Salient Object Segmentation in Images

A THESIS

submitted by

SUDESHNA ROY

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

May 2015

Dedicated To
My parents, my siblings and my teachers

THESIS CERTIFICATE

This is to certify that the thesis titled **Salient Object Segmentation in Images**, submitted by **Sudeshna Roy**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Science**, is a bona fide record of the research work done by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Sukhendu Das
Research Guide
Professor
Dept. of Computer Science and Engineering
IIT-Madras, 600 036

Place: Chennai

Date:

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep and sincere gratitude to my research supervisor Prof. Sukhendu Das whose patient guidance, constant encouragement and excellent advice throughout the course of my MS program has made this thesis possible. His expertise in the area and enthusiasm in the research have been of great value for me. I am also thankful to him for placing the laboratory facilities at my disposal.

I also take this opportunity to express my sincere thanks to my General Test Committee members Prof. P. Srinivasa Kumar, Dr. Madhu Mutyam and Dr. Sunetra Sarkar for their interest, encouragement, valuable suggestions and thoughtful reviews.

I would like to express my gratitude to former HoD Prof. C. Siva Ram Murthy and present HoD Prof. P. Sreenivasa Kumar for providing the best possible facilities to carry out the research work. I am also grateful to Dr. N. S. Narayanaswamy, Dr. Sutanu Chakraborti, Prof. C. Chandra Sekhar, Dr. B. Ravindran and Dr. Krishna Jagannathan (Dept. of Electrical Engineering) for their role in building up the foundation in subjects of Algorithms, Machine Learning, Kernel Methods, Probabilistic Graphical Models, Probability Theory and Optimization which are integral part of my research.

My sincere thanks to Computer Science office and laboratory staffs, Mr. Ravichandran. S, Mr. P. Govindaraju, Mrs. R. Prema, Mr. Balu, Mr. Mani for their valuable cooperation and assistance.

My special thanks to my lab-mates Chiranjoy Chattopadhyay, Suranjana Samanta, Nitin Gupta, Samik Banerjee, Prateek Srivastava, Amit Kumar Maurya and Ankit Srivastava, Geethu M Jacob for being tolerant and cooperative. My cordial thanks to Chiranjoy Chattopadhyay for reviewing my papers and giving important feedbacks and to Suranjana Samanta for technical support.

A special note of thanks to my friend Sarthak Parui for helping me through out my MS career and reviewing my paper and giving valuable feedback on my research.

Finally, I would like to thank my parents, sister and little brother for being a source of encouragement and strength all throughout.

ABSTRACT

KEYWORDS: Saliency; Generic Object Segmentation; Feature Rarity; Background Prior; Objectness.

Saliency is an important property of the human visual perception. Most often the focus of interest of the human eye gets attracted to a region or salient object appearing distinctly in the foreground of a scene. This ability to automatically segment the objects of interest in an image is also useful for many computer vision tasks such as, shape-based formulations, object recognition, indexing and retrieval, image-retargeting, object tracking in videos and so on. This pre-processing step helps to reduce the search space for many other (intermediate or high-level) processes to follow, such as feature-extraction, matching, enhancement, compression etc., there by reducing the computational time.

There are two primary approaches for such algorithms. First, there are methods that find any object of interest based on visual stimuli, without any prior knowledge about its category. These are known as bottom-up methods. Second class of methods (top-down) find category specific objects which are known and learned a priori. The former methods attempts to find regions or objects in an image that are prominent and vividly stand out from the rest of the image. This is a subjective perceptual quality (equivalent to avoiding information overload) in the human visual system (HVS) that has the ability to select regions with important visual information from its bottom-up stimuli. This quality is known as visual saliency and the objects which prominently stand out are considered salient. We concentrate on bottom-up methods of finding salient objects in a scene.

We present two different approaches for bottom-up salient object segmentation from images. The first one relies on basic perceptual cues alone. Whereas, the second one uses generic objectness features along with saliency criteria to segment salient objects from complex natural scenes. In both of these approaches, we first segment the image into small homogeneous patches, known as superpixels. In

the first method, we utilize the low-level perceptual cues such as, rarity of feature, center-bias, boundary prior, and mathematically model them to generate a probability map depicting saliency in an unsupervised framework. Rarity of feature is computed by exploiting graph-based spectral feature rarity and its spatial compactness. Graph-based rarity is computed by obtaining the uniqueness of the spectral features, using the Laplacian of the graph over superpixels. Spatial compactness is obtained using distribution of similar colors over the image. Boundary prior is obtained by statistically modeling (in color space) the set of superpixels near to the boundary of an image. Our method produces a full resolution saliency map, where each pixel is assigned a probability value of being salient.

The second formulation addresses the complex issues in many natural images, where the object cannot be segmented using the low-level perceptual cues alone. Natural images exhibit spatial interactions, as: (i) neighboring superpixels are likely to belong to the same object unless delineated by prominent image edges, (ii) spatially bounded superpixels together generally represent an object part. These dependencies can be captured by a graphical model based approach which, in general, helps in good spatial propagation of low-level saliency cues and different prior information. Hence, to solve this problem of generic object segmentation, we construct a conditional random field (CRF) over superpixels. Since, saliency alone is not sufficient, we exploit saliency in conjunction with different objectness criteria and appearance features, to formulate a multi-criteria energy function. In our algorithm, the edge-cost produces a sub-modular CRF. Thus, we perform an exact inference using graph cut. The CRF parameters are learnt by formulating a max-margin optimization. As the energy cost is a linear function of its parameters, we efficiently estimate the CRF parameters. Hence, both learning and inference are done efficiently so as to perform well on complex situations in natural images.

The performance of both the proposed methods are shown using visual illustrations, while the Precision-Recall-Fmeasure and intersection-over-union metric is used to quantitatively compare the same with recent state-of-the-art techniques using challenging benchmark real-world datasets (e.g., PASCAL 2012, MSRA-B).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CRF	Conditional Random Field
CBIR	Content Based Image Retrieval
EM	Expectation-Maximization
FRBP	Feature Rarity and Background Prior
GMM	Gaussian Mixture Model
IoU	Intersection over Union
MLE	Maximum Likelihood Estimation
MSRA-B	MSRA Salient Object Database, Image set B
PARAM	background Prior And RArity for saliency Modeling
PASCAL	Pattern Analysis, Computational Modeling and Statistical Learning
QP	Quadratic Program
SVM	Support Vector Machine
SED	Segmentation Evaluation Database
VOC	Visual Object Challenge

NOTATIONS

N	Number of superpixels
sp_i	ith Superpixel
c_i	Mean color of superpixel i
p_i	Mean position of superpixel i
G	Graph
V	Vertex set of a graph
E	Edge set of a graph
W	Weight matrix or affinity martix
D	Diagonal matrix
L	Laplacian of a graph
x_i	Spectral feature of superpixel i
r_i	Rarity of superpixel i
v_i	Variance of color of the superpixel i
μ_i	Mean position of the color of superpixel i
D_M	Mahalanobis Distance
π_j	Mixture coefficient of jth Gaussian mode
B_i	Background prior of superpixel i
S_i	Saliency of superpixel i
$\mathcal{P}_e$	Edge Probability value
b_i	Boundedness value of superpixel i
$density_i$	Density of image edges in superpixel i
ed_i	Edge density value of superpixel i
$\mathbf{x}$	Feature vector
$\mathbf{y}$	Labels vector
$\hat{\mathbf{y}}$	Ground truth labels
$\mathbf{w}$	Parameters
$\phi^{(1)}$	Unary potential
$\phi^{(2)}$	Pairwise potential
M	Number of training instances
ξ	Slack variable
$\mathcal{L}$	Loss function

CHAPTER 1

INTRODUCTION

"Vision is the process of discovering from images what is present in the world, and where it is." - David Marr

Modern day life has overwhelming amount visual data and information available and created every minute. This growth in image data has led to new challenges of processing them fast and extracting correct information, so as to facilitate different tasks from image search to image compression and transmission over network. One specific problem of computer vision algorithms used for extracting information from images, is to find objects of interest in an image. Human visual system has an immense capability to extract important information from a scene. This ability enables humans to focus their limited perceptual and cognitive resources on the most pertinent subset of the available visual data, facilitating learning and survival in everyday life. This amazing ability is known as *visual saliency* (Itti *et al.* (1998)). Hence for a computer vision system, it is important to detect saliency so that the resources can be utilized properly to process important information. Applications range from object detection ro Content Based Image Retrival (CBIR), face or human re-identification and video tracking.

1.1 Motivation: What is Saliency?

Saliency is the ability or quality of a region in an image to standout (or be prominent) from the rest of the scene and grab our attention. Saliency can be either stimulus driven or task specific. The former one is known as bottom-up saliency while the later specifies top-down saliency and leads to visual search. Bottom-up saliency can be interpreted as a filter which allows only important visual information to grab the attention for further processing. In our work, we concentrate

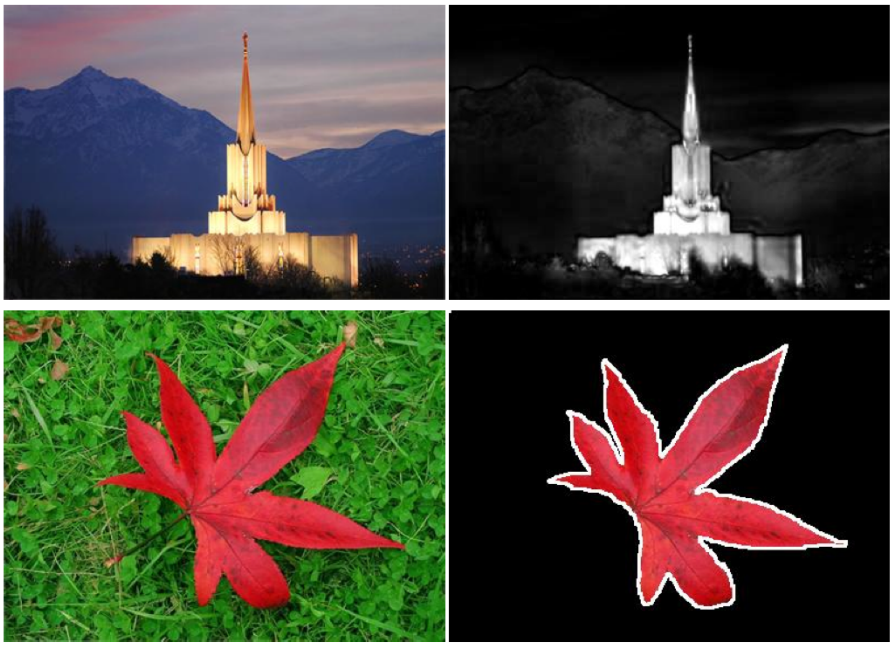


Figure 1.1: The top row shows an example of saliency map generated from the image (left) and the bottom row depicts an ideal segmentation of the object in the image (left).

on bottom-up salient object detection. Saliency is a particularly useful concept when considering bottom-up feature extraction, since one must find what is significant in an image from the scene data alone. In such circumstances, the role of context becomes extremely important. That is to say that saliency can be described as a relative measure of importance. Hence, the bottom-up saliency can be interpreted as its state or quality of standing out (relative to other stimuli) in a scene. As a result, a salient stimulus will often pop-out to the observer, such as a red dot in a field of green dots, an oblique bar among a set of vertical bars, a flickering message indicator of an answering machine, or a fast moving object in a scene with mostly static or slow moving objects. An important direct effect of the saliency mechanism is that it helps the visual perceptual system to quickly filter and organize useful visual information, necessary for object recognition and/or scene understanding.

We propose two different methods for the task. First is based on low-level perceptual features. Second combines the low-level saliency with generic object specific cues in a graphical model based approach. Both the methods are thoroughly evaluated against state-of-the-art methods (Cheng *et al.* (2011); Perazzi *et al.* (2012); Carreira and Sminchisescu (2010)) on challenging benchmark datasets and

found to produce superior results.

1.2 Objective and Scope

The objective of the thesis is to device an efficient salient object detection method that can facilitate as a pre-processing step for many of the previously mentioned tasks. Further, the method must be unsupervised so that it can detect any generic object. Moreover, it has to be computationally efficient to ensure fast processing, considering the huge amount of available data.

As already discussed bottom-up saliency can be characterized by the ability to pop-out in a scene. Consequently, most saliency detection methods in literature (Achanta *et al.* (2009); Goferman *et al.* (2010); Cheng *et al.* (2011); Perazzi *et al.* (2012); Li *et al.* (2013); Yang *et al.* (2013); Jiang *et al.* (2013)) propose a model by exploiting *rarity of features*. But, as also mentioned by Wei *et al.* (2012); Zhu *et al.* (2014) only feature rarity based approach is not enough to extract salient regions from natural images of varying scene conditions. We identify this shortcoming in the rarity of feature based approach and exploit boundary prior as a cue to implement our first method of saliency detection.

Further, class independent object segmentation has recently gained importance in the Computer Vision community (Carreira and Sminchisescu (2010); Endres and Hoiem (2010)). In this context, Alexe *et al.* (2012) had addressed the problem of detecting generic objects and defined objectness properties as likelihood of a region belonging to an object. But it gives bounding boxes rather than pixel level segmentation output. However their precision is very low, as a lot of background regions are proposed as objects. Optimization is based on intersection-over-union criteria (Endres and Hoiem (2010)) to rank the maps, but the results show that the top-most map generally contains almost half of the image. Hence, we propose an algorithm to generate a single map segmenting only the objects of interest, using saliency and objectness on a conditional random field (CRF). The focus of this thesis is on one of the visual capabilities of human - finding objects of interest in images.

Figure 1.2: Different challenges in saliency detection, illustrated using samples from saliency dataset, MSRA B (Achanta *et al.* (2009)).

1.3 Problem Definition and Challenges

The problem we address in the thesis can be defined in short, as:
Given a natural scene, detect one or more regions of interest (ROI) which contain the salient objects in a scene. The method must be unsupervised with no training sample for classes of objects available. Parameters of any optimization function may be learned using a part of another dataset, or verification subset of the same. Although the problem is similar to unsupervised foreground segmentation, it differs in the context of features which is mostly inspired by biological motivation. Some examples of finding objects of interest are presented in Figure 1.1.

This is a challenging task, because objects of interest are detected without any prior knowledge about them purely based on unsupervised stimulus driven perceptual cues. Single features such as, color, brightness, depth alone are not helpful to solve the problem. It suffers from all the challenges that any computer vision problem faces and are illustrated in Figure 1.2. Further, when finding saliency for challenging datasets like PASCAL to facilitate later process of object detection or recognition, segmenting the object of interest becomes even harder. A through study of different method and samples from the dataset reveals the following challenges, apart from the factors already depicted in Figure 1.2:

1. Only a small part of an object is present on the boundary of an image;
2. Objects with large holes, such as cycle wheel;
3. Repeated Distractors in background or foreground.

These are illustrated with respective samples in Figure 1.3.

1.3.1 Assumptions

1. The images are indoor or outdoor natural scenes, captured using optical camera (not X-ray or infrared etc.).
2. Object of interest is not just visible in few pixels on the boundary of the image.
3. Object of interest is generally not hidden behind a large distractor, for example, the image in the third row of Figure 1.3.
4. Not much of both color and texture overlap between object and scene background.

Figure 1.3: Different challenges in saliency detection illustrated with images (left) and respective ground truths (right), from PASCAL dataset (Everingham *et al.* (2012)).

1.4 Contribution of the Thesis

The central contribution of the thesis is pixel accurate localization of the object of interest. The saliency map provided by our proposed methods assign each pixel a saliency value in the range of 0 to 1, depicting their probability of being salient. Hence, it can be easily segmented by simple thresholding mechanism, to obtain the important or salient object. In the work described here, saliency is defined firstly in terms of spatial rarity of image feature, mainly color. Secondly, objectness is used in a graphical model for salient object segmentation. This can change the conventional way of extracting features from the whole image or searching objects in huge 4-dimensional (position, scale and aspect ratio) sliding window search space. It would help simulate the same logistics as human vision and improve both speed and accuracy of the computer vision tasks. Moreover, since we produce a probability value of each pixel being salient, the saliency map can also be utilized for identifying most salient regions for different tasks, for example placing an advertisement in a video. In the following sub-sections, we describe the methods proposed in the thesis in brief.

1.4.1 Saliency Detection based on Low-level Perceptual cues

In the first formulation, we formulate three saliency criteria, namely: (i) graph-based spectral rarity, (ii) spatial compactness and (iii) background prior. Then, a weighted combination of these three measures of saliency, produce a saliency map. A saliency map is represented as a gray scale image, where each pixel is assigned its probability of being salient.

The first two terms named above are based on rarity of feature and the third term correspond to boundary prior. The idea of boundary prior is that the boundary of an image mostly contains background image elements or superpixels and background superpixels are spatially connected among themselves, but not with foreground ones.

Graph-based spectral rarity assigns saliency based on rarity or uniqueness of a superpixel. This measure utilizes the spectral features (as defined by Ng *et al.*

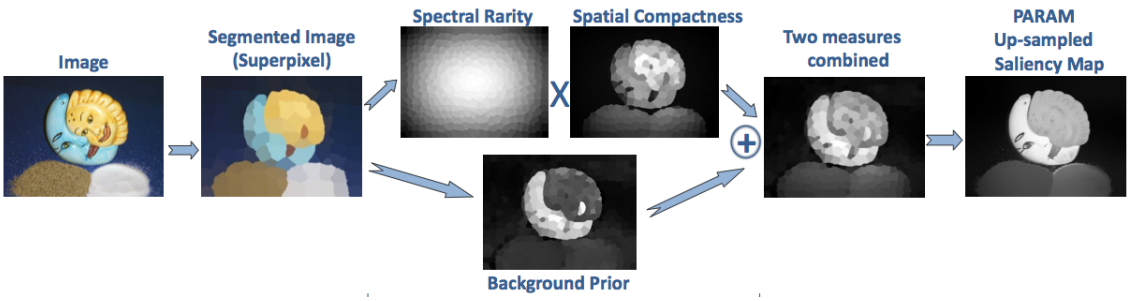


Figure 1.4: Illustration of the sequence of stages of our proposed algorithm (PARAM) for saliency estimation, with an example from MSRA-B Dataset (Achanta *et al.* (2009)).

(2001)) using Laplacian of the superpixel graph.

On the other hand, **spatial compactness** takes into account that a color belonging to a salient object would be grouped at a spatial location and thus the spatial variance of the color would be low. Whereas, background colors are generally distributed over the whole image and score low on spatial compactness.

Our formulation models **background prior** using a Gaussian Mixture Model (GMM). All the superpixels touching the boundary of an image are modeled by GMM in Lab color space. Saliency of a superpixel is measured as the sum of the distances from the GMM modes weighted by the particular mixture coefficient. Since, most of the boundary superpixel would be background, big GMM modes with high value of mixture coefficient belongs to background colors and thus the mentioned distance gives a good measure of saliency.

A non-linear weighted combination of these three different cues are used to compute the final saliency map. Also, binary segmentation maps are generated for quantitative evaluation of performance using an adaptive threshold. An illustration of the complete flow chart of this proposed saliency detection method, which is named as *PARAM* (background Prior And RArity for saliency Modeling), is depicted in Figure 1.4.

1.4.2 Salient Object Segmentation in Natural Images

Next we propose a Salient Object Segmentation method that captures the same visual processing hierarchy as in the human visual system. Our goal is to localize objects independent of its category by formulating an unsupervised algorithm.

Recent saliency detection methods show high performance in saliency datasets, but they fail to perform well when tested using natural image datasets like PASCAL (Everingham *et al.* (2012)). There are two reasons behind this. First, these methods use only low-level perceptual cues such as, center surround operations (Itti *et al.* (1998)), local and global contrast (Goferman *et al.* (2010); Cheng *et al.* (2011)), uniqueness and color distribution (Perazzi *et al.* (2012)) and boundary prior (Yang *et al.* (2013); Wei *et al.* (2012)). Second, there is typically a huge dataset bias which ensures the presence of only a single object at the center of an image. Moreover, in saliency datasets the objects are in high contrast with respect to the background. Hence, this class of methods do not scale up for more natural images such as in the case of PASCAL segmentation dataset (Everingham *et al.* (2012)).

Here, we exploit the saliency (PARAM) described in the previous section along with objectness cues. We formulate two simple but effective objectness factors: geometric constraint and distribution of edges in the image. These two features are respectively modeled as *boundedness* and *edge-density*. To compute these factors we exploit the edge map produced by Dollár and Zitnick (2013). They take a structured learning based prediction on random forest to produce a high-quality edge probability map. Since they do a direct inferencing, the method is computationally efficient. Boundedness captures the extent to which a superpixel is bounded by strong edges on all four directions. Whereas, edge-density computes how much cluttered or smooth a particular superpixel is. As described by Alexe *et al.* (2012), very smooth or highly cluttered regions generally belong to the background.

