Real Time Signboard Detection and Recognition using Salient Object Detection

Thesis submitted to
Visvesvaraya National Institute of Technology, Nagpur
In partial fulfilment of requirement for the award of degree of

Bachelor of Technology

(Electronics and Communication Engineering)

by

Chaitanya Belhekar (BT13ECE010)

Tejas Bujade (BT13ECE013)

Dhananjay Krishna (BT13ECE018)

Tejash Popate (BT13ECE059)

Guide

Prof. Saugata Sinha

Department of Electronics and Communication Engineering
Visvesvaraya National Institute of Technology,
Nagpur – 440010 (India)

April 2017

Certificate

The thesis titled "**Real Time Signboard Detection and Recognition using Salient Object Detection**" submitted by

Mr. Chaitanya Belhekar (BT13ECE010)

Mr. Tejas Bujade (BT13ECE013)

Mr. Dhananjay Krishna (BT13ECE018)

Mr. Tejash Popate (BT13ECE059)

For the award of degree of Bachelor of Technology has been carried out under my supervision at the Department of Electronics and Communication Engineering of Visvesvereya National Institute of Technology, Nagpur. The work is comprehensive, complete and fit for evaluation.

Dr. Saugata Sinha,
Asst. Professor,
Department of Electronics and Communication Engineering,
VNIT, Nagpur.

Head,
Department of Electronics and Communiaction Engineering,
VNIT, Nagpur.

Date:

Declaration

I, hereby declare that the thesis titled "**Real Time Signboard Detection and Recognition using Salient Object Detection**" submitted herein has been carried out by me in the Department of Electronics and Communication Engineering of Visvesvaraya National Institute of Technology, Nagpur. The work is original and has not been submitted earlier as a whole in part for the award of any degree / diploma at this or any other Institute / University.

Chaitanya Belhekar **Dhananjay Krishna**

Tejas Bujade **Tejash Popate**

Date:

Acknowledgements

With immense pleasure and great respect, we express our deep sense of gratitude to Dr. Saugata Sinha Professor Electronics and Communication Engineering for his invaluable guidance, inspiration and constant encouragement throughout the project work.

It was his constant motivation that helped us complete this journey. We have gained immense knowledge under his patronage and are obliged to him for providing not only his valuable guidance, but also the necessary facilities. We take this opportunity to thank our Head of Department Dr. A.S. Gandhi for his support and for making labs available for our project experimentation. The resources provided by VNIT, Nagpur were an asset and made sure we met all our needs.

We would like to thank all the administrative and technical staff of the department for who have advised and helped us in their respective roles. We would like to thank all those who have directly or indirectly contributed to the completion of the project. Their role and support was of vital importance.

Lastly, we would like to thank the lab assistants and non-technical staff for their support and making sure we have access to the lab whenever required.

ABSTRACT

Saliency is an important property of human visual perception. The objects which appear distinctly from the background are noticed by human eye are the focus of interest. This ability is used to detect the most interested object (salient object) and segment it from an image. It 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.

There are many places which are not available on the map, like small shops, restaurants etc. To add these places on the map, one has to go to the maps application and type the name by itself and save it for oneself. But there's a need of an application which can put such places on the map from the pictures of the surrounding area with the sign board of the shop/restaurant in it.

The primary approach used here is salient object detection using Graph based Visual Saliency (GBVS). In this method any object of interest found is based on visual stimuli, without any prior knowledge about its category. This is a bottom-up method. The method attempts to find regions or objects in an image that are prominent and vividly stand out from the rest of the image. These objects in the image are salient objects.

Using this salient object detection algorithm an application is built in which sign boards are detected from the images taken from the camera. The image is processed using this algorithm and the sign boards are considered as the salient objects.

Also, another algorithm for segmenting the salient object (in this case the sign board) from the rest of the image is deployed. This algorithm masks out the non-salient part and cuts out the required board. The location co-ordinates of the sign board is approximated from the image. Just segmenting the board out of the image is not the prime motive, we need to identify as to what is written on it. Now the text from the sign board is extracted using Optical Character Recognition. After this is done the main aim of the project is yet to complete. Then this sign board's text is mapped to that location in the GPS map. In this manner new places could be added to the map using just photographs.

1. TABLE OF CONTENTS

2. TABLE OF FIGURES

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 of visual data and information available and created every minute. 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 [01]. Hence for a computer vision system, it is important to detect saliency so that the resources can be utilized properly to process important information.

Visual saliency is the ability of a vision system (human or machine) to select a certain subset of visual information for further processing. This mechanism serves as a filter to select only the interesting information related to current behaviours or tasks to be processed while ignoring irrelevant information. Recently, salient object detection has attracted a lot of interest in computer vision as it provides fast solutions to several complex processes. Firstly, it detects the most salient and attention-grabbing object in a scene, and then it segments the whole extent of that object. The output usually is a map where the intensity of each pixel represents the probability of that pixel belonging to the salient object.

We have made use of this saliency technique in detecting sign board of shops, restaurants etc., finding the location of the shop/restaurant and then locating it on the map and storing its name as indicated on the board. For example, the picture of a hotel is taken with its sign board with the camera application on the phone. Now it automatically detects the board and then crops it out and extracts the text and points the location on the map. So any place which was not earlier on the map can be easily added to the map with just a photograph.

This salient object detection algorithms helps us detecting the signboard in the image. Thus this signboard can be used in locating its geographical position on map. The image can be of any place like hotels, shops, buildings, monuments having a signboard with its name inscribed on it, which can be used for data extraction and locating the place.

1.1 Motivation

Signboard is a board displaying the name or logo of a place, product or awareness. Sign board detection is important for computer vision applications such as video surveillance and content based visual information retrieval. Previous researches on this topic focused mainly on application specific sign board such as car number plates and traffic signs. Colour segmentation, gradient analysis, and Hough transform are generally used. Some papers also used neural network or symmetry transform. However, these methods are all limited by some application specific conditions, such as the shape of the car plates and the colour of traffic signs.

Signboards serve as navigation aid in an unknown location. The information present on them becomes extremely important for a person who is not a native of that place. For visually impaired people, this information remains unavailable unless a module is employed exclusively to provide the information. Hence signboard detection and Information extraction module focuses to assist the user in navigating seamlessly in an outdoor environment.

In our work, we concentrate on the extraction of information from this signboard and use them to locate the place on map. This will be helpful in discovering different places which are not available on the map, and also help in more accessibility. The signboards, are put up to convey information about the shop/restaurant. Therefore they are specifically designed to be prominent to actively attract human's attention. This would help in locating more and more places and then adding them to the map.

1.2 Objective

The objective of the thesis is to device an efficient method for detecting the signboard and extracting the information and use it to locate the place, with the help of salient object detection. The image consisting of the signboard of the particular place like hotels, monuments, department, building, thus can be used to pinpoint the location of the place. The image consist the coordinate of the place where the image (photograph) is captured in terms of latitude and longitude, this and the signboard information helps us discovering various places that are not marked on the maps. Further, the method must be unsupervised so that it can detect signboard in any image. Moreover, it has to be computationally efficient to ensure fast processing, considering the huge amount of available data.

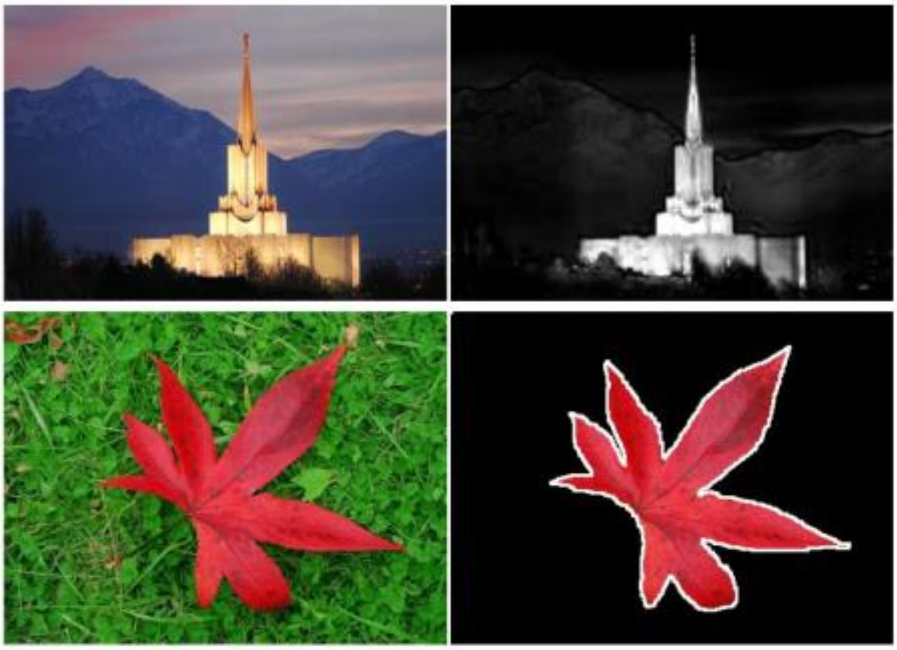

Figure 1-1 The top row shows an example of saliency map generated from the image (left) and the bottom row depicts an ideal segmentation of the object in the image (left).

1.3 Scope

Here, we present a system that can detect generic signs. The signs can be any polygon, not necessarily restricted to traffic signs or car plates. They can also be posters, name plates on doors, advertising board, billboards etc. The only assumption we made is that the sign has to be formed by straight boundaries. The system uses a Graph based visual saliency to detect the most prominent object in the image (sign board). We then develop an efficient detection algorithm and a redundant board detection technique to locate the sign board in the image.

Secondly, this signboard detected using saliency map are used to extract the details using Optical Character Recognition (OCR). This details are then considered in pinpointing the location of the place where the image was captured.

With this application more places can be easily added to the map and more stores and places can be located increasing the accessibility of the places around us.

1.4 Problems and Challenges

The problem we address in the thesis can be defined in short, as:

Given a crude image, detect one or more regions of interest (ROI) which contain the salient objects in a scene. Distinguish the signboard in the salient part in a major problem. Extracting character can be done using OCR, but the problem here lies in quality of image cropped from the salient part. The method must be unsupervised with no training sample for classes of objects available.

The challenging task is detecting salient object in a scene with no training sample classes, because objects of interest are detected without any prior knowledge about them purely based on unsupervised stimulus driven perceptual cues. Single features such as, colour, brightness, depth alone are not helpful to solve the problem. A through study of different method and samples from the dataset reveals the following challenges:

1. Only a small part of an object is present on the boundary of an image;

2. Repeated Distractors in background or foreground.

1.4.1 Assumptions

1. The images are outdoor natural scenes, captured using optical camera (not X-ray or infrared etc.).
2. Object of interest viz. signboard is clearly visible in the image.
3. Object of interest is not just visible in few pixels on the boundary of the image.
4. Object of interest is generally not hidden behind a large distractor, for example,
5. Not much of both colour and texture overlap between object and scene background.
6. The device used for capturing the frame is connected to the internet and is GPS enabled.

1.5 Thesis Overview

We address the problem of signboard detection. We review different methods for detecting the signboard. We analyse the salient object detection method. We focus on the problem of signboard detection using salient object detection. We propose a methods to solve the problem robust 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 in object detection. Moreover, different authors have contributed different methods in salient object detection. The application we proposed for the task completion shows good results in performance.

1.5.2 Chapter 3: Object Detection

In this chapter few methods of object detection are discussed such as edge detection method, template detection, Haar's feature detection method and Viola Jones object detection method.

1.5.3 Chapter 4: Salient Object Detection

Salient object detection is a process in which the most salient object in the image is segmented from its surrounding. The results obtained by object detection using other approaches are not satisfactory. Object detection using saliency technique gives the most efficient segmentation of salient objects. One of the approach is that salient region is that part of the image which is visually more noticeable by virtue of its contrast with respect to its neighbourhood region. Another is using the Boundary prior, Spectral Residual Approach, Contrast prior and Background prior are use along with the objectness proposal (BING).

1.5.4 Chapter 5: Graph Based Visual Saliency

The graph based visual saliency (GB) model used multiple salient regions of different sizes. The approach uses biologically motivated feature selection followed by centre surround operations gives the local gradients, and the combining them gives a final map called as a master map. The method uses a novel application of ideas from graph theory to concentrate mass on activation maps.

