Models for visual saliency in images and videos

Viswanath Gopalakrishnan

2010

Viswanath, G. (2010). Models for visual saliency in images and videos. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/46304>

<https://doi.org/10.32657/10356/46304>

Models for Visual Saliency in Images and Videos

Viswanath Gopalakrishnan

**School of Computer Engineering
Nanyang Technological University
Singapore**

A thesis submitted to Nanyang Technological University
in fulfillment of the requirements for the
Degree of Doctor of Philosophy

December 2010

Abstract

Multimedia data consisting of videos and images are in general very rich in content and carry lot of information. With the digital cameras and camcorders becoming easily available to consumers, the amount of information in the form of multimedia data stored and shared across web has increased exponentially. Human beings are normally interested only in a 'subset' of this information available in images and videos, which they consider as salient. Given the huge amount of data to be processed, it is important to obtain information regarding the regions in an image or video which attract human attention so that the overhead of redundant data can be greatly reduced in multimedia applications. The detection of such salient regions or objects in images and videos is the objective of the proposed research work.

We investigate the role of features like color and texture in deciding the salient regions in an image by analyzing its distributions in the spatial domain and feature domain. A color saliency framework and an orientation saliency framework is proposed to identify salient regions resulting from color and texture. The dominant colors of any image are modeled as mixtures of gaussians and a saliency value is attributed to each gaussian (color) depending on its 'isolation' in the feature domain and its 'compactness' in the spatial domain. The complexity of orientation histogram in different scales for a local patch is proposed as the texture feature. The orientation saliency map is evaluated by computing the spatial distribution of the different orientations and the contrast of the texture complexity to immediate neighborhood. The final saliency map that best describes the salient regions in an image is selected from the color and orientation maps in a systematic way.

Graph representations of images possess great potential in describing the geometric dependencies of various regions in the image. This factor is utilized in our

proposed graph based method for salient object detection. We formulate the problem of salient object detection in images as an automatic labeling problem on the vertices of a weighted graph. The seed (labeled) nodes are first detected using Markov random walks performed on two different graphs that represent the image. While the global properties of the image are computed from the random walk on a complete graph, the local properties are computed from a sparse k -regular graph. An enhanced graph representation of the image that represents the saliency between nodes more accurately, the 'pop-out graph' model, is computed further based on the knowledge of the labeled salient and background nodes. A semi-supervised learning technique is used to determine the labels of the unlabeled nodes by optimizing a smoothness objective label function on the newly created 'pop-out graph' model.

Feature selection plays a crucial role in deciding the salient regions of an image as in any other pattern recognition problem. However, the problem of identifying the relevant features that plays a fundamental role in saliency of an image has not received much attention so far. We introduce an unsupervised feature selection method to improve the accuracy of salient object detection. The noisy irrelevant features in the image are identified by maximizing the mixing rate of a Markov process running on a linear combination of various graphs, each representing a feature. The global optimum of this convex problem is achieved by maximizing the second smallest eigen value of the graph Laplacian via semi-definite programming. The enhanced image graph model, after the removal of irrelevant features, is shown to improve the salient object detection performance on a large image data base with annotated 'ground truth'.

Our proposed research work also focus on the problem of detection of salient motion in videos. Detection of the motion of foreground objects on the backdrop of constantly changing and complex visuals is a crucial task in automated video analysis. The motion of foreground objects, which is termed as salient motion, is marked

by its predictability compared to the more complex unpredictable motion of the backgrounds like fluttering of leaves, ripples in water, smoke filled environments etc. We introduce a novel approach to detect this salient motion based on the control theory concept of 'observability' from the output pixels, when the video sequence is represented as a linear dynamical system. The group of output pixels with maximum saliency is further used to model the holistic dynamics of the salient region. The resulting algorithm is tested on a set of challenging sequences and compared to the state-of-the-art methods to showcase its superior performance on grounds of its computational efficiency and detection capability of the salient motion.

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Deepu Rajan for his invaluable guidance during the period of my thesis work. His insight and knowledge on the potential of research in visual saliency has helped me to choose my research topic. His constant encouragement motivated me to pursue my research passionately while his technical involvement has supported me to carry forward the research. His experience in writing skills has been of great help to me in the presentation of my research work. I would also like to extend my gratitude to Dr. Hu Yiqun for having fruitful collaboration and helping me with his technical expertise on my research topic. I would like to thank all my friends and my colleagues in Centre for Multimedia and Networking (CeMNet) for their technical and non-technical support. Finally, I give my special thanks to the members of my family, my parents and my dearest wife Suma for having stood with me and supported me during the period of my research work.

Contents

1	Introduction	12
1.1	Motivation	14
1.2	Objectives	15
1.3	Contributions	15
1.4	Organization	17
2	Literature Review	18
2.1	Methods for Saliency Detection in Images	18
2.1.1	Biologically Inspired Methods	19
2.1.2	Low-Level Feature Based Methods	20
2.1.3	Methods for Region/Object Segmentation	22
2.1.4	Top-Down Methods	25
2.2	Methods for Saliency Detection in Videos	27
2.2.1	Optical Flow Based Methods	28
2.2.2	Auto-Regressive(AR) Model Based Predictive Methods	29
3	Modeling Image Saliency Using Distributions of Color and Orientation	31
3.1	Introduction	31
3.2	Color Saliency Framework	36

3.2.1	Dominant Hue Extraction	37
3.2.2	Model Fitting in H-S Domain	40
3.2.3	Cluster Compactness (Spatial Domain)	41
3.2.4	Cluster Isolation (H-S Domain)	42
3.2.5	Pixel Saliency	43
3.3	Orientation Saliency Framework	43
3.3.1	Local Orientation Histogram	45
3.3.2	Spatial Variance of Global Orientation	47
3.3.3	Orientation Entropy Contrast	49
3.3.4	Final Orientation Saliency	50
3.4	Saliency Map Selection	52
3.5	Experiments	54
3.5.1	Experiment 1 : Color Saliency	55
3.5.2	Experiment 2 : Orientation Saliency	57
3.5.3	Experiment 3 : Final Saliency Map Selection	60
3.5.4	Comparison with other saliency detection methods	62
3.5.5	Failure cases and analysis	63
3.6	Conclusion	66
4	A Graph Based Method for Salient Region Detection in Images	68
4.1	Introduction	68
4.2	Ergodic Markov chain fundamentals	70
4.3	Graph representation	72
4.4	Node selection	76
4.4.1	Most salient node	76
4.4.2	Other salient nodes	78
4.4.3	Background nodes	79
4.5	‘Pop-Out’ Graph Model	81

4.5.1	Distance factor	83
4.6	Labeling of unlabeled nodes using constrained learning on graphs . .	85
4.7	Experimental results	88
4.7.1	Comparison with other saliency detection methods	96
4.8	Conclusion	97
5	Unsupervised Feature Selection for Salient Object Detection	99
5.1	Introduction	99
5.2	Graph Laplacian and Mixing Rate of Markov Process	102
5.3	Feature Graphs and Irrelevant Features	104
5.3.1	Irrelevant Feature Identification	106
5.4	Experiments	109
5.5	Conclusion	114
6	A Linear Dynamical System Framework for Salient Motion Detection	115
6.1	Introduction	115
6.2	State-Space Representation	117
6.2.1	Dynamic Textures as Linear Dynamical Systems	117
6.2.2	Observability of a State Space Model	119
6.2.3	Distance between AR Models	121
6.3	Pixel Saliency Map Using Observability from Output Pixels	121
6.3.1	Quantitative Measure for Observability	123
6.3.2	Sustained Observability	124
6.4	Region Based Saliency Map Using Salient Patch Dynamics	126
6.4.1	Pole Analysis for Comparison of Region Dynamics	127
6.4.2	Region Based Saliency map using Local Dynamics	130
6.4.3	Region Based Saliency map using Global Dynamics	131

6.5	Experiments	132
6.5.1	Subjective Evaluation	133
6.5.2	Quantitative Evaluation	136
6.5.3	Effect of Region Saliency Maps	137
6.5.4	Limitation	140
6.6	Conclusion	141
7	Conclusions	142
7.1	Summary of Contributions	142
7.2	Future Research	144
A		158
A.1	Observability Measure from Outputs for Discrete-Time, Linear Time-Invariant Systems	158

List of Figures

3.1	The proposed framework for salient region detection in an image. . .	35
3.2	Initialization and fitting of Gaussian clusters. (a) Original image. (b) The dominant hue palette. (c) Initial Gaussian clusters in Hue-Saturation space. (d) Gaussian clusters after EM based parameter estimation. Note that in figures 3.2(c) and 3.2(d) data points lie in Hue - Saturation circle, where hue is represented by the positive angle between the line joining the origin (0,0) to the data point and the X-axis and saturation is represented by the radial distance from the origin (0,0).	39
3.3	(a, c) Original images; (b, d) Color saliency maps.	44
3.4	(a, b) Example images in which color plays no role in saliency.	44
3.5	Orientation histogram from Fourier spectrum. (a) Original image, (b) Centre-shifted Fourier transform, (c) Orientation Histogram.	45
3.6	(a) Original images. (b) Saliency maps extracted by computing spatial variance of global orientation.	48
3.7	(a) Original images (b) Saliency map based on orientation entropy contrast method.	51
3.8	Salient region detection by the color saliency framework. (a, c) Original images. (b, d) Color saliency maps.	56

3.9	(a) Original images. Color saliency maps using (b) intra color spatial variance feature [48], and (c) the proposed color saliency frame work.	57
3.10	(a, c) Original images. (b, d) Saliency map based on spatial variance of global orientation.	59
3.11	(a) Original image. Saliency map based on (b) spatial variance of global orientation (Sal_{gP}) and (c) orientation entropy contrast (Sal_{lP}). (d) Final orientation saliency(Sal_{OP}).	59
3.12	(a) Original images. (b) Binary color saliency map. (c) Binary orientation saliency map. (d) Final saliency map, which corresponds to the binary color or orientation saliency map with higher $SalIndex$	61
3.13	Comparison of average precision values of color saliency framework, orientation saliency framework and the combined framework.	62
3.14	Comparison of average precision, recall and f-measure values of proposed framework, spectral residual method [88], saliency tool box [81] and contrast method [49]. Horizontal axis shows 1) Precision 2) Recall 3) F-Measure	64
3.15	Comparison of f-measure values of proposed framework, spectral residual method [88], saliency tool box [81] and contrast method [49] for different values of α	64
3.16	Saliency map comparisons in Berkeley image database (a) Original image. Saliency maps generated by (b) proposed method, (c) spectral residual method [88], (d) saliency tool box [81], and (e) contrast based method [49].	65
3.17	Failure Cases. (a) Original image. (b) Color saliency map. (c) Orientation saliency map.	66

4.1	Comparison of [29] and proposed method for detection of most salient node. (a) Original images, (b) most salient node detected by proposed method and (c) most salient node detected by [29].	78
4.2	Detection of the most salient node. (a, c) Original images and (b, d) Most salient node spotted with the red region.	79
4.3	Detection of the Salient and Background nodes. (a, c) Original images and (b, d) Salient nodes spotted with red regions and background nodes spotted with the green regions.	81
4.4	Percentage accuracy of salient and background nodes vs. number of nodes detected for an image.	89
4.5	F-Measure surface for all combinations of salient and background node numbers ranging from 1 to 10.	92
4.6	F-Measure variation for different values of σ^2	93
4.7	Results of function fitting using 'soft labeling' and thresholding. (a),(c),(e),(g) Original images (b),(d),(f),(h) The final binary saliency maps.	94
4.8	Bounding box on salient region. (a, c) Original images and (b, d) Red bounding box over the salient region.	95
4.9	Failure examples. (a) Original image, (b) most salient node and (c) the corresponding bounding box.	96
4.10	Comparison of Precision, Recall and F-Measure values of Spectral Residual Method [88], Phase spectrum method [28], Saliency Tool Box [37], and the proposed method. Horizontal axis shows 1) Precision 2) Recall 3) F-Measure.	97
4.11	Comparison of F-Measure values of Spectral Residual Method [88], Phase spectrum method [28], Saliency Tool Box [37], and the proposed method for various values of α	98

5.1	Example images illustrating the relevancy of different features to saliency. The relevant salient features are (a) color (b) texture (c) some combination of color and texture features.	101
5.2	Feature graphs with the same set of nodes showing (a) bottleneck between two groups of nodes (b) no bottleneck. The thicker edges indicate larger affinity between nodes.	106
5.3	Comparison of saliency maps for graph based method in Chapter 4 with the proposed feature selection . (a),(d) Original Images. (b),(e) Saliency maps using method in Chapter 4. (c),(f) Saliency maps of Chapter 4 with the proposed feature selection.	111
6.1	Video sequences with different dynamic backgrounds (a) 'jump' sequence with smoke, (b) 'ocean' sequence with waves and (c) 'skiing' sequence with snow.	122
6.2	Sustained observability calculated over adjacent video buffers. s1 and s2 are saliency maps computed using dynamic texture models from frames 1 to q and 2 to q+1.	125
6.3	Example showing the generation of final saliency map with sustained observability a) 12 th frame of 'cyclists' video (b, c) saliency maps s1 and s2 computed using frame buffers with frames 1 to 21 and 2 to 22 d) final saliency map	126
6.4	Block Diagram detailing different functional modules in evaluation of region based saliency	128
6.5	Poles a , b , c , e , and f inside the unit circle on the complex plane. a* and c* are the complex conjugates of a and c respectively. a , b and c are the poles close to unit circle and hence of oscillatory nature. e and f are poles close to the origin and hence, of fading nature.	129

6.6	Saliency maps obtained using the proposed method on non-consecutive frames of various video sequences showing wildlife. Every row shows frames from a specific video followed by the row of saliency maps of respective frames.	134
6.7	Saliency maps obtained using the proposed sustained observability method on various video sequences in the database from [50] a)'flock' sequence (flock of birds as background) b) saliency map for 'flock' sequence c) 'skiing' sequence (snowfall as background) d) saliency map for 'skiing' sequence e) 'peds' sequence (static camera surveillance) f) saliency map for 'peds' sequence	135
6.8	Average equal error rates for different patch sizes used in the computation of regional saliency maps. X-axis shows the size of the square patch and Y-axis shows the average equal error rate in percentage . .	136
6.9	Improvement in saliency maps using combined pixel and region based methods over pixel only method. The saliency maps shown are binary maps with the same level of threshold applied to both pixel only maps and the combined pixel and region maps. a) original video frame b) Pixel saliency map c) Final saliency map which is a combination of pixel and region based maps	138

List of Tables

2.1	Comparison of different algorithms for saliency detection in images . . .	26
4.1	Average absolute difference of minimum hitting times in three different node pair categories.	90
4.2	F-measure comparison of different stages of the proposed method. . .	92
5.1	F-measure comparison for different methods.	112
6.1	Equal Error Rates (in %) for different algorithms for the dataset in [50] compared to the proposed method (SO).	139
6.2	Average Equal Error Rate comparison of different stages of the pro- posed method	140
•		

Chapter 1

Introduction

Digital cameras and camcorders have become ubiquitous over the past decade and have resulted in millions of images/videos being stored and shared across the web. The explosive growth of multimedia content has thrown open a variety of applications that calls for high-level analysis of the multimedia corpus. Over the past few years the user's requirement of an information retrieval system has moved away from traditional text retrieval systems to efficient and intelligent systems that can search for the required information in images, videos and audio segments. Videos and images in general are very rich in content and convey a lot of information. Given the huge amount of data to be processed, it is natural to explore ways by which a subset of the data can be judiciously selected so that processing can be done on this subset to achieve the objectives of a particular application. The information regarding the regions in an image or a video that can potentially attract the attention of humans is vital in this context as this can be a natural choice for the subset of the multimedia data.

We come across and 'look' at a lot of information in the form of images and videos in our daily life, but 'see' only a subset of it. It is important to differentiate the kind of information that we 'look' as well as 'see' from the kind of data that we 'look'

but do not 'see'. The tendency of humans to neglect many parts of the information they are fed with, can be attributed to the limited processing capability of the human brain. Due to these limitations, the human brain only 'sees' those parts of data that are considered 'relevant'. While the answer to 'what is relevant' may be loaded with subjectivity, it is nevertheless agreed upon that certain common low-level, bottom-up processing governs the task independent visual attention mechanism in humans. Such relevant or salient part of the information available in images and videos can assist many multimedia applications and offload the unnecessary burden of analyzing redundant data.

Salient regions in an image can be defined as the subset of the whole image which is considered relevant by the human visual system (HVS). The brain and the vision system in humans work in tandem to identify such regions. Identification of visual attention regions in images has been a keenly pursued topic by the computer vision community over the past few years. Researchers working on the identification of visual attention regions have generally progressed on two fronts: bottom-up and top-down approaches. Bottom-up approaches resemble fast, stimulus driven mechanisms in pre-attentive vision for salient region selection and are largely independent of the knowledge of the content in the image [37, 54, 64, 24]. For example, it is easy to spot a red dot among a group of green dots or a horizontal line among a bunch of vertical lines. Compared to bottom-up approaches, top-down approaches are goal oriented and make use of prior knowledge about the scene or the context to identify the salient regions [23, 22, 61]. Such methods are context dependent and demand more complete understanding of the scene. For example, context dependent methods will aim at giving more saliency for spotting a pedestrian on the middle of the road than on the walkway as the former is a rather unusual scenario. Salient regions in a video contain the motion of objects that attracts human attention. It can be understood along similar lines as image saliency with the additional factor of motion saliency.

1.1 Motivation

The motivation of this research comes from the usefulness of identifying visual attention regions in many potential applications in multimedia and computer vision. Multimedia consumption on mobile devices requires intelligent adaptation of the content to the real estate available on terminal devices. The detection of salient regions can be effectively utilized for multimedia content adaptation for small screen devices [12]. The zooming functionality can be made more user friendly by automatically adjusting the zoom location to the salient region of the image rather than the users themselves dragging the cursor to the interesting region. Similar intelligence can be brought to cropping of the images with the knowledge of salient parts of the image [70, 75]. Visual attention regions reduce the influence of cluttered background and enhances the performance of image retrieval systems [84]. Salient region detection can be also employed in object detection tasks as the algorithms for object detection can be benefitted from the prior information of visual attention regions. The information regarding the salient parts in an image also finds application in content aware image resizing [5] or image thumbnailing [51]. In such applications the distortion to the salient parts of the image is kept minimum to enhance the user viewing experience. With the exponential growth and easy availability of professionals and amateur videos, it has become imperative to develop intelligent methods to analyze videos and understand its content. The detection of salient motion in a video is quite important in this context as it can guide and help many other related tasks like object tracking and surveillance [96, 90], video retargeting [69] etc. Surveillance and tracking applications will be aided by the information on 'what to follow' with the knowledge of salient motion, while video retargeting applications can use the knowledge of salient motion to adapt multimedia content with minimum distortion. Although there are many applications of salient region detection mentioned above, we focus in this thesis only on developing models and algorithms for salient region detection. Existing al-

gorithms for saliency detection in images either neglect the global aspect of saliency in images or rely on supervised training on a limited dataset. Similarly, the state of the art in motion saliency detection is computationally intensive making it difficult for real time applications. The motivation of our research also derives from trying to address such voids in the existing algorithms in saliency detection.

1.2 Objectives

The primary objectives of the proposed work are to develop computational methods to

- detect and segment salient objects in images.
- detect regions of salient motion in videos in presence of complex background visuals.

We evaluate our proposed methods against extensive experimental data with ground truth information on salient objects in case of images and with the ground truth information on salient motion in the case of videos. The ground truth information is used to compare the proposed work with the state-of-the-art algorithms in saliency detection. This thesis also discusses how different features in an image can be combined in an unsupervised manner so as to enhance the performance of saliency detection algorithms.

1.3 Contributions

A brief overview on the contributions of the proposed thesis is given in this section.

- We introduce a computational method for detecting salient regions in images by modeling the distributions of color and orientation features. The proposed

framework consists of a color saliency framework and an orientation saliency framework. The final saliency map is selected from the either of the frameworks in a systematic way. The classic work from Itti et al. [37] on biologically plausible image saliency also consider color and orientation saliencies, but in a local manner, while our proposed method considers local and global aspects of saliency.

- A novel graph based method for segmenting salient objects in images by making use of random walk based hitting times for automatic labeling of salient and non-salient nodes in the image is introduced in the proposed thesis. A new graph representation of the image that represents the saliency between nodes more accurately - the pop-out graph model - is computed based on the knowledge of the labeled salient and background nodes. A semi-supervised learning technique is used to determine the labels of the unlabeled nodes by optimizing a smoothness objective label function on the newly created pop-out graph model along with weighted soft constraints on the labeled nodes.
- We propose an unsupervised method of feature selection that improves the performance of saliency detection algorithms. The noisy irrelevant features in the image are identified by maximizing the mixing rate of a Markov process running on a linear combination of various graphs, each representing a feature. The global optimum of this convex problem is achieved by maximizing the second smallest eigen value of the graph Laplacian via semi-definite programming.
- We introduce a linear dynamical system framework for the detection of salient motion in videos having complex background visuals. The detection of salient motion is based on the control theory concept of 'observability' from the output pixels, when the video sequence is represented as the output of a linear dynamical system. The group of output pixels with maximum saliency is further used

to model the holistic dynamics of the salient region. The pixel saliency map is bolstered with two region based saliency maps, which are computed based on the similarity of dynamics of the different spatio-temporal patches in the video with the salient region dynamics, in a global and local sense.

1.4 Organization

The thesis is organized into chapters that discuss various computational methods for saliency detection. Chapter 2 gives a brief overview of the literature in this area while the following two chapters discuss two different computational methods to detect and segment salient objects in images. Chapter 3 introduces a computational method for detecting salient regions in images by modeling the distributions of color and texture features while Chapter 4 describes a novel graph based method for segmenting salient objects. Chapter 5 proposes an unsupervised method to combine different features for enhancing the performance of saliency detection algorithms. Chapter 6 proposes a novel method to detect video saliency by modeling videos as output of linear dynamical systems. Chapter 7 concludes the work and also gives directions for future research in saliency detection.

Chapter 2

Literature Review

Detection and segmentation of salient regions in images and videos has been a topic of keen interest to the computer vision community for more than a decade. In this chapter we review the important contributions in this area.

2.1 Methods for Saliency Detection in Images

The methods for saliency detection in images have generally progressed on two main fronts : bottom-up and top-down approaches. Bottom-up approaches are largely independent of the knowledge of the content in the image and in general are fast to implement. On the contrary, top-down approaches are goal oriented and make use of prior knowledge about the scene or the context to identify the salient regions. In this section we will review both these approaches in the context of images. The first three subsections detail various kinds of bottom-up approaches while the last subsection discusses some of the major top-down approaches.

2.1.1 Biologically Inspired Methods

Certain visual attention models are inspired from the advances made in the field of psychology and neurosciences in interpreting the behaviour of the Human Visual System (HVS). The 'Feature Integration Theory' [79] proposed by Treisman is one such work that has helped researchers from the computer vision community to develop computational models for visual attention. Treisman proposed that, in the early vision, primary visual features are encoded in to separate feature maps which are later integrated into a unique saliency map. One of the early works inspired from the 'Feature Integration Theory' is that of Itti *et al.* [37], which evaluated the role of color and orientation features in deciding the salient regions in images. In [37], a center-surround difference operator on Red-Green and Blue-Yellow colors that imitates the color double opponent property of the neurons in receptive fields is used to compute the color saliency map. In addition to the color saliency map, it also computed an orientation saliency map and a luminance saliency map from the local contrasts in orientation and luminance values, respectively. The final saliency map is obtained by combining the color, orientation and luminance saliency maps. The method proposed by Itti is a bottom-up method that resembles fast, stimulus-driven mechanisms in pre-attentive vision for salient region selection. Walther and Koch extended Itti's model to infer the extent of a proto-object from the feature maps in [82], leading to the creation of a saliency toolbox.

Gao *et al.* [24] used the concept of discriminant saliency in a bottom-up method inspired by 'center-surround' mechanisms in pre-attentive biological vision [35]. Saliency of a particular location is equated to the power of the feature set around the location to discriminate between the 'center' and 'surround'. The center-surround discrimination defined in [24] is optimal in a decision-theoretic sense and is shown to produce better predictions of human eye fixations. In [8] Bruce *et al.* proposed a saliency detection

method by quantifying Shannon’s self-information measure on the local image patch. They further demonstrated that the proposed saliency method can be implemented in a neural circuit closely resembling the circuitry existent in the primate visual cortex. Wang *et al.* in [83] modeled early visual features with two sets of sparse coding bases learned from natural images in color and gray similar to the receptive fields in HVS. The input image is filtered using the sparse coding bases producing sub-band feature maps. A random walk technique used on the fully connected graph representation of the feature maps evaluates the total transmitted information of all the graph nodes similar to the model of signal transmission among neurons. The final saliency map is evaluated by accumulating the information transmission in all sub-band feature maps and is compared against a human eye fixation database. In general, biologically motivated methods build computational models for visual attention detection based on contributions from the fields of psychophysics and neurobiology with respect to interpreting the phenomenon of human visual attention. These computational models are primarily intended to predict human eye movements while observing an image and are tested against an existing eye fixation dataset. The complete detection of salient regions or segmentation of salient objects in an image are often not the aim of these methods. Moreover our incomplete understanding of Human Visual System (HVS) is a limitation to such methods.