Next, a graphical model based approach is used for proper spatial propagation of these priors. The image cue priors form the unary term and we formulate a submodular edge cost or pairwise term to specify the CRF (Lafferty *et al.* (2001)). Hence, an exact inference is done efficiently using graph-cut. We employ a margin rescaled algorithm, as explained by Szummer *et al.* (2008), to learn the CRF parameters. It is a max-margin structured learning based approach. The benefit of the formulation is that, it takes into account how far a predicted label is from its ground truth, and the margin is adapted based on how much competing labelings differ from ground truth. Results of the proposed method shows superior performance on PASCAL segmentation dataset (Everingham *et al.* (2012)) when

compared against many recent methods.

In short, the contributions can be described as:

1. Two saliency detection algorithms, which significantly outperform (both quantitatively and qualitatively) several existing algorithms, have been proposed.
2. A unified criteria of saliency, combining boundary prior, rarity and objectness have been proposed.
3. A structured SVM based parameter learning of our graphical model (CRF) based approach, using high-order loss function.
4. Exhaustive experimentation on two real world saliency and object segmentation datasets have been performed to validate the performance of the proposed methods.
5. Publicly available C++ code of the algorithms (Roy and Das (2014)).

1.5 Thesis Overview

We address the problem of saliency detection. We propose two different methods to solve the problem robustly even in challenging natural scenes. Following subsections briefly describes the chapters in the rest of the thesis.

1.5.1 Chapter 2: Literature Survey

In literature there has been many approaches taken by different authors starting from spatial uniqueness to frequency domain analysis. Feature-rarity and boundary prior has proved to be successful cues in saliency detection. Graphical model based approaches show good propagation of low-level priors and results in better performance.

1.5.2 Chapter 3: Saliency Using Low-Level Perceptual Cues

A suitable combination of novel measure of saliency detection is proposed using rarity of features and background prior. Rarity of feature is captured by measuring spectral feature based rarity and spatial compactness of color. background prior,

on the other hand, is modeled using Gaussian mixture model (GMM) of boundary image elements. these two cues have proved to be complementary and equally important, competing state-of-the-art methods.

1.5.3 Chapter 4: Salient Object Segmentation In Natural Images

Object level saliency in natural images is detected to produce a category independent object segmentation. Saliency along with objectness cues are together used on a graphical model to represent a conditional random field (CRF). Graph cut based exact inference produces the segmentation, as the energy is modeled as submodular. CRF parameters are learned using structured max-margin method. This shows better performance than many saliency and category independent segmentation algorithms.

1.5.4 Chapter 5: Conclusion

In this chapter, we present a summary of the work done and contributions. We also discuss the possibilities for improvement and the directions for future work.

CHAPTER 2

LITERATURE SURVEY

The term *saliency* was first proposed by Tsotsos *et al.* (1995) in the context of visual attention. Since then researchers have shown a great interest towards *pre-attentive* or bottom-up saliency detection. Early methods have mostly concentrated on human eye fixation prediction and they have introduced the basic principles of saliency detection. Then, the important problem that has been addressed in literature is salient object detection and segmentation. Since, bottom up saliency is stimulus driven and does not look for any particular object in the scene, it can be used for unsupervised segmentation of all the prominent objects in an image. This leads to a solution of the problem of generic object segmentation. Additionally, literatures in saliency and object segmentation show that graphical model based techniques give efficient modeling and promising result in this field. Hence, we first discuss some relevant state-of-the-art bottom-up saliency detection techniques, followed by category independent or generic object segmentation methods.

2.1 Bottom-up Saliency Models

Bottom up saliency models are mostly inspired by neurophysiology, which adapt the concepts of feature integration theory (FIT) (Treisman and Gelade (1980)) and visual attention (Koch and Ullman (1987)). The very first model which is proposed by Itti *et al.* (1998), henceforth referred as IT in the following chapters, uses three features, namely color, intensity and orientation, similar to the *simple cells* in primary visual cortex. Center-surround differences over these feature channels generate feature maps that are then normalized across scales and linearly combined to give the saliency. Most computational models are based on either spatial or spectral processing. Spatial models use different local or global features, like color, intensity, spatial distance, or a combination. Spectral models

use a spectral domain analysis of the image and inherently use global features. Again, all different saliency methods have mainly two approaches- finding a fixation map or generating a saliency map. Fixation maps (Itti *et al.* (1998); Judd *et al.* (2007)) try to capture the human eye gaze behaviors and eye fixation points. While they are suitable for many different tasks, e.g., finding fixation scan paths, human gaze pattern analysis, advertise placement in a video, they are not applicable to the problem of salient object segmentation in the field of Computer Vision and Pattern Recognition. The other set of methods (Harel *et al.* (2006); Achanta *et al.* (2009); Goferman *et al.* (2010); Cheng *et al.* (2011); Perazzi *et al.* (2012); Li *et al.* (2013); Yang *et al.* (2013)) produce saliency maps where each image pixel is assigned a saliency strength or probability value. A crisp segmentation can be obtained from these maps by simple thresholding. Since, these methods are pertinent to the problem that we try to model, we mostly discuss about them.

2.1.1 Contrast: The Most Important Cue

Most of the models use *contrast* as an important cue. These approaches work well for images which have a simple background and high contrast between background and foreground image elements. Goferman *et al.* (2010) model *context aware* (CA) saliency using both local low-level features and global considerations, as well as visual organization rules and high level features. They have taken overlapping patches at different scales and modeled saliency as distance in color, inversely weighted by distance in position among the patches. The distance between any two patches p_i and p_j is formulated as:

$$d(p_i, p_j) = \frac{d_{color}(p_i, p_j)}{1 + c.d_{position(p_i,p_j)}} \tag{2.1}$$

where c is a constant. Then, saliency of ith pixel at scale r is proposed as,

$$S_i^r = 1 - exp\{-\frac{1}{K}\sum_{k=1}^{K} d(p_i^r, q_k^r)\} \tag{2.2}$$

Goferman *et al.* (2010) take the K most similar patches according to the distance computation in equation (2.1). Then, the final saliency map is given by averaging the saliency value over scales. As a result of such pixel-wise distance based

modeling, edges of the salient regions are highlighted more, rather than the entire salient object.

Cheng *et al.* (2011) have proposed a *region-wise contrast* (RC) based method to compute saliency and use Grabcut algorithm (Rother *et al.* (2004)) to give a refined binary segmentation map based on their saliency. They define the saliency ($S(r_k)$) of kth region r_k as:

$$S(r_k) = \sum_{r_k \neq r_i} w(r_i) D_r(r_k, r_i) \tag{2.3}$$

where $D_r(.,.)$ is the color distance between the two regions and $w(r_i)$ gives higher weightage to bigger regions. The image regions are obtained by segmenting the image using a graph based segmentation method (Felzenszwalb and Huttenlocher (2004)). Being a global contrast based method, it works well for large-scale salient regions.

Since, pixel-wise computation over images is computationally costly, most of the recent methods downsample the image to *superpixels* (Achanta *et al.* (2010)). Superpixel approaches break the image into homogeneous segments and each superpixel can be used as an image element instead of a pixel. Perazzi *et al.* (2012) find uniqueness and distribution of color over superpixels as measures of saliency and name the method as *saliency filter* (SF). They upsample (Dolson *et al.* (2010)) the superpixel-level map to obtain a pixel accurate saliency map. It gives a good precision for focused, large salient regions but fails for natural images.

The *graph based* visual saliency (GB) model proposed by Harel *et al.* (2006) fails to give a smooth object boundary, but works quite well for multiple salient regions of different sizes. In this method an activation map is generated by subtracting feature vectors at different scales (features, similar to Itti *et al.* (1998) are extracted from scale-space pyramid) to give saliency masses. Then, they use a random walker on graph based technique to find the salient regions. But, it gives a blurred map with less precision. Yang *et al.* (2013) and Jiang and Davis (2013) use contrast prior in an optimization framework. Jiang and Davis (2013) formulate the problem as a *facility location* problem which they show to be submodular, and thus efficiently solve it using a greedy approach to detect saliency.

2.1.2 Center Prior

Center prior or *center bias*, implies that the center of an image always attracts more attention (Tseng *et al.* (2009)) and regions at the center become more salient, rather than that in the surroundings. Usually, center prior is realized as a Gaussian map (Achanta *et al.* (2009)). It is either directly combined with other cues as weights (Hou and Zhang (2007); Goferman *et al.* (2010)), or learning-based methods e.g., as proposed by Jiang *et al.* (2013) use it as a feature in their learning framework. Goferman *et al.* (2010) use the center-bias factor as "immediate context" and they redefine their saliency value S_i of a pixel i to $\hat{S}_i$ as:

$$\hat{S}_i = S_i(1 - d_{foci}(i)) \tag{2.4}$$

where, $d_{foci}(i)$ denotes the spatial Euclidean distance of pixel i from the closet focus of attention. The focus of attentions are found by thresholding the initial saliency map ($S_i > 0.8$). The uniqueness term in the method by Perazzi *et al.* (2012) also accounts for center-bias. Jiang and Davis (2013) use center prior as one of the high level priors in their optimization framework. They model the centre prior ($p_l(x)$) as distance (d) of a pixel (x) to the image center $\hat{c}$, as:

$$p_l(x) = e^{(-d^2(x,\hat{c})/\sigma_2)} \tag{2.5}$$

where, $\frac{1}{\sigma_2}$ is computed as expectation over all the pairwise distances.

Table 2.1 note down the different important methods and their features and criterion in the field of saliency detection.

2.1.3 Frequency Domain Analysis

Some methods model the problem of saliency detection using spectral features and perform a frequency domain analysis. Hou and Zhang (2007) represent the log spectrum and Gaussian smoothed inverse Fourier transformed *spectral residual* (SR) component for saliency. They use only the phase information and thus works better for small salient regions in an uncluttered background (Li *et al.* (2013)).

Published Method	Acronym	Reference	Features used	Saliency Criteria	Dataset used
A Model of Saliency Based Visual Attention for Rapid Scene Analysis	IT	Itti *et al.* (1998)	Color(RGB), Intensity, Orientation	A neural network based models combines features linearly at multi-scale	Own attention based natural image set
Saliency Detection: A Spectral Residual Approach	SR	Hou and Zhang (2007)	Log spectra of an image	Inverse Fourier Transform of spectral residual of an image. Spectral residual is given log spectra - amplitude spectra	Own attention based natural image set
Context Aware Saliency Detection	CA	Goferman *et al.* (2010)	Color (Lab) and spatial location	local contrast in color, visual organization rule	MSRA-B (1000 Images)
Frequency Tuned Saliency	FT	Achanta *et al.* (2009)	Color (Lab) Contrast	Distance from mean color after eliminating high frequency intensities	MSRA-B (1000 Images)
Graph Based Visual saliency	GB	Harel *et al.* (2006)	same as IT, Itti *et al.* (1998)	Saliency is found using the equilibrium distribution on a MRF graph of a random walker	Human eye-fixation dataset of 108 images
Global Contrast based Salient Region Detection	RC	Cheng *et al.* (2011)	Color (Lab) and spatial location	Region contrast by sparse histogram comparison	MSRA-B (1000 images)
Saliency Filters	SF	Perazzi *et al.* (2012)	Color(Lab) and spatial location	Uniqueness and spatial distribution of image superpixels	MSRA-B (1000 images)

Table 2.1: A summary of key features in prominent saliency methods from the literature.

Achanta *et al.* (2009) in their *frequency tuned* (FT) model, first omit the very high frequency components as those correspond to background texture or noise artifacts and then compute saliency as the distance from mean color in Lab color space. The saliency value of a pixel at (x, y), is formulated as:

$$S(x, y) = ||\mathbf{I}_{\mu} - \mathbf{I}_{whc}(x, y)|| \tag{2.6}$$

where $\mathbf{I}_{\mu}$ is the mean feature vector (color) and $\mathbf{I}_{whc}$ correspond to the image pixel vector at (x, y) after eliminating the high frequency components. Due to its simplicity the approach is very fast.

A more advanced model (Li *et al.* (2013)) uses *hypercomplex Fourier transform* (HFT, $\mathcal{F}_{\mathcal{H}}[u, v]$ when written in polar form) over different features like Itti *et al.* (1998) and does a spectrum scale space analysis. They create a spectrum scale space $\Lambda = \{\Lambda_k\}$ by smoothing the amplitude spectrum $\mathcal{A}(u, v)$ with Gaussian kernel of different scales. Given an amplitude spectrum Λ_k and the original phase ($\mathcal{P}(u, v)$) and Eigen-axis ($\chi(u, v)$) spectra, the saliency map at scale k is computed as:

$$\mathcal{S}_k = g * ||{\mathcal{F}_{\mathcal{H}}}^{-1}\{\Lambda_k(u, v)e^{\chi\mathcal{P}(u,v)}\}||^2 \tag{2.7}$$

where g is the Gaussian kernel at fixed scale k. Optimal scale is detected by minimizing an entropy, with saliency as probability maps. It gives good results for images with different sizes of salient regions with varying background, but fails to give accurate results often when a single large object is present.

Hou *et al.* (2012) prove that Inverse Discrete Cosine Transform (IDCT) of the sign of DCT of the original image, concentrates the image energy at the location of spatially sparse foreground. The saliency map for each color channel x^i is formed as:

$$\mathbf{s} = g * \sum_i (\bar{x}^i \; o \; \bar{x}^i) \tag{2.8}$$

where, $\bar{x}$ denotes the *image signature* (IS) and defined by the authors as IDCT[sign(DCT(x))]. The operator (. o .) denotes Hadamard product. Simple sum across the color channel gives their final saliency map. This method holds good for only small and sparse salient regions.

2.1.4 Boundary Prior and Connectivity Prior

Another important cue that has come into prominence is *boundary prior* or sometimes termed as *background prior*. This concept again, comes from cognitive science literature (Tatler (2007)), which shows human eye mostly focuses at the center of an image and boundary regions are predominantly occupied by background pixels. Again, the background pixels on the boundary are connected, which is termed as *connectivity prior*. Methods by Wei *et al.* (2012) and Zhu *et al.* (2014) exploit this concept thoroughly to formulate their saliency models. Wei *et al.* (2012) define saliency of an image patch as the shortest-path distance to the image boundary, observing that background regions can easily be connected to the image boundary while foreground regions cannot. The Geodesic saliency of path P is computed as:

$$saliency(P) = min_{P_1=P,P_2,\ldots,P_n=B} \sum_{i=1}^{n-1} weight(P_i, P_{i+1}), s.t. P(P_i, P_{i+1}) \in \mathcal{E} \quad (2.9)$$

where, $\{P_i\}$ is the set of all image patches and B is a virtual background node. $\mathcal{E} = \{(P_i, P_j)|P_i \text{ is adjacent to } P_j\} \cup \{(P_i, B)|P_i \text{ is on image boundary }\}$. Hence, geodesic saliency of patch P is the accumulated edge weight along the shortest path to the background, B. These approaches work better for off-center objects but are still fragile and can fail even when an object only slightly touches the boundary. Moreover, they initially need some hand labeling (see Figure 2.1 reproduced from

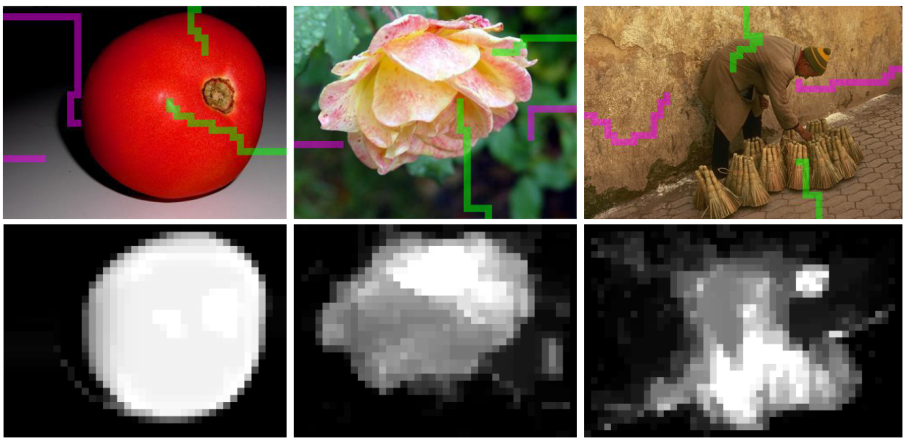


Figure 2.1: Examples from Wei *et al.* (2012) showing the paths of background (in magenta) and foreground (in green) from the boundary in the top row. Bottom row shows saliency maps retrieved by their algorithm.

Wei *et al.* (2012)). Zhu *et al.* (2014) use the same concept in an optimization framework. Yang *et al.* (2013) also use background prior, to rank image element or superpixels taking background and foreground as query, considering the boundary as background. Jiang *et al.* (2013) use contrast from boundary regions as a feature in their regression based learning framework.

2.1.5 Graph-based Modeling

Due their good spatial propagation of saliency information, many authors use graph based methods (Harel *et al.* (2006); Yang *et al.* (2013)). They formulate it as an optimization problem and solve either iteratively or derive a closed-form expression to generate the solution. In graph-based approaches, the image elements (pixels or superpixels) are modeled as the nodes of the graph and the similarity between image elements gives the edge weights. Thus, the image is mapped into a graph $G = (V, E)$, Vs are the node vertices and E is the set of edges with edge weights W. Gopalakrishnan *et al.* (2010) and Alexe *et al.* (2012) model the problem as quadratic energy models. Random walk models have been proposed for saliency detection by Harel *et al.* (2006) and Gopalakrishnan *et al.* (2010). Yang *et al.* (2013) use a manifold ranking technique which was originally proposed by Zhou *et al.* (2004*b*) to find the saliency.

Both random-walk based methods and manifold ranking method derives a closed-form expression from a graph-based optimization. Their closed-from solutions are derived from expressions similar to the *Page Rank* algorithm, $P = D^{-1}W$, where, P is the transition matrix, $W = \{w_{i,j}\}$ is the edge weight matrix and D is a diagonal matrix denoting the degree of each node as diagonal elements. Again, they exploit the Laplacian of the graph, $L = D - W$ to extract the saliency from the energy models. The concept of using the Laplacian of the image graph was taken from the spectral clustering method (Ng *et al.* (2001)). We also exploit this concept to define a rarity term in our perceptual cue based saliency method (see section 3.2.2).

In *manifold ranking* (MR) (Yang *et al.* (2013)), authors rank each of the superpixels, i.e., the nodes in the graph with respect to a given query node $x \in 0, 1$. The authors take an approach similar to semi-supervised learning using local and

global consistency (Zhou *et al.* (2004*a*)) and attempt to learn the optimal ranking function which best describes the relevance between unlabeled node and queries by solving an optimization function:

$$\sum_{i,j} w_{ij} \left(\frac{y_i}{\sqrt{d_{ii}}} - \frac{y_j}{\sqrt{d_{jj}}} \right)^2 + \mu \sum_i (y_i - x_i)^2 \tag{2.10}$$

where μ controls the balance between pairwise smoothness term (first term) and the unary term (second term) and y_i denotes the label of ith superpixel. The closed-form solution is found as (Zhou *et al.* (2004*b*)):

$$\mathbf{y}^* = (D - \alpha W)^{-1}\mathbf{y} \tag{2.11}$$

where, $\alpha = 1/(1+\mu)$. Using this ranking and each of the four boundaries as the queries, they formulate the saliency of each superpixels.

Random walk based methods find the equilibrium distribution of the network or the graph to formulate saliency. Since, these graphs are strongly connected, the chains or paths on the graph are ergodic and a unique equilibrium distribution exists (Harel *et al.* (2006)). Equilibrium distribution reflects the fraction of time a random walker would spend at a particular node, if he has to walk forever. This will automatically assign high value to the nodes surrounded by dissimilar nodes, i.e., the unique or rare nodes. Gopalakrishnan *et al.* (2010) find the equilibrium distribution as:

$$\mathbf{y}^* = (1-\alpha)((I) - \alpha \mathbf{P}^T)^{-1}\mathbf{s} \tag{2.12}$$

where, the vector $\mathbf{s}$ are the probabilities of random jump to the different nodes and $(1-\alpha)$ is the jump probability.

