Detection of Salient Objects in Images Using Frequency Domain and Deep Convolutional Features

Masoumeh REZAEI ABKENAR

A Thesis

In the Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Electrical and Computer Engineering) at

Concordia University

Montreal, Quebec, Canada

June 2019

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Masoumeh REZAEI ABKENAR**
Entitled: **Detection of Salient Objects in Images Using Frequency Domain and Deep Convolutional Features**

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.
Signed by the Final Examining Committee:

_____ Chair
Dr. S.S. Li

_____ External Examiner
Dr. P. Agathoklis

_____ External To Program
Dr. T. Fancott

_____ Examiner
Dr. Y.R. Shayan

_____ Examiner
Dr. M.N.S. Swamy

_____ Examiner
Dr. W.-P. Zhu

_____ Thesis Supervisor
Dr. M.O. Ahmad

Approved by _____
Dr. R.R. Selmic, Graduate Program Director

August 2, 2019

_____
Dr. A. Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Detection of Salient Objects in Images Using Frequency Domain and Deep Convolutional Features

Masoumeh REZAEI ABKENAR, Ph.D.

Concordia University, 2019

In image processing and computer vision tasks such as object of interest image segmentation, adaptive image compression, object-based image retrieval, seam carving, and medical imaging, the cost of information storage and computational complexity is generally a great concern. Therefore, for these and other applications, identifying and focusing only on the parts of the image that are visually most informative is much desirable. These most informative parts or regions that also have more contrast with the rest of the image are called the salient regions of the image, and the process of identifying them is referred to as salient object detection. The main challenges in devising a salient object detection scheme are in extracting the image features that correctly differentiate the salient objects from the non-salient ones, and then utilizing them to detect the salient objects accurately.

Several salient object detection methods have been developed in the literature using spatial domain image features. However, these methods generally cannot detect the salient objects uniformly or with clear boundaries between the salient and non-salient regions. This is due to the fact that in these methods, unnecessary frequency content of the image get retained or the useful ones from the original image get suppressed. Frequency domain features can address these limitations by providing a better representation of the image. Some salient object detection schemes have been developed based on the features extracted using the Fourier or Fourier like transforms. While these methods are more successful in

detecting the entire salient object in images with small salient regions, in images with large salient regions these methods have a tendency to highlight the boundaries of the salient region rather than doing so for the entire salient region. This is due to the fact that in the Fourier transform of an image, the global contrast is more dominant than the local ones. Moreover, it is known that the Fourier transform cannot provide simultaneous spatial and frequency localization.

It is known that multi-resolution feature extraction techniques can provide more accurate features for different image processing tasks, since features that might not get extracted at one resolution may be detected at another resolution. However, not much work has been done to employ multi-resolution feature extraction techniques for salient object detection. In view of this, the objective of this thesis is to develop schemes for image salient object detection using multi-resolution feature extraction techniques both in the frequency domain and the spatial domain.

The first part of this thesis is concerned with developing salient object detection methods using multi-resolution frequency domain features. The wavelet transform has the ability of performing multi-resolution simultaneous spatial and frequency localized analysis, which makes it a better feature extraction tool compared to the Fourier or other Fourier like transforms. In this part of the thesis, first a salient object detection scheme is developed by extracting features from the high-pass coefficients of the wavelet decompositions of the three color channels of images, and devising a scheme for the weighted linear combination of the color channel features. Despite the advantages of the wavelet transform in image feature extraction, it is not very effective in capturing line discontinuities, which correspond to directional information in the image. In order to circumvent the lack of directional flexibility of the wavelet-based features, in this part of the thesis, another salient object detection scheme is also presented by extracting local and global features from the non-subsampled contourlet coefficients of the image color channels. The local features are extracted from

the local variations of the low-pass coefficients, whereas the global features are obtained based on the distribution of the subband coefficients afforded by the directional flexibility provided by the non-subsampled contourlet transform.

In the past few years, there has been a surge of interest in employing deep convolutional neural networks to extract image features for different applications. These networks provide a platform for automatically extracting low-level appearance features and high-level semantic features at different resolutions from the raw images. The second part of this thesis is, therefore, concerned with the investigation of salient object detection using multi-resolution deep convolutional features. The existing deep salient object detection schemes are based on the standard convolution. However, performing the standard convolution is computationally expensive specially when the number of channels increases through the layers of a deep network. In this part of the thesis, using a lightweight depthwise separable convolution, a deep salient object detection network that exploits the fusion of multi-level and multi-resolution image features through judicious skip connections between the layers is developed. The proposed deep salient object detection network is aimed at providing good performance with a much reduced complexity compared to the existing deep salient object detection methods.

Extensive experiments are conducted in order to evaluate the performance of the proposed salient object detection methods by applying them to the natural images from several datasets. It is shown that the performance of the proposed methods are superior to that of the existing methods of salient object detection.

Acknowledgments

I would like to express my sincere gratitude to my supervisor Professor M. Omair Ahmad for his guidance, patience and continuous support during my PhD studies. The time spent in our meetings gave me a deep insight through my research and has been invaluable to me. I feel privileged to have the opportunity to work under your supervision. I would admire your professional and positive attitude forever.

I would like to express my appreciation to my dear siblings and friends without whose support this PhD thesis could have not come into existence.

My special thanks go to my dear mom and dad for their everlasting love. In all my difficult moments, your faith in me makes me the strongest person in the world. I am grateful for all the sacrifices you made for me.

Finally yet importantly, I would like to thank my beloved husband, Mostafa, for his unconditional love, care and encouragement. You have been always there for me, through all my ups and downs. I am lucky to have you in my life.

Contents

List of Figures

List of Tables

List of Symbols

$\boldsymbol{S}$	Gray-level saliency map
$\boldsymbol{S}_b$	Binary saliency map
$\boldsymbol{GT}$	Ground truth
(x, y)	Pixel's spatial location
L	Number of rows in the image
W	Number of columns in the image
T_{adp}	Adaptive threshold
P	Precision
R	Recall
F_β	F-measure
β^2	Relative importance of the precision and recall
$WT(.)$	Discrete wavelet transform
$\boldsymbol{I}^c$	Color channel of the input image
$\boldsymbol{A}$	Approximation subband coefficients in wavelet and non-subsampled contourlet transforms
$\boldsymbol{H}_n^c$	Horizontal subband coefficients in wavelet transform corresponding to channel *c* at level *n*
$\boldsymbol{V}_n^c$	Vertical subband coefficients in Wavelet transform corresponding to channel *c* at level *n*

$\boldsymbol{D}_n^c$	Diagonal subband coefficients in wavelet transform corresponding to channel *c* at level *n*
$IWT(.)$	Inverse wavelet transform
$\boldsymbol{TMAP}^c$	Texture map of the channel *c*
$\boldsymbol{FMAP}^c$	Channel Feature map of channel *c*
$\boldsymbol{FMAP}$	Image Feature map
ω^c	Weight for linear combination, assigned to the map corresponding to channel *c*
$\eta(.)$	Entropy
ε^c	Entropy value of low-pass filtered map
$*$	2D convolution operator
G	Low-pass Gaussian filter
α^c	Strength of the feature map around image center
Z	Gaussian mask
$\|.\|$	Euclidean distance
$\boldsymbol{MFMAP}$	Refined image feature map
$Nr(.)$	Normalization function
BF	Bilateral filtering operation
W_p	Normalization factor in bilateral filter
q	Spatial Location *(x,y)*
$q^{'}$	Spatial Location $(x^{'},y^{'})$
Ω	Set of possible positions in MFMAP
σ_d	Domain standard deviation
σ_r	Range standard deviation
$D_{n,d}$	Bandpass directional subbands in non-subsampled contourlet transform corresponding to level *n* and direction *d*

$NSCT(.)$	Non-subsampled contourlet transform
$INSCT(.)$	Inverse non-subsampled contourlet transform
σ	Standard deviation
$Var(.)$	variance value in a block
Lmap	Local feature map
$\boldsymbol{S}_{Local}$	Local saliency map
Gmap	Global feature map
f_g	Global feature vector
l	Global feature vector length
$p(.)$	Likelihood function
μ	Vector containing the mean values
$\sum$	Covariance matrix
$(.)^T$	Transpose operator
$\|.\|$	Determinant of a matrix
$\boldsymbol{S}_{Global}$	Global saliency map
S_M	Fused map
$\circ$	Hadamard matrix product
k	Number of superpixels
ωf_i	Foreground weight
ωb_i	Background weight
ωs_{ij}	Smoothness weight
$Eblock$	Encoder block
$Dblock$	Decoder block
s	Stride value
L_{BCE}	Binary cross entropy loss function

List of Abbreviations

2D	Two-Dimensional
AMULET	Aggregating Multi-level Convolutional Features method
CNN	Convolutional Neural Network
DCL	Deep Contrast Learning method
DHS	Deep Hierarchical Saliency method
DS	Deep Saliency method
ELD	Encoded Low Level Distance Map method
FCN	Fully Convolutional Network
FT	Fourier Transform
FTU	Frequency-Tuned method
GPU	Graphics Processing Unit
HFT	Hyper-complex Fourier-based method
HVS	Human Visual System
MAC	Multiplication-Accumulation
MAE	Mean Absolute Error
MBD+	Minimum Barrier Distance method
MDF	Multi-scale Deep Features method
MR	Manifold Ranking method
MST	Minimum Spanning Tree method

NSCT	Non-subsampled Contourlet Transform
NSDFB	Non-subsampled Directional Filter Bank
NSP	Non-subsampled Pyramid
RC	Region-based Contrast method
ReLU	Rectified Linear Unit
RFW	Recurrent Fully Convolutional Network
SF	Saliency Filter method
SLIC	Simple Linear Iterative Clustering
SO	Scaling Optimization method
SR	Spectral Residual method
STFT	Short-Time-Fourier Transform
SW	Superpixel-based Wavelet method
Tran. CONV	Transposed Convolution
WAV	Wavelet-based method
WQ	Weighted Quaternion-based method
WT	Wavelet Transform

Chapter 1

Introduction

1.1 General

One of the main characteristics of the human visual system (HVS) is its ability to efficiently detect the visually most conspicuous region, referred to as the salient object, of an image. The salient object is known to be conspicuous and has more contrast with respect to the local and global surrounding regions. It is more distinctive in terms of color, edges, boundaries, etc. Salient object detection methods aim at identifying the salient region using computational algorithms. By detecting the salient parts, any processing task such as object of interest image segmentation [1, 2], adaptive image or video compression [3–5], object recognition [6], object-based image retrieval [7], image retargeting [8,9], video summarization [10, 11], visual tracking [12], image fusion [13], action recognition [14], object class discovery [15], and medical imaging [16, 17], can focus on the major content of an image or a video. In view of this, recently, researches have shown a great deal of interest in developing saliency detection techniques.

Salient object detection methods extract various kinds of features from the image and analyze them in order to assign a saliency value to all the pixels. Based on these values, it is decided whether or not a pixel belongs to the salient region. The output of a saliency detection method is a saliency map, in which each pixel has a saliency value. This value represents the probability of the pixel belonging to the salient region. Thus, a saliency

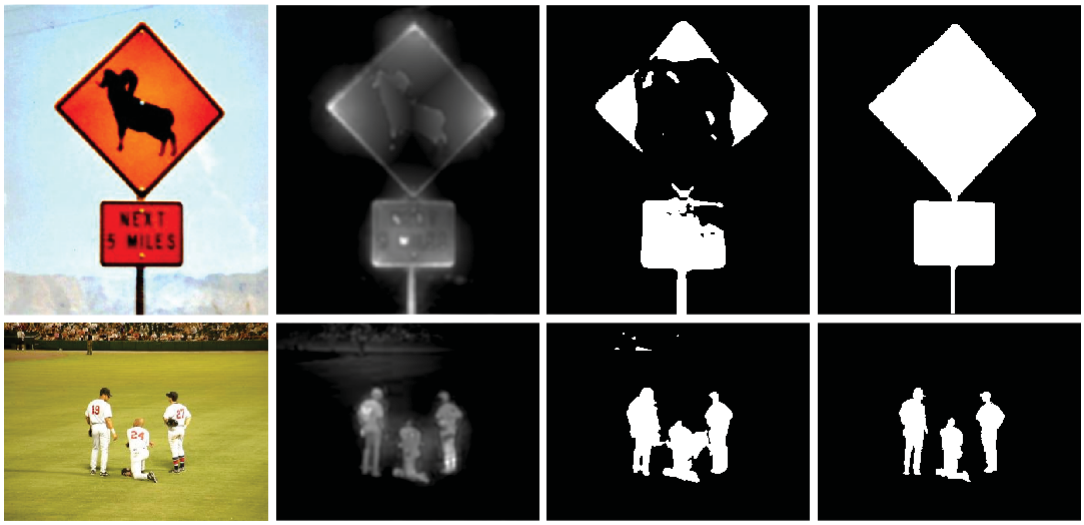


Figure 1.1: (a) Sample natural images. (b) Saliency maps obtained by using the method proposed in [19]. (c) The corresponding binary saliency maps. (d) Ground truth for the saliency maps.

map is inherently a gray-level image. Depending on the probability value of a pixel in the salinecy map, a decision is made whether the pixel belongs to the salient region or not, that is, whether the pixel is assigned a value of 1 and, therefore considered to belong to salient region or a value of 0, in which case it does not belong to the salient region. Thus, the gray-level saliency map is converted into a binary saliency map and compared to the corresponding ground truth. It should be noted that salient object detection methods segment only the salient region from the non-salient region, that is classifying the image pixels/regions into two classes of salient and non-salient, whereas image segmentation methods partition the image into several regions of coherent properties [18]. Figure 1.1 shows examples of natural images, the gray-level and binary saliency maps obtained by using the method proposed in [19], and the ground truth (which is always binary) for the saliency maps.

1.2 A Brief Literature Review of Saliency Detection

In this section, a brief review of the different categories of the research in saliency detection methods and some of the main advances in these categories are presented.

Early studies in salieny detection aimed at predicting human eye-attended image pixels [20, 21]. These methods are called eye-fixation prediction and obtain saliency maps in which the focus is on detection of sparse eye fixation locations rather than detection of the entire salient objects. A comprehensive survey on the eye-fixation prediction schemes can be found in [22]. On the other hand, salient object detection methods, have been developed with an objective to detect the entire salient objects [7, 19, 23–38]. The saliency map obtained by this category of methods is dense and includes connected areas demonstrating the most visually informative object of the image. Due to various applications of salient object detection in image processing and computer vision, our main focus in this thesis is on developing new salient object detection methods that can provide more accurate saliency maps.

In the salient object detection methods, some low-level image features, such as color, edges or texture, or high-level ones, such as semantic features related to the shape or structure of the object, are extracted and a saliency value is assigned to each pixel based on these features. Choosing a feature that is able to provide useful information towards the local and global characteristics of the individual pixels of the image is a challenging task.

To detect the salient objects, several methods have been developed in the spatial domain [7, 23, 26–29, 31–35]. Spatial domain methods first try to extract different features from the image using its pixel values. Then, the feature maps, each using one type of features, are computed by employing a center-surround operation [28], contrast computation [7, 29], or graph-based computations [31–35, 38]. Finally, the various feature maps are normalized and linearly or non-linearly combined to obtain a saliency map. Some of the spatial domain salient object detection methods cannot detect the salient objects from the

background with clear boundaries [7]. This may be due to an inappropriate reduction of the high frequency content of the original image. Some other spatial domain methods obtain saliency maps in which the salient region boundaries are clear, but the entire salient region is not uniformly highlighted, or the textured regions are highlighted regardless of their contrast with their surrounding regions [28]. Other schemes such as the one proposed in [26] can highlight small salient regions but fails in detecting the large ones. From a frequency domain perspective, these limitations are the consequence of retaining an inappropriate range of frequency content from the original image [27].

In order to address the above mentioned limitations of the spatial domain methods, and because of the characteristics of the frequency domain representations of images, several frequency domain salient object detection methods have also been developed [19,24,25,30, 36, 37]. In frequency domain methods, the salient object is detected by following the steps of applying the frequency transform to the input image, modulating the frequency spectrum in order to suppress the background and enhance the salient regions, and finally generating the gray level saliency map through the inverse transform. Early studies of salient object detection in the frequency domain have been mainly based on Fourier transform (FT) [19, 24, 25]. While the FT-based frequency domain methods can successfully detect small salient objects, in images with large salient objects they can only detect the boundary between the salient and non-salient objects rather than detecting the entire salient object. This is due to the fact that in the FT of an image the global contrast features have dominance over the local contrast features. Moreover, it is known that the FT cannot provide simultaneous spatial and frequency localization, and it is not useful for analyzing non-stationary signals such as most of the natural images. The short-time-Fourier transform (STFT) can be utilized to perform local frequency analysis. It segments the signal into narrow spatial intervals (i.e., narrow enough to be considered stationary) and takes the FT of each segment. However, the fixed size of the intervals is the main drawback of the STFT. For low

frequencies, a proper frequency resolution is needed, while for high frequencies, the spatial resolution is more important [39, 40].

It is known that extracting image feature in a multi-resolution manner can provide more comprehensive features for different image processing tasks, since each resolution can extract some features that might not get extracted at other resolutions. A few salient object detection methods have been proposed by employing frequency domain transforms such as the wavelet transform (WT) and the non-subsampled contourlet transform (NSCT) that have the capability of doing a multi-resolution feature extraction [30, 36].

In the past few years, employing deep convolutional neural networks (CNNs) to extract image features has attracted a great deal of interest in different image processing and computer vision applications. These deep networks provide a multi-resolution platform for automatic extraction of low-level and high-level image features pertinent to a large class of images. The reason that such a platform can be utilized to extract both the low-level and high-level features is that they can be trained using ground truth images that are gathered using HVS. In order to effectively extract high-level features using neural networks, we need networks that are deep. Recently, some salient object detection methods have been proposed utilizing CNNs [41–51]. These methods provide substantially improved results over that provided by the conventional schemes. The existing deep salient object detection schemes use the standard convolution, which is a convolution over all the channels in one step. However, performing the standard convolution is computationally expensive specially when the number of channels is increased through the layers of a deep network.

A comprehensive review of the works in salient object detection can be found in the survey papers [18, 52].

1.3 Motivation

Extracting image features that can distinguish the salient regions from the non-salient ones, and then making use of the extracted features to detect the salient regions accurately are the main steps in developing salient object detection schemes. Although there are a number of salient object detection methods, it is still challenging to develop more accurate methods that can obtain saliency maps with uniformly highlighted salient regions and sharp boundaries between the salient and non-salient regions. Despite to capabilities of multi-resolution image feature extraction techniques, not much effort has been made in developing salient object detection schemes by making use of these techniques both in the frequency domain and the spatial domain. Frequency transforms such as the WT and the NSCT can extract multi-resolution image features that could be beneficial in detecting salient objects in images. Also, it has been shown that the deep learning approaches can be used as a multi-resolution spatial domain feature extraction technique for salient object detection. These schemes can automatically extract multi-resolution, low-level, and high-level image features, and obtain significantly improved results compared to the non-deep learning methods. However, the existing deep schemes suffer from the large computational cost. Therefore, developing relatively low complexity deep salient object detection methods is a challenging task.

1.4 Objectives

The objective of this thesis is to develop new schemes for salient object detection in images by extracting multi-resolution multi-level image features using frequency and spatial domain tools, and devising algorithms by making use of the extracted features in order to identify the salient object.

The first part of this thesis is concerned with developing salient object detection schemes

using multi-resolution frequency domain features. First, a salient object detection method is proposed by making use of WT-based image features extracted from the color channels. Also, since, in a color image different features can be extracted from each color channel, it is desirable to investigate an efficient approach to combine the extracted channel features. The core ideas of the proposed salient object detection scheme are recognizing the WT based color channel textural details as suitable features to distinguish salient regions from non-salient ones, and devising a scheme for the weighted linear combination of the color channel features in order to detect and extract the salient objects. A new method based on the entropy of the individual color channels within the image is proposed for determining the weights of the linear combination. Although the WT-based features are effective in representing image features at different resolutions, they suffer from lack of directional flexibility. In order to address this issue, in this part of the thesis another salient object detection scheme is developed by using NSCT-based image features. It is known that this transform is capable of providing a multiscale, multi-directional and translation invariant decomposition of images. In addition, unlike the WT that can capture only the point discontinuities, the NSCT is able to capture the line discontinuities too. The proposed salient object detection method is realized by extracting local features from the local variation of the low-pass coefficients, and global features from the distribution of the directional subband coefficients.