2.1.2 Low-Level Feature Based Methods

Many researchers have undertaken computational approaches based on low-level features for detecting salient points in images even before the advent of the biologically inspired methods. Edge detectors and corner detectors have been proposed as interest point or salient point detectors [30]. Corner features are considered as relatively stable interest points under certain transformations like image rotation [72]. In [47], Lindberg proposed that blob-like structures that are stable in scale-space are likely

candidates to be salient parts in an image. Using the principles from scale-space theory, [47] detected the approximate location, spatial extent and the appropriate scale for the relevant regions in the image. Reisfeld et al. introduced a context free attention operator based on an intuitive notion of symmetry [67]. They assumed symmetric points are more prone to be attentive and developed a symmetry map based on gradients on grey level images. In [32], the idea is extended to color symmetry. The method focused on finding symmetric regions on color images which could be potentially missed by the grey scale algorithms. The underlying assumption in these methods is that symmetric regions are attention regions. However, for general salient region extraction, this is a very tight constraint. In [85], Weijer *et al.* proposed a novel way to incorporate color distinctiveness to saliency. The method used a technique called color boosting based on the analysis of statistics of color image derivatives and is meant to aid the existing salient point detection methods relying on shape distinctiveness. Weijer *et al.* showcased improved performance over Harris detector [30] in identifying the salient points with the help of color boosting. Kadir et al. retrieved ‘meaningful’ descriptors from an image from the local entropy or complexity of the regions in a bottom-up manner [41]. They considered saliency across scales and automatically selected the appropriate scales for analysis.

The low-level feature based methods are mainly intended to get the best ‘description’ of an image with a fixed number of interest points that can be used for further tasks like object matching, tracking etc. For example, the method proposed by Kadir et al. [41] used the local entropy as a clue to saliency by assuming that flatter histograms of a local descriptor correspond to higher complexity (measured in terms of entropy) resulting in higher saliency. The method is shown to be useful in matching fingerprint images by first detecting the salient points which are complex in its neighborhood. However for more generic salient region detection problems, this method

will not always hold well as in cases where a small smooth salient region (low complexity) is surrounded by a highly textured background (high complexity). The low level feature based method in [41] will erroneously show pixels in larger textured region as the salient. All the low -level feature based methods discussed here target to identify pixels with high saliency based on uniqueness of features in the local neighborhood of the pixels. Grouping of such interest points in view of segmenting out a salient region in a complete semantic sense is not an easy task. Hence such methods are primarily targeted towards matching or tracking applications rather than completely identifying or segmenting out salient regions/objects in an image.

2.1.3 Methods for Region/Object Segmentation

The methods discussed so far target to give saliency to pixels based on local neighborhood criteria and neglect the global aspects of saliency. In many cases it can happen that the pixels are not salient with respect to its feature values in its immediate neighborhood, but are part of a salient object or region as seen by an observer. Researchers modeling visual saliency have tried to address this issue in a variety of ways. Zhiwen et al. discussed a rule based algorithm which detects visual attention regions at the object level in [94]. They first performed image segmentation using a real time clustering algorithm and evaluated the confidence factor for each region based on a background measure that combines feature contrast and geometric properties of the regions. The confidence factors are further ranked to detect the hierarchical visual attention regions. In [49], Ma *et al.* proposed a method based on fuzzy event detection to get the 'areas' of attention from the 'points' of attention. A local feature contrast based method is used by [49] to get the initial attention points. Once the attended points are available from the local contrasts, they proposed a fuzzy region growing method to simulate the process of human perception from which the attended areas are extracted. In [33], a new linear subspace estimation method is

described to represent the 2D image after a polar transformation. The polar representation of the features is meant to ensure that all pixels within a homogeneous region are mapped into a single linear subspace that also encodes the location of the region. A new attention measure is then proposed based on the projection of all data to the normal of the corresponding subspace. The methods in [94], [49], and [33] aim to detect salient regions at the object level and try to bridge the gap between traditional visual attention regions and high-level semantics. The methods have an underlying simplified assumption of homogeneity of salient regions and are mostly tested on relatively simpler images with homogenous salient regions.

The role of spectral components of an image in deciding the salient regions is explored in [88] by Hou *et al.* The method worked with the log amplitude spectrum of the image and represented the gist of the scene with the averaged Fourier envelope. The differential spectral components between the original spectrum and the averaged spectrum are considered as 'innovations' corresponding to the salient regions in the scene. The phase spectrum is preserved in the process. This spectral-residue method will work satisfactorily for small salient objects but can fail for larger objects since the algorithm interprets the large salient object as part of the gist of the scene and consequently may fail to identify it. In [28], Guo *et al.* argued that the phase spectrum of the Fourier transform of the image is more effective and computationally more efficient than the spectral residue (SR) method of [88]. The method proposed by them is in principle similar to [88] by using phase spectrum instead of amplitude spectrum and is attractive only from a computational point of view. In [2], Achanta *et al.* discussed a frequency tuned approach with band-pass filters with the aim of highlighting the whole of large salient objects. They wanted to preserve the smooth salient regions (low spatial frequency) as well as the salient boundaries (high spatial frequency) in the final saliency map and hence used a combination of several bandpass

filters with contiguous pass bands. Liu et al., in [48], formulated the salient object detection problem as an image segmentation problem by extracting local, regional and global features from the image. They located salient regions with a bounding box obtained by learning local, regional and global features using conditional random fields. They defined a multi-scale contrast as the local feature, center-surround histogram as the regional feature and color spatial variance as the global feature. The optimal weights to combine different features are learnt from a group of training samples and the salient object detection problem is formulated as a binary labeling problem in a Conditional Random Field (CRF) framework. Khuwuthyakorn et al. [42] recently proposed a method that combines the salient features in an image by training a classifier model with linear Support Vector Machines (SVM). In their proposed method the feature space is partitioned into linearly separable regions with a mixture of SVMs and the parameter estimation is done using the Expectation-Maximization (EM) approach.

2.1.3.1 Salient Object Databases

Another major contribution of [48] is the user annotated database of 5000 images that can be used for objective level comparisons of various saliency detection algorithms. This is the first attempt to baseline the works on image saliency by providing the ground truth data for salient objects. Nine users are employed to draw a rectangle on the images in the database on the region which they thought as salient. The mean rectangle of the nine rectangles are considered as the ground truth. Though the method in [48] provided the ground truth salient regions for 5000 images, it has the drawback that the accuracy of the ground truth is limited to the accuracy of the bounding box. The actual segmentation information is not available in the ground truth. In [2], Achanta *et al.* addressed this problem to some extent by providing the groundtruth segmentation data for the images in the database of [48]. However, they

restricted the ground truth data to 1000 of the 5000 images in the original database. Elazary and Itti used 'LabelMe' database (74000 manually outlined objects in 25000 images) to evaluate how well interesting objects are predicted with visual saliency [20]. The authors considered the 'LabelMe' dataset to provide a good indication of what people would find generically interesting, in the absence of a particular task.

2.1.4 Top-Down Methods

The methods discussed so far use bottom-up strategy for detecting salient regions. Bottom-up methods do not try to understand the context or the visual scene category before trying to estimate the salient regions. In a top-down approach to saliency detection, the more or less complete understanding of the context of the image plays a vital role in the task dependent detection of salient regions. Oliva and Torralba [60] proposed a spatial envelop model to understand the context of the scene before detecting the probable locations of salient object in the scene. The spatial envelope model describes a scene in a holistic manner by attributing every scene with a set of perceptual properties like naturalness, openness, roughness etc. These properties are estimated from the second order statistics and the spatial arrangements of the structures in the scene. The contextual understanding of the scene is further used to estimate the likelihood of finding a particular object in a given scene (e.g. an animal in a road scene is less probable) along with the approximate location and scale of the object (e.g. probability of pedestrians on walkways in a road scene will be high) [61, 78, 62]. The authors argued that, as the saliency map computed using bottom-up methods attribute saliency to all objects having good contrasts unaware of the context, contextual information learned regarding the location and likely appearance of the object will guide to improve the final saliency map.

Gao *et al.* in [23] described a top-down method in which saliency is equated

Table 2.1: Comparison of different algorithms for saliency detection in images

Method	Features	Cons
<i>Itti et al.</i> [37]	Bottom-up, Local, Biologically Plausible, Unsupervised	Local Method
<i>Kadir et al.</i> [41]	Bottom-up, Histogram based local method, Unsupervised	Poor performance with cluttered backgrounds
<i>Ma et al.</i> [49]	Bottom-up, Contrast based local method, Unsupervised	Local Method
<i>Spectral Residual</i> [88]	Bottom-up, Global, Unsupervised	Neglected color factor, Issues with large salient objects
<i>Liu et al.</i> [48]	Bottom-up, Local + Global, Supervised	Dependent on the limited training data
<i>Oliva et al.</i> [61]	Top-down, Scene statistics	Less generic

to discrimination. Those features of a class that discriminate it from other classes in the same scene are defined as salient features. The method assumed that, for a given object, different visual attributes will be salient under different contexts. In [71], Schill *et al.* proposed a hybrid model in which a top-down knowledge based higher cognitive subsystem is used along with a bottom-up module to model the saccadic eye fixations on a scene. In the analysis of a scene, the system proposed in [71] calculates in each step the appropriate eye movement that has to be made to reach a maximum of information about the scene. In [22] Frintrop *et al.* performed a goal directed search in images by extending the bottom -up model in [37] with a top-down counterpart. The top-down module is provided with training images and co-ordinates of salient regions to learn the weights of different features involved in the

image. The features that best represent the target and differentiate it from the rest of the image are weighed large in the training process. The top-down methods are context dependent and by the very definition do not generalize well and is applicable under specific conditions. This is the reason for it being used in conjunction with bottom-up methods most of the times as we have seen in the examples discussed in this section. Table 2.1 provides a quick comparison of the certain major works in image saliency citing their main features and drawbacks.

2.2 Methods for Saliency Detection in Videos

Automated video analysis has been always looked upon as an important yet challenging task by the computer vision community. With the exponential growth of multimedia content including videos, it has become imperative to develop intelligent methods to analyze and understand its content. A 'salient region' in a video can be considered as a spatio-temporal region in the video which is interesting to the observers due to its spatial or temporal properties or both. While the spatial factor for saliency in a video will be similar to images, videos have the additional temporal factor for saliency which can be termed as 'Motion Saliency'. A significant amount of research has been done in understanding motion saliency in videos as it can guide and help many other related tasks like object tracking and surveillance [96, 90], video re-targeting [69] etc. Surveillance and tracking applications will be aided by the information on 'what to follow' with the knowledge of salient motion, while video re-targeting applications can use the knowledge of salient motion to adapt multimedia content with minimum distortion.

2.2.1 Optical Flow Based Methods

Optical flow based salient motion detection techniques make use of the optical flow feature estimated for every pixel over consecutive frames. In [87], Wixson defined the salient motion in a video as the consistent motion of a surveillance target as opposed to the distracting non-consistent motion like oscillation of vegetation in the background. The proposed method in [87] computed the saliency of each pixel based on the straight line distance traveled by it by considering the frame to frame optical flow. Salient motion is related to linear and coherent motions in [86] as well. Such assumption of straight line motion is too restrictive as foreground objects moving with some oscillations will be given less saliency with this method. Mittal *et al.*[56] utilized the optical flow feature along with the standard intensity information to differentiate between foreground and background. They assumed that the optical flow features computed are not perfect with many inherent ambiguities. A hybrid kernel based density estimation method in which the bandwidth of kernel is a function of both sample point and estimation point is used in [56] to address the uncertainties with the optical flow feature. Once the density approximation is built, a classification method is designed to threshold the probability of the new observation for deciding whether it belonged to the foreground or background. In [9], the consistently moving objects are detected as a group of pixels after compensating for the camera motion. The descriptors based on color and motion is used to characterize the 'moving pixels' and pixels with constant photometry and motion are clustered using a mean-shift filtering algorithm and a method for automatic kernel bandwidth selection. The method in [9] computes consistency in motion and photometry only over immediate frames and may not perform well with short term consistent background motions.

2.2.2 Auto-Regressive(AR) Model Based Predictive Methods

Some researchers have approached the foreground salient motion detection problem by representing the video sequence using auto-regressive(AR) models, as AR models approximate the dynamic characteristics of the background well [57]. Monnet *et al.* [57] used an online autoregressive model to predict the dynamic background and detection of salient motion is performed by comparing the prediction with actual observation. Similar to the dynamic texture model proposed in [17], [57] used the most recent 'k' frames in the video sequence to construct a predictive model that captures the most important variations using the subspace analysis of the signal. The differences in the structural appearance and the motion characteristics of the predicted and observed signal is used to detect the foreground from the background. In [95], the continuously changing background is predicted using dynamic texture ARMA model and the foreground objects are detected as outliers to this model. The method evaluates the initial state vector when there are no foreground objects in the scene and iteratively make estimates of the current state vector given the current observation. The foreground mask image is formed by estimating a weighing factor for every output in the formulation for minimization of the error of observed output and predicted output. The assumption of availability of background only scenes is a drawback of the methods discussed in [57] and [95]. Recently, [50] proposed a novel method for spatio-temporal saliency based on biologically motivated discriminant centre-surround saliency hypothesis. They used the dynamic texture model [17] to represent natural scene dynamics and KL divergence between dynamic texture models of center and surround windows is used to evaluate the saliency of a particular location. Though the algorithm performs well on a variety of complex video sequences, the accuracy of the algorithm is dependent on the size of the center and surround windows. Large foreground objects may not display good discriminant centre-surround

saliency in this method. The computational cost of [50] is also large. It is notable that [50] has done a quantitative analysis with other state-of-the-art methods by manually annotating the regions of salient motion over a reasonable number of natural video sequences and evaluating the algorithms on this ground truth.

Chapter 3

Modeling Image Saliency Using Distributions of Color and Orientation

3.1 Introduction

In this chapter we analyze the influence of color and orientation in deciding the salient regions of an image in a bottom-up fashion. The definition of saliency in the chapter revolves around two main concepts. 1) *Isolation (rarity)* and 2) *Compactness*. For colors, isolation refers to the rarity of a color in the color space while compactness refers to the spatial variance of the color in the image space. For orientations, isolation refers to the rarity of the *complexity of orientations* in the image space in a local sense, while compactness refers to the spatial confinement of the orientations in the image space. The role of color and orientation in deciding the salient regions has been studied previously also as can be seen in early works of Itti et al.[37]. Itti's works on color and orientation are motivated by the behavior of neurons in the receptive fields of human visual cortex [46]. The incomplete understanding of human visual

system brings limitations to such HVS based models. In the proposed work, we look at the distribution of color and orientations in the image separately, propose certain saliency models based on their respective distributions and design a robust framework to detect the final salient regions. The proposed framework can be divided into a Color Saliency Framework and an Orientation Saliency Framework. The final saliency map is selected from either one of these models in a systematic manner. Robust objective measures are employed on a large user annotated image database provided by Microsoft Research Asia [48] to illustrate the effectiveness of the proposed method.

The Color Saliency Framework (CSF) is motivated by the fact that the colors in an image play an important role in deciding salient regions. It has been shown in [73] that a purely chromatic signal can automatically capture visual attention. In [39], the contribution of color in visual attention is objectively evaluated by comparing the computer models for visual attention on gray scale and color images with actual human eye fixations. It has been observed that the similarity between the model and the actual eye fixation is improved when color cues are considered in the visual attention modeling. We believe that a color-based saliency framework should consider the global distribution of colors in an image as opposed to localized features described in [37] and [49]. In [37], a center-surround difference operator on Red-Green and Blue-Yellow colors that imitates the color double opponent property of the neurons in receptive fields is used to compute the color saliency map. Similarly in [49], the local neighborhood contrasts of the LUV image is used to generate the saliency map. A fuzzy region growing method is then used to draw the final bounding box on the saliency map. It can be seen that [37] and [49] use highly localized features and the global aspect of colors in the image has not been considered. For example, the saliency of a red apple in the background of green leaves is attributed more to the global distribution of color than to the local contrasts. The saliency of colors based

on its global behavior in the image has been discussed in literature in the works of Liu et al. [48] and Zhai et al.[92]. In [48], it is assumed that lesser the spatial variance of a particular color, the more likely it is to be salient. This assumption may fail when there are many colors in an image with comparable spatial variances but with varying saliencies. Zhai et al. [92] compute the spatial saliency by looking at the 1D histogram of a specific color channel (e.g. R channel) of the image and computing the saliency values of different amplitudes in that channel based on their rarity of occurrence in the histogram. This method focuses on a particular color channel and does not look at the relationship between various colors in the image. In [89], a region merging method is introduced to segment salient regions in a color image. The method focuses on image segmentation so that the final segmented image contains regions that are salient. However, their notion of saliency is related simply to the size of the region - if the segmented region is not large enough, it is not salient. Similarly in [65], a nonparametric clustering algorithm that can group perceptually salient image regions is discussed. The proposed work differs from [89] and [65], as ours is a saliency map generation method while [89] and [65] essentially focus on image segmentation so that the final segmented image contains different regions that are salient.

In the proposed color saliency framework we look at the saliency of colors rather than the saliency of regions. We introduce the concepts of 'compactness' of a color in the spatial domain and is 'isolation' in the H-S domain to define the saliency of the color. *Compactness* refers to the relative spread of a particular color with respect to other colors in the image. *Isolation* refers to the distinctiveness of the color with respect to other colors in the color space. We propose a measure to quantify compactness and isolation of various competing colors and probabilistically evaluate the saliency among them in the image. The saliency of the pixels is finally evaluated using probabilistic models. Our framework considers salient characteristics in the feature space as well as in the spatial domain, simultaneously.

The proposed Orientation Saliency Framework (OSF) identifies the salient regions in the image by analyzing the global and local behavior of different orientations. The role of orientation and of spectral components in an image to detect saliency have been explored in [37] and [88], respectively. In [37], the orientation saliency map is obtained by considering the local orientation contrast between centre and surround scales in a multi-resolution framework. However, a non-salient region with cluttered objects will result in high local orientation contrast and mislead the saliency computation. In [88], the gist of the scene is represented with the averaged Fourier envelope and the differential spectral components are used to extract salient regions. This method works satisfactorily for small salient objects but fails for larger objects since the algorithm interprets the large salient object as part of the gist of the scene and consequently fails to identify it. In [41], Kadir et al. consider the local entropy as a clue to saliency by assuming that flatter histograms of a local descriptor correspond to higher complexity (measured in terms of entropy) and hence, higher saliency. This assumption is not always true since an image with a small smooth region (low complexity) surrounded by a highly textured region (high complexity) will erroneously result in the larger textured region as the salient one. Hence, it is more useful to consider the contrast in complexity of orientations and to this end, we propose a novel orientation histogram as the local descriptor and determine how different its entropy is from the local neighborhood, leading to the notion of orientation entropy contrast. In the proposed orientation saliency framework, global orientation information is obtained by determining the spatially confined orientations and local orientation information is obtained by considering their orientation entropy contrast.

The goal of the proposed method in this chapter is to generate the **saliency map**, which indicates the salient regions in an image. It is not the objective here to segment a salient object, which can be viewed as a segmentation problem. Each of CSF and OSF creates a saliency map such that salient regions are correctly identified by the

Figure 3.1: The proposed framework for salient region detection in an image.

former if color is the contributing factor to saliency and by the latter if saliency in the image is due to orientation. From a global perspective, saliency in an image as perceived by the HVS is derived either from color or orientation. In other words, the visual attention region pops out due to its distinctiveness in color or texture. The final module in our framework predicts the appropriateness of the two saliency maps and selects the one that leads to the correct identification of the salient region. Even if a region in an image is salient due to both color and texture, our algorithm will detect the correct salient region since both the saliency maps will indicate them. Figure 3.1 shows the flowchart of the proposed framework for salient region detection using color and orientation distribution information.

The chapter is organized as follows. Section 3.2 describes the color saliency framework which computes the saliency of component colors of the image and the color saliency map. In Section 3.3, the saliency detection using the orientation distribution and the computation of orientation saliency map is discussed. Section 3.4 discusses the final saliency map selection from the color and orientation saliency maps. Experimental results and comparisons are briefed in Section 3.5. Finally, conclusions are given in Section 3.6.

3.2 Color Saliency Framework

This section describes the general framework used to compute color cluster saliency to extract the attention regions on color images. The first step involves identifying the dominant hue and saturation values in the image. The information from the dominant hues provides a systematic way to initialize the expectation-maximization (EM) algorithm for fitting a Gaussian mixture model (GMM) in the hue-saturation (H-S) space. The dominant color identification can also be done on the YCbCr space. However in HSV space the search process for dominant colors is made efficient

by searching for dominant hues in 1D space and evaluating the dominant saturation by considering the saturation values falling inside every 'hue peak'. If similar search has to be performed in YCbCr space, the search has to be done in 2D space of Cb and Cr. Hence for better computational efficiency HSV space is chosen. Using the mixture of gaussians model, the compactness or relative spread of each color in the spatial domain as well as the isolation of the respective cluster in the H-S domain is calculated. These two parameters determine the saliency of each cluster and finally, the color saliency of each pixel is computed from the cluster saliency.

3.2.1 Dominant Hue Extraction

The colors represented in the H-S space are modeled as a mixture of Gaussians for which the parameters are estimated using the EM algorithm. The critical issue in any EM algorithm is the initialization of the parameters, which in this case includes the number of Gaussians, their corresponding weights, and the initial mean and covariance matrices. If the parameters are not judiciously selected, there is a likelihood of the algorithm getting trapped in a local minimum. This will result in suboptimal representation of the colors in the image.

The initialization problem is solved by identifying the dominant hues in the image using a method similar to [58]. The number of bins in the hue histogram is initially set to 50. This is a very liberal initial estimate of the number of perceptible and distinguishable hues in most color images. In order to include the effect of saturation the hue histogram is defined as

$$H(b) = \sum_{(x,y) \in P_b} S(x,y), \quad (3.1)$$

where $H(b)$ is the histogram value in the b^{th} bin, P_b is the set of image co-ordinates having hues corresponding to b^{th} bin, and $S(x,y)$ is the saturation value at location (x,y) . The histogram is further smoothed using a sliding averaging filter that extends

over 5 bins, to avoid any secondary peaks. The maxima in the final histogram are the dominant hues and the number of dominant hues is set as the number of Gaussian clusters for the EM algorithm. The range of hues around a peak in the hue histogram constitutes a *hue band* for the corresponding hue. The local minimum (valley) between two dominant hues in the hue histogram is used to identify the boundary of the hue band. The saturation values pertaining to the dominant hue belonging to the k^{th} hue band is obtained as

$$S_k = \{S(x, y) : (x, y) \in D_k\} \quad (3.2)$$

where D_k is the set of pixels with hues from the k^{th} hue band and $S(x, y)$ is the saturation at (x, y) . The histogram of S_k contains a maximum which is taken as the dominant saturation value of the k^{th} dominant hue. For each hue band, the covariance matrix of the hue-saturation pair is obtained as

$$C_k = \sum_{(x,y) \in D_k} V_k V_k^T, \quad (3.3)$$

where

$$V_k = \begin{bmatrix} H(x, y) \\ S(x, y) \end{bmatrix} - \begin{bmatrix} H_k^{dom} \\ S_k^{dom} \end{bmatrix}, \quad (3.4)$$

$H(x, y)$ and $S(x, y)$ is the hue and saturation at (x, y) , respectively and H_k^{dom} and S_k^{dom} are the maximum hue and saturation values respectively in the k^{th} band. The histogram strengths of dominant hue values normalized between 0 and 1 are selected as the initial weights of the Gaussian clusters. Thus, we develop a systematic way to initialize the mixture of Gaussians for the EM algorithm. The number of clusters is the number of dominant hues, the initial mean of the Gaussians are the dominant hue and saturation values, the co-variances are determined from eq. (3.3), and the weights are relative strengths of dominant hues.

Figure 3.2 shows an example of the dominant hues extracted and the corresponding Gaussians that initialize the EM algorithm. Figure 3.2(a) is the original image. Figure

3.2(b) shows the palette of dominant hues in the image and the Gaussian clusters that initialize the EM algorithm are shown in figure 3.2(c). In [89], a color region segmentation algorithm for the YUV image is initialized by identifying the dominant Y, U and V values of the image leading to the processing of a 3-D histogram. The proposed method builds the color model for the image around the 1D histogram of hues. The method described in [89] can also be used to initialize our color saliency framework. However, since our probabilistic framework for color saliency is robust to small errors in the initial identification of dominant colors, we have used a simple method based on the 1-D hue histogram.

Figure 3.2: Initialization and fitting of Gaussian clusters. (a) Original image. (b) The dominant hue palette. (c) Initial Gaussian clusters in Hue-Saturation space. (d) Gaussian clusters after EM based parameter estimation. Note that in figures 3.2(c) and 3.2(d) data points lie in Hue - Saturation circle, where hue is represented by the positive angle between the line joining the origin (0,0) to the data point and the X-axis and saturation is represented by the radial distance from the origin (0,0).

3.2.2 Model Fitting in H-S Domain

The EM algorithm is performed to estimate the parameters of the GMM to the H-S space representation of the image. The dominant color extraction in the proposed framework is intended to make a reasonable initialization of the Gaussian clusters. Even if the number of clusters (colors) are estimated to be more than what is required, e.g., two clusters for different shades of green, it must be noted that each pixel in the image belongs to all clusters in a probabilistic sense. This will ensure that the final pixel saliency values computed in our framework will be still correct with the extra clusters present. However, the extra clusters increase the computational complexity of the framework. To reduce the computational complexity, we eliminate some of the extra clusters during EM based parameter estimation of the Gaussian model. This is done by inspecting the determinant of the covariance matrix of a cluster. If it falls below a predetermined threshold, the corresponding cluster is removed and the fitting process is continued. Hence the dominant color module only needs to supply a reasonable estimate of colors in the image. Figure 3.2(d) shows the final fitting of the GMM for the original image in figure 3.2(a).