Most of the methods discussed above show result on MSRA-B saliency dataset (Achanta *et al.* (2009)) and show promising results. But images in the saliency datasets are generally highly biased, as pointed out by Li *et al.* (2014). Here the most significant bias is *center bias* which is an assumption of the saliency algorithms. Again mostly images have focused large objects without much of distractors. But while segmenting objects in natural images these assumptions

may not hold true. Hence, this dataset design bias creates a detrimental effect on benchmarking. Thus such saliency methods fail to perform well on the challenging PASCAL segmentation dataset (Everingham *et al.* (2012)). Hence, in addition to saliency, generic object specific cues are also important for modeling salient object segmentation algorithm.

2.2 Generic Object Segmentation Methods

Class independent object segmentation has recently gained importance in the Computer Vision community (Carreira and Sminchisescu (2010); Endres and Hoiem (2010)). Early methods of object detection were sliding window based and generally produced a bounding box instead of a pixel-accurate map. Sliding window methods perform search over a 4-dimensional search space of position, scale and aspect ratio. This requires a exhaustive search which is computationally very expensive. Hence, category independent object segmentation methods are useful and necessary.

2.2.1 Objectness

Alexe *et al.* (2012) first address the problem of detecting generic objects. The authors sample and rank 100,000 windows per image according to their likelihood of containing an object. This likelihood is called *objectness*. The objectness score is based on multiple cues derived from saliency, edges, superpixels, color contrast. These cues from Alexe *et al.* (2012) are discussed in the following subsections.

Multi-Scale Saliency

The saliency method proposed by Hou and Zhang (2007) is extended to multiscale and is processed for each of the color channels independently. The multiscale saliency map ($MS(w, \theta^s_{MS})$), for a window w at a scale s, is defined as:

$$MS(w, \theta^s_{MS}) = \sum_{\{p \in w | I^s_{MS}(p) \geq \theta_s\}} I^s_{MS}(p) \times \frac{|\{p \in w | I^s_{MS}(p) \geq \theta_s\}|}{|w|} \tag{2.13}$$

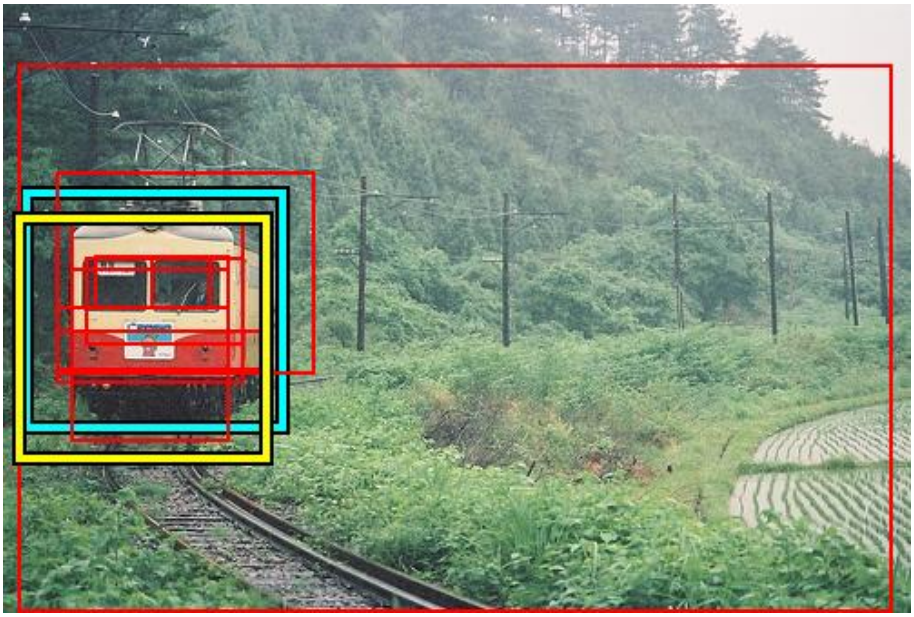


Figure 2.2: Figure shows different candidate bounding boxes from Alexe *et al.* (2012).

where, $I^s_{MS}(p)$ is the saliency for every pixel p and θ^s_{MS} are the scale-specific thresholds. This gives the uniqueness of a particular window.

Edge Density

The edge density (ED) factor captures the density of an edge near the window border. So, it gives a measure that the bounding box is in accordance with the image edges or object boundaries. Density of edgels inside a window w ($Inn(w, \theta_{ED}) = \frac{|w|}{\theta^2_{ED}}$) is computed as:

$$ED(w, \theta_{ED}) = \frac{\sum_{p \in Inn(w,\theta_{ED})} I_{ED(p)}}{Len(Inn(w, \theta_{ED}))} \tag{2.14}$$

where $I_{ED}(p) \in \{0, 1\}$ is the binary edge map, θ_{ED} is a parameter and $Len(.)$ determines the perimeters of the edgels.

Superpixel Straddling

Since, superpixels preserves the object boundaries, a 'good' window should not straddle a superpixel. This idea is presented as superpixel straddling (SS) and

measured as:

$$SS(w, \theta_{SS}) = 1 - \sum_{s \in S(\theta_{SS})} \frac{min(|s\ w|, |s \cap w|)}{|w|} \tag{2.15}$$

where the set of superpixels, $S(\theta_{SS})$, is found using the segmentation method proposed by Felzenszwalb and Huttenlocher (2004) at segmentation scale θ_{SS}. This cue computes the degree of straddling as minimum of the area of a superpixel inside and outside a particular window w.

Color Contrast

The color contrast (CC) term finds how much a window w is distinct from a window surrounding it ($Surr(w, \theta_{CC})$) and is given as the Chi-square distance between their histograms in Lab color space, as:

$$CC(w, \theta_{CC}) = \chi^2(h(w), h(Surr(w, \theta_{CC}))) \tag{2.16}$$

Hence, color contrast actually takes into account that the object window should be sufficiently distinct from the sounding background.

Alexe *et al.* (2012) learn the parameters θ using an MLE (Maximum Likelihood Estimation) approach. Finally the objectness score of a window w is given, using Naive Bayes, as:

$$p(obj|\mathcal{A}) = \frac{p(obj) \prod_{cue \in \mathcal{A}} p(cue|obj)}{\sum_{c \in \{obj,bg\}} p(c) \prod_{cue \in \mathcal{A}} p(cue|c)} \tag{2.17}$$

where, $\mathcal{C}$ = {MS,ED,SS,CC}, the 4 cues and $\mathcal{A} \subset \mathcal{C}$. The proposals tend to fit objects fairly loosely, but the first few hundred are of high quality (see an example in Figure 2.2, reproduced from the publication). The algorithm is fast but gives bounding boxes rather than pixel level segmentation output.

Chang *et al.* (2011) and Jia and Han (2013) use these objectness cues to obtain object-level saliency. Chang *et al.* (2011) propose that objectness and saliency values should be similar for a superpixel that belongs to a salient object. They

minimize an energy function, defined as:

$$E(\mathbf{x}^s, \mathbf{x}^o) = E_s(\mathbf{x}^s) + E_o(\mathbf{x}^o) + \Delta(\mathbf{x}^s, \mathbf{x}^o) \tag{2.18}$$

where E_s and E_o are the energy associated with saliency and objectness respectively and the third term brings the saliency and objectness value close. Whereas, Jia and Han (2013) use objectness as feature of each pixel to find the affinity matrix in a graph based approach and propose their final saliency map, as:

$$\hat{\mathbf{s}} = (diag(\mathbf{G1}) - \mathbf{G})^{+} \left[\mathbf{s} \;\; 1 - \mathbf{s}\right] \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{2.19}$$

where, $\mathbf{1}$ is an N length vector of all ones, elements of matrix $\mathbf{G}$ are computed using the weights between pixels i and j, and $\mathbf{s}$ is the saliency prior vector of length N, the total number of pixels. These two approaches delineate that saliency in conjunction with objectness can be substantially used for detecting object level saliency.

2.2.2 Object Segmentation Proposal

Methods in this class typically give a set of candidate binary maps or masks where each map gives a region in the image, so that each object in the image is represented by at least one of these maps. Both CPMC (Carreira and Sminchisescu (2010)) and Object Proposal (Endres and Hoiem (2010)) start with many seeds to predict a bag of masks as object proposals. Then, they rank order these maps based on precision of representing an object.

Carreira and Sminchisescu (2010) do a constrained parametric graph-cut to minimize the energy over the pixel labels $\{x_1, \ldots, x_k\},\ x_i \in \{0, 1\}$. The energy function is defined as:

$$E^{\lambda}(X) = \sum_{u \in \mathcal{V}} D_{\lambda}(x_u) + \sum_{(u,v) \in \mathcal{E}} V_{uv}(x_u, x_v) \tag{2.20}$$

where, $\lambda \in R$ is a parameter of the unary term, $\mathcal{V}$ are the vertices and $\mathcal{E}$ are the edges in the grid like graph based model. The unary term (D) estimates the

log-likelihood probability of a pixel u belonging to background or foreground using RGB color distribution. Whereas, the smoothness term (V) penalizes assignment of different labels to neighboring pixels u and v, based on the criteria that there is no gPb contour (globalPb as defined by Arbelaez *et al.* (2011)) between them. They minimize equation (2.20) for 30 a-priori defined different values of λ and split the foreground in different maps by a connected component analysis. This gives a large number of diverse maps. After rejecting very small and comparatively high energy segments, they rank the segments using a regressing based technique. For the purpose of ranking, 34 features are used exploiting graph partition properties, region properties (e.g, eccentricity, convexity, Euler number) and Gestalt properties to produce high quality candidate object segments.

Endres and Hoiem (2010) also produce a set of candidate object segments using different seed and graph-cut based energy minimization approach. But, they use different stronger criteria for computing their unary term as affinity for belonging to the same object as the seed chosen. The different features used for computing affinity include, cohesion (color and texture histogram intersection), boundary cues and layout agreement. Then to have a diverse ranking so that high quality segments of different objects are ranked higher, they take a max-margin based structured learning approach (Tsochantaridis *et al.* (2005)) on CRF. Since, these methods give as many as few hundred maps, they give a very good recall. However their precision is very low, as a lot of background regions are proposed as objects. The algorithms requires long computation time, as they rely on the gPb (Arbelaez *et al.* (2011)) edge detector. Although they (Endres and Hoiem (2010)) optimize based on intersection-over-union criteria to rank the maps, the results show that the top-most map generally contains almost half of the image (illustrated later in Chapter 4, Figure 4.1).

2.3 Summary

To conclude, saliency has been a prime research area for a long time. The fundamental ideas of saliency detection are inspired from cognitive science literature, that is, human visual system and perception. The area of work started with eye

fixation generation and fixation maps. Although, eye fixation is a elementary part of visual system, nowadays more interest has grown towards object oriented saliency map generation. This is due to its usability in many Computer Vision applications. Traditional methods use contrast, rarity and background prior for estimating saliency (Borji and Itti (2013)).

On the other hand, generic object segmentation methods have achieved popularity in recent literature. Objectness cues are exploited to achieve generic object detection or segmentation. These methods generate much less number of object proposals compared to the 4-dimensional search space of sliding window based methods. Hence, sliding window based bounding box approaches of object detection are getting replaced by object segmentation proposals. Further, saliency being a strong pre-attentive perceptual cue, objectness along with saliency can give promising object segmentation outputs. This has been the main motivation and focus of our work, in designing the proposed saliency methods proposed in this thesis.

CHAPTER 3

SALIENCY USING LOW-LEVEL PERCEPTUAL CUES

In recent literature, bottom-up visual saliency methods have become very popular and relevant to deal with a lot of computer vision tasks like object detection, object recognition, content based image retrieval, scene understating etc. Bottom-up saliency can be thought as a filter which extracts selected spatial locations of interest, which generally stands out from other locations. It reduces the search space of the problem and helps in extracting correct features for these tasks. It is a perceptual quality of the human visual system, by which humans attend to a subset of the pool of available visual information. Saliency of an image is given by a *saliency map* where we assign a normalized value to an image component or superpixel, denoting its probability of being salient. A salient region in an image is sufficiently distinct from its neighborhood in terms of visual attributes or features, and grabs attention. In this chapter, we concentrate on unsupervised bottom-up saliency detection technique when free-viewing a scene.

In this chapter, we propose a novel unsupervised formulation of saliency measure using appropriate low-level features for discriminating the salient regions from the background. Features are based on color and spatial distance among superpixels. First, a graph based (spectral clustering) rarity approach uses eigenvectors of the Laplacian of the affinity graph. Second, the spatial compactness term is a modified version of the distribution term of Perazzi *et al.* (2012). The third component is color divergence with respect to the superpixels at the border of an image. Wei *et al.* (2012) uses this concept in a semi-supervised algorithm which requires manual intervention. We statistically model the boundary patches using a Gaussian Mixture Model (GMM), and find the distance of all the patches from the modeled background colors in Lab color space. Integration of these priors gives the saliency map.

3.1 Motivation

Most work in the past have defined saliency by, either using spatial features like color, orientation, spatial distances between image patches (Itti *et al.* (1998), Goferman *et al.* (2010), Cheng *et al.* (2011), Perazzi *et al.* (2012)), or using spectral features like amplitude, phase spectrum (Hou and Zhang (2007), Achanta *et al.* (2009), Li *et al.* (2013)) and image energy in the spectral domain Hou *et al.* (2012) or graph based method (Harel *et al.* (2006)). Most of them have defined saliency as rarity of occurrence (or as a surprise) with respect to different local and global features. Color difference in CIELab space is the most distinctive feature used across most of the models. Our spectral clustering based rarity and spatial compactness measures of saliency exploit *rarity of feature* to extract salient regions in an image.

Spectral clustering (Ng *et al.* (2001)) is used in many different applications like, page ranking (Zhou *et al.* (2004*b*)), contour detection (Arbelaez *et al.* (2011)), normalized cut (Shi and Malik (2000)) approaches. A recent paper (Yang *et al.* (2013)) uses the ranking algorithm (Zhou *et al.* (2004*b*)) to find salient regions in images. We do not use any ranking technique (Zhou *et al.* (2004*b*)), nor we cluster the descriptors obtained from the eigenvectors of the graph Laplacian (Shi and Malik (2000)). Instead, we find the rarity using these descriptors itself, since eigenvectors themselves carry information about the superpixels.

Along with the above measures, we exploit the concept of background prior which is similar to the concepts of boundary prior and connectivity prior (Wei *et al.* (2012)). In our work, *background* refers to the non-salient spatial locations in the image. The main idea is that, the distance between background patches will be less and that between background and a salient patch will be more. Recent, cognitive science literature (Tatler (2007)) gives the evidence of boundary prior and shows human fixation happens mostly at the center. This motivates our second component of saliency detection, where we statistically model the boundary patches of an image and use them as background prior in a complete unsupervised formulation to detect saliency. We call our proposed method based on Graph-based Rarity, Spatial Compactness and Background Prior, as PARAM (background Prior And RArity for saliency Modeling).

3.2 Intuitive Understanding of the Method

A brief overview of the concepts that we exploit to propose the saliency detection method, based on low-level perceptual cues are illustrated below. In our approach, we use Lab color space due to its similarity with human perception (Tomasi and Manduchi (1998)).

3.2.1 Abstract the image into Superpixels

We first break our image into superpixels using SLIC superpixel (Achanta *et al.* (2010)) method which makes our method computationally fast. All computations are hence performed on superpixels which are much lesser in number than the set of pixels. Each superpixel is represented by a 5-D vector {labxy}. Thus each patch has its specific color and position.

3.2.2 Graph-based Spectral Rarity

We use the eigenvectors of the normalized Laplacian matrix of the affinity graph with superpixels as nodes. As given in Arbelaez *et al.* (2011), spectral graph theory (Ng *et al.* (2001)) and in particular the Normalized Cuts (Shi and Malik (2000)) criterion provides a way of integrating global image information into the process of grouping similar pixels. Given an affinity matrix W whose entries encode the similarity between pixels, one defines a diagonal matrix, as $D_{ii} = \sum_j W_{ij}$ and solves for the generalized eigenvectors of the linear system: $(D - W)v = \lambda.Dv$, that is, we find the Eigen vectors of the Laplacian matrix. If we look closely, the laplacian matrix provides a measure of the fraction of time a free random walker would spend at each node and what are the most preferable nodes to go from a particular node, considering the edge weights as cost of moving from one node to the other. Hence, as also mentioned by Arbelaez *et al.* (2011), this carries information about the edges in the image. If a random walker has low probability to move from a particular node to another, there is an edge in the image between the two superpixels. The descriptor extracted from the eigenvectors of the normalized Laplacian matrix, when using superpixels as nodes, would capture

the corresponding coarse texture information. Hence, local and global rarity based on these descriptor would give a measure of saliency which takes rarity of textures into account.

3.2.3 Spatial Compactness

We exploit the fact that a salient object would be spatially compact and the background colors will be distributed over the whole image (Goferman *et al.* (2010); Hou and Zhang (2007)). As, human eye can fixate at only one position and vision is centre surround, spatial compactness is an important characteristics of an object to become salient. So, the color belonging to the salient object will be spatially clustered together. Whereas, colors belonging to background will have high spatial variance. Hence, we use spatial variance of color or color compactness as a measure of saliency detection. The less the spatial variance more compact the object is and thus more salient.

3.2.4 Background Prior

Although rarity of feature is a strong cue for saliency, it is alone not sufficient as some previous methods in literature (Perazzi *et al.* (2012); Cheng *et al.* (2011)) which rely only on rarity of feature show. We exploit the concept of *boundary prior* (Wei *et al.* (2012)), which comes from the natural fact that boundary of an image would be mostly occupied by background (Tatler (2007)). Moreover, background will be mostly spatially distributed but homogeneous (in parts, say, the sky above and the grass below, for a natural scene) which results in compact clusters in color (feature) space. Hence, the distance between the background superpixels will be less, but that between background and foreground (salient) superpixels will be high, in 3-D Lab color space. However, occasionally a part of the salient object may exist at the boundary. Hence, it is not justified to consider all the boundary patches as background. To solve this, we statistically model the boundary prior using a GMM. Here, Gaussian modes with large number of pixels, having a large value of mixture coefficient, will generally model the background colors. Whereas, some Gaussian modes which model the few salient object patches, present at the

image boundary, will naturally have low mixture coefficients. Hence, we exploit Mahalonobis distance in color space, between the image patches and these modes, weighted by the corresponding mixture coefficients, to formulate the background prior.

3.2.5 Pixel Accurate Saliency

Finally, we combine the saliency measures yielding a granulated saliency map at superpixel level. To get a pixel accurate saliency map we use the up-sampling technique proposed by (Dolson *et al.* (2010); Perazzi *et al.* (2012)).

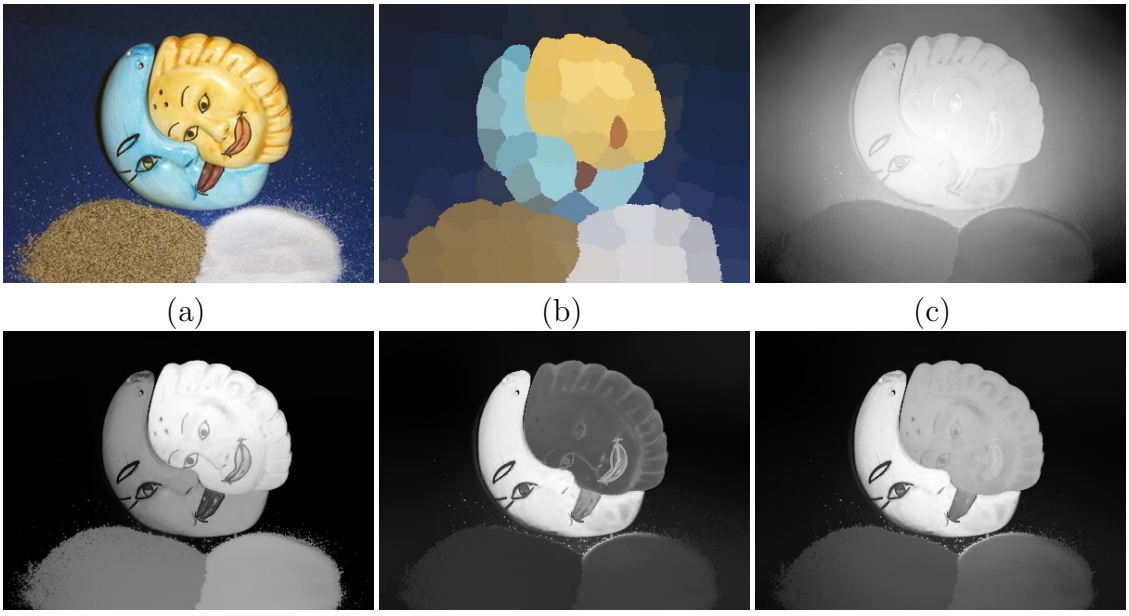
Results of individual components are illustrated in the Figure 3.1, which shows that it finally produces a desired saliency map. Experimental results and performance analysis discussed in Sec. 3.4.2 will reveal the superiority of this model than many recent state-of-the-art methods.

