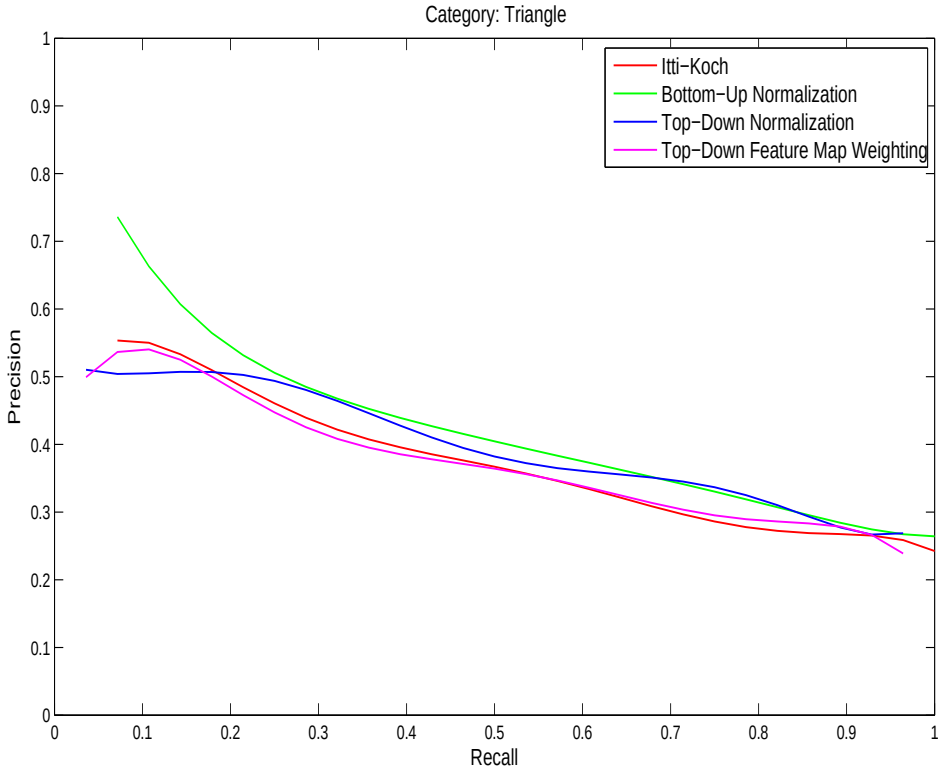


Figure 4.11: : Recall-Precision Curve for Experiment Set_{SM}.Results for category Triangle.

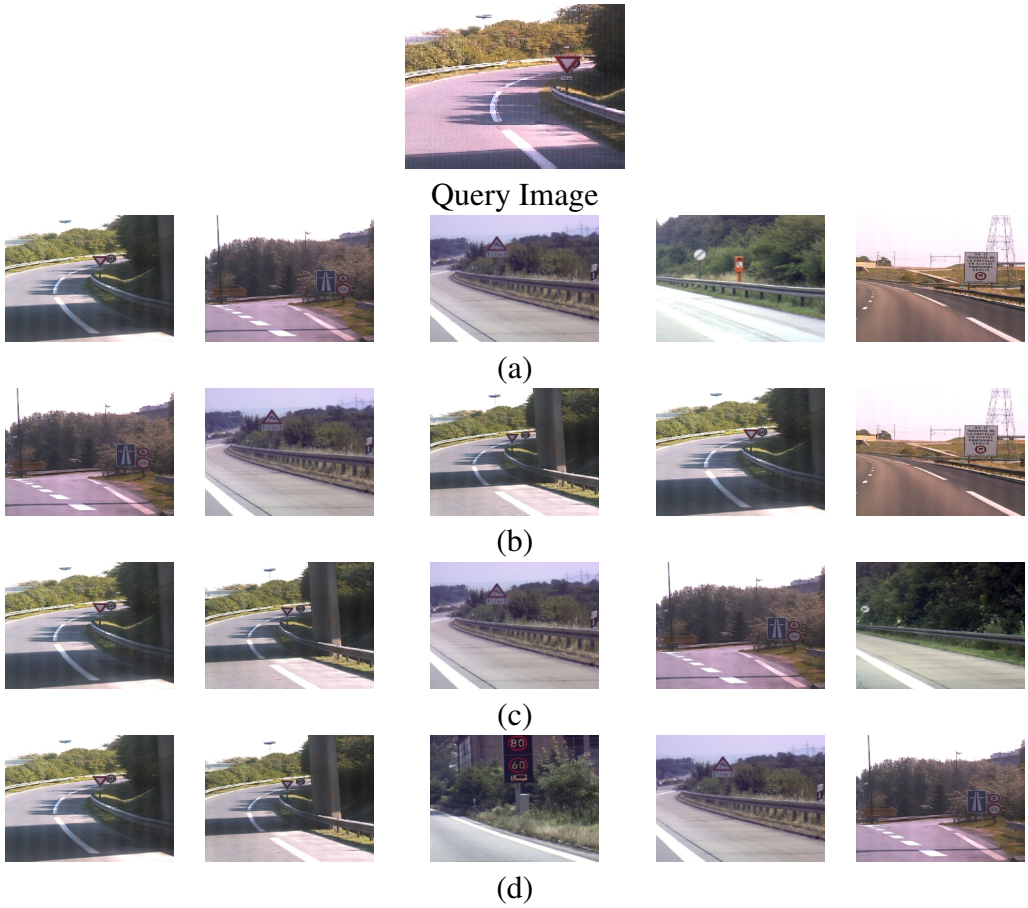


Figure 4.12: The retrieval results of Experiment Set_{SM} for a query image from the Autobahn category (a) Itti-Koch (b) Bottom-Up Normalization (c) Top-down Normalization (d) Top-down Feature Weighting.

The result of Experiment Set_{SM} shows that proposed ABIR system which employs Bottom-up Normalization or Top-down Normalization or Top-down Feature Weighting saliency map computing algorithms give better results compared to the ABIR system which employs Itti-Koch model. Bottom-Up Normalization and Top-Down Feature Weighting give better results for the Autobahn category than the Itti-Koch and Top-Down Normalization. However, the results with all saliency computing algorithms are satisfactory for Autobahn category. The images in the Autobahn category usually contain more than one object, different road signs. More than one region contributes to represent the images in autobahn category. In Figure-4.12, the retrieval results for a query image from the Autobahn category is shown for four saliency map computing algorithms.

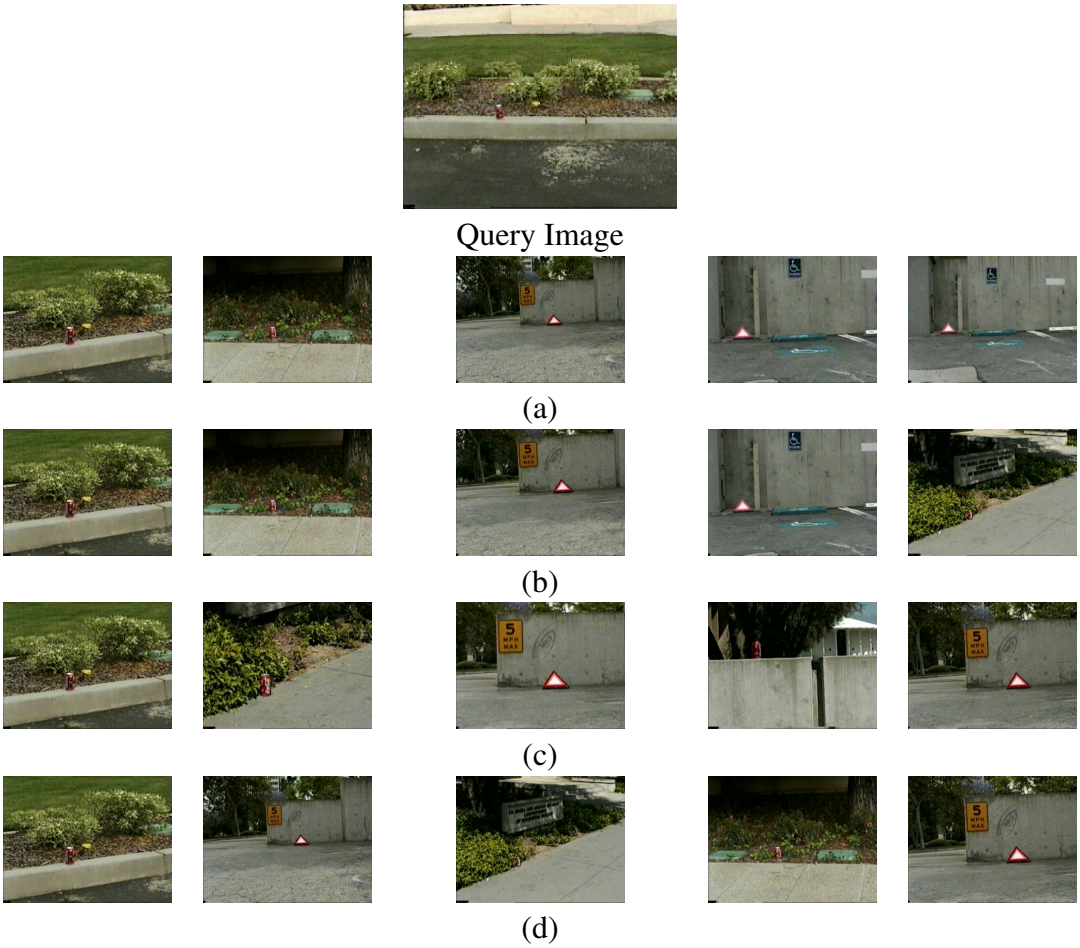


Figure 4.13: The retrieval results of Experiment Set_{SM} for a query image from the Coke Can category (a) Itti-Koch (b) Bottom-Up Normalization (c) Top-down Normalization (d) Top-down Feature Weighting.

ABIR system with Top-Down Normalization give better results compared the other saliency map computing algorithms for Coke Can category. The coke can in the images are visually dominant according to the other regions in the image. However, the normalization operator N_{TD} which utilizes the query image, decreases the saliency of distracter regions in the images. This effect increases the performance results of the ABIR system. In Figure-4.13, the retrieval results for a query image from the Coke Can category is shown for four saliency map computing algorithms.

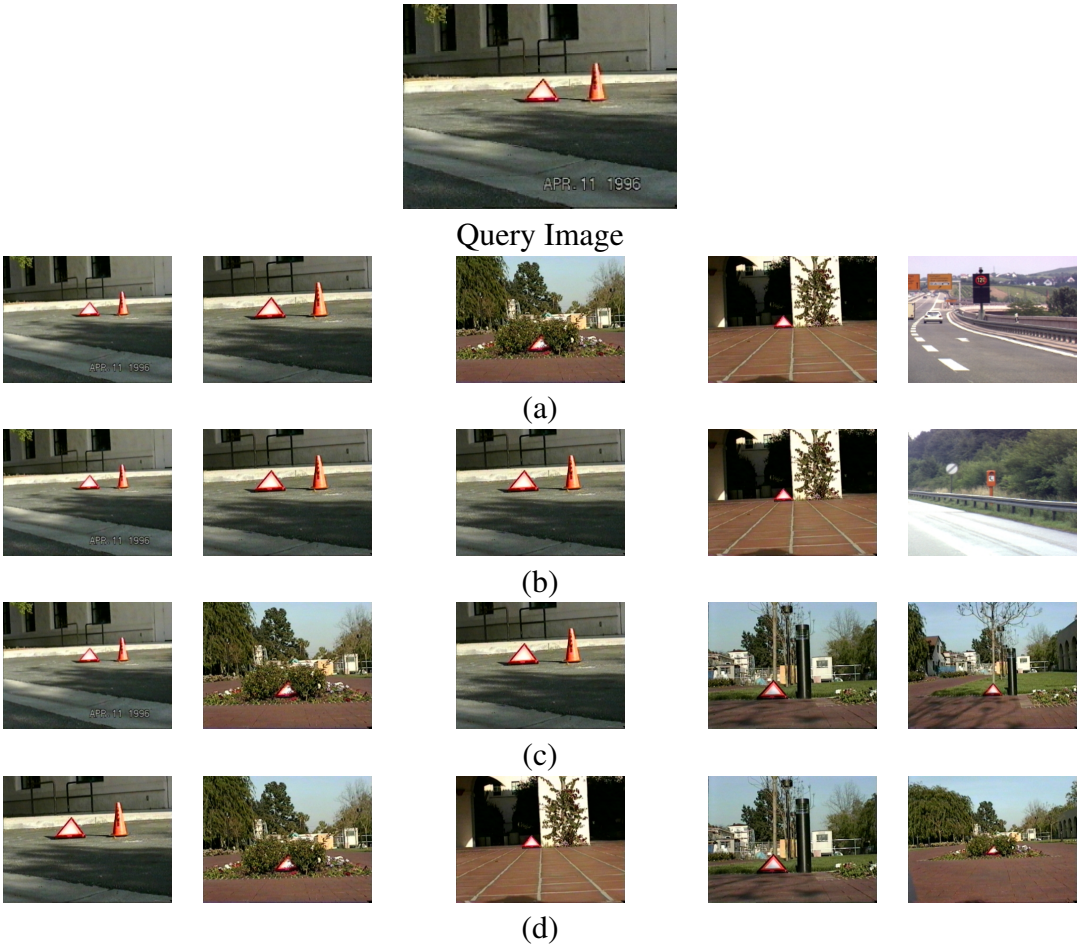


Figure 4.14: The retrieval results of Experiment Set_{SM} for a query image from the Triangle category (a) Itti-Koch (b) Bottom-Up Normalization (c) Top-down Normalization (d) Top-down Feature Weighting.

