IMPROVED SALIENCY OF CLUTTERED IMAGES USING STRUCTURE EXTRACTION FROM TEXTURE

BY:

KABOOL NEUPANE

072/MSI/606

SUPERVISED BY:

Dr. SANJEEB PRASAD PANDAY

A THESIS SUBMITTED TO DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

PULCHOWK CAMPUS

INSTITUTE OF ENGINEERING

TRIBHUVAN UNIVERSITY

LALITPUR, NEPAL

NOVEMBER, 2017

COPYRIGHT

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

The undersigned certify that they have read and recommended to the Department of Electronics and Computer Engineering for acceptance, a thesis entitled "**Improved Saliency of Cluttered Images Using Structure Extraction from Texture**", submitted by **Kabool Neupane** in partial fulfillment of the requirement for the award of the degree of "**Master of Science in Information and Communication Engineering**".

…….

Supervisor, Dr. Sanjeeb Prasad Panday

Department of Electronics and Computer Engineering,

Institute of Engineering, Tribhuvan University.

…….

External Examiner, Dr. Pradip Paudyal

Assistant Director,

Nepal Telecommunications Authority.

…….

Committee Chairperson, Dr. Dibakar Raj Pant

Head of Department,

Department of Electronics and Computer Engineering,

Pulchwok Campus, Institute of Engineering.

Date: ……

DEPARTMENTAL ACCEPTANCE

The thesis entitled **"Improved Saliency of Cluttered Images Using Structure Extraction from Texture"**, submitted by **Kabool Neupane** in partial fulfillment of the requirement for the award of the degree of **"Master of Science in Information and Communication Engineering"** has been accepted as a bonafide record of work independently carried out by him in the department.

……

Dr. Dibakar Raj Pant

Head of Department,

Department of Electronics and Computer Engineering,

Pulchowk Campus, Institute of Engineering,

Lalitpur, Nepal.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor **Dr. Sanjeeb Prasad Panday** for his encouragement, suggestions and continuous guidance throughout the course of my thesis work.

I am thankful to our program coordinator **Dr. Basanta Joshi** for providing us with a suitable platform for thesis work. I would also like to thank **Dr. Dibakar Raj Pant**, Head of Department of Electronics and Computer Engineering, Pulchowk Campus, for his precious support and guidance. I am highly indebted to **Prof. Dr. Subarna Shakya** for his insights and opinions regarding the thesis work.

I would also like to thank all of my classmates and faculty of Department of Electronics and Computer Engineering for providing me their views and ideas regarding the thesis work.

ABSTRACT

Automatic salient region detection in an image is a useful technique that can assist in many computer vision tasks such as image segmentation and object recognition. It allows processing of input images without prior information of its contents. In this thesis work, a salient object detection approach which aims in limiting the distractions caused by small patterns in the background and foreground of an image is considered. First, a structure extraction algorithm is employed to smooth the textures present in the input image while preserving its structure information. Second, the structure image is segmented into perceptually homogenous regions by using graph-based image segmentation. Third, saliency map is computed according to color contrast and complimentary priors. The performance of the proposed method is compared against the prior method (without structure extraction) with the help of salient object detection datasets. Quantitative evaluation of the saliency maps are conducted using receiver operating characteristics (ROC) curve, overlap ratio (OR), weighted F-measure score and structure similarity (SSIM) index. The proposed method obtained a weighted-F score of 0.643 as compared to the 0.607 obtained by the prior method in the case of Microsoft Research Asia (MSRA10k) dataset. The same evaluation measure was improved from 0.509 to 0.534 in the case of Extended Complex Scene Saliency Dataset (ECSSD) dataset. Similarly, the SSIM index score obtained was 0.751 and 0.648 as compared to 0.539 and 0.516 obtained by the prior method for each datasets respectively. The comparison of the proposed method with three other existing salient region detection methods is also shown. The proposed method obtained a competitive score for MSRA10k dataset images whereas it obtains the highest score considering OR, weighted-F measure and SSIM index in the case of ECSSD dataset with a score of 0.551, 0.534 and 0.648 respectively.

Keywords: Saliency detection, bottom-up method, structure extraction

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AUC	Area Under the Curve
CIE	Commission Internationale de l'Eclairage
ECSSD	Extended Complex Scene Saliency Dataset
FN	False Negative
FP	False Positive
FPR	False Positive Rate
MAE	Mean Absolute Error
MSRA	Microsoft Research Asia
OR	Overlap Ratio
ROC	Receiver Operating Characteristics
SPD	Symmetric Positive Definite
SSIM	Structure Similarity
TN	True Negative
TP	True Positive
TPR	True Positive Rate

CHAPTER 1

INTRODUCTION

1.1 Background and motivation

Modern day has astounding amount of visual data and information available and is generated every minute. The growth in image data has led to new challenges of processing these data fast and extracting proper information. This extracted information helps to facilitate different tasks such as image retrieval, object detection and object recognition. One peculiar 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 capacity to extract crucial information from a scene. This ability permits humans to center their perceptual and cognitive resources on the most appropriate subset of the available visual data, aiding them to learn and survive in everyday life. This amazing ability is known as visual saliency. Motivated from this fact, it is worthwhile to design a computer vision system that is able to detect saliency such that resources can be utilized properly for processing important visual data.

Saliency is the quality of a region in an image to standout (or be noticeable) from the rest of the scene and grab our attention. There are two primary approaches for salient object detection, bottom-up saliency and top-down saliency. Bottom-up methods find any object of importance based on low-level cues, without any prior information about its kind and category. Top-down method is task dependent and include context and feature based cues amongst others. For instance, while searching for a car, we would call upon our knowledge of what a car looks like and would drive our attention. We can further classify bottom-up algorithms into local and global schemes. Local contrast based methods examine the rarity of image regions with respect to (small) local neighborhoods. On the other hand, global contrast based methods evaluate saliency of an image region using its contrast with respect to the whole image.

1.2 Problem statement

Human visual system has a tremendous capacity to draw out important information from a scene. However, the same task is still a challenge for computer vision systems. Hence, designing a system that can handle visual data efficiently to extract meaningful information is necessary.

When dealing with images containing cluttered background, small patterns in foreground and background of the image adversely affects saliency detection. Such patterns commonly occur in natural images. Therefore, implementation of a suitable algorithm that would address this situation is necessary.

1.3 Objectives

- To implement structure extraction algorithm for suppressing small scale patterns in the input image while retaining its structure information
- To implement salient region detection algorithm using color contrast and complimentary priors
- To compare the performance of the method with and without the use of structure extraction by using evaluation measures like OR, weighted-F measure and SSIM

1.4 Scope of the work

Extracting information from images to find objects of interest is one of the fundamental tasks in computer vision system. The proposed framework can assist in detecting a salient region from an image. It also addresses one of the challenges in saliency detection which is detecting a salient region from an image containing cluttered background. It does so by first obtaining the structure image from the given input image. The structure image is then used to compute the saliency map.

1.5 Organization of the thesis

The rest of the thesis is organized as follows. A survey of related work on salient region detection is presented in Chapter 2. In Chapter 3, theory related to the proposed method is discussed. The proposed framework for obtaining the saliency map along with related algorithms for each stage is explained in Chapter 4. In Chapter 5, the quantitative results obtained for the proposed method is compared against the results obtained using the prior method. Conclusion and possible future recommendation for the proposed method is presented in Chapter 6. An example of generating a symmetric positive definite Laplacian matrix from a given input image along with saliency maps obtained from the considered methods are presented in Appendix A and Appendix B respectively.

CHAPTER 2

LITERATURE REVIEW

Saliency stems from visual uniqueness and is usually attributed to variations in image at features like color, intensity and texture. Various approaches involving pre-attentive bottom-up saliency region detection are either biologically motivated, purely computational, or involve both aspects. The works related to local contrast based method is discussed at first. It is followed by discussion of work involving global contrast based approach and their limitations.

A visual attention system inspired by behavior and neuronal architecture of early primate was presented in the paper [1]. The author presented a method to combine multi-scale image features into a single saliency map. The model uses three features, namely color, intensity and orientation, similar to the simple cells in primary visual cortex. Ma and Zhang [2] proposed an alternative algorithm to local contrast analysis for generating saliency maps, which are extended using fuzzy growth model. At first, saliency map is generated based on local contrast analysis. It is followed by extraction of objects from the saliency map by using fuzzy growing methods which simulates human perception.

Liu et al. [3] proposed a model to find multi-scale contrast by linearly combining contrast in a Gaussian image pyramid. The authors proposed a set of novel features including multi-scale contrast and color spatial distribution to describe a salient object. Goferman et al. [4] proposed context aware saliency which aims at detecting the image regions that represent the scene. It simultaneously models local low-level clues, global considerations and high-level features to highlight the salient objects.

These methods tend to produce higher saliency values near edges instead of uniformly highlighting salient objects. A global contrast based method, which separates a large-scale object from its surroundings, is desirable over local contrast based methods producing high saliency values at or near object boundaries. Global considerations enable assignment of comparable saliency values across similar image regions, and can uniformly highlight entire objects.

Zhai and Shah [5] defined pixel-level saliency by contrast to all other pixels. The authors used image histograms to calculate color saliency. In their proposed model, both spatial and temporal saliency maps are constructed in case of video sequences and fused in a dynamic fashion to produce the overall spatio-temporal attention model. However, for efficiency they use only luminance information, thus ignoring distinctiveness clues in other channels.