The second part of this thesis deals with investigating salient object detection using the multi-resolution features extracted by a deep CNN. The proposed salient object detection network has an encoder-decoder architecture, in which both the encoder and decoder parts are designed based on depthwise separable convolution operations rather than the standard convolution. In the design of the proposed network a skip connection is established between an encoder layer an a judiciously chosen decoder layer that essentially has been aimed to improve the performance of the network by combining multi-level image features

and controlling the gradient vanishing problem. In view of performing the lightweight depthwise separable convolution in the proposed network, its computational complexity is reduced to an extent that makes it possible to raise the complexity slightly up for allowing the utilization of the skip connections.

Extensive experiments are conducted in order to evaluate the performance of the proposed salient object detection schemes and to compare them with other existing works by applying them to the natural images from several datasets.

1.5 Organization of the Thesis

The thesis is organized as follows: In Chapter 2, a brief review of different tools that are used to extract image features in this thesis is presented. It includes an introduction to the WT, the NCST, and the CNNs. Also, different metrics that are utilized to evaluate the performance of the proposed salient object detection methods are described in this chapter.

In chapter 3, a salient object detection method is developed by utilizing multi-resolution image features extracted from the WT decomposition of the three color channels. A scheme is proposed to linearly combine the channel feature maps in which the weights for the linear combination are determined by making use of the entropy of the channel feature maps and a Gaussian kernel, utilizing the fact that the salient objects are generally clustered and scene-centric. Several experiments are conducted on sets of natural images to evaluate the performance of the proposed method.

In chapter 4, using the multi-resolution image features obtained by applying the NSCT, a salient object detection scheme is devised by extracting local and global features from the NSCT coefficients of the three color channels at different scales. The local features are extracted from the local variations of the low-pass coefficients whereas the global features are obtained based on the distribution of the directional subband coefficients. Experiments are carried out to evaluate the effectiveness of the proposed salient object detection method.

In chapter 5, a deep salient object detection network is designed using the lightweight depthwise separable convolution. The proposed network is developed by extracting muli-level and multi-scale image features in the layers of the network and exploiting the fusion of extracted image features through judicious skip connections between the layers. The proposed deep salient object detection network is aimed at providing a good performance while reducing the computational cost. A number of experiments are conducted to evaluate the proposed network and to compare it with the existing deep salient object detection schemes in terms of performance and computational complexity.

Finally, in Chapter 6, some concluding remarks and scope for further research are presented.

Chapter 2

Background Material

In this chapter, a brief review of the background material required for the development of the proposed salient object detection schemes in subsequent chapters is presented.

2.1 Two Dimensional Discrete Wavelet Transform

The WT has been developed as a powerful tool for different signal processing applications [53]. It was first presented as the foundation of the multi-resolution theory which is concerned with the signal analysis at more than one resolution [54, 55]. The main idea behind the multi-resolution theory is that the features that can not be detected at one resolution, may be easy to detect at another. Therefore, the multi-resolution analysis is capable of detecting the local and global features of the image [56]. Moreover, unlike the FT which is based on sinusoid basis functions with an infinite length, the WT employs basis function of varying frequency and limited duration. These functions provide space-frequency localization [57].

In image processing applications, we deal with two-dimensional (2D) discrete wavelet transform. It consists of scaling functions, used to generate approximations of an image each differing by a scale of $1/2$ in resolution, and wavelet functions, used to encode the differences between successive approximations. By applying the WT, the image is represented as a linear combination of the scaling functions and the wavelets [53]. The WT

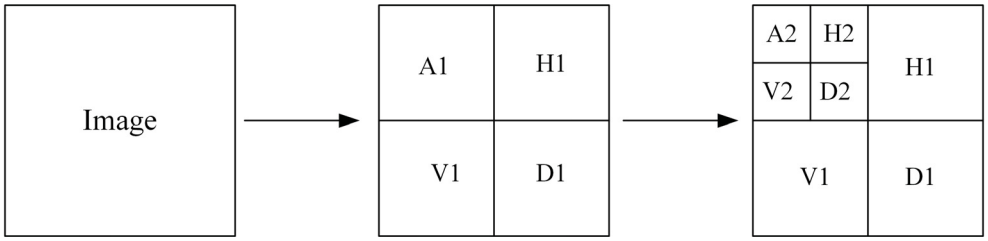


Figure 2.1: Illustration of a 2-level WT decomposition: A, H, V, and D represent approximation, horizontal, vertical and diagonal subbands, respectively.

decomposition of an image at each scale comprises the approximation or the low-pass subband and the oriented details (horizontal, vertical, and diagonal) or the high-pass subbands. Figure 2.1 shows an schematic of a 2-level WT decomposition process.

2.2 Non-subsampled Contourlet Transform

As mentioned in Chapter 1, the NSCT [58] has been proposed in order to obtain directional flexibility in the multi-resolution representation of images. This transform is composed of two filtering stages including non-subsampled pyramid (NSP) and non-subsampled directional filter bank (NSDFB). Multi-resolution property is realized by using two-channel non-subsampled filter bank resulting in low- and high-frequency images at each NSP decomposition level. Subsequently, to capture the singularities in the image, NSP iteratively decomposes the low-frequency component. As a result, NSP provides $J+1$ sub-images comprising 1 low- and J high-frequency image components having the same size as that of the original image, where J denotes the number of decomposition levels. Finally, multi-directional decomposition is achieved by applying NSDFB to high-frequency images at each scale. NSDFB is constructed by combining the directional fan filter banks and then used to produce flexible directional sub-images with the same size as that of the original image. It is not only able to capture the smooth contours effectively in a flexible number of directions, but is also invariant under translations. Figure 2.2 shows a schematic

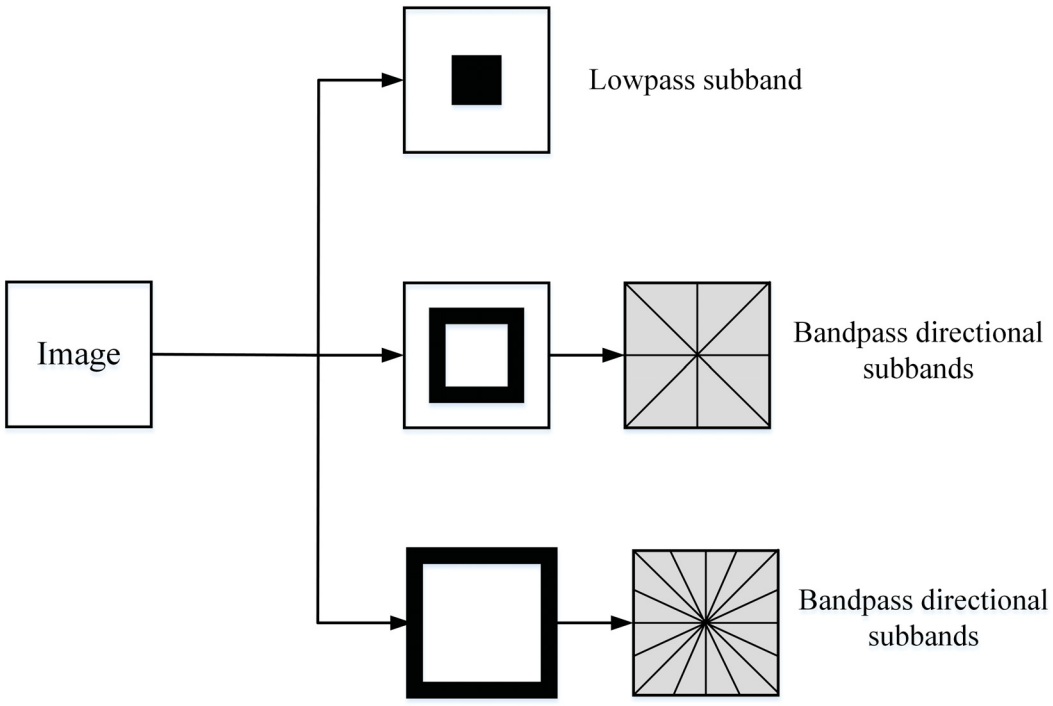


Figure 2.2: Illustration of the NSCT decomposition.

of the NSCT decomposition process.

2.3 Deep Convolutional Neural Networks

Neural networks provide a platform for automatically extracting both low-level and high-level image features from the raw images. The reason that such a platform can be utilized to extract both the low-level and high-level features is that they can be trained using ground truth images that are gathered using HVS. In order to effectively extract high-level features using neural networks, we need networks that are deep. Deep neural networks remained practically infeasible in view of the requirement of large processing power for their implementation. However, the advent of modern graphics processing units (GPUs) paved the way for implementing deep neural networks. Consequently, training of deep neural networks using very large image datasets to learn the ground truth images have

become possible. With the high computational capability provided by the modern GPUs, Krizhevsky et al. in 2012 were able to implement a deep CNN with 8 layers and millions of parameters for the purpose of image classification [59]. They trained the network using the ImageNet dataset [60] containing 1.2 million images. Subsequently, larger and deeper CNNs have been proposed [61–63] and widely used for different tasks, such as semantic segmentation [64], edge detection [65], object detection [66, 67], and pedestrian detection [68].

Use of convolution operations in deep neural networks is motivated by three main factors [69]. First, computing the output using a CNN requires fewer parameters and operations compared to that required by the traditional neural networks. This is accomplished by using weight matrices, called kernels, that are only a fraction of the input image size. Second, in a CNN each element of the kernel is used at every position of the input, while in a traditional neural network each element of the kernel is used exactly once. This characteristic of the CNNs referred to as parameter sharing reduces the storage requirements of the model. Third, the convolution operation provides the flexibility of working with images of different sizes.

A CNN consists of an input layer, an output layer, and a number of convolutional layers. A typical convolutional layer of a CNN performs three main tasks. The first task of a convolutional layer is a convolution of local regions (a spatial extent) of the image, called receptive fields, with a single kernel. The use of a single kernel over all the receptive fields of the image results in a linear activation map with only one kind of feature. Therefore, in order to obtain activation maps of different kinds of features, different kernels are applied resulting in a set of activation maps or a tensor of features. Adding non-linearity to the network improves its ability to learn the ground truth images more accurately. Therefore, the second task of a convolutional layer is application of a non-linear activation function, such as the rectified linear unit (ReLU), to each of the linear activation maps. The third task is

modification in the dimension of the non-linear activation maps in order to extract features at different levels and modify the number of parameters. Depending on the application of the network, the output of the layer may have to be down-sampled using a function such as max-pooling, or up-sampled by applying a function such as deconvolution [69]. Number of the convolutional layers in a CNN depends on the application for which the network is designed. By employing more convolutional layers, multi-level image features are extracted.

2.4 Metrics Used to Evaluate the Performance of Salient Object Detection Methods

In this section, we review different metrics used to evaluate the performance of a salient object detection method.

As mentioned earlier, the output of a salient object detection method is a gray-level saliency map, $\boldsymbol{S}$. Salient object detection can be considered as binary classification of salient and non-salient regions in the saliency map. Since in this problem the ground truth, $\boldsymbol{GT}$, is a binary image, the gray-level saliency map, $\boldsymbol{S}$, should be converted into a binary saliency map, $\boldsymbol{S}_b$, using a threshold value. Two different schemes are used to convert the gray level saliency map into a binary map; fixed and adaptive threshold binarization.

In the fixed threshold binarization, the threshold value is varied from 0 to 255 and it is applied to the gray-level saliency map. The binary saliency map corresponding to each threshold value is then compared to the ground truth and the precision, P, and recall, R, values are computed. Precision, also called positive predictive value, is the fraction of retrieved instances that are relevant, while recall, also known as sensitivity, is the fraction of relevant instances that are retrieved. Precision and recall are computed as

$$P = \frac{\sum_x \sum_y S_b(x,y) GT(x,y)}{\sum_x \sum_y S_b(x,y)}, \\ R = \frac{\sum_x \sum_y S_b(x,y) GT(x,y)}{\sum_x \sum_y GT(x,y)}, \tag{2.1}$$

where (x, y) represents a pixel's spatial location. The average precision and recall values for images in each dataset are shown on a precision-recall curve.

The use of the fixed thresholds is not sufficient to evaluate the performance of a method, since it is image independent. Thus, an adaptive image dependent threshold is also utilized. In the adaptive threshold binarization, the threshold is defined to be twice the mean of the saliency values of all the pixels in each gray-level saliency map as given by [27]

$$T_{adp} = \frac{2}{LW} \sum_x \sum_y S(x,y). \tag{2.2}$$

where L and W are, respectively, the numbers of rows and columns in the saliency map. The adaptive threshold is then applied to the gray-level saliency map to obtain the corresponding binary saliency map. For each saliency map, the precision and recall values are then computed using (2.1). Then, as a combination of the precision and recall metrics, the F-measure value, F_β, defined as

$$F_\beta = \frac{(1+\beta^2)PR}{(\beta^2 P) + R}, \tag{2.3}$$

where β^2 is a positive parameter specifying the relative importance of the precision and recall, is computed. In order to be consistent in comparing the performance of the proposed method with that of the existing works in the literature, we set $\beta^2 = 0.3$ as suggested in [27].

In order to further evaluate the performance of the proposed method, the mean absolute

error (MAE) between the binary saliency map using the adaptive threshold binarization and the ground truth is obtained as

$$MAE = \frac{1}{LW} \sum_{x} \sum_{y} \left| S_b(x, y) - GT(x, y) \right| . \tag{2.4}$$

The MAE value evaluates directly how similar is the binary saliency map to the ground truth.

2.5 Summary

In this chapter, the background material that is required for the investigation carried out in the succeeding chapters has been described. First, an introduction to the WT has been presented. Then, the NSCT and its properties in representing the image has been briefly reviewed. Next, an introduction to the general structure and characteristic of the CNNs has been presented. Finally, several metrics that are used to evaluate the performance of a salient object detection method have been discussed.

Chapter 3

Salient Object Detection Using Wavelet-based Image Features

3.1 Introduction

In view of the fact that the WT is capable of performing a multi-resolution spatial and frequency localized analysis, it could be an appropriate tool for extracting oriented details of an image and detecting salient regions of different sizes in various images. In [30], the WT is applied to the image and at each decomposition level a feature map is generated by setting the low-pass coefficients of that level to 0 and applying the inverse WT. Then, a local saliency map is computed by linear combination of the feature maps obtained at various levels, while a global saliency map is computed based on the distribution of the feature maps. However, since the last decomposition level consists of all the detail coefficients of the previous levels, no additional information is obtained by applying the inverse transform after each decomposition level. Also, generating these feature maps increases the computational complexity of the method.

In this chapter, utilizing the WT-based image features, a new salient object detection method is proposed [70, 71]. In Section 3.2, the proposed scheme is developed by devising the various steps involved in obtaining the channel feature maps, their associated weights, and the image feature map that lead to obtaining the final saliency map. In Section 3.3,

experiments are carried out in order to demonstrate the effectiveness of the proposed salient object detection scheme and compare its performance with that of some of the existing schemes in the literature. Finally, in Section 3.4, the work of this chapter is summarized and the significant features of the proposed scheme are highlighted.

3.2 Proposed Salient Object Detection Method

The proposed method attempts to detect salient regions of images using WT-based textural features of the three color channels and incorporates them using an efficient weighting scheme. In this method, the input image is converted to the CIELAB color space, having luminance, red/green, and blue/yellow channels, denoted by L, a, and b, respectively, since, this color space is perceptually more uniform than the RGB color space. The proposed method consists of the steps of extracting feature maps corresponding to the three color channels using WT decomposition, linearly combining the three channel feature maps based on the concept of entropy and a border avoidance criterion, modifying the feature map by making use of the centers of gravity of the image [72], and obtaining the final saliency map by bilateral filtering [73] of the modified image feature map. In the following, these four steps are described in detail.

3.2.1 Wavelet-based Feature Maps of Color Channels

In the proposed technique, the WT is utilized to extract textural details of the image at different scales, $n = 1, ..., N$, where depending on the size of the input image and size of the wavelet filter used, N is the maximum level of WT decomposition. The operation of WT carried out individually to the L, a and b channels of the image yields,

$$\left\{ \boldsymbol{A}_N^c, \{\boldsymbol{H}_n^c, \boldsymbol{V}_n^c, \boldsymbol{D}_n^c\}_{n=1,...,N} \right\} = \mathrm{WT}_N(\boldsymbol{I}^c), \tag{3.1}$$

where $\text{WT}_N(.)$ denotes the wavelet transform at level N, $\boldsymbol{I}^c$, $c \in \{L, a, b\}$, represents the color channels of the input image, $\boldsymbol{A}_N^c$ is the approximation coefficients at the coarsest level, N, and $\boldsymbol{H}_n^c, \boldsymbol{V}_n^c, \boldsymbol{D}_n^c$ are the matrices of appropriate sizes representing the horizontal, vertical and diagonal sub-band coefficients of the channel c at level n, respectively.

To extract textural details from the WT decomposition of the image, the low-pass coefficients at the coarsest decomposition level, N, are set to 0, and the 2D signal is reconstructed by applying the inverse WT as

$$\boldsymbol{TMAP}^c = \text{IWT}_N \left(\boldsymbol{A}_N^c = 0, \{\boldsymbol{H}_n^c, \boldsymbol{V}_n^c, \boldsymbol{D}_n^c\}_{n=1,\ldots,N} \right), \tag{3.2}$$

where $\text{IWT}_N(.)$ denotes the inverse wavelet transform at level N, and $\boldsymbol{TMAP}^c$ is the texture map of the channel c.

The channel feature map of a location (x, y), $\text{FMAP}^c(x, y)$ for the channel c, is obtained by enhancing the high-intensity values and suppressing the low-intensity values of the channel texture map at that location, $\text{TMAP}^c(x, y)$, as

$$\text{FMAP}^c(x, y) = \left(\text{TMAP}^c(x, y)\right)^2. \tag{3.3}$$

In order to investigate the effect of the number of decomposition levels used on the quality of the feature maps, we stop decomposition at different levels, $m = 1, \ldots N$, and construct a feature map for each m by setting the low-pass coefficients of the mth level to 0 and reconstructing the 2D signal by using the detail coefficients of all the levels from 1 to m. Figure 3.1 shows an example of the feature maps for the three channels resulting from a 200×150 color image after each of the decomposition levels from $m = 1$ to $m = 8$, which is the maximum decomposition level for the image of this size, by employing the Daubechies wavelets (Daub.7). It is seen from this figure that, as the number of decomposition levels is increased, the proposed method results in increasingly improved textural

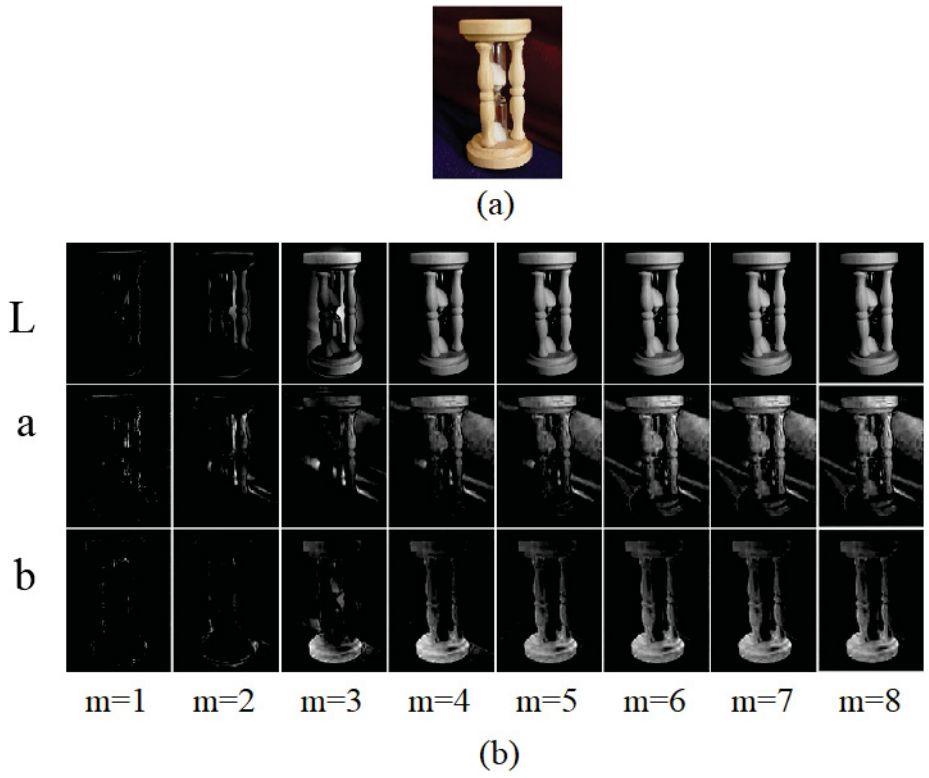


Figure 3.1: (a) Original 200×150 color image, (b) Channel feature maps obtained after m decomposition levels for L, a and b channels.

features for each of the channels. It is noted that for $m < N$ the number of sub-bands used to construct the channel feature maps is a subset of the number of sub-bands used for constructing these channel feature maps for $m = N$. Thus, a linear combination of the feature maps for $m = 1, ...N$, for a given c, cannot be expected to improve the quality of the channel feature maps over that constructed by using $m = N$ levels of decompositions. Thus, we advocate to construct the channel feature maps only after the last level of decomposition, i.e., $m = N$, which should also result in reduced computational complexity.