When we analyze the results for triangle category, we observe less intuitive results than the other two categories. The main reason for this poor results is that the triangle object is also existed in autobahn category. Therewith, there are different objects which have salient regions such as car or dustbin in images of this category. The bottom-up normalization gives a higher retrieval performance compared to other normalization algorithms for this category. The salient regions which do not contain a triangle are rarely existed in images. Therefore, bottom-up normalization can decrease the saliency of these regions by considering all the images in the database. In Figure-4.14, the retrieval results for a query image from the triangle category is shown for all the saliency map computation algorithms.

4.2.2 Analyzing the Number of Salient Regions

The method for extracting the salient regions from the saliency maps are adopted from [88] as explained in Chapter 3. In this section, we analyze the effect of the number of salient regions on the performance of the ABIR system for four saliency map extraction algorithms, namely Itti-Koch, Bottom-up Normalization, Top-down Normalization and Top-down Feature Map Weighting. These experiments are performed by varying the number of regions for query and dataset images between 1 to 5. The following set of experiments are used:

Experiment Set_{RN}*: {*

> ***Saliency Computation****: 'Itti-Koch', 'Bottom-Up Normalization' 'Top-down Normalization', 'Top-down Feature Weighting'*
>
> ***Salient Region Extraction****:* $QR = 1, 2, \cdots, 5$ *and* $DR = 1, 2, \cdots, 5$
>
> *Feature Extraction: 'Color Histogram'*
>
> *Similarity Matching: 'p=AV'}.*

During the experiment Set_{RN}, we analyze the number of regions for each proposed saliency map computing algorithms separately. The average precision results for the first five retrieved images are tabulated in the Table-4.2 for Itti-Koch, Table-4.3 for Bottom-Up Normalization, Table-4.4 for Top-Down Normalization and Table-4.5 for Top-Down Feature Weighting. As it can be observed from the tables, the number of extracted regions highly affects the retrieval results of ABIR system.

The highest precision result is achieved for Autobahn category, when the number of extracted regions is 4 or 5 for Itti-Koch model. The number of regions for Coke Can category is 2-3 for the higher precision results. The highest precision results for Triangle category is given when the number of region of dataset images is 4 and the number of region of query images is 5 as seen from Table-4.2.

The Bottom-Up normalization gives the higher precision result for Autobahn category when the number of region of dataset images is 5 and the number of region of query

Table 4.2: Prescision (Pre) and Recall (Rec) results of ABIR system for Itti-Koch Model.

(a) Category : Autobahn

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.36	0.04	0.40	0.05	0.56	0.07	0.52	0.06	0.60	0.07
DR=2	0.44	0.05	0.68	0.08	0.80	0.10	0.84	0.10	0.76	0.09
DR=3	0.48	0.06	0.68	0.08	0.84	0.10	0.80	0.10	0.80	0.10
DR=4	0.6	0.07	0.88	0.10	0.92	0.11	0.88	0.10	0.92	0.11
DR=5	0.72	0.09	0.84	0.10	0.92	0.11	0.88	0.10	0.88	0.10

(b) Category : Coke Can

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.72	0.09	0.84	0.1	0.8	0.1	0.8	0.1	0.84	0.1
DR=2	0.76	0.09	0.84	0.1	0.76	0.09	0.72	0.09	0.64	0.08
DR=3	0.72	0.09	0.88	0.1	0.84	0.1	0.68	0.08	0.76	0.09
DR=4	0.6	0.07	0.68	0.08	0.68	0.08	0.64	0.08	0.6	0.07
DR=5	0.6	0.07	0.84	0.1	0.76	0.09	0.72	0.09	0.64	0.08

(c) Category : Triangle

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.28	0.05	0.28	0.05	0.56	0.1	0.48	0.09	0.44	0.08
DR=2	0.28	0.05	0.4	0.07	0.48	0.09	0.48	0.09	0.48	0.09
DR=3	0.36	0.06	0.48	0.09	0.52	0.09	0.44	0.08	0.44	0.08
DR=4	0.44	0.08	0.6	0.11	0.64	0.11	0.6	0.11	0.68	0.12
DR=5	0.48	0.09	0.52	0.09	0.52	0.09	0.56	0.1	0.56	0.1

(d) Average

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.45	0.06	0.51	0.07	0.64	0.09	0.6	0.08	0.63	0.08
DR=2	0.49	0.06	0.64	0.08	0.68	0.09	0.68	0.09	0.63	0.08
DR=3	0.52	0.07	0.68	0.09	0.73	0.1	0.64	0.08	0.67	0.09
DR=4	0.55	0.07	0.72	0.1	0.75	0.1	0.71	0.1	0.73	0.1
DR=5	0.6	0.08	0.73	0.1	0.73	0.1	0.72	0.1	0.69	0.09

Table 4.3: Prescision (Pre) and Recall (Rec) value for Bottom-Up Normalization Algorithm.

(a) Category : Autobahn

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.36	0.04	0.52	0.06	0.56	0.07	0.56	0.07	0.64	0.08
DR=2	0.36	0.04	0.76	0.09	0.76	0.09	0.76	0.09	0.76	0.09
DR=3	0.40	0.05	0.80	0.10	0.84	0.10	0.88	0.10	0.88	0.10
DR=4	0.56	0.07	0.80	0.10	0.92	0.11	0.88	0.10	0.84	0.10
DR=5	0.56	0.07	0.88	0.10	0.92	0.11	0.92	0.11	0.96	0.11

(b) Category : Coke Can

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.76	0.09	0.84	0.10	0.80	0.10	0.84	0.10	0.84	0.10
DR=2	0.72	0.09	0.92	0.11	0.80	0.10	0.72	0.09	0.72	0.09
DR=3	0.72	0.09	0.76	0.09	0.76	0.09	0.68	0.08	0.72	0.09
DR=4	0.68	0.08	0.80	0.10	0.72	0.09	0.72	0.09	0.72	0.09
DR=5	0.64	0.08	0.76	0.09	0.68	0.08	0.64	0.08	0.64	0.08

(c) Category : Triangle

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.32	0.06	0.40	0.07	0.44	0.08	0.40	0.07	0.44	0.08
DR=2	0.20	0.04	0.44	0.08	0.40	0.07	0.28	0.05	0.28	0.05
DR=3	0.32	0.06	0.52	0.09	0.60	0.11	0.48	0.09	0.44	0.08
DR=4	0.36	0.06	0.56	0.10	0.64	0.11	0.56	0.10	0.56	0.10
DR=5	0.48	0.09	0.56	0.10	0.60	0.11	0.68	0.12	0.68	0.12

(d) Average

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.48	0.06	0.59	0.08	0.60	0.08	0.60	0.08	0.64	0.09
DR=2	0.43	0.05	0.71	0.09	0.65	0.09	0.59	0.07	0.59	0.09
DR=3	0.48	0.06	0.69	0.09	0.73	0.10	0.68	0.08	0.68	0.09
DR=4	0.53	0.07	0.72	0.10	0.76	0.10	0.72	0.08	0.71	0.09
DR=5	0.56	0.08	0.73	0.10	0.73	0.10	0.75	0.08	0.76	0.10

Table 4.4: Prescision (Pre) and Recall (Rec) value for Top-Down Normalization Algorithm.

(a) Category : Autobahn

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.32	0.04	0.48	0.06	0.48	0.06	0.44	0.05	0.4	0.05
DR=2	0.4	0.05	0.6	0.07	0.64	0.08	0.6	0.07	0.68	0.08
DR=3	0.6	0.07	0.7 2	0.09	0.84	0.10	0.84	0.10	0.84	0.10
DR=4	0.56	0.07	0.84	0.10	0.92	0.11	0.92	0.11	0.84	0.10
DR=5	0.6	0.07	0.84	0.10	0.88	0.10	0.92	0.11	0.84	0.10

(b) Category : Coke Can

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.88	0.10	0.80	0.10	0.8	0.10	0.88	0.10	0.88	0.10
DR=2	0.64	0.08	0.76	0.09	0.68	0.08	0.8	0.10	0.84	0.10
DR=3	0.64	0.08	0.84	0.10	0.72	0.09	0.72	0.09	0.76	0.09
DR=4	0.64	0.08	0.80	0.10	0.72	0.09	0.6	0.07	0.76	0.09
DR=5	0.6	0.07	0.84	0.10	0.68	0.08	0.68	0.08	0.76	0.09

(c) Category : Triangle

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.28	0.01	0.28	0.01	0.32	0.01	0.36	0.01	0.36	0.01
DR=2	0.48	0.01	0.60	0.02	0.64	0.02	0.56	0.02	0.60	0.02
DR=3	0.28	0.01	0.52	0.02	0.56	0.02	0.56	0.02	0.60	0.02
DR=4	0.28	0.01	0.56	0.02	0.52	0.02	0.64	0.02	0.60	0.02
DR=5	0.28	0.01	0.48	0.02	0.40	0.02	0.52	0.02	0.56	0.02

(d) Average

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.49	0.05	0.52	0.05	0.53	0.05	0.56	0.06	055	0.05
DR=2	0.51	0.04	0.65	0.06	0.65	0.06	0.65	0.06	071	0.07
DR=3	0.51	0.05	0.69	0.07	0.71	0.07	0.71	0.07	073	0.07
DR=4	0.49	0.05	0.73	0.07	0.72	0.07	0.72	0.07	073	0.07
DR=5	0.49	0.05	0.72	0.07	0.65	0.07	0.71	0.07	072	0.07

Table 4.5: Prescision (Pre) and Recall (Rec) value for Top-Down Feature Map Weighting Algorithm.