When the modeling is complete, the probability of a pixel at location (x, y) belonging to the i^{th} color cluster is found as [48]

$$p(i|I_{x,y}) = \frac{w_i \mathcal{N}(I_{x,y}|\mu_i, \Sigma_i)}{\sum_i w_i \mathcal{N}(I_{x,y}|\mu_i, \Sigma_i)} \quad (3.5)$$

where w_i , μ_i , and Σ_i are the weight, mean value and covariance matrix respectively, of the i^{th} cluster.

The spatial mean of the i^{th} cluster is

$$\mu_i^{sp} = \begin{bmatrix} M_x(i) \\ M_y(i) \end{bmatrix} \quad (3.6)$$

where

$$M_x(i) = \frac{\sum_{(x,y) \in A_{xy}} p(i|I_{x,y}) \cdot x}{\sum_{(x,y) \in A_{xy}} p(i|I_{x,y})} \quad (3.7)$$

and

$$M_y(i) = \frac{\sum_{(x,y) \in A_{xy}} p(i|I_{x,y}) \cdot y}{\sum_{(x,y) \in A_{xy}} p(i|I_{x,y})}, \quad (3.8)$$

where $I_{x,y}$ is the H-S vector at location (x, y) and A_{xy} is the set of 2D vectors representing the integer pixel locations of the image.

As noted earlier, color saliency in a region is characterized by its compactness and isolation. Next, we define distance measures that quantify these two parameters.

3.2.3 Cluster Compactness (Spatial Domain)

The colors corresponding to the dominant hues and their respective saturation values will appear in the spatial domain as clusters that are spread to different extents over the image. It should be noted that these clusters of pixels in the image belong to different component colors in a probabilistic manner. The less the colors are spread, the more will be its saliency. The colors in the background have larger spread compared to the salient colors. This spread has to be evaluated by the distance in the spatial domain between the clusters rather than by intra color spatial variance alone, as in [48]. The reason is that there can be colors with similar intra spatial variance, but their compactness or relative spread may be different. The relative spread of a cluster is quantified by the mean of the distance of the component pixels of the cluster to the centroids of the other clusters. With the calculation of relative spread, we can not only eliminate the background colors of large spatial variance, but also bias the colors occurring towards the centre of the image with more importance, in case the other competing colors in the image possess similar spatial variance. The distance, d_{ij} , between cluster i and cluster j quantifies the relative compactness of cluster i

with respect to cluster j , i.e.,

$$d_{ij} = \frac{\sum_{X \in A_{xy}} \|X - \mu_j^{sp}\|^2 \cdot P_i(X)}{\sum_{X \in A_{xy}} P_i(X)}, \quad (3.9)$$

where μ_j^{sp} is the spatial mean of the j^{th} cluster, A_{xy} is the set of 2D vectors representing the integer pixel locations of the image, $P_i(X)$ is the probability that the H-S value in the spatial location $X \in A_{xy}$ falls into the cluster i .

Finally, the compactness $COMP_i$ of cluster i is calculated as the inverse of the sum of the distances of cluster i from all the other clusters, i.e.,

$$COMP_i = \left(\sum_j d_{ij} \right)^{-1}. \quad (3.10)$$

Next we look at the clusters in H-S domain representation of the image to calculate the isolation or remoteness of the component colors.

3.2.4 Cluster Isolation (H-S Domain)

Isolation of a color refers to how distinct it is with respect to other colors in the image. The distinction naturally contributes to the saliency of a region. The inter-cluster distance in H-S space quantifies the isolation of a color. As the image is represented as a mixture of Gaussians in the H-S space, the most isolated Gaussian is more likely to capture the attention. The inter-cluster distance between clusters i and j that quantifies the isolation of a particular color is defined as

$$h_{ij} = \frac{\sum_{Z \in S_{xy}} \|Z - \mu_j^{hs}\|^2 \cdot P_i(Z)}{\sum_{Z \in S_{xy}} P_i(Z)}, \quad (3.11)$$

where μ_j^{hs} is the mean of the H-S vector of the j^{th} cluster, S_{xy} is the set of 2D vectors representing the hue and the corresponding saturation values of the image, and $P_i(Z)$ is the probability that the H-S value falls into the cluster i .

The isolation ISO_i of cluster i is calculated as the sum of the distances of cluster i from all other clusters in the H-S space, i.e.,

$$ISO_i = \sum_j h_{ij}. \quad (3.12)$$

3.2.5 Pixel Saliency

The saliency for a particular color cluster is computed from its compactness and isolation. A more compact and isolated cluster is more salient. Hence, the saliency Sal_i of a color cluster i is

$$Sal_i = ISO_i \times COMP_i. \quad (3.13)$$

The saliency values of the color clusters are normalized between 0 and 1. The top down approach, taking the global distribution of colors into account, results in propagating the cluster saliency values down to the pixels. The final color pixel saliency Sal_{Cp} , on pixel p in location $X \in A_{xy}$ is computed as

$$Sal_{Cp} = \sum_i Sal_i \cdot P_i(X), \quad (3.14)$$

where $P_i(X)$ is the probability that the H-S values at $X = (x, y)$ falls into the cluster i . The pixel saliencies are once again normalized to obtain the final saliency map.

Figure 3.3 shows preliminary results to demonstrate the effectiveness of the proposed color saliency framework, when color is indeed the contributing factor to saliency. Figures 3.3(a) and 3.3(c) show original images and Figures 3.3(b) and 3.3(d) show their corresponding color saliency maps. The color saliency maps are in agreement with what can be considered as visual attention regions for the HVS.

3.3 Orientation Saliency Framework

From a global perspective, certain colors or orientation patterns provide important cues to saliency in an image. It is this aspect of saliency that we wish to capture in

Figure 3.3: (a, c) Original images; (b, d) Color saliency maps.

the proposed salient region detection framework. In cases where color is not useful to detect salient regions, it is some property of the dominant orientation in the image that allows such regions to be identified. For example, in the images in figure 3.4, there is no salient color, whereas the wings of the bird in figure 3.4(a) and the orientation of the buildings in the background in figure 3.4(b) act as markers to aid in the detection of salient objects. Note that in the second case, the dominant orientation actually contributes to the non-salient region so that the structure in the foreground is detected as the salient region.

Figure 3.4: (a, b) Example images in which color plays no role in saliency.

Determination of saliency in an image due to orientation has been previously studied in works like [37] and [38]. However, these methods consider orientation only as a local phenomenon. In the proposed orientation saliency framework, we model both the local as well as the global manifestations of orientation. In particular, we propose a novel orientation histogram of image patches that characterizes local orientations and their distribution over the image is used to model the global orientation. Furthermore, the orientation entropy of each patch is compared against the neighborhood patches and a high contrast is deemed to be more salient. The final orientation saliency map is obtained by a combination of the global orientation information obtained from the distributions and the local orientation information obtained from the entropy contrasts.

Figure 3.5: Orientation histogram from Fourier spectrum. (a) Original image, (b) Centre-shifted Fourier transform, (c) Orientation Histogram.

3.3.1 Local Orientation Histogram

Local gradient strengths and Gabor filters are two commonly used methods for constructing the orientation histogram of an image [15, 19]. We introduce a new and effective method to compute the orientation histogram from the Fourier spectrum of the image as a prelude to our orientation saliency framework. The absolute values of

the complex Fourier coefficients indicates the presence of the respective spatial frequencies in the image and their strengths in a specific direction of the centre shifted Fourier transform is an indication of the presence of orientations in a perpendicular direction [25]. Hence, the Fourier transform of an 8×8 patch in an image is computed and the orientation histogram is found as

$$H(\theta_i) = \sum_{\tan^{-1}(m^*/n^*) \in \theta_i} \log(|F(m, n)| + 1), \quad (3.15)$$

where $|F(m, n)|$ is the absolute Fourier magnitude in location (m, n) , (m^*, n^*) are shifted locations with respect to the centre of Fourier window for the purpose of calculating the orientation of the spectral components and are obtained by subtracting the centre of the Fourier window from (m, n) , and θ_i is the set of angles in the i^{th} histogram bin. The orientations ranging from -90° to 90° are divided into bins of 10° each. Logarithm of the Fourier transform is used to reduce the dynamic range of the coefficients. Since the Fourier transform is symmetric over 180° , only the right half of the centre shifted Fourier transform is used to build the orientation histogram.

In order to remove the dependency of the high frequency components on the patch boundary, the orientation histogram is calculated as the average histogram on the patch over three different scales of the image, $\sigma_i \in \{0.5, 1, 1.5\}$. The scale space is generated by convolving the image with Gaussian masks derived from the three scales. Figure 3.5 shows a simple example in which Fourier transform is used to obtain the orientation histogram of the image. Figure 3.5(a) shows an image from the Brodatz texture album. Its centre shifted Fourier transform is shown in Figure 3.5(b), and the orientation histogram is shown in Figure 3.5(c). The orientation histogram peaks at two different angles with one direction being dominant over the other as it is evident from the original image. It is noted that for a patch containing a smooth region there will be no dominant orientation and the histogram will be almost flat. Hence the orientation histogram is modified by subtracting the average orientation

magnitude followed by normalization between zero and one. Thus, in the strict sense, the orientation histogram loses the property of a histogram, but it does provide a useful indication of the dominant orientations present in the image.

3.3.2 Spatial Variance of Global Orientation

The local orientation histogram is used to analyze the global behavior of the various orientations in the histogram. The orientations that exhibit large spatial variance across the images are less likely to be salient. For example, in the images of Figure 3.6, the orientations in the salient regions are confined in a local spatial neighborhood as compared to the rest of the image. The dominant orientations in the background contributed by the door and by the ground in the images in figure 3.6(a) have a large spatial variance and hence are not salient. This property can be objectively measured by analyzing the global behavior of the different orientations ranging from -90° to 90° . The global spatial mean of orientation θ_i is computed as

$$M_x(\theta_i) = \frac{\sum_{(x,y)} x \cdot H(\theta_i)}{\sum_{(x,y)} H(\theta_i)}, \quad (3.16)$$

$$M_y(\theta_i) = \frac{\sum_{(x,y)} y \cdot H(\theta_i)}{\sum_{(x,y)} H(\theta_i)}, \quad (3.17)$$

where x and y are the image co-ordinates. The spatial variance of orientation θ_i in the x and y directions are then computed as

$$Var_x(\theta_i) = \frac{\sum_{(x,y)} (x - M_x(\theta_i))^2 \cdot H(\theta_i)}{\sum_{(x,y)} H(\theta_i)}, \quad (3.18)$$

and

$$Var_y(\theta_i) = \frac{\sum_{(x,y)} (y - M_y(\theta_i))^2 \cdot H(\theta_i)}{\sum_{(x,y)} H(\theta_i)}, \quad (3.19)$$

and the total spatial variance $Var(\theta_i)$ is calculated as

$$Var(\theta_i) = Var_x(\theta_i) + Var_y(\theta_i). \quad (3.20)$$

$Var(\theta_i)$ is a direct measure of the spatial confinement of the i^{th} orientation and hence, its relative strength among the various orientations is used as a direct measure of saliency of the respective orientations. In other words, the saliency of an orientation θ_i , denoted by $Sal(\theta_i)$, is its total spatial variance $Var(\theta_i)$ and it is normalized so that the orientation having least spatial variance receives a saliency of one and that with maximum variance receives a saliency of zero. Finally, the saliency of a local patch P is calculated as

$$Sal_{gP} = \sum_i Sal(\theta_i).H_P(\theta_i) \quad (3.21)$$

where $H_P(\theta_i)$ is the contribution of orientation θ_i in the patch P .

Figure 3.6: (a) Original images. (b) Saliency maps extracted by computing spatial variance of global orientation.

We evaluate the effectiveness of the proposed global salient orientation algorithm on the images in Figure 3.6(a). As expected, the first saliency map in figure 3.6(b) shows the boy as the salient region ignoring the background orientation that has a larger spatial variance while the second saliency map in figure 3.6(b) shows the vehicle

as the saliency region while excluding the ground in the background.

3.3.3 Orientation Entropy Contrast

The spatial variance of the orientations helps in detecting salient objects by virtue of their confinement in the spatial domain and the exclusion of other dominant orientations due to their large variance. However, images in which there are several dominant orientations from the foreground and the background that are confined and hence result in low spatial variance, will compete for saliency according to the method in the previous section. In this case, it is required to determine which among those competing orientations has higher saliency. We next introduce the orientation entropy contrast to address this issue.

The entropy of the orientation histogram of a patch is considered to be a measure of the complexity of the orientation in that patch. Higher entropy indicates more uncertainty in the orientations; this can happen in smooth as well as in highly textured regions with dominant orientations in several directions. Similarly, low entropy is an indication of strong orientation properties contributed by edges and curves. The orientation entropy E_P of patch P with orientation histogram H_P is calculated as

$$E_P = - \sum_{\theta_i} H_P(\theta_i) \cdot \log(H_P(\theta_i)). \quad (3.22)$$

The orientation complexity is also calculated at three different scales, $\sigma_i \in \{0.5, 1, 1.5\}$, to reduce the dependencies on the selected patch size. Hence the three dimensional vector $E_P = [E_{\sigma_1}, E_{\sigma_2}, E_{\sigma_3}]$ represents the final entropy vector on patch P of the image, where E_{σ_i} represents the orientation entropy calculated at scale σ_i .

Saliency of a local patch is now calculated based on the norm of the vector difference between its entropy and the entropy of the neighboring patches. Such a measure of entropy contrast is better than [37] since it is not biased towards edges, smooth regions or textured regions. Any region with such characteristics can be salient de-

pending on the entropy of the neighboring regions. Hence, the final criterion of salient region selection is how discriminating a particular region is, in terms of its orientation complexity. Furthermore, in the spatial domain, as the Euclidian distance between the patches increases, the influence of the entropy contrast decreases. Integrating the distance factor into the saliency computation, we obtain the saliency of a patch P , Sal_{lP} , based on local orientation entropy contrast as

$$Sal_{lP} = \sum_{Q, Q \neq P} \frac{\|E_P - E_Q\|}{\|D_{PQ}\|}, \quad (3.23)$$

where E_Q is the orientation entropy vector of patch Q and $\|D_{PQ}\|$ is the Euclidian distance between the centers of patches P and Q . For a particular patch of size 8×8 , we consider all the remaining patches in the image as neighborhood patches. However, their contribution to the orientation saliency is inversely weighted by the Euclidean distance between the neighborhood patch and the particular patch under consideration. In this way, the major contribution to the entropy contrast will come from spatially neighboring patches. The saliency of a patch is normalized similar to section 3.3.2.

Figure 3.7(a) shows two images in which there are dominant textures in the background that are confined to a spatial region so that their spatial variance is small. However, the contrast in entropy of such regions with the surrounding regions is not very high. On the other hand, the foreground objects exhibit high distinctiveness in its orientation complexity with the surrounding regions. These objects are rightly determined as salient regions in the salient maps shown in Figure 3.7(b).

3.3.4 Final Orientation Saliency

The two approaches adopted for orientation based saliency complement each other. The first method using spatial variance of global orientation searches for spatially confined orientations and in the process detects salient objects as well as cluttered

Figure 3.7: (a) Original images (b) Saliency map based on orientation entropy contrast method.

non-salient regions. The cluttered regions will have low contrast of orientation entropy and hence will be eliminated by the second method that used orientation entropy contrast. At the same time, long non-salient edges in the images like horizon, fence etc. can possess high orientation entropy contrast depending on the neighborhood. But those orientations will have high spatial variance and hence will be eliminated by the global orientation variance method. For example, in the boy image in figure 3.6(a), the latch and the planks on the wooden door has high orientation entropy contrast, but is eliminated by the first method. Similarly, in the images in figure 3.7(a) the background is rich in orientations that are confined in their spatial neighborhood. This misleads the global variance computation, but the orientation entropy contrast rightly eliminates such regions. The final orientation saliency of patch P is computed as

$$Sal_{OP} = Sal_{gP} \times Sal_{lP}. \quad (3.24)$$

It will have a maximum value on locations where both the methods agree and vary proportionately in regions that favor only one of the methods.

3.4 Saliency Map Selection

The color saliency framework and orientation saliency framework produce two different saliency maps. Previous methods such as [37] and [34] fuse the individual saliency maps corresponding to each feature based on weights assigned to each map. Our premise is that *salient regions in most images can be globally identified by certain colors or orientations*. Since saliencies occur at a global level rather than at a local level, it is not feasible to combine the saliency maps through weights, for this may result in a salient region detection with less accuracy than when individual maps are considered. We introduce a systematic way by which the correct saliency map can be estimated so that the attribute due to which saliency is present - color or orientation - is faithfully captured.

The saliency maps are binarized by Otsu's thresholding method [63]. The saliency map in which the salient regions are more *compact* and form a *connected* region will be the one that better identifies the salient region in the image. The *compactness* of a region is objectively evaluated by considering the saliency map as a probability distribution function and evaluating its variance as

$$SalMap_{Var} = SalMap_{VarX} + SalMap_{VarY}, \quad (3.25)$$

where

$$SalMap_{VarX} = \frac{\sum_{(x,y)} (x - Sal_x)^2 \cdot Sal(x, y)}{\sum_{(x,y)} Sal(x, y)}$$

and

$$SalMap_{VarY} = \frac{\sum_{(x,y)} (y - Sal_y)^2 \cdot Sal(x, y)}{\sum_{(x,y)} Sal(x, y)} \quad (3.26)$$

are the spatial variances of the binary salient regions in the x and y directions, respectively, and

$$Sal_x = \frac{\sum_{(x,y)} x \cdot Sal(x, y)}{\sum_{(x,y)} Sal(x, y)},$$

and

$$Sal_y = \frac{\sum_{(x,y)} y \cdot Sal(x,y)}{\sum_{(x,y)} Sal(x,y)}, \quad (3.27)$$

are the spatial means of the salient regions in the x and y directions, respectively, with $Sal(x,y)$ denoting the saliency value at location (x,y) .

The *connectivity* of the saliency map is measured by considering the number of salient pixels in the 8-point neighborhood of a salient pixel p in the binary saliency map and is evaluated as

$$SalMap_{Con} = \frac{\sum_{p \in ISal} \sum_{(x,y) \in N_p} \mathbf{I}_i(p_{xy})}{|\{ISal\}|} \quad (3.28)$$

where $\{ISal\}$ is the set of salient pixels in the saliency map, $|\{ISal\}|$ is the cardinality of the set, $\mathbf{I}_i(p_{xy})$ is the indicator function denoting whether the pixel p at location (x,y) is a salient pixel, N_p is the set of co-ordinates in the 8-point neighborhood of pixel p .

The saliency map with high connectivity and less spatial variance is estimated to represent a meaningful salient object. Hence, the final measure, *SalIndex*, used to select between color and orientation saliency maps is

$$SalIndex = \frac{SalMap_{Con}}{SalMap_{Var}}. \quad (3.29)$$

The saliency map with higher *SalIndex* is chosen as the final saliency map. The reason for binarizing the saliency maps was to facilitate the selection of either the color saliency map or the orientation saliency map through efficient computation of the spatial variance and the connectivity of each saliency map. This computation will not be reasonable on the original (non-binary) saliency maps since the saliencies take on continuous values in the grayscale and the notion of connectivity of a region is not clear in such a case. Moreover, the saliency values in both the maps have different dynamic ranges and therefore it is better to binarize the maps in order to compare the spatial variance. As demonstrated in the next section, our experiments on a large database of images annotated with 'ground truth', validate the above method.

3.5 Experiments

The experiments are conducted on a database of about 5,000 images provided by Microsoft Research Asia [48]. As described previously, the MSRA database used in this chapter contains the ground truth of the salient region marked as bounding boxes by nine different users. Similar to [48], a saliency probability map (SPM) is obtained by averaging the nine users' binary masks. The performance of a particular algorithm is objectively evaluated by its ability to determine the salient region as close as possible to the SPM. We carry out the evaluation of the algorithms based on Precision, Recall and F-Measure. *Precision* is calculated as ratio of the total saliency (sum of intensities in the saliency map) captured inside the SPM to the total saliency computed for the image. *Recall* is calculated as the ratio of the total saliency captured inside the SPM to the total saliency of the SPM. *F-Measure* is the overall performance measurement and is computed as the weighted harmonic mean between the precision and recall values. It is defined as

$$F\text{-Measure}_\alpha = \frac{(1 + \alpha) * Precision * Recall}{(\alpha * Precision + Recall)}, \quad (3.30)$$

where α is real and positive and decides the importance of precision over recall while computing the *F-Measure*. While absolute value of precision directly indicates the performance of the algorithms compared to ground truth, the same cannot be said for recall. This is because, for recall we compare the saliency on the area of the salient object inside the SPM to the total saliency of the SPM. However, the salient object need not always fill the SPM. Yet, calculation of recall allows us to compare our algorithm with others in a relative manner. Under these circumstances, the improvement in precision is of primary importance. Therefore, while computing the F-measure, we weight precision more than recall by assigning $\alpha = 0.3$. However, in order to substantiate the fact that a particular choice of α does not affect the comparative performance of our proposed algorithm, we also display the F-measure values of the

proposed algorithm and other competing methods for α ranging from 0.2 to 0.8. The following subsections discuss the various intermediate and final results of the proposed framework and present comparisons with some existing popular techniques for salient region detection.

3.5.1 Experiment 1 : Color Saliency

First, we evaluate the color saliency framework, preliminary results for which were presented in section 3.2.5. Figures 3.8(a) and 3.8(c) show some of the original images selected from the MSRA database of 5,000 images. Figures 3.8(b) and 3.8(d) show the color saliency maps. It can be observed that color plays a crucial role in deciding salient regions in the images shown in the figures 3.8(a) and 3.8(c). The color saliency maps in figure 3.8(b) and 3.8(d) are shown without any threshold. We see that the salient regions are very accurate in relation to what the HVS would consider to be salient.

In Figure 3.9, a comparison with the color spatial variance feature described in [48] is illustrated. As shown in Figure 3.9(b), there are many clusters that have comparable variances for the original images in figure 3.9(a). Hence, if intra-color spatial variance is used as feature to model the color saliency, the neighboring regions of the salient regions are also extracted. The proposed color saliency framework detects a more accurate salient region as shown in Figure 3.9(c). Thus, if the saliency in an image is dominated by color content, the global color modeling proposed in this chapter detects salient regions with sufficient accuracy. A center-weighted spatial variance feature is also introduced in [48] to eliminate regions with small variances close to the boundaries of the image since they are less likely to be the part of a salient region. But this assumption may not be true in all cases since salient regions can be present along the image borders. In our method, the color regions occurring near the image boundary having less color spatial variance may have large inter-

Figure 3.8: Salient region detection by the color saliency framework. (a, c) Original images. (b, d) Color saliency maps.

cluster distance in the spatial domain and will be automatically eliminated without any center-weighting. However, if they are isolated in the hue-saturation space, they are more likely to be salient and the proposed color saliency framework will detect such regions.

Figure 3.9: (a) Original images. Color saliency maps using (b) intra color spatial variance feature [48], and (c) the proposed color saliency frame work.

3.5.2 Experiment 2 : Orientation Saliency

We use the same database of images as was used in the previous experiment to test the effectiveness of the orientation saliency framework. All images are resized to 128*128 before giving as input to Orientation Saliency Framework. We have conducted experiments with patch sizes 4*4 , 8*8 and 16*16 on the MSRA dataset. It has been found that the F-measure of 8*8 patch size is better by 3% compared to that of 16*16 patch size and is better by 6% compared to that of 4*4 patch size. As all images are resized to same size before orientation saliency evaluation, the experimental choice of a certain patch size implies that certain structural patterns are discerning in most of the images (in the dataset used) within the chosen patch size. We have also performed experiments by evaluating orientation histograms under different image scales($\sigma_i =$

0.5,1.0,1.5,2.0,2.5). It is observed that the F-measure accuracy variation is negligible with the addition of more scales with a selection of 3 scales ($\sigma_i = 0.5, 1.0, 1.5$) performing better than other selections by over 1%.

Initially, we consider the methods involving spatial variance of the orientations and the orientation entropy contrast separately in order to ascertain their individual performance. Figure 3.10 shows some images in which the salient regions are extracted based on the spatial variance of the dominant orientations. Figure 3.10(a) and 3.10(c) show the original images that contain salient objects having compact orientations and the extracted salient regions are shown in Figure 3.10(b) and 3.10(d). In all of these images, the salient region cannot be distinguished by color since the contrast between the salient object and the background is too little for it to pop out.

Although the contrast in orientation entropy has been shown to be useful for detection of salient regions in some images in Figure 3.7, we observe that it serves more as a critical complement to the spatial variance of orientation in order to determine the final orientation saliency. We illustrate this fact through results in Figure 3.11. The original images are shown in Figure 3.11(a), and the orientation saliency maps obtained by the spatial variance of the orientation and the contrast in orientation entropy methods are shown in Figure 3.11(b) and Figure 3.11(c). Individually, each of the approaches returns saliency maps that are different; however, when they are combined according to equation (3.24), those regions in which both the approaches are in agreement are re-inforced so that the final orientation saliency map contains the accurate salient region, as shown in Figure 3.11(d).

The improvement obtained by combining the two orientation saliency determination frameworks is quantified by computing the precision values achieved for the MSRA database. The average precision obtained by the global variance method, the entropy contrast method and the final orientation saliency obtained by combi-

Figure 3.10: (a, c) Original images. (b, d) Saliency map based on spatial variance of global orientation.