3.2.2 Image Feature Map

After constructing channel feature maps for each of the L, a, and b channels, an image feature map, $\boldsymbol{FMAP}$, is obtained through a weighted linear combination of the channel

feature maps as follows

$$\boldsymbol{FMAP} = \sum_{c \in \{L,a,b\}} \omega^c \boldsymbol{FMAP}^c, \tag{3.4}$$

where ω^c is the weight assigned to the feature map corresponding to the channel c.

In determining the values of weights, ω^c, we focus on two aspects of a desirable feature map. A desirable feature map should have a cluster of pixels with high values of the gray levels corresponding to the salient region and the rest of pixels with low values. The other consideration is the fact that since salient objects are generally center biased, weights should be determined so as to provide less importance to the borders of the channel feature map.

In order to take into consideration the first aspect of a desirable feature map, we use the entropy of the channel feature maps. A channel feature map having only two levels of pixels would have an entropy value lower than that of a channel feature map having a larger number of gray levels of the pixels. Figure 3.2 shows an example of synthetic images with pixel values of 2 and 4 gray levels and their entropy values. Thus, the channel feature map with a smaller entropy should be assigned a larger weight and vice-versa. However, the problem with determining a weight based on the entropy value of the channel feature map is that it does not take into account the spatial distribution of the pixel gray levels across the map. Figure 3.3 shows an example of two binary images of size 20×20 pixels. The number of pixels with a given gray level are the same in both the images. Despite the fact that distribution of these two gray levels are different in the two images, they both have the same entropy value. In order to use the entropy value for determining the weights, we would like to increase the entropy value in a situation in which the pixel gray levels of the channel feature map is more scattered. Therefore, we propose the channel feature map to undergo a low-pass filtering, a process through which the number of gray levels in the two images with different distributions will increase. Accordingly, their entropy value

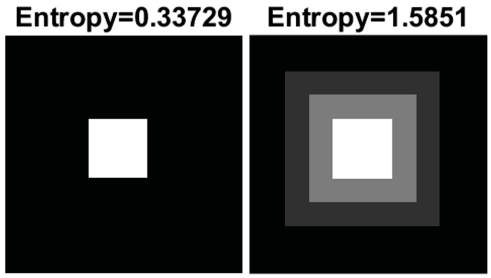


Figure 3.2: Synthetic images with (a) two gray levels of 0 and 1, and (b) four gray levels of 0, 0.25, 0.5, 1 and their entropy values.

will also increase. However, the increase in the entropy value of the image with a scattered distribution is larger than that of the image with a clustered distribution. Figure 3.4 shows the same two images as shown in Figure 3.3 after they are filtered using a 3×3 low-pass Gaussian filter, along with their entropy values. It is noted that after low-pass filtering, the entropy value of the image with the scattered distribution of pixels with the gray level of unity is increased much more than that of the image in which the pixels with gray level of unity are clustered together. Thus, the entropy value of the low-pass filtered channel feature map can be considered as a parameter in assigning weights to the channel feature maps. The entropy value is computed as

$$\varepsilon^c = \eta\left(\boldsymbol{FMAP}^c * \boldsymbol{G}\right), \tag{3.5}$$

where $\eta(.)$ is the entropy, $\boldsymbol{G}$ is a low-pass Gaussian filter, and $*$ represents the 2D convolution of the two associated matrices. The standard deviation of the Gaussian filter should be large enough so as to include sufficient number of neighboring pixels around each pixel in the filtering process.

In order to take into consideration the fact that in most of the natural images the salient region is located close to the center rather than on or near to the borders, in the linear combination of (3.4) a smaller weight is assigned to a channel feature map with a strong

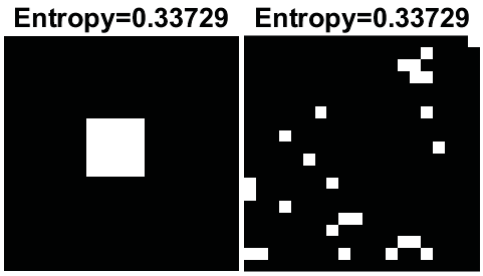


Figure 3.3: Synthetic images of size 20×20 pixels, each of two images has 25 pixels with the gray level of 1 and 375 pixels with the gray level of 0.

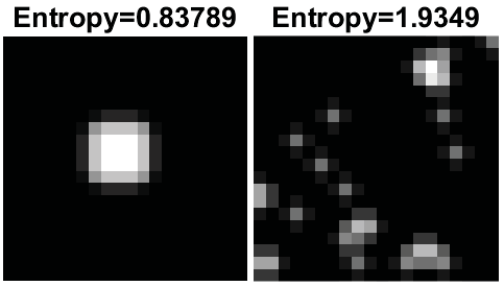


Figure 3.4: Images of 3.3 after low-pass filtering and their entropy values.

response at the border. In this work, utilizing the border-avoidance criterion in [19], the strength of the channel feature map is computed with a greater emphasis given to a salient-like region at the center rather than at the border of the image, as

$$\alpha^c = \sum_x \sum_y Z(x,y).Nr(\text{FMAP}^c(x,y)), \tag{3.6}$$

where $Z(x,y)$ is the (x,y)th element of a Gaussian mask $\boldsymbol{Z}$, of the same size as that of the channel feature map and its entries normalized to a maximum value of 1, and $Nr(\text{FMAP}^c(x,y))$ represents the (x,y)th element of the normalized image feature map $\boldsymbol{FMAP}^c / \sum_x \sum_y \text{FMAP}^c(x,y)$.

Thus, in order to emphasize the presence of a salient region at the center and de-emphasize any possible salient-like regions at the border of the image, we propose the weights in (3.4) to be chosen as

$$\omega^c = \left(\alpha^c/\varepsilon^c\right)^4, \tag{3.7}$$

where the power 4 in the expression for this weight is empirically set. The image feature map is then computed through the linear weighted combination of channel feature maps given by (3.4).

3.2.3 Modification of the Image Feature Map by Considering Centers of Gravity

The image feature map obtained as above is further refined by taking centers of gravity into account. Centers of gravity are defined as one or several pixels about which the visual form of the image is organized [72]. The regions surrounding the centers of gravity attract our attention. Thus, the saliency value at the locations around the centers of gravity should be greater than those of the locations that are far away. In view of this, first the pixels whose intensity values in the feature map exceed a certain threshold are identified as centers of gravity in the same way as in [74]. Then, all the other pixels are weighted according to their Euclidean distances from the closest center of gravity in order to obtain a refined image feature map, as

$$\begin{aligned} &MFMAP(x,y) = \\ &FMAP(x,y)\left(1 - \min\left\{\|(x,y),(\acute{x},\acute{y})\| \,|(\acute{x},\acute{y}) \in \text{Centers of Gravity}\right\}\right), \end{aligned} \tag{3.8}$$

where $\|.\|$ represents the Euclidean positional distance between the pixel at location (x, y) and the center of gravity at $(\acute{x}, \acute{y})$.

In order to have a smooth saliency map, in most of the existing saliency detection methods, a low-pass Gaussian filter is applied to the saliency map in the last step [19, 30].

It is known that in Gaussian low-pass filtering, the pixel value of filtered image at a given location is computed as the weighted average of pixel values in a neighborhood specified by the standard deviation of the filter. The weight decreases as the distance of a pixel from the neighborhood center increases. Since the nearby pixels are likely to have the same intensity values, it is appropriate to average them together. However, the idea fails at the edges where there is an abrupt change in the intensity values of the neighboring pixels, and it results in blurred edges. Thus, low-pass filtering of the saliency map destroys the borders of the salient region.

In this work, in order to smooth the modified image feature map, $\boldsymbol{MFMAP}$, while preserving the strong edges between salient and non-salient regions, a bilateral filter [73] is applied to it. Two pixels at an edge which are close to each other spatially could be very different in terms of their intensity values. The basic idea of bilateral filtering is considering both spatial closeness and intensity similarity of the pixels in assigning the weights in the filtering process. Thus, bilateral filtering can preserve high-contrast edges while removing low-contrast or gradual changes. A simple case of bilateral filtering is shift-invariant filtering, in which a Gaussian closeness filter and a Gaussian similarity filter are simultaneously used. In this work, the saliency map, $\boldsymbol{S}$, is obtained by bilateral filtering of the modified image feature map as

$$\boldsymbol{S} = \mathrm{BF}\left(\boldsymbol{MFMAP}\right), \tag{3.9}$$

where BF (.) denotes the bilateral filtering operation. The value of a pixel at location q of the $\boldsymbol{MFMAP}$ after bilateral filtering is obtained as

$$BF\left[MFMAP_q\right] = \frac{1}{W_q}\sum_{q^{'}\in\Omega} G_{\sigma_d}\left(\left\|q-q^{'}\right\|\right) G_{\sigma_r}\left(\left|MFMAP_q - MFMAP_{q^{'}}\right|\right) MFMAP_{q^{'}}, \tag{3.10}$$

where W_q is a normalization factor given by

$$W_q = \sum_{q^{'}\in\Omega} G_{\sigma_d}\left(\left\|q-q^{'}\right\|\right) G_{\sigma_r}\left(\left|MFMAP_q - MFMAP_{q^{'}}\right|\right), \tag{3.11}$$

where q denotes a location (x, y) in $\boldsymbol{MFMAP}$, $q^{'}$ denotes a location $(\acute{x}, \acute{y}) \in \Omega$, Ω being the set of possible positions in $\boldsymbol{MFMAP}$, and G_{σ_d} and G_{σ_r} are the Gaussian closeness (domain) and similarity (range) functions with the standard deviations of σ_d and σ_r, respectively.

The proposed wavelet-based salient object detection scheme is summarized in Algorithm 3.1.

3.2.4 Evaluating the Impact of the Main Phases of the Proposed Method

In this section, the effect of each phase of the proposed method is studied. To this end, first, the wavelet-based channel feature maps (as given in (3.3)) for a sample image obtained by employing the proposed method. Figure 3.5 shows the three color channels and the channel feature maps obtained. It can be seen that for this sample image the channel feature map corresponding to the channel a can represent the salient region better than represented by the maps corresponding to the channels L and b. It is also seen that some details of the salient region has been captured by the b channel feature map, but not by the L channel feature map.

Algorithm 3.1: Salient Object Detection Using Wavelet-based Features

Input: RGB Image $\boldsymbol{I}$

1 Convert the input image , $\boldsymbol{I}$, from RGB color space $c \in \{R, G, B\}$ to CIELAB color space $c \in \{L, a, b\}$

2 **for** $c \in \{L, a, b\}$ **do**

3 Perform the wavelet transform

4 Set low-pass coefficients to 0

5 Obtain channel texture map by performing the inverse wavelet transform

6 Obtain channel feature map using (3.3)

7 Calculate the weight of the channel feature map based on small entropy and border avoidance criteria using (3.7)

8 Obtain image feature map by linearly combining the 3 channel feature maps, obtained in Step 6, using the weights calculated in Step 7

9 Identify centers of gravity in the image feature map obtained in Step 8

10 Refine image feature map based on each pixel's distance to the centers of gravity using (3.8)

11 Apply bilateral filtering

Output: Saliency Map $\boldsymbol{S}$

The effects of the succeeding steps that use the three channel feature maps given in Figures 3.5 (d), (f) and (h) are depicted in Figure 3.6. Figure 3.6 (a) is simply the average of the three channel maps, that is, a map obtained by linearly combining the three maps with equal weights. Figure 3.6 (b) shows the image feature map obtained through a weighted linear combination of the three channel feature maps. A comparison of the maps in Figures 3.6 (a) and (b) clearly shows a positive effect of the proposed weighting scheme, as described in Section 3.2.2, on the image feature map. It is seen that after applying the linear combination using the proposed weights given by (3.7), the non-salient regions are suppressed effectively compared to the simple averaging of the channel feature maps. Figure 3.6 (c) shows the modified image feature map obtained by taking into consideration the centers of gravity in the image feature map of Figure 3.6 (b). A comparison of the maps in Figures 3.6 (b) and (c) shows that the saliency values of the salient regions have increased, while those of the non-salient regions have decreased or remained unchanged. However, there are still some regions in the background which have been wrongly detected as salient.

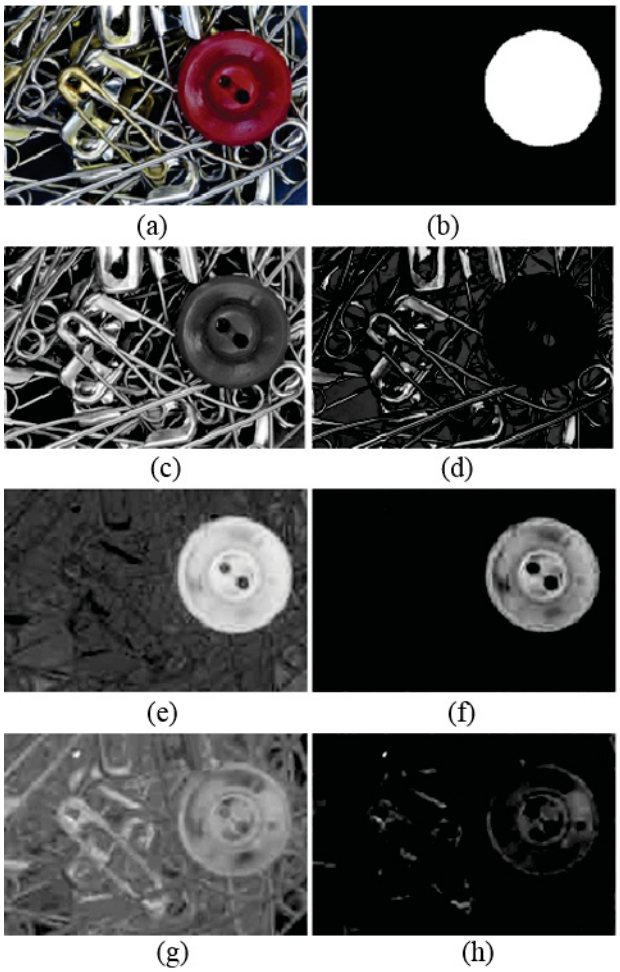


Figure 3.5: (a) Original image. (b) Ground truth. (c) L color channel. (d) L channel feature map. (e) a color channel. (f) a channel feature map. (g) b color channel. (h) b channel feature map.

Finally, Figure 3.6 (d) shows the gray-level saliency map by applying the bilateral filtering on the modified image feature map of Figure 3.6 (c). It is seen that, after the bilateral filtering the wrongly detected regions in the background are further suppressed, and that the salient region has become more uniform. It is noted that enhancement in the saliency map has been achieved while still preserving the boundary between the salient and non-salient regions.

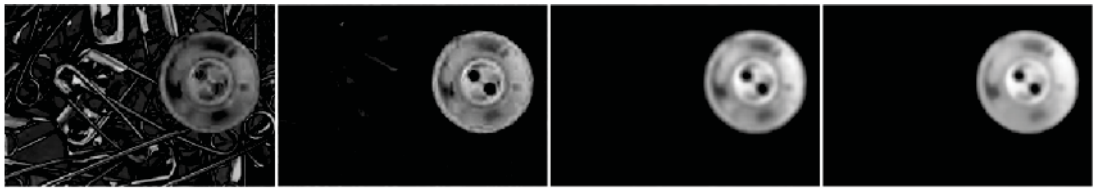


Figure 3.6: The maps obtained after the application of the proposed method. (a) Average of the channel feature maps. (b) Weighted linear combination of the channel feature maps. (c) Refined map by considering centers of gravity. (d) The final saliency map after bilateral filtering.

3.3 Experimental Results

Performance of the proposed salient object detection method is evaluated on six commonly used datasets of natural images, namely, MSRA-1000 [27] (1,000 images), MSRA-10K [34] (10,000 images), HKU-IS [44] (4,447 images), PASCAL-S [75] (850 images), DUT-OMRON [31] (5,172 images), and CSSD [76] (200 images) datasets.

In the proposed method, the Daubechies wavelets (Daub.7) are used to extract the features. The size of the filter results in a trade-off between the time complexity and promising quality of the saliency maps. The standard deviation of the low-pass Gaussian filter, $\boldsymbol{G}$, in (3.5) is set to $\sigma = 0.02 \times \frac{L+W}{2}$, and the standard deviation of the Gaussian mask, $\boldsymbol{Z}$, in (3.6) is set to $\acute{\sigma} = 0.25 \times \frac{L+W}{2}$, where L and W are, respectively, the number of rows and columns in the original image. The threshold to identify centers of gravity in (3.8) is chosen to be 0.8. The standard deviations of the Gaussian closeness and similarity filters in the bilateral filter given by (3.10) are set to $\sigma_d = \frac{\min(L,W)}{16}$ and $\sigma_r = \frac{\max(\boldsymbol{S})-\min(\boldsymbol{S})}{10}$, respectively.