Achanta et al. [6] proposed a frequency tuned method that directly defines pixel saliency using the color differences from the average image color. This method exploits features of color and luminance while boundaries are preserved by retaining substantially more frequency content from the original image. This approach, however, only considers average color, which can be insufficient to analyze complicated variations common in natural images.

Perazzi et al. [7] proposed a contrast based filtering algorithm and defined superpixel-level saliency by combining two contrast measures. Two measures of contrast are computed that rates the uniqueness and the spatial distribution of the elements. From the element contrast, a saliency measure is derived that produces a saliency map which uniformly covers the objects of interest and consistently separates foreground and background. However, this approach ignores spatial relationships across image regions, which can be crucial for reliable and coherent saliency detection.

More and more complimentary priors are being used along with the bottom up method to improve the detection of salient region. Boundary prior [8], semantic and center prior [9] have been used in saliency detection. Center prior is based on assumption that object near the image center is more likely to be salient. In this thesis work, the problem of detecting a salient region from an image containing cluttered background is addressed using structure extraction from texture algorithm. Center prior and spatial prior is incorporated into the proposed framework for improving the detection of salient region in the input image.

CHAPTER 3

RELATED THEORY

3.1 Human visual system and modeling

Human visual system modeling requires complete and accurate knowledge about the entire visual pathways. In present, only certain aspects of vision are understood and so, human visual system models have been developed in order to simplify the behaviors of what is a very complicated system.

There are two types of photoreceptor cells: rods and cones, and they have different functions. Rods are found primarily in the periphery of the retina and are not sensitive to color. It is sensitive only to light and dark or to black and white. Cones are located in the center of the retina and are used to differentiate color at normal levels of light. There are three types of cones that differ in the wavelengths of light they absorb; they are usually called short or blue (S), middle or green (M), and long or red (L). The photoreceptor cells convert the light energy into neural signals. From the three primaries given by cones and the intensity given by rods, the color is eventually encoded as one luminance channel (white-black) and two other channels: one for red-green (R-G) and the other for blue-yellow (B-Y) cones. The color opponent cells respond best to local color contrast.

Many cells in the human visual system and mainly in the visual cortex have been proven to be selectively sensitive to certain types of signals such as patterns of a particular frequency or orientation. The visual cortex is made from the combination of several areas: V1 (or primary visual cortex), V2, V3, V4, and V5. Neurons in the visual cortex respond to visual stimuli that appear within their receptive field. The receptive field of one neuron is the region within the entire visual field which causes a response from that neuron. First visual areas (for example V1 area) have neurons with simpler tuning that will respond to stimuli falling in their receptive fields such as vertical lines or textures with particular spatial frequencies. In later visual areas, neuronal cells have complex tuning that is much more complicated to simulate [10]. By utilizing this information, different models are implemented to mimic functions of human visual system.

3.2 Sparse matrix

A sparse matrix is an n×n matrix which is mostly zeros as the matrix grows in size. The interest in sparse matrices started in the late 1950s when researchers in linear programming and electrical power system began to solve linear equations using a computer. It was found during the research that when realistic problems were modeled by system of linear equations the resulting matrices were sparse with highly structured non-zero patterns. That is, the non-zero elements of these sparse matrices were arranged in a definite pattern. The formulated problems had variables of size thousands (and higher) but whose coefficients were mostly zeros. It was realized that sparsity had to be exploited for such problems to be solved on a computer. Structure and sparsity are both exploited in order to save computation time and space [11].

3.3 Sparse symmetric positive definite matrices

Many real, physical systems are modeled using symmetric positive definite (SPD) matrices. A matrix ''A'' is said to be symmetric if it is equal to its transpose (A^T). Similarly, a matrix is said to be SPD if all of its eigen values are real and positive. When the eigen values of a SPD matrix is positive then all the pivots are also positive. The determinant of such matrix, calculated by using the product of the pivots, is then real and positive. Hence, linear systems involving these matrices can be easily solved using different direct and iterative methods. As opposed to positive definite matrices, we have to consider for dynamic storage and numerical instability in the case of indefinite matrices. An example of a 6×6 sparse SPD matrix is given below with the help of matrix A. It has eigen values 1.0000, 1.0564, 1.2469, 1.5721, 1.6523, and 2.4835.

$$A[6,6] = \begin{bmatrix} 1.3364 & -0.2903 & -0.0461 & 0 & 0 & 0 \\ -0.2903 & 1.3290 & 0 & -0.0387 & 0 & 0 \\ -0.0461 & 0 & 1.29210 & -0.2032 & -0.0429 & 0 \\ 0 & -0.0387 & -0.2032 & 1.4741 & 0 & -0.2323 \\ 0 & 0 & -0.0429 & 0 & 1.6951 & -0.6522 \\ 0 & 0 & 0 & -0.2323 & -0.6522 & 1.8845 \end{bmatrix}$$

In this thesis work, for an image of size 400×300, a sparse symmetric positive definite matrix of size 120,000×120,000 containing around 598600 non-zero elements is obtained.

3.4 Graph theory and representation of an image

A graph is a set of elements and a set of pair wise relationship between these elements of the graph. The elements are called nodes or vertices and the relationship between these nodes is called an edge. Mathematically, a graph G can be defined using two sets as G= (V, E) where the set V consists of vertices and the set E consists of edges in the graph G respectively. The i-th vertex of the graph is denoted by $v_i \in V$ and the edge between the i-th and j-th vertex is denoted as $e_{ij} \in E$. The graph can be viewed as weighted. Given the graph G, the weight of an edge incident to two vertices is denoted by $w(v_i, v_j)$ or w_{ij}.

In the graph based approach, the image pixels are modeled as nodes of the graph and the similarity between these elements give the edge weights. For a color image, the similarity between the pixels can be calculated using Equation 3-1.

$$Weight = \sqrt{(R1 - R2)^2 + (G1 - G2)^2 + (B1 - B2)^2} \quad (3\text{-}1)$$

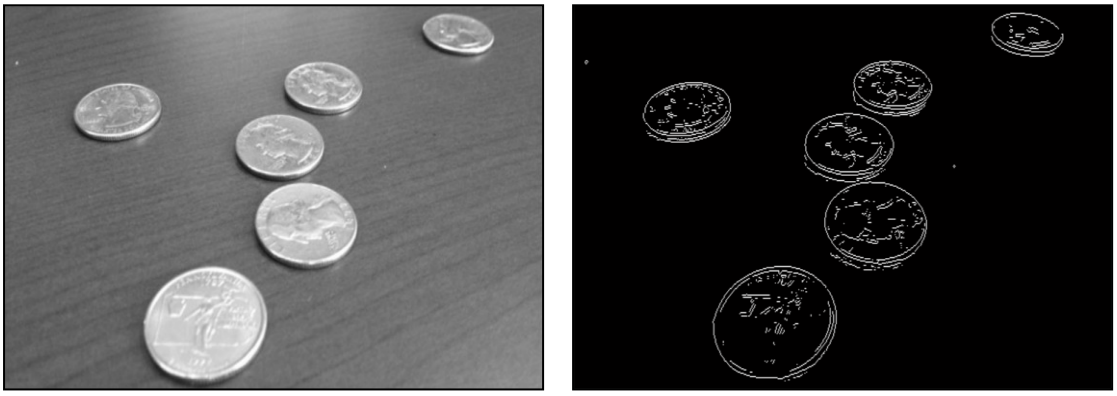
Where, R, G and B are the red, green and blue components of pixels for color image.

3.5 Image segmentation

Image segmentation is the segregation of an image into a set of regions such that each region represents a purposeful area having similar features and properties. When a region of interest doesn't cover the whole image, the segmentation can be used to partition the image into foreground and background regions. It breaks down an image into regions for further analysis. For example, we can segment a human face from an image and use it further for person identification. Segmentation of an image can be used for a change in representation. That is, the pixels are organized into regions such that each region carries more perceptual information.

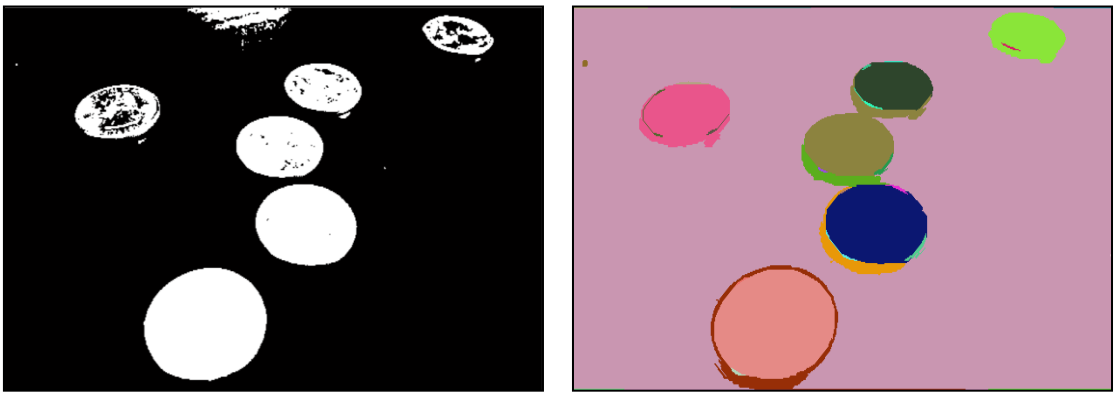
The regions of image segmentation should be consistent and uniform with respect to some characteristics such as gray level, color and texture. The adjacent regions should be different in the sense that pixels belonging to these two regions should be dissimilar based on some predefined criteria. The segmentation approaches can be divided into two types based on the properties of the image.