(a) Category : Autobahn

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.40	0.05	048	0.06	0.60	0.07	0.56	0.07	0.60	0.05
DR=2	0.52	0.06	064	0.08	0.84	0.10	0.84	0.10	0.84	0.08
DR=3	0.52	0.06	084	0.10	0.88	0.10	0.88	0.10	0.88	0.10
DR=4	0.80	0.10	080	0.10	0.96	0.11	0.92	0.11	0.96	0.10
DR=5	0.84	0.10	092	0.11	0.96	0.11	1	0.12	0.96	0.10

(b) Category : Coke Can

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.72	0.09	0.80	0.10	0.72	0.09	0.68	0.08	0.76	0.09
DR=2	0.80	0.10	0.88	0.10	0.76	0.09	0.76	0.09	0.72	0.09
DR=3	0.72	0.09	0.96	0.11	0.92	0.11	0.76	0.09	0.76	0.09
DR=4	0.76	0.09	0.76	0.09	0.76	0.09	0.68	0.08	0.76	0.09
DR=5	0.60	0.07	0.84	0.10	0.76	0.09	0.60	0.07	0.72	0.09

(c) Category : Triangle

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.24	0.04	0.32	0.06	0.48	0.09	0.44	0.08	0.36	0.06
DR=2	0.36	0.06	0.40	0.07	0.48	0.09	0.48	0.09	0.32	0.06
DR=3	0.36	0.06	0.48	0.09	0.56	0.10	0.44	0.08	0.44	0.08
DR=4	0.44	0.08	0.56	0.10	0.52	0.09	0.48	0.09	0.52	0.09
DR=5	0.48	0.09	0.56	0.10	0.68	0.12	0.52	0.09	0.60	0.11

(d) Average

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.45	0.06	0.53	0.07	0.60	0.08	0.56	0.08	0.57	0.08
DR=2	0.56	0.07	0.64	0.08	0.69	0.09	0.69	0.09	0.63	0.08
DR=3	0.53	0.07	0.76	0.10	0.79	0.10	0.69	0.09	0.69	0.09
DR=4	0.67	0.09	0.71	0.10	0.75	0.10	0.69	0.09	0.75	0.10
DR=5	0.64	0.09	0.77	0.10	0.80	0.11	0.71	0.09	0.76	0.10

images is 5. The highest precision result for Coke Can category is given when the number of region of dataset images is 2 and the number of region of query images is 2. The highest precision result for Triangle category is given when the number of region of dataset images is 4 or 5 and the number of region of query images is 5 as seen from Table-4.3.

The Top-Down normalization gives the highest precision result for Autobahn category, when the number of region of dataset images is 4 or 5 and the number of region of query images is 4. The highest precision result for Coke Can category is given when the number of region of dataset images is 1 and the number of regions of query images does not affect the results so much. The highest precision result for Triangle category is given when the number of region of dataset images is 4 and the number of region of query images is 4as seen from Table-4.4.

When we analyze the results for Top-Down Feature Weighting, the best results for all categories are given by this algorithm. It gives the highest precision result for Autobahn category when the number of region of dataset images is 5 and the number of region of query images is 4. The highest precision result for Coke Can category is given when the number of region of dataset images is 3 and the number of regions of query images is 2. The highest precision result for Triangle category is given when the number of region of dataset images is 5 and the number of region of query images is 3 as seen from Table-4.5.

Let us explore the tables, precision results are as expected for each categories according to the properties of images. The better precision results are given with high number of regions for Autobahn category: the background regions are important for this category. For example, when we extract only one region both from query and dataset images the average precision is 0.4 for Top-Down Feature Weighting. But the result is dramatically changed to 1.0 when the number of region of dataset images is 5 and the number of region of query images is 4. The highest precision results are given for Coke Can category when the numbers of regions are decreased: there is only one object of interest in all images which is a red coke can.

All the results which are discussed here indicate that: the number of regions extracted from the images affects performance of the proposed ABIR system. Therefore, we

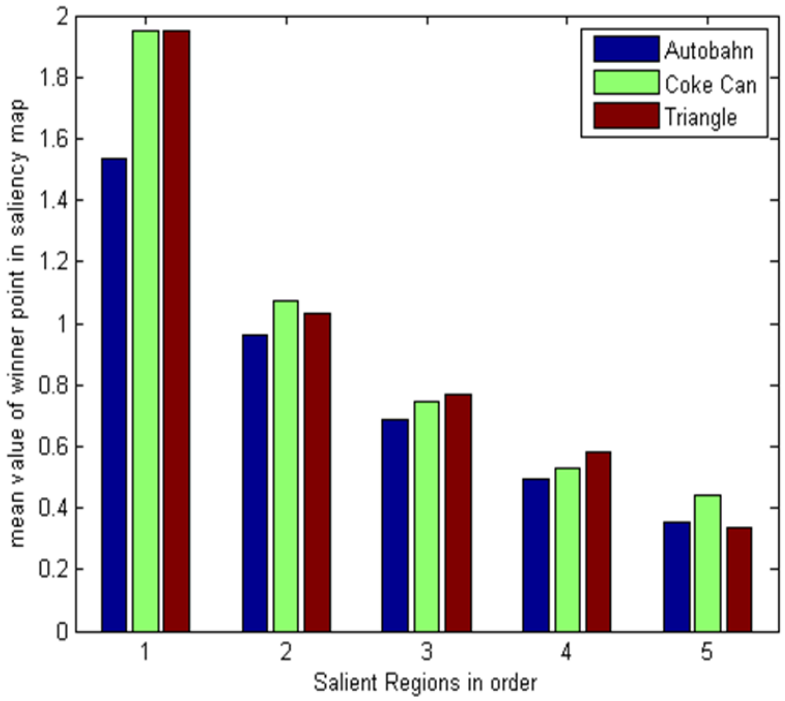


Figure 4.15: Salient regions ordered between the most salient and least salient regions

analyze the saliency value of each region according to the categories as shown in Figure-4.15. The saliency value of each region is the winner point of the saliency map for that region. In other words, saliency value of each region is the pixel value of fixation points in the saliency map. As it is shown from the Figure-4.15, there is no characteristic change in the values to determine number of regions extracted from saliency map of images. Hence, all regions are salient according to the saliency map and we cannot differentiate salient regions which are from the object of interest or background by using saliency values.

4.2.3 Time Complexity Analysis of Saliency Map Computation and Region Extraction

Computational complexity of a computer vision algorithm is important for real-life applications. In this section, we analysis the computation time of algorithms for saliency map computation and region extraction together. The average time required to compute saliency map and region extraction for algorithms is shown in Table-4.6.

Table 4.6: Average Saliency Map computation and region extraction times of an image for algorithms in seconds for different number of salient regions (DR)

	DR=1	DR=3	DR=5
Itti-Koch	1.61	1.90	2.20
Bottom-up Normalization	3.14	3.28	3.66
Top-Down Normalization	1.62	1.82	2.41
Top-Down Feature Map Weighting	1.63	1.88	2.38

Average time is calculated over all 112 images in STIM dataset. All algorithms are executed on a computer running Windows7 with 8GB RAM and core i7 2.8 GHz Intel CPU. The Itti-Koch model is the fastest algorithm followed by top-down normalization algorithm and top-down feature map weighting algorithm. However, bottom-up normalization algorithm is slow comparing to the other saliency computation algorithms. Bottom-up normalization algorithm computes the normalization parameters firstly, then compute the saliency map of the image. For this reason, computation time of bottom-up normalization algorithm is depends on the number of images in the dataset. Furthermore, number of salient regions affects the computation time of algorithms.

4.2.4 Analysing Feature Extraction from Salient Regions

After the regions are extracted from saliency maps, the next step is to extract features from each salient regions. A region can be represented with a variety of features, among the color, texture and shape feature. How to select and combine these features is an important problem in computer vision applications. The general idea in literature is to concatenate all the feature types to form a feature vector. We proposed two algorithms to solve feature space design problem by considering saliency of regions: Saliency-based Feature Integration and Saliency-based Feature Selection. The idea behind the algorithms is to weight or select features according to the saliency value of the region. The details of algorithms are explained in Chapter 3. In this section, we perform a set of experiments to test the proposed algorithms. For this aim, the following experiment sets Set_{FE} are expressed as below:

Experiment Set_{FE}*: {*

Table 4.7: Feature Extraction precision results for the first 5 retrieved images in the database.

Features	Autobahn	Coke-Can	Triangle	Average
Color	0.80	0.60	0.24	0.71
Intensity	0.76	0.44	0.24	0.48
Orientation	0.68	0.48	0.80	0.65
Feature Concatenation	1	0.56	0.48	0.69
Feature Selection Algorithm (Algorithm 5)	0.96	0.68	0.68	0.76
Feature Weighting Algorithm (Algorithm 4)	0.96	0.72	0.60	0.76

(a)

Saliency Computation: 'Bottom-Up Normalization'

Region Extraction: QR=5, DR=5

***Feature Extraction**: ,'Feature Concatenation', 'Saliency-based Feature Integration' ,'Saliency-based Feature Selection'}.*

Feature Extraction step: Basic low-level features are selected for each region by considering feature maps of Bottom-up Normalization algorithms. Recall from the Chapter-3 that, features used in design of feature space for regions are:

Color: mean red, green and blue values of regions

Intensity: mean intensity value of regions

Orientation: entropy of regions

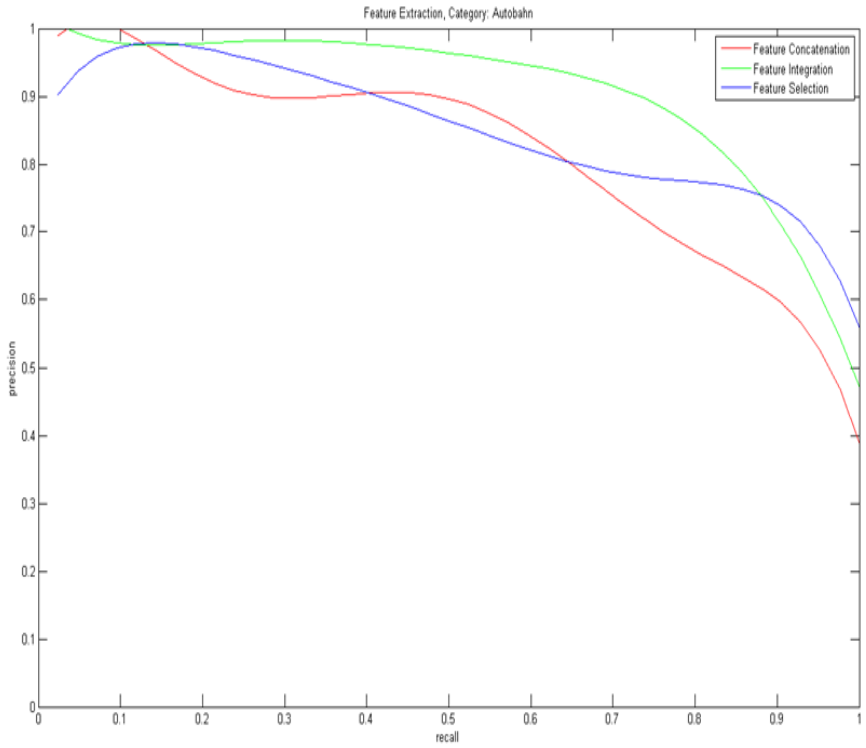


Figure 4.16: Precision-recall curve of feature extraction methods employed to ABIR system for Autobahn category

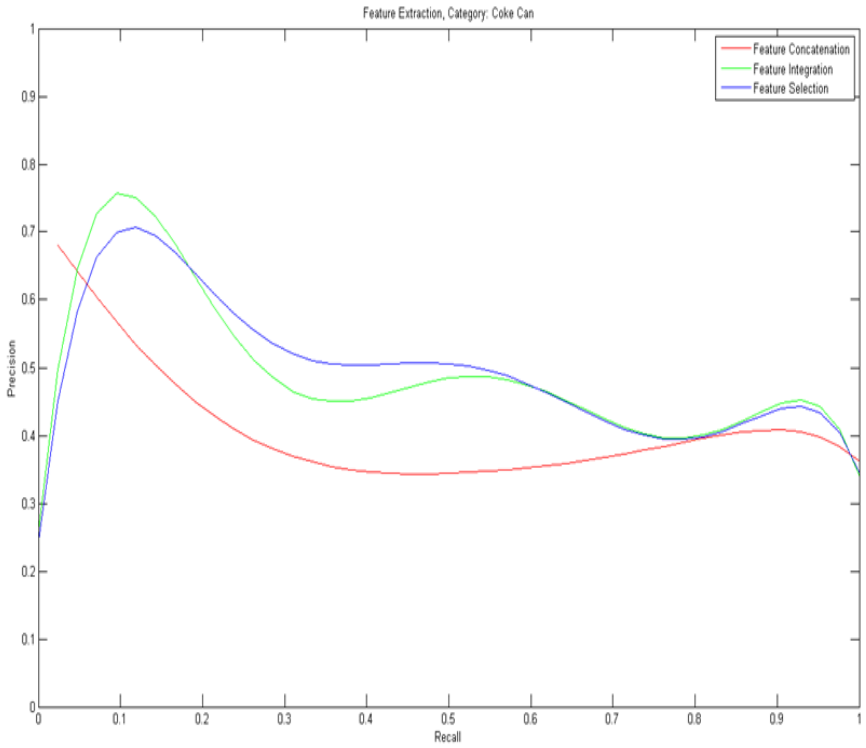


Figure 4.17: Precision-recall curve of feature extraction methods employed to ABIR system for Coke can category