Figure 3.11: (a) Original image. Saliency map based on (b) spatial variance of global orientation (Sal_{gP}) and (c) orientation entropy contrast (Sal_{IP}). (d) Final orientation saliency (Sal_{OP}).

nation through equation 3.24 are 0.61, 0.46 and 0.67 respectively. As expected, the combination of the two methods results in the highest precision.

3.5.3 Experiment 3 : Final Saliency Map Selection

The previous experiments showed the results obtained when color saliency and orientation saliency frameworks are applied separately to the selected image data set. Now we move on to the final stage of our proposed framework. We convert the color and orientation saliency maps to binary maps for the purpose of final saliency calculation as described in section 3.4. The saliency map with higher *SalIndex* as defined in equation (3.29) is selected as the final saliency map. Figure 3.12(b) and Figure 3.12(c) shows the binary color and orientation saliency maps respectively of the original images shown in Figure 3.12(a). The final saliency map in Figure 3.12(d) is selected based on the *SalIndex* of the binary color and orientation maps. It can be verified from the images in Figure 3.12(d) that the final saliency map selection based on *SalIndex* is justified. The first image shown in Figure 3.12(a) is salient both in color as well as in orientation. However, based on the *SalIndex* value, the orientation saliency map is selected as the final saliency map in the first image. The next two images are color salient while the last two are orientation salient. The color and orientation saliency maps of the images detect the salient regions accordingly. Recall that we model the global saliency according to the premise that salient regions in most images are either color salient or orientation salient. Even if the saliency is due to both color and orientation, the proposed method will indicate the correct salient region as shown in the first image of Figure 3.12(a).

Figure 3.13 shows the comparison of average precision for images in MSRA database obtained with the color saliency framework, orientation saliency framework and when the final saliency map is selected to be one of color or orientation. It is clear that the selection results in better performance than the individual methods. This is due

Figure 3.12: (a) Original images. (b) Binary color saliency map. (c) Binary orientation saliency map. (d) Final saliency map, which corresponds to the binary color or orientation saliency map with higher *SalIndex*.

to the fact that the color saliency framework and the orientation saliency framework complement each other quite well. The final map is selected mostly from the framework that returns the true salient region, as can be seen in the examples in Figure 3.12.

Figure 3.13: Comparison of average precision values of color saliency framework, orientation saliency framework and the combined framework.

3.5.4 Comparison with other saliency detection methods

The proposed method is an unsupervised, stand-alone method for saliency detection. Hence we have selected the state-of-the art unsupervised methods which do not include any segmentation algorithm for saliency detection for comparison. The performance of the proposed framework is compared with Saliency Tool Box (STB) (latest version available at www.saliencytoolbox.net) [81], contrast based method [49] and spectral residual method [88]. The user annotation in the MSRA database facilitates objective comparisons of different algorithms. The precision, recall and F-measure of various methods computed for the MSRA database is shown in Figure 3.14. The recall of all methods keep generally lower values due to the reasons explained at the beginning of this section. It can be seen from the figure that the proposed method performs consistently better with respect to precision, recall and

F-measure. Figure 3.15 shows f-measure values of all methods for various values of α ranging from 0.2 to 0.8. It can be seen from figure 3.15 that the proposed method performs consistently better irrespective of the choice of α . We have also compared the performance of the proposed method with the other existing methods in the Berkeley image database [52] as well. The saliency map comparisons of the proposed method and other methods on selected images from Berkeley database is shown in Figure 3.16. It is evident that the proposed method outperforms the other methods. For example, in row (6), our method detects the salient flower regions correctly while the spectral residual method does not take color into account and fails. The proposed method has rightly picked up the red color region as the salient region. Similarly, in row (13), disconnected salient regions corresponding to the dolphins have been accurately detected. It is not possible to compute the precision of saliency maps for the Berkeley image database since no ground truth is available for the salient region; instead the database contains the ground truth only for image segmentation results. Note that the proposed method is independent of the size of the salient object as it looks for salient features. The regions containing the salient features get more saliency in the saliency map. If such regions are small as a consequence of a small salient object, the saliency map will indicate that. The same logic holds true for large salient objects.

3.5.5 Failure cases and analysis

As stated previously, the proposed framework detects salient regions under the assumption that from a global perspective, certain colors or orientations contribute to defining a region as salient. This assumption may fail in those images in which saliency is decided by more complex psycho-visual aspects influenced by prior training in object recognition. Figure 3.17 shows two such examples. The bird in figure 3.17(a) is not salient due to its color or orientation as shown in the corresponding maps in figure 3.17(b) and (c), but it is still a salient object due to the object recog-

Figure 3.14: Comparison of average precision, recall and f-measure values of proposed framework, spectral residual method [88], saliency tool box [81] and contrast method [49]. Horizontal axis shows 1) Precision 2) Recall 3) F-Measure

Figure 3.15: Comparison of f-measure values of proposed framework, spectral residual method [88], saliency tool box [81] and contrast method [49] for different values of α .

Figure 3.16: Saliency map comparisons in Berkeley image database (a) Original image. Saliency maps generated by (b) proposed method, (c) spectral residual method [88], (d) saliency tool box [81], and (e) contrast based method [49].

nition capability of the trained human brain. For similar reasons, the reptile in figure 3.17(a) is salient; the color and orientation maps do not indicate it as a salient region.

Figure 3.17: Failure Cases. (a) Original image. (b) Color saliency map. (c) Orientation saliency map.

3.6 Conclusion

We have presented a framework for salient region detection which consists of two parts - color saliency and orientation saliency. For most images, a region turns out to be salient due to these two important factors and this fact is captured in the proposed framework. The properties of the color and orientation distributions of the image are explored in the image domain as well as in the feature domain. This is in contrast to previous works which consider either one of the spaces. The new notion of isolation of a color in the color space of the image and its relation to saliency has been explored. The choice of HSV space is done for dominant color identification for the purpose of initializing the Gaussian Mixture Model. Experiments were conducted with MSRA dataset using HSV and YCbCr colorspace. It has been observed that the change of colorspace do not result in a significant difference in F-Measure (1%

increase in F-measure with the choice of HSV color space). In terms of orientation, we propose a new mechanism to determine local and global salient orientations and combine them to yield an orientation salient region. As opposed to previous works which fuse saliency maps through weights, we propose a simple selection of either color or orientation saliency map. The robustness of the proposed framework has been objectively demonstrated with the help of a large image data base.

Chapter 4

A Graph Based Method for Salient Region Detection in Images

4.1 Introduction

In this chapter, we consider the properties of random walks on graph representation of images and its relationship to saliency. We formulate the problem of salient object detection in images in a bottom-up manner as an unsupervised learning problem on the vertices of a weighted graph. In the weighted graph representation of an image, the vertices of the graph correspond to fixed size patches in the image and the edge strengths correspond to the feature similarities between the patches. Such a graph representation of an image captures the geometry of different components in the image. The proposed work in this chapter deals with random walks on global and local graph representation of the image and automatically identify the labels of certain seed nodes (patches) that are more likely to form the part of salient object and the background. A constrained learning framework is further used to get the labels of rest of the unlabeled nodes in the graph.

Previous works on detecting salient regions from images represented as graphs

include [14] and [29]. In [14], Costa presents two models in which random walks on graphs enable the identification of salient regions by determining the frequency of visits to each node at equilibrium. While some results are presented on only two synthetic images, there is no evaluation of how the method will work on real images. A similar approach in [29] uses edge strengths to represent the dissimilarity between two nodes; the strength between two nodes decreases as the distance between them increases. Here, the most frequently visited node will be most dissimilar in a local context. A major problem when looking for local dissimilarities of features is that cluttered backgrounds will yield higher saliencies as such backgrounds possess high local contrasts. The proposed method differs from that in [29] in the following way. We evaluate the ‘isolation’ of nodes in a global sense and identify nodes corresponding to ‘compact’ regions in a local sense. A robust feature set based on color and orientation is used to create a fully connected graph and a k -regular sparse graph to model the global and local characteristics, respectively, of the random walks. The behavior of random walks on the two separate graphs are used to identify some salient and background nodes in the image. With the knowledge of these labeled nodes we use learning techniques with specific constraints to fit a classifying function that identifies all the salient parts in the image. Lastly, it is to be noted that the main objective in [29] is to predict human fixations on natural images as opposed to identifying salient regions that correspond to objects, as illustrated in this chapter. Thus the main contributions of the chapter are

- We develop measures based on random walks on a graphical representation of an image to model the behavior of salient regions.
- The seed nodes belonging to salient and background regions are identified in an automatic way by modeling the global and local properties of salient regions.
- A ‘pop-out’ graph model of the image is developed for salient object detection based on the knowledge of the identified seed nodes.

- We apply a constrained learning mechanism to learn the salient objects in the image, based on the seed nodes and the ‘pop-out’ graph model.

The remainder of the chapter is organized as follows. Section 4.2 reviews some fundamental properties of ergodic Markov chains. Section 4.3 details the feature set used to create the graphs and presents the analysis of random walks on global and k -regular graphs. In Section 4.4, we describe the method of finding some salient nodes and background nodes based on random walk hitting times. Section 4.5 describes the evolution of a more accurate graph model to represent saliency between nodes based on the already identified salient and background nodes. Section 4.6 describes a constrained learning method to predict the labels of the unlabeled nodes of the final graph model. Experimental results are presented in Section 4.7 and conclusions are given in Section 4.8.

4.2 Ergodic Markov chain fundamentals

In this section, we review some of the fundamental results from the theory of Markov chains [3], [27], [59]. *A Markov chain is a sequence of random variables $X_1, X_2, \dots, X_n$ with the Markov property, namely that, given the present state, the future and past states are independent.* Formally,

$$Pr(X_{n+1} = x | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_1 = x_1) = Pr(X_{n+1} = x | X_n = x_n) \quad (4.1)$$

where, $x_1, x_2, \dots, x_n$ are known values of the past states and $Pr(X|Y)$ is the conditional probability. A Markov chain having N states is completely specified by the $N \times N$ transition matrix P , where p_{ij} is the probability of moving from state i to state j and an initial probability distribution on the states. An ergodic Markov chain is one in which it is possible to go from every state to every state, not necessarily in a single step. A Markov chain is called a regular chain if some power of the transition matrix

has only positive elements. Hence every regular chain is ergodic but the converse is not true. In this chapter, the Markov chains are modeled as ergodic but not necessarily regular.

A random walk on a graph is a process that begins at some vertex, and at each time step moves to another vertex. When the graph is unweighted, the vertex the walk moves to is chosen uniformly at random among the neighbors of the present vertex. When the graph is weighted, it moves to a neighbor with probability proportional to the weight of the corresponding edge. A random walk starting at any given state of an ergodic Markov chain reaches equilibrium after a certain time; this means that the probability of reaching a particular state from any state is the same. This equilibrium condition is characterized by the equilibrium probability distribution of the states, π , which satisfies the relation

$$\pi \cdot P = \pi \quad (4.2)$$

where π is the $1 \times N$ row vector of the equilibrium probability distribution of N states in the Markov chain. In fact, π is the left eigen vector of the stochastic matrix P corresponding to the eigenvalue one and hence is easy to compute.

The matrix W is defined as the $N \times N$ matrix obtained by stacking the row vector π , N times. For regular Markov chains, W is the limiting case of the matrix P^n as n tends to infinity. The fundamental matrix Z of an ergodic Markov chain is defined as

$$Z = (I - P + W)^{-1} \quad (4.3)$$

where I is the $N \times N$ identity matrix. The fundamental matrix Z can be used to derive a number of interesting quantities involving ergodic chains including the hitting times. We define $E_i(T_i)$ as the expected number of steps taken to return to state i if the Markov chain is started in state i at time $t = 0$. Similarly, $E_i(T_j)$ is the expected number of steps taken to reach state j if the Markov chain is started in state i at

time $t = 0$. $E_i(T_j)$ is known as the *hitting time* to state j from state i . $E_\pi(T_i)$ is the expected number of steps taken to reach state i if the Markov chain is started in the equilibrium distribution π at time $t = 0$, i.e., the hitting time to state i from the equilibrium condition is $E_\pi(T_i)$.

The three quantities $E_i(T_i)$, $E_i(T_j)$ and $E_\pi(T_i)$ can be derived from the equilibrium distribution π and the fundamental matrix Z as [3]

$$\begin{aligned} E_i(T_i) &= \frac{1}{\pi_i} \\ E_i(T_j) &= E_j(T_j) \cdot (Z_{jj} - Z_{ij}) \\ E_\pi(T_i) &= E_i(T_i) \cdot Z_{ii} \end{aligned} \tag{4.4}$$

where π_i is the i^{th} element of the row vector π and Z_{ii} , Z_{jj} , Z_{ij} are the respective elements of the fundamental matrix Z . Equation 4.4 shows that the first return time of a node is inversely related to its equilibrium probability. This is also true intuitively since the node with a high equilibrium probability will be visited more often and hence will have less return time. For detailed proofs of the results in eq. (4.4), please refer to [3].

4.3 Graph representation

We represent the image as a graph $G(V, E)$, where V is the set of vertices or nodes and E is the set of edges. The vertices (nodes), $v \in V$, are patches of size 8×8 on the image while the edges, $e \in E$, represent the connection between the nodes. An edge between node i and node j is represented as e_{ij} , while $w(e_{ij})$ or simply w_{ij} represents the weight assigned to the edge e_{ij} based on the similarity between the feature set defined on node i and node j .

The important roles played by color and orientation in deciding the salient regions in an image has been well documented [37]. Hence, we consider these two features to

be defined on each node. The image is represented in YCbCr domain and the Cb and Cr values on each node are taken as the color feature. The motivation for choosing YCbCr is that it is perceptually uniform and is a better approximation of the color processing in the human visual system. As for orientation, we consider the complexity of orientations in a patch rather than the orientations itself. A salient region can be highly textured or smooth and the saliency is indicated by how its complexity is different with respect to the rest of the image. Hence, it is more useful to consider the contrast in complexity of orientations and to this end, we use the feature derived from the orientation histogram entropy at different scales similar to the one described in Chapter 3. Recall that in [41], only the local entropy is computed while Chapter 3 calculates the orientation histogram of a local patch and its complexity is calculated as the entropy of histogram. The orientation entropy E_P of the patch (or node in the graph) P having the orientation histogram H_P is calculated as

$$E_P = - \sum_i H_P(\theta_i) \log H_P(\theta_i) \quad (4.5)$$

where $H_P(\theta_i)$ is the histogram value of the i^{th} orientation bin corresponding to the orientation θ_i . We calculated the orientation entropy at five different scales $\rho_i \in \{0.5, 1, 1.5, 2, 2.5\}$ to capture the multi-scale structures in the image. The scale space is generated by convolving the image with Gaussian masks derived from the five scales. When we consider the orientation complexity at different scales, the dependency of the feature set on the selected patch size is reduced. Hence, the seven dimensional vector $x = [Cb, Cr, E_{\rho_1, \dots, \rho_5}]$ represents the feature vector associated with a node on the graph. Here Cb and Cr are calculated as the average values over the 8×8 image patch and E_{ρ_i} represents the orientation entropy calculated at scale ρ_i .

The weight w_{ij} of the edge connecting node i and node j is

$$w_{ij} = e^{\frac{-||x_i - x_j||^2}{\sigma^2}}. \quad (4.6)$$

where x_i and x_j are the feature vectors attributed to node i and node j respectively.

The value of σ_1 is fixed to unity in our experiments.

In the proposed framework, the detection of salient regions is initiated by the identification of the most ‘salient’ node in the graph. In doing so, we wish to incorporate both global as well as local information into the salient node identification mechanism. Clearly, such a technique is better compared to considering either global or local information only. The image is represented as a fully connected (complete) graph and a k -regular sparse graph to capture the global and local characteristics, respectively, of the random walk. In the complete graph, every node is connected to every other node so that there is no restriction on the movement of the random walker as long as the strength of the edge allows it. Hence, it is possible for the random walker to move from one corner of the image to the other corner in one step depending on the strength of the edge between the nodes. Spatial neighborhood of the nodes is given no preference and this manifests the global aspect of features in the image. The $N \times N$ affinity matrix, A^g , that captures the global aspects of the image features is defined as

$$A_{ij}^g = \begin{cases} w_{ij}, & i \neq j \\ 0, & i = j. \end{cases} \quad (4.7)$$

The degree d_i^g of a node i , is defined as the sum of all weights connected to node i and the degree matrix D^g is defined as the diagonal matrix with the degrees of the node in its main diagonal, i.e.,

$$d_i^g = \sum_j w_{ij},$$

$$D^g = \text{diag} \{d_1^g, d_2^g, \dots, d_N^g\}. \quad (4.8)$$

The row weights of the affinity matrix A^g needs to be divided by the degree of the corresponding node to obtain the probability transition matrix. Hence the transition matrix for the random walk on the fully connected graph, P^g , is obtained as

$$P^g = D^{g^{-1}} A^g. \quad (4.9)$$

The equilibrium distribution π^g and the fundamental matrix Z^g can be obtained from the transition matrix P^g according to equations (4.2) and (4.3). This leads to the computation of the hitting times of the complete graph, viz., $E_i^g(T_i)$, $E_i^g(T_j)$ and $E_\pi^g(T_i)$, according to equation (4.4).

The characteristics of the features in a local area in the image are encoded into the sparse k -regular graph where every node is connected to a limited k other neighboring nodes. We choose a particular patch and the 8 patches in its spatial neighborhood to study the local properties of the random walk so that $k = 8$. Thus, the random walker is restricted to a local region in the image while its path is determined by the features in that region. In this configuration, therefore, a random walker at one corner of the image cannot jump to the opposite corner, but has to traverse through the image according to the strengths of the edges. Such random walks capture the properties of the local features of the image. The $N \times N$ affinity matrix, A^l , that captures the local aspects of the image features is defined as

$$A_{ij}^l = \begin{cases} w_{ij}, & j \in N(i) \\ 0, & otherwise \end{cases} \quad (4.10)$$

where $N(i)$ denotes the nodes in the spatial neighborhood of node i . The degree matrix D^l and the transition matrix P^l of the k -regular graph can be obtained from the affinity matrix A^l similar to equations (5.3) and (4.9). Further we calculate the equilibrium distribution π^l , fundamental matrix Z^l and the hitting times $E_i^l(T_i)$, $E_i^l(T_j)$, $E_\pi^l(T_i)$ for the k -regular graph in a similar manner as that for the complete graph.

4.4 Node selection

Having represented the image as a graph, the next task is to identify the node that corresponds to a patch that most likely belongs to the salient region of the image - this node is called the most salient node. The selection of the most salient node is based on the hitting times of the random walker computed on both the complete graph as well as the k -regular graph. The most salient node is then used to identify some other salient nodes and some background nodes. The salient nodes and the background nodes together enable a seeded extraction of the salient region.

4.4.1 Most salient node

The most salient node in the image should globally pop-out in the image when compared to other competing nodes. At the same time it should fall on a compact object in the image in some local sense. The global pop-out and compactness properties are reflected in the random walks performed on the complete graph and the k -regular graph, respectively. When a node is globally a pop-out node, what it essentially means is that it is isolated from the other nodes so that a random walker takes more time to reach such a node. On the other hand, if a node is to lie on a compact object, a random walker should take less time to reach it on a k -regular graph. We now elaborate on these concepts.

We calculate the global isolation of a node by the time taken to access the node when the Markov chain is in equilibrium. In [29], Harel et al. identify the most frequently visited node as the most salient node. This is directly measured from π_i or from the first return time $E_i(T_i)$ and it characterizes the dissimilarity of the node in a local sense, i.e., their activation map encodes how different a particular location in the image is, compared to its neighborhood. As mentioned earlier, we believe that the isolation of a node should be computed in the *global* sense and hence, the measure in

[29] will not characterize the global isolation of node i . A better characterization will be by the sum of hitting times from all other nodes to node i on a complete graph, i.e.,

$$H_i = \sum_j E_j^g(T_i). \quad (4.11)$$

Since the edge strengths represent the similarity between nodes, a higher value of H_i indicates higher global isolation of the node. However, in the computation of H_i , the hitting times from all the other nodes to node i are given equal preference. The measure can be further improved if the hitting times from the most frequently visited nodes are given priority over hitting times from less frequently visited nodes. This property is inherent in $E_\pi^g(T_i)$, which is the time taken to access the node i when the Markov chain starts from equilibrium. The equilibrium distribution directly depends on the frequency of access to different nodes, and hence $E_\pi^g(T_i)$ gives priority to hitting times from most frequently visited nodes over hitting time from less frequently visited nodes. Hence, the global isolation of a node is measured by $E_\pi^g(T_i)$.

Next, we ensure that the most isolated node falls on a compact object by looking at the equilibrium access time of nodes in the k -regular graph. The local random walk under Markov equilibrium will reach the nodes corresponding to compact structures faster as it is guided by strong edge strengths from the neighborhoods. Hence a low value of local random walk equilibrium access time, $E_\pi^l(T_i)$ ensures that the respective node falls on a compact surface. Considering both the global and local aspects, saliency of node i , $NSal_i$, is defined as

$$NSal_i = \frac{E_\pi^g(T_i)}{E_\pi^l(T_i)}. \quad (4.12)$$

The most salient node can be identified as the node that maximizes $NSal_i$, i.e.,

$$N_s = \arg \max_i NSal_i. \quad (4.13)$$

Figure 4.1 compares the detection of the most salient node between [29] and the proposed method. While the method in [29] generates fixations produced on the

image by the algorithm, for comparison with our method, we use their definition of the least visited node indicated by the first return time of the random walk as the most salient node. However, it must be noted that in [29], the edge strengths represent the dissimilarity between nodes as opposed to similarity in our case. Hence, it turns out that the random walker’s most frequently visited node in [29] is actually the least frequently visited node using our definition of edge strength. Figures 4.2(a) and 4.2(c) show some more example images from the database and figures 4.2(b) and 4.2(d) show the most salient node marked with a red region in the respective images.

Figure 4.1: Comparison of [29] and proposed method for detection of most salient node. (a) Original images, (b) most salient node detected by proposed method and (c) most salient node detected by [29].

4.4.2 Other salient nodes

The accuracy of seeded salient region extraction could be further improved if more labeled salient nodes are available. Intuitively, the saliency measure calculated in eq. (4.12) could be used for this purpose, e.g., by selecting n nodes with highest $NSal_i$. However, it is possible that a salient object is non-homogeneous in appearance in which case, selection of nodes with high $NSal_i$ could result in sampling the nodes from a single homogeneous part of the salient object. We need to sample the salient object as uniformly as possible for a more accurate extraction of the salient region.

Figure 4.2: Detection of the most salient node. (a, c) Original images and (b, d) Most salient node spotted with the red region.

Hence, it is desirable to select nodes with high saliency but those that are not quite similar to the already identified salient node(s). This condition is formulated as

$$N_{s_1} = \arg \max_j (E_{N_s}^g(T_j) \cdot NSal_j) \quad (4.14)$$

where N_{s_1} is the first salient node detected in addition to the maximum salient node. Similarly, the n^{th} salient node is sampled as

$$N_{s_n} = \arg \max_j ((E_{N_s}^g(T_j) \cdot E_{N_{s_1}}^g(T_j) \dots \\ \dots E_{N_{s_{n-1}}}^g(T_j)) \cdot (NSal_j)^n), \quad (4.15)$$

where $s_1, s_2, \dots, s_{n-1}$ are the salient nodes identified prior to the n^{th} salient node. Note that the node distance measures are multiplied by the saliency measures to enforce the constraint that the newly identified node is on the salient object.

4.4.3 Background nodes

After identifying the salient nodes in the image we move on to identify certain nodes in the background. This is to facilitate the extraction of the salient regions using the

background nodes as negative labels along with the positive labels (salient nodes). The most important feature of a background node is obviously the less saliency of the node as calculated from eq. (4.12). Moreover, the background nodes have the property that it is at a large distance from the most salient node, N_s . Hence the first background node, N_{b_1} , is calculated as

$$N_{b_1} = \arg \max_j \frac{E_{N_s}^g(T_j)}{NSal_j} \quad (4.16)$$

The background in an image is, more often than not, inhomogeneous, e.g. due to clutter or due to regions having different feature values. In the cat image in figure 4.2(c), for example, the background consists of regions with different colors. Similarly, the tiger image in figure 4.2(c) has multiple backgrounds with different colors as well as texture. Our goal is to capture as much of these variations as possible by locating at least one background node in each of such regions. Hence, while maximizing the distance of a node to the maximum salient node, we impose an additional condition of maximizing the distance to all background nodes identified so far. This will ensure that the newly found background node falls on a new region. Even if there are no multiple backgrounds, i.e. the background is relatively homogeneous, the algorithm will only place the new node in the same background region. This does not affect the performance of the algorithm. Thus, the n^{th} background node, N_{b_n} , is identified as

$$N_{b_n} = \arg \max_j \frac{E_{N_s}^g(T_j) \cdot E_{N_{b_1}}^g(T_j) \dots E_{N_{b_{n-1}}}^g(T_j)}{(NSal_j)^n} \quad (4.17)$$

The above equation can be viewed as a product of n terms, where each term is of the form $\frac{E_{N_s}^g(T_j)}{NSal_j}$.