3.3.1 Performance Comparison

Performance of the proposed salient object detection method is evaluated both subjectively and objectively and compared to that of the six other frequency domain salient object detection methods, namely, superpixel-based wavelet method (SW) [70], weighted

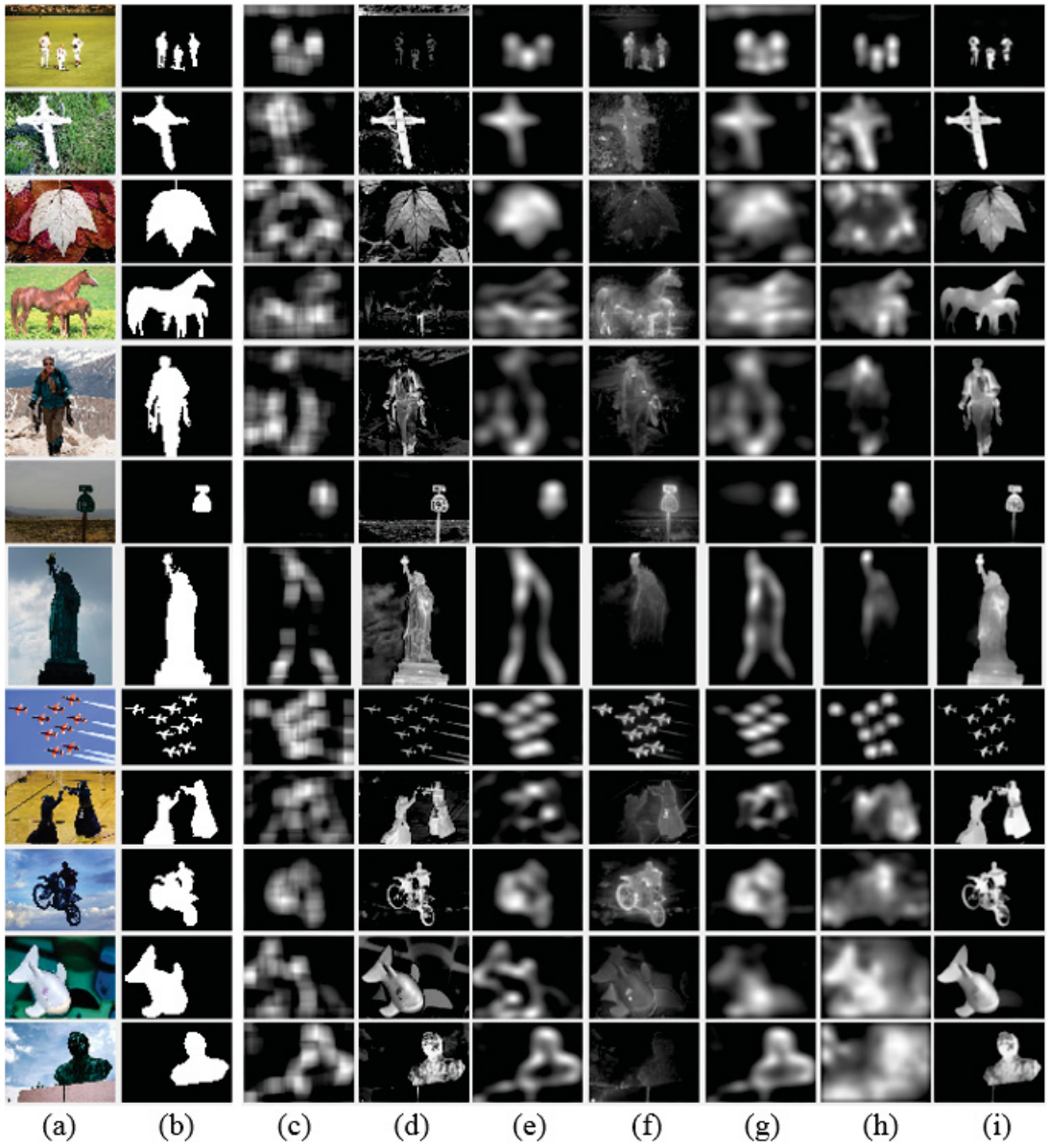


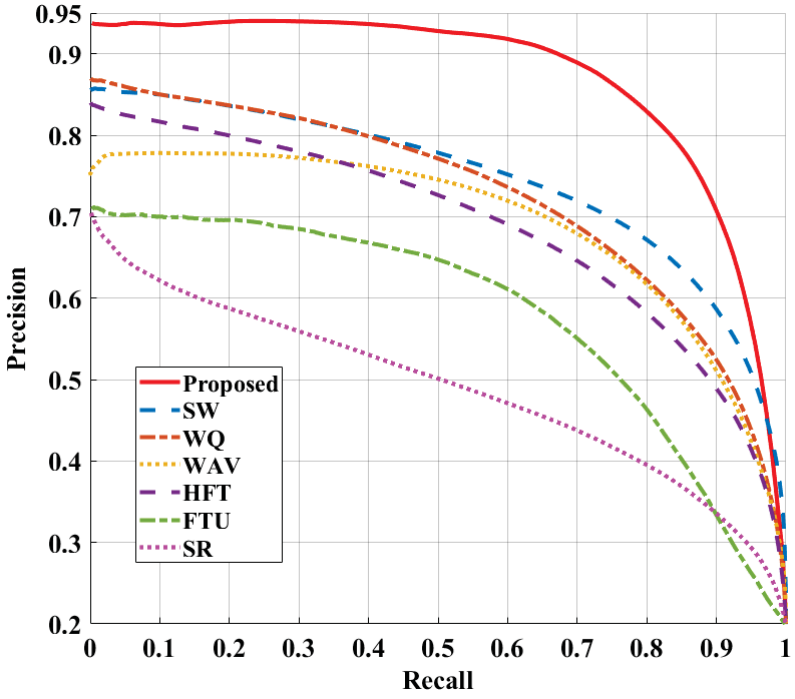
Figure 3.7: Saliency maps obtained by applying the proposed and other methods on sample images from the datasets in [27, 34, 76, 77]. (a) Original image. (b) Ground truth. (c) Spectral residual method (SR) [27]. (d) Frequency-tuned method (FTU) [19]. (e) Hyper-complex Fourier-base method (HFT) [19]. (f) Wavelet-based method (WAV) [30]. (g) Weighted quaternion-based method (WQ) [78]. (h) Superpixel-based wavelet method (SW) [70]. (i) Proposed method [71] .

quaternion-based method (WQ) [78], wavelet-based method (WAV) [30], hyper-complex Fourier-based method (HFT) [19], frequency-tuned method (FTU) [27], and spectral residual method (SR) [24].

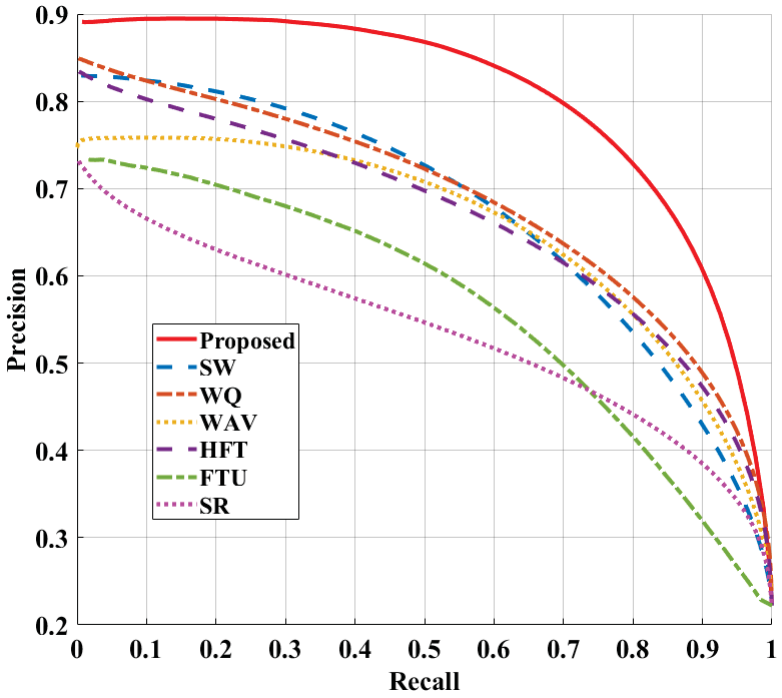
Figure 3.7, shows the saliency maps obtained by utilizing the proposed method as well as the other methods for some sample images. It is seen form this figure that the saliency maps obtained using the proposed method are more similar to the ground truth in comparison to the other methods. Also, in the saliency maps obtained, the salient regions are uniformly highlighted with a sharp boundary. It is seen from Figure 3.7 that some other methods are not able to detect the entire salient object. For instance SR [24] detects edges of the salient object. In the saliency maps obtained by FTU [19], the non-salient regions are not clean. Some other methods such as HFT [19] are not successful in detecting salient regions of large size (see sample images in rows 7, 11 and 12 of Figure 3.7). The saliency maps obtained by the method WAV [30] are very blurred and the salient region is not detected accurately.

Figure 3.8-3.10 depict the average precision-recall curves obtained by applying the proposed and the other schemes with the fixed thresholds. It is seen from this figure that the proposed scheme outperforms all the other schemes in terms of the precision-recall performance. For the entire range of the fixed thresholds on each of the six datasets, the proposed method obtains the largest values of precision and recall compared to the other methods. There are two other wavelet-based methods amongst the methods considered for comparison, namely SW [70] and WAV [30], which have smaller precision and recall values compared to those of the proposed method.

Figure 3.11 and 3.12 show the average precision, recall and F_β values obtained using the adaptive threshold given by (2.2). It is seen from this figure that the proposed method provides the largest values for the precision metric amongst all the methods regardless of the datasets on which the methods are applied. In terms of the recall metric, the proposed

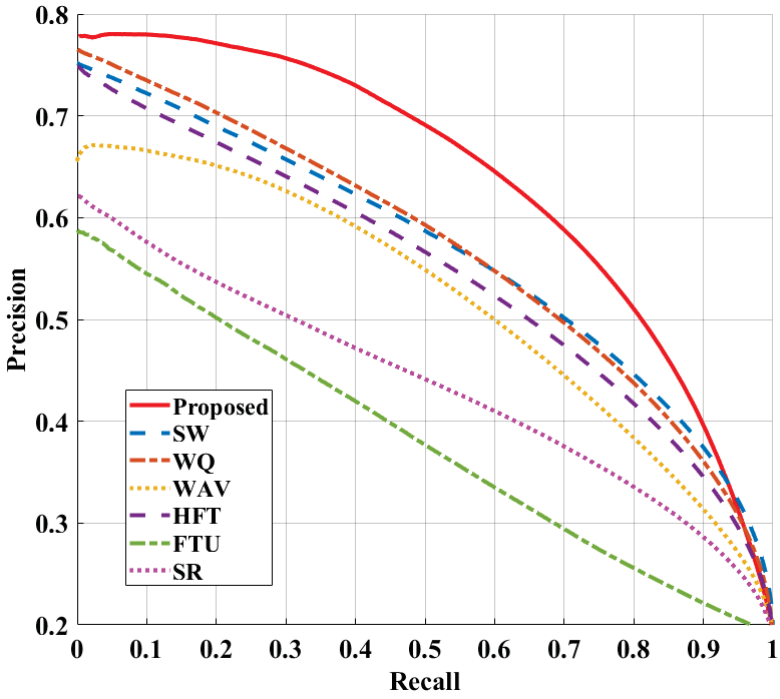


(a)

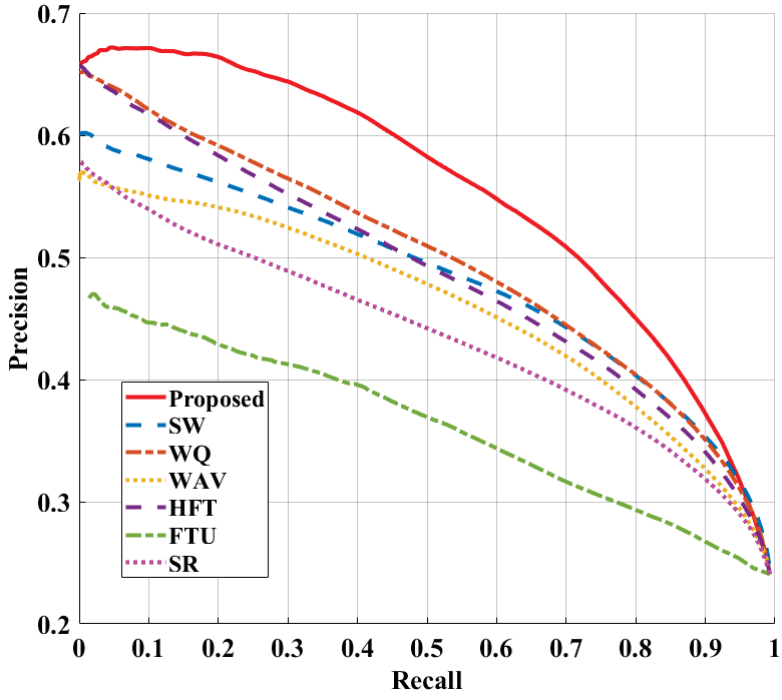


(b)

Figure 3.8: Precision-recall curves obtained by applying the proposed and other salient object detection methods on images in (a) MSRA-1000 and (b) MSRA-10K datasets.

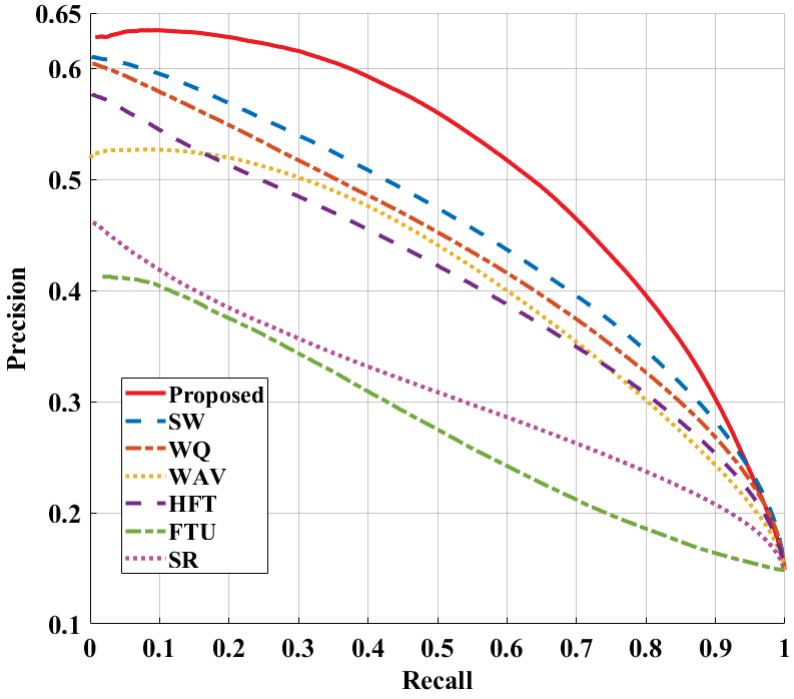


(a)

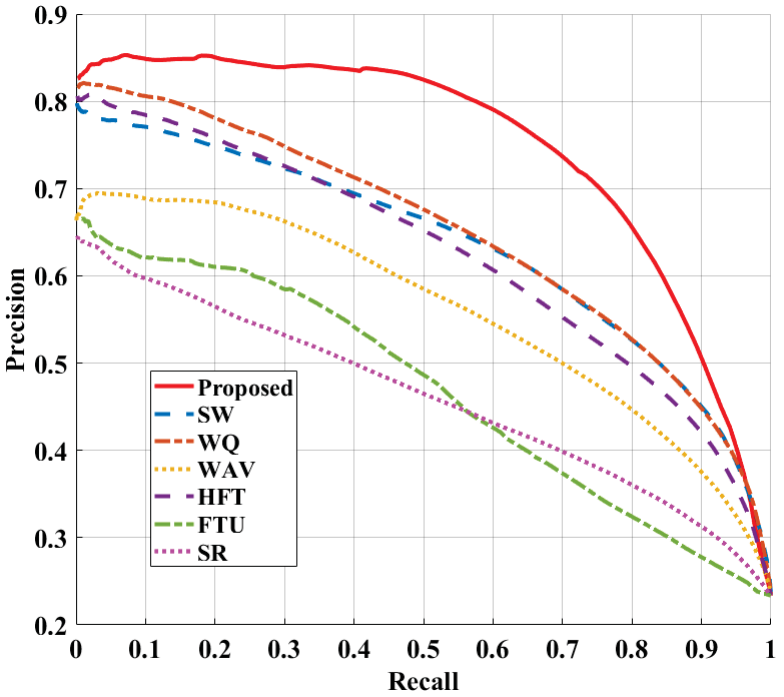


(b)

Figure 3.9: Precision-recall curves obtained by applying the proposed and other salient object detection methods on images in (a) HKU-IS and (b) PASCAL-S datasets.



(a)



(b)

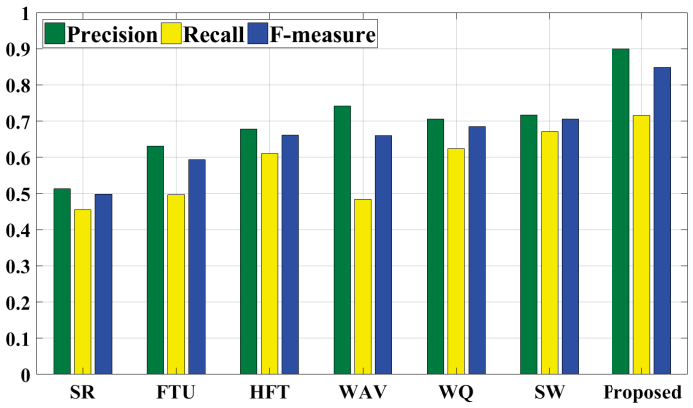
Figure 3.10: Precision-recall curves obtained by applying the proposed and other salient object detection methods on images in (a) DUT-OMRON and (b) CSSD datasets.

scheme is second only to SW scheme in the cases of the HKU-IS, PASCAL-S and DUT-OMRON datasets. However, the SW scheme has provided larger values for the recall metric in comparison to that of the proposed method at the expense of smaller values for the precision metric. Since, both of the precision and recall values are considered in computing the F_β metric, the overall performance of a salient object detection method can be evaluated using the F_β metric. It is seen from Figures 3.11 and 3.12 that for all the six datasets, the proposed method provides the largest value for the F_β metric amongst all the methods. To make the comparison simpler, the F_β values obtained by applying different methods are also given in Table 3.1.

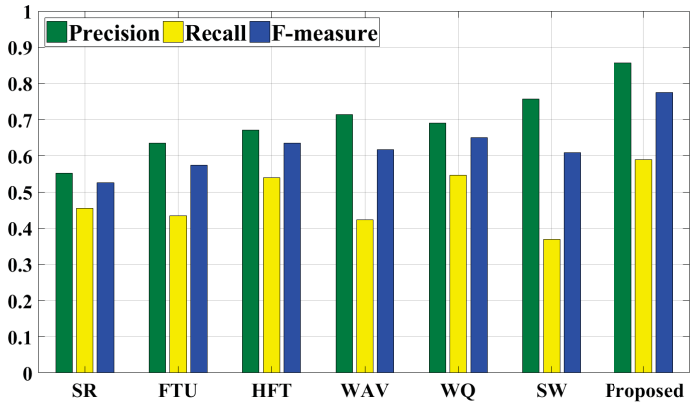
The MAE values obtained using the proposed method as well as by using the other methods are also given in Table 3.1. It is seen from this table that the proposed method provides the lowest value for the MAE metric amongst all the methods, irrespective of the datasets used in our experiments, indicating a strong similarity between the saliency maps obtained by applying the proposed method and the ground truth.

As seen from Figure 3.7, the saliency maps obtained using the proposed method are more similar to the ground truth in comparison to those obtained using the other methods, thus indicating the superiority of the proposed method. In addition, as seen from Figures 3.8-3.12, and Table 3.1, the proposed salient object detection method outperforms the other methods in terms of precision-recall performance for the fixed thresholds, and F_β and MAE values for the adaptive threshold. This performance improvement can be attributed to the incorporation of the wavelet-based textural feature maps that are able to represent the saliency-related information of each color channel, and their efficient linear combination using the proposed weighting scheme.