1.5.5 Chapter 6: Signboard Detection Algorithm.

The approach started with detecting salient region from the input image (i.e. finding the saliency map) using GBVS algorithm. Then, region of interest from the saliency map had been found out to obtain informative part of the image which also contains the signboard and all other salient objects in the image. From this, signboard had been detected using the proposed algorithm and OCR algorithm (optical character recognition) was implemented to extract the required data from the main input image in order to meet the requirements of project.

1.5.6 Chapter 7: Optical Character Recognition (OCR)

The OCR offers a simple way to add text recognition feature to an extensive array of applications. It is a procedure of translating a document into electronic system bit by bit. It is widely used as a method of data entry from printed paper information records, business cards, mail, printouts of static information, or any suitable documentation.

1.5.7 **Chapter 8: Experiment and Results**

This chapter includes the final outputs of the overall algorithm which contains the cropped signboard and detected alphanumeric data from various test input image considered depending on the parameters considered for observation

1.5.8 Chapter 8: Discussion and Conclusion

In this chapter a summary of work done is presented. As well as the possibility of improvement and the area of future scope is discussed.

1.5.9 Chapter 9: References

In this chapter, a list of papers used as references for understanding and implementation of algorithms used in the project

2. LITERATURE SURVEY

In recent years there has been a surge of interest in context modelling for numerous applications in computer vision. The basic motivation behind these efforts is generally the same - attempting to enhance current image analysis technologies by incorporating information other than the image itself.

In the early work of Galleguillos and Belongie [02], it refers to three main types of contextual information that can be exploited in computer vision:

(1) The context which refers to the likelihood of an object being found in some scenes but not in others, can be expressed in terms of the corresponding object's probability of co-occurrence with other objects and the probability of occurrence in certain scenes;
(2) The position (spatial) context which corresponds to the likelihood of finding an object in some positions and not others with respect to other objects in the scene
(3) The size (scale) context which exploits the fact that objects have a limited set of size relations with other objects in the scene.

A terminology was proposed by Heitz and Koller [03] who introduced a "Things and Stuff" (TAS) context model. In their work, the terms "stuff" and "things" are used to distinguish "material" that is defined by a homogeneous or repetitive pattern of fine scale properties, but has no specific or distinctive spatial extent or shape (stuff) from "objects with specific size and shape" (things).

Rabinovich and Belongie [04] proposed a classification of contextual models for computer vision and object recognition, consisting of models with contextual inference based on the statistical summary of the scene and models representing the context in terms of relationships among objects in the image. Only recently, object hierarchy context has drawn much research attention [05]. The object hierarchy is the further research of object co-occurrence context under the assumption that objects should be related with a semantic hierarchy. With the increased number of object categories, object relationship is naturally exhibited as a hierarchical structure. Context modeling with hundreds or thousands of object categories seeks to model this relationship with high level semantic structure or learned from data [06].

2.1 Object Detection

While object recognition is a long-standing and widely studied problem, most attention until recently has been paid to the recognition of textured objects, for which discriminative local appearance features, invariant to changes in viewpoint and illumination, can be readily extracted [07, 08, 09]. These objects are often assumed to have piece-wise planar surfaces. Their appearance variations can be therefore modelled by a simple geometric transformation (e.g. similarity), which can be reliably determined from the rich textural information. Candidate object locations in the scene are typically determined by identifying so-called interest points or interest regions, a strategy which drastically reduces the overall computational cost compared to exhaustive image search.

2.2 Salient Object Detection

In the past 25 years, over 65 kinds of visual saliency models have been proposed [10], most of which might be able to evaluate the saliency of scene texts. Visual saliency models can be roughly classified into three catalogues:
a) Bottom-up model, b) top-down model, and c) hybrid model.

The bottom-up model of visual saliency [11], [12] considers three low level channels (intensity, colour and orientation) as the feature to identify the salient locations. It consists of three stages, including

1. Feature maps calculation,
2. Conspicuous maps calculation, and
3. Saliency map calculation.

Bottom-up model of visual saliency is task-independent as it does not use any prior information (such as the size and/or shape of the object) to identify the salient objects. In contrast to bottom-up model, top-down model of visual saliency is task-dependent and use prior contextual knowledge to guide to the objects. This kind of model is based on the fact that the context of the scene governs how a person's attention changes while searching for an object [13]. For example, while scanning for pedestrians, we mainly focus our attention on the bottom of the scene and pay less attention on the top part. However top-down model requires at least a basic image understanding technology, which is still a difficult problem to be solved.

The hybrid model of visual saliency combines a bottom-up and a top-down model by using a Bayesian framework [10]. The top-down model is used to indicate the probability of finding the target at the given place, while the bottom-up model verifies the target.

In 1998, inspired by the behaviour and the neuronal architecture of the early primate visual system, Itti et al [11] implemented the first complete visual saliency model of Koch & Ullman [14]. This model uses three low level channels (colour, intensity and orientation) as features and calculates saliency map, which is defined as the degree of difference between an object and its neighbours, for each channel. The degree of difference is measured by the centre-surround operation, which is implemented as the subtraction of images that derive from the same image and are at different scales. In this implementation, Itti et al. down-sampled image from 1 (1:1) to 8 (1:256) scales, and defined images at $c\in\{2,3,4\}$ scales as centre and images at s=c+δ scales, with $s\in\{c,8\}$ for each c and $\delta\in\{3,4\}$, as surround, which is experimentally proved that can maximize the difference. Fig. 2-1 illustrates the meanings of parameter c and δ. The final saliency map is obtained by combining the three saliency maps

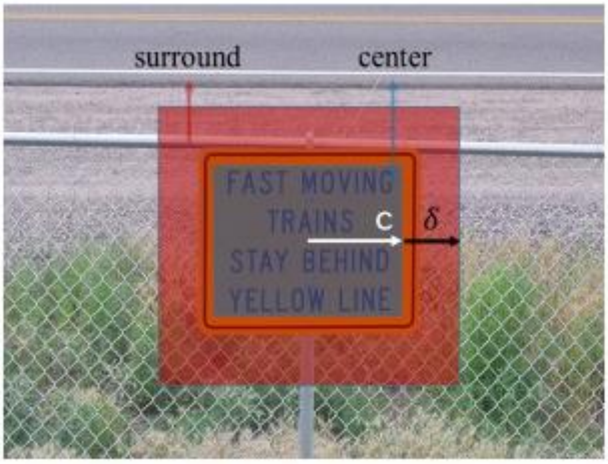

Figure 2-1 Centre Surround Histogram

Like Itti's visual saliency model, Harel et al.'s [12] graph-based visual saliency model (GBVS) utilizes the same features; unlike Itti's visual saliency model, it makes use of the self-information (entropy) of the object instead of using the simple centre-surround operation while processing the feature maps and combining feature maps into activation maps (conspicuous maps) and final saliency map. In order to calculate the self-information, a Markov chain is applied to construct the full connected directed graph which joins all pixels (nodes). Weight of each edge is defined as the dissimilarity and the distance of the two pixels. The more dissimilar as well as the further apart two pixels are, the smaller is the value of the weight. Equilibrium distribution is employed to ensure that, for a given pixel (node), the total weights of the outbound edges is 1. The saliency of the pixel is calculated as the self-information via Shannon formula.

Combining visual saliency with statistical methods, Torralba et al. [13] proposed a hybrid visual saliency model. This model uses Gabor filters to calculate the orientation features (local features) based on which GIST and principal component analysis (PCA) technologies are used to calculate global features for distinguishing different natural scenes. It consists of two components: 1) bottom up component, which is defined as the inverse of the probability of finding the given local feature under the given natural scene and is used to evaluate the saliency, and 2) top down component, which is trained to classify where the object may be by using global features and Gaussian mixture model (GMM), and is used to filter objects that are less like the target. The final saliency map results from the product of the two components. Fig. 2.2 shows the details of the architecture of Torralba's contextual guidance model.

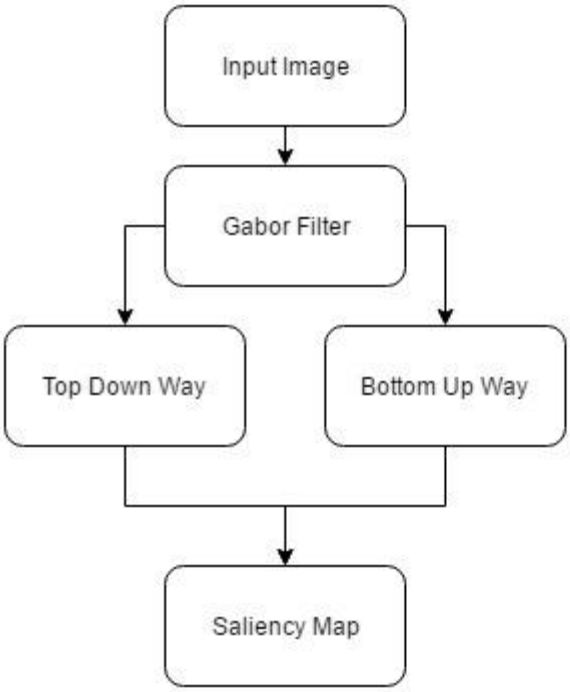

Figure 2-2 Simplified architecture of Torralba's contextual guidance model

Zhang et al. [15] also proposed a hybrid visual saliency model based on Bayesian framework. This model aimed at finding potential targets that might be important for survival, such as food and predators. Unlike the above mentioned models, in which model the saliency of an object is calculated based on the current viewing image, this kind of visual saliency model calculates the saliency of an object based on the images that have been viewed before (past experience). It defines the bottom up saliency of a given object as the probability of it having been seen in the past, that is, the less rare been seen, the more salient it is. This definition is reasonable considering the problem of survival, especially in a dangerous situation. The prior top down knowledge is defined as how the object is similar with the target. All probability distributions are trained off-line on a prepared natural scene image database. The final saliency map is constructed via combining the bottom up saliency component and the top down component. We use Nich et al.'s [16] C++ implementation of fast saliency, which is optimized for robot vision by the use of difference of box (DoB) filters and estimating a Laplacian distribution of unit variance. In the later parts of this paper, we use the term fast saliency to represent Zhang's visual saliency model.

Achanta et al. [17] proposed a novel saliency model in the frequency domain for saliency region detection. This model is mainly to find the largest salient objects in a full resolution (the same resolution as the original input image), in which all the salient regions are uniformly highlighted. All the requirements are satisfied by feeding two thresholds to the difference-of-Gaussian (DoG) filter: 1) w_{lc} for cutting off the low frequency to emphasize the largest saliency objects, and 1) w_{hc} for disregarding high frequencies that are raised from texture, noise and blocking artifacts. The saliency map is calculated as the absolute difference between the arithmetic mean pixel value of the original image and the DoG filtered original input image. The reason to include Achanta's frequency-tuned model in this study are two-fold: first, it obtained the best results in [17]; second, it is based on a different theoretical foundation than the other four models.

2.3 Generic Object Segmentation Methods

Class independent object segmentation has recently gained importance in the Computer Vision community. 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 an exhaustive search which is computationally very expensive. Hence, category independent object segmentation methods are useful and necessary

2.3.1 Objectness

Alexe et al. [18] 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, colour contrast. These cues from [18] are discussed in the following subsections.