Figure 4.3 shows examples of the salient and background nodes extracted using the proposed method as red and green spots respectively. As expected, the salient nodes sample the salient object in a uniform way to represent the inhomogeneity in the salient regions. For example, in the girl image in fourth row of Figure 4.3(d), the

Figure 4.3: Detection of the Salient and Background nodes. (a, c) Original images and (b, d) Salient nodes spotted with red regions and background nodes spotted with the green regions.

salient nodes are placed on the face, hand, shirt and trousers of the girl in the original image shown in fourth row of Figure 4.3(c). The background nodes are pushed away from the salient nodes and more or less represent the inhomogeneity in the background. For example, in the car image in second row of Figure 4.3(d), the background nodes are placed in the sky, wall and the street regions of the original image shown in the second row of Figure 4.3(c). The number of salient nodes (including the most salient node) and background nodes is fixed to 8 in the experiments.

4.5 ‘Pop-Out’ Graph Model

The identification of salient nodes and background nodes can be used as a partial labeling on the graph with $l \in \{1, -1\}$ and $l \ll u$, where l is the set of labeled nodes and u is the set of unlabeled nodes. With the knowledge of partial labeling on the graph, learning techniques can be employed to predict the labels of the unlabeled nodes in the graph. The success of such learning techniques depends on how

accurately the relationships between various nodes of the graph represent the salient object and background. For example in such an ideal ‘pop-out’ graph, the nodes belonging to salient object of the image should show high affinity between them. The nodes belonging to background should also show high affinity among them. However, there should not be large affinity between nodes falling on salient regions and nodes in the background. This behavior is not guaranteed in the affinity matrices in Section 4.3. In images with multiple backgrounds, the affinity matrices A^g and A^l , which are based on feature affinities could show low affinity between two background nodes. Also for salient objects that are non-homogeneous, the affinity matrices can show low affinity among different parts of the salient object due to feature dissimilarity. In this section we introduce a new graph representation of the image, the ‘pop-out’ graph model, which will take care of these limitations.

The ‘pop-out’ graph model is created under the following assumptions.

1) If a node falls on a salient object, it should maintain high hitting times to all of the labeled background nodes. No assumption is made on the hitting time of the node to the salient nodes.

2) If a node falls in the background, it should maintain a low hitting time to at least one of the labeled background nodes, provided we have captured the various backgrounds faithfully.

Hence, the minimum hitting time of node N_1 to any of the labeled background nodes is calculated as

$$T_{N_1} = \min(E_{N_1}^g(T_{N_{b1}}), E_{N_1}^g(T_{N_{b2}}), \dots, E_{N_1}^g(T_{N_{bn}})). \quad (4.18)$$

The affinity between two nodes N_1 and N_2 in the ‘pop-out’ graph model is calculated with the minimum hitting times T_{N_1} and T_{N_2} as

$$w_{N_1 N_2}^p = e^{\frac{-|T_{N_1} - T_{N_2}|^2}{\sigma^2}}. \quad (4.19)$$

The value of σ^2 is empirically decided in our experiments. If both N_1 and N_2 belong

to the background, T_{N_1} and T_{N_2} will have comparatively closer low values, which will result in a higher $w_{N_1 N_2}^p$ and hence high affinity between the nodes. Similarly, if both N_1 and N_2 belong to the salient region, T_{N_1} and T_{N_2} will have comparatively closer large values, which will also result in high affinity between the nodes. If N_1 belongs to salient object and N_2 to background or vice versa, the difference between T_{N_1} and T_{N_2} will be large resulting in a lower value of $w_{N_1 N_2}^p$ and consequently low affinity between the nodes.¹ The $N \times N$ ‘pop-out’ affinity matrix, A^p , that captures the pop-out aspects of the image is constructed similar to eq. (4.7).

4.5.1 Distance factor

The affinity matrix of the ‘pop-out’ graph model indicates the affinities between nodes based on their pop-out characteristics and is independent of the spatial location of the nodes. However it can so happen that some unlabeled nodes falling in the background may also exhibit large hitting times to the labeled background nodes. Such nodes can be considered as ‘locally salient’ in the background (for example, the hole on the wall in the car image in the second row of figure 4.3(c)), but is not of interest as they do not form a part of the salient object. The ‘pop-out’ affinity matrix will indicate high affinity between these ‘locally salient’ background nodes and the nodes on the salient object by virtue of their high hitting times to the labeled background nodes. As the locally salient nodes do not fall on the salient object we can reduce these affinities if we scale it by a ‘distance factor’ as well while creating the affinity matrix.

The affinity matrix, A^p , has to be redesigned so that the affinities between the

¹The affinity between two salient nodes will not be as high as affinity between two background nodes owing to the larger values of T_{N_1} and T_{N_2} in the former. Yet, these affinities will be higher than the affinity between a salient node and a background node, and hence they will be able to represent the ‘pop-out’ characteristics of the image. See the Experiments section for a statistical validation of this property using the ground truth labeling.

nodes decreases as the ‘distance’ between them increases. Euclidean distance between nodes is not a good measure due to the following reason. Recall that we introduce the distance factor to reduce the affinity between the ‘locally salient’ background nodes and the nodes on the salient object. Consider the case of three nodes N_x , N_y and N_z in which nodes N_x and N_y fall on the salient object while N_z is a ‘locally salient’ node in the background. It cannot be guaranteed that the distance between N_z and N_x is always larger than distance between N_x and N_y in a Euclidian sense. Hence we consider a random walk based distance measure to re-evaluate the node affinities in matrix A^p .

A sparse k-regular affinity matrix A^{lp} is created from A^p similar to eq. (4.10). The local random walks on A^{lp} will traverse through the local neighborhoods guided by the edge strengths. Note that, as the edge strengths in A^{lp} represents the ‘pop-out’ similarity between the neighboring nodes, the ‘locally salient’ nodes in background, like N_z , will take more time to reach the nodes on the salient object. The ‘distance’ between N_x and N_y will be relatively less under the new random walk on A^{lp} .

The distance between node N_1 and node N_2 is given by $D_{N_1N_2} = Dist_{N_1N_2} + Dist_{N_2N_1}$, where $Dist_{N_1N_2} = E_{N_1}^{lp}(T_{N_2})$ is the hitting time from N_1 to N_2 and $Dist_{N_2N_1} = E_{N_2}^{lp}(T_{N_1})$ is the hitting time from N_2 to N_1 .

Before redesigning the weights in A^p based on the distance measure, $D_{N_1N_2}$, we also have to closely examine the behavior of background nodes. The distance measure is introduced to increase the affinity between nodes that are spatially close in a local random walk sense. This assumption will hold good for the nodes falling on the salient object as the salient object forms a compact region in the spatial domain. However the distance factor will affect the affinities between nodes falling in the background region as there are no constraints on the spatial location of background nodes. Hence the affinity matrix, A^p , has to be refined to consider the following conditions.

- 1) If both nodes N_1 and N_2 fall on a salient object, they will possess large T_{N_1}

and T_{N_2} . The distance factor has to be introduced in this case.

2) If nodes N_1 and N_2 show low T_{N_1} and T_{N_2} , it indicates both the nodes fall in background and hence distance factor should be avoided.

3) If nodes N_1 and N_2 show large difference in T_{N_1} and T_{N_2} , distance factor is avoided as there has to be low affinity between the nodes irrespective of the spatial location.

To this end, we introduce a measure called the ‘Salient Object Factor’, $SF_{N_1N_2}$, between nodes N_1 and N_2 calculated as

$$SF_{N_1N_2} = \min(T_{N_1}, T_{N_2}). \quad (4.20)$$

The above equation indicates that $SF_{N_1N_2}$ will be large only when both nodes fall on salient regions. The ‘Salient Object Factor’ between all pairs of nodes are normalized between 0 and 1 to facilitate further processing. The new weight between nodes N_1 and N_2 scaled by the distance factor, $(D_{N_1N_2})^{SF_{N_1N_2}}$, is calculated as

$$w_{N_1N_2}^{pd} = \frac{w_{N_1N_2}^p}{(D_{N_1N_2})^{SF_{N_1N_2}}}. \quad (4.21)$$

A value of zero for $SF_{N_1N_2}$ implies no scaling of original weight while $SF_{N_1N_2} = 1$ implies scaling of original weight by a factor of $D_{N_1N_2}$. Thus a ‘pop-out’ graph model, A^{pd} , is created in similar terms to eq. (4.7) using the weights in eq. (4.21). In this model, the affinity between two nodes indicates the chances of them falling on a spatially compact object which shows pop-out properties from the background(s).

4.6 Labeling of unlabeled nodes using constrained learning on graphs

The label values on the unlabeled nodes are predicted by fitting a classifying function $\mathbf{f} : V \rightarrow \mathbb{R}$. In [97], the function $\mathbf{f}$ that maps the vertex set V of the graph to real

numbers is required to minimize the quadratic energy function $E(\mathbf{f})$ given by

$$E(\mathbf{f}) = \frac{1}{2} \sum_{ij} w_{ij}^{pd} (f(i) - f(j))^2, \quad (4.22)$$

subject to the labeling constraints on the labeled nodes. It can be inferred from eq. (4.22) that the $\mathbf{f}$ that minimizes $E(\mathbf{f})$ will not vary much over the nodes having high affinities. Along with this smoothness assumption, the function $\mathbf{f}$ is constrained to assume constant values over the labeled nodes. Hence, the label optimization problem in [97] can also be expressed as

$$\begin{aligned} \text{minimize } E(\mathbf{f}) = & \begin{bmatrix} \mathbf{f}_l^T & \mathbf{f}_u^T \end{bmatrix} L^{pd} \begin{bmatrix} \mathbf{f}_l \\ \mathbf{f}_u \end{bmatrix} \\ \text{subject to } & \mathbf{f}_l = \mathbf{y} \end{aligned} \quad (4.23)$$

where $\mathbf{f}_l$ and $\mathbf{f}_u$ denote the function values over labeled and unlabeled nodes, respectively, and L^{pd} is the Laplacian matrix of the ‘pop-out’ affinity graph A^{pd} re-arranged with respect to labeled and unlabeled nodes as

$$A^{pd} = \begin{bmatrix} A_{ll} & A_{lu} \\ A_{ul} & A_{uu} \end{bmatrix}. \quad (4.24)$$

A_{ll} , A_{lu} , A_{ul} and A_{uu} are the respective block matrices over labeled and unlabeled nodes. $\mathbf{y}$ is the label value which is fixed as 1 for salient nodes and -1 for background nodes. The optimization problem in eq. (4.23) essentially minimizes a quadratic function of $\mathbf{f}$ subject to some linear constraints.

The optimization in eq. (4.23) can be considered as ‘hard labeling’ since it enforces the already identified labels. It is clear that the performance of salient region extraction depends on the accuracy of the extracted salient and background nodes. These nodes are found in an automatic manner in contrast to manual labeling used in a typical learning scenario with constraints. Hence, in order to address potential errors in the labeled nodes, we adopt a ‘soft labeling’ approach to label the unlabeled

nodes which allows us to assign different weights or confidence measures to the already identified labels while optimizing the energy function $E(\mathbf{f})$. This 'soft-labeling' method is motivated by the semi-automatic image segmentation methods discussed in [18] and [16]. The energy function $E(\mathbf{f})$ is now re-written as

$$E(\mathbf{f}) = [\mathbf{f}_l^T \mathbf{f}_u^T] L^{pd} \begin{bmatrix} \mathbf{f}_l \\ \mathbf{f}_u \end{bmatrix} + \lambda \sum_{i=1}^L k^i (f^i - y^i)^2, \quad (4.25)$$

where $y^i \in \{1, -1\}$ is the pre-assigned label to the i^{th} node and L is the total number of labeled nodes. f^i is the function value over i^{th} node, $\mathbf{f}_l = [f^1, f^2, \dots, f^L]^T$, and k^i is the confidence of the label for i^{th} node. Minimizing the energy function in eq. (4.25) implies finding a smooth labeling on the graph A^{pd} using the Laplacian matrix L^{pd} and at the same time respecting the already assigned labels on the nodes according to predetermined confidence weights.

As evident from equations (4.15) and (4.17), the salient and background labels are identified based on their distances to the most salient node along with their distances to the already identified nodes of the same label. The background labels arise from their large distances from the most salient node and their low saliency values. However, for salient nodes, the criterion of a large distance from the most salient node is enforced along with the condition of high saliency values. As explained previously, such a criterion is enforced to sample the salient nodes as uniformly as possible over the salient region. Clearly, the condition for salient nodes is a weaker one compared to background nodes leading to more accurate labeling of the background nodes than the salient nodes. Hence, in the design of the proposed 'soft label' energy function, we assign equal priority to all background labels while the confidence weights for the salient labels are varied by a measure called the 'Background Affinity'.

The 'Background Affinity' of a labeled salient node s_i is defined as the inverse of $T_{N_{s_i}}$, defined in eq. (4.18). Recall that $T_{N_{s_i}}$ is the minimum of the hitting times of s_i to all the labeled background nodes. A large $T_{N_{s_i}}$ implies less background affinity and

vice versa. The confidence on a labeled salient node decreases with the increase in its background affinity. The values of $T_{N_{s_i}}$ for all labeled salient nodes is normalized between 1 and 2. Each background node is assigned a confidence weight which is the average weight of the salient nodes. Thus eq. (4.25) can be re-written using the diagonal matrix K with the confidence weights $k^1, \dots, k^L$ as

$$E(\mathbf{f}) = [\mathbf{f}_l^T \mathbf{f}_u^T] L^{pd} \begin{bmatrix} \mathbf{f}_l \\ \mathbf{f}_u \end{bmatrix} + \lambda [\mathbf{f}_l - \mathbf{y}]^T K [\mathbf{f}_l - \mathbf{y}]. \quad (4.26)$$

Since the energy function $E(\mathbf{f})$ in eq. (4.26) is quadratic in $\mathbf{f}_l$ and $\mathbf{f}_u$, its minimum can be found by equating the partial derivatives of $E(\mathbf{f})$ with respect to $\mathbf{f}_l$ and $\mathbf{f}_u$ to zero and solving for the variables to finally obtain

$$\mathbf{f}_u = \lambda Q^{-1} L_{uu}^{-1} A_{ul} P K \mathbf{y}, \quad (4.27)$$

where $P = (L_{ll} + \lambda K)^{-1}$ and $Q = I - L_{uu}^{-1} A_{ul} P A_{lu}$.

L_{uu} and L_{ll} are unnormalized block Laplacian matrices calculated similar to [97] as

$$L_{uu} = D_{uu} - A_{uu} \quad (4.28)$$

$$L_{ll} = D_{ll} - A_{ll}$$

where D_{uu} and D_{ll} are respectively the degree matrices of block affinity matrices A_{uu} and A_{ll} . λ is empirically set to 1 in the experiments. The final saliency map is obtained by pulling those values of $\mathbf{f}_u$ greater than zero to one and thresholding those values of $\mathbf{f}_u$ less than zero to zero.

4.7 Experimental results

The experiments are conducted on a database of about 5,000 images available from [48]. The database used in this work contains the ground truth of the salient region marked as bounding boxes by nine different users. The median of nine boxes

is selected as the final salient region. It has to be noted that there is no large difference of opinions among the annotating users regarding the salient object for all the 5000 images in the database. In other words, the database contains single salient object/region in all images. We have shown some results of identifying the salient nodes and background nodes in Figure 4.3. In order to evaluate the robustness of the detection of the salient and background nodes, we calculate the Node Accuracy Percentage, which is the percentage of nodes consistent with the ground truth in the entire dataset. The salient nodes falling inside the user bounding box and the background nodes falling outside the user bounding box are considered as accurate. Figure 4.4 shows the accuracy of salient and background nodes for different node numbers. The robustness of the node selection method is evident from the high percentage accuracy of salient and background nodes as can be seen from Figure 4.4. However, it can be observed that as the node numbers are increased the accuracy of both salient and background nodes falls off slightly.

Figure 4.4: Percentage accuracy of salient and background nodes vs. number of nodes detected for an image.

The strategy for 'pop-out' graph modeling and affinity calculation between any two nodes N_1 and N_2 using the minimum hitting times T_{N_1} and T_{N_2} as described in Equation (4.19) is statistically validated with the knowledge of ground truth user

Table 4.1: Average absolute difference of minimum hitting times in three different node pair categories.

Node pair category	Average absolute difference of T_N
Background - Background	52.1
Salient - Salient	68.4
Salient - Background	78

bounding box on the image. After evaluating the minimum hitting time T_N for all the nodes in the image, we evaluated the absolute differences of T_N between all pairs of nodes. Further, the pairs of nodes are classified into three different categories - Background-Background, Salient-Salient and Salient-Background. This is done with the knowledge of the user bounding box for the image. This exercise is carried out for all the images in the dataset. The average difference between minimum hitting times for the three different categories is listed in Table 4.1. Note that larger the absolute difference in minimum hitting times, lesser the affinity between the nodes according to Equation (4.19). Thus the method of 'pop-out' affinity design is validated with the salient-background node pair category showing the least affinity among all three categories.

The objective evaluation of the algorithm is carried out based on Precision, Recall and F-Measure. *Precision* is calculated as ratio of the sum of intensities in the saliency map captured inside the user annotated rectangle to the total saliency computed for the image. *Recall* is calculated as the ratio of the total saliency captured inside the user annotated window to the area of the user annotated window. *F-Measure* is the overall performance measurement and is computed as the weighted harmonic mean between the precision and recall values. It is defined as

$$f\text{-Measure}_\alpha = \frac{(1 + \alpha) * Precision * Recall}{(\alpha * Precision + Recall)}, \quad (4.29)$$

where α is real and positive and decides the importance of precision over recall. The

recall values are in general low compared to precision values since the area of the user-annotated window is in the denominator of recall computation. In our experiments we fix the value of α as 0.3.

Table 4.2 shows the comparison of the F-Measure values at different stages of the evolution of the ‘pop-out’ graph model. The F-Measure values are the results of the experiments carried over the entire dataset of 5000 images. A connected component labeling that includes the salient nodes is carried out to extract the salient object. The proposed method performs well after the identification of the seed nodes due to the following reasons.

- The distance weighted ‘pop-out’ affinity graph improves the accuracy of the graph model with respect to salient object detection.
- Function fitting based on ‘soft labeling’ prioritizes correct labeled nodes over possibly incorrect labeled nodes to improve the final performance.

Table 4.2 shows ‘soft labeling’ only marginally increases the F-Measure compared to hard labeling since the partial labeling of the salient and background nodes were quite accurate already. Introduction of multiple salient nodes and increase in the number of labeled nodes also improved the F-Measure. However, the optimal number of labeled nodes varies from image to image. Increase in number of labeled nodes can increase the accuracy in some images while it can introduce wrong labels in certain other images. Figure 4.5 shows the F-Measure surface generated for all combinations of the number of salient and background nodes ranging from 1 to 10. The best F-Measure for the MSRA dataset is obtained as 0.694 with 5 salient nodes and 6 background nodes with the parameter σ_2 set to the value 1500. The F-Measure variation of proposed method with respect to the parameter σ_2 is shown in figure 4.6

Figure 4.7 shows examples of the saliency map extracted using the proposed algorithm. Figures 4.7(a), 4.7(c), 4.7(e) and 4.7(g) shows the original images and Figures 4.7(b), 4.7(d), 4.7(f), 4.7(h) shows the extracted binary salient region. The size of

Figure 4.5: F-Measure surface for all combinations of salient and background node numbers ranging from 1 to 10.

Table 4.2: F-measure comparison of different stages of the proposed method.

Method	F-Measure
Proposed method with distance weighted ‘pop-out’ affinity graph A^{pd} and ‘hard labeling’ (Salient Nodes =1, Background Nodes = 4)	0.66
Proposed method with multiple salient nodes, distance weighted ‘pop-out’ affinity graph A^{pd} and ‘hard labeling’ (Salient Nodes = 8, Background Nodes = 8)	0.685
Proposed method with multiple salient nodes, distance weighted ‘pop-out’ affinity graph A^{pd} and ‘soft labeling’ (Salient Nodes = 8, Background Nodes = 8)	0.69

Figure 4.6: F-Measure variation for different values of σ

the patches decides the final smoothness of edges in the saliency map. Bigger patches result in lesser computational complexity of the proposed method but at the same time cause the edges of the final saliency map to be jagged. It can be verified from the results in Figure 4.7 that the proposed salient region extraction method works satisfactorily for images with multiple backgrounds and non-homogenous salient objects. Many images shown in Figure 4.7 have multiple backgrounds as can be verified with the couple image in the first row of Figure 4.7(g) or the dog image in the fifth row of Figure 4.7(a). Similarly we have obtained almost accurate saliency maps for non-homogenous salient objects as can be verified from the horse cart image in the sixth row of Figure 4.7(g) and the lady with the poster image in the last row of Figure 4.7(a).

In Figure 4.8, we show further examples of the proposed salient region extraction algorithm with the salient image marked with a red bounding box. The original images are in figures 4.8(a) and 4.8(c) and the corresponding salient regions are marked in figures 4.8(b) and 4.8(d). The final bounding box over the salient region can be used in applications like cropping and zooming for display on small screen

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4.7: Results of function fitting using 'soft labeling' and thresholding. (a),(c),(e),(g) Original images (b),(d),(f),(h) The final binary saliency maps.

devices.

Figure 4.8: Bounding box on salient region. (a, c) Original images and (b, d) Red bounding box over the salient region.

We also show some failure examples of the proposed method for salient region detection in Figure 4.9. The underlying reason for the failures seem to be the similarity of features on the salient object with the background which affects the random walk, e.g., the branches and the bird in the first row and the fish and the rock in the second row. However, as noted earlier, the general framework of the algorithm allows for more robust features to be utilized. In the third row, the features caused the building in the background to be detected as the salient region; however, this failure

Figure 4.9: Failure examples. (a) Original image, (b) most salient node and (c) the corresponding bounding box.

has opened up the question of what effect, if any, does depth have on saliency since it is evident that there is a large variation in the depth field of the image.

4.7.1 Comparison with other saliency detection methods

The proposed method being an unsupervised for saliency detection is compared against the state-of-the art unsupervised methods. We have compared the saliency map of the proposed random walk based method with the saliency maps generated by the saliency toolbox (STB)[37], the spectral residual method based on [88] and the phase spectrum based on [28].

As noted, the intensities of the saliency map are used in the computation of precision and recall. The final saliency maps obtained in our case and in the salient tool box are binary; however, in the spectral residue method and in the phase method, the saliency maps are not binary. Hence, we binarize the saliency maps of the spectral residue and phase spectrum method using Otsu's thresholding technique, so that the effect of intensity values is applied equally on all the saliency maps. Figure 5.3 shows the precision, recall and F-Measure marked as 1, 2 and 3 on the x-axis, respectively,

Figure 4.10: Comparison of Precision, Recall and F-Measure values of Spectral Residual Method [88], Phase spectrum method [28], Saliency Tool Box [37], and the proposed method. Horizontal axis shows 1) Precision 2) Recall 3) F-Measure.

for the proposed method as well as for the other methods. It can be seen that the proposed method outperforms all the other methods with respect to precision recall and F-Measure. The variation of F-measure values with the parameter α ranging from 0.2 to 0.8 is shown in figure 4.11. It can be seen from figure 4.11 that the proposed method outperforms other competing methods irrespective of the choice of α .

4.8 Conclusion

We have presented an automatic method to extract salient regions in images through random walks on a graph modeled by color and orientation features. It provides a generic framework that can be enriched with more robust feature sets. The proposed method captures saliency using both global and local properties of a region by carrying out random walks on a complete graph and a k-regular graph, respectively. The accuracy of the automatic labeling module and the subsequent function fitting module depends on the accuracy of the graphs with respect to the representation of the salient

Figure 4.11: Comparison of F-Measure values of Spectral Residual Method [88], Phase spectrum method [28], Saliency Tool Box [37], and the proposed method for various values of α

object. The graphs are built on orientation entropy features on different scales and color features by giving equal priority to all individual features. It could be the case that salient regions can ‘pop-out’ by virtue of only its texture or color properties, or a combination of both. Identifying this exact ‘mixture’ to represent the graph of the image is a subject of great interest. Such an ideal graph will improve the accuracy of most salient node detection and subsequently increase the correctness of later parts of the proposed algorithm. However, it is observed that the detection of the most salient node in the proposed random walk is also quite robust. This could be effectively used for zooming, where the co-ordinate of the zoom fixation is directly available from the most salient node. The robustness of the proposed framework has been objectively demonstrated with the help of a large image data base and comparisons with existing popular salient region detection methods demonstrate its effectiveness.

Chapter 5

Unsupervised Feature Selection for Salient Object Detection

5.1 Introduction

Feature selection plays a crucial role in deciding the salient regions of an image as in any other pattern recognition problem. However the problem of identifying the relevant features that play a fundamental role in identifying saliency in an image has not received much attention. In this chapter, we introduce an unsupervised feature selection method to improve the accuracy of salient object detection. Feature selection, which is essentially selecting a subset of features from the original feature set (consequently representing an image in lower dimension), is quite important in many classification problems like face detection that involve high dimensional data. There has not been much work reported on unsupervised feature selection in contrast to supervised feature selection as the former is a far more challenging problem due to the lack of any label information. Feature selection methods can be generally classified as filter, embedded and wrapper methods depending on how the feature selection strategy is related to the classification algorithm that follows. He et.al[31] and Mitra

et.al [55] discuss unsupervised 'filter' methods of feature selection in which the feature selection strategy is independent of the classification algorithm. The problem of simultaneously finding a classification and a feature subset in an unsupervised manner is discussed in the 'embedded' feature selection methods in [45] and [93]. Law et.al [45] use a clustering algorithm based on Gaussian Mixture Models along with the feature selection, while Zhao et.al [93] optimize an objective function to find a maximum margin classifier on the given data along with a set of feature weights that favor the features which most 'respect' the data manifold. In [68], a 'wrapper' strategy for feature selection is discussed in which the features are selected by optimizing the discriminative power of the classification algorithm used.