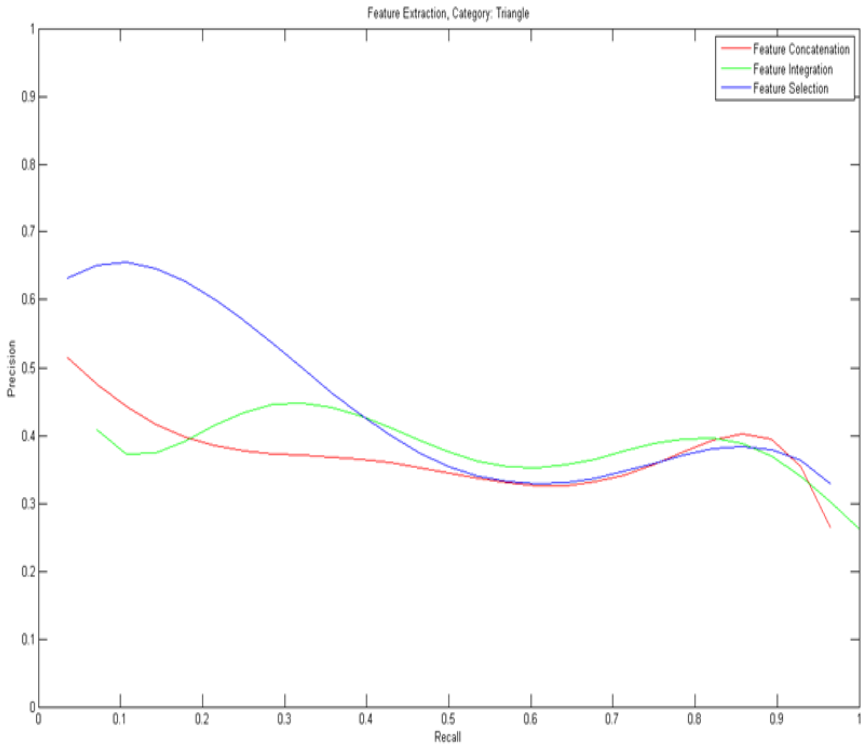


Figure 4.18: Precision-recall curve of feature extraction methods employed to ABIR system for Triangle category

Table-4.7 shows the average precision results of experiments for each category. Number of retrieved images is five. Firstly, color, intensity and orientation features are tested individually. Then, these features are concatenated in a feature vector named as Feature Concatenation in experiments shown in Table-4.7. Finally, the results of Saliency-based Feature Integration and Saliency-based Feature Selection algorithms are shown in the same table. The Recall-Precision curves for feature Concatenation, Feature Integration and Feature Selection is given for each category in Figure-4.16, Figure-4.17 and Figure-4.18. As it is seen from the Table-4.7 and figures, Feature Selection and Feature Integration algorithms give higher retrieval performance than the feature concatenation and using single feature types. Also, the results of Saliency-based Feature Integration and Saliency-based Feature Selection algorithms with different number of regions are tabulated in Table-4.8 and Table-4.9. The highest precision results are given by the proposed algorithms. The number of extracted regions affects the retrieval results as seen in Table-4.8 and Table-4.9. However, this effect is decreased compared with previous results especially for Autobahn category.

The aim of selecting basic features is to show the importance of using saliency values of regions when designing a feature space. The experimental results claim this idea, even though using rather basic feature types, such as average color, average intensity and entropy, promising retrieval results are given by the system. The more representative feature spaces can be designed by using the idea behind the proposed algorithms.

4.2.5 Similarity Matching of Salient Regions

The last step of the propose ABIR system is to match the similar salient regions. Similarity between two images is measured by the similarity of salient regions of between the query and database images. We define a similarity matching criteria based on IRM (Integrated Region Matching)[27]. We use saliency value of regions to compute a significance matrix as explained in Chapter 3. In this section, a set of experiments are done to test the system with a variety of significance criteria. For this aim, the following experiment set is realized:

Experiment Set$_{match}$*:* {

Table 4.8: Prescision (Pre) and Recall (Rec) value of the ABIR system with the Feature Integration Method

(a) Category : Autobahn

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.88	0.10	0.96	0.11	0.92	0.11	0.92	0.11	0.92	0.11
DR=2	0.96	0.11	0.92	0.11	0.96	0.11	1	0.12	1	0.12
DR=3	0.92	0.11	0.96	0.11	1	0.12	1	0.12	1	0.12
DR=4	0.88	0.11	0.96	0.11	0.96	0.11	0.96	0.11	0.96	0.11
DR=5	0.96	0.11	0.92	0.11	1	0.12	0.96	0.11	0.96	0.11

(b) Category : Coke Can

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.64	0.08	0.44	0.05	0.56	0.07	0.64	0.08	0.72	0.09
DR=2	0.52	0.06	0.64	0.08	0.56	0.07	0.52	0.06	0.72	0.09
DR=3	0.6	0.07	0.6	0.07	0.56	0.07	0.68	0.08	0.68	0.08
DR=4	0.48	0.06	0.4	0.05	0.48	0.06	0.48	0.06	0.56	0.07
DR=5	0.4	0.05	0.52	0.06	0.48	0.06	0.56	0.07	0.72	0.09

(c) Category : Triangle

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.28	0.05	0.36	0.06	0.48	0.09	0.4	0.07	0.4	0.07
DR=2	0.52	0.09	0.28	0.05	0.48	0.09	0.4	0.07	0.56	0.10
DR=3	0.44	0.08	0.44	0.08	0.44	0.08	0.48	0.09	0.6	0.11
DR=4	0.4	0.07	0.44	0.08	0.56	0.10	0.56	0.10	0.68	0.12
DR=5	0.4	0.07	0.36	0.06	0.36	0.06	0.36	0.06	0.6	0.11

(d) Average

	QR = 1		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.60	0.08	0.59	0.08	0.65	0.09	0.65	0.09	0.68	0.09
DR=2	0.67	0.09	0.61	0.08	0.67	0.09	0.64	0.08	0.76	0.10
DR=3	0.65	0.09	0.67	0.09	0.67	0.09	0.72	0.10	0.76	0.10
DR=4	0.59	0.08	0.60	0.08	0.67	0.09	0.67	0.09	0.73	0.10
DR=5	0.59	0.08	0.60	0.08	0.61	0.08	0.63	0.08	0.76	0.10

Table 4.9: Prescision (Pre) and Recall (Rec) value of the ABIR system with Feature Selection Method

(a) Category : Autobahn

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.96	0.11	1	0.12	0.96	0.11	0.96	0.11	1	0.12
DR=2	0.88	0.10	1	0.12	1	0.12	1	0.12	1	0.12
DR=3	0.92	0.11	1	0.12	1	0.12	1	0.12	1	0.12
DR=4	0.88	0.10	0.92	0.11	0.96	0.11	0.96	0.11	0.96	0.11
DR=5	0.84	0.10	0.84	0.10	0.96	0.11	0.96	0.11	0.96	0.11

(b) Category : Coke Can

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.6	0.08	0.52	0.05	0.48	0.07	0.52	0.08	0.52	0.09
DR=2	0.64	0.06	0.56	0.08	0.56	0.07	0.56	0.06	0.76	0.09
DR=3	0.64	0.07	0.52	0.07	0.52	0.07	0.52	0.08	0.68	0.08
DR=4	0.52	0.06	0.56	0.05	0.64	0.06	0.6	0.06	0.68	0.07
DR=5	0.44	0.05	0.56	0.06	0.56	0.06	0.6	0.07	0.68	0.09

(c) Category : Triangle

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.36	0.06	0.4	0.07	0.32	0.06	0.24	0.04	0.44	0.08
DR=2	0.4	0.07	0.4	0.07	0.5	0.092	0.48	0.09	0.44	0.08
DR=3	0.36	0.06	0.6	0.11	0.44	0.08	0.52	0.09	0.48	0.09
DR=4	0.32	0.06	0.52	0.09	0.52	0.09	0.48	0.09	0.56	0.10
DR=5	0.28	0.05	0.48	0.09	0.44	0.08	0.52	0.09	0.68	0.12

(d) Average

	$QR = 1$		2		3		4		5	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
DR=1	0.64	0.08	0.64	0.08	0.59	0.08	0.57	0.07	0.65	0.09
DR=2	0.64	0.08	0.65	0.09	0.69	0.09	0.68	0.09	0.73	0.10
DR=3	0.64	0.08	0.71	0.10	0.65	0.09	0.68	0.09	0.72	0.10
DR=4	0.57	0.07	0.67	0.09	0.71	0.09	0.68	0.09	0.73	0.10
DR=5	0.52	0.07	0.63	0.08	0.65	0.09	0.69	0.09	0.77	0.11

Table 4.10: Precision results of Experiment Set_{match} for various criteria where $p = AP$, $p = RFxIPF$, $p = AVxIPF$, $p = AV$.

Dataset	AP	$RFxIPF$	$AVxIPF$	AV
STIMautobahn	0.80	0.92	0.88	0.96
STIMcoke	0.44	0.52	0.56	0.64
STIMtriangle	0.52	0.28	0.28	0.68
Average	0.58	0.57	0.57	0.76

Saliency Computation: 'Bottom-Up Normalization'

Region Extraction: QR=5, DR=5

Feature Extraction: 'Color Histogram'

***Similarity Matching**: $p =' Saliency(AV)'$,'AP','RFxIPF', 'AVxIPF'}*

Different significance criteria are used for similarity matching, where $p = AP$ (Area Percentage), $p = RFxIPF$ (Region Frequency-Inverse Picture Frequency), $p = AVxIPF$ (Attention Value-Inverse Picture Frequency) and $p = AV$ (Attention Value). The regions are clustered to compute the $RFxIPF$ criteria. K-means algorithm is used to cluster the regions of images. Different K values are tested and the precision results are given when K value is chosen as 30.

The average precision results of experiment Set_{match} are given in Table-4.10. Then, we calculate average precision of each category by evaluating the top 5 returned results. Various significance assigning methods are employed in the experiments. As seen from the table computing significance matrix using only attention values of the regions to match the region out performs the other methods. Emphasizing the significance of relatively more salient regions improves the retrieval performance.

The analysis of the significance matrix S provides us to interpret the results more clearly. For that purpose, the average difference of the diagonal trace of the significance matrix S are computed by taking the same number of salient regions extracted from the images. The reason for analyzing the diagonal trace is that the most distinctive regions are expected as the first extracted regions according to the suggested model. In this situation, while measuring the similarity between two images, the first extracted region of the first image is matched with the first extracted region in the

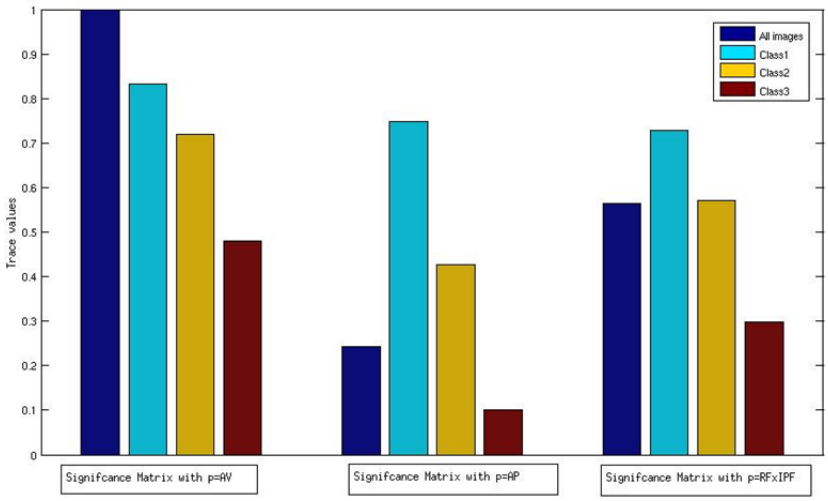


Figure 4.19: Trace of Significance matrix

second image as if these two images are visually similar with each other. The trace of the normalized S matrix for different p values are shown in Figure-4.19. The computed values are parallel to the experimental results. The experimental results show that the trace of the S matrix is higher than the other methods in the case of applying AV suggested in this work as a significant measure. By investigating the computed results shown in Figure-4.19, we can say that the higher precision performance leads to higher tracing values than the other p values .