Multi-Scale Saliency

The saliency method proposed by Hou and Zhang [19] is extended to multiscale and is processed for each of the colour channels independently. The multiscale saliency map for $(MS(w, \theta_{MS}^s))$, a window w at a scale s, is defined as:

$$MS(w, \theta_{MS}^s) = \sum_{\{p \in w | I_{MS}^s(p) \geq \theta_s\}} I_{MS}^s(p) \times \frac{|\{p \in w | I_{MS}^s(p) \geq \theta_s\}|}{|w|}$$

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

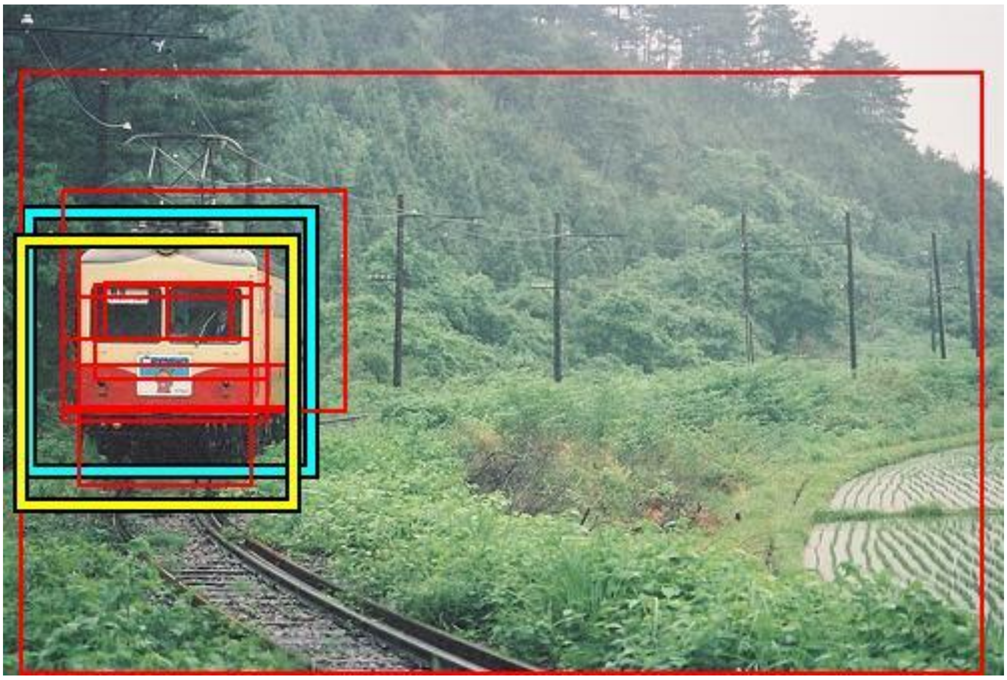

Figure 2-3 Different candidate bounding boxes from Alexe et al. (2012)

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 edges inside a window w($Inn(w, \theta_{ED}) = \frac{|w|}{\theta_{ED}^2}$) is computed as:

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

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

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:

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

where the set of superpixels, $S(\theta_{SS})$, is found using the segmentation method 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.

2.4 Bottom-up Saliency Models

Bottom up saliency models are mostly inspired by neurophysiology, which adapt the concepts of feature integration theory (FIT) and visual attention (Koch and Ullman [20]). The very first model which is proposed by Itti et al. [01], henceforth referred as IT in the following chapters, uses three features, namely colour, intensity and orientation, similar to the simple cells in primary visual cortex. Centre-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 colour, 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 [01] try to capture the human eye gaze behaviours 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. [21]; Achanta et al. [22]) 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.

2.4.1 Graph Based Visual Saliency

A new bottom-up visual saliency model, Graph-Based Visual Saliency (GBVS), is proposed. It consists of two steps: first forming activation maps on certain feature channels, and then normalizing them in a way which highlights conspicuity and admits combination with other maps. The model is simple, and biologically plausible insofar as it is naturally parallelized. This model powerfully predicts human fixations on 749 variations of 108 natural images, achieving 98% of the ROC area of a human-based control, whereas the classical algorithms of Itti & Koch [01] achieve only 84%.

2.5 Summary

The approach is used because

1. The contrast prior approach by Cheng et al and the boundary prior by Wei et al alone are not efficient in obtaining the objects out of the image
2. The BING algorithm is highly computational and is used for high level image processing.
3. GBVS method can detect any generic object as it is unsupervised (bottom up). Moreover, fast processing is ensured by its high computational efficiency, considering the large amount of available data.
4. GBVS predicts human fixations more reliably than the standard algorithms.
5. The application 'sign board detection' do not requires complex algorithms, hence highly complex algorithms are not used and hence GBVS is used.

3. OBJECT DETECTION

Object detection plays an important role in image processing. They are tremendously used in different applications. An image consists of various data, some of it is important in giving information while some of it is uninformative. Object detection is a process of extracting an information from the image. The information may be important or unimportant. Different methods are used for object detection.

3.1 Edge detection

Edges are the boundaries of the object. An enclosed boundary forms an object. Thus if the edges are detected then the object can be identified. Edge detection in image processing is used to detect the boundaries of the objects in the images. It is done by detecting the brightness level in the images. The discontinuities in brightness leads to the recognition of edge. The detected edges leads to the image segmentation. For an object it is likely to happen that the contrast is same. Thus gradients are computed and when there is a sharp change in the gradients then it is considered as an edge. Thus all the edges are computed. If they are connected and forms an enclosed region then it is the object.

Different edge detection algorithms are:

- Canny Edge Detection,
- Sobel Edge Detection,
- Prewitt Edge Detection.

3.2 Template matching

A template is a part of image which needs to be found. In this technique the target image is compared with the template. When the object in the image is matched with the template the output is obtained. The template matching approach is based on level histogram method. The level histogram method is easy to operate and the errors and accuracy has gone qualitative analysis. This approach compares the RGB histogram of both the template and sub image regions of the image. After comparing, when the histogram matches the most that object is the object of interest. A histogram is a graph of intensity of pixels to total number of pixels corresponding to that histogram. . The template is moved all over the image in search of the target sub image and thus the computational cost is very high

3.3 Object detection using Haar-cascade Classifier

This approach uses features for the detection of the objects. The feature calculation was computationally expensive while working with only image intensities which made it slow in most cases. However in this method feature considers neighbouring rectangular regions at a particular location in a detection window, adds up the pixel intensities in each region and calculates the difference between these sums. This difference is then used to classify subparts of an image. The cascade classifiers are used to detect the objects. The system moves windows over the image in search of object. The current location of the window is labelled as positive or negative by the each stage of the classifier. Positive means object has been founded or negative means that the specified object has not been founded in the image. For the negative result the window moves to the next region while for positive result the region goes to the further levels of classification until the final result is positive and thus the object is found.

3.4 The Viola Jones object detection

The main motivation behind this was face detection. Viola Jones object detection is the first object detection framework for high detection rates. It is very robust algorithm because of its very high detection rate (true-positive rate) & very low false-positive rate always. This algorithm is a Real time approach for practical applications at least 2 frames per second must be processed. The objects are only detected not recognized for example face detection only (not recognition) .The goal is to distinguish faces from non-faces (detection is the first step in the recognition process).

4. SALIENT OBJECT DETECTION

The term saliency was first proposed by Tsotsos et al. [23] in the context of visual attention. Salience is defined as something that is particularly noticeable or important. The term visual saliency is related to salience which means it is the perceptual quality which makes the some regions of the world being noticed from their surrounding and immediately the attention is grabbed. Human vision has a property to process the relevant parts of the image while discarding the others.

In image processing saliency is dealt with images. An image constitutes of different types of regions. It has many objects along with the background. Saliency is related to images in order to extract important data from the image. The area of interest has to be considered while removing the other regions in the image.

Considering an example if the image contains a water bottle and there is nothing in its surrounding then the bottle is the region of the image which gives most of the information. This object in the image is called as the salient object. The background gives no information and hence is of no importance. Thus salient object is also defined as the part of the image which gives the maximum information.

A lot of attention is received by this from computer vision owing to its use in object detection, recognition and cropping. Saliency detection has a greater application in image segmentation. The important regions of the image are segmented by removing the uninformative regions.

4.1 Object detection using saliency

Object detection is a technology in image processing which deals with detecting the instances of objects of certain class like human, houses, statues etc. Object detection is one of the most important task in various applications of image processing. Applications like face detection, or tracking a cricket ball, or keeping a track of object's movement. Salient object detection is a process in which the most salient object in the image is segmented from its surrounding. The results obtained by object detection using other approaches are not satisfactory. Object detection using saliency technique gives the most efficient segmentation of salient objects in the image.

Object detection in saliency generally works in three stages

1. Prediction
2. Saliency map formation
3. Normalizing/Combining.

4.2 Saliency map

Saliency map is an image which represents pixel's unique behaviour in the image. An image with lot of pixel values makes the image difficult to analyse in further processing. The representation of an image into a simplified form that is more meaningful and easier to analyse is the goal of a saliency map. The different visual features that contribute to attentive selection of a stimulus (colour, orientation, movement etc.) are combined into one single topographically oriented map, the Saliency map.

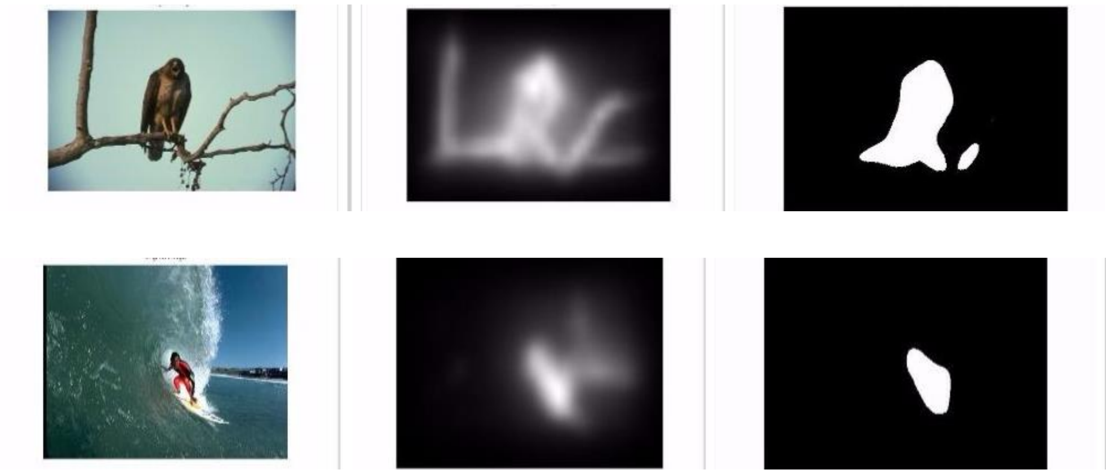

Figure 4-1 (a) Original Image, (b) GBVS Map, (c) Saliency Map

4.3 Salient Object Detection Algorithms

Many approaches have been found in order to implement salient object detection.

4.3.1 Contrast Prior

One of the approach is that salient region is that part of the image which is visually more noticeable by virtue of its contrast with respect to its neighbourhood region. The standard algorithms are given by Cheng et al. (2011) who proposed a region-wise contrast (RC) based method to compute saliency. Also Ming-Ming Cheng, proposed "Global contrast based salient region detection. A high contrast difference is assumed between the image and the background. An image region having a certain contrast level in a region while in the salient region the contrast level drastically changes and that region can be considered as the salient region. Then saliency maps are found using filters that determines contrast levels with reference to a predefined scale. The saliency values are given to each pixels in the saliency maps. These are the weights assigned to each pixel. These maps are further combined to get a final saliency map.

4.3.2 Boundary Prior

Another is using the boundary prior. The methods of Wei et al. (2012) and Zhu et al. (2014) imposed this concept completely to formulate saliency models. It is assumed that the background is present along with the salient object and is not cropped. In this approach the focus is less on objects and more on the background. Two types of priors are being exploited namely boundary prior and connectivity prior. It fails even if the object slightly touches the boundary of the image. A shortest distance between the image patch and the virtual boundary node is calculated and they are connected by an edge of certain weight. These background weights are obtained and pixels with weights in a range are combined. Thus giving the background connectivity.

4.3.3 Objectness Measure

The two approaches the contrast prior and background prior are use along with the objectness proposal in this scheme. Cheng et al. proposes a binarized version of normed gradient features (BING) which can be tested using few atomic operations to generate Objectness proposals. Objectness proposal is used to find the most dominant object in the image. The image contains small number of windows that are likely to contain the object. This reduces the time required for the classifiers to detect the object. Thus Binarized Normed Gradients approach (BING) is used. Using this objectness proposals foreground region is found instead of background regions and the objects are separated from the background. Later superpixels are calculated which gives the idea of how likely a pixel is attached to the foreground and using these weights foreground connectivity is calculated. It requires BING for computation of objectness proposal.

The generic objects with well-defined closed boundary can be discriminated by looking at the norm of gradients, with a suitable resizing of their corresponding image windows in to a small fixed size. Based on this observation and computational reasons, the window is resized to 8×8 and the norm of the gradients are used as a simple 64D feature to describe it, for explicitly training a generic objectness measure.