The graph based method discussed in Chapter 4 gives equal weight to all features while the saliency factor may be attributed only to a smaller subset of features. The relevancy of different features in different images is illustrated in Figure 5.1. The saliency of the image in Figure 5.1(a) is mainly derived from the color feature, in Figure 5.1(b) from texture features while the salient parts in Figure 5.1(c) is due to some combination of color and texture features. Hence it can be concluded that some of the features used in the method in Chapter 4 are irrelevant and only add noise to the graph structure of the image thereby reducing the accuracy of the salient object detection. In this context, we propose an unsupervised filter method for identifying irrelevant features to improve the performance of salient region detection algorithms. Feature combination strategies in the context of saliency detection have been previously studied in [43] and [40]. The data-driven approaches of feature combination mentioned in Kienzle et al.'s [43] and Judd et al.'s [40] work are supervised methods that learns a classifier over the labeled data to formulate an efficient way of combining different features. The labeled data used in both methods is an eye fixation dataset obtained using eye-tracker. The objective of the above mentioned works is to learn a classifier on the eye fixation dataset so as to predict human eye fixations on unseen

images. The objective of our unsupervised method of feature selection is to find the best combination of features on an image so as to boost the performance of a salient object detector running on the same image.

Figure 5.1: Example images illustrating the relevancy of different features to saliency. The relevant salient features are (a) color (b) texture (c) some combination of color and texture features.

We relate the problem of identifying irrelevant features for salient object detection to maximizing the mixing rate of a Markov process running on a weighted linear combination of ‘feature graphs’. A semi-definite programming formulation is used to maximize the second smallest eigen value of the graph Laplacian to maximize the Markov process mixing rate. The effectiveness of the proposed unsupervised filter method of feature selection (elimination) method is demonstrated by using it in conjunction with the graph based salient object detection method in Chapter 4 resulting in improved performance on a large annotated dataset.

The chapter is organized as follows. Section 5.2 reviews some of the theory related to the graph models, Markov process equilibrium and the mixing rate of Markov pro-

cesses. In Section 5.3 we discuss the relation between irrelevant feature identification and maximization of mixing rate of Markov process. The detailed optimization algorithm is also provided in Section 5.3. Section 5.4 details the experiments done and the comparative results showing performance improvement. Finally conclusions are given in Section 5.5.

5.2 Graph Laplacian and Mixing Rate of Markov Process

In this section we review some of the theory related to graph Laplacians and mixing rate of a Markov process running on the graph. Consider an undirected graph, $G(V, E)$, where V is the set of n vertices and E is the set of edges in the graph. The graph G is characterized by the affinity matrix A where

$$A_{ij} = \begin{cases} w_{ij}, & i \neq j \\ 0, & i = j. \end{cases} \quad (5.1)$$

w_{ij} is the weight between node i and node j based on the affinity of features on the respective nodes.

The Laplacian matrix L of the graph G is defined as

$$L = D - A \quad (5.2)$$

where

$$\begin{aligned} D &= \text{diag} \{d_1, d_2 \dots d_n\}, \\ d_i &= \sum_j w_{ij}, \end{aligned} \quad (5.3)$$

and d_i is the degree of node i and is the total weight connected to node i in the graph G .

Now, consider a symmetric continuous-time Markov process running on the graph $G(V, E)$. A continuous-time Markov process is a stochastic process, $X(t): t \geq 0$, that satisfies the Markov property. The Markov property states that at any time $s \geq t \geq 0$, the conditional probability distribution of the process at time s given the whole history of the process up to and including time t , depends only on the state of the process at time t . The state space of the Markov process is the vertex (node) set V , and the edge set E represent the transition strengths. The edge e_{ij} connecting nodes i and j is associated with a transition rate w_{ij} which is the respective weight entry in the affinity matrix, A , of the graph $G(V, E)$. This implies that the transition from node i to node j is faster when the affinities between the respective nodes are larger. Consequently, a Markov process entering node i at any specific time t will take more time to move into the nodes that possess lesser affinities with node i . The mixing time of a Markov process is the time taken to reach the equilibrium distribution π starting from some initial distribution $\pi(0)$.

Let $\pi(t) \in \mathbf{R}^n$ be the n dimensional vector showing probability distributions on the n states at time t for a symmetric Markov process starting from any given distribution on the graph. The evolution of $\pi(t)$ is given by [59],

$$\frac{d\pi(t)}{dt} = Q \cdot \pi(t) \quad (5.4)$$

where Q is the transition rate matrix of the Markov process defined as

$$Q_{ij} = \begin{cases} q_{ij}, & i \neq j \\ -\sum_{j, j \neq i} q_{ij}, & i = j. \end{cases} \quad (5.5)$$

q_{ij} is the *transition rate* or rate of going from state i to state j . As explained previously, we equate this rate to the affinity between nodes i and j , w_{ij} . Hence the Laplacian matrix L and the transition rate matrix Q is related as $Q = -L$. Consequently, the solution for the distribution $\pi(t)$ at time t starting from the initial

condition $\pi(0)$ is given as

$$\pi(t) = e^{-tL}\pi(0). \quad (5.6)$$

The Laplacian L is a symmetric and positive semi-definite matrix. The eigen values of L are such that $\lambda_1 = 0 \leq \lambda_2 \leq \dots \leq \lambda_N$. λ_2 is also known as the *algebraic connectivity* of the graph G . If the graph G is connected, i.e. any node of the graph is accessible from any other node through the edge set E , then $\lambda_2 > 0$ [13]. The probability distribution $\pi(t)$ will tend to the equilibrium distribution π as $t \rightarrow \infty$. The *total variation distance* (maximum difference in probabilities assigned by two different probability distributions to the same event) between distributions $\pi(t)$ and π is upper bounded by λ_2 as [76]

$$\sup \|\pi(t) - \pi\|_{tv} \leq (1/2)n^{0.5}e^{-\lambda_2 t} \quad (5.7)$$

where $\|\cdot\|_{tv}$ is the total variance distance. Equation (5.7) implies that a larger value of λ_2 will result in a tighter bound on the total variation distance between $\pi(t)$ and π and consequently result in a faster mixing of the Markov process. The second smallest eigen value λ_2 can be expressed as a function of some Laplacian matrix L as [6]

$$\lambda_2(L) = \inf \{x^T L x \mid \|x\| = 1, \mathbf{1}^T x = 0\} \quad (5.8)$$

where $\mathbf{1}$ is the vector with all ones. As $L\mathbf{1} = 0$, $\mathbf{1}$ is the eigen vector of the smallest eigen value, $\lambda_1 = 0$, of L . Equation (5.8) implies that λ_2 is a concave function of L , as it is the point wise infimum of a family of linear functions of L . Hence maximization of λ_2 as a function of the Laplacian matrix L is a convex optimization problem for which a global maximum is guaranteed.

5.3 Feature Graphs and Irrelevant Features

We now relate the maximization of Markov process mixing rate to the identification of irrelevant features for salient object detection. In salient region detection method that

use multiple features (e.g. graph based method in Chapter 4), all features are given equal importance. Our method for unsupervised feature selection aims to identify the important features by representing the image with several ‘feature graphs’, which implies that each feature is encoded into a separate image graph model. The nodes in the ‘feature graphs’ represents 8×8 patches in the image similar to the approach in Chapter 4. The affinity between nodes in a feature graph depends on the respective features. If there are M features available for every node in the image, we have M different feature graphs in which the affinity between node i and node j for the m^{th} feature is defined as equation (5.1) in which

$$w_{ij}^m = e^{-\|(f_i^m - f_j^m)\|^2}, \quad (5.9)$$

where f_i^m and f_j^m are the values of m^{th} feature on node i and j respectively.

It is evident that the connectivity in the M feature graphs will be different. The relevant features or the features that contribute to the salient parts of the image will have feature graphs with ‘bottlenecks’. The term ‘bottleneck’ implies a large group of edges which show very less affinity between two groups (clusters) of nodes. Such bottlenecks arise from the fact that the feature values for a certain feature on the salient part of an image will be different from the feature values on a non-salient part like the background. For example, in Figure 5.2(a), if the feature graph is attributed to the color feature, then the nodes connected by thick lines (indicating strong affinity) fall into two groups, each having a different color value. One group might belong to the salient object and the other to the background and both of them are connected by thin edges (indicating weak affinity). This group of thin edges signifies the presence of a bottleneck. This phenomenon of clustering of nodes does not occur in feature graphs attributed to irrelevant features and hence such graphs will not have bottlenecks. This is illustrated in Figure 5.2(b) for the same set of nodes as in Figure 5.2(a), but for a different feature. Here the affinities for the irrelevant feature vary smoothly (for example, consider the shading variations of color in Figure 5.1(b)) and hence do not

result in a bottleneck in its respective feature graph. Therefore, a Markov process running on an irrelevant feature graph mixes faster compared to a Markov process in a relevant feature graph since in the relevant feature graph the transition time between two groups of nodes connected by very weak edge strengths will be always high. Now consider all possible graph realizations resulting as a linear combination of the feature graphs. When we maximize the Markov mixing rate of a linear combination of these feature graphs, what we are essentially doing is to identify the irrelevant features that does not form any kind of bottleneck.

Figure 5.2: Feature graphs with the same set of nodes showing (a) bottleneck between two groups of nodes (b) no bottleneck. The thicker edges indicate larger affinity between nodes.

5.3.1 Irrelevant Feature Identification

We adopt a semi-definite programming method similar to [76] for maximizing the second smallest eigen value of the Laplacian. However, in [76] a convex optimization method is proposed to find the fastest mixing Markov process among all possible edge weights of the graph realization $G(V, E)$ to give a generic solution. In our proposed method, searching the entire family of graph realizations that gives the fastest mixing

rate will not give us the required specific solution. Hence we restrict our search to all possible linear combinations of the feature graphs. The maximization problem of the second smallest eigen value is formulated as

$$\begin{aligned}
 & \text{maximize} && \lambda_2(L) \\
 & \text{subject to} && L = \sum_k a_k \cdot L_k, \quad \sum_k a_k = 1 \\
 & && a_k \geq 0, \quad \forall k
 \end{aligned} \tag{5.10}$$

where L_k is the Laplacian matrix of the k^{th} feature graph and a_k is the weight associated with the respective feature graph. Now we establish that the constraints on matrix L in Equation (5.10) make it a valid candidate for a Laplacian matrix.

Theorem 1: *If $L_1, L_2, \dots, L_M$ are Laplacian matrices of M feature graphs and if $L = \sum_k a_k \cdot L_k$ under the conditions $\sum_k a_k = 1$ and $a_k \geq 0$, then L is also a Laplacian matrix.*

Proof:

$$\begin{aligned}
 \sum_j L_{ij} &= \sum_j \sum_k a_k \cdot L_{kij} = \sum_k \sum_j a_k \cdot L_{kij} \\
 &= \sum_k a_k \sum_j L_{kij} = 0
 \end{aligned}$$

Also $L_{ij} = \sum_k a_k \cdot L_{kij}$ is negative for $i \neq j$ as $a_k \geq 0$ and $L_{kij} \leq 0 \quad \forall k$ and for $i \neq j$. Similarly all diagonal elements of L are positive proving that L represents the Laplacian matrix of some graph resulting from the linear combination of feature graphs.

The optimization problem in Equation (5.10) can be equivalently expressed as

$$\begin{aligned}
 & \text{maximize} && s \\
 & \text{subject to} && sI \preceq L + \beta \mathbf{1}\mathbf{1}^T \\
 & && L = \sum_k a_k \cdot L_k, \quad \sum_k a_k = 1, \quad a_k \geq 0, \quad \forall k, \beta > 0
 \end{aligned} \tag{5.11}$$

where the symbol $\preceq$ represents matrix inequality. The equivalent optimization problem maximizes a variable s subject to some matrix inequality constraints and some linear equality constraints. The term $\beta \mathbf{1}\mathbf{1}^T$ removes the space of matrices with the eigen vector $\mathbf{1}$ and eigen value 0 from the space of matrices L [7]. Hence maximizing the second smallest eigen value of L is equivalent to maximizing the smallest eigen value of $L + \beta \mathbf{1}\mathbf{1}^T$.

Theorem 2: *The maximum s that satisfies the matrix inequality constraint in Equation (5.11) is the minimum eigen value of the matrix variable $L + \beta \mathbf{1}\mathbf{1}^T$.*

Proof: Let $P = L + \beta \mathbf{1}\mathbf{1}^T$. The smallest eigen value of P can be expressed as ,

$$\lambda_{min} = \inf\left(\frac{x^T P x}{x^T x}\right).$$

Assume $P \prec \lambda_{min} I$. This implies that, $P - \lambda_{min} I$ is negative definite.

$$\Rightarrow x^T (P - \lambda_{min} I) x < 0 \text{ for any } x,$$

$$\Rightarrow x^T P x < \lambda_{min} x^T x,$$

$$\Rightarrow \frac{x^T P x}{x^T x} < \lambda_{min}$$

which is false since the minimum value of $\frac{x^T P x}{x^T x}$ is λ_{min} .

Hence $P \succeq \lambda_{min} I$ and the maximum s that satisfies the constraint $P \succeq sI$ is λ_{min} .

The inequalities in Equation (5.11) can also be expressed as

$$sI - (L + \beta \mathbf{1}\mathbf{1}^T) \preceq 0. \quad (5.12)$$

Since $L = \sum_k a_k L_k$, the matrix inequality in Equation (5.12) can be expressed as

$$sI + a_1(-L_1) + a_2(-L_2) + \dots a_M(-L_M) + \beta(-\mathbf{1}\mathbf{1}^T) \preceq 0. \quad (5.13)$$

which is a standard Linear Matrix Inequality (LMI) constraint in an SDP problem formulation [6] as all the matrices in Equation (5.13) are symmetric. The Equation (5.11) can be seen as the minimization of a linear function subject to an LMI and

Algorithm 1 Feature Weight Calculation for Salient Object Detection

Require: M features on n Nodes of the image

- 1: Calculate the feature graph of every feature with the $n \times n$ affinity matrix A_m for m^{th} feature according to Equations (5.9) and (5.1.)
 - 2: Calculate the Laplacian matrices of M features $L_1, L_2, \dots, L_M$ according to Equations (5.2) and (5.3)
 - 3: Formulate the SDP problem in Equation (5.11) with input data as $L_1, L_2, \dots, L_M$ and variables as $s, \beta, L, a_1, a_2, \dots, a_M$ and solve for $a_1, a_2, \dots, a_M$ using a standard SDP solver.
 - 4: Find the feature weights as $\mathbf{r} = \mathbf{1} - \mathbf{a}$ and normalize the weights.
-

linear equality constraints and can be effectively solved by standard SDP solvers like CVX or SeDuMi. Our proposed algorithm for finding the weights of different features is presented in Algorithm 1.

5.4 Experiments

The effectiveness of the proposed algorithm for unsupervised feature selection is experimentally verified by using it in conjunction with the random walk based method for salient object detection described in Chapter 4, where a random walk is performed on a simple graph model for the image with equal weight for all features. The proposed method will provide the feature weights according to feature relevancy resulting in a much more accurate graph model in the context of salient object detection. The final image graph model is decided by the graph affinity matrix A , which is evaluated as

$$A = \sum_i r_i \cdot A_i \quad (5.14)$$

where $r_1, r_2, \dots, r_M$ are the final feature weights evaluated in Algorithm 1 and $A_1, A_2, \dots, A_M$ are the feature affinity matrices.

The proposed feature selection method falls in the class of unsupervised filter method of feature selection where the feature selection is independent of the classification algorithm that follows. Hence we compare the proposed method with two unsupervised filter methods of feature selection explained in [31] and [55]. The Laplacian score based feature selection method [31] assigns a score to every feature denoting the extent to which each feature 'respects' the data manifold. In [55], an unsupervised filter method is proposed to remove redundant features by evaluating a feature correlation measure for every pair of features. The features are clustered according to their similarity and the representative feature of each set is selected starting from the most dissimilar feature set. However, as our proposed method weighs different features according to their importance, we use the feature correlation measure in [55] to weigh the features while comparing with the proposed method.

We used the same dataset of 5000 images used in Chapter 4 which was originally provided by [48]. The dataset contains user annotations as rectangular boxes around the most salient region in the image. The performance of the algorithm is objectively evaluated with the ground truth available in the data set. We use the same seven dimensional feature set used in Chapter 4. The features are color (Cb and Cr) and texture (orientation complexity values at five different scales). According to our algorithm, there will be seven different feature graphs (transition matrices), one for each feature. The optimal weights of features are obtained as explained in Algorithm 1 with semi-definite programming in Matlab using the cvx package [26]. Using the optimal feature weights, a new graph model for the image is designed according to Equation (5.14), which is used with the method in Chapter 4 to detect the salient regions.

Figure 5.3 shows saliency maps on some example images for which the original algorithm in Chapter 4 has resulted in wrong or partially correct saliency maps. The original images are shown in figure 5.3(a), 5.3(d) and figures 5.3(b) and 5.3(e)

Figure 5.3: Comparison of saliency maps for graph based method in Chapter 4 with the proposed feature selection . (a),(d) Original Images. (b),(e) Saliency maps using method in Chapter 4. (c),(f) Saliency maps of Chapter 4 with the proposed feature selection.

Table 5.1: F-measure comparison for different methods.

Method	F-Measure
Random walk method in Chapter 4	0.69
Chapter 4 method with Laplacian score [31] based feature selection	0.66
Chapter 4 method with correlation based feature selection [55]	0.66
Chapter 4 method with proposed unsupervised feature selection	0.72

show the saliency map generated by Chapter 4, in which the image graph model is designed with equal importance to all features, whether relevant or irrelevant . Figure 5.3(c) and 5.3(f) shows the saliency maps resulting from the proposed feature selection method used with graph based method in Chapter 4. It can be seen that the saliency maps are more accurate in Figure 5.3(c) and 5.3(f) where the features are weighted according to their relevance. For example, consider the first image in Figure 5.3(a). The color feature plays a good role in saliency for this image and the proposed algorithm detects the salient region more accurately by giving more weight to color feature. On the contrary, for the aerial image showing water and land in the third row of 5.3(d), the color features do not have much role in deciding the salient region and the proposed algorithm gets a more accurate saliency map by giving more priority to the texture feature. In most of the real life images it is some specific mixture of various features (e.g. a combination of color and orientation entropy feature at some scale) that proves to be relevant in the context of salient object detection. The proposed algorithm intends to model this rightly weighted combination of features as can be testified from the results for other images in Figure 5.3. It has to be also noted that, though our proposed method assigns soft weights to different features according to their relevance, a hard decision on the features as in a selection of a feature subset can be made by applying some heuristics to the soft weights (for example, select the minimum number of features that satisfies 90% of the total soft weights assigned).

The ground truth available for the dataset allows us to objectively compare various algorithms using precision, recall and F-measure. Precision is calculated as ratio of the total saliency in the saliency map captured inside the user annotated rectangle (sum of intensities inside user box) to the total saliency computed for the image (sum of intensities for full image). Recall is calculated as the ratio of the total saliency captured inside the user annotated window to the area of the user annotated window. F-measure, which is the harmonic mean of precision and recall, is the final metric used for comparison. Table 5.1 lists the F-measure computed for various feature selection methods on the entire dataset. We compare the proposed unsupervised feature selection algorithm with the Laplacian score based feature selection algorithm [31] and the correlation based feature selection algorithm [55]. It can be seen from the table that the proposed method significantly improves the performance of the random walk method. The unsupervised feature selection methods in [31] and [55] are not tailored for salient object detection. Hence the performance of these methods are inferior to the original algorithm in Chapter 4 where all features are given equal weightage. In order to prove the statistical significance of the F-Measure results, a 't-test' is conducted for the research hypothesis - 'The average F-Measure with feature selection is higher than the average F-Measure without feature selection'. The values of F-measure for all samples with and without proposed feature selection are collected and its mean and variance is calculated. The t-test resulted in a 't' value of 6.1 on 5000 samples in both cases. This corresponds to a false alarm rate or Type I Error less than 5% which again implies that the tests are statistically significant.

It has to be noted that the methods proposed by [48] and [80] claims better f-measure than the proposed graph-based method with feature selection on the same data set. However this increase in f-measure can be attributed to two major factors as discussed below.

1)[48] and [80] reports an f-measure of around 0.79 with the MSRA database. The method of computation of f-measure in [48] and [80] is different from the method in the proposed work. In [48] and [80], the rectangular box constructed over the binary saliency map is considered as the final output of the algorithm. This rectangular box is compared against user drawn box to compute the f-measure while in our work the final segmented image is compared with the user box to calculate the f-measure. We evaluated the f-measure of our proposed algorithm in the same manner as is done in [48] and [80] and got an improved f-measure of 0.77 for the entire MSRA dataset.

2)The method in [48] is a supervised method in which 1000 annotated images were used to train a salient region detector which was tested on the rest of 4000 images. In [80], after obtaining the saliency map, a graph cut based segmentation algorithm is used to segment the image. Some segments with higher saliency are combined to form final saliency map using the computationally efficient but non-exhaustive sub-window search method in [44]. The accuracy of [80] depends on the accuracy of segmentation algorithm and on the techniques used to combine different segments to obtain the final salient object. Hence as an unsupervised and independent method (like [82] and [28]) that provides a good f-measure on a large dataset, the graph based approach is attractive and demands attention for future enhancements.

5.5 Conclusion

We have proposed an unsupervised feature selection method for salient object detection by identifying the irrelevant features. The problem formulation is convex and is guaranteed to achieve the global optimum. The proposed feature selection method can be used as a filter method to improve the performance of salient object detection algorithms. The power of the proposed method is demonstrated by using it in conjunction with the graph based algorithm in Chapter 4 for salient object detection.

Chapter 6

A Linear Dynamical System Framework for Salient Motion Detection

6.1 Introduction

The problem of identifying salient regions in images has received much attention over the past few years [88, 48]. There has been significant progress showing encouraging results that could be quantified, thanks to datasets with ground truth containing salient objects. The approaches mainly involve the identification of ‘rare’ features in an image that potentially contribute to visual saliency. However, when detecting saliency in videos, it is important to consider the inherent temporal information in addition to the spatial features in a single frame. This leads us to the notion of ‘motion saliency’, which could be the primary factor that contributes to the overall saliency in videos. The detection of salient motion in videos finds applications in object tracking and surveillance [96, 90], video retargeting [69] and activity recognition [74]. For example, surveillance and tracking applications are aided by the information on

‘what to follow’, while video retargeting applications can use the knowledge of salient motion to resize video frames with minimum distortion.

In the proposed method for the detection of saliency in videos, the sequence of frames in a video is represented as a Multi-Input Multi-Output (MIMO) Auto-Regressive(AR) state-space model. We explore the relationship between the output of the model and the states of the model to evaluate the relevance of every output pixel. The dynamics of the salient region is modeled by analyzing the spatio-temporal patch around the most relevant set of output pixels within a linear dynamical system framework. A distance measure [53] is used to compare the AR model around the most relevant spatio-temporal patch with other models. The proposed method does not require any kind of explicit background modeling. The contributions in this chapter are two-fold:

- We introduce the novel concept of evaluating the saliency of the output pixels in a video frame by exploring its relationships to the hidden states when the video sequence is represented as a MIMO state-space model. The concept of ‘observability’ from the output pixels is shown to provide a good clue to the salient motion in natural videos.
- The dynamics of salient region is modeled by fitting a DT model to the spatio-temporal patch having maximum pixel saliency. This model is further used for comparison with the other patch models in the spatio-temporal volume to compute the region based saliency maps. The final saliency map is computed from the pixel saliency maps and the region based saliency maps

The chapter is organized as follows. Section 6.2 reviews the theory of dynamic textures, observability of a state space model and non-Euclidian distance measures

for AR processes. Section 6.3 discusses the computation of pixel saliency map by first establishing the relationship between the observability of the linear dynamical system and salient motion in a video. The method of computing saliency map by evaluating observability from different outputs is further elaborated in Section 6.3. Section 6.4 details a local and global approach to the computation of region based saliency maps based on salient region dynamics. Section 6.5 presents the experimental results, which contain subjective as well as objective evaluations and comparisons with other state-of-the-art methods. Finally, conclusions are given in Section 6.6.

6.2 State-Space Representation

In this section, we present a review of (i) the state space representation of natural video sequences, (ii) the concept of observability of state space models that originated from control systems theory, and (iii) the non-Euclidian distance measures between AR processes.

6.2.1 Dynamic Textures as Linear Dynamical Systems

In [17], Doretto et. al proposed a state space model known as dynamic textures, in which the frames in a video sequence are represented as the output of a linear dynamical system. The output of the model at time t is the lexicographically ordered frame vector observed at time t , $y(t) \in \mathbb{R}^m$, which is the result of a hidden state Markov process driven by a Gaussian observation input noise. The state-space model can be represented as

$$\begin{aligned}x(t+1) &= \mathbf{A}x(t) + v(t) \\ y(t) &= \mathbf{C}x(t) + w(t)\end{aligned}\tag{6.1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the state transition matrix that describes how the state vector $x(t) \in \mathbb{R}^n$ evolves. $v(t)$ is the input Gaussian observation noise which is an IID realization from the density, $\mathcal{N}(0, \mathbf{Q})$, where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is the covariance matrix of the zero-mean Gaussian process. $w(t)$ is the output observation noise sampled from $\mathcal{N}(0, \mathbf{R})$, where $\mathbf{R} \in \mathbb{R}^{m \times m}$ is the covariance matrix. The $m \times n$ appearance matrix $\mathbf{C}$ relates the current state to the observed output. The system identification problem proposed by [17] estimates the matrices $\mathbf{A}$, $\mathbf{C}$, $\mathbf{Q}$ and $\mathbf{R}$ from the measurements $y(1), y(2) \dots y(\tau)$.