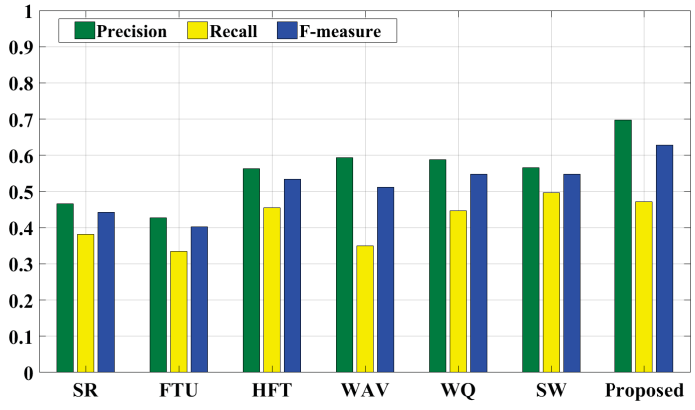
Using features that cannot suitably distinguish the salient regions from the non-salient ones, such as the spectral residual in the SR method [24], has resulted in inaccurate saliency maps (as seen from Figure 3.7 (c) rows 3, 5 and 7, for some test images), small precision



(a)

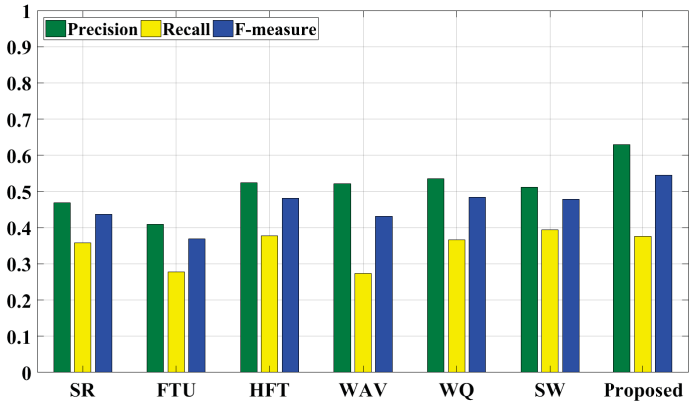


(b)

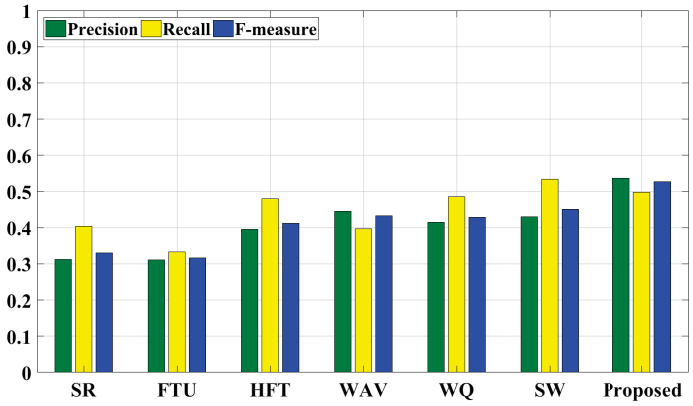


(c)

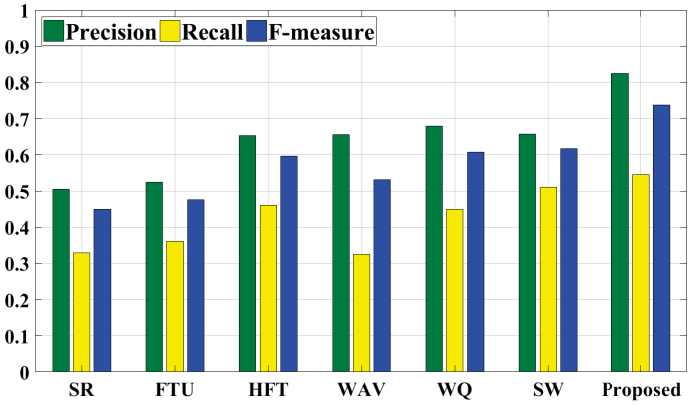
Figure 3.11: Precision, recall, and F_β obtained by applying the proposed and other salient object detection methods on images in (a) MSRA-1000, (b) MSRA-10K, and (c) HKU-IS datasets.



(a)



(b)



(c)

Figure 3.12: Precision, recall, and F_β obtained by applying the proposed and other salient object detection methods on images in (a) PASCAL-S, (b) DUT-OMRON, and (c) CSSD datasets.

and recall values, and large MAE values (as seen from Figures 3.8-3.12, and Table 3.1). In the FTU method [27], the use of only the color components as features, while ignoring other effective features such as textures, has led to the detection of the non-salient regions incorrectly as the salient regions (Figure 3.7 (d) rows 3, 5-7, 9, and 11). In the HFT method [19], the features have been extracted using the hyper-complex Fourier transform, which is more suitable for extracting global irregularities. As a result, this method has detected only the borders of the salient region and has failed in detecting the entire salient object in images with large salient regions (see rows 7, 11, and 12 of Figure 3.7 (e)). In the SW [70] and the WAV [30] methods, the wavelet-based features have been used. The wavelet transform is particularly suitable in representing the image details. However, since all the details employed in these methods are not related to the salient regions, they have failed in detecting the salient regions in some of the test images (see Figures 3.7 (f) and (h), rows 3, 11, and 12). As seen from Figures 3.8-3.12, and Table 3.1, these two methods have yielded smaller precision and recall values and larger MAE values compared to the proposed method. In the WT decomposition, the first decomposition levels extract edges rather than textures, while the coarsest decomposition level consists of all the saliency-related textural details of the image. In the proposed method, the extraction of a feature map only after the wavelet decomposition at the coarsest level has resulted in more accurate saliency maps, larger precision and recall values, and smaller MAE values as seen from Figures 3.7 (i), 3.8-3.12, and Table 3.1, respectively.

The way the extracted features are utilized has a significant impact on the salient object detection. In the HFT method [19], the maps have been generated at different levels but only one of them has been selected as the final saliency map, resulting in not necessarily an accurate detection of the salient objects (see Figure 3.7 (e), rows 4 and 9). In the WQ method [78], using a set of pre-specified weights to linearly combine the features has led to an inaccurate detection of the salient regions in some images (see Figure 3.7 (g), rows

Table 3.1: F_β and MAE values obtained by applying the proposed and the other methods on images in the six datasets

Dataset	**MSRA-1000**		**MSRA10K**		**HKU-IS**	
Metric	F_β	**MAE**	F_β	**MAE**	F_β	**MAE**
Proposed [71]	**0.8487**	**0.0785**	**0.7755**	**0.1213**	**0.6284**	**0.1458**
SW [70]	0.7061	0.1238	0.6091	0.1701	0.5482	0.1692
WQ [78]	0.6846	0.1346	0.6505	0.1625	0.5480	0.1672
WAV [30]	0.6606	0.1442	0.6166	0.1716	0.5119	0.1703
HFT [19]	0.6610	0.1413	0.6357	0.1657	0.5338	0.1722
FTU [27]	0.5940	0.1622	0.5741	0.1885	0.4018	0.2155
SR [24]	0.4982	0.1887	0.5260	0.1990	0.4428	0.1946
Dataset	**PASCAL-S**		**DUT-OMRON**		**CSSD**	
Metric	F_β	**MAE**	F_β	**MAE**	F_β	**MAE**
Proposed [71]	**0.5432**	**0.2170**	**0.5274**	**0.1362**	**0.7376**	**0.1516**
SW [70]	0.4773	0.2410	0.4506	0.1627	0.6168	0.1903
WQ [78]	0.4806	0.2363	0.4291	0.1635	0.6076	0.1922
WAV [30]	0.4286	0.2411	0.4329	0.1513	0.5313	0.2071
HFT [19]	0.4787	0.2397	0.4119	0.1689	0.5958	0.1938
FTU [27]	0.3662	0.2766	0.3160	0.2047	0.4752	0.2334
SR [24]	0.4355	0.2530	0.3303	0.1921	0.4500	0.2262

3-5). In the proposed method, the channel feature maps have been combined by applying a weighting scheme, in which a larger weight is assigned to a channel feature map that can represent the salient region more efficiently.

3.4 Summary

In this chapter, a salient object detection method has been proposed by using a new weighted linear combination of the wavelet-based feature maps. The textural features of the image have been extracted using the wavelet coefficients of the three color channels, and an effective feature map fusion scheme based on the concept of entropy and border

avoidance criterion has been proposed. In order to take into consideration both the spatial and intensity information of the pixels, in this scheme, the entropy value of the low-pass filtered map has been utilized. The map thus obtained has been further refined based on the image centers of gravity. Finally, unlike most of the existing methods, a bilateral filter has been applied to the resulting map in order to smooth it while preserving the sharp boundaries between salient and non-salient regions.

Several experiments have been carried out by applying the proposed scheme on the images from several datasets in order to evaluate its performance and to compare it to other existing methods. It has been shown that the saliency maps obtained using the proposed method is more similar to the salient regions detected by HVS. The proposed method provides the values of precision, recall and F_β higher than those provided by the other methods.

Chapter 4

Salient Object Detection Using Feature Extraction in the Non-subsampled Contourlet Domain

4.1 Introduction

Despite the advantages of the WT, it is known that the wavelets are optimal only in representing point discontinuities, but not very effective in capturing line discontinuities corresponding to the directional details in the image [79]. In order to circumvent the lack of directional selectivity of the wavelet, other multi-resolution and multi-directional representations, such as the NSCT [58], have been developed. A salient object detection method has been proposed in [36] by fusing local and global information from the NSCT coefficients of the images. However, it has not fully taken into account the effective visual features, such as texture and structure, since these attributes provide significant information towards the local and global characteristics of the image and are known to be of great importance in salient object detection.

In this chapter, a salient object detection method is proposed utilizing the non-subsampled contourlet coefficients of the three color channels of the CIELAB color space [80, 81]. The proposed method is realized by extracting local and global features from the non-subsampled contourlet coefficients of the color channels. A saliency map is obtained based

on a linear combination of the local features and the distribution of the global features. In order to provide a better preservation of the structure and boundary of the objects and to obtain a more uniformly highlighted salient region, the saliency map is abstracted using an optimization framework.

Section 4.2 presents the details of developing the proposed salient object detection method using the NSCT domain features. Simulation results are provided in Section 4.3. Finally, Section 4.4 concludes the work presented in this chapter.

4.2 Proposed Salient Object Detection Method

In this section, the various steps of the proposed salient object detection method are presented.

The proposed scheme involves generating local and global feature maps. The local feature maps are generated in order to detect pixels that are salient in a limited neighborhood, while the global feature maps detect pixels that are salient in the entire image. The local features are extracted from the local variations of the low-pass coefficients whereas the global features are obtained based on the distribution of the directional subband coefficients. To extract the image visual features effectively, the saliency map thus obtained is finally abstracted into meaningful regions by applying an optimization framework.

The input image is first converted to the CIELAB color space. Then, the NSCT of different decomposition levels, $m = 1, ..., N$, is applied to each color channel in order to extract local and global features from the non-subsampled contourlet coefficients. At each decomposition level, each channel is decomposed into a low-pass subband and a number of bandpass directional subbands as

$$\left\{ \boldsymbol{A}_m^c, \boldsymbol{D}_{n,d}^c \right\} = \text{NCST}_m \left(\boldsymbol{I}^c \right), \tag{4.1}$$

where NSCT$_m(.)$ denotes the non-subsampled contourlet transform at level m, $\boldsymbol{I}^c$, $c \in \{L, a, b\}$, represents the color channels of the input image, $\boldsymbol{A}_m^c$ denotes the low-pass subband, $\boldsymbol{D}_{n,d}$ denotes set of bandpass directional subbands corresponding to the level $n = 1, ..., m$ and the direction $d = 1, ..., q$ of color channel c.

By taking advantage of such multi-scale and directional decomposition, in the proposed method, two feature maps are generated, namely, a local feature map based on the statistical parameters of the low-pass coefficients and a global feature map by utilizing the directional subband coefficients.

4.2.1 Local Saliency Map Using Textural Features

In this section, the local saliency map generation algorithm used in the proposed method is described. Our approach to this end is first applying $m = 1, ..., N$ level decomposition to each color channel. Then, after each level of decomposition the low-pass subband is divided into blocks of $b \times b$ pixels, and the local variance of each block, Var(.), is considered as the local feature map value of the block as given by

$$\text{Lmap}_m^c(x, y) = \text{Var}\left\{A_m^c(x + i, y + j)\right\}_{\{i,j=\pm 3,\pm 2,\pm 1,0\}}. \tag{4.2}$$

Figure 4.1 shows the local feature maps for the color channels of a sample image after the mth decomposition level. It is seen from this figure that the local feature maps represent various textural details of the image at different levels.

The local saliency map, $\boldsymbol{S}_{Local}$, is then obtained by linearly combining the local feature maps $\boldsymbol{Lmap}_m^c$ as

$$\boldsymbol{S}_{Local} = \sum_{c,m} \omega_m^c \boldsymbol{Lmap}_m^c. \tag{4.3}$$

As explained earlier in Chapter 3, the weights are assigned according to the entropy values of the local feature map and pixel intensity values around the center of the map, in

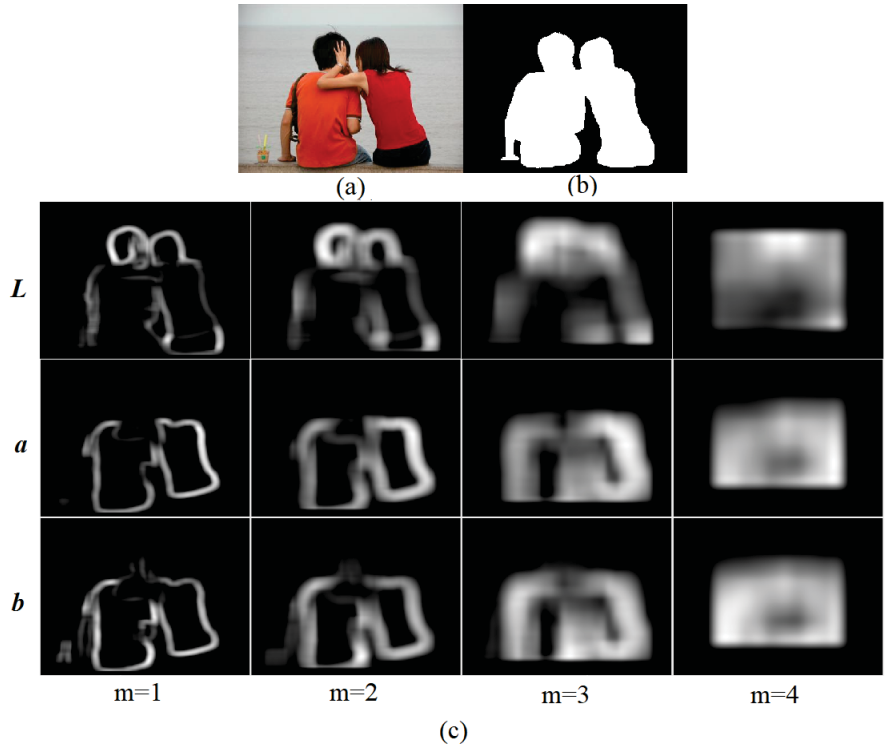


Figure 4.1: (a) A sample image, (b) the corresponding ground truth, and (c) local feature maps obtained after $m = 1, 2, 3, 4$ level of NSCT decomposition for the L, a and b color channels.

Algorithm 4.1: Determination of Local Saliency Map

Input: Image Color Channels in CIELAB color space $\boldsymbol{I}^c$, $c \in \{L, a, b\}$

for $c \in \{L, a, b\}$ **do**
  for $m = 1, ..., N$ **do**
    Perform an m level non-subsampled contourlet decomposition
    Obtain local feature map based on the variance of $b \times b$ blocks in the low-pass subband
    Calculate the weight of the local feature map based on small entropy and border avoidance criteria using (4.4)
  Obtain local saliency map by linearly combining the local feature maps corresponding to different decomposition levels of the 3 color channels, obtained in Step 4, using the weights calculated in Step 5

Output: Local Saliency Map $\boldsymbol{S}_{Local}$

such a way that a larger weight is assigned to a local feature map having small value of entropy and large values around the center of the map. Therefore, using (3.5) and (3.6), the weights in (4.3) are calculated as

$$\omega_m^c = (\alpha_m^c / \varepsilon_m^c)\,. \tag{4.4}$$

The algorithm to obtain the local saliency map is summarized in Algorithm 4.1 .

4.2.2 Global Saliency Map Using Bandpass Directional Subbands

In this section, the global saliency map generation algorithm used in the proposed method is described. In order to generate the global feature map, after each decomposition level, we set the low-pass coefficients to 0 for each color channel, and apply the inverse non-subsampled contourlet transform, INSCT(.), to the bandpass coefficients as

$$\boldsymbol{Gmap}_m^c = \text{INSCT}_m\left(\boldsymbol{A}_m^c = 0, \boldsymbol{D}_{n,d}^c\right). \tag{4.5}$$

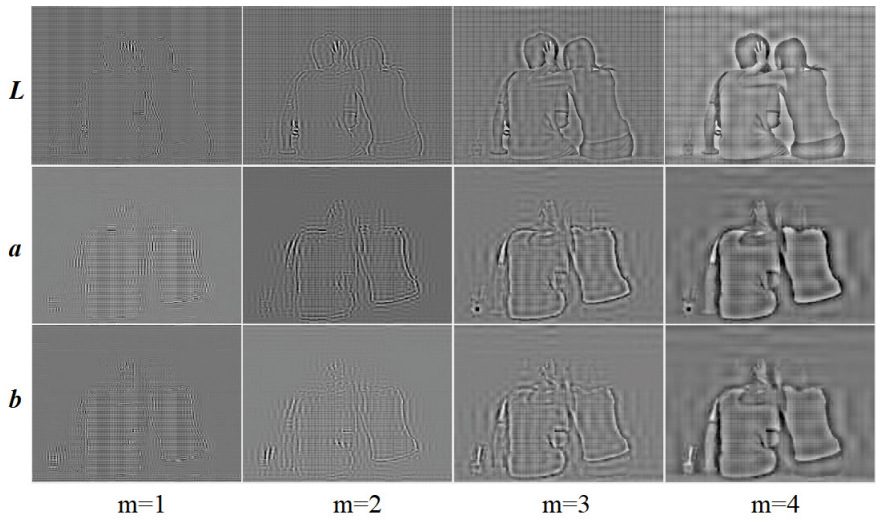


Figure 4.2: Global feature maps obtained after $m = 1, 2, 3, 4$ level of NSCT decomposition for the L, a and b color channels.

The resulting inverse transform are the global feature maps. The global feature maps corresponding to the different levels of decomposition and each channel for the image in Figure 4.1 are shown in Figure 4.2.

The elements of the global feature maps are then used to form global feature vectors, each corresponding to one pixel of the image. For each pixel, a global feature vector, $f_g(x, y)$, with a length of $l = 3X$ (X being the number of the global feature maps for each channel) is constructed. If a particular vector is less similar to the others, the corresponding pixel is different from the other pixels and thus it can be considered to be more salient. In view of this, the distribution of the global feature vectors are modeled by a Gaussian distribution [82] and the global saliency of each pixel is defined as the likelihood of finding its global feature vector amongst all the vectors, as

$$p\left(f_g(x,y)\right) = \frac{1}{\sqrt{(2\pi)^l \left|\Sigma\right|}} \exp\left\{\frac{-1}{2}\left(f_g(x,y)-\mu\right)^T \Sigma^{-1}\left(f_g(x,y)-\mu\right)\right\}, \tag{4.6}$$

where μ is a vector containing the means of the global feature maps, Σ is an $l \times l$ covariance

Algorithm 4.2: Determination of Global Saliency Map

Input: Image Color Channels in CIELAB color space $\boldsymbol{I}^c$, $c \in \{L, a, b\}$

for $c \in \{L, a, b\}$ **do**
    for $m = 1, ..., N$ **do**
        Perform an m level non-subsampled contourlet decomposition
        Set low-pass coefficients to 0
        Obtain global feature map by performing the inverse non-subsampled contourlet transform
Construct a global feature vector for each image pixel using the global feature maps corresponding to different decomposition levels of different color channels
Model the distribution of the global feature vectors by a Gaussian distribution
Obtain global saliency map value at each pixel location from the likelihood of finding its global feature vector using (4.7)

Output: Global Saliency Map $\boldsymbol{S}_{Global}$

matrix, $(.)^T$ is the transpose operator, and $|.|$ denotes the determinant of a matrix. Using (4.6), the global saliency map, $\boldsymbol{S}_{Global}$, is computed as

$$S_{Global}(x, y) = \log \left(p \left(f_g(x, y) \right)^{-1} \right)^{0.5}. \tag{4.7}$$

The algorithm to obtain the global saliency map is summarized in Algorithm 4.2 .