4.3.4 Spectral Residual approach

Saliency detection can be modelled by 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. The phase information is only used and thus works for salient regions which are small and in an orderly background. The image consists of background region and the object. This approach mainly focusses on the background's properties using the power of log spectrum. It is based on the fact that the log spectra share similarities between the images and these leads to redundancies. In this approach the redundancies in the log spectrum are reduced. The log spectrum is computed by using the Fourier transform. This reduces the complexity in the computation part of the image. A threshold is set and according to that threshold the image is divided into two different regions. The region interest i.e. which contains the object is cropped out. It is a simple method in terms of complexity but the output obtained from this approach fails to give an object boundary owing to complex and unorderly regions

4.3.5 Graph based Visual Saliency

The ability of the human eye to fix on relevant objects and discarding the remaining uninformative part is commonly used in vision community. The idea to keep a fixation on important information is the main idea behind it.

The graph based visual saliency (GB) model proposed by Harel et al. (2006) works quite well for multiple salient regions of different sizes. It is similar to Itti et al. (1998). The approach uses biologically motivated feature selection followed by centre surround operations gives the local gradients, and the combining them gives a final map called as a master map. The two quantities as self-information and surprise are defined for the saliency. On operating on a feature map a saliency map is computed for each number of feature channels.

The three stages of saliency include

1. Extraction- a feature map is made over an image map. It gives the region of image which is likely to contain the object
2. Activation- by working on a feature map an activation or saliency map is computed. It describes the image in such a way that analyses can be done easily
3. Normalization/ combination- the maps are normalized first and then they are combined into a single map where a connected object is found.

5. GRAPH-BASED VISUAL SALIENCY

In saliency the differentiation between the features is of prime importance. The interesting information in the image is likely to be contained in saliency map. High saliency regions are the objects of our interest while low saliency are the parts which are the background in the image. A graph based approach called as Graph based visual saliency (GBVS) is used to find the saliency at each and every location of the image where the local information of the image is not being utilized.

The three stages of the model of visual saliency are extraction, activation and normalization followed by combination.

After extracting the features from the image, method consists of two steps i.e. the activation – forms a saliency map on the feature vectors and normalizing the saliency map or (maps) and then combining into a single map. It is simple and biologically feasible. Using some criteria the aim is to find the notable locations where the image is informative. The algorithm allows to include pre-eminent points away from the boundaries. In order to get efficient saliency computations a different kind of approach on which the use of topological structures, parallel nature of graph algorithm and computational power.

Markov chains are defined over the image maps and the equilibrium locations over the map are treated as activation and saliency values. The activation/saliency map and the normalization are computed by using non uniformity. In this case, like current methods only the somehow connected features are not connected.

Figure shows the original image with its corresponding map, the overlapped mapped on the original image and the final image in which salient object is found. The image input is processed first by computing the feature map by linear filtering and then followed by some elementary nonlinearity.

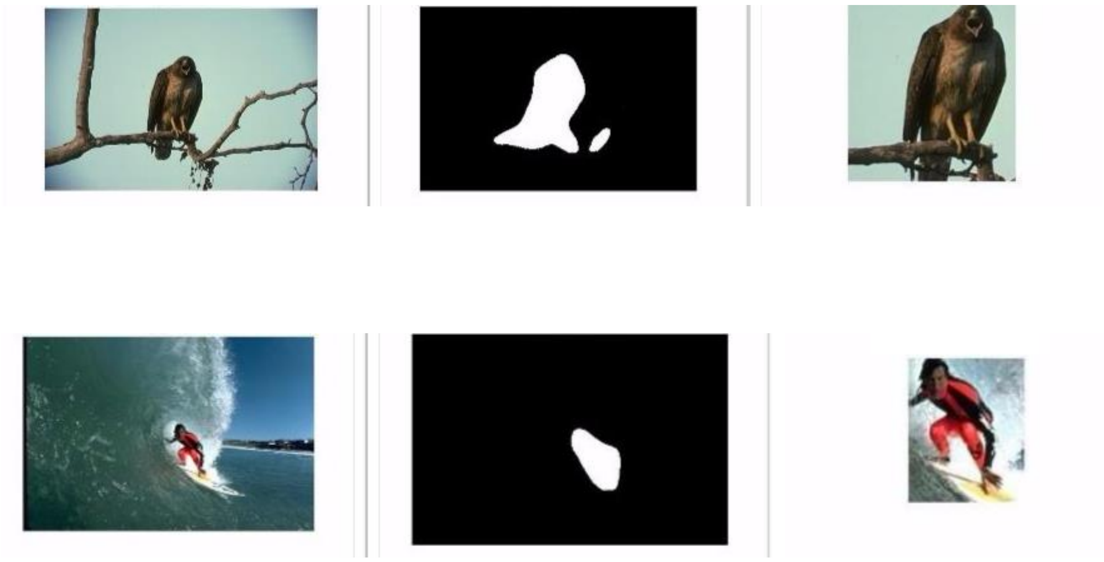

Figure 5-1 (a)Original Map, (b)Saliency map after applying GBVS, (c)Cropped output image.

5.1 <u>Computing a feature map</u>

A feature is computed by centre surround operations which is similar to visual receptive fields. These visual neurons are sensitive to the small regions of virtual space (the centre), while stimuli presented in broader weaker region concentric with centre i.e. the surround inhibit neural response. The architecture sensitive to local spatial discontinuities is sensitive to find the locations which are different from the surrounding.

Centre surround is the difference between coarse and fine scales for the pixel values. If the centre is a pixel at x={1,2,3} and delta ={5,6} then the surround is a pixel at value s=del+x. using various several scales for del and x gives the extraction of feature map.

1

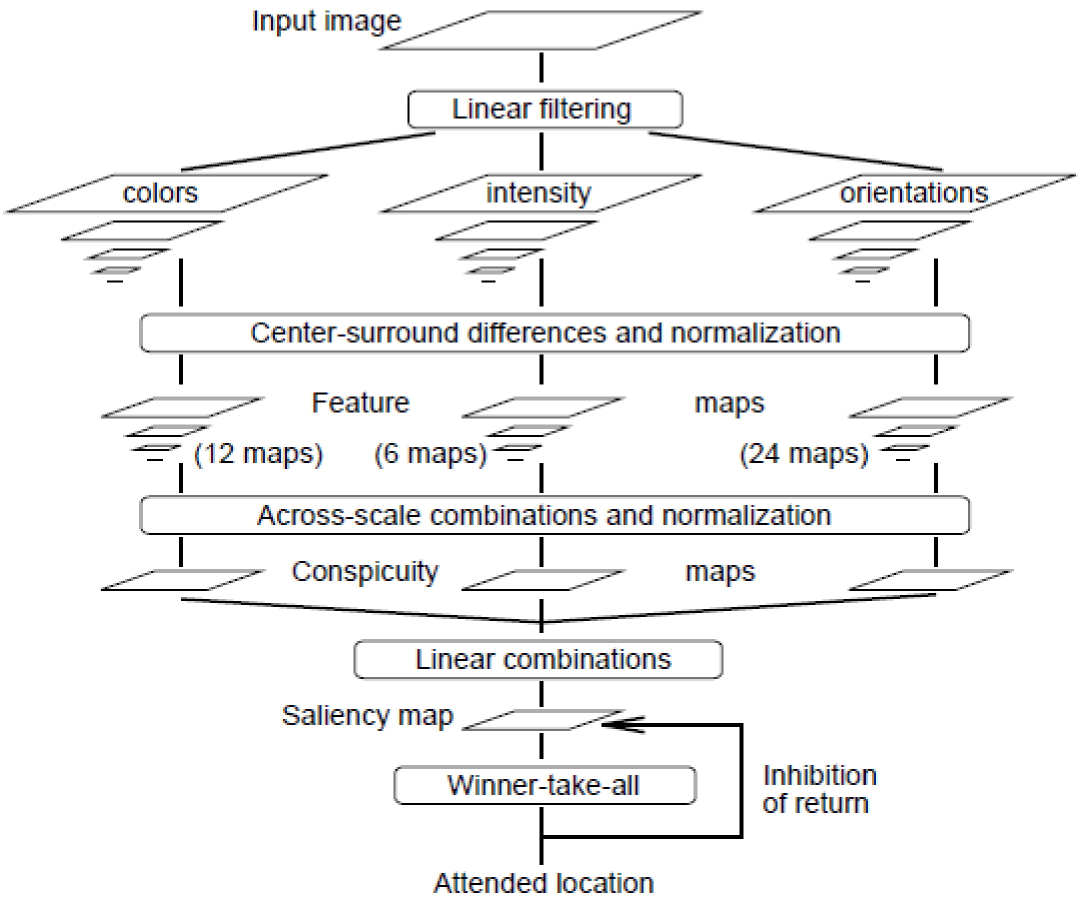

Figure 5-2 General Architecture of the Model

Considering an image given as an input, initially the noise is removed by doing linear filtering. The next step is to extract the feature map from the image.

5.1.1 <u>Extraction of early visual features</u>

The image I which is the intensity image is found by taking the average of the red (r) blue (g) and green (b) channel. I is used to find a Gaussian pyramid I (□). The r, g, b channels are normalized by the value of I in order to decouple hue from intensity. Normalization is applied at the places where the intensity is larger than one tenth of its maximum over the entire image.

Other than this colour channels are created given by $R=r-(g+b)/2$ for red $G=g-(r+b)/2$ for green $B=b-(r+g)/2$ for blue and $Y=(r+g)/2-|r-g|/2-b$ for yellow (as zero is set for negative values). The four colour gradients are defined as R (□), G (□), B (□), Y (□).

Using centre surround method the difference between the centre and surrounding is calculated so as to get the feature map. As shown in the figure the first set consists of intensities which tells that the difference between the darker to brighter or brighter to darker. Both of these are calculated simultaneously.

$$I(x,s)=|I(x)\Theta I(s)|$$

Another maps are constructed for the colour scheme in the image. The colours exhibit opponency for the pairs as red-green, blue-yellow, green-red, yellow-blue in visual cortex of the human. In the small region i.e. is the centre the neurons of humans are exhibited by one of the colour and inhibited by its opponent pair and conversely for the surrounding region.

The feature maps RG(x, s) are created by simultaneously for red-green and green-red which are given by

$$RG(x, s) = (R(x)-G(x)) \Theta (R(s)-G(s))$$

Similar is for blue-yellow and yellow-blue double opponency. The information about orientation is obtained by using Gabor pyramids $O(\square,\theta)$, $\square$= 0.8 is the scale and the orientations are θ =(0^0 45^0 90^0 135^0). The feature maps of orientation encrypt in a group, orientation contrast between the surround and centre scales.

$$O(x, s, \theta)=|O(x, \theta)\Theta O(s, \theta)|.$$

In all total 42 maps are computed in which 6 are of intensities, 12 are for colours and 24 are for orientation. After the features are extracted the next step is forming an activation or saliency map from multiple feature maps.

5.2 Computing an activation map

The high values of activation are found when locations in a feature map: M: $[n]^2 \rightarrow$ R, are somehow unusual in its neighbourhood. In this case the activation maps A: $[n]^2 \rightarrow$ R are computed using the Markovian approach.

5.2.1 A Markovian approach.

The definition of dissimilarity is quality of being unlike or unique. In image processing it is defined as the distance between one and the ratio of two quantities, measured on a logarithmic scale.

$$d((i,j)\|(p,q)) \triangleq \left|\log \frac{M(i,j)}{M(p,q)}\right|$$

It is defined as the dissimilarity between M(i, j) and M(p, q) which can be called as a more organic approach. Instead of logarithmic ratio the difference can be used |*M(i, j)- M(p, q)*|.

Connecting every node of the lattice M, with all other n-1 nodes gives a fully-connected directed graph,

A weight is assigned to the directed edge from node (i, j) to node (p, q) will be

$$w_1((i,j),(p,q)) \triangleq d((i,j)\|(p,q)) \cdot F(i-p, j-q)$$

$$F(a,b) \triangleq \exp\left(-\frac{a^2+b^2}{2\sigma^2}\right)$$

In M domain the weight of edge from one node to other is proportional to the dissimilarity and closeness between the two. Also the weight of edge is exactly same in the opposite direction. Now on the fully connected graphs the Markov chain is defined by normalizing the weights of the out bounded edges of each node to 1 and an equivalent relation is drawn between nodes & states, and weights of edges & probabilities of transition.