3.3 Algorithm for Saliency Map Estimation

We formulate two new measures of saliency detection, using graph-based rarity, spatial compactness of color and statistical model of boundary colors. The overall process of saliency computation is described in the following subsections:

3.3.1 Pre-processing

We first represent the image using superpixels by exploiting the concept of SLIC superpixel segmentation (Achanta *et al.* (2010)) in five-dimensional {labxy} space. We fix the target number of clusters to be 400, for all the experiments, to the SLIC superpixel algorithm (see appendix A.1.1) which yields N superpixels, $\{sp_i\}_{i=1}^{N}$. The benefit of SLIC segmentation is that, it produces compact homogeneous color patches as clusters. This helps the next stages of our algorithm. Each superpixel, sp_i has color in CIELab space c_i and position p_i. In the following subsection, we describe the analytical measures for saliency computation.



(a) (b) (c)

(d) (e) (f)

Figure 3.1: (a) Image from MSRA B dataset (Achanta *et al.* (2009)); (b) Superpixel abstraction of image in (a); Saliency detected by: (c) Graph-based rarity, (d) Spatial compactness, (e) Background prior; and finally the (f) proposed Saliency Map.

3.3.2 Saliency Computation

Our measure of saliency has broadly two components for salient object detection. The first one is a feature rarity based approach and is given by graph-based spectral rarity and spatial compactness of the salient object. Whereas, the second component of saliency detection utilizes the concept of boundary prior and connectivity prior (Wei *et al.* (2012)).

Graph-based Spectral Rarity

Given an image, we define a graph G = (V, E) whose nodes are the superpixels and edges E are the 8-neighbors of a superpixel. The edges are weighted by an affinity matrix W = $[w_{ij}]_{NXN}$. Let, D = $diag\{d_{11}, ..., d_{NN}\}$, be a diagonal matrix, where $d_{ii} = \sum_j w_{ij}$. The normalized Laplacian of the graph G, is given by, L = $I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. Let, $\{u_1, ..., u_k\}$ are the eigenvectors corresponding to largest k eigenvalues of L. We form the matrix $X_{N\times k}$ by stacking the eigenvectors in columns. Now we take normalized row vectors of $X_{N\times k}$ as the descriptor for each superpixel. Let, the k-dimensional descriptors be $\{x_1, ..., x_N\}$ and spatial position

of superpixel sp_i be p_i. We find rarity of sp_i, using the following formulation:

$$r_i = \sum_{j=1}^{N} ||x_j - x_i||^2 \exp\left(-k_r ||p_j - p_i||^2\right) \tag{3.1}$$

where, $||.||$ implies Euclidean distance. k_r controls the spatial extend of rarity. If k_r is infinite, it becomes a global rarity measure. k_r is set to 8.0 in all the experiments (as specified in Perazzi *et al.* (2012)). The affinity matrix W is defined as, $w_{ij} = \exp\left(-||c_i - c_j||^2\right), i, j \in V$, where c_i is the color feature of sp_i in CIELab space. Figure 3.1 (b) shows an example of the saliency map generated using only r_i as saliency probability of superpixel sp_i.

Spatial Compactness

To compute the spatial compactness, let us first define *spatial variance of color* (v_i) of a superpixel sp_i, with color in CIELab space c_i and position p_i as, how much similar colored patches are distributed over the image. A salient color is expected to be spatially compact and thus will be close to the spatial mean position of the particular color (Perazzi *et al.* (2012)). Thus, v_i is computed as,

$$v_i = \sum_{j=1}^{N} ||p_j - \mu_i||^2 . \exp\left(-k_c ||c_j - c_i||^2\right) \tag{3.2}$$

where, μ_i, the weighted mean position of color c_i, denotes the mean position of a particular color, c_i, weighted by the difference in color with other similar colored patches, as:

$$\mu_i = \frac{\Sigma_{j=1}^{N} p_j . \exp\left(-k_c ||c_j - c_i||^2\right)}{\Sigma_{j=1}^{N} \exp\left(-k_c ||c_j - c_i||^2\right)} \tag{3.3}$$

k_c controls the sensitivity of color similarity, while computing their spatial mean position. k_c is set to $\frac{1}{20^2}$ (as in Perazzi *et al.* (2012)), in all the experiments. High value of k_c implies that, only when the colors of the patches are very similar, it would contribute to the computation of μ for that particular color.

If the spatial variance of color for superpixel sp_i is less, it corresponds to a salient region, and not the background, as background colors are generally dispersed over the entire image. Thus, for a salient superpixel sp_i, its mean (μ_i) will be spatially near to p_i and also to all other p_js ($j \neq i$) belonging to similar (i.e., $\forall j | c_j \simeq c_i$)

colored patches in 2-D spatial space. Also, only those patches for which $c_j \simeq c_i$, contribute significantly to the summation terms of v_i. So, for a salient superpixel sp_i, v_i will be small as p_js are close to μ_i making $||p_j - \mu_i||$ small, $\forall j | c_j \simeq c_i$. Hence, lower the value of v_i, more salient is the superpixel sp_i.

Hence, our first component of saliency for superpixel sp_i, using feature rarity is given as:

$$F_i = \exp(-k.v_i).r_i \tag{3.4}$$

Figure 3.1 (c) shows the saliency map generated using only spatial compactness measure ($\exp(-k.v_i)$ for i^{th} superpixel). Large value of F_i indicates greater saliency. k is the scale of the exponent and set to 3 in all the experiments, as done in Perazzi *et al.* (2012).

Background Prior

The feature rarity based criteria specified above is not enough to find salient object in various types of images, specially with objects near boundary. We assume that boundary superpixels are less likely to be salient and recent studies in cognitive science (Tatler, 2007) reveals the same. We model the boundary superpixels using a Gaussian Mixture Model (GMM) in CIELab-color space and find the Mahalanobis distance (D_M) of all the superpixels from the means of the Gaussians. The distance of a superpixel from the boundary superpixel modes in Lab-color space is proportional to its saliency. Whereas, background superpixels are mostly homogeneous and thus the distances of background patches from these GMM means will be lesser. Again, boundary is mostly occupied by non-salient background superpixels. So, the GMM modes with large value of mixture coefficient (π_j) refer to a non-salient color.

Following above, the second component of saliency measure of superpixel sp_i, using background prior is formulated as,

$$B_i = \sum_{j=1}^{K} \pi_j . D_M(c_i, \mu_{Gj}) \tag{3.5}$$

where, $D_M(x, y)$ denotes the Mahalanobis distance between x and y, c_i is the color of i^{th} superpixel, μ_{Gj} is the mean of jth Gaussian mode, π_j is the weight or

mixture coefficient of the j^{th} Gaussian and K is the number of GMM components used to model the distribution of the boundary superpixels in CIELab space. We dynamically compute the optimal value of K maximizing the cluster compactness of the boundary patches. Cluster compactness can be defined as $\sum_i \sum_j (x_j - \mu_i)$. It denotes how compact (in corresponding feature space) all the clusters are for any particular clustering obtained. It is a goodness measure of the clustering. To find the optimal K (number of GMM components), we iterate over number of clusters starting from 1 to 4, considering all four boundaries, and find cluster compactness using k-means clustering. K is taken as the number of cluster for which the cluster compactness is best. Figure 3.1 (d) shows the saliency map generated using only background prior, B_i as saliency probability of i^{th} superpixel sp_i.

3.3.3 Saliency by Up-Sampling to Image Resolution

Saliency of each pixel is obtained as a weighted linear combination of saliency of its surrounding image elements, sp_js, using the idea proposed by Dolson *et al.* (2010). In our work, S_j, the saliency value of sp_j, is sum of F_j and B_j ($S_j = F_j + B_j$) and we use the same formulation as used by Perazzi *et al.* (2012) (also see appendix A.2 for details). The saliency of ith pixel is computed as,

$$\tilde{S}_i = \sum_{j=1}^{N} w_{ij} S_j \tag{3.6}$$

where, $w_{ij} = \frac{1}{Z_i} exp(-\frac{1}{2}(\alpha||c_i - c_j||) + (\beta||p_i - p_j||))$
Z_i is the normalizing factor, so that $\sum_{i=1}^{N} w_i = 1$. α and β are the parameters which controls the sensitivity of up-sampling to color and position respectively.

The whole procedure is illustrated with an example in Figure 3.2. The figure also shows that our two components of saliency measures are complementary to each other and the combined measure produces an improved saliency map than the individual components themselves. Figure 3.3 shows the contributions of the individual saliency priors, demonstrated using Precision-Recall (defined in section 3.4.2) curves. It illustrates the performances by excluding r_i from equation (3.4) (termed as 'without Rarity' in figure), and only exploiting equation (3.4) without

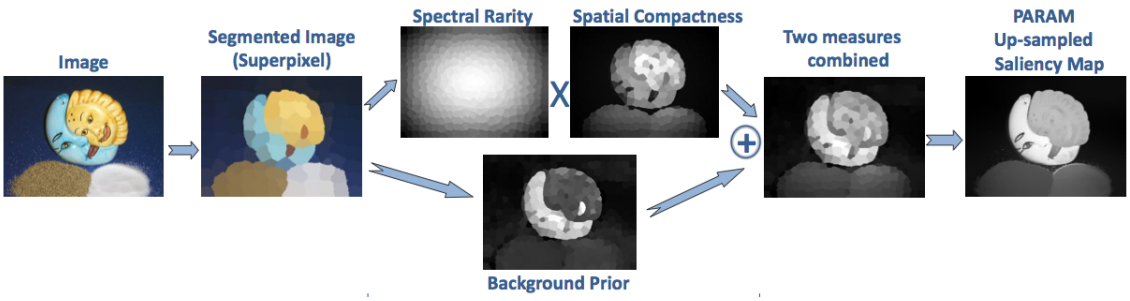
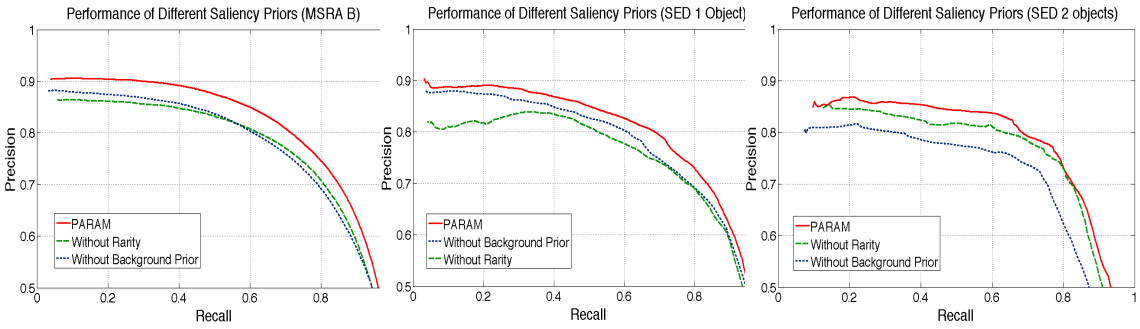


Figure 3.2: Illustration of the different stages of our proposed algorithm for saliency estimation with an example from MSRA-B Dataset (section 3.4.1).



(a) (b) (c)

Figure 3.3: Performance curves illustrating the importance of the different components of our proposed method (PARAM) for saliency computation (eqn. (3.6), using Precision vs Recall metric on: (a) MSRA-B; (b) SED1 and (c) SED2 datasets.

the background prior (equation (3.5)) measure (termed as 'without Background Prior' in the figure). It also shows the performance of the combined final saliency map ($PARAM$), on different datasets (for details of experimentations, see Section 3.4.2). The figure illustrates that excluding either the rarity prior or the background prior measure deteriorates the performance of the method, and thus quantitatively establishes the importance of these measures.

3.4 Results and Experimentation

The method proposed in this chapter is evaluated on popular saliency and object segmentation datasets. The performance of the method is measured using precision, recall and f-measure metrics which are described in the following subsections.

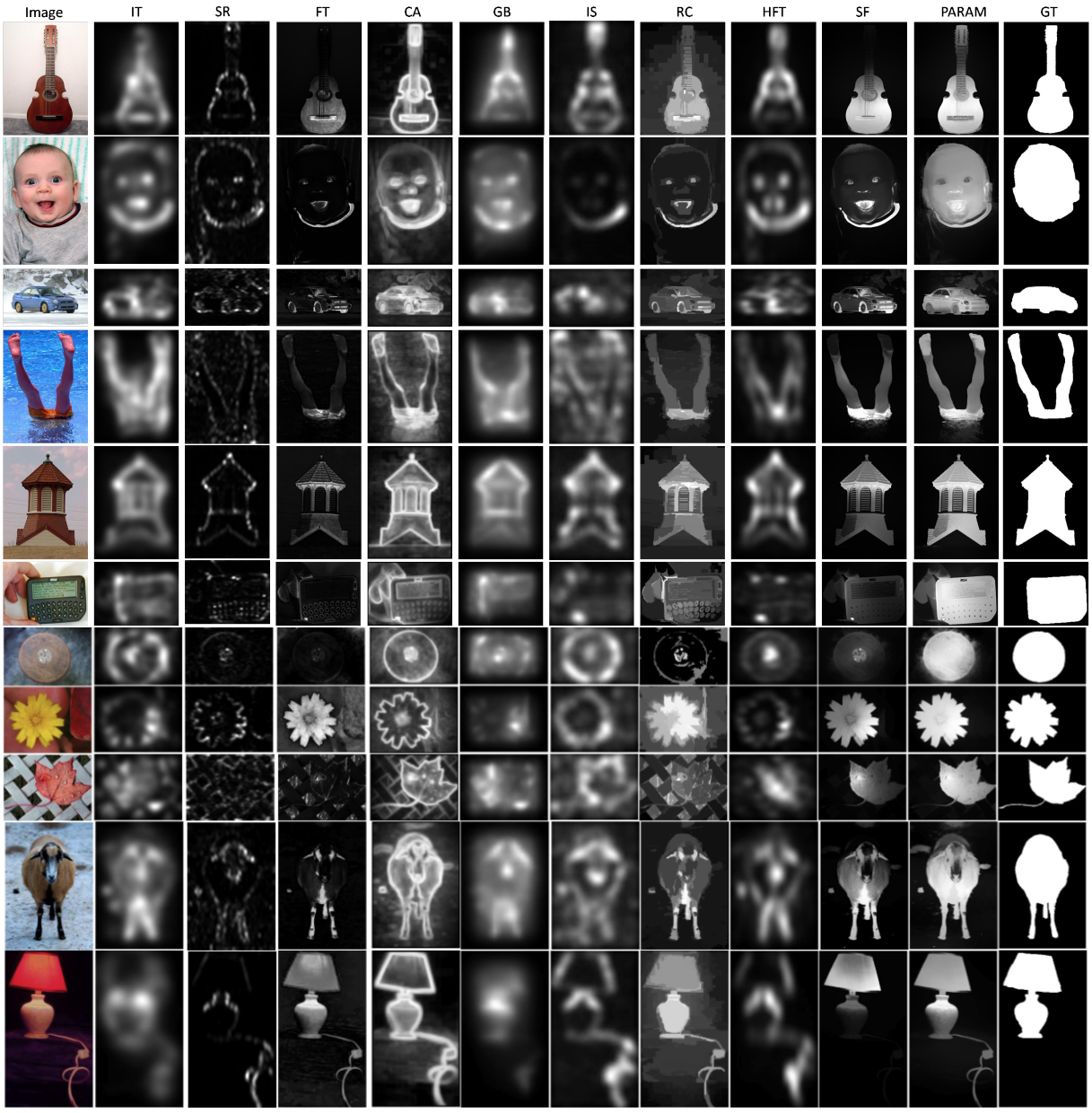


Figure 3.4: Visual comparison of the results of nine state-of-the-art methods along with our proposed method (PARAM) of saliency estimation, on eleven different samples of MSRA-B dataset. *PARAM* consistently performs better for different types of images including indoor, outdoor natural scenes, when compared with the ground truth (GT) given in the last column.

3.4.1 Datasets

We evaluate the performance of our proposed method (*PARAM*) using the following two benchmark datasets.

MSRA-B

MSRA-B[1] has 5000 images with their ground truth masks as given in the work by Achanta *et al.* (2009) and Jiang *et al.* (2013). Images are of numerous kind including indoor, outdoor natural scenes, humans, animals with different types of contrast and color variance. This makes the dataset diverse and challenging.

SED

Segmentation Evaluation Dataset (SED)[2] has two parts, SED1 and SED2. SED1 has 100 images with a single salient object. SED2 images has 100 images with 2 salient objects of different size and color. Ground truth masks for all the images are publicly available.

3.4.2 Experimental Results and Performance Analysis

We compare the performance of our proposed method, *PARAM*, with 9 state-of-the-art methods. Here, IT denotes Itti *et al.* (1998), SR is Hou and Zhang (2007), CA is Goferman *et al.* (2010), FT is Achanta *et al.* (2009), RC is Cheng *et al.* (2011), IS is Hou *et al.* (2012), GB is Harel *et al.* (2006), HFT is Li *et al.* (2013) and SF denotes Perazzi *et al.* (2012). We use this set of acronyms for the rest of the thesis to refer to these prior published works.

Figures 3.4 - 3.7 show results of these 9 different state-of-the-art saliency detection methods along with our proposed saliency method, *PARAM*. Figure 3.4 gives a visual illustration of *PARAM* and the 9 other methods. Saliency map provided by PARAM is closest to the ground truth (denoted by GT) and highlights the overall

[1] `http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm`

[2] `http://www.wisdom.weizmann.ac.il/~vision/Seg_Evaluation_DB/dl.html`

salient object uniformly, giving better results than the existing state-of-the-art methods.

Quantitative Performance Evaluation

We quantitatively evaluate the performance of our method (*PARAM*) using precision, recall rate similar to the Achanta *et al.* (2009), Cheng *et al.* (2011), Hou and Zhang (2007).

Precision measures what part of the predicted output is correct and is given by the formula,

$$Precision = \frac{|S \cap G|}{|S|} = \frac{tp}{tp + fp} \tag{3.7}$$

where, S is the predicted output and G is the ground truth map. Both S and G have 1's as the salient pixels and 0's as background. tp and fp are true positive

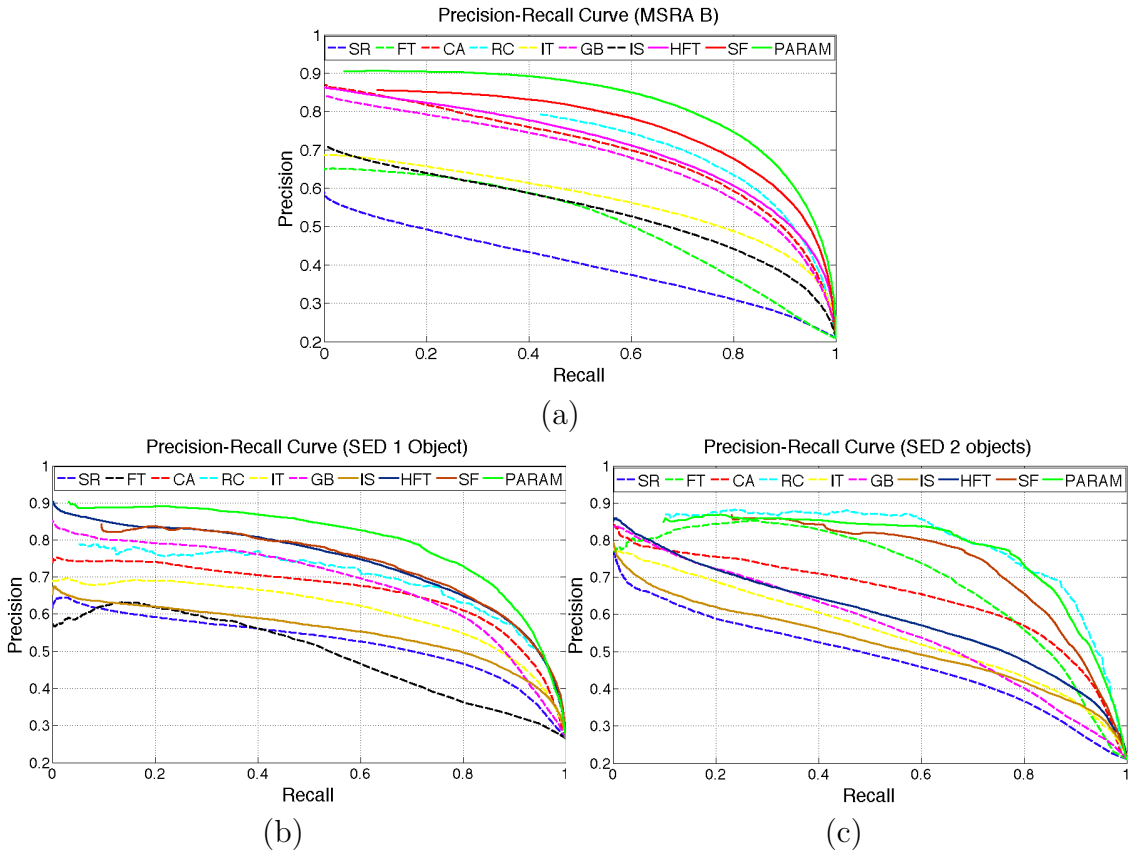


Figure 3.5: Performance analysis of 9 different state-of-the-art methods along with our proposed method (PARAM) using Precision vs Recall metric on: (a) MSRA-B; (b) SED1 and (c) SED2 datasets. It shows that PARAM out-performs all the methods on MSRA-B and SED1 datasets, and is the second best for SED2 dataset. This figure is best viewed in color.