The discontinuity based approach segments an image based on the discontinuity. In edge detection based technique, the discontinuities in intensity of pixels are detected. It is then connected to form the boundary of a region. An example of edge detection is shown with the help of Figure 3-1. In similarity detection based approach, the image is segmented into regions, each having a set of comparable pixels. The techniques that fall under this approach include segmentation using thresholding, region growing techniques and region splitting and merging. Clustering technique also follows this methodology and divides an image into regions (clusters) having similar attributes. An example of segmentation for image shown in Figure 3-1 (a) using thresholding technique and graph based technique are shown with the help of Figure 3-2.



a) b)

Figure 3-1: Edge detection using Sobel filter a) Input image b) Edge detected by filter



a) b)

Figure 3-2: Image segmentation using a) Thresholding technique b) Graph based technique

3.6 CIELAB color space

It is the color space defined by CIE (Commission Internationale de l'Eclairage) based on one channel for Luminance (L) and two color channels (A and B). CIELAB (Lab or L*a*b*) is an opponent color system which correlates with the discovery that somewhere between the optical nerve and the brain, the retinal color stimuli are translated into distinctions between light and dark, red and green, and blue and yellow [12]. These three values are indicated in the three axes L*, a* and b* respectively. The values on the vertical L* axis runs from 0 to 100 whereas the values on the horizontal a* and b* axes runs from -128 to +127. The color axes are based on the fact that a color can't be both red and green, or both blue and yellow, because these colors oppose each other. The advantage of using CIELAB color space is that the distance between individual colors corresponds to the perceived color differences. The CIELAB color model is shown with the help of Figure 3-3.

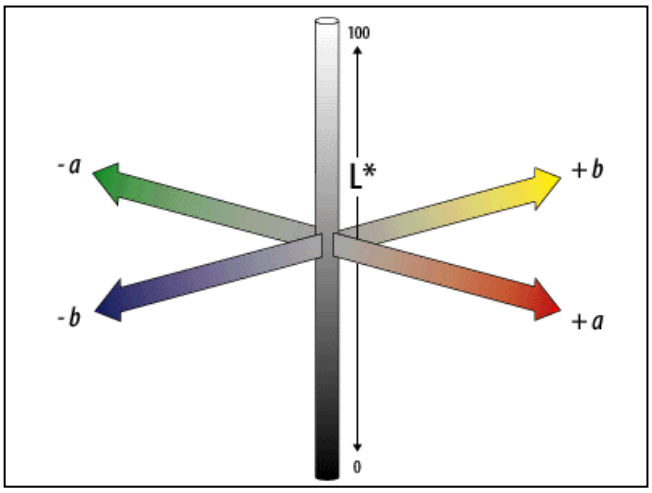


Figure 3-3: CIELAB color space
[Source: http://dba.med.sc.edu/price/irf/Adobe_tg/models/cielab.html]

3.7 Color distance metric

Most of the bottom up method uses contrast as an important cue. This approach works well when there is high contrast between the foreground and background image elements. The distance or difference between two colors is a metric of interest. It can be used to quantify the color contrast between two regions in an image.

Euclidean distance between two colors in a device independent color space is frequently used for salient region detection and can be expressed using Equation 3-2.

$$distance = \sqrt{(L_1 - L_2)^2 + (A_1 - A_2)^2 + (B_1 - B_2)^2} \quad (3\text{-}2)$$

Where L, A and B are the components of each channel in CIELAB color space.

3.8 Superpixel representation of an image

Images are represented as a grid of pixels. The pixel grid system of representation of image is by no means a natural representation of the image. If a single pixel is taken from an image and presented to a person, it will neither have any semantic meaning nor tell which part of the image it represents. So, it would intuitively make sense to explore the semantic meanings of an image which is formed by grouping a set of pixels. When we perform a grouping of pixels we arrive at superpixels which would carry more perceptual and semantic meaning as opposed to the single pixel [13]. It has the following advantages.

- It helps in reducing the complexity of images from a hundred of thousands of pixels to only a few hundred of superpixels
- The pixels that belong to a superpixel represent some perceptual and semantic meaning and share some common features such as color or texture distribution
- The use of superpixel algorithms over segment the image. It means that important boundaries in the image can be easily found.

An example of an input image and the corresponding boundaries obtained after using a superpixel algorithm is shown with the help of Figure 3-4.

Figure 3-4: Superpixel representation of an image and its boundaries

CHAPTER 4

METHODOLOGY

4.1 Proposed framework

The proposed framework for obtaining the final saliency map for a given input image is shown with the help of Figure 4-1. The input image is collected from datasets designed for evaluation of salient object detection. Structure extraction algorithm is then applied to suppress the small scale patterns and textures in the input image. The structure image is then segmented using graph-based image segmentation. Finally, a saliency map is obtained as output by using color contrast and complimentary priors.

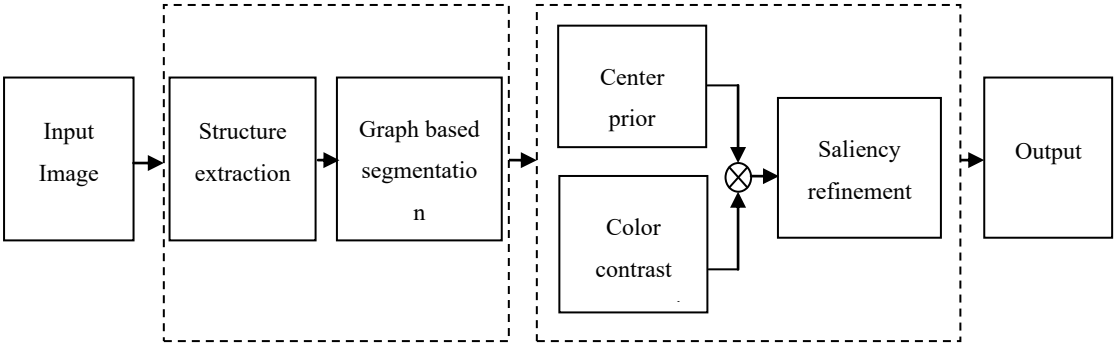


Figure 4-1**:** Proposed framework for obtaining the saliency map

4.2 Data collection

Different datasets are available for evaluation of salient object detection. These dataset consists of the original image as well as the ground truth with salient object labeling. Datasets that are used for salient object detection during this thesis work are given below.

- Extended Complex Scene Saliency Dataset (ECSSD) [14]
- Microsoft Research Asia (MSRA10k) dataset [15]

The MSRA10K benchmark dataset comprises of per-pixel ground truth annotation for 10, 000 MSRA images. Although images from MSRA10K have a large variety in their content, background structures are primarily simple.

On the other hand, ECSSD consists of 1000 images that represent the situations that natural images generally fall into and includes many meaningful but structurally complex images for evaluation.

4.3 Structure extraction

When human detect salient regions in natural scenes, small scale textures, such as grass and foliage, do not catch attention at first glance. For example, we are attracted by the black and white ball when seeing Figure 4-2 (a) at first glance, while we ignore the small variations of grass blades surrounding the ball. Hence, structure extraction is employed to smooth out such textures which in turn help in obtaining a better saliency result.

(a) Input image (b) Structure image

Figure 4-2**:** Structure image for a given input image

In the paper [16], the authors showed that with regard to relative total variation, texture and main structure exhibit completely different properties, making them decomposable. The advantage of using this method is that assumption regarding regularity or symmetry of the texture patterns is not required. The input image (I) can be decomposed into structure image (S) and its texture component (T) using structure extraction algorithm. The objective function is given with the help of Equation 4-1.

$$\arg\min_{S}\left[\sum_{p}(S_p - I_p)^2 + \lambda\left(\frac{D_x(p)}{L_x(p)+\varepsilon} + \frac{D_y(p)}{L_y(p)+\varepsilon}\right)\right] \tag{4-1}$$

Where, I is the input image, p is the index for image pixels and S is the resulting structure image. The effect of removing texture from an image is introduced by the regularizer weighted by λ which is called the relative total variation.

The objective function in Equation 4-1 is non-convex. Its solution thus cannot be obtained trivially. An optimization procedure using iterative method is followed. It involves solving a series of linear Equation of the form given by Equation 4-2.

$$Ax = B \tag{4-2}$$

Where, A is a symmetric positive definite Laplacian matrix. B is initialized using the input image. The result x obtained is the required structure image. In the successive iterations, the structure image obtained in the previous iteration (x^{n-1})is replaced for B. The steps for obtaining the matrix A are shown in Appendix A with the help of an example.

Windowed total variation and windowed inherent variation which is combined to form the relative total variation is discussed in the following section.