The node with high dissimilarities with the neighbouring nodes will have the distribution function with high mass at that node which reflects the fraction of time spend at each node and for the nodes which have similar M values will have unlikely transitions into sub graph. It is a measure of activation derived from pairwise contrast.

The approach is also termed as organic as the nodes are related with neurons exists in a connected network and their fast processing nature in the areas where additional processing is required. The computations are held at each node simultaneously in a synchronous manner at each and every step the incoming mass are simply added and then forwarded the required partitions of mass to neighbours.

5.2.2 Normalizing and combination of activation maps

This is the interesting area of study. The aim of this step is to concentrate the mass on activation maps. Before combining if mass are not concentrated on the activation maps then the map formed after combining is uniform and thus very less information is contained i.e. the area of interest is not obtained.

The very aim of the GBVS is to concentrate their mass on prime locations in the image. An approach used to compute it follows that only the activation map of interested locations are considered which are further normalized. Then a graph is constructed for nodes. An edge is introduced between two nodes (i, j) and (p,q) with a weight equal to

$$w_2((i,j),(p,q)) \triangleq A(p,q) \cdot F(i-p, j-q).$$

The weight of outbound edges of each node are again normalized to 1 and the obtained result is considered as Markov chain leads to compute equilibrium distribution over the nodes. At these nodes the mass will flow preferentially with high activation. Thus the mass is concentrated and the algorithm works in parallel.

5.3 Saliency maps

The saliency map is used to represent the saliency at each and every location and the attended locations are selected based on saliency distribution level. The feature maps are combined to provide a bottom up input to the saliency map. They face difficulties such as salient objects present in the image are masked by noise or due to presence of other many non-salient objects.

Thus a map normalization operator is used N(.) in which

1. The values in map are normalized to a particular range to eliminate the modality dependant amplitude difference range
2. To find the global maxima (M) and average of all the local maxima (m).
3. The map is multiplied globally by $(M-m)^2$

The N(.) compares responses with the activation spots and the local maxima of activity are considered ignoring the homogenous regions. The maximum activity is compared with the average of the overall activation spots and the difference leads to the result that how much the most active region is different from the other average regions. Higher the difference that region is highly different from its surrounding and hence is highly selected while lower the difference the probability of selecting it in the map is less as both of them are unique.

Next step is adding them to form a saliency map. The feature maps are combined into three conspicuity maps which are for intensity, orientation and colour but are at a scale of 4. The three conspicuity maps are normalized and added together to get the final saliency map.

6. SIGNBOARD DETECTION ALGORITHM

After considering the assumptions, the algorithm for extracting signboard is explained in the following steps:

- Step I: Calculate saliency map using GBVS method.
- Step II: Extract region of interest (ROI) within saliency map using morphological operations.
- Step III: Removing the noise in the saliency map by considering regions having high saliency values using threshold.
- Step IV: Apply saliency map as mask to obtain ROI from the original image.
- Step V: Label the connected regions from saliency map and crop region of the image having largest area v.i.z. the required region for the signboard.
- Step VI: Use optical character recognition (OCR) method to detect/extract alphanumeric data from the cropped region obtained in the previous step. Data obtained is saved for further purposes.
- Step VII: Get the GPS location from the EXIF (Exchangeable image file format) metadata provided with the test input image to map the data recorded in previous step on Google Static Maps API/().

These steps are explained in details as follows:

Step 1: GBVS method was applied on the test input image where activation maps of each computed feature of the image were calculated, then normalized to obtain saliency map. The result is shown in fig. (b)&(c), where the fig. (b) shows the resized saliency map and fig.(c) shows overlapping of saliency map on the test image. The GBVS approach is more useful for images containing number of prominent images. Thus even if the image contains more than one signboards, they all will be considered prominent objects and considered in the saliency map. Mathematical approach of GBVS method was explained in the previous chapter (Chapter 5).

6.1 Region of Interest Detector

Step 2: Morphological image processing is used in the consideration of region of interest (ROI).The scalable framework provided in [24] for region of interest detector uses object location priors from saliency map for graph-cut. Bag-of-features (also known as bag-of-words, e.g. [25], [26]) has emerged as a powerful framework in the context of object and scene classification. The segmented objects are classified using bag-of-features followed by scene classification using decision trees. Fig. (1) shows the whole procedure.

A. Saliency Guided Object Discovery

The output of saliency detector (GBVS) is a probability map of pixels being salient in the range of [0, 1]. The existing state-of-art algorithms [12] are compared using automatic ROI detection. Fig 6-2 shows the comparison between GBVS and Itti et al. model. Graph based visual saliency (GBVS) [12] had consistent good localization on the database which means that a high proportion of overlap of detected regions is obtained. Following equation measures the overlapping,

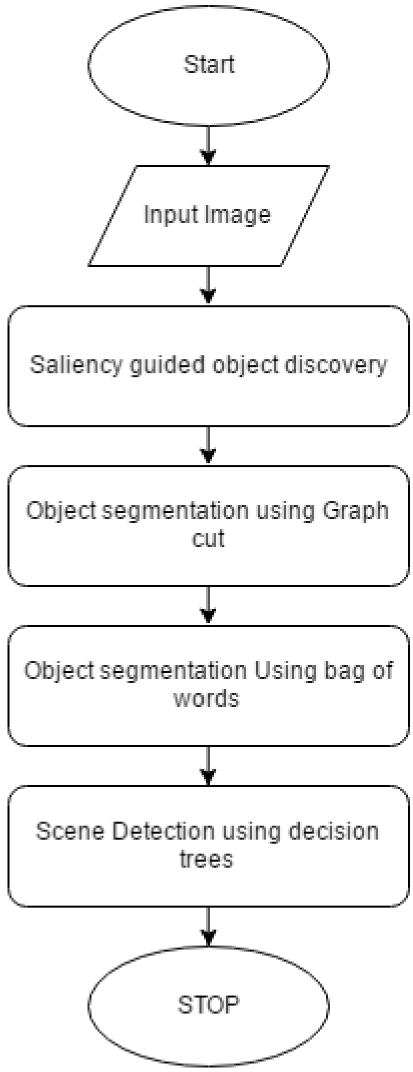

Figure 6-1 Flowchart for Detecting Region of Interest

$$O = \frac{|Rs \cap Rg|}{|Rg|}$$

Where, Rs –output of ROI detection and Rg-ground truth value. GBVS achieved 64.4% average overlapping with ground truth while in case of Itti et al model, it was only 44.1%.

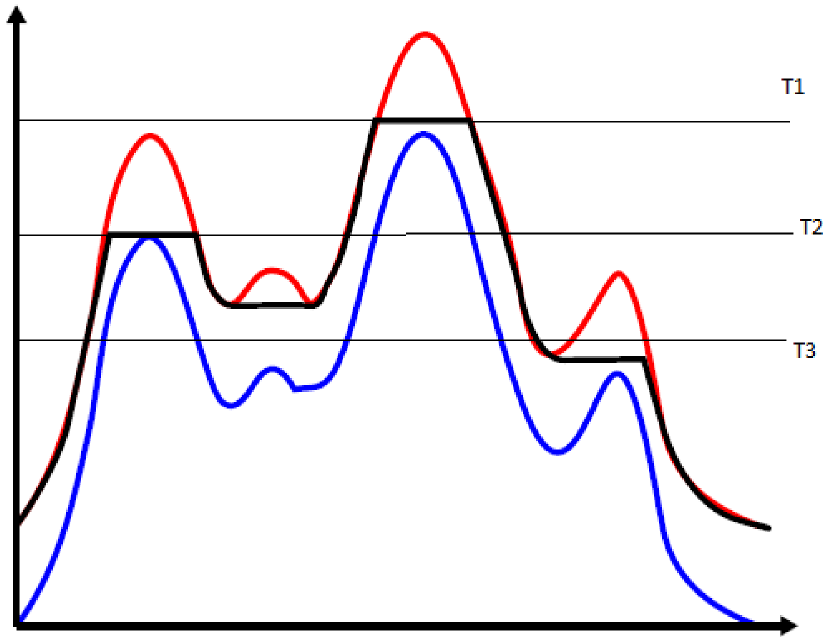

Figure 6-2 First, original saliency map (red) is eroded, then eroded image (blue) is dilated using original saliency map(red) in order to get final reconstructed image (black).

Conversion of saliency map into binary mask is required for obtaining rough object locations. This can be achieved by simply setting threshold for saliency map, but due to large variance of peaks of same scene , regional maxima of saliency map is calculated by using morphological operation known as "opening by reconstruction" followed by finding regional maxima. This operation consists of two steps, first, erosion is performed on the saliency map using 20-pixel disk-shaped structuring element. Undesirable small peaks which are smaller than the structuring element are removed whereas the desirable peaks remain. In second step, morphological reconstruction is performed on the eroded image where original image is considered as a mask for reconstruction. Here, the dilation of the eroded image is repeated till its contours fit under mask image. This results in creation of flat peaks at regions of high saliency. The output of this step gives set of regions which will be used as foreground markers for segmentation by graph cut.

B. Object Segmentation by Graph-cut

This gives segmentation based on minimization problem for Markov Random Field. The idea is to separate the background and foreground pixels with minimum cost. Each pixel has its pairwise matrix where each pixel is connected to its 4 or 8 neighbours. Also, each pixel is connected to its background or foreground node with some respective associated weight. Graph-cut also requires background markers apart from the foreground ones which were obtained in the previous step. The least salient regions with saliency value less than 0.1 threshold are used to select the background markers. Using k-means, clustering is performed on marked foreground and background pixels. And for each cluster, Gaussian Mixture Models are fitted [27]. The probability of a pixel for being a background and foreground is calculated by,

$$\mathrm{Px} = \sum_{\mathrm{k}} \mathrm{Wk}\exp(-|\mathrm{Ix} - \mathrm{Ck}|)$$

Where C_k -kth Gaussian centre, I_x –RGB value of the pixel, W_k –proportion of the marked pixels belonging to the kth centre. From this, edge weights of all pixel nodes to foreground and background node are obtained. For further minimization, max-flow/min-cut procedure is used.

$$\mathrm{E(f)} = \sum_{\mathrm{p \in P}} \mathrm{Dp(fp)} + \sum_{\mathrm{p,q \in N}} \mathrm{Vp, q(fp, fq)}$$

Where, D(_)- edge cost of pixels to background and foreground nodes

V(_,_)- edge cost of each pixel to its neighbours.

Fast and precise object segmentation of regions is achieved is achieved from graph-cut method. This is critical as object classifier is sensitive to the segmentation quality. Graph-cut is performed on each foreground object separately for obtaining segmented images of individual objects.

C. Object Classification

This stage consist of two steps, a) feature representation, b) classification based on the bag-of-features model [28]. The following block diagram represents the whole stage of object classification:

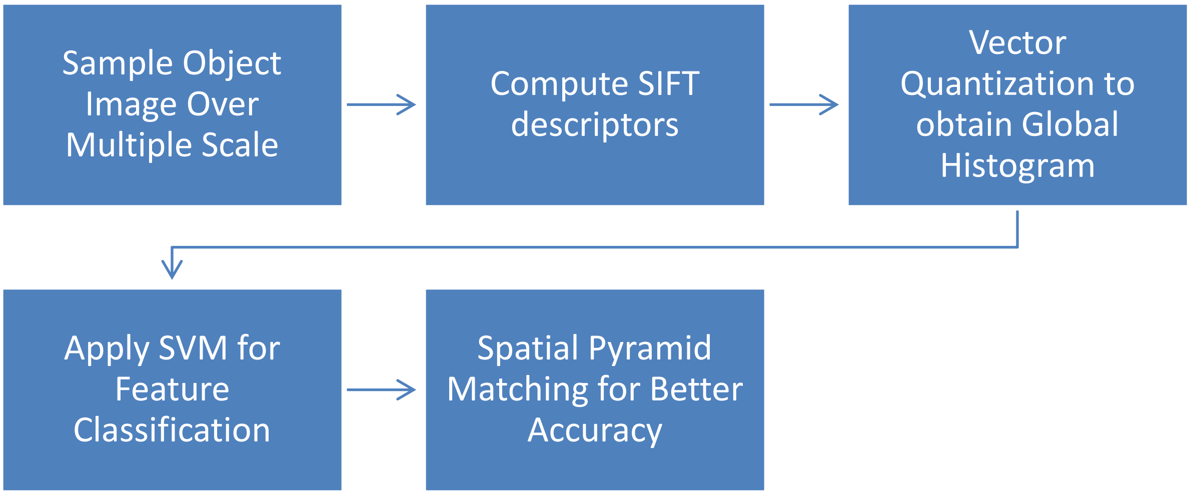

Figure 6-3 Object Classification Overall Process

D. Scene Classification

The images are labelled depending upon the occurrences of objects within them. A decision tree is formed based on labelled images. The presence/absence of objects in the image are treated as parameters for the decision tree. This helps in distinguishing a particular scene from others. This step is not used in our actual project as most of the time, region having largest area is considered for further processing.