4.2.6 Comparison of Segmented Regions with Salient Regions

Image segmentation is the process of partitioning an image into homogeneous segments according to the predefined homogeneity predicate. Image segmentation is an intermediate process between low-level features and high-level abstraction to separate objects and background. Image segmentation methods are not generic; solutions depend on the application and domain of the image. However, object-based image retrieval systems generally rely on the image segmentation methods to represent images with local regions [5].

The proposed ABIR system extracts salient regions from the saliency map of images without employing segmentation to the image plain. We assume that salient regions

are necessary to represent object-based images. However, segmentation-based image retrieval systems represent the images with the all regions extracted from the image. We employ a set of experiments to compare regions extracted by using visual attention and regions extracted by using segmentation to demonstrate the power of our ABIR system.

The regions are extracted by using Normalized Cut segmentation method [64]. We employ the following CBIR algorithm to use the regions extracted from the normalized cut segmentation. The steps of the algorithm explained as below:

Step1: The regions of query and database images are extracted by employing normalized cut segmentation algorithm.

Step2: The features of regions are extracted by using color histogram

Step3: The similarity between images are computed by using IRM method. Area percentage ($p = AP$) is used as significant criteria.

The following set of experiment for normalized cut regions is described by:

Experiment $Set_N cut$ {

***Region Extraction**: 'Normalized Cut Method'*

Feature Extraction: 'Color Histogram'

Similarity Matching: 'Area Percentag' }

As seen from the Experiment $Set_N cut$, saliency map computing step is removed. Other steps are the same as the experiments performed in the proposed ABIR system for the sake of fair comparison. Table-4.11 shows the retrieval results with average precision values. The average precision of each category is calculated by evaluating the top 5 returned results. Our system significantly gives better results than the system by using $N - cut$ segmentation: salient regions represent the objects in the database.

Table 4.11: Precision results for the CBIR system with Normalized cut segmentation and ABIR

	N-Cut	ABIR
STIMautobahn	0.88	0.96
STIMCoke	0.58	0.64
STIMtriangle	0.40	0.68
Average	0.62	0.76

4.3 Experiments with Sival Dataset

In the previous section, we analyse the proposed ABIR system through the experiments in STIM datasets. In this section, the proposed system is compared with the state of the art content-based image retrieval systems. The systems are tested in SIVAL dataset which is created for a localized CBIR system, ACCIO! [74]. In SIVAL dataset, the scenes and objects are highly diverse and often complex. Furthermore, the objects may occur anywhere spatially in the image and also may be photographed at a wide-angle or close up or with different orientations.

We have chosen AUC (the area under the ROC curve) as an evaluation measure in order to be consistent with the existing works which employ SIVAL dataset. Moreover, AUC can be considered as a reasonable measure of the retrieval results which is equivalent to the probability that a randomly chosen positive image will be ranked higher than a randomly chosen negative image. Unlike the precision-recall curve, the ROC curve is insensitive to ratio of positive to negative examples in the image repository. The confidence interval in experiments is set to 95 per cent. The proposed system is compared with the region-based and localized image retrieval systems in the literature which are developed for object-based image databases similar to our proposal. The regions can be extracted from the blocks of images. More commonly way is to employ segmentation methods to the images to extract meaningful regions for region-based CBIR systems. On the other hand, the goal of localized image retrieval systems are to extract interested part of the image by using a set of positive and negative query images. Our Attention-based image retrieval system is more likely similar to localized image retrieval systems considering salient regions. We compared our system with following studies:

SIMPLIcity [60]: SIMPLIcity is proposed by J. Wang. It is a standard region-based CBIR system, where the regions are extracted by dividing the image into predefined blocks. The system ranks the images in the dataset according to their similarity to the query image based on the integrated region matching (IRM) algorithm [27].The SIMPLIcity system uses the area percentage as significance criteria.

SBN [85]: SBN (Single-Blob with Neighbours) is a localized system which applies multiple instance learning. The system solved multiple instance learning with Diverse Density (DD) framework. A bag is labelled positive if it contains at least one positive instance, otherwise it is labelled as negative.

ACCIO! [12], [74]: ACCIO! Algorithm is used to identify the desired local content, re-weight features and the rank images in the image dataset. The algorithm uses randomly selected positive and negative labelled images as query set in conjunction with multiple instance learning algorithms.

GMIL [84]: GMIL (Graph-based Multiple Instance Learning) is an object-based image retrieval system which uses multiple-instance semi-supervised learning. Semi supervised learning is used to construct query set from a large set of unlabelled images.

The steps of our Attention-based image retrieval (ABIR) system used in this section are described with the following experiment set below:

Experiment Set_{SIVAL}*: {*

Saliency Computation*: 'Top-Down Feature Weighting'*

Region Extraction: QR=1, DR=2

Feature Extraction: '125-bin RGB Color Histogram'

Similarity Matching: 'p=AV'}.

Top-Down Feature Weighting algorithm is used in saliency map extraction step of ABIR system by considering detailed analysis in STIM dataset. The objects and backgrounds in SIVAL images are complex. For this reason, the Bottom-up normalization

algorithm is not convenient for SIVAL dataset. The normalization operatorN_BU utilize all the images in the dataset to compute global maxima of feature maps. In this case, the number of salient regions would be increased. A top-down algorithm which utilizes query image is more convenient for SIVAL images, since there is only one object of interest in images. We use the Top-down Feature Weighting algorithm considering this property of images.

Table-4.12 shows the average AUC values over 30 independent runs for all categories. Our system provides only one query image for each category. The retrieval performances of GMIL, ACCIO! and SBN systems using only one positive query images are provided to compare our system in equal conditions. The results of GMIL, ACCIO!, SIMPLIcity and SBN are directly referred from [84].

Our ABIR system outperforms the other systems for one positive query images in average AUC performance of all the categories. SIMPLIcity searches for images that holistically match query images, and may prefer images with similar background but different object, since the target object on average occupy only 10-15% of the image. SBN suffers from the lack of using a region-based approach and has less flexibility in varying the neighbour weights which prevents it from recognizing the same object that occur in scenes with different lighting. The first part of the table indicates the categories where the proposed ABIR system give higher AUC performance than the compared systems. The ACCIO! and GMIL gives better result in some categories such as Stripednotebook, checkeredscarf, feltflowerrug and greenteabox. When we examine these categories, we can see that either the representative features for objects of interest are texture. Our ABIR system cannot perform well enough to localize these objects because the color histogram are only extracted from the salient regions. On the other hand, our system gives prominent results for the categories containing a visually salient object such as Goldmedal, or Spritecan.

The scope of the proposed ABIR system is images which contain at least one interest of salient object. Saliency of objects should be represented by predefined low level features which are normalized color, intensity and orientation. However, saliency of objects can be formalized by top-down effects in real life objects, such as context information, expectation, expected locations. The proposed ABIR system give better

Table 4.12: Average AUC values over 30 independent runs for all categories

	GMIL	ACCIO	SIMPLIcity	SBN	ABIR
banana	55.8	52.8	50.8	49.6	62,0045
bluescrunge	55.5	50.4	53.9	46.7	60,5801
cokecan	66.6	65.6	55	47	67,0638
dataminingbook	68.1	53.9	55.6	46	69,0558
dirtyworkgloves	60.9	60.7	56.6	51.8	63,5233
fabricsoftenerbox	76.9	70.8	51.4	52.1	77,7978
glazedwoodpot	66.4	58	55.6	49.3	76,6942
goldmedal	58.2	57.3	57.1	51.6	77,0389
juliespot	68	58.9	55.4	49.6	69,0910
largespoon	54.6	53.2	-	55.3	56,3514
rapbook	61.6	56.4	57	51.2	67,2427
smileyfacedoll	61.4	60	55.1	46.7	74,5173
spritecan	62.4	60.8	54.4	48.6	73,8897
translucentbowl	61	61.2	54.6	45.8	77,3355
wd40can	63.5	65.4	55.8	49.4	75,0230
ajaxorange	67.1	56.3	54.6	52.4	66.7380
apple	59.6	52.2	50.5	45	57,5031
candlewithholder	64.7	62.5	-	54.1	64,5933
cardboardbox	64.6	61.6	56.6	50.6	63,5184
checkeredscarf	75.4	78.2	65.2	58	69,1209
dirtyrunningshoe	73.5	78.6	61.9	57.6	73,3475
feltflowerrug	75.8	69.2	55.4	45.9	66,2271
greenteabox	74.6	65.6	53.5	50.1	72,9908
stripednotebook	61.6	56.3	56.9	51.9	58,1154
woodrollingpin	56.5	60.2	55.1	50	53,9206
Average	64.6	61	55.6	50.3	67,7300

performance with simple objects such as banana, bluescrunge, translucentbowl, gold medal and smileyfacedoll. On the other hand, the retrieval results of ABIR system is poor with complex objects such as stripednotebook and checkeredscarf categories.

4.4 Itti-Koch and SIFT comparison

Itti-Koch is a computational model of human visual system how it processes and precepts the visual world. The model is based on a cognitive model which is called 'Feature Integration Theory'. On the other hand, SIFT is a method of extracting visual features which indicate the interesting points in an image. Both itti-koch and SIFT have similar and different aspects according to their computation and usage. The mechanism of itti-koch is explained in background chapter. Main process in Itti-Koch model is extraction of feature maps. In this step, Itti-Koch and SIFT have similar processes. A multi scale computation is used in both systems.

The main differences of both model is how they interpret the whole image. The output of itti-koch model is a saliency map which indicates the salient regions in the image and this map is also used in how to sift the attendance to the different regions in the image. Beside this, SIFT gives a number of points and its features which are distinctive points in the image. In the other words, Itti-koch gives regions and SIFT gives points. Because of this property, their usages in applications areas are different. However, both of them use the same idea that how a pixel is different from its neighbours. We give a comparison of SIFT and Itti-Koch in Table-4.13.

In addition, we compared both models with a set of experiments. Firstly, we applied Lowe's SIFT algorithm. As it known that, the problem in SIFT is number of descriptors extracted from images. We have reduced number of descriptors by using regions extracted from Itti-Koch model. The descriptors which are in these regions are used in the second experiments. Finally, we computed a map by using SIFT descriptors of the images for using with ABIR.

The SIFT map is visualization of SIFT descriptors of an image. Each descriptor is drawn by using its orientation and sigma. A Gaussian function with the sigma extracted from the SIFT descriptors is fitted for each location of SIFT descriptors.

Table 4.13: Comparison of SIFT and Itti-Koch

SIFT	Itti-Koch
Scale-invariant	Center-surround difference between different scales
No color	Use color features
Orientation descriptors	Orientation feature maps (Gabor)

All SIFT descriptors of an image are used when computing SIFT map. For example, the number of SIFT descriptors extracted form an image is X. In this case, X Gaussian functions calculated from Algorithm-6 with the original sigma values are applied to the locations of the in the map image:

In Figure-4.20, sample SIFT maps and saliency maps are shown. The second row shows the saliency map of Itti-Koch model and SIFT maps are shown in the next row. As seen from the Figure-4.20, SIFT map is coarse than saliency map.