The local and global saliency maps obtained for the sample image in Figure 4.1 are shown in Figures 4.3 (a)-(b). It is seen from this figure that the global saliency map provides some meaningful information, especially around the edges of the salient object, which cannot be detected well only by using the local saliency map. This is due to the fact that global saliency map comprises of the statistical relation among the global feature maps. An image region is salient if it is different from its surrounding regions and also it stands out in the entire image. In view of this, after computing the local and global saliency maps, the pixels which are both locally and globally salient are considered as the final salient pixels. Thus, the local and global saliency maps are merged using a fusion algorithm that involves a Hadamard product followed by a process that normalizes the pixels in this last map in the range [0,1]. The purpose of the fusion algorithm is to give prominence to the pixels that

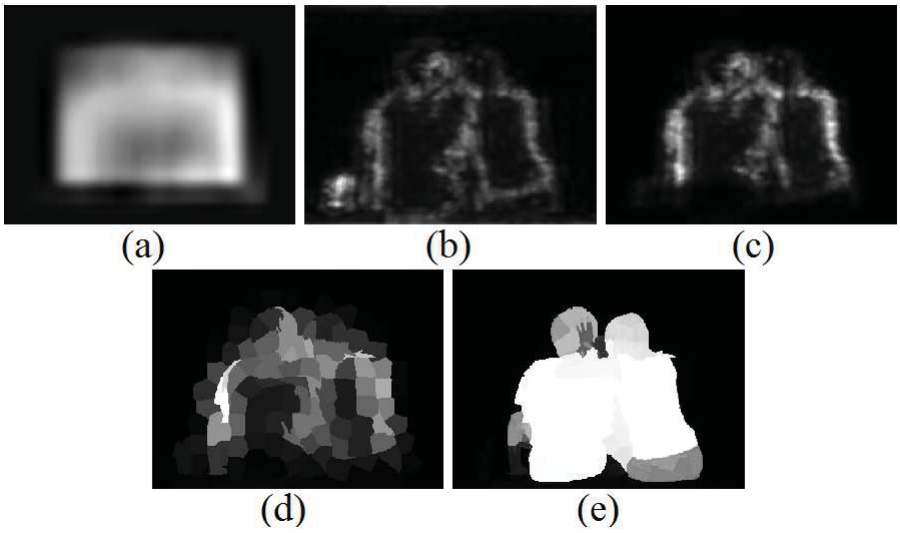


Figure 4.3: (a) Local and (b) global saliency maps, (c) map obtained by fusion of local and global saliency maps, (d) saliency map averaged over superpixels, (e) saliency map after abstraction.

have large values both in the local and global saliency maps, whereas to assign small values to pixels that have small values in either of the local and global maps or in both. The fused map S_M is given by

$$\boldsymbol{S}_M = Nr(\boldsymbol{S}_{Local} \circ \boldsymbol{S}_{Global}), \tag{4.8}$$

where $Nr(.)$ represents the normalization process of the pixels of the associated map in the range [0,1] and $\circ$ denotes the Hadamard matrix product. The fused map obtained is shown in Figure 4.3 (c).

4.2.3 Image Saliency Map Abstraction

It is known that a salient object detection technique aims at detecting the most distinctive regions or objects rather than indivisual pixels. In view of this, the saliency map is further abstracted into perceptually meaningful regions. More specifically, the input image is segmented into k superpixels using the simple linear iterative clustering (SLIC) algorithm [83], and the saliency value of each superpixel, is replaced by the average of the

saliency values of all the pixels belonging to that superpixel.

Figure 4.3 (d) shows the saliency map averaged over superpixel for the sample image in Figure 4.1. The saliency map abstraction is implemented using the optimization technique of [32] in which the saliency values of superpixels are used as the foreground weights. In this optimization problem, the cost function given by

$$\sum_{i=1}^{k} wb_i \, S_i^2 + \sum_{i=1}^{k} wf_i \left(S_i - 1\right)^2 + \sum_{i,j} ws_{ij} \left(S_i - S_j\right)^2, \tag{4.9}$$

where S_i is the optimized saliency value of the superpixel i, and wf_i, wb_i, and ws_{ij} denote the foreground, background and smoothness weights, respectively, is minimized. The cost function contains three terms corresponding to the pixels in the foreground and background, and the smoothness of the pixels within the two regions. The foreground term tries to assign a saliency value of 1 to the superpixels with a larger value in the saliency map. On the other hand, the background term aims at assigning a saliency value of 0 to the superpixels with a strong boundary connectivity. Boundary connectivity is measured in terms of the spatial distance and the CIELAB color distance between a superpixel and the superpixels located around the image boundary. The smoothness term tries to assign the same saliency values to the superpixels within the foreground region or the background region that are spatially close to each other.

Since in most of the images, image boundary pixels belong to the background, the degree of contrast of a pixel from the boundary pixels, provides a measure of its belonging to the salient region. Therefore, we generate a boundary contrast map by computing each pixel's distance to the mean color and covariance matrix of the boundary pixels. The abstracted saliency map and the boundary contrast map are then pixel-wise added.

The saliency maps are finally refined employing a post-processing algorithm. First, to smooth the saliency map while keeping the details of object boundaries, a morphological smoothing step, composed of a reconstruction-by-dilation operation followed by

Algorithm 4.3: NSCT-based Salient Object Detection Using Local and Global Features

Input: RGB Image ***I***

1 Convert the input image , ***I***, from RGB color space $c \in \{R, G, B\}$ to CIELAB color space $c \in \{L, a, b\}$
2 Obtain local saleincy map using Algorithm 4.1
3 Obtain global saliency map using Algorithm 4.2
4 Fuse the local and global saliency maps using (4.8)
5 Perform abstraction on the fused map using (4.9)
6 Obtain boundary contrast map by computing each pixel's distance to the mean color and covariance matrix of the boundary pixels
7 Obtain the saliency map by pixel-wise adding the abstracted saliency map obtained in Step 5 and the boundary contrast map obtained in Step 6

Output: Slainecy Map ***S***

a reconstruction-by-erosion [84] is employed. The contrast between the salient and non-salient regions is then enhanced using a sigmoid function.

Figure 4.3 (e) shows the saliency map after the abstraction for the sample image in Figure 4.1. It is seen from this figure that the salient region is detected more precisely and more uniformly after the abstraction.

The proposed NSCT-based salient object detection scheme is summerized in Algorithm 4.3 .

4.3 Experimental Results

In this section, performance of the proposed salient object detection method using the NSCT-based features is compared with that of the some of the more recent salient object detection schemes.

In the proposed scheme, the image is decomposed into 4 scales and 4 directional subbands in each scale by using NSCT. It should be noted that the larger the number of directional subbands used, the greater is the number of features available. However, after

a certain number of decomposition levels, the improvement in the performance becomes insignificant. Hence, a further increase in the number of the decompositions used only adds to the computational complexity. In the proposed scheme, both the number of decomposition levels and the number of directional subbands are empirically set to 4 in order to provide the best performance. The block size in local feature extraction, b, is set to 7, and the number of superpixels in the saliency map abstarction, k, is set to 200.

4.3.1 Performance Comparison

Performance of the proposed method is compared to six more recent schemes for salient object detection, namely, saliency filters method (SF) [29], manifold ranking method (MR) [31], saliency optimization method (SO) [32], region-based contrast method (RC) [34], minimum barrier distance method (MBD+) [33], and minimum spanning tree method (MST) [35].

The saliency maps obtained using the proposed method as well as that of the other methods for some test images are shown in Figure 4.4. It can be seen from this figure that in general the saliency maps obtained by the proposed method is more similar to the ground truth as compared to that provided by the other methods. It is seen that the SF scheme can not detect the entire salient object (Figure 4.4 (c)) and the MR method can not detect the salient object uniformly (rows 2, 4, and 6 of Figure 4.4 (d)). In the saliency maps obtained by the SO and RC schemes some non-salient regions are incorrectly detected as salient (row 10 of Figure 4.4 (e) and row 6 of Figure 4.4 (f)). The MBD+ and MST schemes provide more accurate saliency maps but they still fail to detect the salient objects in some images such as the image in the row 4 of Figures 4.4 (g)-(h).

Figures 4.5-4.7 depict the average precision-recall curves obtained by applying the proposed salient object detection method and the other methods to the images in the six datasets when the threshold is fixed. It is seen from this figure that the proposed method

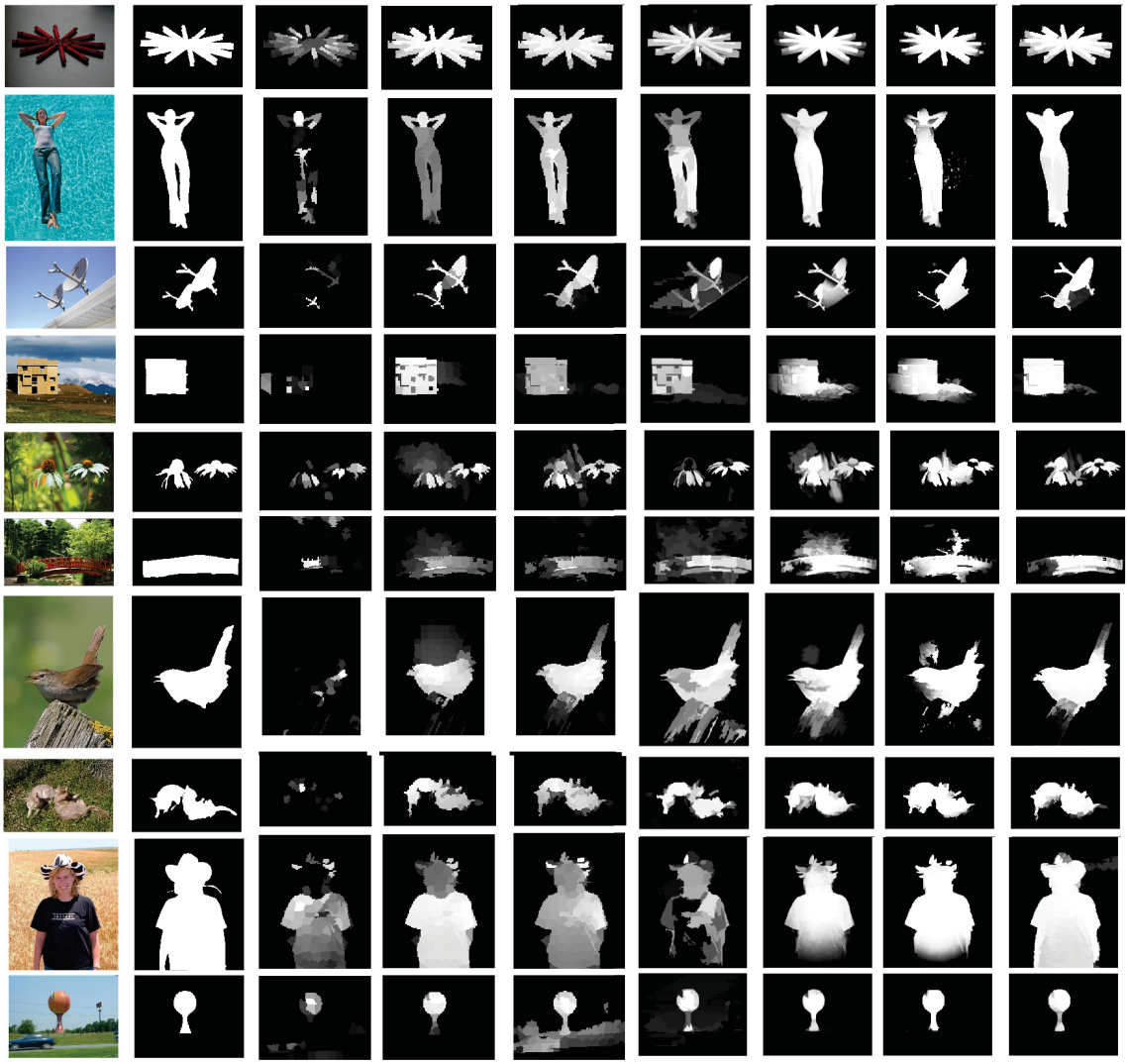


Figure 4.4: Saliency maps obtained by applying the proposed method and the other methods on test images from the datasets in [27, 31, 34, 44, 75]. (a) Original image. (b) Ground truth. (c) SF [29]. (d) MR [31]. (e) SO [32]. (f) RC [34]. (g) MBD+ [33]. (h) MST [35]. (i) Proposed method [81].

generally outperforms all the other methods in terms of precision-recall performance when applied on the images in all but the PASCAL-S dataset, where the MBD+ method provides the best performance and MST, SO, and MR schemes each provides almost the same performance as that of the proposed method.

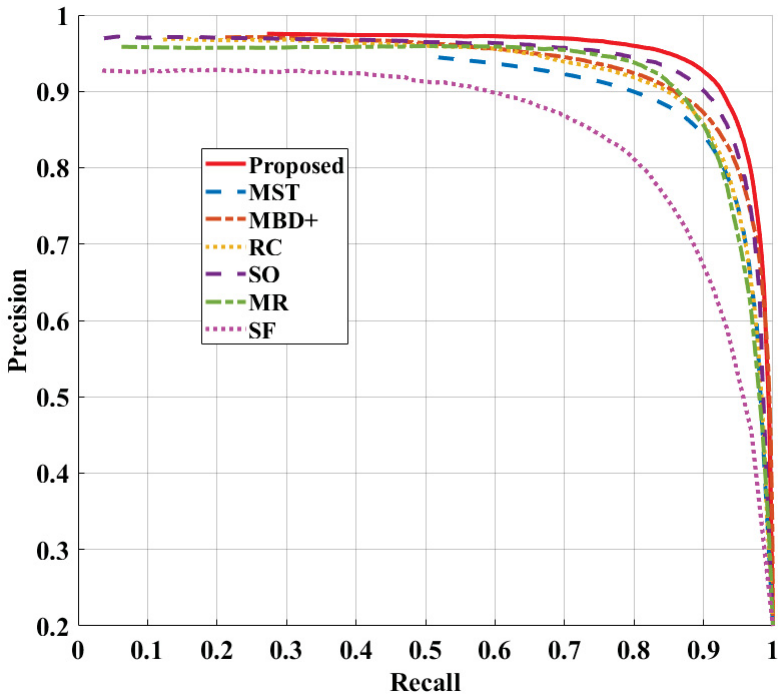
The average precision, recall and F_β values obtained using the proposed method with the adaptive threshold as well as that of the other methods are calculated and shown in Figures 4.8 and 4.9. It is seen from these figures that the proposed method provides the largest values of precision and F_β for all the datasets except for the PASCAL-S, where its F_β value is lower by 1.75% in comparison to that of the best value provided by the MBD+ scheme. The recall values provided by the proposed method are the largest for the MSRA-1000 and MSRA-10K datasets and competitive to the largest values for the other datasets.

The F_β and MAE values obtained using the proposed method as well as that of the other methods are also presented in Table 4.1. It is seen from this table that the proposed method yields the smallest MAE values when applied to images in all but the PASCAL-S dataset. For images in the PASCAL-S dataset, MBD+ is the leading method while the proposed method provides MAE value that is 1.63% larger than that provided by the MBD+ scheme.

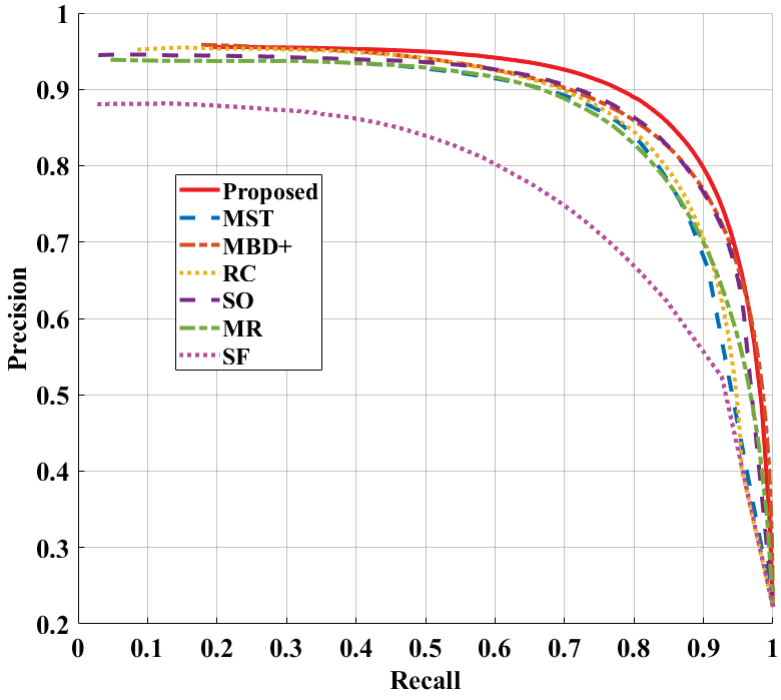
Overall the proposed scheme provides the best performance for most of the datasets on which the various schemes have been experimented.

4.4 Summary

In this chapter, a salient object detection method has been proposed using multi-scale and directional selectivity properties of the NSCT. The local features have been extracted from the local variations of the low-pass coefficients whereas the global features have been obtained based on the distribution of the directional subband coefficients. The final saliency map has been obtained by combining the local and global features and shown to efficiently

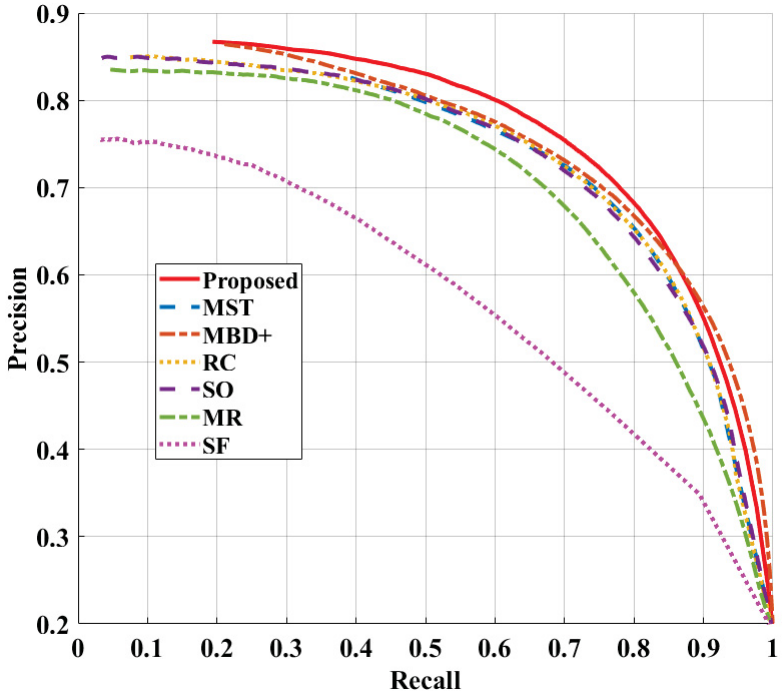


(a)

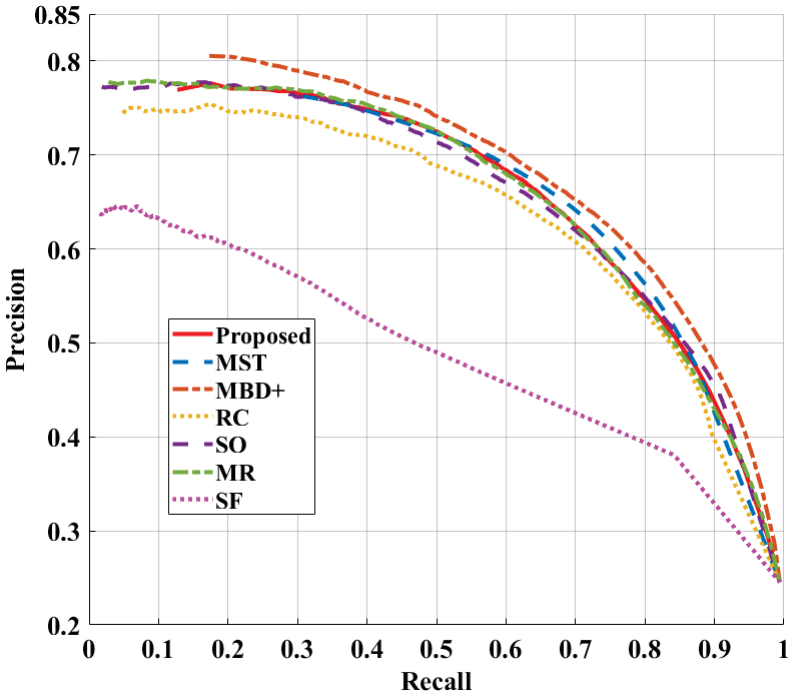


(b)

Figure 4.5: Precision-recall curves obtained by applying the proposed and other salient object detection methods on images in (a) MSRA-1000 and (b) MSRA-10K datasets.