4.3.1 Windowed total variation

A general pixel-wise windowed total variation measure is given with the help of Equation 4-3 and Equation 4-4.

$$D_x(p) = \sum_{q \in R(p)} g_{p,q} . \left|(\partial_x S)_q\right| \tag{4-3}$$

$$D_y(p) = \sum_{q \in R(p)} g_{p,q} . \left|(\partial_y S)_q\right| \tag{4-4}$$

$$g_{p,q} = \exp\left(-\frac{(x_p - x_q{}^2) + \left(y_p - y_q\right)^2}{2\sigma^2}\right) \tag{4-5}$$

Where, q belongs to R(p), the rectangular region centered at pixel p. $g_{p,\,q}$ is a weighting function given by Equation 4-5. $D_x(p)$ and $D_y(p)$ are windowed total variations in the x and y directions for pixel p, which count the absolute spatial difference within the window R(p).

4.3.2 Windowed inherent variation

To help distinguish prominent structures from the texture elements, besides D, a windowed inherent variation is used. It is expressed with the help of Equation 4-6 and Equation 4-7.

$$L_x(p) = \left| \sum_{q \in R(p)} g_{p,q} . (\partial_x S)_q \right| \tag{4-6}$$

$$L_y(p) = \left| \sum_{q \in R(p)} g_{p,q} . (\partial_y S)_q \right| \tag{4-7}$$

4.4 Graph based image segmentation

After the structure extraction step, the structure image is segmented into regions using graph-based image segmentation [17] method. It uses a graph-based representation of the image. For a given pixel p, its 8-neighbours are used during the computation. The implementation maintains the segmentation using union by rank and path-compression.

The input is a graph G = (V, E), with n vertices and m edges. The output is a segmentation of V into components S = (C_1,...,C_r). The output is generated by following the steps given below.

1. Sort the edges by non-decreasing edge weight
    - $\text{Weight(RGB image)} = \sqrt{(R1 - R2)^2 + (G1 - G2)^2 + (B1 - B2)^2}$
2. Start the segmentation where each vertex is in its own component
3. Repeat 4 for each sorted edges and update the threshold after a merge operation
4. Components containing node1 and node2 of the graph is merged if and only if
    - weight(node1) <= threshold(parent_node1)
    - weight(node2) <= threshold(parent_node2)
    - node1 and node2 belong to two different components of the graph

4.5 Saliency-map computation

Saliency map for each region is computed by following the procedure [15] described as follows. For a region r_k, its saliency value is computed by measuring its color contrast to all other regions in the image. It is expressed with the help of Equation 4-8.

$$S(r_k) = \sum_{r_k \neq r_i} w(r_i) D_r(r_k, r_i) \tag{4-8}$$

Where, $w(r_i)$ is the weight of region r_i and $D_r(\cdot,\cdot)$ is the color distance metric between the two regions. The number of pixels in r_i is used as $w(r_i)$ to emphasize color contrast to bigger regions. The color distance between two regions r_1 and r_2 is defined using the Equation 4-9.

$$D(r_1, r_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f(c_{1,i}) f(c_{2,j}) D(c_{1,i}, c_{2,j}) \tag{4-9}$$

Where, $f(c_{k,i})$ is the probability of the i-th color $c_{k,i}$ among all n_k colors in the k-th region r_k with k= {1, 2}. $D(c_i,c_j)$ is the color distance metric between colors c_i and c_j in the CIELAB color space. To reduce the number of colors needed to consider, each color channel is divided into 12 different levels.

4.5.1 Spatial weight and center-bias

To further incorporate spatial information, a spatial weighting term and a center-bias term is introduced in Equation 4-8. The resulting equation is given using Equation 4-10.

$$S(r_k) = w_s(r_k) \sum_{r_k \neq r_i} exp\left(\frac{-D_s(r_k, r_i)}{\sigma_s^2}\right) w(r_i) D_r(r_k, r_i) \tag{4-10}$$

Where, $D_s(r_k, r_i)$ is the spatial distance between regions r_k and r_i , and σ_s controls the strength of spatial weighting. In this implementation, σ_s^2 =0.4 is used. The spatial distance between two regions is defined as the Euclidean distance between the centroids of the respective regions.

This spatial weight increases the effects of closer regions and decreases the effects of farther regions. Similarly, $w_s(r_k)$ in Equation 4-10 is the center bias term. We use $w_s(r_k) = \exp(-9d_k^2)$, where d_k is the average distance between pixels in region r_k and the center of the image. Thus, $w_s(r_k)$ gives a high value if region r_k is close to the center of the image and it gives a low value if the region is a border region away from the center.

4.5.2 Saliency refinement

As the final step, the saliency map is refined by using color space smoothing. We replace the saliency value of each color by the weighted average of the saliency values of similar colors (measured by CIELab color distance). It is given with the help of Equation 4-11.

$$S(c) = \frac{1}{(m-1)T} \sum_{i=1}^{m} (T - D(c, c_i))\ S(c_i) \tag{4-11}$$

Where, $T = \sum_{i=1}^{m} D(c, c_i)$ is the sum of distances between color c and its m nearest colors. The value of m is chosen to be equal to n/10, where n is the total number of colors in the image. This varying smoothing weight assigns a larger weight to colors closer to c in the color feature space.

4.6 Tools

Different computational tasks in this research are computed using Matrix Laboratory (MATLAB 2013b) and Python (Python 3.5).

CHAPTER 5

RESULT AND DISCUSSION

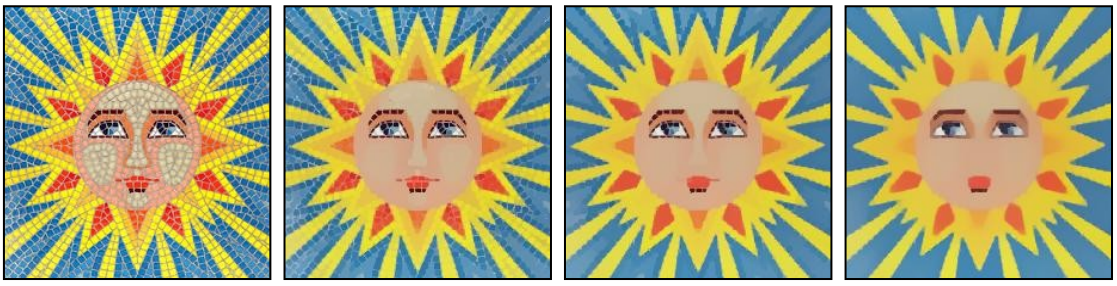
5.1 Experimental results

To demonstrate the performance of the saliency detection algorithm two datasets, ECSSD and MSRA10k, are used. The images in MSRA10k dataset have a resolution of 400×300 and the images in ECSSD dataset have a resolution of 400×267. At first, structure extraction algorithm is applied to the input image. The main parameter required for the structure extraction algorithm is lambda (λ). The parameter λ controls the smoothness of the output image. The output images for different values of this parameter (with σ fixed) are shown with the help of Figure 5-1.

(a) Input image (b) λ=0.01, σ=1 (c) λ=0.03, σ=1 (d) λ=0.05, σ=1

Figure 5-1: Structure extraction results for different values of the parameter λ

In contrast, spatial parameter σ in Equation 4-5 controls the window size for gaussian weighting function. It is used for specifying the size of the texture elements. The value of σ is decreased after completing each iteration for preserving edge sharpness. The output images for different values of this parameter (with λ fixed) are shown using Figure 5-2.



(a) Input image (b) σ=0.5, λ=0.01 (c) σ=2, λ=0.01 (d) σ=5, λ=0.01

Figure 5-2: Structure extraction results for different values of the parameter σ

In the structure extraction algorithm, the third parameter required is the number of iterations (no_iter). The algorithm requires 3-5 iterations to suppress the small scale patterns and textures in the image. The intermediate results up to three iterations are shown using Figure 5-3.

(a) Input image (b) First iteration (c) Second iteration (d) Third iteration

Figure 5-3**:** Structure image in different iterations

After completing the structure extraction, the structure image is then segmented using graph-based image segmentation. The segmentation is maintained using a disjoint-set forest with union by rank and path compression. The segmentation starts with each vertex in its own component. If a merge operation occurs between two components of a graph, parent of the component with higher rank is selected as the representative of that new component. The rank of a parent increases only if two components having a parent of equal rank are merged. The main parameters required by the segmentation algorithm are sigma (σ) and the k.

The parameter sigma is used by the gaussian filter to preprocess the image. The parameter k sets a scale of observation, in that a larger k causes a preference for larger components. The results for using different values of k are shown with the help of Figure 5-4.

(a) Input (b) k = 280 (c) k = 800 (d) k = 1500

Figure 5-4**:** Segmentation using different values of k

After completing the image segmentation, saliency map for the input image is calculated. The commonly employed assumption is that, regions which stand out from the surrounding ones, catch our attention and should be labeled salient. Therefore, saliency for a region is computed by measuring its color contrast to all other regions in the image. The values of the parameters used during the implementation of proposed framework are presented with the help of Table 5-1.

Table 5-1: Parameters used during implementation of framework

(λ, σ , no_iter) structure extraction	(σ, k) segmentation	σ_s^2 spatial weight
(0.03, 1, 4)	(0.6, 250)	0.4

Natural images consist of small scale textures, such as grass and foliage and can be smoothed by the structure extraction algorithm by using the parameters specified in Table 5-1. The dataset image does not explicitly contain big textures and rather contains small patterns for a given natural image. Therefore a lower value of the parameter σ and a total of 4 iterations are used by the structure extraction algorithm. The values used for the parameters λ and σ are 0.3 and 1 respectively. Structure images obtained using these values are smooth without any blurring around the edges for the dataset images.