The SIFT map is used as saliency map in experiments. In Table-4.14, results average precision results for STIM dataset are given. As seen from the results, the performances of SIFT map and ABIR are almost same. The reason for that, after saliency or SIFT map extraction, the same steps are used: number of regions (5 regions), features of regions (color histogram) and similarity measure (saliency) are same for both experiments. The background information is distinctive for some images especially for Autobahn category. So, SIFT Map and ABIR shows nearly same performances

Algorithm 6 The Algorithm of SIFT Map

1: X = number of SIFT descriptors of image I_x
2: $Sigma$: vector which stores the sigma value of each SIFT descriptors of I_x.
3: $Location$: vector which stores x and y coordinate of each SIFT
4: $Gaussian(sigma, location, image_plane)$: applies a Gaussian function to the image plane.
5: $sift_mape$: is an empty image with the same size of I_x.
6: **for** i=1:X **do**
7: $Gaussian(Sigma(i), Location(i), sift_map)$
8: **end for**
9: $returnsift_map$.

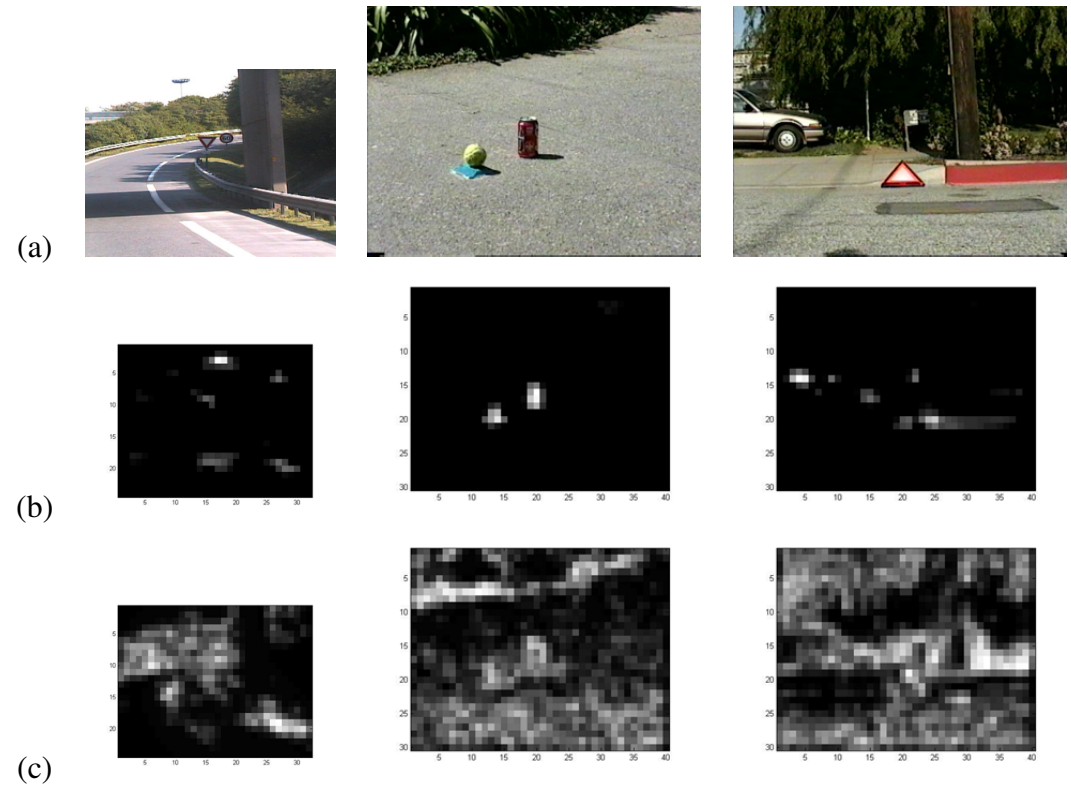


Figure 4.20: (a) Original images, (b) Saliency Map from the Itti-Koch Model (c) SIFT map.

for this dataset.

In Table-4.15, results of experiments in SIVAL dataset are given. In this experiments number of regions for query is set to 1 and number of regions for dataset images set to 2. When we analysis the results our ABIR system outperforms the SIFT, reduced SIFT and SIFT map. Although the numbers of regions are decreased, the performance of Reduced SIFT is not poor. However, the results of original SIFT and Reduced SIFT is interesting in some categories. For example, the result for FabricsoftnerBox, datamininbook, rapbook outperforms the all methods. But as expected, our method gives better result for visually salient objects.

4.5 Experiments with Satellite Images

Spatial resolutions of satellite images are very high compared to the standard image databases. For example, a satellite image in our Airport database covers a 5 square

Table 4.14: SIFT, Reduced SIFT and SIFTmap results for STIM dataset

(a)

Dataset	SIFT	Reduced SIFT	SIFT Map	ABIR
STIMautobahn	0.72	0.88	0.94	0.96
STIMCoke	0.24	0.24	0.66	0.64
STIMtriangle	0.52	0.52	0.68	0.68
Average	0.49	0.54	0.76	0.76

(b)

Dataset	SIFT	Reduced SIFT	SIFT Map	ABIR
STIMautobahn	0.09	0.10	0.11	0.11
STIMCoke	0.03	0.03	0.08	0.08
STIMtriangle	0.09	0.09	0.12	0.12
Average	0.07	0.08	0.10	0.10

Table 4.15: SIFT, Reduced SIFT and SIFTmap results for SIVAL dataset

	ABIR	*SIFT*	*Reduced_SIFT*	*SIFT_MAP*
ajaxorange	0.6674	0.7355	0.8053	0.5563
apple	0.5750	0.5958	0.623	0.5317
banana	0.6200	0.5860	0.5859	0.4388
bluescrunge	0.6058	0.5391	0.5316	0.4507
candlewithholder	0.6459	0.6000	0.6368	0.5342
cardboardbox	0.6352	0.6424	0.6162	0.494
checkeredscarf	0.6912	0.4405	0.5226	0.6639
cokecan	0.6706	0.6927	0.5646	0.7151
dataminingbook	0.6906	0.8287	0.4683	0.65
dirtyrunningshoe	0.7335	0.6456	0.5918	0.6493
dirtyworkgloves	0.6352	0.6765	0.6539	0.4916
fabricsoftenerbox	0.7780	0.8556	0.9052	0.5385
feltflowerrug	0.6623	0.7699	0.7198	0.5277
glazedwoodpot	0.7669	0.5585	0.5898	0.5213
goldmedal	0.7704	0.6031	0.6327	0.5486
greenteabox	0.7299	0.6858	0.6813	0.5198
juliespot	0.6909	0.5651	0.6094	0.5975
largespoon	0.5635	0.5237	0.635	0.3859
rapbook	0.6724	0.7748	0.8115	0.5441
smileyfacedoll	0.7452	0.6419	0.6255	0.5665
spritecan	0.7389	0.7050	0.912	0.6043
stripednotebook	0.5812	0.5165	0.5046	0.426
translucentbowl	0.7734	0.4885	0.5261	0.5522
wd40can	0.7502	0.5781	0.6138	0.5383
woodrollingpin	0.5392	0.5465	0.5333	0.5136
Average 67,7300	0.6318	0.636	0.5424	

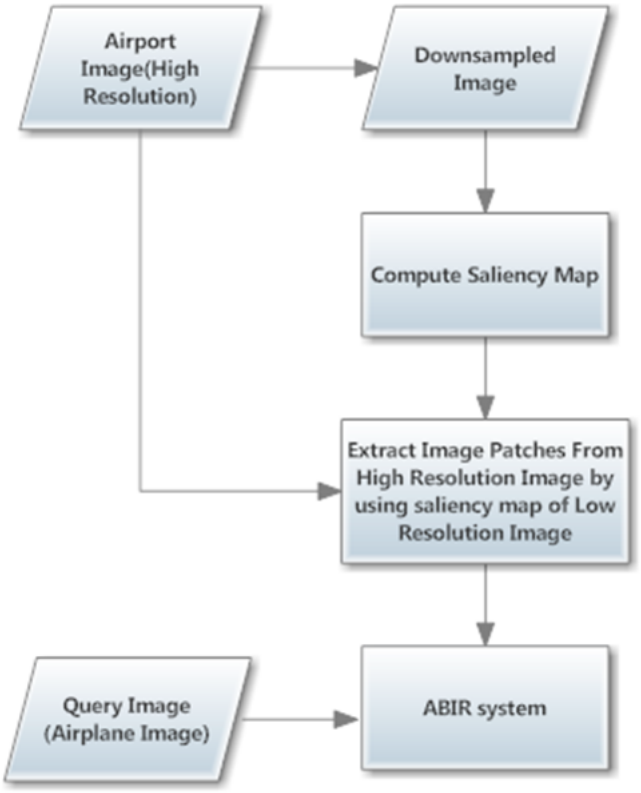


Figure 4.21: Setup for composing an image database from single High resolution sattellite Image.

km area. This size of image is not appropriate to process with our ABIR system. The objects can not be attended in high resolution. Therefore, low resolution versions of images are used to analyse coarsely which means to extract "salient areas" rather than "salient objects". Afterwards, "salient areas" are processed at original resolution to extract "salient objects". We can divide the high resolution images into a set of images by using saliency map of low resolution images as described as shown in Figure-4.21. A sample image and image patches which are extracted as described is shown in Figure-4.22.

There are five high resolution satellite images in our airport database. Each airport images processes separately where the goal is to localize airplanes in the images by the ABIR system. 9-12 image patches are extracted from each high resolution images by using saliency map of low resolution images. Table 4.16 tabulates the number of patches extracted from the high resolution of images and total area percentage of patches. The total area percentage equals to $\frac{total\#pixelinallpatches}{total\#pixelsinoriginalimage}$. Query images are

cropped from the high resolution images manually. Query images contain only one airplane. We have only one query image for each airport images.

Table 4.16: Number of patches extracted from high resolution Airport images and the total area percentage of patches to the original high resolution images

Image Name	Number of patches extracted from images	The total area percentage of patches
Airport-1	12	45
Airport-2	9	42
Airport-3	10	42
Airport-4	12	60

We compute the saliency map of image patches by using Top-Down Feature Map Weighting algorithm. The number of regions is determined by monitoring after how many fixations the serial scanning of the saliency map starts to cycle. Hence, numbers of regions which are extracted from image patches are varying according to the saliency of the images. For example, an image patch and the corresponding query image are shown in Figure-4.23. In the left image, the regions are extracted by using Itti-Koch model. The regions of right images are extracted by using Top-Down Feature Map Weighting algorithm considering the query image. As seen from the images, number of regions is the same (13 regions) in both images. However, the top-down feature map weighting algorithm localizes the airplanes better than Itti-Koch model. The feature maps equally contribute to the saliency map for Itti-Koch model. On the other hand, feature maps which are dominant for airplanes are emphasized by using Top-Down Feature Map Weighting algorithm.

The results of experiments are tabulated in Table-4.17. Image patches covers the "salient areas" in the images. The airplanes cannot be obtained if the airplanes are not in the "salient areas". However, nearly all airplanes can be obtained by the image patches as seen from the results. For example, one airplane is missed by the image patches for Airport-1 image. There are 49 airplanes in the original images and 48 airplanes in all image patches. Also, 140 regions are extracted from image patches and 35 of them contains airplane. The color histogram is used to match image regions. The precision results prominent, since those regions which contains airplane are visually similar to regions of query image.

4.6 Summary

This chapter gives the experimental results of the developed Attention-based Image Retrieval (ABIR) system. The proposed ABIR framework is tested over a subset of STIM image database and SIVAL image databases. The proposed ABIR system has four major steps: Saliency Map computing, Region Extraction, Feature Extraction and Similarity Matching. The each step is analysed in details with the proposed algorithms in STIM databases. SIVAL database is employed to compare our system with the CBIR systems in literature. Furthermore, we use our system on satellite images to localize airplanes on airports. We conclude that using visual attention give promising experimental results on object-based databases.