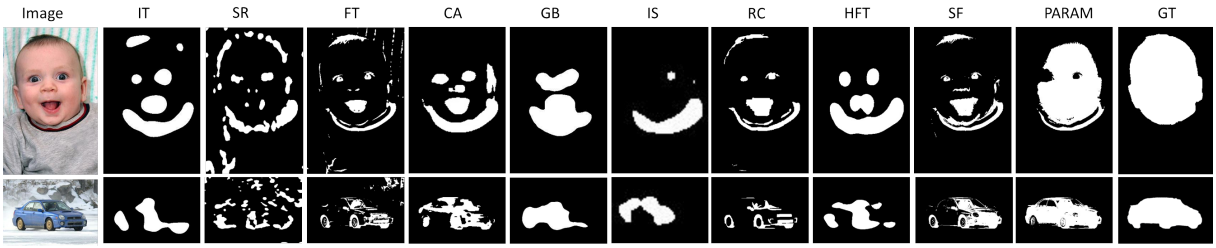


Figure 3.6: Visual comparison of the Adaptive Cut binary maps of the nine state-of-the-art methods and our proposed method (PARAM), on two samples of MSRA-B dataset, with the ground truth as given in the last column.

and false positive respectively.

Recall measures what part of the correct output has been predicted and is given by the formula,

$$Recall = \frac{|S \cap G|}{|G|} = \frac{tp}{tp + fn} \tag{3.8}$$

where fn is false negative. In order to generate the precision-recall curve we threshold the saliency map by $\{0, \ldots, 255\}$ values and compute the precision and recall similar to Achanta *et al.* (2009). We have compared our method with all the above mentioned 9 state-of-the-art methods. We do not compare with Jiang *et al.* (2013), as being a training based method, it has output for only their test set which contains 2000 images from MSRA-B. Our method, *PARAM*, clearly out-performs all the methods on MSRA-B (Figure 3.5 (a)) and SED1 (Figure 3.5 (b)) datasets. Only for the SED2 (Figure 3.5 (c)) dataset, *PARAM* is not a clear winner. This is mainly due to the occasional presence of two objects only on the boundary. Such a scenario is not biologically plausible to become salient for human vision either.

We take the *adaptive threshold* (T_a) as twice of average saliency (S_{avg}) and create a binary map, which is proposed as Adaptive Cut in Achanta *et al.* (2009).

Method	IT	FT	GB	CA	RC	IS	HFT	SF	PARAM
Time (s)	0.41	**0.13**	1.63	128.05	0.21	2.20	0.76	0.23	**0.23**

Table 3.1: Average runtime (in seconds per image) of different competing methods of estimating saliency.

S_{avg} is obtained as:

$$S_{avg} = \frac{1}{WH} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} S(x, y) \tag{3.9}$$

where, S is the full resolution saliency map of width W and height H.

Figure 3.6 shows the binary maps or adaptive cuts, on 2 images from MSRA-B dataset, which are generated using adaptive threshold, from the saliency maps obtained from the 9 different state-of-the-art and our proposed method, *PARAM*. From these binary maps we calculate specific values of precision, recall and the f-measure as in Achanta *et al.* (2009), for the 9 methods along with *PARAM*.

Given Precision and Recall, **F-measure** is computed as:

$$F_\beta = (1 + \beta^2) \frac{Precision \times Recall}{\beta^2 Precision + Recall} \tag{3.10}$$

We use $\beta^2 = 0.3$ as also used by Achanta *et al.* (2009) and Perazzi *et al.* (2012).

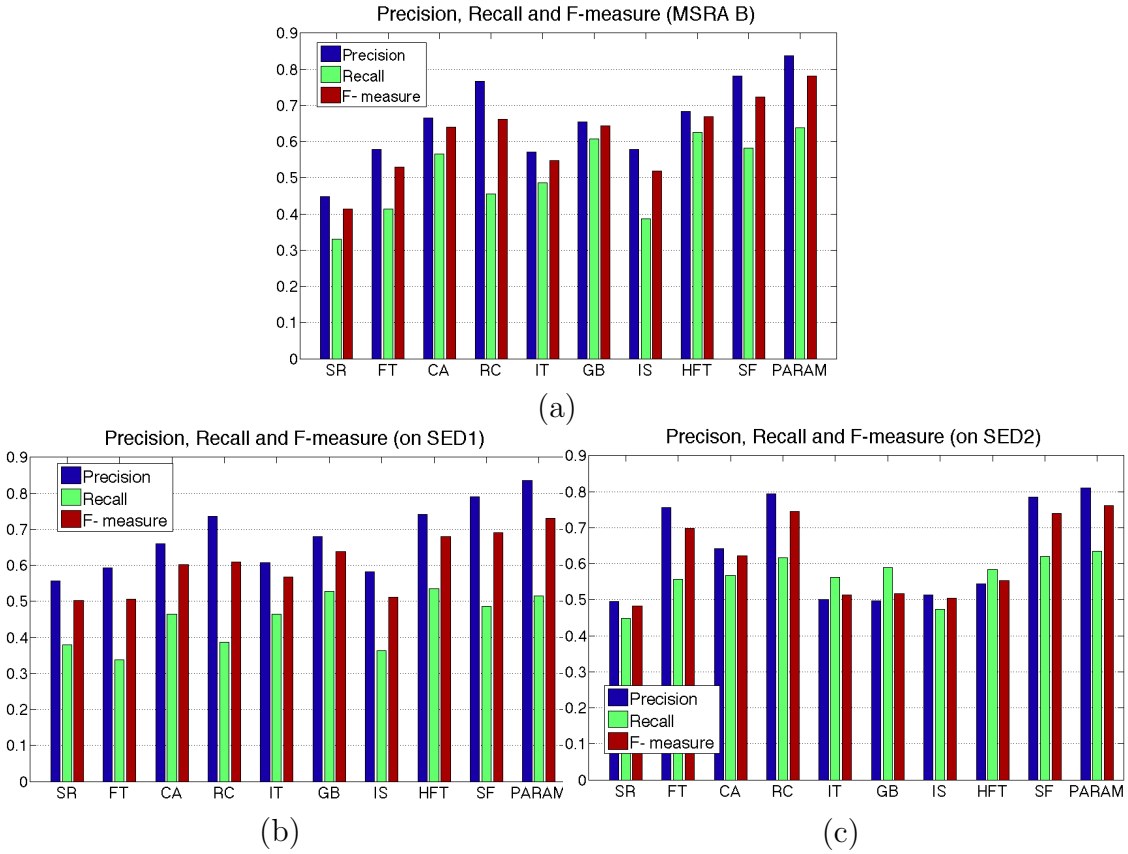


Figure 3.7: Precision, Recall & F-measure using adaptive cut, on (a) MSRA-B; (b) SED1 and (c) SED2 datasets, show that our method (PARAM) performs better than all the 9 state-of-the-art methods for all the datasets. RC performs close to PARAM only in case of SED2 dataset.

The bar charts in Figure 3.7 show that the adaptive cuts generated using *PARAM* saliency output produce the best result for all the three performance measures: Precision, Recall and F-measure. This implies that *PARAM* produces better segmentation than the 9 competing state-of-the-art methods. Only RC (Cheng *et al.* (2011)) has comparable (near, but marginally better) performance to PARAM for SED2 Dataset (see Figure 3.5 (c)).

Efficiency

Although our method has more than one saliency priors to be computed, it is time efficient and can be easily used as preprocessing step for different applications. This is mainly due to the fact that the saliency computation by our method is performed on image patches (or superpixels) which are much lesser in number than the set of pixels. Moreover, parallel computation of the priors is also possible. We compare the running time of our implementation (in C++) with other competing methods. We use the Matlab implementation from authors for Itti *et al.* (1998), Goferman *et al.* (2010), Achanta *et al.* (2009), Li *et al.* (2013), Harel *et al.* (2006), Hou *et al.* (2012) and C++ implementation of Cheng *et al.* (2011), Perazzi *et al.* (2012) on a intel core 2 extreme 3.00 GHz CPU with 4 GB RAM. Table 3.1 lists the average running time of 8 competing methods along with *PARAM*. In case of Hou and Zhang (2007), we get the results from the publicly available executable of Cheng *et al.* (2011), and we do not have the time efficiency information for the same. The work (FT) proposed in Achanta *et al.* (2009) is the fastest, but performs much inferior (refer Figures 3.4 - 3.7). RC (Cheng *et al.* (2011)) is as fast as our method and performs comparable on SED2 dataset, but the results are not at par with us on MSRA-B and SED1 datasets.

3.5 Discussion

In this chapter, we have presented a bottom-up saliency estimation method for images purely based on low-level perceptual cues. We have proposed a novel graph-based feature rarity computation, utilizing the concepts of spectral clustering (Ng *et al.* (2001)). It shows that eigenvectors of the Laplacian of the affinity matrix

of the graph, taking image elements as nodes gives a good measure of rarity. This term actually also accounts for center prior. Additionally, we exploit spatial compactness of color and use the cue of boundary prior by statistically modeling the background in color space. We show, both qualitatively (Figure 3.1) and quantitatively using Precision-Recall metric (Figure 3.3), that these two priors compliment each other. We also give a comparative study of the performance of our method with 9 state-of-the-art methods, using three different measures of evaluation on two popular real-world benchmark datasets. Since, our method is not just restricted to global spatial feature rarity, but also utilizes the boundary cue as well as spectral clustering based feature rarity, it gives better performance (Figure 3.5 and 3.7) and in most of the cases accurately detects the salient object.

CHAPTER 4

SALIENT OBJECT SEGMENTATION IN NATURAL IMAGES

Segmenting objects in a scene, without prior knowledge about the class of the object, is a significant and highly challenging problem. Bottom-up salient object detection methods attempt to predict a probability map with pixel-accurate object locations, but fail in most natural images. We propose an efficient method that uses saliency in conjunction with objectness cues to predict the likelihood of a region belonging to an object in an image. Further, a CRF based approach with these likelihood priors perform a salient object segmentation using graph cut. Our method is orders of magnitude faster than various state-of-the-art algorithms and can be employed as a pre-processor to different high-level computer vision tasks. We compare our method against saliency methods as well as category independent object proposal methods on the PASCAL 2012 segmentation dataset (Everingham *et al.* (2012)). Our method shows a 21% increase in performance using intersection-over-union score when compared with the top 10 masks of the recent category independent object proposal methods (Carreira and Sminchisescu (2010); Endres and Hoiem (2010)).

4.1 Motivation

Human visual system has an amazing capability to localize objects even before recognizing them. This comes from the ability to select regions with important visual information during early vision. This ability of the human visual system is known as *visual saliency*. Again, cognitive science literature describes that spatial groupings of a small set of simple primitives give the early description of an image (Treisman and Gormican (1988)). Thus, localization of multiple objects in an image happens as a part of early visual processing. This indicates that saliency can be substantially utilized for localizing objects, imitating the human

visual system. Salient object segmentation can then be successfully used as a pre-processing step to accomplish low-level tasks, e.g., shape-based feature extraction, and high-level vision tasks such as, object recognition, scene understanding, object tracking and so on.

Category dependent object localization algorithms work only for a predefined set of objects and are practically infeasible given the huge number of classes that exist in reality. Studies show that human beings can localize objects even when the identification or recognition system is impaired (Goodale *et al.* (1991)). There has been thorough research in class-specific object detection and localization. Sliding window approaches (Viola and Jones (2004); Felzenszwalb *et al.* (2010)) try to find objects at different windows at different scales and orientations. Therefore, these methods incur huge computational cost, due to the 4-D search space of sliding window approach. Further, state-of-the-art segmentation methods (Shi and Malik (2000); Arbelaez *et al.* (2011); Felzenszwalb and Huttenlocher (2004)) are often not suitable to extract object-specific image regions. Hence, it is important to device such a system which can localize objects without any prior knowledge about it. Later, more features can be learned from these extracted object regions, and recognition algorithms can also be benefited from this spatial filtering.

We propose a Salient Object Segmentation method that captures the same visual processing hierarchy as in the human visual system. Our goal is to localize objects in an image independent of its category. Our method of salient object segmentation uses *saliency* and *objectness* feature as two important cues to generate a single salient object segmentation map. The aim of the map is to depict all the object regions with high probability values. Natural images exhibit spatial interactions, e.g., neighboring pixels are likely to belong to the same object or spatially bounded regions by image contours generally represent an object part. Graph-based methods can capture these dependencies and provide good spatial propagation of saliency information. Hence, we employ a graph-based approach and model our method as a conditional random field (CRF) based optimization approach. We perform all necessary processing at the superpixel level. To determine the superpixels of an image, we use SLIC algorithm (Achanta *et al.* (2010)) which preserves primitive information like color, object boundary and edges (refer Section 3.3.1). We incorporate low-level perceptual cues within the saliency

prediction method. Additionally, the objectness factors are incorporated based on a geometric constraint and distribution of edges in the image. *Objectness*, as first defined by (Alexe *et al.* (2012)), describes the features that predict the likelihood of a superpixel belonging to any object. Thus objectness in combination with saliency gives the likelihood of each superpixel belonging to a salient object and forms the unary potential of CRF in the proposed method. Since, many objects are roughly homogeneous in appearance as discussed by Endres and Hoiem (2010), CRF smoothness constraint gives a benefit. In the following two sections, we first describe the image cues, namely saliency and objectness, followed by our graphical model formulation, inference and learning methods. We have tested our method on the challenging PASCAL 2012 dataset (Everingham *et al.* (2012)) and it shows better performance than other existing saliency as well as object proposal methods, in terms of both F-measure and intersection-over-union score.

4.2 Image Cues

The aim of our approach is to segment all the salient objects in an image. Saliency methods generally produce a probabilistic saliency map. Therefore, to segment the objects from an image, we use objectness criteria in conjunction with saliency. We characterize objectness by two constituent factors, geometric constraint and distribution of edges in the image. These two features are respectively modeled as boundedness and edge-density. To find the image cues, an image is first segmented into a set of N superpixels, $\{sp_i\}$, i = 1,. . . ,N using the SLIC algorithm by Achanta *et al.* (2010) (see appendix A.1.1). Then all the image cues are computed over superpixels as described in the following subsections.

4.2.1 Saliency as a Cue

Motivated by biological factors of human vision, as described in section 4.1, we employ saliency as a primary factor in our algorithm. Saliency detection methods mostly rely on low-level cues, such as, center-surround response (Itti *et al.* (1998)), frequency domain features (Achanta *et al.* (2009); Hou *et al.* (2012); Li *et al.* (2013)) or local and global contrast based information (Goferman *et al.*

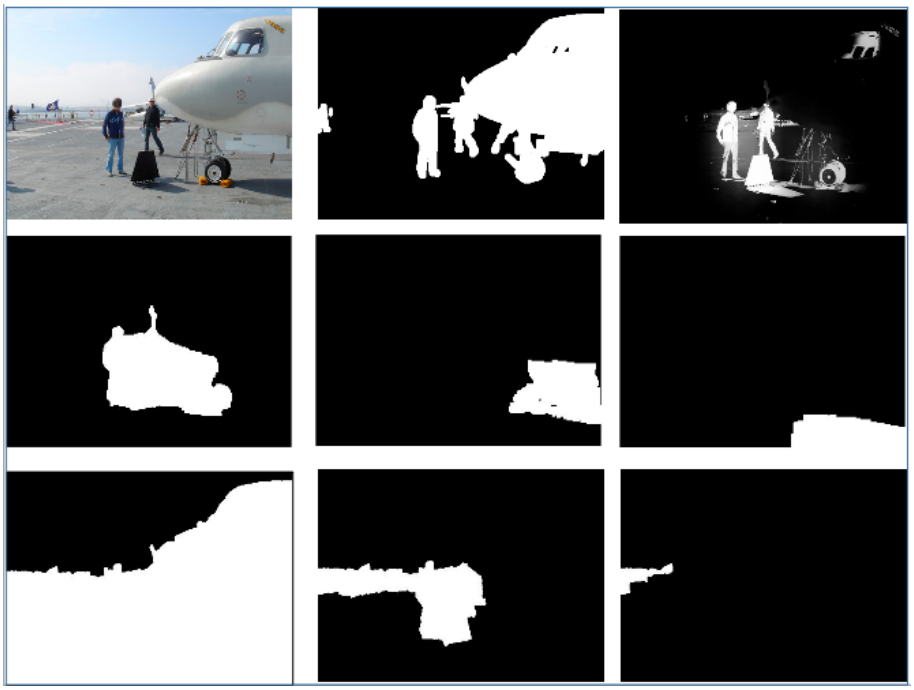


Figure 4.1: Examples of category independent models on a sample image from PASCAL VOC 2012 segmentation dataset Everingham *et al.* (2012). From left to right, first row shows the image, its binarized ground truth and output of proposed method (refer Section 4.3). Second and third row show top 3 ranked maps of CPMC Carreira and Sminchisescu (2010) and Object Proposal Endres and Hoiem (2010) methods respectively.

(2010); Cheng *et al.* (2011); Perazzi *et al.* (2012); Yang *et al.* (2013); Liu *et al.* (2011)), as discussed in Section 2.1. All these methods attempt to detect the rare or unique information in an image and represent that as salient. This kind of approach is known as feature rarity based approach. Our proposed saliency method described in Section 3.3 and also that in Wei *et al.* (2012) show that the feature rarity alone is not enough to describe the salient object in complex scenes. Therefore we have introduced a background prior term. Background prior argues that most of the area in boundary are occupied by non-salient regions and these regions are connected with each other. The criteria formulated for these pair of propositions are termed as boundary prior and connectedness prior. Further, distance from boundary (in feature space) gives a measure of saliency, as also described in Section 3.3.2. These methods, which have exploited background prior for saliency detection, have given considerably improved performance in terms of localizing region belonging to objects, as shown in literature.