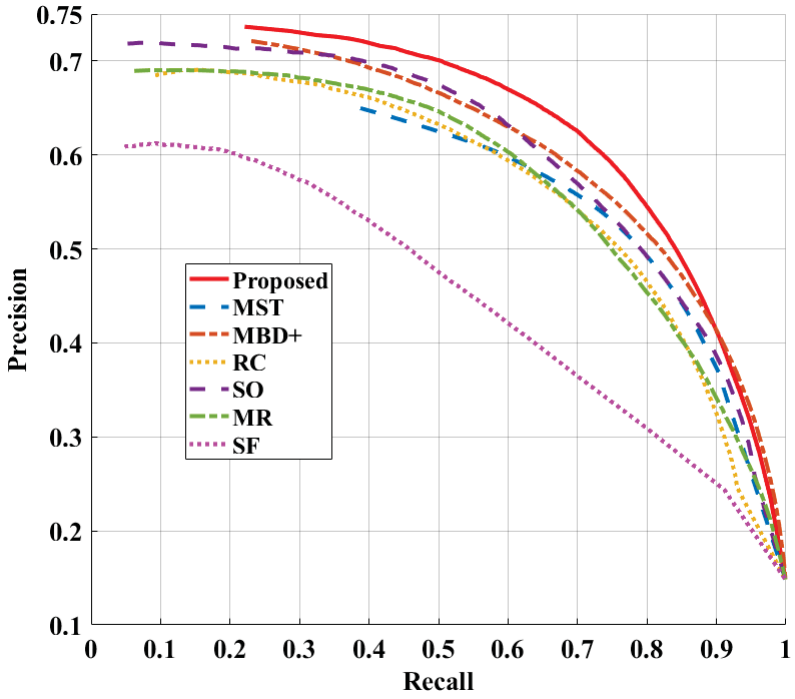


(a)

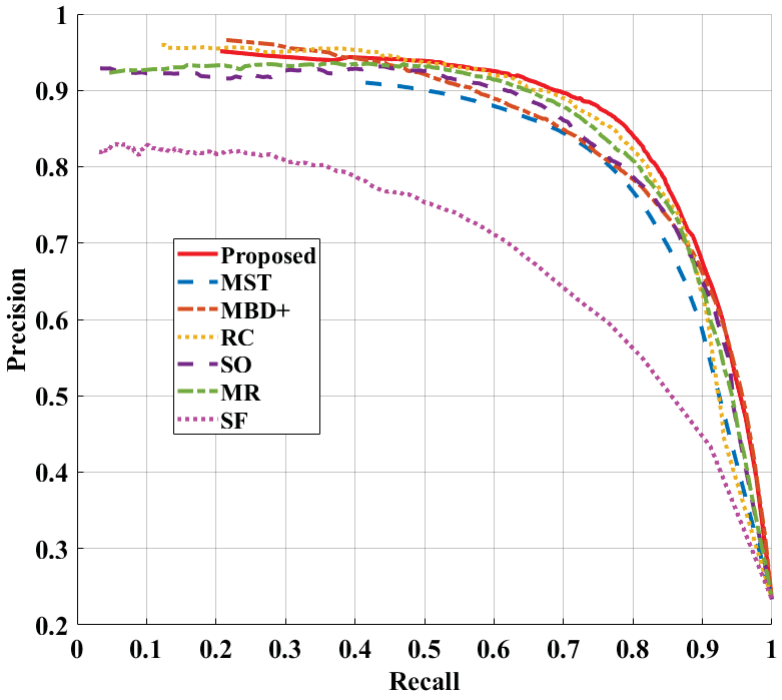


(b)

Figure 4.6: Precision-recall curves obtained by applying the proposed and other salient object detection methods on images in (a) HKU-IS and (b) PASCAL-S datasets.

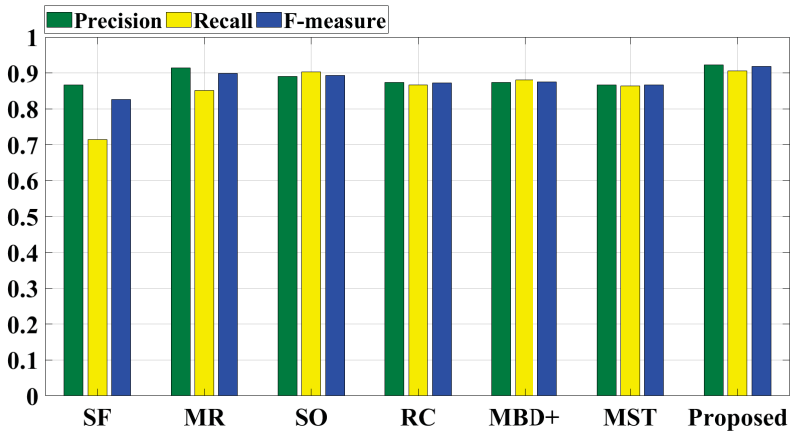


(a)

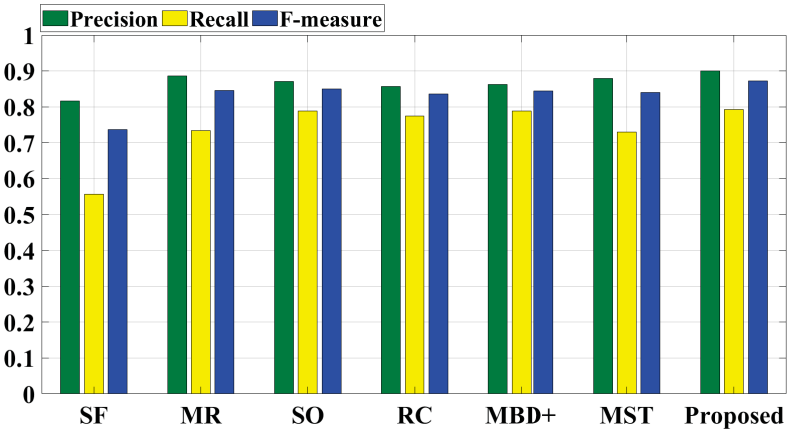


(b)

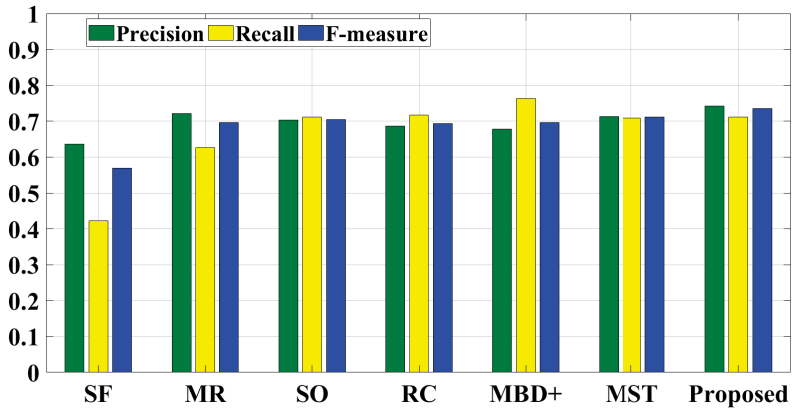
Figure 4.7: Precision-recall curves obtained by applying the proposed and other salient object detection methods on images in (a) DUT-OMRON and (b) CSSD datasets.



(a)

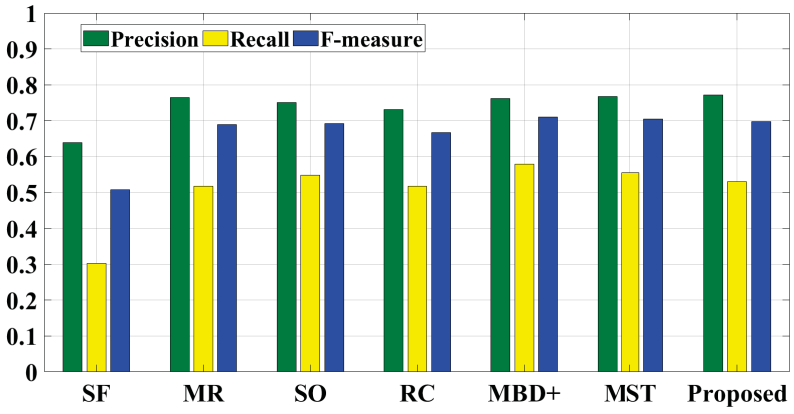


(b)

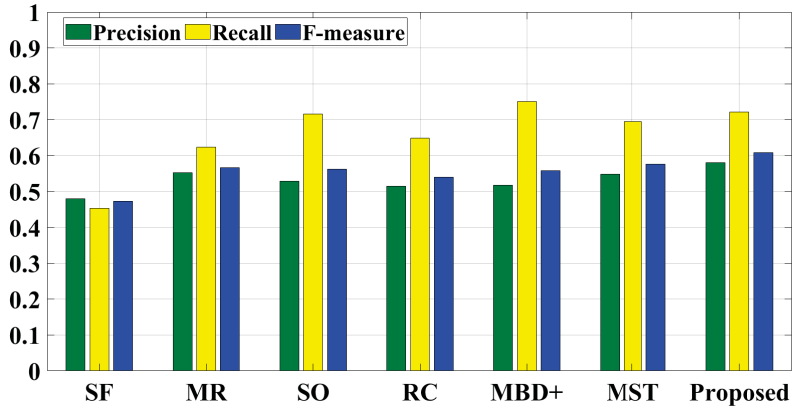


(c)

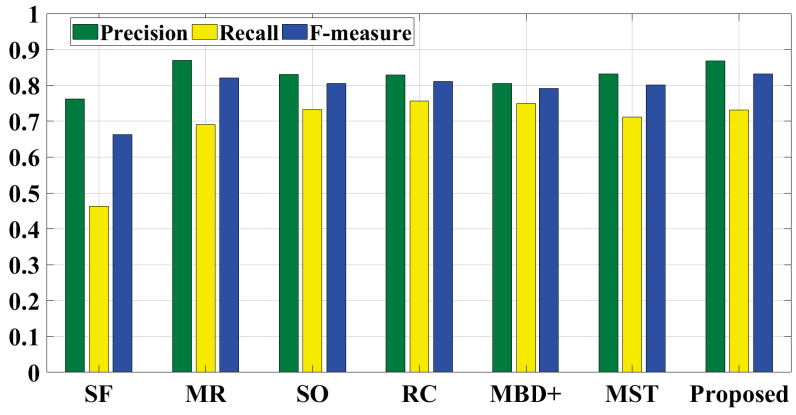
Figure 4.8: Precision, recall, and F_β obtained by applying the proposed and other salient object detection methods on images in (a) MSRA-1000, (b) MSRA-10K, and (c) HKU-IS datasets.



(a)



(b)



(c)

Figure 4.9: Precision, recall, and F_β obtained by applying the proposed and other salient object detection methods on images in (a) PASCAL-S, (b) DUT-OMRON, and (c) CSSD datasets.

Table 4.1: F_β and MAE values obtained by applying the proposed and the other methods on images in the six datasets

Dataset	**MSRA-1000**		**MSRA10K**		**HKU-IS**	
Metric	F_β	**MAE**	F_β	**MAE**	F_β	**MAE**
Proposed [81]	**0.9182**	**0.0332**	**0.8729**	**0.0718**	**0.7354**	**0.1117**
MST [35]	0.8664	0.0562	0.8400	0.0893	0.7117	0.1216
MBD+ [33]	0.8750	0.0517	0.8446	0.0828	0.6957	0.1235
RC [34]	0.8723	0.0517	0.8365	0.0845	0.6933	0.1233
SO [32]	0.8930	0.0415	0.8505	0.0775	0.7046	0.1168
MR [31]	0.8990	0.0450	0.8454	0.0835	0.6967	0.1238
SF [29]	0.8260	0.0822	0.7365	0.1322	0.5697	0.1618
Dataset	**PASCAL-S**		**DUT-OMRON**		**CSSD**	
Metric	F_β	**MAE**	F_β	**MAE**	F_β	**MAE**
Proposed [81]	0.6976	0.1872	**0.6082**	**0.1218**	**0.8321**	**0.1053**
MST [35]	0.7053	0.1858	0.5759	0.1435	0.8005	0.1174
MBD+ [33]	**0.7100**	**0.1842**	0.5580	0.1449	0.7911	0.1161
RC [34]	0.6674	0.2000	0.5404	0.1463	0.8114	0.1090
SO [32]	0.6920	0.1895	0.5619	0.1345	0.8054	0.1126
MR [31]	0.6886	0.1868	0.5670	0.1314	0.8199	0.1116
SF [29]	0.5080	0.2442	0.4733	0.1537	0.6631	0.1775

represent the pixels that are locally and globally distinctive. In addition, the structure and boundary of the image objects have been preserved by abstracting the saliency maps into meaningful image regions.

Experimental results have shown that the proposed method outperforms a number of more recent schemes in terms of precision, recall, F_β and MAE metrics.

Chapter 5

Salient Object Detection Using Deep Convolutional Features

5.1 Introduction

Recently, some salient object detection schemes have been proposed by taking advantage of the CNNs. In these schemes, a salient object detection model is automatically learned based on low level and high level image features extracted by the different layers of a deep CNN. However, the existing deep salient object detection schemes suffer from high computational cost. In view of this, in this chapter a low complexity deep salient object detection network is proposed by making use of the depthwise separable convolution [85]. The chapter is organized as follows. In Section 5.2 a brief literate review of the existing deep salient object detection networks is presented. In Section 5.3, the proposed deep salient object detection method is developed by devising the various steps involved in designing the architecture of the network. The network training process and the hardware and software platforms for its implementation are also described in this section. In Section 5.4, several experiments are conducted in order to study the impact of using the depthwise separable convolution and that of the skip connections on the performance of the proposed network. In this section, the performance of the proposed scheme is also compared with that of some of the state-of-the-art schemes. Finally, in Section 5.5, the work of this chapter

is summarized and the significant features of the proposed method are highlighted.

5.2 A brief Literature Review of Salient Object Detection Schemes Using Deep Convolutional Neural Networks

The initial deep salient object detection schemes [41–45] have employed the same networks as the ones originally proposed for the image classification problem, in which a fully connected dense layer was used as the last layer of the network to produce a class score value for the classification of the input image. Such a network when used for salient object detection, this last fully connected layer has then been used to provide a single saliency value for each image segment (e.g., patch, superpixel, or object proposal). As such, in these methods [41–45], the image is first partitioned into segments and the training/prediction process is repeated for each image segment, which is a computationally expensive overhead in these methods. Also, the performance of the network is affected by the accuracy of the segmentation algorithm and its parameters such as the number of segments. Moreover, due to the segment-level prediction, the saliency maps obtained by these methods are spatially non-uniform and contain undesired abrupt changes around the boundary of the segments.

More recent CNN-based approaches for salient object detection have developed schemes that generate a saliency map for the entire image rather than for individual image segments [46–51]. In these methods, an end-to-end mapping is employed for which a fully convolutional network (FCN) [64] is well-suited. Thus, in these schemes, the output layer that makes predictions of the pixel saliency values of the input image is also a convolutional layer. Even though in these schemes, the saliency maps are obtained for the entire image, they still make use of an over-segmented version of the input image. The architectures used in these schemes have two parts to carry out the downsampling and upsampling operations,

respectively. In [46–48], the first part of the network is a FCN that obtains a saliency map at a reduced resolution level. The second part of the network, using this low-dimension saliency map, obtains a saliency map with a resolution which is the same as that of the input image. Also, it is in this part that the over-segmented version of the input image is used to refine the saliency map further. In the scheme of [49], the operations of both down-sampling and upsampling are done using a FCN, but both the input image and a saliency prior map of its over-segmented version are fed to the first part of the network from the very beginning using, respectively, two different convolutional filters. A combination of the outputs from the two filters then flows through the rest of the layers of the first part as well as through all the layers of the second part of the network to produce the final saliency map. It is to be noted that the schemes of [46–49] still suffer from the overhead of additional computational complexity, as the schemes of [41–45] do, due to the segmentation operation. It is also to be noted that in the schemes of [46–49], the effect of the gradient vanishing problem associated with deep networks is only partially compensated in view of making use of an over-segmented version of the input image.

It is known that skip connections are effective in handling the gradient vanishing problem, which is a very important issue during the training process of deep networks. These connections facilitate an efficient flow of information through the network, which results in a better training and an improved performance of the network. In a deep salient object detection network, the shallower layers are capable of extracting low level appearance features of the image, whereas the deeper layers extract high level semantic features. Fusing the features from different levels using skip connections has been used to improve the performance of deep salient object detection methods [50, 51]. These connections feed the features extracted by the layers at different levels of a deep network in order to prevent loss of information during sub sampling and to provide the receiving convolutional layers with lower level features that otherwise are not directly available to them. However, adding skip

connections increases the complexity of the network. The above methods employ these connections without due regard to the increase in the complexity of the network relative to the resulting improvement in the performance. In general, having more features available by employing skip connections seems to be useful in improving the performance but they may not necessarily improve the performance in a salient object detection problem in the same proportion as the increase in the network complexity. Therefore, a judicious choice of the pairs of layers to be involved in the skip connections could be useful in determining the trade off between the performance and complexity of a network used for salient object detection.

The existing deep salient object detection schemes use the standard convolution, which is a convolution over all the channels in one step. However, performing the standard convolution in a deep network is computationally expensive specially when the number of channels is increased as more number of filters are used in a layer of the network. This problem gets even worse if the two sets of channels are concatenated through a skip connection. A factorized form of convolution called depthwise separable convolution [86] has been recently shown to be effective in designing deep networks for different applications [87, 88]. A depthwise separable convolution consists of a standard convolution performed independently over each input channel followed by a 1×1 convolution. This type of convolution has been shown to provide almost the same performance as that provided by the standard convolution for image classification, object detection, and semantic segmentation applications [88] with a significantly lower computational complexity. Thus, utilizing this convolution could be beneficial in reducing the number of parameters and the computational complexity in a deep salient object detection network, thereby making the use of the skip connections affordable. In addition, performing the depthwise separable convolution in two separate steps, the depthwise and pointwise convolutions, provides an opportunity for inserting a non-linearity in a way that allows some of the negative-valued

features to go through the network.