Let $\mathbf{Y}_1^\tau = [y(1), y(2), \dots y(\tau)] \in \mathbb{R}^{m \times \tau}$ and $\mathbf{X}_1^\tau = [x(1), x(2), \dots x(\tau)] \in \mathbb{R}^{n \times \tau}$ be the matrices stacked with observed output vectors and the hidden state vectors, respectively. The subscripts and the superscripts indicate the first and the last time instants, respectively, in the output or state space. If the SVD decomposition of the stacked up output observation vectors is given by $\mathbf{Y}_1^\tau = U\Sigma V$, then $\mathbf{C}(\tau)$ and $\mathbf{X}_1^\tau$ can be estimated as [50]

$$\begin{aligned} \mathbf{C}(\tau) &= U \\ \mathbf{X}_1^\tau &= \Sigma V, \end{aligned} \tag{6.2}$$

where $\mathbf{C}(\tau)$ is the estimate of the appearance matrix $\mathbf{C}$ using τ observations. The state transition matrix at time τ , $\mathbf{A}(\tau)$ is determined as the unique solution to the linear problem $\mathbf{A}(\tau) = \text{argmin}_A \|\mathbf{X}_2^\tau - A\mathbf{X}_1^{\tau-1}\|_F$ where $\|\cdot\|_F$ is the Frobenius norm. The variance of state estimates are further used to estimate the input noise covariance matrix $\mathbf{Q}$. When a video sequence is modeled as a linear dynamical system, it captures the holistic representation of motion in the scene. The appearance component of the scene is encoded into the matrix $\mathbf{Y}_1^\tau$ while the motion component is encoded into the state sequence [10]. Further, the columns of $\mathbf{C}(\tau)$ consists of the principal components of the video sequence during $t = 1$ to τ and the output at time τ , $y(\tau)$, is the linear combination of columns of $\mathbf{C}(\tau)$ weighted by the appropriate state estimates from the vector $x(\tau)$.

6.2.2 Observability of a State Space Model

Observability of a state space model measures how well the internal states of a system can be inferred from the knowledge of its outputs. Consider the following more generic form of equation (6.1):

$$\begin{aligned} x(t+1) &= \mathbf{A}x(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}x(t) + \mathbf{D}u(t). \end{aligned} \quad (6.3)$$

Formally, the state-space model in equation (6.3) is said to be observable if any unknown initial state $x(0)$ can be uniquely determined from the knowledge of output $y(t)$ together with the knowledge of the input $u(t)$ in a finite time interval $[0, t_1]$. In other words, observability relates the observed outputs of a system to its hidden internal states. If a system is not observable, it implies that the current values of some of its internal states cannot be estimated from its current or future outputs. The observability of a linear time invariant system can be easily determined from the rank conditions on the $nm \times n$ ‘observability matrix’, O , formed by the $n \times n$ matrix $\mathbf{A}$ and $m \times n$ matrix $\mathbf{C}$ as [4],

$$O = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix}. \quad (6.4)$$

The system described by $(\mathbf{A}, \mathbf{C})$ is completely observable when the $nm \times n$ observability matrix O has full column rank n . Note that any invertible linear transform, $\mathbf{T} \in \mathbb{R}^{n \times n}$, acting on the states of the system can result in an equivalent representation with a different set of system parameters: $\bar{\mathbf{A}} = \mathbf{TAT}^{-1}$, $\bar{\mathbf{B}} = \mathbf{TB}$, $\bar{\mathbf{C}} = \mathbf{CT}^{-1}$, $\bar{\mathbf{D}} = \mathbf{D}$ [11]. This *equivalence class* of systems will have the same transfer function matrix

that relates the inputs and outputs of the system. Also, the eigen values of $\mathbf{A}$ and $\bar{\mathbf{A}}$ are the same for any invertible $\mathbf{T}$.

If the pair $(\mathbf{A}, \mathbf{C})$ is unobservable, there is an equivalent representation of the system with a non-singular transform $\mathbf{T}$ such that

$$\hat{\mathbf{A}} = \mathbf{TAT}^{-1} = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}, \quad \hat{\mathbf{C}} = \mathbf{CT}^{-1} = \begin{bmatrix} C_1 & 0 \end{bmatrix}$$

where $A_1 \in \mathbb{R}^{n_0 \times n_0}$ and $C_1 \in \mathbb{R}^{m \times n_0}$ form the observable pair (A_1, C_1) [4]. The n_0 eigenvalues of A_1 are called the observable modes or eigenvalues of the system and the $n - n_0$ eigenvalues of A_2 are the unobservable modes or eigenvalues. All observable eigen values are present in the system transfer function $C(sI - A)^{-1}B$. Observability of a system remains unchanged in all forms of its equivalent representations. Hence, observability of eigen values of $\mathbf{A}$ is the deciding factor that determines the system observability irrespective of a specific realization among all possible realizations in the equivalence class. In a MIMO system, the observability from a specific output to a particular eigen value indicates the influence of that particular mode on the values of the specific output. A mode that is not observable from a specific output will not be present in the equation of that particular output. If a mode is not observable from any of the outputs of a MIMO system, that mode is an unobservable mode for the system. A state space system is said to be completely observable if all the eigen values of $\mathbf{A}$ are observable. We can summarize this short discussion by listing three interpretations of observability: i) observability of an LDS system is determined from the rank of the observability matrix O , ii) observability can also be attributed to state variables or eigen values of the state transition matrix $\mathbf{A}$, and iii) observability can be measured from the outputs of the system to any eigen value of $\mathbf{A}$. We are specifically interested in the third interpretation since the observability from the outputs of a system has a strong relation to salient motion as will be seen in the next section.

6.2.3 Distance between AR Models

The dynamic texture (DT) models are auto-regressive in its nature and consequently the distance between two different DTs is not Euclidian. In [53], the distance measure between two stable AR models M_1 of order n_1 and having poles at α_i and M_2 of order n_2 and having poles at β_j is shown to be

$$d(M_1, M_2)^2 = \ln \frac{\prod_{i=1}^{n_1} \prod_{j=1}^{n_2} |1 - \alpha_i^* \beta_j|^2}{\prod_{i,j=1}^{n_1} (1 - \alpha_i^* \alpha_j) \prod_{i,j=1}^{n_2} (1 - \beta_i^* \beta_j)}, \quad (6.5)$$

α_i^* and β_i^* are complex conjugates of α_i and β_i . The poles of the DT model are the eigen values of the state transition matrix $\mathbf{A}$ [4].

6.3 Pixel Saliency Map Using Observability from Output Pixels

We represent the video sequence as a linear dynamical system that can be described by the state space equations of eq.(6.3). Consider the video sequences with various dynamic backgrounds shown in figure 6.1. The ‘jump’ sequence in figure 6.1(a) has smoke filled environment as the background, while the ‘ocean’ sequence in figure 6.1(b) has waves in the sea as the dynamic background. The snowfall and the snow thrown out from the skiing gear is the complex dynamic background for the ‘skiing’ video sequence in figure 6.1(c).

For the human visual system, the salient or foreground motion in these three different scenarios are the cyclist jumping, the people walking and the movement of the skiers, respectively. This inference is made in the presence of complex and dynamic backgrounds in all the three cases. The salient motion is considered as one that is more ‘predictable’ for a human observer compared to the relatively unpredictable backgrounds. For example, the smoke moving in the background of the ‘jump’ sequence has haphazard motion, while the cyclist has a regular motion. Similarly, the

Figure 6.1: Video sequences with different dynamic backgrounds (a) ‘jump’ sequence with smoke, (b) ‘ocean’ sequence with waves and (c) ‘skiing’ sequence with snow.

movement of skiers is more robust and predictable compared to the snow fall and the motion of snow thrown out while skiing. Though waves in figure 6.1(b) are less complex in its motion, the motion related to people walking becomes relatively simpler compared to the motion of the waves. Thus, we attach the notion of ‘predictability’ or ‘regularity’ to salient motion. It is less restrictive than the notion of linear motion as in [87], [86] and it can handle various kinds of motion. Also, we intend to focus only on the relative predictability of different motions.

As explained in Section 6.2.1, the dynamic texture model represents a scene of τ frames in a holistic manner using the state space framework where the matrix $\mathbf{C}$ is the appearance matrix which encodes the visual appearance of the texture and the states $x(t)$ encode the motion of the entire sequence [10]. The observability from a specific pixel in output $y(t)$ is a measure of inference that can be made on the values of the state variables, which are a holistic characterization of the motion in the video sequence. Pixels belonging to complex unpredictable backgrounds will have

less observability to state variables compared to pixels belonging to regions of simple predictable motions. This is due to the fact that we are modeling the video sequence over a period of time (τ frames), and the motion components belonging to complex unpredictable regions are hard to estimate from its holistic characterization. Pixels belonging to static regions will have zero observability as they do not contribute to the inference of any motion and hence the inference of state variables. Hence, observability from a pixel computed under the framework in equation (6.3) can be taken as a clue for the importance of the pixel with regards to how much influence it holds in describing the holistic motion of the scene.

6.3.1 Quantitative Measure for Observability

In Section 6.2.2 we reviewed the basic definition of observability and a simple rank based method to evaluate whether a system is observable or not. However such methods only help to make a binary decision on the observability of a system and do not provide any insight into the observability from different outputs to the eigenvalues of state transition matrix $\mathbf{A}$. Recall that observability is evaluated with respect to eigenvalues of $\mathbf{A}$ since it is invariant under various equivalent system representations. In [77], Tarokh discusses a simple and computationally efficient method to evaluate a measure of observability for the i^{th} eigen value λ_i of the state transition matrix $\mathbf{A}$ for continuous-time, time-invariant systems as

$$m_{oi} = |\delta_i| [e_i^* \mathbf{C}^T \mathbf{C} e_i]^{\frac{1}{2}}, \quad (6.6)$$

where e_i is the right eigen vector of the transition matrix $\mathbf{A}$ corresponding to λ_i and δ_i is calculated as

$$|\delta_i| = \left| \prod_{j=1, i \neq j}^n (\lambda_i - \lambda_j) \right|. \quad (6.7)$$

It can be further concluded from equation (6.6) that the angle between a specific row of output matrix $\mathbf{C}$ and the right eigen vector e_i is a measure of observability from the respective output. If all the row vectors of the output matrix $\mathbf{C}$ are orthogonal to the right eigen vector e_i , that specific λ_i will have zero observability [77]. We are particularly interested in the contribution of any eigen value λ_i to a specific output which can be interpreted as the visibility of that eigen value from the specific output. In our method, every output corresponds to a pixel p in the output frame since $y(t)$ is the output frame lexicographically ordered as the $m \times 1$ vector. Hence, we evaluate the observability measure from a pixel p to λ_i as

$$m_{pi} = \mathbf{C}[p] \times e_i \quad (6.8)$$

where $\mathbf{C}[p]$ is the p^{th} row of $\mathbf{C}$. An intuitive explanation for equation (6.8) for discrete-time time-invariant systems is given in Appendix A.1. The total observability from output p , denoted by m_p , is calculated as the sum of its observability measures to all the eigen values of $\mathbf{A}$ i.e.,

$$m_p = \sum_{i=1}^n m_{pi} \quad (6.9)$$

We evaluate the observability from all the pixels in the output in a similar fashion and normalize the values between 0 and 1 to obtain the pixel saliency map.

6.3.2 Sustained Observability

The observability measure evaluated for pixels in the output vector provides a good clue for salient motion. However, in this process some other pixels in the dynamic background can also possess some observability as can be seen in the 'cyclists' video sequence whose 12th frame is shown in figure 6.3(a). Figure 6.3(b) and (c) show the saliency maps for adjacent frame buffers wherein some part of the dynamic back-

Figure 6.2: Sustained observability calculated over adjacent video buffers. $s1$ and $s2$ are saliency maps computed using dynamic texture models from frames 1 to q and 2 to $q+1$.

ground also acquires observability. The idea of sustained observability is to lock on to points of consistent salient motion irrespective of the potential errors that would have happened due to the sub-optimal system modeling using [17] for a single buffer. Pixels having 'sustained observability' are those that maintain high observability over successive saliency maps computed for adjacent frame buffers. Figure 6.2 shows two such buffers each of length q - the first buffer is from frame 1 to frame q and the second is from frame 2 to frame $q + 1$.

The dynamic texture models are evaluated for the adjacent buffers resulting in saliency maps $s1$ and $s2$. The final saliency map is computed by multiplying the two maps and normalizing it to $[0, 1]$. Thus, we give more saliency to those pixels that hold the observability across the adjacent models. As seen in figure 6.3, although different regions in the dynamic background show some saliency in individual binary saliency maps $s1$ and $s2$, the final saliency map in figure 6.3(d) correctly identifies points that show high observability consistently.

Figure 6.3: Example showing the generation of final saliency map with sustained observability a) 12th frame of 'cyclists' video (b, c) saliency maps s1 and s2 computed using frame buffers with frames 1 to 21 and 2 to 22 d) final saliency map

6.4 Region Based Saliency Map Using Salient Patch Dynamics

The pixel saliency map computed in Section 6.3 evaluates the importance of each pixel in inferring the holistic dynamics of the scene. However, the interior of large homogeneous salient objects that occupy a majority of the pixels in the frame can appear to be relatively static over the frame buffer. Consequently, they will be marked as less salient due to their apparent lesser involvement in the holistic dynamics of the scene. This drawback of the pixel saliency map is addressed by computing region based saliency maps. In the proposed method 'regions' correspond to spatio-temporal patches of fixed size selected from the spatio-temporal volume of the frame buffer. The dynamics of different regions are compared with the dynamics of the most salient region/patch, which can be inferred from the pixel saliency map. However, the dy-

namics of the most salient region hold very specific information of the region dynamics and may not always generalize well to the dynamics of the salient object as a whole. Hence, we also compare the dynamics of different regions to the dynamics of the whole of spatio-temporal volume (frame buffer). This can be considered as a combination of local and global approaches in deciding region saliency. Finally, to compare the dynamics of different models we rely on a distance measure similar to the one discussed in equation 6.5. Figure 6.4 shows a block diagram describing various functional modules in the evaluation of region based saliency maps. The following subsections discuss the computation of region based saliency maps in detail.

6.4.1 Pole Analysis for Comparison of Region Dynamics

According to equation (6.5), the distance between two AR models contains knowledge of the poles of the two models. In the proposed method, we intend to compare the dynamics of two different models which are encoded into their respective state transition matrices. Therefore, their dynamics are reflected in the eigen values (poles) of the respective matrices. The stable poles of the LDS lie inside the unit circle in the complex plane. In [91], Yuan et al. show that in a dynamic texture synthesis process, the poles that lie near to the unit circle influence the output with its ‘oscillatory’ nature while the influence of poles that lie close to origin ‘fades’ away as time progresses. They conclude that poles of ‘oscillatory’ nature (close to the unit circle) contribute to a more accurate modeling of the dynamic texture. Thus, in order to compare system dynamics, we need a distance measure that prioritizes poles that are oscillatory in nature over poles of fading nature. We formulate such a distance measure by considering the numerator of equation (6.5).

To analyze this in detail, consider the simple case of comparing the dynamics of two models with poles **a** and **b** on the complex plane shown in figure 6.5. The two poles have magnitudes close to 1 (hence are oscillatory in nature) and are close

Figure 6.4: Block Diagram detailing different functional modules in evaluation of region based saliency

Figure 6.5: Poles **a**, **b**, **c**, **e**, and **f** inside the unit circle on the complex plane. **a*** and **c*** are the complex conjugates of **a** and **c** respectively. **a**, **b** and **c** are the poles close to unit circle and hence of oscillatory nature. **e** and **f** are poles close to the origin and hence, of fading nature.

to each other. Let θ is the approximate angle made by the poles with the real-axis and let $|1 - \mathbf{a}^*\mathbf{b}|$ be the distance factor between the poles, where $\mathbf{a}^*$ is the complex conjugate of pole **a**. The product $\mathbf{a}^*\mathbf{b}$ consists of the product of the magnitudes of $\mathbf{a}^*$ and **b** and sum of their phase angles. The sum of the phase angles are close to zero. Hence, the product $\mathbf{a}^*\mathbf{b}$ lies close to the real-axis near $1 + j0$, which implies that $|1 - \mathbf{a}^*\mathbf{b}|$ is close to zero. Next, consider the pole **c** that also has an oscillatory nature but lying far away from **b**. The maximum distance of **c** from **b** will be when the former makes an angle $180 + \theta$ with the real axis. The distance measure between them is $|1 - \mathbf{c}^*\mathbf{b}|$, where the product $\mathbf{c}^*\mathbf{b}$ has a phase angle $180 - \theta + \theta = 180^\circ$. This implies that $\mathbf{c}^*\mathbf{b}$ lies on the negative real-axis and far from $1 + j0$ resulting in a large value for $|1 - \mathbf{c}^*\mathbf{b}|$. Thus, for poles of oscillatory nature, the factor $|1 - \alpha^*\beta|$ is sensitive to the relative location of the two poles, where α is the pole of first model and β is the pole of second model.

Now, consider distance of the poles $\mathbf{e}$ and $\mathbf{f}$ to the pole $\mathbf{b}$. As can be verified from figure 6.5 poles $\mathbf{e}$ and $\mathbf{f}$ lie close to the origin. The product $\mathbf{e}^*\mathbf{b}$ and $\mathbf{f}^*\mathbf{b}$ will also lie close to the origin (because the magnitude is low) and the distance measure $|1 - \mathbf{e}^*\mathbf{b}|$ and $|1 - \mathbf{f}^*\mathbf{b}|$ is comparatively insensitive to the relative location of the poles lying close to the origin. Therefore, by computing only the factor $\prod_{i=1}^{n_1} \prod_{j=1}^{n_2} |1 - \alpha_i^* \beta_j|$ for the distance measure, we can prioritize poles of oscillatory nature over poles of fading nature resulting in a better comparison of system dynamics.

6.4.2 Region Based Saliency map using Local Dynamics

The first step in determining the region based saliency map is to understand the nature of the dynamics of the most salient region in the frame buffer. The pixel saliency map contains the most salient pixel having the maximum sustained observability to the frame dynamics. We select a spatio-temporal patch of fixed size with largest average saliency around the most salient pixel. Since this spatio-temporal patch is deemed to fall on the salient object, it can be modeled as a DT that is a fair approximation of the dynamics of the salient object. We use the same sub-optimal method proposed by [17] and described in Section 6.2.1 to model the DT over the salient patch. The DT model around the salient patch is represented as M_s with parameters (A_s, B_s, C_s, D_s) . The eigen values of the state transition matrix A_s captures the information regarding the holistic dynamics in the salient region. This salient dynamics constrained to the local region of maximum saliency is compared with spatio-temporal patches selected from the frame buffer. The frame buffer is divided into N non-overlapping spatio-temporal patches of fixed size and each patch is modeled as described in Section 6.2.1. The i^{th} patch is represented by model M_i with parameters (A_i, B_i, C_i, D_i) . The distance between the i^{th} patch and the maximum salient patch is calculated as

$$D_L(M_i, M_s) = \prod_{i=1}^n \prod_{j=1}^n |1 - \alpha_i^* \beta_{s_j}| \quad (6.10)$$

where α_i are the poles of the i^{th} patch and β_{s_j} are the poles of the maximum salient patch. Both models have the same order n . The region based saliency map is computed using the distance measure $D_L(M_i, M_s)$ for N patches, normalized between 0 and 1. A higher value of $D_L(M_i, M_s)$ implies less saliency for patch M_i . This region based saliency map is local in nature since it compares the dynamics of a patch with the local dynamics of the maximum salient patch.

6.4.3 Region Based Saliency map using Global Dynamics

As mentioned at the beginning of section 6.4, the dynamics of the most salient region is local in nature and may not always agree with the dynamics of the entire salient object. We assume that the salient object as a whole should have dynamics ‘similar’ to the salient region but need not match its dynamics exactly. The basis of this assumption is the existence of minor deviations in the model when generalizing the dynamics of a part of the object to the dynamics of the whole object. This problem is addressed by a global approach to comparing the dynamics in the spatio-temporal buffer. Specifically, we consider the same N spatio-temporal patches of the local approach, but compare its dynamics with the model of the entire frame buffer. However, the poles of the frame buffer model are prioritized according to its affinity towards the poles of the salient patch so as to give more importance to the ‘types’ of motion occurring around the most salient region. The distance of the i^{th} pole, β_{F_i} , of frame buffer to all poles of salient patch is

$$D_i = \prod_{j=1}^n |1 - \beta_{F_i}^* \beta_{s_j}|. \quad (6.11)$$

The distances D_i are normalized between 0 and 1. A larger distance D_i for a pole β_{F_i} in the frame buffer model implies that β_{F_i} is less probable to represent some motion originating from the salient object. Hence the priority of β_{F_i} is calculated as $Pr_i = 1 - D_i$. In the computation of distance measure between the N spatio-temporal

patches and the frame buffer model, we prioritize the poles of frame buffer such that the distance measure of frame buffer poles with less priority has least contribution to the total distance measure. This objective is achieved by re-designing the distance factor equation between frame buffer and spatio-temporal patches as

$$D_G(M_i, M_F) = \prod_{i=1}^n \prod_{j=1}^n (|1 - \alpha_i^* \beta_{F_j}|)^{Pr_j}. \quad (6.12)$$

It should be noted that we have now introduced two levels of priority for the system poles. The first level prioritizes poles of oscillatory nature of both models while computing the distance measure and the second level prioritizes poles of the frame buffer model only while computing the global distance. The global region based saliency map is computed using the distance measure $D_G(M_i, M_F)$ for N patches normalized between 0 and 1. The final saliency map is computed as the product of three maps - pixel saliency map, local region based saliency map and global region based saliency map. The final saliency map is also normalized between 0 to 1.

6.5 Experiments

The performance of the proposed algorithm is tested both subjectively and objectively. The objective evaluation is aided by a dataset for salient motion detection available at [1]. It consists of 18 video sequences that are captured under diverse situations, e.g. surveillance videos with static backgrounds, sports videos with complex dynamic backgrounds and others also with dynamic backgrounds that could interfere with the detection of salient motion. The ground truth of salient motion is manually annotated. This dataset was used by Mahadevan et al. in [50] where a quantitative evaluation of their algorithm in terms of Equal Error Rate (EER) was presented together with comparisons with other state-of-the-art methods. We also test our algorithm on some wildlife sequences, which we believe is also an appropriate and

challenging data since wildlife videos usually contain complex dynamic backgrounds due to moving vegetation etc. In the experiments, the buffer size is 23 frames ($q = 23$) and the number of states is fixed as 10 ($n = 10$). The patch size used for the evaluation of the region saliency map is experimentally chosen as 13×13 .

6.5.1 Subjective Evaluation

First, we present results for subjective evaluation on the wildlife sequences. The motion of the animals is more robust and salient compared to background motion of, say, vegetation or water. Examples of distracting background motions are the splashing of water in the first row of figure 6.6 and the motion of background vegetation in the third row. The figure shows 4 frames from each sequence and their corresponding saliency maps. As shown in the figure, the proposed algorithm has been able to identify the object containing the salient motion. Of particular interest is the polar bear sequence where a cub emerges from behind the bear and is also correctly identified by the algorithm as salient. There is no requirement that there should be only one salient object at any time.

We also illustrate the performance of the algorithm on three videos from the salient motion dataset, for subjective evaluation. Figure 6.7 shows some of the frames and their corresponding saliency maps obtained by combining pixel and regional (local and global) saliency maps. The 'flock' sequence in figure 6.7(a) is captured with a static camera with the flock of birds as the dynamic background while the 'skiing' sequence in figure 6.7(c) has snow fall in the background with some camera motion. The 'peds' sequence in figure 6.7(e) is a typical surveillance video watching pedestrians on a road. The saliency maps in figure 6.7(b), (d) and (f) show that the proposed method detects the regions of salient motion accurately under different capturing conditions. Note that no thresholds are applied to the saliency maps shown in figure 6.7; what is shown are actual saliency values of pixels obtained using the proposed method.

Figure 6.6: Saliency maps obtained using the proposed method on non-consecutive frames of various video sequences showing wildlife. Every row shows frames from a specific video followed by the row of saliency maps of respective frames.

Figure 6.7: Saliency maps obtained using the proposed sustained observability method on various video sequences in the database from [50] a) 'flock' sequence (flock of birds as background) b) saliency map for 'flock' sequence c) 'skiing' sequence (snow-fall as background) d) saliency map for 'skiing' sequence e) 'peds' sequence (static camera surveillance) f) saliency map for 'peds' sequence

6.5.2 Quantitative Evaluation

In this section, we present results of the quantitative evaluation of salient motion detection on the manually annotated ground truth data from [50]. The final saliency maps are thresholded at a large number of values to compute the false alarm rate, α , and the detection rate, β . The performance of the algorithm is evaluated by the Equal Error Rate (EER), which is the error at which the false alarm rate is equal to the miss rate, i.e., $\alpha = 1 - \beta$. Lesser the EER, better is the performance of the algorithm.