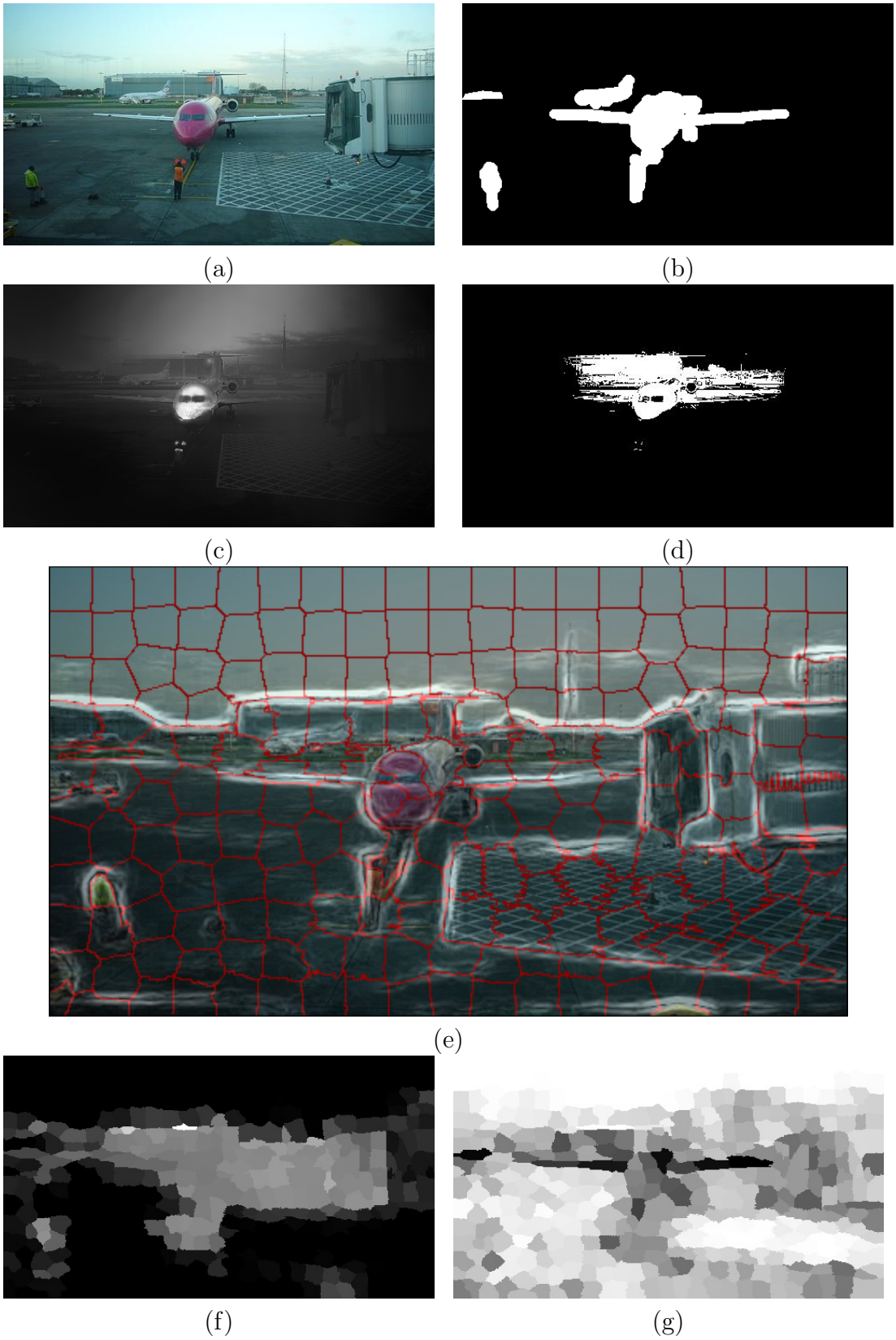


Figure 4.2: The figure shows the (a) Image with (b) it's binary ground truth and (c) saliency map of PARAM (Chapter 3); (d) extracted airplanes by the proposed Salient Object Segmentation method. The bottom row illustrates the objectness factors with (e) the edge map Dollár and Zitnick (2013) on superpixeled image and the two cues: (f) boundedness and (g) edge-density.

All the methods discussed above are bottom-up or stimulus driven methods and do not exploit any prior information about any specific object. There exists on the other hand, top-down saliency methods (Zhang *et al.* (2008); Yang and Yang (2012)) which learn the features of the objects to be found and given that object class as an input, the object becomes salient within an image (Yang and Yang (2012)). However, similar to category dependent object segmentation models these methods cannot localize an object before recognizing it. So, they do not conform with the goal of our work.

For the purpose of our salient object segmentation task we require a bottom-up saliency method which better predicts object regions employing both the feature rarity and boundary prior cues. We have shown in Section 3.4.2 that our proposed method (PARAM, section 3.3) uses these two factors effectively and in a time efficient way. As we are proposing our method to work as a pre-processing stage for different computer vision tasks, computational efficiency is very important without compromising on performance. Hence, we use the map produced by PARAM as the saliency cue. The saliency detection method, PARAM produces a probabilistic saliency map on superpixels which is then upsampled to pixels. We use this superpixel saliency map and denote the saliency probability value of ith superpixel as s_i.

Figure 4.2(c) shows the saliency map of PARAM on a sample from PASCAL VOC 2012 segmentation dataset. Clearly, the saliency map visually delineates that the rare features are depicted as salient, although not very accurately. PARAM uses compactness in color space and distinctness from boundary as the prominent cues for saliency. Hence, only the purple head of the airplane is filtered as salient. Wings on the other hand are completely ignored, due to color similarity with sky on the boundary. Furthermore, since white color is not detected as a compact color in the image, the white airplane in the background is also not identified. Moreover, partly the sky is predicted as salient which is not correct. More specifically, PARAM exploits color information in great detail, but shape and edge informations are not utilized. As, human eye is most sensitive to color and brightness, the proposed saliency method (PARAM) performs good on saliency datasets, but fails when tested on natural image datasets such as PASCAL VOC

dataset (Everingham *et al.* (2012)).

4.2.2 Objectness Features

Objects typically have well defined boundaries (Alexe *et al.* (2012)) and many objects are mostly aggregation of regions that are homogeneous in appearance, as also mentioned by Endres and Hoiem (2010). Superpixels preserve object boundaries, as superpixel algorithms (e.g., Achanta *et al.* (2010)) group the pixels with homogeneous color and texture as a superpixel. So, there should be no superpixel straddling by edges in an image (Alexe *et al.* (2012)). Recently, Dollár and Zitnick (2013) have described an efficient edge detection algorithm. Their method gives a high-quality edge probability map using structured learning (Nowozin and H. (2011)) prediction on random forest. Since, they do a direct inferencing, the method is computationally efficient than all competing edge detection methods. We use this algorithm to generate an edge map. Next we compute the boundedness and edge-density factors.

Boundedness

We first define the strong edges in an image as the pixels with high probability values ($> T = 0.8$) in an edge map. As there should not exist any superpixel straddling, the pixels with high edge probability values mostly correspond to object boundaries. We define boundedness of a superpixel sp_i as b_i, based on the extent to which it is bounded by strong edges on all four directions. Boundedness is calculated at pixel level first and then averaged to superpixels. Edge contours on an edge map may be discontinuous and often exist with small gaps. Due to this, some pixels which are visually bounded and belong to some object may not be bounded on all four directions, producing a low score on boundedness. But all of the rest of the pixels within the particular superpixel would not score low, if that superpixel belongs to an object. Hence, averaging over all the pixels makes it insensitive to noise in the edge contour. Moreover, as it can handle the discontinuities in the object boundary in an edge map, a computationally expensive high quality edge map (Arbelaez *et al.* (2011)) is not required.

Boundedness of superpixel sp_i is formulated as:

$$b_i = \frac{1}{|sp_i|} \sum_{p \in sp_i} (l_{p(x,y)} + t_{p(x,y)} + r_{p(x,y)} + d_{p(x,y)}) \\ .\mathcal{I}(l_{p(x,y)}, t_{p(x,y)}, r_{p(x,y)}, d_{p(x,y)}) \tag{4.1}$$

where, $l_{p(x,y)}, t_{p(x,y)}, r_{p(x,y)}, d_{p(x,y)} \in [0, 1]$ denote the strength of the left, top, right and bottom boundaries (defined later) respectively, obtained from the edge probability map of the particular pixel p with spatial location (x, y). $|sp_i|$ denotes the number of pixels within ith superpixel. Also,

$$\mathcal{I}(l_{p(x,y)}, t_{p(x,y)}, r_{p(x,y)}, d_{p(x,y)}) = \\ \begin{cases} 1, & if\ l_{p(x,y)}, t_{p(x,y)}, r_{p(x,y)}, d_{p(x,y)} > 0 \\ 0, & otherwise \end{cases} \tag{4.2}$$

is an indicator function and represents whether the pixel is close-bounded. For all pixels the boundedness values dictated by the edge strength of it's boundaries are estimated.

The edge map gives a probability map where each pixel value denotes the strength of an edge passing through that particular pixel. Let the edge map value for a pixel at spatial location (x, y) be $\mathcal{P}e_{(x,y)}$. We use a dynamic programming approach to compute the boundedness in order of number of pixels, and then recursively define the strength of the left boundary as:

$$l_{p(x,y)} = \begin{cases} l_{p(x-1,y)} & if\ \mathcal{P}e_{(x,y)} < T \\ 0 & if\ x = 0 \\ \mathcal{P}e_{(x-1,y)} & otherwise \end{cases} \tag{4.3}$$

Similarly, boundary strengths, $t_{p(x,y)}$, $r_{p(x,y)}$, $d_{p(x,y)}$ can be computed. For the whole image all the values of $l_{p(x,y)}$, $t_{p(x,y)}$, $r_{p(x,y)}$, $d_{p(x,y)}$ are computed only once in O(number of pixels) time. Thus while calculating b_i, these values are accessed in O(1). High boundedness value implies that most of the pixels in the superpixel are bounded by strong edges, and is likely to belong to an object.

Edge Density

The distribution of edges in an image is captured as a cue using our edge-density term. Since objects are mostly homogeneous in appearance (Endres and Hoiem (2010)), there should be less edges inside the superpixels belonging to an object. High density of edges in a region generally implies a cluttered background, e.g., rippling river, grass or forest. Again, very low density or overly smooth regions which are also not bounded should be part of background, e.g., clear sky. As no strong edge crosses over a superpixel (i.e., image boundaries are respected by superpixels) there should be only weak edges inside a superpixel. So, we compute the density of the edges within a superpixel sp_i, as:

$$density_i = \frac{1}{|sp_i|} \sum_{p(x,y) \in sp_i} \mathcal{P}e_{(x,y)} \tag{4.4}$$

Now, we compute the mean μ_d and standard deviation σ_d of the set of densities, $\{density_i\}$, i = 1,. . . ,N. For ith superpixel, the edge-density ed_i, is calculated as:

$$ed_i = \begin{cases} 1 - density_i & if\ |\mu_d - density_i| < \sigma_d \\ 0 & otherwise \end{cases} \tag{4.5}$$

We have noticed that superpixels with high edge-density value have less probability for belonging to an object. Thus, we use it as a negative prior in our energy formulation as mentioned in the next section.

Bottom row of the Figure 4.2 shows a superpixel-level image with the edge map superposed on it, along with boundedness cue and edge-density results on an example from PASCAL Segmentation dataset (Everingham *et al.* (2012)). In Figure 4.2(e), red boxes show the superpixels and white lines portrays the edge map. It can be seen that the strong edges mostly depicting the object boundaries, are mostly respected by superpixels as well. Exploiting the edge map we generate the boundedness map and edge-density map as illustrated in Figure 4.2(f) and 4.2(g) respectively.

4.3 Salient Object Segmentation

CRF (Conditional Random Fields) described in the work of Lafferty *et al.* (2001), has the ability to concisely represent dependencies among multiple random variables. Thus it can capture the structure of the problem efficiently. Hence, we formulate a CRF over superpixels to estimate the MAP (Maximum a Posteriori) value of each of the superpixels belonging to an object. Now each superpixel has three features, as computed in the previous section, along with color. In the following subsections, we discuss the CRF formulation and the label prediction task.

4.3.1 Preliminaries: Random Field Model

Let $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{N}$ be the feature vector set and $\mathbf{y}$ be the segmentation labels of all the superpixels, where N is the total number of superpixels. Conditional random field model takes the form,

$$P(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z} e^{-E(\mathbf{y},\mathbf{x};\mathbf{w})} \tag{4.6}$$

where $\mathbf{w}$ are parameters and Z is the partition function. The energy term E generally decomposes over nodes $\mathcal{V}$ (set of superpixels) and edges $\mathcal{E}$ (8-neighborhood of each of the superpixel). We consider the energy E with node and edge features as $\phi^{(1)}$ and $\phi^{(2)}$ respectively, resulting in the following formulation:

$$E(\mathbf{y}, \mathbf{x}; \mathbf{w}) = \sum_{i \in \mathcal{V}} \phi_i^{(1)}(y_i, \mathbf{x_i^{(1)}}; \mathbf{w_1}) + \lambda \sum_{(i,j) \in \mathcal{E}} \phi_{i,j}^{(2)}(y_i, y_j, \mathbf{x_i^{(2)}}, \mathbf{x_j^{(2)}}; \mathbf{w_2}) \tag{4.7}$$

where, $\mathbf{w_1}$ and $\mathbf{w_2}$ are the parameters in node and edge potential respectively. Node potentials represent negative log-likelihood. Thus we first compute different features, such as, saliency, boundedness, edge-density for the node potential or the unary term. We then define the edge cost or the pairwise smoothness term to fully specify the CRF.

Now we can consider the inference and learning tasks. The two tasks of our interest are: (i) the test-time prediction of labels of a likely segmentation and (ii)

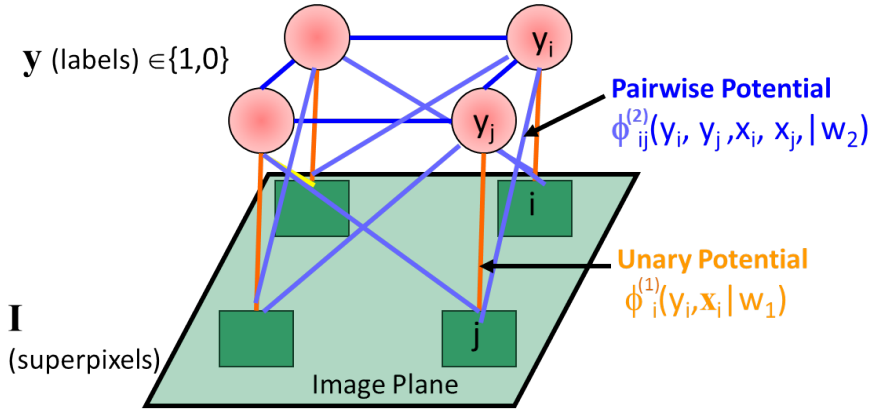


Figure 4.3: Graphical model showing a basic CRF model. Green boxes are superpixels and red circles represents the hidden layer of labels. Orange and blue lines depict corresponding node potentials and edge potentials respectively.

the parameter learning task in which we have annotated training data to compute an optimal set of parameters $\mathbf{w} = [\mathbf{w_1}\ \mathbf{w_2}]^T$. Figure 4.3 represents a basic CRF model. It pictorially illustrates the dependencies among the superpixels and the hidden layer of labels.

4.3.2 Salient Object Likelihood

The node potentials are obtained by combining the image cues that are defined in Section 4.2. The image features for node potential of ith node (superpixel) is denoted by $\mathbf{x_i^{(1)}}$ and the parameters by $\mathbf{w_1}$. s_i, b_i and ed_i are the image features, as described in equations (3.6), (4.1) and (4.5) respectively. These features correspond to the likelihood of a superpixel being part of an object. So, it penalizes when a superpixel with strong likelihood is assigned a background label or a less likely superpixel is assigned a foreground label. Here, a particular label $y_i \in \{0, 1\}$ is assigned for foreground and background respectively. Hence, the node potential is written as:

$$\begin{aligned}\phi_i^{(1)}(y_i, \mathbf{x_i^{(1)}}; \mathbf{w_1}) = \underbrace{w_s((1-y_i)(1-s_i) + y_i s_i)}_{saliency} + \underbrace{w_b((1-y_i)(1-b_i) + y_i b_i)}_{boundedness} \\ + \underbrace{w_e((1-y_i)ed_i + y_i(1-ed_i))}_{edge\ density}\end{aligned} \tag{4.8}$$

Hence, $\mathbf{w_1} = [w_s \ w_b \ w_e]^T$. The method to learn the optimal values of these parameters are discussed in section 4.3.4.

4.3.3 Edge Cost

Edge cost enforces agreement between adjacent superpixels in an image. If two adjacent superpixels are similar in appearance, they should have the same label, otherwise the objective function is penalized. $\mathbf{x}_i^{(2)}$ is the feature that accounts for the pairwise term of ith superpixel. Here $\mathbf{x}_i^{(2)}$ is color in Lab space, denoted by c_i. We express the pairwise or smoothness term as:

$$\phi_{i,j}^{(2)}(y_i, y_j, \mathbf{x_i^{(2)}}, \mathbf{x_j^{(2)}}) = |y_i - y_j| e^{-(-k_c||c_i - c_j||^2)} \tag{4.9}$$

Hence, similarity in Lab color space specifies the edge cost. Here, k_c dictates the sensitivity of color similarity and is given a constant value in all our experiments.

4.3.4 Superpixel Label Prediction: Inference Problem

Now that the random field is fully specified, we have two tasks, inference and parameter learning.

Inference

The edge-cost defined by our model leads to a sub-modular CRF and hence, we perform an exact inference using graph cut. This makes our method time efficient. With more complex edge-cost and approximate inference technique, the label prediction task will become computationally inefficient. To improve the results, we take a feedback from the first graph cut output and combine that with the unary term and again repeat the process (perform graph cut) to generate a better segmentation. In subsequent iterations, unary prior is taken as the exponentiated and normalized intersection of the original unary prior (as calculated in eqn. 4.8) and the result of graph cut from previous iteration. Hence, the unary prior is refined by the predicted labels in each iteration, by using the graph cut result as a feedback for better prediction. Experimentally, a maximum of 8 iterations are

performed for all images.

Parameter Learning

As per our formulation, the parameter vector $\mathbf{w}$ can be considered as $\mathbf{w} = [\mathbf{w_1}\ \mathbf{w_2}]^T = [w_s\ w_b\ w_e\ \lambda]^T$. This gives us a energy function E which is linear in $\mathbf{w}$ and can be written as $\mathbf{w}^T\phi$. This is important because linearity in $\mathbf{w}$ ensures a convex learning problem and can be solved efficiently. We take a simple max-margin approach for parameter learning, given as follows:

$$\begin{aligned} &minimize\ \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{M}\sum_{n=1}^{M}\xi_n \\ &subject\ to,\ \mathbf{w}^T\mathbf{v} = 1 \\ &\qquad \mathbf{w}^T\phi_n - \mathbf{w}^T\hat{\phi}_n \leq \mathcal{L}(\mathbf{y}_n, \hat{\mathbf{y}}_n) + \xi_n \\ &\qquad \xi_n \geq 0 \\ &\qquad w_2 > 0 \end{aligned} \tag{4.10}$$

where,

$\mathcal{L}(\mathbf{y}_n, \hat{\mathbf{y}}_n) = \frac{False\ Positive + False\ Negative}{True\ Positive + False\ Positive + False\ Negative}$, is the loss function.

n ranges over the M training instances, $\hat{\mathbf{y}}$ is the ground truth labeling and $\mathbf{w}^T\hat{\phi}$ represents the ground truth energy. ξ represents the slack variable (refer to appendix B.2 for details), and $\mathbf{v} = [1\ 1\ 1\ 0]^T$. The formulation is a structured learning approach. It is similar to the structured SVM approach (Tsochantaridis *et al.* (2005), see appendix B.1) and follows the margin rescaled algorithm (Szummer *et al.* (2008)). The second part of the objective function in equation 4.10, gives a penalty if the calculated parameters do not lead to an energy close to ground truth, which ideally should have minimum energy. Thus, minimizing this objective function leads to estimating the optimal parameters for which the energy for predicted labels is close to ground truth energy. The first constraint normalizes $\mathbf{w}$. This is necessary so that the unary term alone does not dominate the summation cost function, given in eqn. 4.7. It also gives a lower bound to $\mathbf{w}$. $w_2 > 0$ makes sure that the problem remains submodular and we can hence do an exact inference via graph cut. This is an important point for the efficiency of the algorithm.

Since, we are using a high-order loss function $\mathcal{L}$ (1- IoU), in equation 4.10, the margin-rescaled structured SVM formulation is not a simple Quadratic Programming (QP) problem any more, unlike normal SVM formulation. So, to solve the problem, we take a simple iterative approach similar to EM (Estimation-Maximization). In the following algorithm we give step-by-step details of our approach.

Algorithm 4.1 Finding optimal **w**

Input:

- input-labeling pairs $\{(x_n, y_n)\}$ of M training instances

- initial parameters: $w = w^0$

Repeat until w is unchanged (within a tolerance)

1. Run graph cuts to find the MAP labeling $\mathbf{y_n}$
2. Estimate the loss function $\mathcal{L}$
3. Given $\mathcal{L}$, update the parameters w by optimizing the formulation in equation 4.10

Since, the objective function as well as all the constraints are convex in step 3, we have used cvx toolbox (Grant and Boyd (2008)) to solve the optimization problem.

After we perform inference, we obtain a binary map at superpixel level. This superpixel map can be thought as a low-resolution image and we upsample it to full resolution pixel accurate map (as described in Dolson *et al.* (2010)). Now each pixel has a label $\in [0, 1]$. Pixel label values depict the probability value of that pixel belonging to a salient object. The top right image in Figure 4.1 shows an example of our proposed upsampled map. The upsampling algorithm is again a fast implementation and performs in linear time. We threshold this upsampled salient object probability map to generate a salient object segmentation mask. Refer to section 3.3.3 and appendix A.2 for details of upsampling method. Here, the threshold is taken as the median of maximum and minimum probability values. This method of producing the resultant masks are used for all the experiments in Section 4.4 and a few results are presented in Figure 4.4.