The parameters specified for the image segmentation decomposes an image into a number regions for performing region wise color contrast. Total number of regions obtained for original and the structure image using the image segmentation parameters are shown with the help of Figure 5-5.

Original image Total regions= 382 Structure image Total regions= 175

Figure 5-5: Total number of regions for original image and structure image

From the results shown in Figure 5-5, it can be seen that by smoothing the input image using structure extraction algorithm, the total number of regions which has to be evaluated for obtaining the saliency maps can be reduced. The smoothing process changes small scale patterns like grass into a smoothed homogenous region.

5.2 Evaluation measures

In this thesis work, four evaluation measures are used to evaluate the saliency results. The evaluation measures used are receiver operating characteristics (ROC) curve, overlap ratio (OR), weighted F-measure score and structure similarity (SSIM) index.

The ROC curve is a plot of true positive rate (TPR) versus the false positive rate (FPR) by varying the threshold of the saliency maps in the range [0,255]. The TPR and FPR are calculated using the Equation 5-1 and Equation 5-2.

$$TPR = \frac{|B \cap G|}{|G|} \tag{5-1}$$

$$FPR = \frac{|B \cap \bar{G}|}{|\bar{G}|} \tag{5-2}$$

Where $\bar{B}$ and $\bar{G}$ denote the inversion of the binary mask B and the ground truth G respectively. The binary mask is obtained by binarizing the saliency maps using the thresholds.

OR is defined as the overlapping ratio between the segmented object mask S' and the binary ground truth G. The OR is obtained using Equation 5-3.

$$OR = \frac{|S' \cap G|}{|S \cup G|} \tag{5-3}$$

Where, S' is obtained by binarizing S using an adaptive threshold i.e. twice the mean values of S.

Weighted F-measure [18] is an evaluation measure for overcoming the flaws of comparing a non-binary saliency map with a binary ground truth with the use of multiple thresholds. It does so by extending the four basic quantities true positive (TP), true negative (TN), false positive (FP) and false negative (FN) in terms of the weighted error. At first, an error term is defined as given by Equation 5-4.

$$E = |G - D| \tag{5-4}$$

Where, G is the binary ground truth and D denotes the non-binary map which is evaluated against the ground truth. The error term in Equation 5-4 is equally weighted. Hence, to incorporate the dependency between pixels and also to weight pixels of varying importance, the error term is weighted. It is shown with the help of Equation 5-5.

$$E^w = \min(E, EA) \,.\, B \tag{5-5}$$

Where, min gives the minimum for two arguments. The matrix A that captures the dependency between the foreground pixels and the matrix B is used for incorporating pixels of varying importance. The four quantities are written in terms of the weighted error as shown in Equation 5-6, Equation 5-7, Equation 5-8 and Equation 5-9.

$$TP^w = (1 - E^w).G \tag{5-6}$$

$$TN^w = (1 - E^w).(1 - G) \tag{5-7}$$

$$FP^w = (1 - E^w).G \tag{5-8}$$

$$FN^w = E^w.G \tag{5-9}$$

Finally, the weighted F-measure is computed using the weighted version of precision and recall. It is calculated using Equation 5-10 where a value of 1 is used for β.

$$F_\beta = \frac{(1 + \beta^2) Precision^w . Recall^w}{\beta^2 . Precision^w + Recall^w} \tag{5-10}$$

SSIM [19] is used for accessing perceptual image quality and separates the task of similarity measurement of an image with a reference image into comparison of luminance, contrast and structure. It is different with respect to mean absolute error (MAE) that treats each pixel independently and individually. SSIM tries to capture the way humans perceive images using different statistics to estimate luminance, contrast and structural differences. It is calculated using Equation 5-11.

$$SSIM(x,y) = l(x,y).c(x,y).s(x,y) \tag{5-11}$$

Where, l(x,y), c(x,y) and s(x,y) are luminance, contrast and structure comparison function respectively for the given saliency map and the corresponding ground truth. The value of SSIM lies in the range [-1, 1] where a value of 1 indicates a perfect match between the image and the reference image. The luminance, contrast and structure comparison function is given with the help of Equation 5-12, Equation 5-13 and Equation 5-14 respectively.

$$l(x,y) = \frac{2u_x u_y + C_1}{u_x^2 + u_y^2 + C_1} \tag{5-12}$$

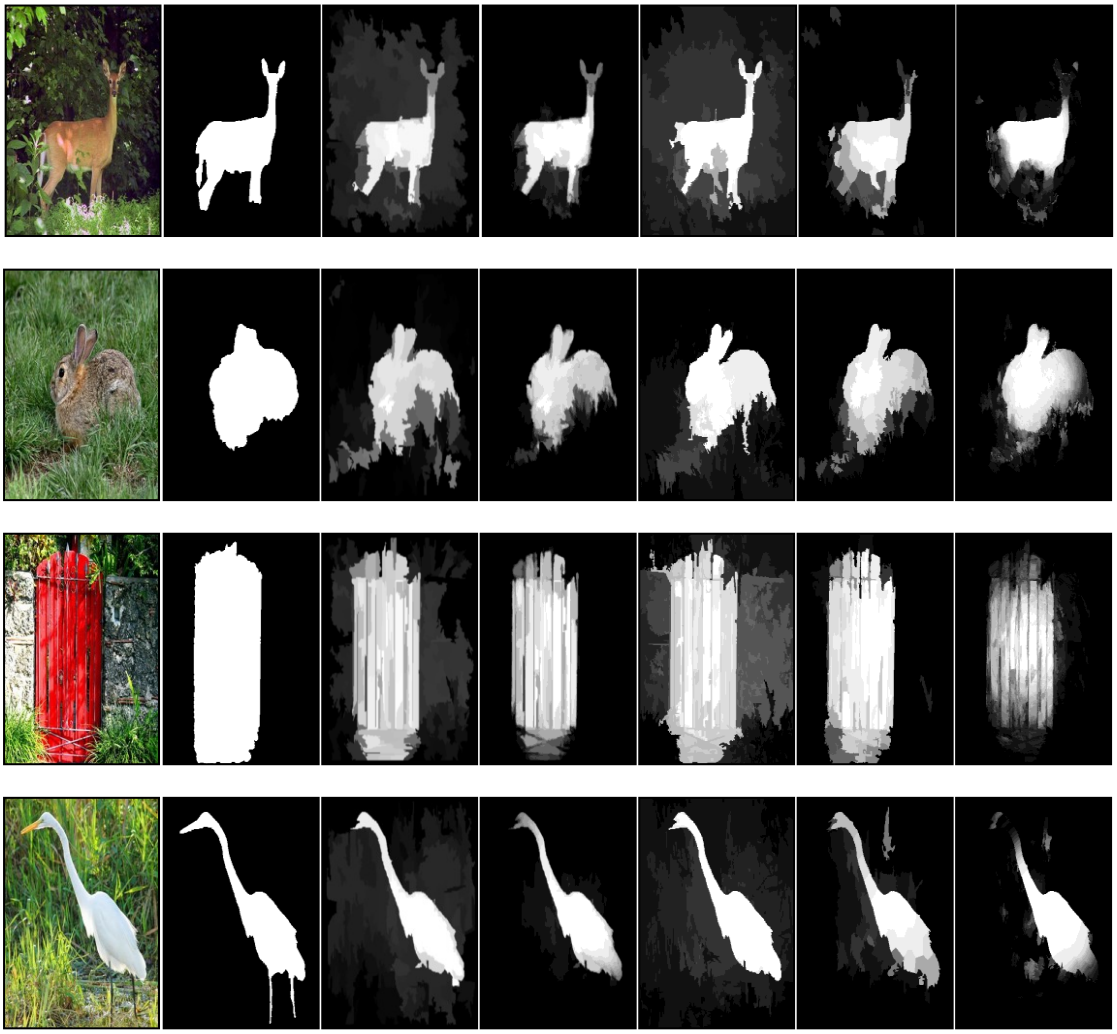
$$c(x,y) = \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{5-13}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \tag{5-14}$$

Where, u_x and u_y are the mean intensity for the image x and y respectively. σ_x and σ_y is the standard deviation which is used for obtaining an estimate of contrast similarity. The correlation term along with standard deviation is used for obtaining an estimate of structural similarity. The values C_1, C_2 and C_3 are used to prevent instability when the denominator gets close to zero. The value of C_1 is taken as $C_1 = (K_1.L)^2$ where, L is the dynamic range of pixel values and $K_1 << 1$ is a small constant. $C_2 = (K_2.L)^2$ and $K_2 << 1$ and $C_3 = C_2/2$ is used respectively.