Step 3:

The previous step gives ROI in the saliency map image in grayscale. This image is then converted into binary image and will be used for future operations. Now, to remove further noise and small component values, opening and closing operation are performed simultaneously on the same image by considering 20 pixel disk structuring element of radius value ‘5’. The operation results are shown in Fig 6-4. Then, pixel values of obtained result are then compared with the predefined threshold value which was considered on observation basis taken on image database. Result obtained after performing these operations shows the regions having most of the higher values of saliency. Thus, in other way, new saliency map in the form of binary image is obtained which will be used as mask in next step.

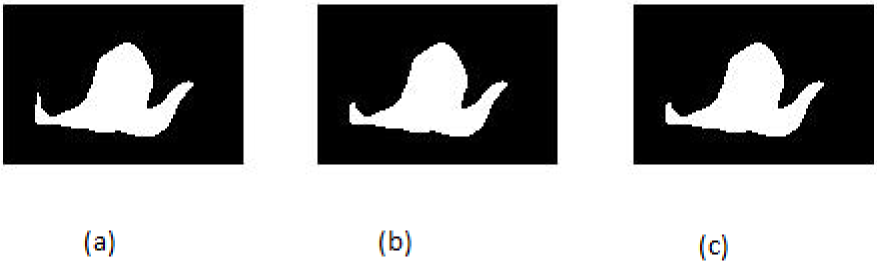

Figure 6-4 (a) Binary Saliency map, (b) Opening Operation, (c) Closing Operation

Step 4:

The binary image obtained in the last step will be used as binary mask for obtaining part of the image that contains one or more salient object. This reduces the image area considered for computations. Binary mask used is of the same size as that of original input image. Masking used is much simpler than other masking algorithms. Only the location of pixels having value of '1's in binary mask will be taken in account for the extracting most informative uncropped image from the original image. Resultant image will have same size as that of original image. Mathematically, this step is as follows,

$$G(x, y, z) = \begin{cases} F(x, y, z), & \text{for } f(x, y, z) = 1 \\ 0, & \text{otherwise} \end{cases}$$

Where, F(x,y,z)- represents matrix of original test image,

f(x,y,z)- represents matrix of binary mask,

G(x,y,z)- represents matrix of informative image.

6.2 Signboard Cropping

Step 5:

Figure 6-5 Process for Cropping Signboard

Finding Connected Regions:

The basic steps in finding the connected components are:

1. Search for the next unlabelled pixel, p.
2. Use a flood-fill algorithm [29] to label all the pixels in the connected component containing p.
3. Repeat steps 1 and 2 until all the pixels are labelled.

The attributes of detected connected regions (area, connectivity, size, etc.) are then saved in structure variable.

Flood-fill Algorithm [29](Iterative):

Flood fill (I, u, v, label)

step.1 Generate empty queue q

step.2 ENQUEUE (q, (u, v))

step.3 Until queue q is not empty do-

step.4 (x, y) _ DEQUEUE (q)

step.5 If (x, y) is inside the image and label of (x,y) = 1 then

step.6 set I (x, y) label

step.7 ENQUEU (q, (x-1, y+1))

step.8 ENQUEU (q, (x, y+1))

step.9 ENQUEU (q, (x+1, y+1))

step.10 ENQUEU (q, (x-1, y))

step.11 ENQUEU (q, (x+1, y))

step.12 ENQUEU (q, (x-1, y-1))

step.13 ENQUEU (q, (x, y-1))

step.14 ENQUEU (q, (x+1, y-1))

step.15 return

- Label Connected Regions:

  The connected regions found in previous step are assigned intensity values. Pixels of same connected region will have same intensity value. Pixels in different connected region will have different intensity value. For eg. If there are three different connected regions say R_1, R_2, R_3 then pixels in R_1 will have intensity value of I_1, that of R_2 will have I_2 and that of R_3 will have I_3.

- Finding the Properties of Connected Regions:
  Image region can have properties such as area, connectivity, centre of mass, orientation bounding box, eccentricity, perimeter, etc. Out of these, area of the regions are considered.
- Cropping Region of Signboard Area:
  Region having largest area is found out using Matlab commands and that region is cropped out showing the required signboard. As the signboard is the major salient object in the test input image, it will have largest connected area. If an image contains more than just one signboard then it will follow following approach:
  a) Calculate height (h_i) and width (w_i) of each connected region (R_i) bounding box of each connected region.
  b) For i= 1 to N(no of connected regions in an image)

     If $w_i > h_i$

     Perform OCR operation.

     Increment i

     End
  c) Stop.

6.3 Accessing the GPS location

Step 7:

This step involves extracting GPS location data from the EXIF (Exchangeable image file format) metadata provided with the test input image and mapping it on google map using Zoharby Bar –Yehuda's MATLAB toolbox.

EXIF data: Modern digital camera has the capability to record information, along with many other camera settings, right into the photographs. These settings can then be later used to organize photographs, perform searches and provide vital information to photographers about the way a particular photograph was captured. This stored data is called "EXIF Data" and it is comprised of a range of settings such as ISO speed, shutter speed, aperture, white balance, camera model and make, date and time, lens type, focal length, GPS location and much more.

```
loca =

           GPSVersionID: [2 2 0 0]
         GPSLatitudeRef: 'N'
            GPSLatitude: [21 7 21.7478]
        GPSLongitudeRef: 'E'
           GPSLongitude: [79 3 2.8137]
         GPSAltitudeRef: 0
            GPSAltitude: 266
           GPSTimeStamp: [12 58 47]
            GPSMapDatum: 'WGS-84'
    GPSProcessingMethod: 'ASCII   gps'
           GPSDateStamp: '2017:03:27'
```

Figure 6-6 EXIF data obtained from a test image

Coordinate format conversion: The numerical values for latitude and longitude can occur in a number of different formats

- degrees minutes seconds: 21° 7′ 21.7478″ N 79° 3′ 2.8137″ W
- degrees decimal minutes: 21° 7.22′ N 79° 3.933′ W
- decimal degrees: 40.446° N 79.982° W

There are 60 minutes in a degree and 60 seconds in a minute. Therefore, to convert from a degrees minutes seconds format to a decimal degrees format,

Decimal degrees = degrees + minutes/60 + seconds/60.

MATLAB Toolbox: It uses the Google Maps API to plot a map in the background of the current figure. It assumes the coordinates of the current figure are in the WGS84 datum, and uses a conversion code to convert and project the image from the coordinate system used by Google into WGS84 coordinates. The zoom level of the map is automatically determined to cover the entire area of the figure. Additionally, it has the option to auto-refresh the map upon zooming in the figure, revealing more details as one zooms in.

As shown in Fig (), the obtained GPS location (latitude and longitude) is in the format of D˚M'S'' which is then converted into D.D˚ format of geographic locations. These values are used to mark the GPS location in the google map using Zoharby Bar-Yehuda MATLAB toolbox.

6.3 Database

For the validation of the algorithm we need to have sets of the images which satisfy all our assumptions. For this we accumulated a set of images to form a database for the verification of the algorithm. The proposed solution should uphold its parameters on every image and able to produce an efficient output.

For the database we need to divide the categories of image using some parameters. For this database we have considered 3 types of parameters which are as follows:

1. Angle – The database images consist of a set of images of the same object of interest captured from different angles.
2. Time – The frames are captured at different time period like evening, night, morning, to get different aspect of images.
3. Object of Interest – Here we capture the frames with more than one object of interest, with different orientation of object of interest to check the limitations of the project.

With this database we can analyse different parameters, and test the algorithm to its full extent.

7. OPTICAL CHARACTER RECOGNITION (OCR)

Optical character recognition is electronic alteration of images of handwritten, keyed or printed script into text which is machine-encoded, whether from a scanned article, a picture of a document, a text on signs and billboards or from subtitle text overlaid on an image. It is widely used as a method of data entry from printed paper information records, business cards, mail, printouts of static information, or any suitable documentation. It is a common method of digitizing printed typescripts so that they can be electronically corrected, searched, put in storage more compactly, exhibited on-line, and used in machine developments such as "cognitive computing", "machine translation", text-to-speech, key data and "text mining". OCR is a field of study in "pattern recognition", "artificial intelligence" and "computer vision."

Early versions needed to be trained with images of each character, and worked on one font at a time. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common, and with support for a variety of digital image file format inputs.[2] Some systems are capable of reproducing formatted output that closely approximates the original page including images, columns, and other non-textual components.

7.1 Text Recognition Using the OCR Function

Identifying text in images is advantageous in many computer vision applications such as image search and document analysis. The OCR function offers a simple way to add text recognition feature to an extensive array of applications.

The procedure of translating a document into electronic system bit by bit. This electronic translation is achieved through a process called "imaging", here a document is scanned and an electronic illustration of the original is created. The imaging process contains recording variations in light intensity reflected from the document as a matrix of dots. The light/ colour values of each dot is kept in binary digits, one bit being necessary for each dot in a black/white scan and up to 32 bits for a scan in colour. OCR takes this data further by translating this electronic data, firstly a bitmap, into machine-readable, editable text.

7.2 Components of an OCR System

An OCR system comprises of numerous components. The first step is to digitize the analog document using an optical scanner or take a picture using a camera. When the regions comprising text are positioned, each symbol is extracted using a segmentation method. The extracted symbols is then pre-processed, removing noise, to enable the extraction of features in the next step. The uniqueness of each character is found by equating the extracted features with descriptions of the character classes attained through an earlier learning phase.

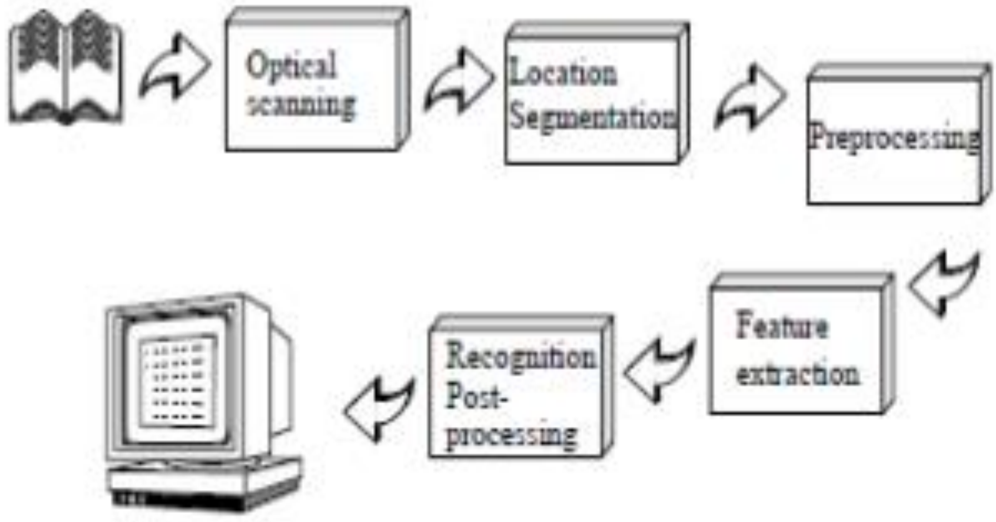

Figure 7-1 Components of an OCR system

7.2.1 Optical scanning

In OCR optical scanners are used, which commonly comprises of a transport mechanism and a sensing device that transforms light intensity into grey levels. A picture can also be taken using a camera. When performing OCR, it is general practice to translate the multilevel image into a bi-level image of black and white. This process is often known as thresholding. A fixed threshold is used, where grey-levels above this threshold is said to be white and levels below are said to be black.