4.4 Experiments and Results

We measure the performance of our approach and compare the same with recent saliency detection as well as object proposal methods, on both object segmentation and saliency datasets. We evaluate the performance using both F-measure and intersection-over-union score. All the results presented are obtained using the optimal learned parameter set, generated using the optimization process in algorithm 4.1 solving equation (4.10).

4.4.1 PASCAL Segmentation Dataset

We use the segmentation part of the PASCAL VOC 2012 dataset (Everingham *et al.* (2012)) to evaluate the proposed method. It has a segmentation part which has 2,913 images with object specific segmentation ground truth. To extract all the objects in an image we generate a binarized ground truth map as the second image in the top row of Figure 4.1 shows. We perform training on 1,464 images in the training set and testing is done on 1,449 images from the validation set.

Qualitative results on images from this dataset are presented in Figure 4.4. VOC 2012 has images from 20 different classes, from which we have shown images from 12 different classes to illustrate the performance. The figure shows that our method performs better than the recently proposed saliency methods, namely, SF (Perazzi *et al.* (2012)), MR (Yang *et al.* (2013)), PARAM (proposed in Section 3.3) and top 3 masks of CPMC (Carreira and Sminchisescu (2010)), OP (Endres and Hoiem, 2010). Examples in first five rows clearly depicts the superiority of the segmentation by the proposed method. All the methods fails to perform well on the samples in last three rows. The example in the last row shows a failure case of our approach too. In this case, objects are not salient by the features used in the saliency methods. In addition, as the edge map mostly captures the circles in the stand at the front, the boundedness and the edge-density cues also fail. Again, it is qualitatively visible that precision of CPMC and OP is generally very low even for top ranked masks and same has been found in the quantitative analysis presented in the next subsection.

Figure 4.4: Visual results of our Salient Object Segmentation method and different saliency and low-level object proposal methods, on some samples of PASCAL VOC 2012 segmentation dataset (Everingham *et al.*, 2012). It clearly demonstrates the superiority of our segmentation. GT denotes the binarized ground truth masks.

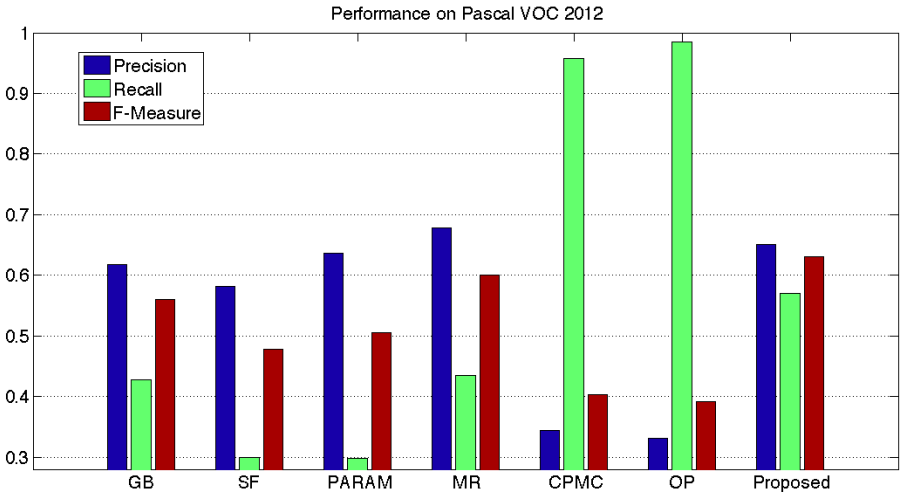


Figure 4.5: Precision Recall F-measure on PASCAL VOC 2012 segmentation dataset Everingham *et al.* (2012). It demonstrate that proposed method keeps a good Precision-Recall balance and outperform in terms of F-measure.

4.4.2 F-measure and Intersection-over-Union Score

F-measure (Eqn. 3.10) score is widely used in all saliency detection methods where β is taken as 0.3 (as in Perazzi *et al.* (2012); Achanta *et al.* (2009)). Figure 4.5 shows the precision-recall-fmeasure values on VOC 2012 segmentation dataset (Everingham *et al.*, 2012), for different competing saliency methods. We also compare with the category independent object proposal methods, viz, CPMC (Carreira and Sminchisescu (2010)) and OP (Endres and Hoiem (2010)) (where top 10 masks are used). Competing saliency methods used are GB (Harel *et al.* (2006)), SF (Perazzi *et al.* (2012)), MR (Yang *et al.* (2013)) and PARAM (Section 3.3) for comparison. Clearly, in figures, in terms of F-measure our proposed method outperforms all the other methods. In terms of precision and recall also our method is comparable to very recent saliency methods Yang *et al.* (2013) and our saliency detection method PARAM. CPMC and OP give high recall values as they generate a number of maps, but are very low on precision. This implies that they propose a lot of non-object parts of an image as object regions, even when just the top 10 maps are considered (see Figures 4.1 and 4.5).

Intersection-over-Union (IoU) score is computed as,

$$IoU = \frac{|Predicted\ Map \cap Ground\ Truth|}{|Predicted\ Map \cup Ground\ Truth|}$$

Method Name	IoU Score
CPMC	0.3319
OP (Object Proposal)	0.3266
Proposed	**0.4097**

Table 4.1: Intersection-over-Union score of top 10 object maps of category independent generic object segmentation methods, viz., CPMC Carreira and Sminchisescu (2010), OP Endres and Hoiem (2010) and our proposed method of Salient object segmentation, on the PASCAL 2012 segmentation dataset.Our proposed method produces much better segmentation results.

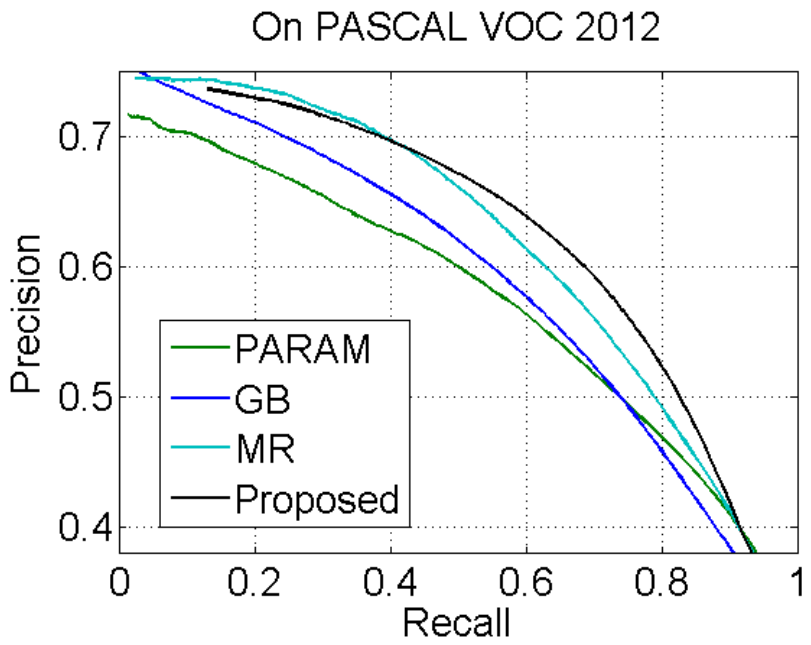


Figure 4.6: Precision Recall Curve of proposed method and other competing saliency methods on PASCAL VOC 2012 dataset (Everingham *et al.* (2012)).

We compute the IoU score of CPMC, OP and our method against the binarized ground truth on PASCAL segmentation dataset. Results are presented in Table 4.1 . We use the 10 top-ranked maps of CPMC and OP to compute the IoU score, considering all the objects. Table 4.1 shows that the single map of our method produces 21% better object segmentation maps in terms of IoU score, compared to the object proposal methods.

Figure 4.6 illustrates the performance for different methods with a Precision-Recall curve (see Section 3.4.2 for process of generation) on the PASCAL VOC 2012 (Everingham *et al.* (2012)) dataset. We compare the saliency map of the competing methods, such as, the saliency method proposed in Chapter 3 (PARAM),

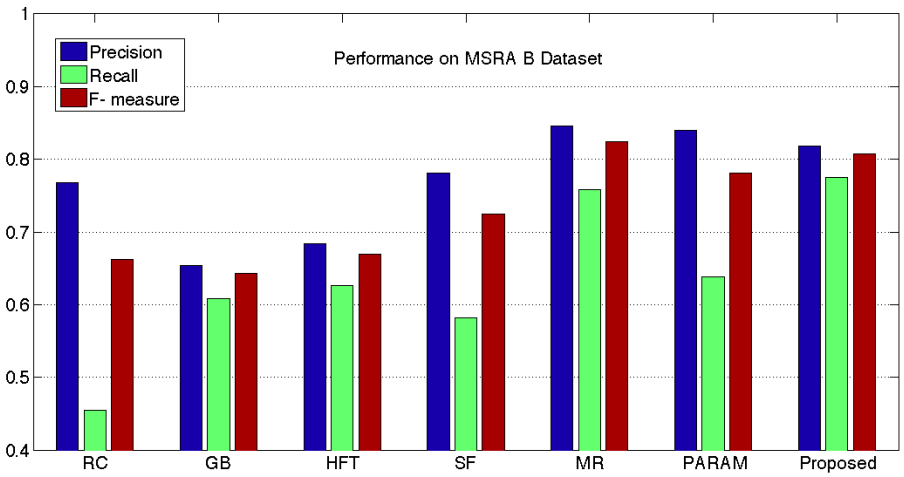


Figure 4.7: Precision Recall F-measure for proposed method and six other saliency methods on MSRA-B saliency dataset (Achanta *et al.* (2009)).

MR (Yang *et al.* (2013)) and GB (Harel *et al.* (2006)) with our CRF-based upsampled map. Our proposed method gives visibly better precision than all other methods for higher values of recall. For very low values of recall MR (Yang *et al.* (2013)) is marginally better. This is also reflected in the bar plot in Figure 4.5. This can be interpreted as, if MR (Yang *et al.* (2013)) attempts to find most of the salient object regions (i.e., for high recall), it incorporates a lot of false positives (i.e., the precision falls).

We do not compare with a few other detection proposal methods, such as Objectness (Alexe *et al.* (2012)), SelectiveSearch (van de Sande *et al.* (2011)) or BING (Cheng *et al.* (2014)), as they only generate bounding box proposals and do not produce fine segmentation (foreground masks) results. SelectiveSearch (van de Sande *et al.* (2011)) emphasizes on recall and does not intend for a pixel-accurate map as they aim for an object recognition application.

4.4.3 Performance on Saliency Dataset

We also compare the performance with recent state-of-the-art saliency techniques also, on a saliency dataset (as in section 3.4.2). For this purpose we use MSRA-B dataset (Achanta *et al.* (2009)). It has 5000 images with publicly available pixel accurate binary ground truth masks. Figure 4.7 shows the precision-recall-fmeasure values on MSRA-B dataset, for our method along with that for other recent saliency methods, viz, RC (Cheng *et al.* (2011)), GB (Harel *et al.* (2006)),

HFT (Li *et al.* (2013)), SF (Perazzi *et al.* (2012)), MR (Yang *et al.* (2013)) and PARAM (Section 3.3). It illustrates the result on MSRA-B dataset with the parameters learned using PASCAL segmentation dataset. Although, the recent saliency method MR performs marginally better with little higher precision, it provides low recall when tested on natural image dataset like PASCAL segmentation dataset (see Figure 4.4), i.e., it fails to extract many object regions and thus not always the best for the current task.

4.4.4 Computational Efficiency

Our proposed method of Salient Object Segmentation is deemed to work as a pixel-level pre-processing step for different computer vision tasks and must be suitable for a live system. Hence, computational efficiency is of real importance. However, since related category independent object proposal methods use much complex procedure of generating a bag of outputs from different seeds and ranking them, they are less time efficient. Our proposed method has three components to be computed for the prediction task, viz., saliency, edge map and objectness cues. All of these happen in either order of pixels or even less (order of superpixels). The edge map (Dollár and Zitnick, 2013) extraction method has described time efficiency in their paper, and is completely suitable for a real-time or online system. Also, the proposed saliency method (PARAM) described in Section 3.4.2, uses fast computations as proposed by Perazzi *et al.* (2012) and does all the operations at superpixel level and is thus making it computationally efficient. Our computation of boundedness and edge-density is in order of number of pixels, compared to the objectness cues of Alexe *et al.* (2012) which are costly. Hence, our method is time efficient and is suitable as a precomputing technique.

We avoided a direct run-time comparison as shown in the previous chapter (Table 3.1), as it may be inappropriate to compare with methods such as CPMC (Carreira and Sminchisescu (2010)) or OP (Endres and Hoiem (2010)). These methods produce around 1000s of binary map proposals, whereas our method is way much faster and produces a single map. *Average run-time* required per image (on the PASCAL segmentation dataset) for the entire pipeline of our proposed

method is **0.42** secs, whereas the same for OP is 3 min 16 secs (using i7, 16GB RAM). Complexity of Graph-cut is $O(V * E^2) = O(8^3.V^3)$ [8-neighborhood is considered] = $O((\text{number of superpixels})^3)$.

Hence, complete run-time complexity can be given as:

Run-time for (PARAM) saliency + Objectness criteria + grpah-cut minimization = O(number of superpixels) + O(number of pixels) + $O((\text{number of superpixels})^3)$ = $O((\text{number of superpixels})^3)$ (approx).

In comparison, the unary term or 'affinity' calculation for each seed itself in OP (Endres and Hoiem (2010)) is much complex (larger) than O(number of pixels) depending upon the number of regions and boundaries present in the image.

4.5 Discussion

We attempt to solve the problem of category independent salient object segmentation using a multi-criteria objective function. We propose a time efficient approach which performs better than the recent state-of-the-art methods. Motivated by saliency and category independent object segmentation methods, we propose to predict a segmentation which captures all the salient objects in an image. We device two objectness factors which are computed in linear time and used with saliency as the priors. We demonstrate that graph-based methods can be used efficiently both in terms of inference and learning parameters. Proposed method can be easily utilized as a pre-processing step for many high-level computer vision tasks.

CHAPTER 5

Conclusion

In the thesis, we have described two competent salient object detection techniques. In the following, we summarize the two efficient methods and discuss how they have contributed to the recent progress in Computer Vision area. We discuss possible applications and important future extensions possible from our work.

5.1 Thesis Summary

In summary, we propose two novel image saliency algorithms. The first method concentrates in finding high-level saliency information. It shows that rarity and compactness of a color are visually important cues to attract human attention. It explores another important factor known as boundary prior. Further, it shows that the rarity based measures and boundary prior are two complementary components in detecting salient object in images.

The second method utilizes object-level constraints over saliency prior devised in first method. Then the consistency among salient regions are enforced by a proper spatial propagation of the saliency and our novel objectness information, using a CRF with image elements (superpixels) as nodes and their exponentially weighted color difference as edge-weights. Hence, the unary terms are designed using saliency and objectness features and pairwise term depicts the color similarity based edge-weights. CRF parameters are learned using a max-margin approach.

Results on three real-world datasets exhibit the superiority of proposed methods. These methods can provide significant advantage in many high-level Computer Vision tasks, such as object recognition, object detection, video segmentation and retrieval, summarization, re-identification etc. Unsupervised bottom-up algorithms to detect salient regions (segments) in images will act as a good initializer for fast convergence during learning activities from videos, object shapes, scene recognition, posture identification and so on.

5.1.1 Limitations

One main limitation comes from the issue related to dataset bias. Every dataset has some bias of feature distributions, depending upon the different sets of image samples it contains. For example (comparing the images from MSRA-B and PASCAL VOC datasets), in images from MSRA-B the objects are mostly homogeneous in color and occupy a large portion of the image. Whereas, images from PASCAL VOC dataset contain objects which may be visible in small parts or a single object is present with lot of variability in intensity at different parts. Hence, evidently the set of parameters that are learned on the PASCAL VOC dataset may not help the saliency detection algorithms to perform well on images in MSRA-B dataset, compared to some method which targets to learn the parameter set from MSRA-B dataset itself and vice-versa. In addition to that, since saliency is the major focus of our algorithms, as per definition of saliency in section 1.1, proposed methods may fail to catch the objects which are less visible to the human eye. Specifically, very small part of an object present at the boundary of an image, ignored by the human vision, may also be ignored by our proposed methods. Again, we rely on boundedness of the superpixels; that is, a superpixel belonging to an object is likely to have contours surrounding it. Hence, very thin parts of an object, such as spokes of a cycle wheel, may also be ignored. But the optimistic point is, although proposed methods may fail to capture all the objects (specially which are slightly visible or not clearly visible due to less contrast) in an image or internals of an object (some parts such as cycle spokes), it produces correct results for the important or visible objects which also would be visible to a human eye and also gives a good estimate of the holistic shape of the object. This is very important for further processing of the image for different tasks, e.g., object recognition and detection.

5.2 Some Reflections and Future Work

There has been a paradigm shift in Computer Vision in recent years. Researchers are inclined towards developing algorithms similar to human visual perception. These approaches have been shown to perform more accurately and intuitively.

Segmentation aimed towards object detection has become popular. Intelligent CBIR, CVBR tasks often need fast and efficient detection of objects in images or videos. This is where saliency has been found to be very useful and similar to the human visual system. Our first method follows these lines of thought. Again to solve this problem of detecting objects from an image, segmentation or generating object proposal is a natural solution.

Hence, a fundamental shift in object detection approaches have occurred. Object detection algorithms have started to use segmentation and object proposals instead of the costly sliding window based approach. So, we correctly employ salient and objectness feature and contribute to these kind of novel approaches for detecting the object of interest. In addition to detecting the salient object in an image, our proposed segmentation based approaches can be aptly used for detecting the boundary or salient contour of an object (useful for sketch based retrieval).

Our methods and approaches can be further extended in different directions. Firstly, our method of saliency map generation (refer Section 3.3) can also be used to find the most salient part of an image or finding the important fixation regions for different applications, such as placing advertisement in a video frame. Secondly, our object level saliency map (refer Section 4.3) can be refined to generate a small set of accurate object proposals so as to capture all the different objects in an image and hence can capture their boundary information. The boundary information can be used for shape based processing of images.

Extension to a top-down saliency model can form a nice scope of future work. Here the system needs a pre-defined task or goal and a few training samples under each category of objects. Task specific top-down saliency (as in humans) will need large research effort from the academic community. Here the context and environment dictates the goal, e.g., locate food if you are hungry or if you have lost your key, look for small metallic objects. Training using eye-fixation data forms an integral part of such top-down saliency detection modules. Hence the methods are generally supervised, unlike our proposed bottom-up unsupervised algorithms.

As a final thought, most supervised methods get biased on the type of training

dataset being used. Making an unbiased saliency detection process by incorporating deep learning, DA (domain adaptation), transfer learning based methods, or try using "ImageNet" dataset with partially hand-labeled or weakly labeled ground truth, may form some future directions of work, which need a lot of analytical and experimental studies. Future researchers may need to estimate attention, saliency and recognition under a unified umbrella for the design of an automatic visual registration system.

APPENDIX A

Image Processing Techniques

A.1 Superpixels

For all our approaches mentioned in Chapter 3 and 4, we first abstract the image into superpixels or perceptually uniform regions. The image abstraction we use is an adaption of SLIC algorithm (Achanta *et al.* (2010)) and was proposed in the work of Perazzi *et al.* (2012). Superpixels are locally compact edge aware clusters of pixels and the adaptation guarantees connectivity also.

A.1.1 SLIC Superpixels

The SLIC algorithm (Achanta *et al.* (2010)) generates superpixels by clustering pixels based on their color similarity and proximity in the image plane and uses a simple linear iterative clustering algorithm, very similar to K-means, in the five-dimensional *[labxy]* space. Given input K, start with K superpixel cluster centers in 5D space with k = [1,K] at regular grid intervals and the distance measure is given by sum of Euclidean distance in 3D *lab* space and 2D *xy* space, normalized by the grid interval.

Given a desired number of approximately equally-sized superpixels K for an image with N pixels, roughly there will be superpixel center at every grid interval $S = \sqrt{N/K}$. Initially K regularly spaced cluster centers are sampled and seeds are placed at the location corresponding to the lowest gradient position in a 3 neighborhood, to avoid placing the seeds on edge. Image gradients are computed taking both color and intensity information into account:

$$G(x, y) = ||\mathbf{I}(x + 1, y) - \mathbf{I}(x - 1, y)||^2 + ||\mathbf{I}(x, y + 1) - \mathbf{I}(x, y - 1)||^2 \qquad \text{(A.1)}$$

where $\mathbf{I}(x, y)$ is the *lab* vector of pixel at (x, y). Now all the pixels are clustered

based on these seeds or cluster centers in an approach similar to K-means algorithm.