5.3 Result for dataset images

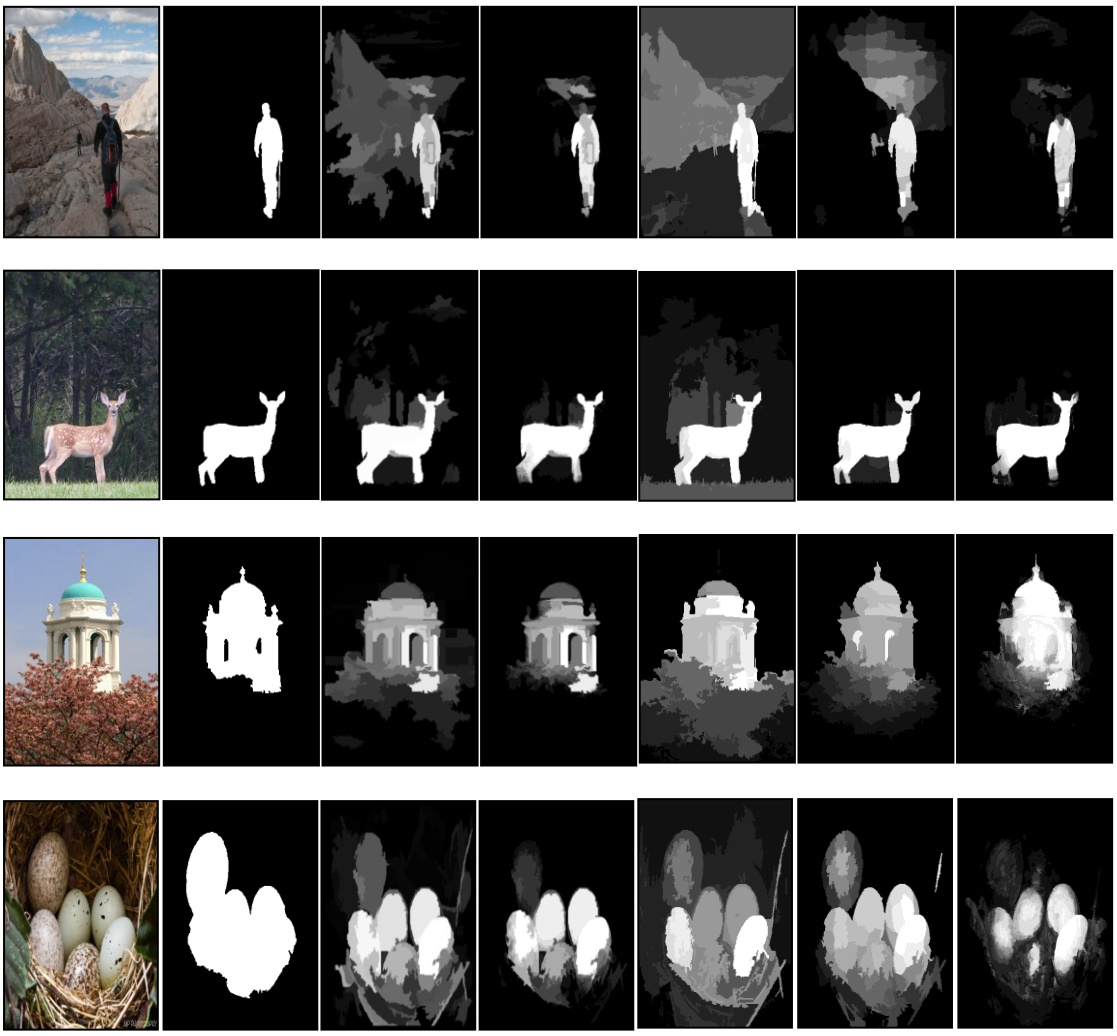
The saliency maps obtained by using the proposed method and the prior method [15] for MSRA10k dataset are shown using the Figure 5-6. The first column of the figure shows the input images taken from the dataset. The second column shows the ground truth for the corresponding input images. The third and the fourth columns show the results obtained using the prior method and the proposed method respectively. The last three columns show the results obtained using HS [14], RBD [20] and DSR [21] methods respectively.



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

Figure 5-6: Saliency results for MSRA10k images. (a) Input image (b) Ground truth (c) Without structure extraction (d) With structure extraction (e) HS (f) RBD (g) DSR

The same experiments are carried out on the images from ECSSD dataset and the results using the proposed method along with prior [15], HS [14], RBD [20] and DSR [21] methods is shown with the help of Figure 5-7.



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

Figure 5-7: Saliency results for ECSSD images. (a) Input image (b) Ground truth (c) Without structure extraction (d) With structure extraction (e) HS (f) RBD (g) DSR

5.4 Quantitative results using evaluation measures

The proposed method is compared with the prior method [15] by using the images from MSRA10k and ECSSD dataset. ROC curve is plotted for both methods for each dataset. Figure 5-8 shows the quantitative comparison of saliency maps in terms of ROC curve for MSRA10k dataset. The ROC curve for ECSSD dataset is shown with the help of Figure 5-9.

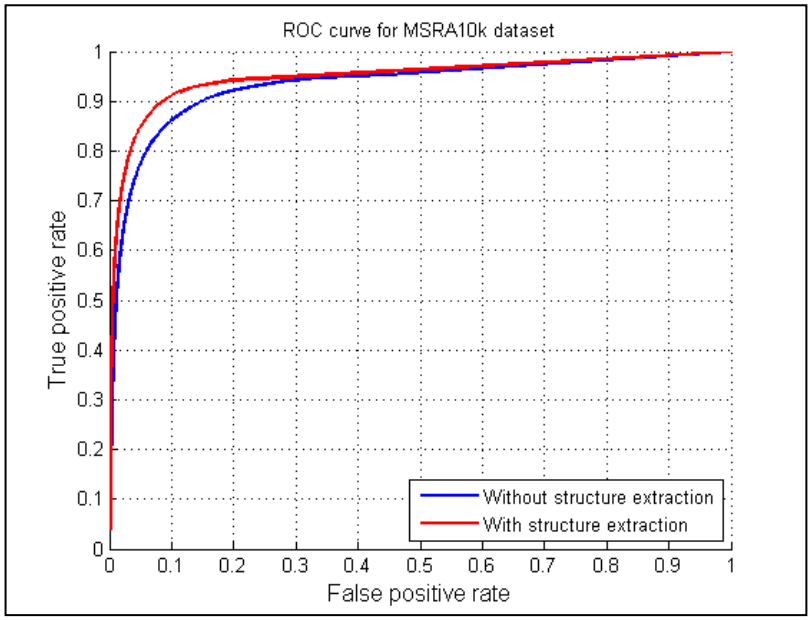


Figure 5-8: Quantitative comparison in terms of ROC curve for MSRA10k dataset

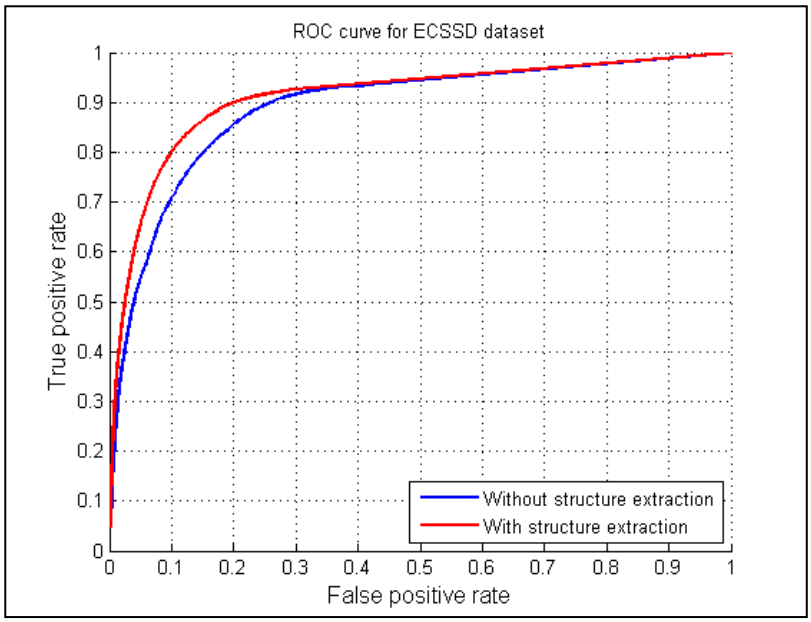


Figure 5-9: Quantitative comparison in terms of ROC curve for ECSSD dataset

The comparison of proposed method with prior method [15] along with HS [14], RBD [20] and DSR [21] methods in terms of three evaluation metrics is performed. The results obtained in terms of weighted F-measure, OR and SSIM for each dataset is shown in Table 5-2 and Table 5-3 respectively.

Table 5-2: Result on MSRA10k dataset in terms of OR, weighted-F and SSIM

MSRA10k dataset	Metric	Proposed method	Prior method [15]	HS [14]	RBD [20]	DSR [21]
	OR	0.686	0.681	0.655	**0.715**	0.653
	weighted-F	0.643	0.607	0.604	**0.685**	0.656
	SSIM	**0.751**	0.539	0.406	0.713	0.633

Table 5-3: Result on ECSSD dataset in terms of OR, weighted-F and SSIM

ECSSD dataset	Metric	Proposed method	Prior method [15]	HS [14]	RBD [20]	DSR [21]
	OR	**0.551**	0.528	0.458	0.525	0.531
	weighted-F	**0.534**	0.509	0.454	0.513	0.514
	SSIM	**0.648**	0.516	0.233	0.560	0.497

The bold entries in the table highlights the best score obtained for each evaluation metric for the given dataset. The results from Table 5-2 and Table 5-3 show that the proposed method performs competitively in the case of MSRA10k dataset whereas it obtains the best result among all other methods in the case of ECSSD dataset. The results of these tables are shown visually using a bar chart with the help of Figure 5-10, Figure 5-11 and Figure 5-12 respectively.