7.2.2 Location and segmentation

The isolation of characters or words is Segmentation. The majority of OCR algorithms segment the words into isolated characters which are recognized separately. This segmentation is performed by isolating each connected component, which is each connected black area.

7.2.3 Pre-processing

The image resulting from the scanning procedure may comprise of a certain amount of noise. The smoothing indicates both thinning and filling. Filling removes slight breaks, holes and gaps in the digitized symbols, while thinning shrinks the width of the line.

7.2.4 Feature Extraction

The procedures for extraction of such features are often divided into three main groups, where the features are found from:

- Structural analysis.
- The distribution of points.
- Transformations and series expansions.

mat2cell command in MATLAB is used for the extraction of image in the form of a cell for associating with the saved templates.

7.2.5 Template-matching and correlation techniques

These procedures are dissimilar from the others in that no features are actually extracted. In its place, the matrix comprising the image of the input character is directly matched with a set of pattern characters representing each possible class.

7.2.6 Post Processing

It involves error detection, grouping and correction techniques. The result of plain symbol recognition on a document, is a set of separate symbols. Even the best recognition schemes will not give 100% percent accurate identification of all characters, but some of these errors may be distinguished or even corrected by the use of context.

7.3 The OCR function in MATLAB.

Before the recognition procedure, the function converts true colour or grayscale input images into a binary image. The function uses Otsu's thresholding technique for the conversion. For best results from OCR, the height of a lowercase ‘x', or equivalent character in the input image, must be bigger than 20 pixels. From either the horizontal or vertical axes, eliminate any text rotations bigger than +/- 10 degrees, to improve recognition results.

The function returns text recognized in the rectangular regions as an array of objects.

7.3.1 Otsu's method

It is used to automatically accomplish clustering based image thresholding, or, the reduction of a grey level image to a binary image. The algorithm assumes that the picture contains two classes of pixels following bi-modal histogram (foreground pixels and background pixels), it then computes the optimum threshold separating the two classes so that their collective spread (intra-class variance) is minimal, or equivalently (because the sum of pairwise squared distances is constant), so that their inter-class variance is maximum. Consequently, Otsu's method is roughly a one-dimensional, discrete analog of Fisher's Discriminant Analysis.

7.4 Algorithm

1. Compute and probabilities and histogram of each intensity level.
2. Set up initial $\mu_i(0)$ and $\omega_i(0)$.
3. Step through all possible thresholds t=1,… maximum intensity.
    1. Update μ_i and ω_i .
    2. Compute $\sigma_b^2(t)$.
4. Desired threshold corresponds to the maximum $\sigma_b^2(t)$.

7.4.1 How the OCR function works?

ocr(I) returns an ocrText object containing optical character recognition data from the input image, I. The object contains recognized text location, text and a metric indicating the assurance of the recognition result. Using reasonable heuristics on text strings, such as spacing, height similarity and alignment, the extracted strokes are then processed to form tight rectangular bounding boxes around the corresponding text strings.

OCR system is highly dependent on the input quality, this makes it difficult to evaluate and compare different systems. Still, recognition rates are often represented, and usually give as the percentage of characters correctly classified.

Three different performance rates are investigated by the following:

- Rejection rate
- Recognition rate
- Error rate.

8. EXPERIMENT AND RESULTS

The method proposed in this chapter is evaluated on popular saliency and object segmentation datasets.

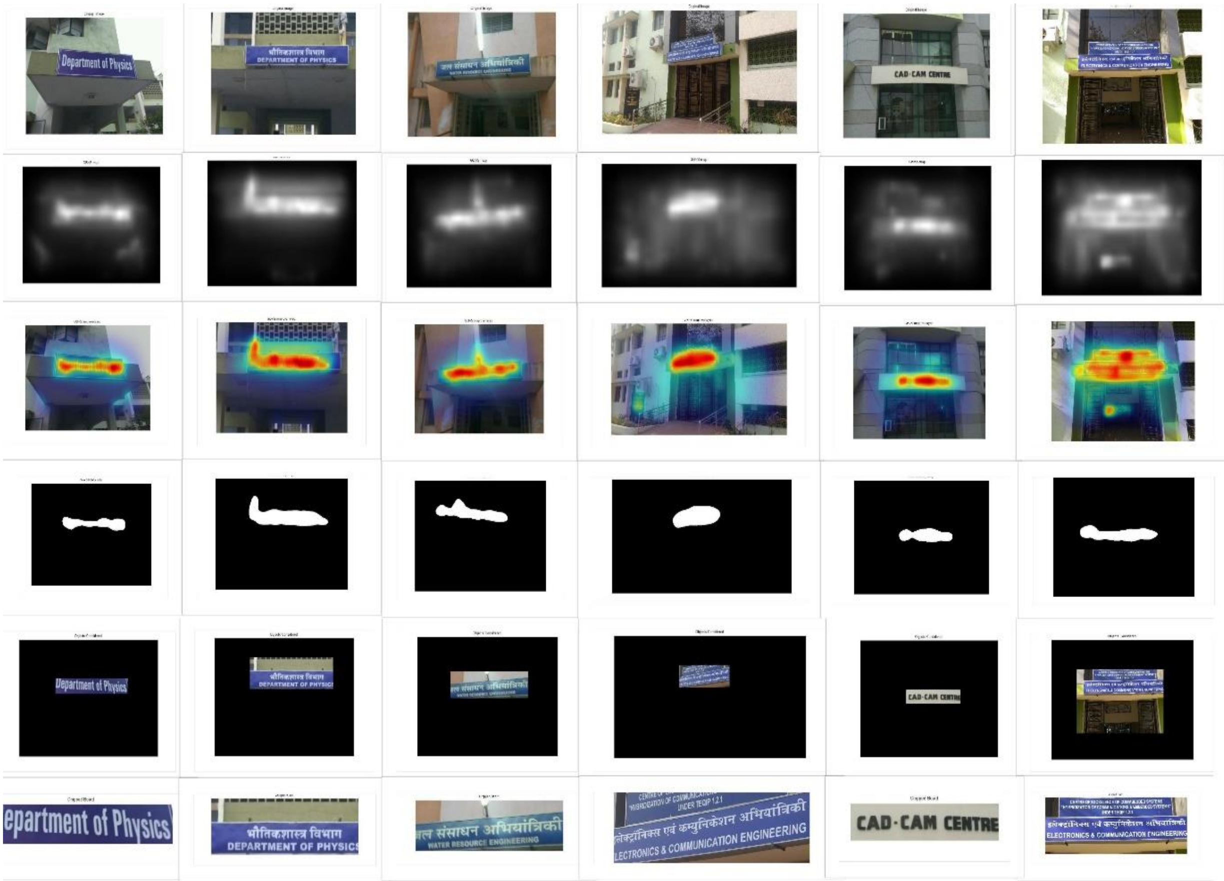

Figure 8-1 Results

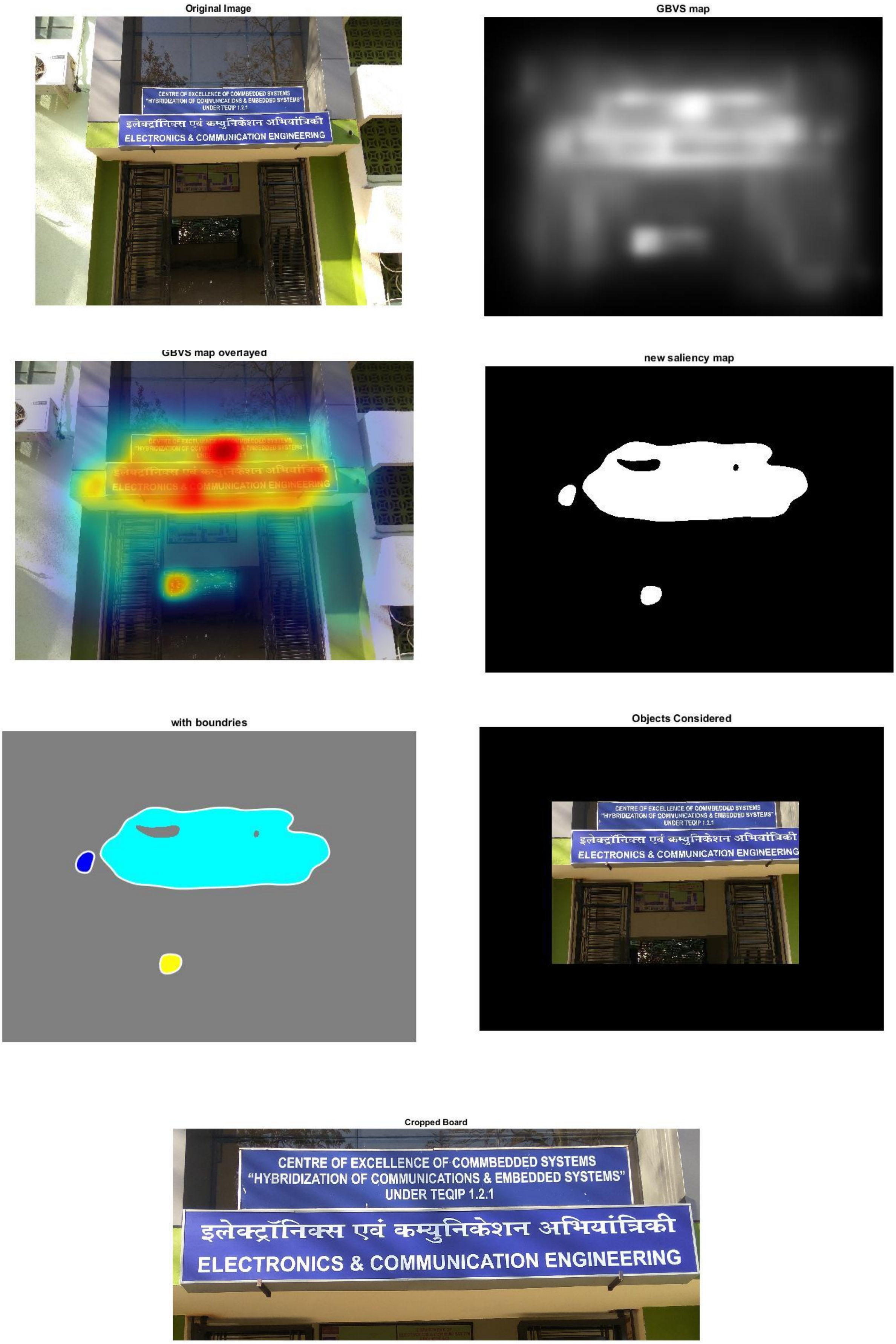

Figure 8-2 Results

9. DISCUSSION AND CONCLUSION

9.1 Discussion

We attempt to solve the problem of category independent salient sign board detection. We propose a time efficient approach which performs better than the recent state of art methods. The picture is clicked by the camera of a mobile phone which is then sent to cloud and processed by saliency technique. Inspired by the saliency, we propose detection and cropping of all salient objects in the image. After processing the image and extracting the sign board the characters are extracted and are linked with the location of sign board on the GPS map. Proposed application can be used in locating the position of the sign board on GPS map.

9.2 Conclusion

In thesis we have described a salient object detection technique and its application. In the following, we summarize the object detection method with its application and how they contribute to recent progress in Computer vision area. We discuss possible applications and important future extensions possible from our work.

9.3 Thesis summary

In summary, we propose an application using image saliency algorithm. Initially a photograph of a sign board is taken using an application. The location of the sign board is integrated with the image. The image is then sent to cloud system for further processing. The image is imported and here comes the saliency algorithm.

The method concentrates on finding the high level saliency information. The contrast feature map of intensity channel, the luminance variance (centre–surround) is computed in a local neighbourhood of the same scale. The next steps of the GBVS model are to form the activation map from each feature map based on graph theory (i.e., by using the adjacency matrix of the graph and iterations of the Markov matrix), and normalizing of the activation map is also based on the Markov matrix. After finding the saliency map the next step is to crop the object (sign board) from the original image. There is a possibility of having more than one salient object in an image. The object is thus selected on the basis of largest area of connected regions. Once the object (sign board) is cropped, it is linked the GPS map by extracting the location from the image.