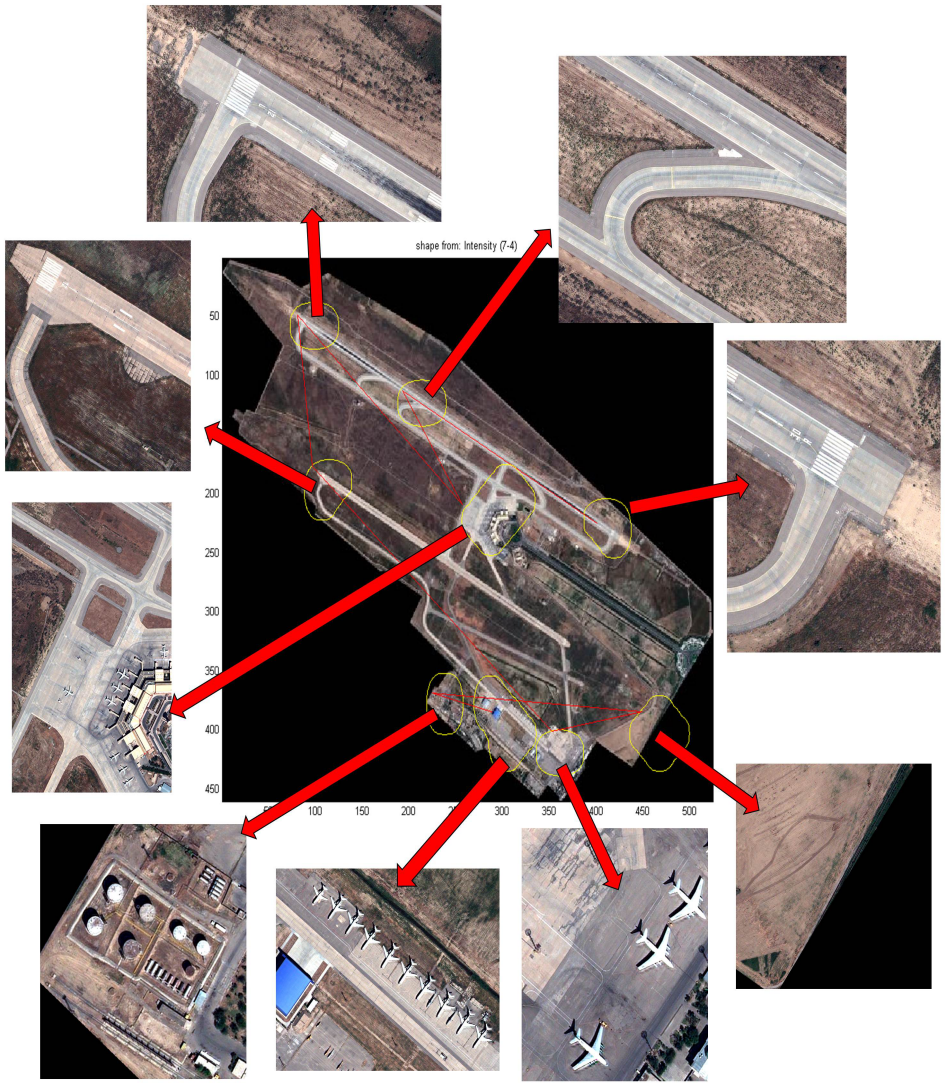


Figure 4.22: Shows the downsampled images with extracted "salient areas". From each region, image patches are extracted from original size of image.

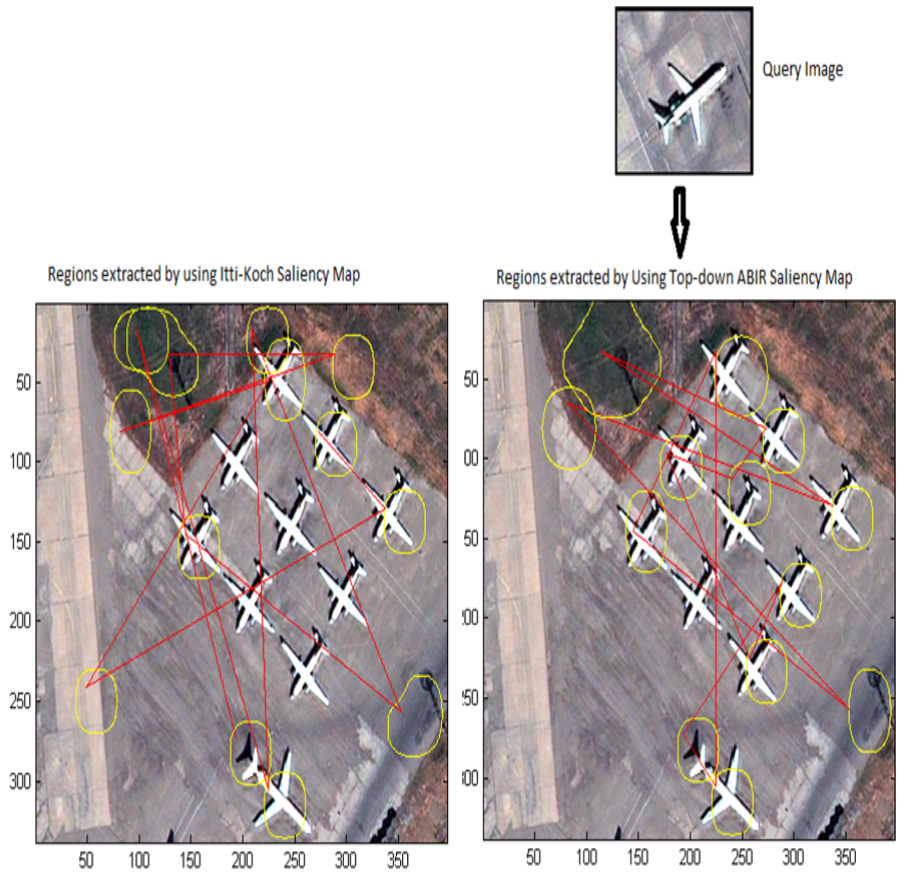


Figure 4.23: An image patch and the corresponding query image

Table 4.17: Results for Airplane Localization

Image Name	# of Airplanes	Number of Airplanes extracted from Image patches	# of Regions extracted from image patches	# of Regions contain an Airplane	Precision for the five regions
Airport-1	49	48	140	35	100
Airport-2	28	25	75	18	80
Airport-3	29	28	92	17	100
Airport-4	13	9	116	5	40

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

In this thesis, we proposed a new approach to the CBIR problem, motivated by human visual attention. This approach formalizes the image retrieval problem as retrieving images which have salient regions similar to salient regions of the query image in an image database. We assume that salient regions compose of object(s) or area(s) that we are looking for them in an image.

The proposed method can be summarized as follows:

- First, salient regions are extracted from the saliency map of images. Saliency map indicates the locations of salient regions in an image. We propose three algorithms to compute saliency map based on Itti-Koch visual attention model. These algorithms are namely Bottom-Up Normalization, Top-Down Normalization and Top-Down Feature Map Weighting. Bottom-up Normalization algorithm introduces a new normalization operator, N_{BU}, to compute saliency map of images according to the global maxima of feature maps of all images in the database. The Top-down Normalization algorithm normalize feature maps of images in the database by using global maxima of query image. The last algorithm, Top-down Feature Map Weighting, uses query image to emphasis salient regions of query image in the database images by giving weights to the feature maps of the database images. After computing saliency maps, salient regions are extracted by the method proposed in [9] by using winner maps of most salient point in saliency map.

- The next step is the feature extraction of salient regions, after obtaining salient

regions from saliency map of images. For that purpose, we propose Saliency-Based Feature Integration and Saliency-Based Feature Selection algorithms for feature extraction step. The Saliency-based Feature Integration theory concatenate different feature types, such as average color, intensity and entropy, by giving weights computed from saliency of region. The Saliency-based Feature Selection algorithm selects a feature type in a set of feature types by utilizing saliency of region.

- Finally, the similarities between the query image and database images in salient regions are measured. We propose a new method by adopting IRM [27] called Saliency-Based Similarity Matching. While IRM method matches the images based on region sets according to a significance measure between the regions, our proposed method uses saliency value of regions as a significance measure to match images.

The proposed system is tested on STIM, SIVAL and an airport databases. A set of experiments are carried out to evaluate steps of proposed system on relatively small STIM database. The experimental results show that the proposed saliency computation algorithms give better performance than Itti-Koch by using saliency-based similarity matching algorithm. Bottom-up Normalization, Top-Down normalization and Top-down Feature Map Weighting algorithms are specific to CBIR problem, where the Itti-Koch model is designed for a single image. Feature extraction experiments show that Saliency-based Feature Integration and Saliency-based Feature Selection have better performance on STIM database than classic feature concatenation in the context of our ABIR system.

The performance of proposed ABIR system is compared with selected state of the art CBIR systems namely Simplicity [60], SBN [85], ACCIO! [12] and GMIL [84] on SIVAL databases. There are 25 categories with complex objects and cluttered backgrounds on SIVAL database. The experimental results show that, the proposed ABIR system has better performance in average in terms of all categories of SIVAL than these four systems.

Lastly, we test our ABIR system on high resolution satellite images to localize airplanes in an airport image. The airplanes are salient objects in an airport image.

However, the high resolution of images makes them difficult to be tested with our system. For this reason, we employ Itti-Koch model to low resolution version of the original image to extract salient areas. The image patches which contain salient areas are extracted from the high resolution image. Hence, we obtain a set of high resolution image patches. We test our ABIR system on these patches. We eliminate the homogen areas such as large green fields with these method. The experimental results are promising to localize airplanes in an airport image.

We investigate the relation of SIFT local descriptors and Itti-Koch model with a set of experiments. SIFT is powerful method to find invariant points in an image. Experimental results show that the performance of SIFT is poor if the background is cluttered and the object of interest is small relatively to the image background. However, proposed ABIR system gives better performance to localize object of interest in an object-based image database such as SIVAL.

The proposed ABIR system is attractive, because it reflects the human perception to CBIR problem. We consider query image and database together to formalize the ABIR system. The saliency information of images is employed to all steps of the system. The salient regions are extracted from saliency map of images without applying segmentation to the images, which is an important property of the proposed system.

5.1 Future Directions

The work presented in this thesis introduces a new aspect which considers human perception to formalize CBIR problem. It is my hope that our proposed ABIR system will be inspired and improved by the further study in computer vision. There are still a lot of problem to be solved in image retrieval. Therefore, it leaves a number of issues open-ended for future research:

- The number of regions extracted from the images depends on the image database. We use a predefined value for number of salient regions which is determined by observing saliency maps of sample images from the database. Determining number of salient regions automatically is an open question.

- Computation time of Bottom-up normalization algorithm depends on number of images in the dataset. The algorithm is not usable for huge dataset for real-life applications in this format. A system learns the normalization parameters from a training set should be developed to solve time complexity of the algorithm.

- The number of salient regions extracted from the images are more than one region. Some of the salient regions are not region of interest. In some images, background can include salient regions. We use query image to compute the saliency map of database images in Top-down Normalization and Top-down Feature Map Weighting algorithms. These algorithms try to emphasis salient regions according to the query image, however, they can not eliminate other salient regions such as other salient objects or background. Supervised methods can be used to distinguish salient regions.

- Saliency-based Feature Integration and Saliency-based Feature Selection algorithms do not consider the some aspects of feature design such as size and number of feature vectors. We choose rather basic features to represent region of interest. However, there are more complicated features in literature. Also, these algorithms could be used with only saliency-based systems. Another feature direction is feature space design. The proposed Saliency-based feature integration and saliency-based feature selection algorithms can be improved by more complex features in literature.

- The proposed ABIR system considers the CBIR problem in a bottom-up manner. We use query image in Top-down Normalization and Top-down Feature weighting algorithm, however, we do not know what we are looking for in a query image. We try to emphasis the saliency of query image into the database images. A top-down visual attention model for ABIR system can be realized if we learn the object of interest in the query image. Semi-supervised learning with a set of positive and negative query images can be used to achieve a top-down visual attention model.