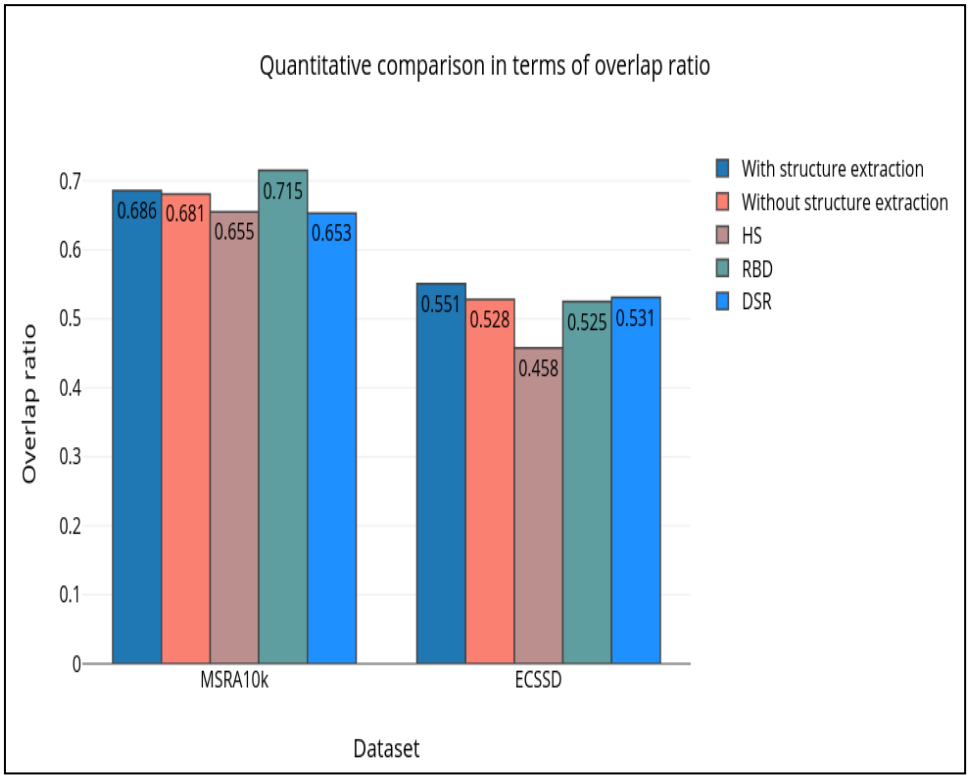


Figure 5-10: Quantitative comparison in terms of OR for both datasets

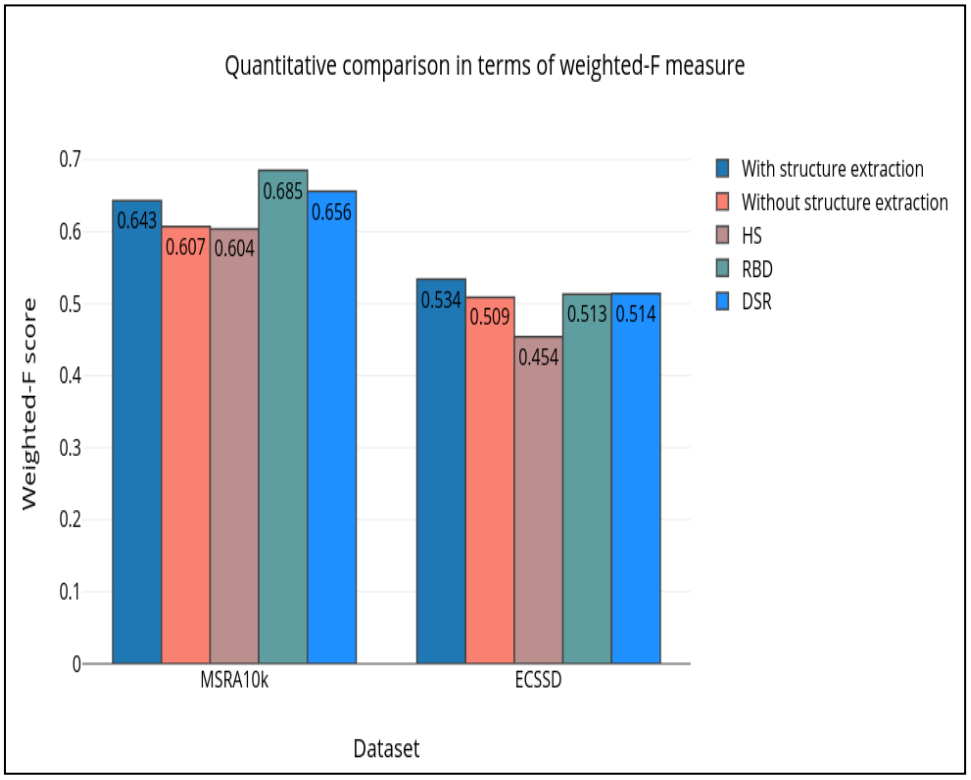


Figure 5-11: Quantitative comparison in terms of weighted-F for both datasets

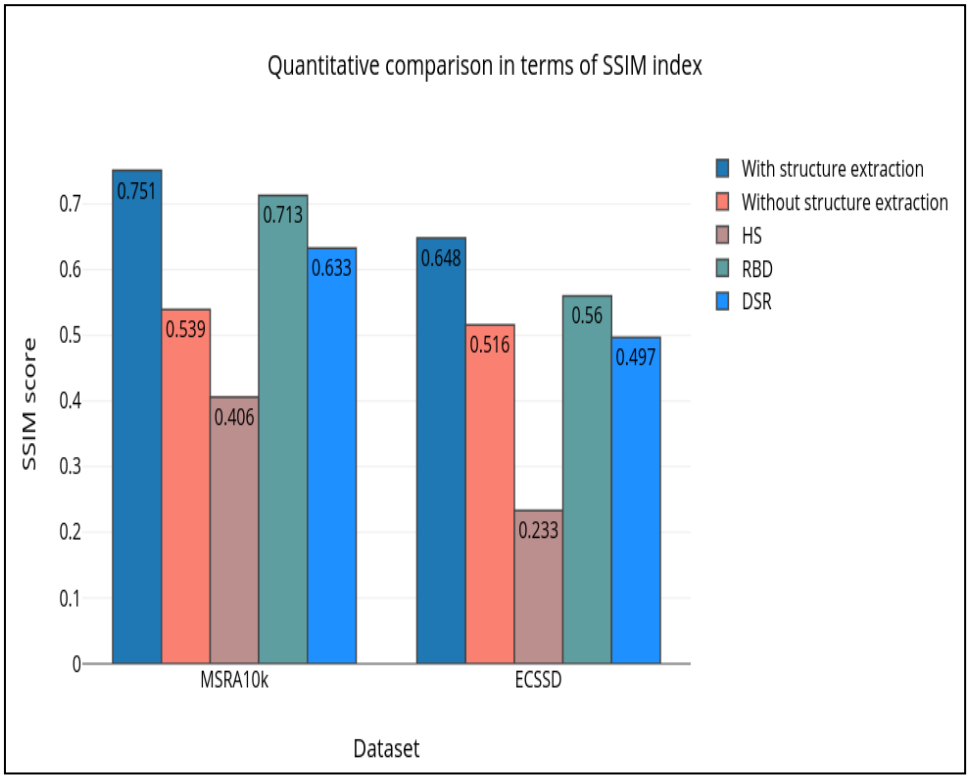


Figure 5-12: Quantitative comparison in terms of SSIM index for both datasets

Figure 5-10 shows the quantitative comparison of the saliency maps obtained by using the considered methods in terms of overlap ratio using a bar chart. Figure 5-11 and Figure 5-12 shows the quantitative comparison in terms of weighted-F measure and SSIM respectively. In the case of MSRA10k dataset, RBD [20] scores the highest considering both OR and weighted-F measure with a score of 0.715 and 0.685 respectively. The lowest score is obtained by DSR [21] considering OR and HS [14] when considering weighted-F measure with a score of 0.653 and 0.604 respectively. The proposed method obtains a competitive score of 0.686 and 0.643 for OR and weighted-F respectively. In the case of ECSSD dataset, the lowest score is obtained by HS [14] considering both OR and weighted-F measure with a score of 0.458 and 0.454 respectively. The proposed method on the other hand, obtains the best score considering all three evaluation measures for the ECSSD dataset.

5.5 Discussion

The experiments were conducted using different images from standard datasets involving various scenes. The saliency results obtained are shown using Figure 5-6 and Figure 5-7 respectively for each dataset. The result of parameter settings for algorithms at each stage is provided in Section 5-1. The results for more images (containing single and multiple objects) are presented in Appendix B.

To objectively evaluate the results, the saliency maps obtained from the proposed method along with prior method (including three other salient region detection methods) are evaluated using OR, weighted-F measure and SSIM index score. It can be observed from Table 5-2 that the proposed method obtains a competitive score in terms of weighted-F measure for MSRA10k dataset with a score of 0.643. The highest score is obtained by RBD [20] method with a score of 0.685 for the same dataset. The proposed method obtained a SSIM score of 0.751 for MSRA10k dataset and is the highest among all other considered methods for this evaluation measure. This method also obtains the best score in terms of all three evaluation measures (OR, weighted-F and SSIM) for the case of ECSSD dataset with a score of 0.551, 0.534 and 0.648 respectively as given in Table 5-3.