Figure 6.8: Average equal error rates for different patch sizes used in the computation of regional saliency maps. X-axis shows the size of the square patch and Y-axis shows the average equal error rate in percentage

The first experiment in this section is to study the influence of patch size on the accuracy of salient motion detection. Figure 6.8 shows the variation of average equal error rates with respect to different patch sizes used in computing regional saliency. It can be seen from the figure that the average equal error rate decreases with the increase in patch size, reaches the minimum value for a patch size of 13×13 and then increases with the increase in patch size. Thus we use 13×13 as the

patch size in all our experiments. Table 6.1 shows the equal error rates (in %) for saliency maps obtained with different methods for various video sequences available in the dataset. The methods used for comparison include (1) Discriminant Saliency Method (DS) [50] (2) 'Surprise' model from Itti and Baldi (Su) [36] (3) Non-parametric Kernel Density Estimator (KDE) method [21] (4) Linear Dynamical Model of Monnet et. al. (Mo) [57] (5) modified Gaussian Mixture Model (GMM) method [98]. The proposed sustained observability method is indicated as **SO** in the table. The last row of the table shows the average EER over all the sequences. The EER for the methods that we compare with, are reproduced from [50]. On comparison with GMM, KDE and 'Surprise' model, the proposed algorithm and DS [50] method has the advantage of simultaneous modeling of spatial and temporal components by making use of the Dynamic Texture (DT) model. The proposed method does not require any background estimation to be performed while errors in estimating highly dynamic backgrounds reduce the accuracy of methods in GMM, KDE and Monnet et.al. As seen from table 6.1 the proposed method outperforms all the methods except for DS. However, the former has significant computational advantages over the latter. According to [50], the computational complexity of the DS method is 37 seconds for processing a single video frame of size 240×320 pixels with a Matlab implementation on a PC with 3 Ghz CPU and 2 GB RAM. The proposed sustained observability method takes only about 1.3 seconds for processing the video frame of same size with the Matlab implementation on a 2.66 GHz dual core CPU with 2 GB RAM.

6.5.3 Effect of Region Saliency Maps

In this section, we consider the effectiveness of region saliency maps in improving the salient motion detection results. In table 6.2, EERs are shown for the pixel saliency map (S_P), for pixel saliency map used in conjunction with local region saliency map (S_{PL}) and with global region saliency map (S_{PG}), and lastly, for the final saliency

Figure 6.9: Improvement in saliency maps using combined pixel and region based methods over pixel only method. The saliency maps shown are binary maps with the same level of threshold applied to both pixel only maps and the combined pixel and region maps. a) original video frame b) Pixel saliency map c) Final saliency map which is a combination of pixel and region based maps

Table 6.1: Equal Error Rates (in %) for different algorithms for the dataset in [50] compared to the proposed method (**SO**).

	DS	Su	KDE	Mo	GMM	SO
skiing	3	26	46	11	36	4.3
bottle	2	5	38	17	25	4.3
surf	4	30	36	10	23	6.8
hockey	24	28	35	29	39	23.9
cyclists	8	41	44	28	36	9.2
land	3	31	54	16	40	14.2
birds	5	19	20	7	23	11
zodiac	1	19	20	3	40	2.3
chopper	5	13	43	8	35	4.1
peds	7	37	17	11	11	7.5
flock	15	23	33	31	34	17
traffic	3	46	39	9	34	5.5
boat	9	9	13	15	15	15.5
freeway	6	43	21	31	25	7.9
jump	15	25	51	23	39	20.6
ocean	11	42	19	11	30	10.1
surfers	7	24	25	10	35	6.9
rain	3	10	23	17	15	3.4
Avg EER	7.6 %	26.2%	33.1 %	16 %	29.7 %	9.7 %

Table 6.2: Average Equal Error Rate comparison of different stages of the proposed method

	S_P	S_{PL}	S_{PG}	S_{PLG}
Avg EER	11.6 %	10%	10.1 %	9.7 %

map that combines pixel saliency map with local and global region saliency map (S_{PLG}). As expected, the EER decreases when the pixel saliency map is used in conjunction with the region saliency maps. Moreover, the best performance of 9.7% is when pixel saliency maps and the region saliency maps are considered together. The improvement in performance of the final saliency map is attributed to the fact that the local and global regional saliency maps re-inforce each other as they address the local and global aspects of the motion of the salient object. Figure 6.9 shows the comparison of binary saliency maps of 'flock' and 'land' sequences. The same threshold is applied to both the saliency maps for subjective comparison. The final saliency map in Figure 6.9(c) which is a combination of pixel and region based saliency maps performs better than the pixel saliency map in Figure 6.9(b).

6.5.4 Limitation

The limitation of the proposed algorithm arises from the underlying assumption of the relation between salient motion and holistic motion modeled by the frame buffer. The method assumes complex, unpredictable motion in the background as non-salient and such motion gets less observability from the output pixels in the state-space framework. However, in the specific case of a camera panning an object moving at slow-medium pace over a long period of time, the motion of background objects due to camera egomotion becomes quite robust and predictable and the proposed

method identifies this motion as salient. For example, in the 'land' sequence from the dataset, the camera follows the landing of an aircraft while the buildings in the static background appear to move slowly in the opposite direction. For the proposed method such a motion in the background is a slow and predictable regular motion and hence, will be salient along with the aircraft in foreground. However the DS method uses contrast of centre and surround windows and identifies the motion of aircraft as salient. Hence the 'land' sequence has high EER for the proposed method and low EER for the DS method.

6.6 Conclusion

We have modeled videos as a linear dynamical system to identify salient motion under complex unpredictable backgrounds. The key idea is to relate the observability of the output of the system to the eigen values of the state transition to the notion of predictable motion in a video. Saliency maps are obtained at two levels - pixel and regions - which are then merged to obtain the final saliency map. The proposed method is simple and efficient in that it requires no background estimation or understanding of motion in the system. The method is computationally efficient and is shown to perform better when compared to most of the state-of-the-art methods.

Chapter 7

Conclusions

In this thesis we have proposed some computational methods for saliency detection in images and videos. For saliency detection in images we have proposed two different methods - one based on distribution of features and the other based on graph models. We have introduced a novel way to evaluate salient motion in videos in a linear dynamical system framework. In this chapter we summarize the contributions from this research work and also look forward to future research issues in the related areas.

7.1 Summary of Contributions

(i) We have made two separate contributions in addressing the problem of image saliency. The first method described in Chapter 3 models the distribution of various features in an image and aims to identify the salient features among them based on their properties in the spatial as well as feature domains. The orientation entropy feature in this chapter is a novel contribution and has been proved quite useful in identifying salient regions. The feature saliency is further carried to the spatial domain to get the saliency maps depicting 'regions' with salient features. One drawback of the proposed method in Chapter 3 is that it results in identifying the most salient regions

in an image and need not always succeed in segmenting a salient object completely. This is attributed to the fact that this method selects only the subset of features that best describes the salient parts of the image and can potentially leave out some other parts of the salient object. In contrast to the method in Chapter 3, the graph based method in Chapter 4 aims to segment out the salient object(s) in a complete semantic sense. The semi-supervised segmentation module after the automatic detection of salient and background nodes achieves this purpose to a great extent. It has to be noted that the features used to build the initial graph model in Chapter 4 is inspired from the salient region detection method proposed in Chapter 3. The graph based method works on image patches considering them as nodes and the final labeling available for the nodes is the label information for different patches. Hence the accuracy of final segmentation information is limited by the patch size. Smaller patch size results in better accuracy but scales up the computational complexity as the method involves a number of eigenvector computations.

(ii) The unsupervised feature selection method discussed in Chapter 5 is primarily motivated by the need to improve the performance of the graph based method discussed in Chapter 4 . However this feature selection method can also be considered as an unsupervised strategy to improve the performance of the saliency detection algorithms in general. The underlying philosophy of this method is that, given an image and a set of features for the image, the algorithm constructs feature graphs using different features and maximizes the mixing rate of weighted linear combination of feature graphs to prioritize relevant features over irrelevant features. The weights of the features can be used to improve any saliency detection algorithm that follows. The method optimizes features based on their graph structure and hence includes the geometry of the image in the process of feature selection. Thus the contribution in Chapter 5 is significant due to following aspects - 1) it is an unsupervised method for

feature selection, 2) it is tailored for salient object detection and 3) it considers the geometry of the image while selecting the relevant features.

(iii) The salient motion detection method discussed in Chapter 6 is a novel attempt to solve the salient motion detection problem using the concepts from system modeling and control theory. We consider the salient motion in a video as those kind of motions that are 'easy to observe' for humans with their robust and predictable characteristics. Interestingly, the control theory concept of 'observability' used in the form of a quantitative measure from the output pixels in a linear dynamical system framework helps in evaluating the saliency of the pixels. The pixel saliency is used to study the dynamics of salient region and match it with the dynamics of other patches in the frame buffer. The region based saliency maps found in this manner help to improve the accuracy of pixel saliency maps. The saliency maps evaluated in Chapter 6 do not completely segment out the salient objects in motion. However the maps provides vital clues on the regions of salient motion which will be very useful in surveillance and video retargeting applications.

7.2 Future Research

Now we discuss some of the potential areas of future research related to saliency in images and videos.

i) The potential of combining supervised or unsupervised learning methods with top-down methods for understanding contextual saliency of object categories is a rather unexplored arena in salient object detection. The interaction of various object categories and the relation between them in an image provide vital clues to detect saliency. For example, majority of the images available in the web with object category cattle will be having most naturally occurring scene categories like open fields around them. After learning the object interactions using topic models like pLSA (Proba-

bilistic Latent Semantic Analysis), the prediction of an event like finding the cattle on a highway will be unlikely and will be regarded as salient. The contextual information can be learned online as and when new data is available. Directed graphical models can be used to stress the significance of the relative order of different categories. For example, detecting a person under the roof of a house and on top of the roof of a house are two separate events in which the latter is more unlikely and demands more attention.

ii) Detection of unlikely events or anomaly in an observed video pattern is an open problem which is related to video saliency. Anomaly detection in crowd moving pattern has extensive applications in surveillance and maintenance of security in sensitive regions like airports. Holistic video modeling approaches like dynamic texture modeling is useful in analyzing the pattern of movement of such large number of objects, as individual recognition of objects is more difficult in such cases. However more work needs to be done for view-invariant or shift-invariant modeling of dynamic textures for semantic identification of events. Segmentation of different dynamic textures (e.g. fire in swaying vegetation) and segmentation of dynamic textures from regular activity scenes are further open issues which will be easier to resolve with a better semantic understanding of dynamic textures in the video sequence.

iii) Context aware summarization of images and videos is another open research problem with respect to detection of saliency in images and videos. When a saliency detection algorithm targets to predict the human eye fixations or detect the most salient object in an image or video, the context aware summarization problem intends to find the subset of the image or video which best describes it so that it could be used for applications like image/video retargeting. The motivation of finding a good summary of the image or video is to minimize the distortions in applications

like content aware image/video resizing or content aware compression of videos and images.

iv) The feature selection problem for salient object detection discussed in Chapter 5 has many open challenges. Chapter 5 proposed to identify the relevant features by first identifying the weights for irrelevant features so as the problem can be formulated in a convex programming framework with a global optimum. The identification of irrelevant features achieves the purpose of a filtering operation of noisy features while a framework for direct identification of relevant features can enhance the existing feature selection problem. The formulation of such a relevant feature identification algorithm into an efficient framework as in the convex problem in Chapter 5 is an open issue. Supervised methods for feature selection using human eye fixation dataset [66] is an unexplored area of research in saliency detection. Online learning methods using eye tracking technologies like tobii eye tracking technology offer large scope in the study of relevance of different features in deciding the saliency in images.

Bibliography

- [1] http://www.svcl.ucsd.edu/projects/background_subtraction.
- [2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1597–1604, 2009.
- [3] D. Aldous and J. A. Fill. Reversible markov chains and random walks on graphs, <http://stat-www.berkeley.edu/users/aldous/RWG/book.html>.
- [4] P. J. Antsaklis and A. N. Michel. *Linear Systems*. McGraw-Hill Higher Education, 1997.
- [5] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3):10, 2007.
- [6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [7] Stephen Boyd. Convex optimization of graph laplacian eigenvalues. *Proceedings International Congress of Mathematicians*, 3:1311–1319, 2006.
- [8] N. D. Bruce and J. K. Tsotsos. Saliency based on information maximization. In *Conference on Neural Information Processing Systems (NIPS)*, pages 155–162, 2005.

- [9] A. Bugeau and P. Perez. Detection and segmentation of moving objects in complex scenes. *Computer Vision and Image Understanding*, 113(4):459–476, April 2009.
- [10] A.B. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I: 846–851, 2005.
- [11] Chi-Tsong Chen. *Linear System Theory and Design*. Oxford University Press, Inc., New York, NY, USA, 1998.
- [12] L. Q. Chen, X. Xie, X. Fan, W. Y. Ma, H. J. Zhang, and H. Q. Zhou. A visual attention model for adapting images on small displays. *Multimedia Syst.*, 9(4):353–364, 2003.
- [13] Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, February 1997.
- [14] Luciano da Fontoura Costa. Visual saliency and attention as random walks on complex networks, arXiv preprint, 2006.
- [15] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.
- [16] L. Ding and A. Yilmaz. Interactive image segmentation using probabilistic hypergraphs. *Pattern Recognition*, 43(5):1863–1873, May 2010.
- [17] G. Doretto, A. Chiuso, Y.N. Wu, and S. Soatto. Dynamic textures. *IJCV*, 51(2):91–109, February 2003.

- [18] O. Duchenne, J.Y. Audibert, R. Keriven, J. Ponce, and F. Segonne. Segmentation by transduction. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [19] A. G. Dugue and A. Oliva. Classification of scene photographs from local orientations features. *Pattern Recognition Letters*, 21(13-14):1135–1140, December 2000.
- [20] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3), 2008.
- [21] A.M. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proc. IEEE*, 90:1151–1163, 2002.
- [22] S. Frintrop, G. Backer, and E. Rome. Goal-directed search with a top-down modulated computational attention system. In *DAGM05*, pages 117–124, 2005.
- [23] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *Conference on Neural Information Processing Systems (NIPS)*, pages 481 – 488, 2004.
- [24] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. In *IEEE Int. Conf. on Computer Vision*, 2007.
- [25] R.C. Gonzalez and R.E. Woods. *Digital Image Processing, Third Edition*. Prentice Hall.
- [26] M. Grant and S. Boyd. Cvx: Matlab software for disciplined convex programming (web page and software). <http://stanford.edu/~boyd/cvx>, June 2009.
- [27] C. M. Grinstead and L. J. Snell. Introduction to probability, American Mathematical Society, 1997.

- [28] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [29] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Conference on Neural Information Processing Systems (NIPS)*, pages 545–552, 2006.
- [30] C. Harris and M. Stephens. A combined corner and edge detector. *Proc. Fourth Alvey Vision Conf*, 15:147–151, 1988.
- [31] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *Conference on Neural Information Processing Systems (NIPS)*, 2005.
- [32] G. Heidemann. Focus-of-attention from local color symmetries. *IEEE Transactions on Pattern Analysis Machine Intelligence (PAMI)*, 26(7):817–830, 2004.
- [33] Y. Hu, D. Rajan, and L. T. Chia. Robust subspace analysis for detecting visual attention regions in images. In *ACM Multimedia*, pages 716–724, 2005.
- [34] Y. Hu, X. Xie, W. Y. Ma, L. T. Chia, and D. Rajan. Salient region detection using weighted feature maps based on the human visual attention model. In *5th Pacific-Rim Conference on Multimedia*, pages 993–1000, 2004.
- [35] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.*, 28:229–289, 1965.
- [36] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I: 631–637, 2005.

- [37] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.
- [38] J.S. Joseph and L.M. Optican. Involuntary attentional shifts due to orientation differences. *Perception and Psychophysics*, 58:651–665, 1996.
- [39] T. Jost, N. Ouerhani, R. Wartburg, R. Müri, and H. Hügli. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107–123, 2005.
- [40] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *International Conference on Computer Vision (ICCV)*, 2009.
- [41] T. Kadir and M. Brady. Saliency, scale and image description. *Int. J. of Computer Vision*, 45(2):83–105, 2001.
- [42] P. Khuwuthyakorn, A. Robles Kelly, and J. Zhou. Object of interest detection by saliency learning. In *European Conference on Computer Vision (ECCV)*, pages II: 636–649, 2010.
- [43] W. Kienzle, F. A. Wichmann, B. Scholkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. In *Neural Information Processing Systems (NIPS)*, 2006.
- [44] C. Lampert, M. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [45] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis Machine Intelligence (PAMI)*, 26:1154–1166, 2004.

- [46] A.G. Leventhal. The neural basis of visual function: Vision and visual dysfunction,. *Boca Raton, Fla.: CRC Press*, 4, 1991.
- [47] T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *Int. J. of Computer Vision*, 11(3):283–318, 1993.
- [48] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Y.Shum. Learning to detect a salient object. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [49] Y. F. Ma and H. J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381, 2003.
- [50] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *PAMI*, 32(1):171–177, January 2010.
- [51] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *International Conference on Computer Vision (ICCV)*, pages 2232–2239, 2009.
- [52] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.
- [53] R.J. Martin. A metric for arma processes. *IEEE Transactions on Signal Processing*, 48(4):1164 – 1170, April 2000.

- [54] O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis Machine Intelligence (PAMI)*, 28(5):802–817, 2006.
- [55] P. Mitra, C. A. Murthy, and S. K Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24:301–312, 2002.
- [56] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages II: 302–309, 2004.
- [57] Antoine Monnet, Anurag Mittal, Nikos Paragios, and Visvanathan Ramesh. Background modeling and subtraction of dynamic scenes. In *International Conference on Computer Vision (ICCV)*, pages 1305–1312, 2003.
- [58] B. S. Morse, D. Thornton, Q. Xia, and J. Uibel. Image based color schemes. In *Int. Conf. on Image Processing*, volume 3, pages III–497 – III–500, 2007.
- [59] J. Norris. Markov chains, Cambridge University Press, Cambridge, 1997.
- [60] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001.
- [61] A. Oliva, A. Torralba, M. S. Castelhana, and J. M. Henderson. Top-down control of visual attention in object detection. In *Int. Conf. on Image Processing*, pages I: 253–256, 2003.
- [62] A. Oliva, A. Torralba, M. S. Castelhana, and J. M. Henderson. Saliency, objects and scenes: Global scene factors in attention and object detection. In *Vision Sciences Society Annual Meeting*, 2004.

- [63] N. Otsu. A threshold selection method from grey-level histograms. *SMC*, 9(1):62–66, January 1979.
- [64] S. J. Park, J. K. Shin, and M. Lee. Biologically inspired saliency map model for bottom-up visual attention. In *Biologically Motivated Computer Vision Second International Workshop*, pages 418 – 426, 2002.
- [65] E. J. Pauwels and G. Frederix. Finding salient regions in images: nonparametric clustering for image segmentation and grouping. *Computer Vision and Image Understanding*, 75(1-2):73–85, 1999.
- [66] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.S. Chua. An eye fixation database for saliency detection in images. In *European Conference on Computer Vision (ECCV)*, pages IV: 30–43, 2010.
- [67] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context free attentional operators: the generalized symmetry transform. *Int. J. of Computer Vision*, 26:119–130, 1995.
- [68] Volker Roth and Tilman Lange. Feature selection in clustering problems. In *In Advances in Neural Information Processing Systems 16*. MIT Press, 2004.
- [69] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Improved seam carving for video retargeting. *ACM Trans. Graph.*, 27(3):1–9, 2008.
- [70] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. F. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI*, pages 771–780, 2006.
- [71] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetsche. Scene analysis with saccadic eye movements: Top-down and bottom-up modeling. *J. Electron. Imaging*, 10(152), 2001.

- [72] Cordelia Schmid and Roger Mohr. Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19:530–535, 1997.
- [73] R. J. Snowden. Visual attention to color: Parvocellular guidance of attentional resources ? *Psychological Science*, 13(2), March 2002.
- [74] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis Machine Intelligence (PAMI)*, 22(8):747–757, 2000.
- [75] F. Stentiford. Attention based auto image cropping. In *The 5th International Conference on Computer Vision Systems, Bielefeld*, 2007.
- [76] Jun Sun, Stephen Boyd, Lin Xiao, and Persi Diaconis. The fastest mixing markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Review*, 48(4):681–699, 2006.
- [77] M. Tarokh. Measures for controllability, observability and fixed modes. *IEEE Transactions on Automatic Control*, 37(8):1268 – 1273, August 1992.
- [78] A. Torralba. Modeling global scene factors in attention. *Journal of Optical Society of America. A, Special Issue on Bayesian and Statistical Approaches to Vision*, 20(7):1407–1418, 2003.
- [79] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.
- [80] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.
- [81] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, 2006.

- [82] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Network*, 19:1395–1407, 2006.
- [83] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010.
- [84] X. J. Wang, W. Y. Ma, and X. Li. Data-driven approach for bridging the cognitive gap in image retrieval. In *2004 IEEE International Conference on Multimedia and Expo,,* pages 2231–2234, 2004.
- [85] J. Weijer, T. Gevers, and A. D. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis Machine Intelligence (PAMI)*, 28(1):150–156, 2006.
- [86] R.P Wildes. A measure on motion salience for surveillance applications. In *International Conference on Image Processing (ICIP)*, pages 183–187, 1998.
- [87] L.E. Wixson. Detecting salient motion by accumulating directionally-consistent flow. *PAMI*, 22(8):774–780, August 2000.
- [88] X.Hou and L.Zhang. Saliency detection: A spectral residual approach. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2007.
- [89] C.M. Kuo Y.H.Kuan and N.C. Yang. Color-based image salient region segmentation using novel region merging strategy. *IEEE transactions on Multimedia*, 10(5):832–845, 2008.
- [90] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006.

- [91] L. Yuan and H. Y. Shum. Synthesizing dynamic texture with closed-loop linear dynamic system. In *European Conference on Computer Vision (ECCV)*, pages 603–616. Springer, 2004.
- [92] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *ACM Multimedia*, pages 815–824, 2006.
- [93] B. Zhao, J. Kwok, F. Wang, and C. Zhang. Unsupervised maximum margin feature selection with manifold regularization. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.
- [94] Y. Zhiwen and H.S. Wong. A rule based technique for extraction of visual attention regions based on real-time clustering. *IEEE Transactions on Multimedia*, 9(4):766–784, 2007.
- [95] Jing Zhong and Stan Sclaroff. Segmenting foreground objects from a dynamic textured background via a robust kalman filter. In *International Conference on Computer Vision (ICCV)*, pages 44–50, 2003.
- [96] J. Zhu, Y. Lao, and Y.F. Zheng. Object tracking in structured environments for video surveillance applications. *IEEE Transactions on circuits and systems for video technology*, 20(2):223–235, February 2010.
- [97] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning (ICML)*, pages 912–919, 2003.
- [98] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *International Conference on Pattern Recognition (ICPR)*, pages II: 28–31, 2004.

Appendix A

A.1 Observability Measure from Outputs for Discrete-Time, Linear Time-Invariant Systems

Consider the discrete-time LTI system given in Equation 6.3. The response of the linear discrete-time system at time k is given as [4]

$$y(k) = CA^k x(0) + \sum_{j=0}^{k-1} CA^{k-(j+1)} Bu(j) + Du(k) \quad (\text{A.1})$$

where $k > 0$ and A, B, C, D are system parameters explained in Equation 6.3 and $y(k)$ is the system output at time k .

It can be seen from Equation A.1, the relation between eigenvalues of matrix A and output $y(k)$ is governed by the factors CA^r , where r is some integer value between 1 and k . The eigen decomposition of the square matrix A is given as $Q \Lambda Q^{-1}$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$. The matrix A and A^r can be represented as

$$\begin{aligned} A &= \sum_{i=1}^{i=n} \lambda_i \cdot q_i * f_i^T, \\ A^r &= \sum_{i=1}^{i=n} (\lambda_i)^r \cdot q_i * f_i^T, \end{aligned} \quad (\text{A.2})$$

where λ_i is the i^{th} eigen value, q_i and f_i the i^{th} columns of matrices Q and Q^{-1} respectively. Hence $CA^r = C \sum_{i=1}^{i=n} (\lambda_i)^r . q_i * f_i^T = \sum_{i=1}^{i=n} (\lambda_i)^r . (Cq_i) * f_i^T$. Thus the vector (Cq_i) becomes the crucial factor in deciding the presence of (λ_i) in the multi-dimensional output equation. If a specific row, say p , of C is orthogonal or near orthogonal to q_i , the estimate that can be made about λ_i from the p^{th} output in the vector $y(k)$ will be poor. Hence the angle between the rows of C and the vector e_i is chosen as the measure of the observability from output pixels to the eigen value λ_i .

List of Publications

Conference Papers

1. V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs to model saliency in images. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 1698 - 1705, held in Miami, F.L. ,USA on June 2009
2. V. Gopalakrishnan, Y. Hu, and D. Rajan. Unsupervised feature selection for salient object detection , In Asian Conference on Computer Vision (ACCV), held on November 2010 in Queenstown, NewZealand
3. V. Gopalakrishnan, Y. Hu, and D. Rajan. Sustained Observability for Salient Motion Detection, In Asian Conference on Computer Vision (ACCV), held on November 2010 in Queenstown, NewZealand

Journal Papers

1. V. Gopalakrishnan, Y. Hu, and D. Rajan. Salient region detection by modeling distributions of color and orientation. IEEE Transactions on Multimedia, 11(5):892- 905, Aug. 2009
2. V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs for salient object detection in images. IEEE Transactions on Image Processing 19(12) 3232 – 3242 Dec. 2010
3. V. Gopalakrishnan, D. Rajan and Y. Hu. A Linear Dynamical System Framework for Salient Motion Detection. IEEE Transactions on Circuits and Systems for Video Technology (Submitted).