- The evaluation of proposed saliency map computing algorithms are realized with retrieval performance of ABIR system. An evaluation with human subjects

should be informative to see drawbacks of algorithms and develop a ABIR system which gives better performance.

REFERENCES

[1] J. Theeuwes, A.F. Kramer, and A. Kingstone. Attentional capture modulates perceptual sensitivity. *Psychonomic Bulletin & Review*, 11(3):551–554, 2004.

[2] J.M. Wolfe. Visual search. *Attention*, 1:13–73, 1998.

[3] Treisman A. and Gelade G. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136, 1980.

[4] J.M. Wolfe and G. Gancarz. Guided search 3.0: A model of visual search catches up with jay enoch 40 years later. *Documenta Ophthalmologica Proceedings Series*, 60:189–192, 1997.

[5] S. Frintrop, E. Rome, and H. I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept.*, 7(1):6:1–6:39, January 2010.

[6] S. Frintrop. Vocus:a visual attention system for object detection and goal-directed search. Phd. Thesis in Fraunhofer Institute for Autonomous Intelligent Systems of Bonn University, 2001.

[7] F. W. M. Stentiford. An attention based similarity measure with application to content based information retrieval. In *In Proceedings of the Storage and Retrieval for Media Databases conference, SPIE Electronic Imaging*, volume 5021, 2003.

[8] L. Itti and C Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, feb 2001.

[9] Walther D. and Koch C. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395 – 1407, 2006.

[10] J Li and J.Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1075 – 1088, sept. 2003.

[11] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254 –1259, nov 1998.

[12] R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J.E. Fritts. Localized content based image retrieval. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 227–236, 2005.

[13] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[14] H. Bay, T. Tuytelaars, and L. Gool. Surf: Speeded up robust features. In A. Leonardis, Ho. Bischof, and A. Pinz, editors, *Computer Vision ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer Berlin Heidelberg, 2006.

[15] K. Yan and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513, june 2004.

[16] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008.

[17] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 257–263, 2003.

[18] A. W. M. Smeulders, M. Worring, S. Santini, and R. Gupta, A.and Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, December 2000.

[19] Datta R., Joshi D., Li J., and Wang J.Z. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2), 2008.

[20] W. James. *The Principles of Psychology Vol.1*. New York, NY, US: Henry Holt and Co., 2005.

[21] K.R. Cave and J.M. Wolfe. Modeling the role of parallel processing in visual search. *Cognitive psychology*, 22(2):225–271, 1990.

[22] Max Velmans. Is human information processing conscious? *Behavioural and Brain Sciences*, 14(4):651–726, 1991.

[23] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):802 –817, may 2006.

[24] Charles E.C., Howard E.E., and Steven Y. Visual attention: Bottom-up versus top-down. *Current Biology*, 14(19):R850 – R852, 2004.

[25] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 37–44, June 2004.

[26] O. Marques, L.M. Mayron, G.B. Borba, and H.R. Gamba. An attention-driven model for grouping similar images with image retrieval applications. *EURASIP Journal on Applied Signal Processing*, 2007(1):116–116, 2007.

[27] Jia Li, James Z Wang, and Gio Wiederhold. Irm: integrated region matching for image retrieval. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 147–156. ACM, 2000.

[28] H. Deubel and W.X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research*, 36(12):1827–1837, 1996.

[29] W. Styles. *Attention, Perception, and Memory: An Integrated Introduction*. Psychology Press, 2005.

[30] M.I. Posner, C.R. Snyder, and B.J. Davidson. Attention and the detection of signals. *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General*, 109(2):160, 1980.

[31] M. Begum and F. Karray. Visual attention for robotic cognition: a survey. *IEEE Transactions on Autonomous Mental Development*, 3(1):92–105, 2011.

[32] J. M. Wolfe, K. R. Cave, and S. L. Franzel. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419–433, aug 1994.

[33] J.M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

[34] A. Johnson and R.W. Proctor. *Attention: Theory and practice*. Sage Publications, Incorporated, 2003.

[35] M. Corbetta, G.L. Shulman, et al. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):215–229, 2002.

[36] M.I. Posner and S.E. Petersen. The attention system of the human brain. Technical report, DTIC Document, 1989.

[37] A. Treisman and S. Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychological review*, 95(1):15, 1988.

[38] J. M. Marr and L. Vania. Representation and recognition of the movements of shapes. *Proc. R. Soc. Lond. B*, 214:501–524, 1982.

[39] D. Marr. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, 1982.

[40] S.E. Palmer. *Vision science: Photons to phenomenology*, volume 1. MIT press Cambridge, MA, 1999.

[41] C. Koch and S. Ulman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[42] Simone Frintrop, Gerriet Backer, and Erich Rome. Goal-directed search with a top-down modulated computational attention system. In *Pattern Recognition*, pages 117–124. Springer, 2005.

[43] M.I. Posner and Y. Cohen. Components of visual orienting. In H. Bouma and D.G. Bouwhuis, editors, *Attention and Performance X: Control of LAnguage Processes*, volume 10, pages 531–556. Lawrence Eplbaum Assoc.,Hillsale,NJ, 1984.

[44] M.Z Aziz and B. Mertsching. Color saliency and inhibition using static and dynamic scenes in region based visual attention. In Lucas Paletta and Erich Rome, editors, *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, volume 4840 of *Lecture Notes in Computer Science*, pages 234–250. Springer Berlin Heidelberg, 2007.

[45] T. Horowitz and J. Wolfe. Memory for rejected distractors in visual search? *Visual Cognition*, 10(3):257–298, 2003.

[46] S. Frintrop and A.B. Cremers. Top-down attention supports visual loop closing. In *Proc. of European Conference on Mobile Robotics (ECMR 2005)*. Citeseer, 2007.

[47] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10(1):161–169, 2001.

[48] L. Itti and C. Koch. Comparison of feature combination strategies for saliency-based visual attention systems. In *Electronic Imaging'99*, pages 473–482. International Society for Optics and Photonics, 1999.

[49] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005.

[50] J. K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(03):423–445, 1990.

[51] Shishir K. Shandilya and Nidhi Singhai. A survey on: Content based image retrieval systems. *International Journal of Computer Applications*, 4(2):22–26, July 2010. Published By Foundation of Computer Science.

[52] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[53] R. Datta, W. Ge, J. Li, and J. Z. Wang. Toward bridging the annotation-retrieval gap in image search by a generative modeling approach. In *Proceedings of the 14th annual ACM international conference on Multimedia*, MULTIMEDIA '06, pages 977–986, New York, NY, USA, 2006. ACM.

[54] V.N. Gudivada and V.V. Raghavan. Content based image retrieval systems. *Computer*, 28(9):18 –22, sep 1995.

[55] Li W.J. and Yeung D.Y. Localized content-based image retrieval through evidence region identification. In *IEEE Conference on Computer Vision and Pattern Recognition.*, pages 1666–1673, june 2009.

[56] T. Kato. Database architecture for content-based image retrieval. *Proc. SPIE 1662, Image Storage and Retrieval Systems*, 1662:112–123, 1992.

[57] M. J. Swain and Ballard D. H. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[58] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460 – 473, june 1978.

[59] A Jain and Vailaya A. Shape-based retrieval: A case study with trademark image databases. *Pattern Recognition*, 31:1369–1390, 1998.

[60] Z.J. Wang, J. Li, and Wiederhold Y.G. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:947–963, 2001.

[61] J. Wang, G. Wiederhold, O. Firschein, and S. Wei. Content-based image indexing and searching using daubechies' wavelets. *International Journal on Digital Libraries (IJODL)*, 1(4):311–328, March 1998.

[62] Y Chen, J.Z. Wang, and R. Krovetz. Clue: cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing*, 14(8):1187 –1201, aug. 2005.

[63] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026 – 1038, aug 2002.

[64] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888 –905, aug 2000.

[65] K. Barnard, P. Duygulu, and D. Forsyth. Clustering art. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 434–441, 2001.

[66] P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 97–112, London, UK, UK, 2002. Springer-Verlag.

[67] D. Liu and T. Chen. Probabilistic relevance feedback for image retrieval with binary feature vectors. In *IEEE Conf. Acoustics, Speech, and Signal Processing*, 2005.

[68] D. Hoiem, R. Sukthankar, H. Schneiderman, and L. Huston. Object-based image retrieval using the statistical structure of images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 490–497, june 2004.

[69] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615 –1630, oct. 2005.

[70] D.G. Lowe. Object recognition from local scale-invariant features. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150 –1157, 1999.

[71] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45:83–105, 2001.

[72] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988.

[73] D. Zhang, A. Wong, M. Indrawan, and G. Lu. Content-based image retrieval using gabor texture features. In *IEEE Pacific-Rim Conference on Multimedia, University of Sydney, Australia*, 2000.

[74] R. Rahmani, S.A. Goldman, Hui Zhang, S.R. Cholleti, and J.E. Fritts. Localized content-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1902 –1912, nov. 2008.

[75] Feng S., Lang C., and Xu D. Localized content-based image retrieval using saliency-based ,graph learning framework. In *IEEE 10th International Conference on Signal Processing (ICSP)*, pages 1029 –1032, oct. 2010.

[76] F. Li, Q Dai, W Xu, and G. Er. Multilabel neighborhood propagation for region-based image retrieval. *IEEE Transactions on Multimedia*, 10(8):1592 –1604, dec. 2008.

[77] K Gao, S Lin, Y Zhang, Sheng Tang, and H Ren. Attention model based sift keypoints filtration for image retrieval. In *Seventh IEEE/ACIS International Conference on Computer and Information Science*, pages 191 –196, may 2008.

[78] Fu H., Chi Z., and Feng D. Attention-driven image interpretation with application to image retrieval. *Pattern Recognition*, 39(9):1604 – 1621, 2006.

[79] O. Marques, L.M. Mayron, G.B. Borba, and H.R. Gamba. Using visual attention to extract regions of interest in the context of image retrieval. In *Proceedings of the 44th annual Southeast regional conference*, pages 638–643. ACM, 2006.

[80] FWM Stentiford. An estimator for visual attention through competitive novelty with application to image compression. In *Picture Coding Symposium*, pages 25–27, 2001.

[81] S. Feng, D. Xu, X. Yang, and A. Wu. A novel region-based image retrieval algorithm using selective visual attention model. In *Advanced Concepts for Intelligent Vision Systems*, pages 235–242. Springer, 2005.

[82] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16:321–328, 2004.

[83] R. Rahmani and S.A. Goldman. Missl: Multiple-instance semi-supervised learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 705–712. ACM, 2006.

[84] C. Wang, L. Zhang, and H.J. Zhang. Graph-based multiple-instance learning for object-based image retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 156–163. ACM, 2008.

[85] O. Maron and A.L. Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, volume 15, pages 341–349, 1998.

[86] D. Parkhurst, K. Law, E. Niebur, et al. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–124, 2002.

[87] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–27, 1985.

[88] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition- a gentle way. In *Biologically Motivated Computer Vision*, pages 251–267. Springer, 2002.

[89] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, 2003.

[90] J.Z. Wang and Y. Du. Scalable integrated region-based image retrieval using irm and statistical clustering. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 268–277. ACM, 2001.

[91] Wei Li, Shiming Xiang, Haibo Wang, and Chunhong Pan. Robust airplane detection in satellite images. In *18th IEEE International Conference on Image Processing (ICIP)*, pages 2821–2824, 2011.

[92] J.D. Still, V.J. Dark, and D.J. Parkhurst. Viewpoint invariant object features attract overt visual attention. *Journal of Vision*, 7(9):445–445, 2007.

CURRICULUM VITAE

Gülşah Tümüklü Özyer received her B.Sc. degree from Erciyes University in Computer Engineering Department in 2001. She was a visitor researcher at Penn State University, USA in James Z.Wang Research Group between February 2007 and March 2008. She is currently a Ph.D. student at Department of Computer Engineering of Middle East Technical University, Turkey. Her research interests include computer vision, image processing and pattern recognition.