For the distance measure in clustering Euclidean distance in *Lab* color space is the perceptually most meaningful measure, but only for small distance. If spatial distance is so high that it exceeds this perceptual color distance limit, it overweights color similarity and as a result it may create superpixels which do not respect region boundaries. Therefore, instead of using simple Euclidean distance, distance measure D_S is defined as:

$$\begin{aligned} D_S &= d_{lab} + \frac{m}{S} d_{xy} \\ where, & \\ d_{lab} &= \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2} \\ d_{xy} &= \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \end{aligned} \tag{A.2}$$

Hence, D_S or the distance is taken as sum of *lab* distance and normalized *xy* distance. Here m controls the compactness of a superpixel and given a constant value ($m = 10$). This keeps good balance between color similarity and spatial proximity. To segment into superpixels iteratively cluster centers are updated until the residual error (L1 distance between prevu=ious centers and recomputed centers) goes below a certain threshold. Finally, connectivity within the superpixel is ensured by relabeling disjoint superpixels with the labels of largest neighboring cluster.

A.2 Up-sampling

Every time we abstract the image into superpixels after desired computation it needs to be computed at pixel level so as to generate a complete full resolution output map. This is done by upsampling algorithm. A superpixeled image can be thought as a low resolution image and we upsample it to pixels. It basically does a d dimensional filtering which can be mathematically expressed as:

$$\hat{c_i} = \sum_{j=1}^{n} f(|p_i - p_j|).c_i \tag{A.3}$$

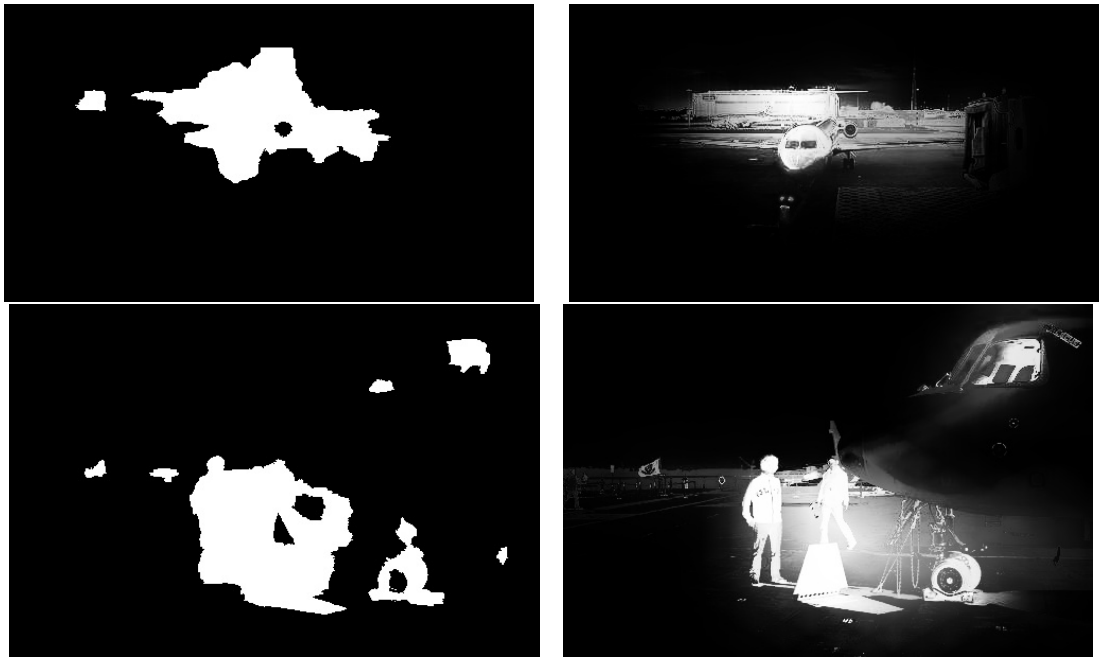


Figure A.1: Some example of upsampling to illustrate the importance of upsampling. Left column shows image before upsampling and right column has the corresponding upsampled images.

where, $\hat{c_i}$ is the color value of pixel i with surrounding image elements $j = \{1, \ldots, n\}$ and p_i, c_i are the pixel position and color respectively. In the most common case, function f can be any kernel. Most commonly used in case of denoising is, $f(x) = e^{-\frac{|x|^2}{2\sigma^2}}$, the standard Gaussian function with a standard deviation σ. Moreover, it uses a accelerated d dimensional filter as proposed by Adams *et al.* (2010).

The upsampling algorithm (Dolson *et al.* (2010)) can be summarized in three steps:

1. 5-D feature (RGBXY) are sampled at the resolution of the image
2. and then blurred with values of neighboring node
3. Then value of each pixel is queried using location, along the original position space

This pipeline greatly accelerates a bilateral filter and also Adams *et al.* (2010) show that a 5D representation is more accurate than just filtering 3D illuminance volume. Some example of upsampling from Chapter 4 are presented in Figure A.1. This illustrates result of our method before and after upsampling.

APPENDIX B

Structured Prediction

While working on posteriors and building autonomous systems, we not only want to sense the environment but also recognize and interact with it. A structured prediction on MRF is a fantastic mathematical tool which just not try to solve an individual task, but tries to give a solution of all these tasks together in a holistic way. This is important because we want to propagate uncertainty of these different tasks instead of only making a single decision. There are four things that we need for a structured prediction, data, good learning and inference algorithm and proper representation. Representation is very important, that is, what are the random variables and connection among them.

B.1 Max-Margin Method: Structured SVM

A general formulation (Ramanan, 2013) of a max-margin method or SVM which encompasses various common problems, such as binary classification, regression, and structured prediction. Given training data where the i^{th} example is described by a set of N_i vectors $\{x_{ij}\}$ and a set of N_i scalars $\{l_{ij}\}$, where j varies from 1 to N_i, we wish to solve the following optimization problem:

$$\arg\min_{w} L(w) = \frac{1}{2}||w||^2 + \sum \max_{j \in N_i}(0, l_{ij} - w^T x_{ij}) \tag{B.1}$$

This can also be written as the following Quadratic Program (QP) as:

$$\begin{aligned} &\arg\min_{w,\xi} \ \frac{1}{2}||w||^2 + \sum_{i}^{N} \xi_i \\ &s.t. \ \forall i, j \in N_i, \ \ w^T x_{ij} > l_{ij} - \xi_i \end{aligned} \tag{B.2}$$

B.1.1 Structured SVM

A linearly-parametrized structural predictor produces a label of the form

$$Label(x) = \arg\max_{y \in Y} w^T \phi(x_i, y) \tag{B.3}$$

where Y represents a (possibly exponentially-large) structured output space. The associated learning problem is given by a dataset $\{x_i, y_i\}$ where $x_i \in R^N$ and $y \in Y$:

$$\begin{aligned} &\arg\min_{w,\xi \geq 0} \frac{1}{2}||w||^2 + \sum_i \xi_i \\ &s.t\ \forall i, h \in Y,\ \ w^T \phi(x_i, y_i) - w^T \phi(x_i, h) \geq loss(y_i, h) - \xi_i \end{aligned} \tag{B.4}$$

One can define $N_i = |Y|$, $x_{ij} = \phi(x_i, y_i)\phi(x_i, j)$ and $l_{ij} = loss(y_i, j)$, where $j = h$ is interpreted as an index into the output space Y.

B.2 Margin-Rescaled Approach

The first approach that was proposed (Tsochantaridis *et al.* (2005)) for the case of arbitrary loss functions, is to re-scale the slack variables according to the loss incurred in each of the linear constraints. Intuitively, violating a margin constraint involving a $y \neq y_i$ with high loss $\Delta(y_i, y)$ should be penalized more severely than a violation involving an output value with smaller loss. This can be accomplished by multiplying the margin violation by the loss, or equivalently, by scaling the slack variable with the inverse loss. The slack-rescaled approach is formulated as:

$$\begin{aligned} &\arg\min_{w,\xi \geq 0} \frac{1}{2}||w||^2 + \sum_i \xi_i \\ &s.t\ \forall i, h \in Y,\ \ w^T \phi(x_i, y_i) - w^T \phi(x_i, h) \geq 1 - \frac{\xi_i}{\Delta(y_i, y)} \end{aligned} \tag{B.5}$$

This formulation has the advantage that the loss δ and slack penalty share the same scale. In particular, it enforces the same default margin of 1 for all examples. In contrast, the margin-rescaled formulation (Tsochantaridis *et al.* (2005), and

was proposed by B. Taskar and Koller (2004) for the special case of the Hamming loss) requires large margins of labelings differing significantly from the ground truth, which could cause the algorithm to focus on assigning high energies to poor labelings, rather than assigning low energies to labelings close to ground truth. The margin constraints for margin-rescaled formulation, thus becomes the following:

$$\forall i, h \in Y, \;\; w^T \phi(x_i, y_i) - w^T \phi(x_i, h) \geq \Delta(y_i, h) - \xi_i \tag{B.6}$$

As high order loss functions are used, the problem remain no longer QP and becomes intractable. Many different approximate and iterative solutions are given in this context (Tarlow and Zemel (2011); Pletscher and Kohli (2012)).

REFERENCES

1. **Achanta, R.**, **S. Hemami**, **F. Estrada**, and **S. Susstrunk**, Frequency-tuned salient region detection. *In CVPR*. 2009.

2. **Achanta, R.**, **A. Shaji**, **K. Smith**, **A. Lucchi**, **P. Fua**, and **S. Susstrunk** (2010). SLIC Superpixels. Technical report, EPFL.

3. **Adams, A.**, **J. Baek**, and **M. A. Davis**, Fast high-dimensional filtering using the permutohedral lattice. *In Computer Graphics Forum (EG 2010 Proceedings*. 2010.

4. **Alexe, B.**, **T. Deselaers**, and **V. Ferrari** (2012). Measuring the objectness of image windows. *TPAMI*, **34**(11), 2189–2202.

5. **Arbelaez, P.**, **M. Maire**, **C. Fowlkes**, and **J. Malik** (2011). Contour detection and hierarchical image segmentation. *TPAMI*, **33**(5), 898–916.

6. **B. Taskar, C. G.** and **D. Koller**, Max-margin markov networks. *In NIPS*. 2004.

7. **Borji, A.** and **L. Itti** (2013). State-of-the-art in visual attention modeling. *TPAMI*, **35**(1), 185–207.

8. **Carreira, J.** and **C. Sminchisescu**, Constrained parametric min-cuts for automatic object segmentation. *In CVPR*. 2010.

9. **Chang, K.-Y.**, **T.-L. Liu**, **H.-T. Chen**, and **S.-H. Lai**, Fusing generic objectness and visual saliency for salient object detection. *In ICCV*. IEEE, 2011.

10. **Cheng, M.-M.**, **G.-X. Zhang**, **N. J. Mitra**, **X. Huang**, and **S.-M. Hu**, Global contrast based salient region detection. *In CVPR*. 2011.

11. **Cheng, M.-M.**, **Z. Zhang**, **W.-Y. Lin**, and **P. Torr**, Bing: Binarized normed gradients for objectness estimation at 300fps. *In CVPR*. 2014.

12. **Dollár, P.** and **C. L. Zitnick**, Structured forests for fast edge detection. *In ICCV*. 2013.

13. **Dolson, J.**, **B. Jongmin**, **C. Plagemann**, and **S. Thrun**, Upsampling range data in dynamic environments. *In CVPR*. 2010.

14. **Endres, I.** and **D. Hoiem**, Category independent object proposals. *In ECCV*. Springer, 2010.

15. **Everingham, M.**, **L. Van Gool**, **C. K. I. Williams**, **J. Winn**, and **A. Zisserman** (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

16. **Felzenszwalb, P. F.**, **R. B. G.**, **D. McAllester**, and **D. Ramanan** (2010). Object detection with discriminatively trained part-based models. *TPAMI*, **32**(9), 1627–1645.

17. **Felzenszwalb, P. F.** and **D. P. Huttenlocher** (2004). Efficient graph-based image segmentation. *IJCV*, **59**(2), 167–181.

18. **Goferman, S.**, **L. Zelnik-manor**, and **A.Tal**, Context-aware saliency detection. *In CVPR*. 2010.

19. **Goodale, M. A.**, **A. D. Milner**, **L. Jakobson**, and **D. Carey** (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, **349**(6305), 154–156.

20. **Gopalakrishnan, V.**, **Y. Hu**, and **D. Rajan** (2010). Random walks on graphs for salient object detection in images. *Image Processing, IEEE Transactions on*, **19**(12), 3232–3242.

21. **Grant, M.** and **S. Boyd** (2008). Cvx: Matlab software for disciplined convex programming.

22. **Harel, J.**, **C. Koch**, and **P. Perona**, Graph-based visual saliency. *In NIPS*. 2006.

23. **Hou, X.**, **J. Harel**, and **C. Koch** (2012). Image signature: Highlighting sparse salient regions. *TPAMI*, **34**(1), 194–201.

24. **Hou, X.** and **L. Zhang**, Saliency detection: A spectral residual approach. *In CVPR*. 2007.

25. **Itti, L.**, **C. Koch**, and **E. Niebur** (1998). A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, **20**(11), 1254–1259.

26. **Jia, Y.** and **M. Han**, Category-independent object-level saliency detection. *In ICCV*. IEEE, 2013.

27. **Jiang, H.**, **J. Wang**, **Z. Yuan**, **Y. Wu**, **N. Zheng**, and **S. Li**, Salient object detection: A discriminative regional feature integration approach. *In CVPR*. 2013.

28. **Jiang, Z.** and **L. S. Davis**, Submodular salient region detection. *In CVPR*. IEEE, 2013.

29. **Judd, T.**, **K. Ehinger**, **F. Durand**, and **A. Torralba**, Learning to predict where humans look. *In CVPR*. 2007.

30. **Koch, C.** and **S. Ullman** (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. **188**, 115–141.

31. **Lafferty, J. D.**, **A. McCallum**, and **F. C. N. Pereira**, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In ICML*. 2001.

32. **Li, J.**, **M. D. Levine**, **X. An**, **X. Xu**, and **H. He** (2013). Visual saliency based on scale-space analysis in the frequency domain. *TPAMI*, **35**(4), 996–1010.

33. **Li, Y.**, **X. Hou**, **C. Koch**, **J. Rehg**, and **A. Yuille**, The secrets of salient object segmentation. *In CVPR*. 2014.

34. **Liu, T.**, **Z. Yuan**, **J. Sun**, **J. Wang**, **N. Zheng**, **X. Tang**, and **H. Shum** (2011). Learning to detect a salient object. *TPAMI*, **33**(2), 353–367.

35. **Ng, A. Y.**, **M. I. Jordan**, and **Y. Weiss**, On spectral clustering: Analysis and an algorithm. *In NIPS*. 2001.

36. **Nowozin, S.** and **L. C. H.** (2011). Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*.

37. **Perazzi, F.**, **P. Krahenbuhl**, **Y. Pritch**, and **A. Hornung**, Saliency filters: Contrast based filtering for salient region detection. *In CVPR*. 2012.

38. **Pletscher, P.** and **P. Kohli**, Learning low-order models for enforcing high-order statistics. *In International Conference on Artificial Intelligence and Statistics*. 2012.

39. **Ramanan, D.** (2013). Dual coordinate solvers for large-scale structural svms. *arXiv preprint arXiv:1312.1743*.

40. **Rother, C.**, **V. Kolmogorov**, and **A. Blake** (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. **23**, 309–314.

41. **Roy, S.** and **S. Das** (2014). Saliency detection (PARAM) VISAPP 2014. http://www.cse.iitm.ac.in/ sudeshna/VISAPP_Project_page/ index.html.

42. **Shi, J.** and **J. Malik** (2000). Normalized cuts and image segmentation. *TPAMI*, **22**(8), 888–905.

43. **Szummer, M.**, **P. Kohli**, and **D. Hoiem**, Learning crfs using graph cuts. *In ECCV*. 2008, 582–595.

44. **Tarlow, D.** and **R. Zemel**, Big and tall: Large margin learning with high order losses. *In Workshop on Inference in Graphical Models with Structured Potentials, CVPR*. 2011.

45. **Tatler, B. W.** (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, **7**(14).

46. **Tomasi, C.** and **R. Manduchi**, Bilateral filtering for gray and color images. *In ICCV*. 1998.

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

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

49. **Tseng, P.-H.**, **R. Carmi**, **I. G. Cameron**, **D. P. Munoz**, and **L. Itti** (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*, **9**(7), 4.

50. **Tsochantaridis, I.**, **T. Joachims**, **T. Hofmann**, and **Y. Altun** (2005). Large margin methods for structured and interdependent output variables, 1453–1484.

51. **Tsotsos, J. K.**, **S. M. Culhane**, **W. Y. Kei Wai**, **Y. Lai**, **N. Davis**, and **F. Nuflo** (1995). Modeling visual attention via selective tuning. *Artificial intelligence*, **78**(1), 507–545.

52. **van de Sande, K. E. A.**, **J. R. R. Uijlings**, **T. Gevers**, and **A. W. M. Smeulders**, Segmentation as selective search for object recognition. *In ICCV*. 2011.

53. **Viola, P.** and **M. J. Jones** (2004). Robust real-time face detection. *IJCV*, **57**(2), 137–154.

54. **Wei, Y.**, **F. Wen**, **W. Zhu**, and **J. Sun**, Geodesic saliency using background priors. *In ECCV*. 2012.

55. **Yang, C.**, **L. Zhang**, **H. Lu**, **X. Ruan**, and **M.-H. Yang**, Saliency detection via graph-based manifold ranking. *In CVPR*. 2013.

56. **Yang, J.** and **M.-H. Yang**, Top-down visual saliency via joint crf and dictionary learning. *In CVPR*. 2012.

57. **Zhang, L.**, **M. H. Tong**, and **e. a. Marks, T. K.** (2008). Sun: A bayesian framework for saliency using natural statistics. *JOURNAL OF VISION*, **8**(32).

58. **Zhou, D.**, **O. Bousquet**, **T. Lal**, **J. Weston**, and **B. Scholkopf**, Learning with local and global consistency. volume 16. 2004*a*.

59. **Zhou, D.**, **J. Weston**, **A. Gretton**, **O. Bousquet**, and **B. Schlkopf**, Ranking on data manifolds. *In NIPS*. 2004*b*.

60. **Zhu, W.**, **S. Liang**, **Y. Wei**, and **J. Sun**, Saliency optimization from robust background detection. *In CVPR*. 2014.

LIST OF PAPERS BASED ON THESIS

1. Sudeshna Roy and Sukhendu Das "Multi-criteria Energy Minimization with Boundedness Edge-density and Rarity for Object Saliency in Natural Images" *ACM proceedings of Ninth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, IISC Bangalore, Karnataka, India, December 14-18, 2014, DOI: 10.1145/2683483.2683538.

2. Sudeshna Roy and Sukhendu Das "Saliency Detection in Images using Graph-based Rarity, Spatial Compactness and Background Prior" *Proceedings of International Conference on Computer Vision, Theory and Applications (VISAPP)*, January 5-8, (2014), Lisbon, Portugal, SciTePress, 2014, pp 523-530, DOI: 10.5220/0004693605230530.

3. Sudeshna Roy and Sukhendu Das "Spatial Variance of Color and Boundary Statistics for Salient Object Detection" *IEEE Proceedings of National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, December 19-21, (2013), IIT Jodhpur, India, pp 1-4, DOI: 10.1109/NCVPRIPG.2013.6776270.

CURRICULUM VITAE

Name : SUDESHNA ROY

Educational Qualification:

2010 Bachelor of Technology (B. Tech)

Institute	:	Institute of Engineering & Management, Kolkata West Bengal University of Technology
Specialization	:	Information Technology

2015 Master of Science (MS)

Institute	:	Indian Institute of Technology, Madras
Specialization	:	Computer Science and Engineering
Date of registration	:	July 16, 2012

GENERAL TEST COMITTEE

Chairperson:

Prof. P. Sreenivasa Kumar
Dept of Computer Science and Engineering
IIT Madras

Guide:

Prof. Sukhendu Das
Dept of Computer Science and Engineering
IIT Madras

Members:

Dr. Madhu Mutyam
Dept of Computer Science and Engineering
IIT Madras

Dr. Sunetra Sarkar
Dept of Aerospace Engineering
IIT Madras