The results from these experiment show that the extraction of structure image from the input image helps in improving the saliency of the result. The improvement in results can be attributed to extraction of image's structure information before performing saliency detection. It helps in suppressing small patterns in the background which could be highlighted in saliency detection process. The proposed method addresses the problem of salient region detection for images containing cluttered background. This could be seen from the results shown in Table 5-3 for the case of ECSSD dataset. On the other hand, the score obtained by the proposed method is competitive for the case of MSRA10k dataset with a score of 0.686, 0.643 and 0.751 for OR, weighted-F measure and SSIM index score respectively.

CHAPTER 6

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

In this thesis work, a method to perform salient region detection by incorporating the extraction of structure image from the given input image is presented. The task of salient region detection can be broken down into two major steps. The first step consists of obtaining the structure image from the given input image such that small scale textures in the image are smoothed out while the main structures are retained. In the second step, the image is segmented into perceptually homogenous regions by using graph based image segmentation and the saliency map is computed using color contrast and complementary priors. Experimental results on two datasets show that the proposed method achieves better results when compared with the results obtained from the prior method without structure extraction. The proposed method improved the weighted-F score from 0.607 to 0.643 in the case of MSRA10k dataset when compared to the prior method. The same evaluation measure was improved from 0.509 to 0.534 in the case of ECSSD dataset. The proposed method obtained a SSIM index score of 0.751 and 0.648 as compared to 0.539 and 0.516 obtained by the prior method for each datasets respectively. It also obtained the best score considering OR, weighted-F measure and SSIM for ECSSD dataset with a score of 0.551, 0.534 and 0.648 respectively among all considered methods.

6.2 Limitation

Limitations exist for the proposed method for detecting a salient region in an image. The cues and low level features that are involved during the detection of salient region are limited and hence results in an inconsistent salient map. The proposed approach in this thesis work is mainly targeted for images containing cluttered backgrounds. However, there are still more challenges when it comes to salient region detection. Other challenges include presence of shadows, variation within an object and the case when background and foreground is not distinct with respect to color. These challenges in saliency detection still pose as a limitation for the proposed method.

6.3 Future Work

Researchers in the field of computer vision are motivated towards developing an algorithm similar to human visual system and have been shown to perform more intuitively. In the future, combining eye gaze pattern along with the proposed method can be explored for selecting multiple subjects within an image that is consistent with human visual system. In addition, inclusion of robust high-level priors like detection of human faces in an image can prove beneficial for saliency detection using the proposed method. Similarly, an investigation of algorithm guided by edge information for performing salient region detection can be considered and combined with the proposed method. This can help in situations where background and foreground have similar colors and color contrast alone cannot efficiently perform salient region detection.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 20, pp. 1254-1259, 1998.

[2] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 374-381.

[3] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang*, et al.*, "Learning to detect a salient object," *IEEE Transactions on Pattern analysis and machine intelligence,* vol. 33, pp. 353-367, 2011.

[4] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 1915-1926, 2012.

[5] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proceedings of the 14th ACM international conference on Multimedia*, 2006, pp. 815-824.

[6] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*, 2009, pp. 1597-1604.

[7] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 733-740.

[8] Y. Hu, Q. Zhou, G. Gao, Z. Yao, W. Ou, and L. J. Latecki, "Robust background exclusion for salient object detection," in *Wireless Communications & Signal Processing (WCSP), 2016 8th International Conference on*, 2016, pp. 1-5.

[9] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 853-860.

[10] C. Oprea, C. Paleologu, I. Pirnog, and M. Udrea, "Saliency Detection Making Use of Human Visual Perception Modelling," 2010.

[11] D. O'Connor, "Report on the Dublin matrix theory conference, March 1984: An introduction to sparse matrices," *Linear Algebra and its Applications,* vol. 68, pp. 271-272, 1985.

[12] E. Damiani and J. Jeong, *Multimedia Techniques for Device and Ambient Intelligence*: Springer US, 2009.

[13] X. Ren, "Superpixel: Empirical Studies and Applications," ed.

[14] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155-1162.

[15] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 37, pp. 569-582, 2015.

[16] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM Transactions on Graphics (TOG),* vol. 31, p. 139, 2012.

[17] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision,* vol. 59, pp. 167-181, 2004.

[18] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 248-255.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing,* vol. 13, pp. 600-612, 2004.

[20] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2814-2821.

[21] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2976-2983.

APPENDICES

Appendix A: Steps for obtaining the matrix A

$$Image[2,3,3] = \begin{bmatrix} (3 & 7 & 5) & (1 & 2 & 5) & (6 & 4 & 3) \\ (0 & 9 & 2) & (8 & 7 & 6) & (7 & 6 & 5) \end{bmatrix}$$

PART- I

At first the difference of pixels in x and y direction is taken. It is followed by zero padding.

$$wx = \begin{bmatrix} (-2 & -5 & 0) & (5 & 2 & -2) & (0 & 0 & 0) \\ (8 & -2 & 4) & (-1 & -1 & -1) & (0 & 0 & 0) \end{bmatrix}$$

$$wy = \begin{bmatrix} (-3 & 2 & -3) & (7 & 5 & 1) & (1 & 2 & 2) \\ (0 & 0 & 0) & (0 & 0 & 0) & (0 & 0 & 0) \end{bmatrix}$$

The square root of ($wx^2 + wy^2$) is taken. The sum of R, G and B components is taken and divided by 3. All the zeros are replaced by 0.001 and 1/(each element) is computed.

$$wto = \begin{bmatrix} 0.2502 & 0.1849 & 0.6000 \\ 0.2143 & 1.0000 & 1000 \end{bmatrix}$$

PART- II

The original image is passed through the gaussian smoothing filter that results into the following:

$$\begin{bmatrix} (0.3843 & 0.5692 & 0.4200) & (0.4229 & 0.5879 & 0.4440) & (0.4190 & 0.5481 & 0.4230) \\ (0.3886 & 0.5773 & 0.4195) & (0.4279 & 0.5965 & 0.4440) & (0.4241 & 0.5563 & 0.4235) \end{bmatrix}$$

The difference of pixels in x and y direction is taken. It is followed by zero padding.

$$gx = \begin{bmatrix} (0.0387 & 0.0187 & 0.0240) & (-0.0039 & -0.0398 & -0.0210) & (0 & 0 & 0) \\ (0.0393 & 0.0192 & 0.0245) & (-0.0038 & -0.0402 & -0.0205) & (0 & 0 & 0) \end{bmatrix}$$

$$gy = \begin{bmatrix} (0.0043 & 0.0082 & -0.00044) & (0.0050 & 0.0086 & 0.000052) & (0.0051 & 0.0082 & 0.000532) \\ (0 & 0 & 0) & (0 & 0 & 0) & (0 & 0 & 0) \end{bmatrix}$$

The absolute value of each element of gx and gy is taken. Then the sum for R, G and B components is taken and divided by 3. All zero elements are replaced by 0.001 and 1/(each element) is computed. It results into matrices wtbx and wtby.

$$wtbx = \begin{bmatrix} 36.9 & 46.4 & 1000 \\ 36.1 & 46.5 & 1000 \end{bmatrix}$$

$$wtby = \begin{bmatrix} 232 & 219.7 & 217.4 \\ 1000 & 1000 & 1000 \end{bmatrix}$$

The matrix wto is multiplied with wtbx and wtby respectively to obtain the required weight matrix given by weight-x and weight-y.

$$weight - x\ [2,3] = \begin{bmatrix} 9.2200 & 8.5710 & 600 \\ 7.7362 & 46.5 & 1000000 \end{bmatrix}$$

$$weight - y\ [2,3] = \begin{bmatrix} 58.05 & 40.63 & 130.44 \\ 214.3 & 1000 & 1000000 \end{bmatrix}$$

The last column and last row of these weight matrices is replaced with zeros respectively and then multiplied with parameter lambda. The result is put column-wise to obtain matrix B.

$$B[6,2] = \begin{bmatrix} -0.0461 & -0.2903 \\ -0.0387 & 0 \\ -0.0429 & -0.2032 \\ -0.2323 & 0 \\ 0 & -0.6522 \\ 0 & 0 \end{bmatrix}$$

Using the matrix B and few other operations we obtain the sparse matrix given by A.

$$A[6,6] = \begin{bmatrix} 1.3364 & -0.2903 & -0.0461 & 0 & 0 & 0 \\ -0.2903 & 1.3290 & 0 & -0.0387 & 0 & 0 \\ -0.0461 & 0 & 1.29210 & -0.2032 & -0.0429 & 0 \\ 0 & -0.0387 & -0.2032 & 1.4741 & 0 & -0.2323 \\ 0 & 0 & -0.0429 & 0 & 1.6951 & -0.6522 \\ 0 & 0 & 0 & -0.2323 & -0.6522 & 1.8845 \end{bmatrix}$$

The Eigen values of the matrix ‘A’ are: 1.0000, 1.0564, 1.2469, 1.5721, 1.6523, and 2.4835.

Appendix B: Comparison of results for more images

The figure below shows the saliency maps obtained by using the proposed method, prior method [15], HS [14], RBD [20] and DSR [21] methods respectively. The first column of the figure shows the input images taken from the dataset. The second column shows the ground truth for the corresponding input images. The third, fourth, fifth, sixth and seventh columns show the results obtained by using prior method [15], proposed method, HS [14], RBD [20] and DSR [21] methods respectively.