The results on the real world images exhibit superiority of proposed application with the image processing method. These methods can provide significant advantage in many high-level Computer Vision tasks. 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.

9.4 Limitations

One main limitation comes from the issue related to object orientation in the image. Every image has some bias of feature distributions, depending upon the different sets of samples it contains. As the orientation of the object changes the sign boards are less precisely detected. The straight boards with angle 0^0 are most accurately detected but as the angle increases the probability of detection decreases. Following figure shows the above limitation.

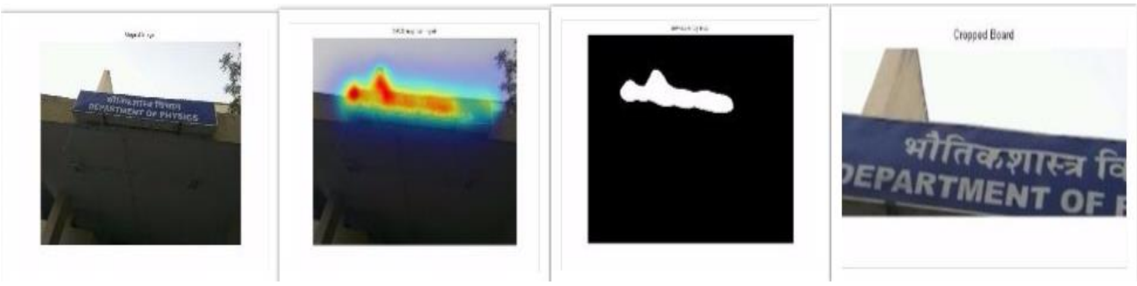

Figure 9-1 (a)Original image, (b)GBVS map overlayed, (c)Saliency map, (d)Cropped image .

Secondly when multiple objects are present, and area of the object which is not required is largest then that object is cropped out. The object of interest is not cropped.

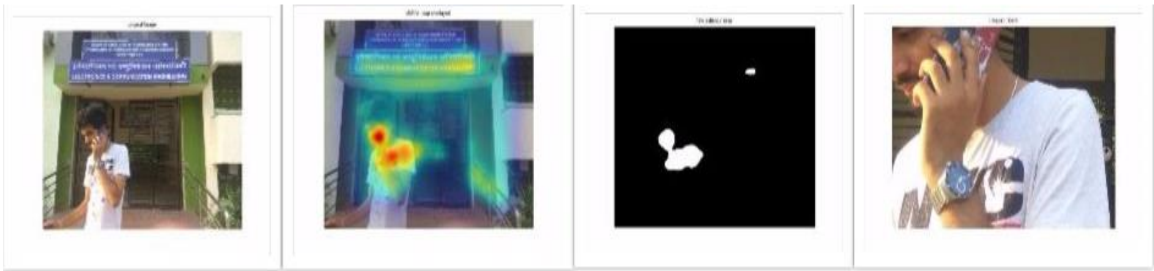

Figure 9-2 (a)Original image, (b)GBVS map overlayed, (c)Saliency map, (d)Cropped image

9.5 <u>Future Scope</u>

Future work will be focussed on more reliable and fast processing of the algorithm. All the limitations can be nullified by more efficient saliency map algorithm. The orientation problem can be solved by image filtering and alignment techniques. More work can be done on application part. We propose to implement the same algorithm on various signboard like road signs, number plate, etc.

In application part we would like to focus on detailing and improving the ease of access by developing a clean working mobile application which can be helpful in various location marking system. This thesis can be further used in many other machine learning applications like self-driving car, home delivery system.

Many multi-national companies around the world are developing machine learning techniques. This project can be used in development of such devices. Self-driving car can use such algorithm to detect the road signs.

10. REFERENCES

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

[02] C. Galleguillos and S. Belongie, “Context based object categorization: A critical survey,” Comput. Vis. Image Understanding, vol. 114, no. 6, pp. 712–722, 2010.

[03] G. Heitz and D. Koller, “Learning spatial context: Using stuff to find things,” in Proc. Eur. Conf. Comput. Vis., 2008, pp. 30–43

[04] A. Rabinovich and S. Belongie, “Scenes vs. objects: A comparative study of two approaches to context based recognition,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog., 2009, pp. 92–99

[05] M. J. Choi, J. Lim, and Torralba, “Exploiting hierarchical context on a large database of object categories,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2010, pp. 129–136.

[06] J. Deng, A. C. Berg, and L. Fei-Fei, “Hierarchical semantic indexing for large scale image retrieval,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2011, pp. 785–792.

[07] D.G. Lowe. Object recognition from local scale-invariant features. In ICCV, volume 2, pages 1150–1157, 1999.

[08] Tinne Tuytelaars and Krystian Mikolajczyk. Local invariant feature detectors: A survey. Found. Trends. Comput. Graph. Vis., 3(3):177– 280, July 2008.

[09] A. Collet, M. Martinez, and S. Srinivasa. The MOPED framework: Object Recognition and Pose Estimation for Manipulation. I. J. Robotic Res., 30(10):1284–1306, 2011.

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

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

[12] Harel J, Koch C, Perona P (2007) Graph-based visual saliency. In: Advances in Neural Information Processing Systems 19. MIT Press pp. 545–552.

[13] Torralba A, Oliva A, Castelhano MS, Henderson JM (2006) Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. Psychological Review 113:766–786

[14] Koch C, Ullman S (1987) Shifts in selective visual attention: Towards the underlying neural circuitry. In: Vaina LM, editor, Matters of Intelligence, Springer Netherlands. pp.115–141

[15] Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision 8:32.

[16] Butko N, Zhang L, Cottrell G, Movellan JR (2008) Visual saliency model for robot cameras. In: IEEE International Conference on Robotics and Automation. pp. 2398–2403.

[17] Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1597–1604.

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

[19] Hou, X. and L. Zhang, Saliency detection: A spectral residual approach. In CVPR. 2007.

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

[21] Harel, J., C. Koch, and P. Perona, Graph-based visual saliency. In NIPS. 2006.

[22] Achanta, R., S. Hemami, F. Estrada, and S. Susstrunk, Frequency-tuned salient region detection. In CVPR. 2009.

[23] 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.

[24] Bharath, Ramesh, et al. "Scalable scene understanding using saliency-guided object localization." Control and Automation (ICCA), 2013, 10th IEEE International Conference on. IEEE, 2013.

[25] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2, pp. 2169–2178, IEEE, 2006.

[26] A. Bosch, A. Zisserman, and X. Munoz, “Scene classification via plsa,” Computer Vision–ECCV 2006, pp. 517–530, 2006.

[27] Y. Li, J. Sun, C. Tang, and H. Shum, “Lazy snapping,” in ACM Transactions on Graphics (ToG), vol. 23, pp. 303–308, ACM, 2004

[28] E. Nowak, F. Jurie, and B. Triggs, “Sampling strategies for bag of- features image classification,” in Computer Vision ECCV 2006 (A. Leonardis, H. Bischof, and A. Pinz, eds.), vol. 3954 of Lecture Notes in Computer Science, pp. 490–503, Springer Berlin Heidelberg, 2006

[29]N. Bhargava, P. Trivedi, A. Toshniwal and H. Swarnkar, "Iterative Region Merging and Object Retrieval Method Using Mean Shift Segmentation and Flood Fill Algorithm," *2013 Third International Conference on Advances in Computing and Communications*, Cochin, 2013, pp. 157-160.

11. APPENDIX

11.1 Matlab Code:

```
clc
close all;
clear;
[file,path]=uigetfile({'*.jpg;*.bmp;*.png;*.tif'},'Choose an
image');%to choose image file from file explorer/image database
s=[path,file];
data=imfinfo(s);
img=imread(data.Filename);
img1=imresize(img,[400 800]);
%imshow(img1);
% if(size(img1,3)==3)
%     img1=rgb2gray(img1);
% end
%%
out_gbvs = gbvs(img);
subplot(3,3,1);
imshow(img);
title('Original Image');
subplot(3,3,2);
imshow( out_gbvs.master_map_resized );
title('GBVS map');
subplot(3,3,3);
show_imgnmap(img,out_gbvs);
title('GBVS map overlayed');

%% region of interest/ Morphological image processing
I=out_gbvs.master_map_resized;
% Opening by reconstruction

    se = strel('disk',5,0);
    Ie = imerode(I, se);
    Iobr = imreconstruct(Ie, I);

    if mean2(Iobr)==0 %rare case when sal. map is completely eroded
```

```
        fgm = I > 0;
        return;
    end

    % Regional maxima selects flat peaks
    fgm = imregionalmax(Iobr,18);

    % Discard regions with very low saliency values; Extra step not
    % included in the paper.
    discard_thresh = 0.5;

    labelimg = bwlabel(fgm);
    s = regionprops(labelimg, I, 'MeanIntensity');
    avg_sal = [s.MeanIntensity];
  % avg_sal = rescale(avg_sal,0,1);

    idx = find(avg_sal > discard_thresh);

    if ~isempty(idx)
    fgm = ismember(labelimg,idx);
    end

 subplot(3,3,4);
 imshow(fgm);title('ROI')

%%
%
%th=graythresh(I);
%h1=im2bw(I,th);
X1=im2bw(I);
se=strel('disk',5);
%opening and closing to remove small values
X2=imopen(X1,se);
X3=imclose(X2,se);

%% new saliency map
[r1 c1]=size(X3);
I1= zeros(r1,c1);
I2=zeros(r1,c1,3);
for i= 1:r1
```

```
    for j=1:c1
        if I(i,j)>=0.65
            I1(i,j)=1;
        end
        j=j+1;
    end
    i=i+1;
end
subplot(3,3,5);imshow(I1);title('new saliency map');
%detremining the boundries of objects
[B,L] = bwboundaries(I1,'noholes');

% Display the label matrix and draw each boundary
subplot(3,3,6);imshow(label2rgb(L, @jet, [.5 .5 .5]));title('with
boundries');
hold on
for k = 1:length(B)
  boundary = B{k};
  plot(boundary(:,2), boundary(:,1), 'w', 'LineWidth', 2)
end

%% considering area having objects only
 img2=im2double(img);
[rows, columns] = find(I1);
topRow = min(rows);
bottomRow = max(rows);
leftColumn = min(columns);
rightColumn = max(columns);
width=rightColumn-leftColumn+1;
height=bottomRow-topRow+1;
for k=1 :3
    for i=topRow:bottomRow+1
        for j=leftColumn:rightColumn+1
                 I2(i,j,k)=img2(i,j,k);
                 j=j+1;
        end
    i=i+1;
    end
k=k+1;
end
```

```
subplot(3,3,7);
imshow(I2);title('Objects Considered');

%% cropping the board area

CC=bwconncomp(I1);
LL=labelmatrix(CC);
SS=regionprops(LL);
allArea=[SS.Area];
maxArea=max(allArea);
mem1=find(allArea==maxArea);
h5=SS(mem1,1).BoundingBox;
I3=imcrop(img,h5);
subplot(3,3,8);imshow(I3);title('Cropped Board');

%% location detection
loca=struct;
loca=data.GPSInfo;
lat=loca.GPSLatitude;
lon=loca.GPSLongitude;
x=dms2degrees(lon);
y=dms2degrees(lat);
figure();
plot(x,y,'.b','MarkerSize',20);
plot_google_map('APIKey','AIzaSyCoWw98OWhxLya9Q7bUjaBAV22NN8p4HCk','Sc
ale',2,'MapType','hybrid','Showlabels',1);
%%
%filtering/masking
% th=graythresh(ans);
% ObjectMask=~im2bw(ans,th);
% cc = bwconncomp(ObjectMask);
% stats = regionprops(cc ,h1,'Area','BoundingBox');
% A = [stats.Area];
% [~,biggest] = max(A);
% ObjectMask(labelmatrix(cc)~=biggest) = 0;
% ObjectMask = imfill(ObjectMask,'holes');
% subplot(2,3,4);
% imshow(ObjectMask);title('mask');
%%
```