5.3 Proposed Deep Salient Object Detection Network Using Depthwise Separable Convolution

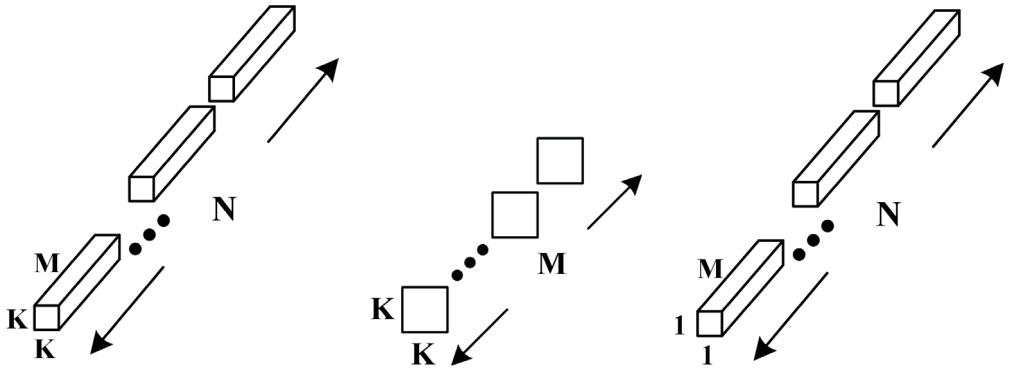
In order to address the limitations of the existing deep salient object detection schemes, in this section, using the lightweight depthwise separable convolution, an end-to-end deep salient object detection network is developed by exploiting the fusion of multi-level and multi-scale image features through judicious skip connections between the layers. The proposed deep salient object detection network is aimed at providing good performance with a much reduced complexity. The proposed salient object detection network follows the framework of a general encoder-decoder architecture, in which both the encoder and decoder parts are designed based on the depthwise separable convolutional blocks. In the encoder part, an efficient lightweight network [88], including strided convolutional layers with short skip connections is adopted to extract image features of different levels. In the decoder part, a number of transposed convolutional layers are used to extract features at levels that are even higher than those extracted by the encoder. At the same time, these decoder layers progressively increase the resolutions of the extracted features to that of the input image. A long skip connection is made between a layer in the encoder part and the layer having the same resolution in the decoder part of the network. Such a skip connection compensates the effect of the sparsity in the map of the decoder layer, which is created due to the upsampling operation, by combining such a layer with the corresponding map in the encoder part, which obviously does not have the sparsity problem of the same intensity. Since the proposed network employs the depthwise separable convolution, the network complexity is reduced to an extent that makes it possible to raise the network complexity slightly up for allowing the skip connections to be made between an encoder layer and

a decoder layer in order to improve the performance of the method. Moreover, the use of the depthwise separable convolution allows the placement of the non-linear activation unit immediately after the first part of the operation rather than following the complete convolution operation, as is done in the case of the standard convolution. This allows the passage of at least some of the negative-valued features through the network, which is not possible with use of the standard convolution.

In the following, first, the depthwise separable convolution is briefly discussed as the backbone of the proposed method. Then, the architecture of the proposed deep salient object detection network comprising its encoder-decoder structure and skip connections is presented. Afterward, the training and implementation details of the network are explained.

5.3.1 Depthwise Separable Convolution

The basic idea of the depthwise separable convolution is to split the standard convolution, which operates on all the channels in one step, into two convolutions, namely depthwise convolution and pointwise convolution. The depthwise separable convolution is capable of reducing the computational complexity of a deep network over that of using the standard convolution. Assume that a convoltional layer takes an input map of size $L \times W \times M$ and generates an output feature map of size $L \times W \times N$ (assuming a unity stride, $s = 1$, and padding of the border pixels). Thus, the number of channels changes from M to N between the input and the output of the convolutional layer. If the standard convolution is performed in this layer, N convolutional kernels each of size $K \times K \times M$ are applied to the input map to generate the output map. Therefore, the computational cost of a single layer using the standard convolution is $K^2 \times M \times N \times L \times W$. If the convolutional layer performs a depthwise separable convolution, then in its first part (the depthwise convolution), M convolutional kernels each of size $K \times K$ are applied to the M input channels individually, and the M outputs each of size $L \times W$ are generated. The



(a) Standard Convolution (b) Depthwise Convolution (c) Pointwise Convolution

Figure 5.1: Kernel structures involved in implementing (a) the standard, and (b-c) depthwise separable convolutions.

computational cost of this part would be $K^2 \times M \times L \times W$. In the second convolution (the pointwise convolution), the outputs of the depthwise convolution are linearly combined and a new channel space is generated. More specifically, N convolutional kernels each of size $1 \times 1 \times M$ are applied to the outputs of the depthwise convolution and an output map of size $L \times W \times N$ is generated. The computational cost of the pointwise convolution is $N \times M \times L \times W$. Therefore, the computational cost of a single layer using the depthwise separable convolution is $M \times L \times W \times (K^2 + N)$. Thus, by splitting the convolution into two steps of channel-wise filtering and linear combining, the computational cost of a single layer is reduced by the factor of $K^2N/(K^2 + N)$ over that of the standard convolution. Figure 5.1 shows the kernel structures involved in implementing the standard convolution and the depthwise separable convolution.

5.3.2 Encoder-Decoder Structure of the Proposed Network

The proposed network consists of an encoder part, a decoder part, and a number of skip connections each between a layer in the encoder and the layer having the same resolution in the decoder. Figure 5.2 shows the general architecture of the proposed network.

Recently, a network called MobileNetV2 [88] has been proposed and its effectiveness has been evaluated in the applications of image classification, object detection, and semantic segmentation. In view of its very lightweight character resulting from the use of the depthwise separable convolution, we adopt a part of this network as the encoder part of our proposed network for salient object detection.

As seen from Figure 5.2, the encoder part consists of 8 blocks, Eblocks 1 to Eblock 8. Moving from Eblock 1 to Eblock 8, the spatial resolution of the feature maps is decreased by a factor of 5. Eblock 1 performs a standard convolution with a kernel of size 3×3, while all the other blocks of the encoder consist of one, two, three, or four bottleneck units. The encoder employs two types of bottlenecks, referred to as Bottleneck 1 and Bottleneck 2, shown as in Figures 5.3 (a) and (b), respectively. Either of these types of bottleneck units consists of a 1×1 convolution with a non-linear activation functions (ReLU) to increase the number of input channels, a depthwise convolution with a 3×3 kernel and a non-linear activation function, and a pointwise convolution that reduces the number of channels back to that of the input to the bottleneck. The difference between the two types of bottlenecks is that Bottleneck 1 uses a residual connection between the output of the pointwise convolution layer and the input to the unit, which cannot be done in Bottleneck 2 because of spatial incompatibility resulting from the convolution operation with stride 2, $s = 2$, or from changing the number of channels.

The output of the encoder, that is, that of Eblock 8, which is a feature map of size $7 \times 7 \times 320$, is first filtered using a convolutional layer performing a depthwise separable convolution. The output resulting from this convolution is then used as the input to the decoder part of the proposed salient object detection network.

The decoder part of the proposed salient object detection network consists of 5 decoder blocks, Dblock 1 to Dblock 5, in which image features of levels that are higher than those that are yielded by the encoder are extracted while the spatial resolution of the feature maps

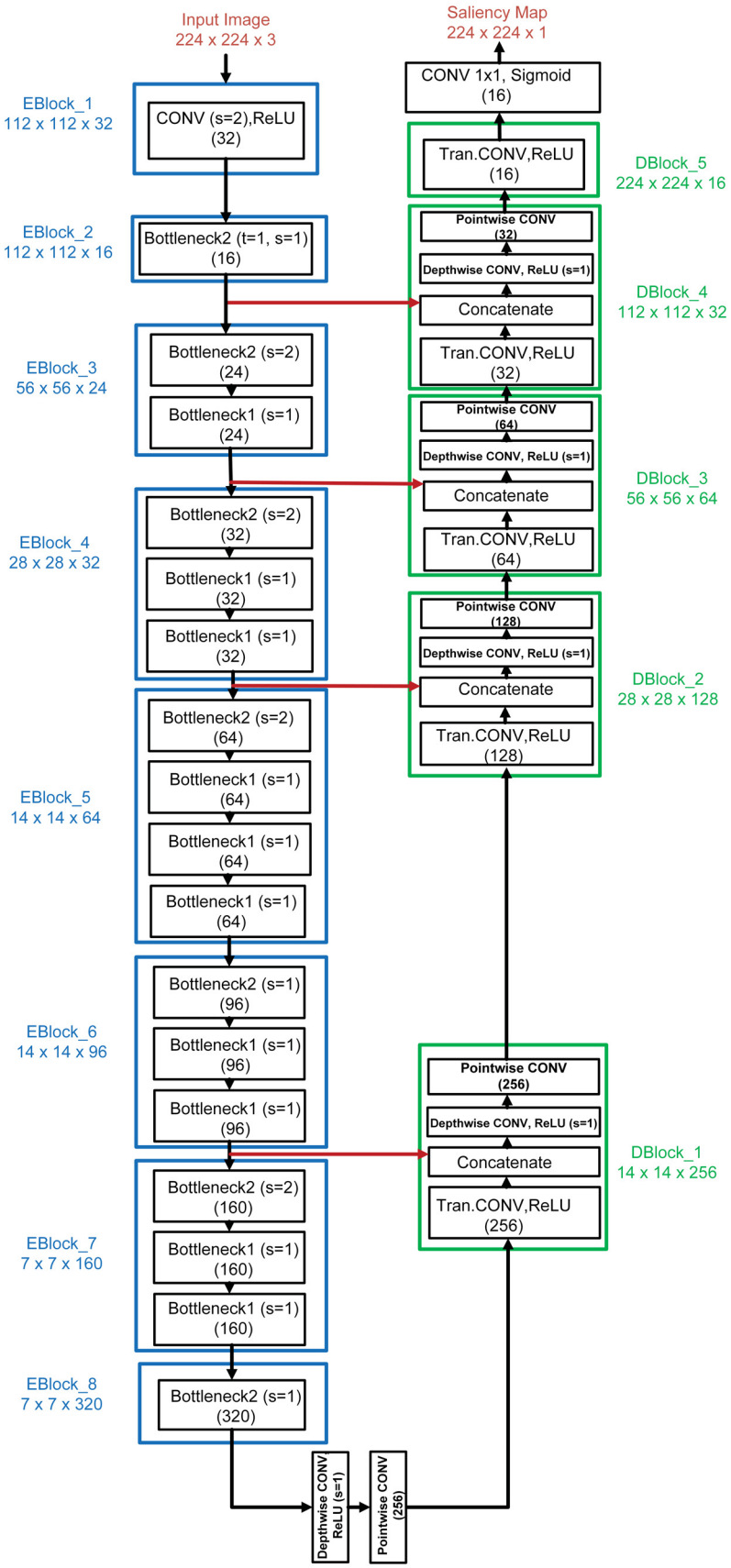


Figure 5.2: Architecture of the proposed low complexity network for salient object detection (Size of a block's output is indicated next to it.).

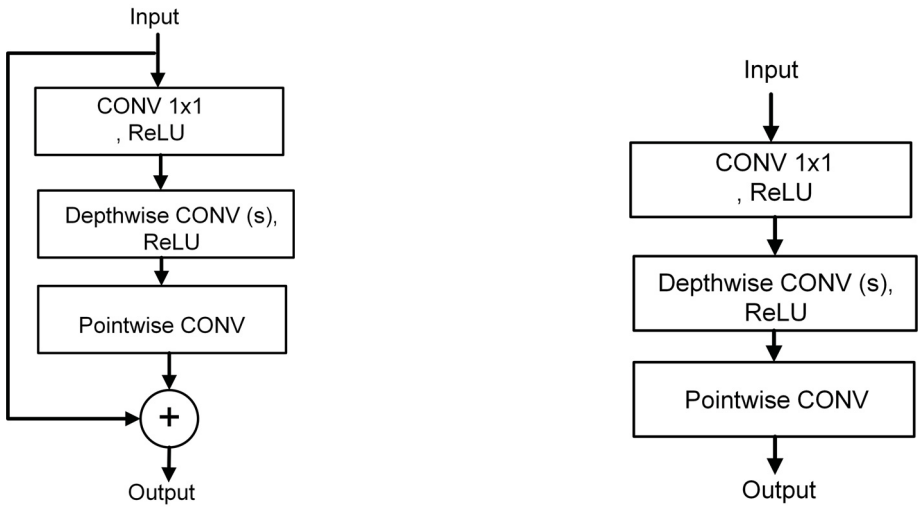


Figure 5.3: (a) Bottleneck 1, and (b) Bottleneck 2 units [88] for the architecture in Figure 5.2.

is progressively increased.

Each of the first four decoder blocks, Dblocks 1 to Dblock 4, consists of a transposed convolutional layer with the kernel size of 3×3 and a stride value of 2, $s = 2$, a concatenation operation, and a depthwise separable convolutional layer consisting of a depthwise convolution with a stride value of 1 , $s = 1$, and a pointwise convolution. In each decoder block, first, the spatial resolution of the feature maps is increased by a factor of 2 through the transposed convolution. The upsampled feature maps are then concatenated with the feature maps of the same spatial resolution from the encoder. Finally, features of higher levels compared to those extracted by the encoder layers are extracted using the depthwise separable convolution. The last decoder block, Dblock 5, consists of only a transposed convolutional layer with the kernel size of 3×3 and a stride value of 2, $s = 2$.

It is known that when a standard convolution is performed in layers of a deep network, ReLUs are introduced following the convolution operation for injecting non-linearity in the network in order to improve its end-to-end mapping capability. This, however, is achieved

at the expense of loosing information contained in the negative features discarded through the non-linear operation of ReLU. Use of the depthwise separable convolution provides an opportunity to rectify this problem partially. In this case, the ReLU unit is placed in between the depthwise and pointwise convolutions rather than introducing it following the complete depthwise separable convolution operation.

The last part of the network is a 1×1 convolutional layer that reduces the dimensionality of the set of feature maps obtained by Dblock 5 and generates a single saliency map. A sigmoid activation function is employed in this layer.

5.3.3 Skip Connections between the Encoder and Decoder Layers

Use of the skip connections in deep networks improve their performance by providing them with a capability to curtail the gradient vanishing problem. Our objective is to do this in such a way as not to adversely affect the computational complexity of the proposed network. As seen in the previous section, the proposed network already has short residual skip connections in the Bottleneck 1 units of the encoder part of the network. However, having additional skip connections between a layer from the encoder and that from the decoder should further improve the network's capability in handling the gradient vanishing problem.

The upsampling operation in the layers of the decoder part of the network causes increasingly more sparsity in the feature maps as we go through the deeper layers of the decoder. On the other hand, the feature maps in the layers of the encoder are not sparse. Therefore, a skip connection between a layer from the encoder part to a layer in the decoder part should remedy this problem of sparsity in the feature maps of the decoder layers through the concatenation of the sets of channels of the two layers involved in a skip connection. However, each skip connection would increase the depthwise dimensionality (i.e., the number of channels) of the decoder layer involved in the connection, and hence,

inevitably the complexity of the convolution operation of that layer. Consequently, we should next consider as to which pairs should be employed for the skip connections so as to curtail unnecessary growth in the complexity resulting from the increase in the number of channels in the decoder layer.

If the spatial dimensions of the features maps of the two layers involved in a skip connection are not matched, a dimensionality adjustment is needed before carrying out the concatenation operation. This dimensionality adjustment adds to the computational complexity of the method. In addition, if the spatial dimension of the two sets of feature maps from the layers participating in a concatenation are made compatible by resizing the spatial dimension of the feature maps with lower dimensionality, such maps will become even more sparse, and hence, concatenation would not be helpful in reducing the sparsity of the maps in the decoder layers. Therefore, in the proposed network, the best solution is achieved by having a skip connection between a layer in the decoder and a layer from the encoder having the same spatial dimension. By doing so, the sparsity problem in the decoder feature maps are reduced while avoiding the computational cost involved in adjusting the dimensionality of the concatenating layers.

It should be noted that the convolutional layer of the last decoder block, namely Dblock 5, is not made to participate in a skip connection. The reason for this is the fact that such a skip connection will hinder with the purpose of this block, which is to extract high level semantic features.

Impact of employing the skip connections on the performance of the network is studied in more details in Section 5.4.

5.3.4 Loss Function

The proposed network is trained using a binary cross entropy loss function. For a predicted saliency map, $\boldsymbol{S}$, and its corresponding pixel-level annotated ground truth, $\boldsymbol{GT}$,

the binary cross entropy loss function, L_{BCE}, is calculated as

$$L_{BCE}(\boldsymbol{S}, \boldsymbol{GT}) = \sum_{x}\sum_{y} \left[GT(x,y)\ log[S(x,y)]\ +\ [1 - GT(x,y)]\ log[1 - S(x,y)]\right], \tag{5.1}$$

where (x, y) represents a pixel's spatial location.

5.3.5 Implementation Details

The proposed network is implemented in Python on the publicly available Keras library [89] with Tensorflow backend [90]. The network is trained using the Adam optimizer [91]. The learning rate is set to 10^{-4} and is reduced by a factor of 10 if no improvement is observed in the validation performance for 5 consecutive epochs. Dropout [92] and batch normalization [93] are used during the training process. The weights of encoder layers are initialized with those of the MobileNetV2 network [88] that was trained using the images in ImageNet dataset [59], while the weights of all the other layers are randomly initialized.

The MSRA-B dataset [28] containing $5,000$ images and their corresponding pixel-level ground truths is used in the training phase, where 80% of the images are used for training and 20% for validation. Data augmentation techniques has been shown to be effective in training deep networks. In this work, we apply random shift-scale-rotate and horizontal flip to all the training images, thus effectively utilizing resulting $12,000$ images for the training of the network. All the images are resized to 224×224 pixels. The network is trained for 100 epochs with an early stopping method, while a mini batch of 10 images are utilized in each training iteration.

The network is trained on a machine equipped with an i7-8750H 2.21 GHz CPU and an NVIDIA GeForce GTX 1060 GPU.

5.4 Experimental Results

In order to evaluate the performance of the proposed method, several experiments are conducted on images in four challenging and widely used datasets, namely, HKU-IS [44], PASCAL-S [75], DUT-OMRON [31], and CSSD dataset [76]. It should be mentioned that since some images are in common between the MSRA-B dataset, which is used for training, and the two datasets of MSRA-1000 and MSRA-10K, the proposed network is not evaluated on images from these two datasets.

5.4.1 Impact of Employing Skip Connections between the Encoder and Decoder Layers on the Performance of the Proposed Network

In this section, the effect of employing the skip connections between the encoder and decoder layers on the performance of the proposed salient object network is studied. To this end, the backbone network, i.e., the encoder and decoder parts without any skip connection, is trained and then applied to the images of the four datasets. The F_β and MAE values obtained from the proposed network with and without the skip connection, are given in Table 5.1. Also, in order to study the computational cost, for each version of the network, the number of parameters and the number of multiplication-accumulation (MAC) operations required to generate a saliency map for an image of size $224 \times 224 \times 3$ are calculated and presented in Table 5.2. It is seen from Tables 5.1 and 5.2 that employing the skip connections results in a significantly improved performance of the network. For instance, for images in the HKU-IS dataset, the F_β value is increased by 4.18% and the MAE value is decreased by 20.74% through the use of the skip connections. The performance improvement by employing the skip connections in the proposed scheme is achieved at the expense of a slight increase in the computational cost. Specifically, the number of parameters and

Table 5.1: Comparison of the proposed salient object detection network with and without the skip connections between the encoder and decoder layers when applied to the images of the different datasets

Dataset	**HKU-IS**		**PASCAL-S**	
Performance Metric	F_β	**MAE**	F_β	**MAE**
Backbone Network (without skip connections)	0.8345	0.0569	0.7824	0.1124
Proposed Network (with skip connections)	0.8694	0.0451	0.8050	0.1042
Dataset	**DUT-OMRON**		**CSSD**	
Performance Metric	F_β	**MAE**	F_β	**MAE**
Backbone Network (without skip connections)	0.6994	0.0744	0.8904	0.0636
Proposed Network (with skip connections)	0.7260	0.0681	0.9152	0.0527

Table 5.2: Comparison of the number of parameters and the number of MACs for the proposed salient object detection network with and without the skip connections between the encoder and decoder layers

Method	**No. of Parameters** ($\times 10^6$)	**No. of MACs** ($\times 10^9$)
Backbone Network (without skip connections)	2.97	0.609
Proposed Network (with skip connections)	3.01	0.630

that of the MAC operations are increased only by 1.35% and 3.45%, respectively. The impact of employing the skip connections is also evaluated in terms of the precision-recall curves which are shown in Figures 5.4 and 5.5. It is seen from these figures that employing the skip connections results in a better precision-recall performance for the images of all the four datasets.

5.4.2 Impact of Employing the Depthwise Separable Convolution

In order to investigate the impact of using the depthwise separable convolution on the computational complexity and performance of the proposed network, we implement the same network architecture as that of the proposed network by using the standard convolution operation, and then compare its performance with the proposed network.

It should be pointed out that for the architecture performing the standard convolution, the pre-trained encoder layers are not available for initialization. Therefore, to make a fair comparison, in this section, both versions of the network are trained using random initialization. In order to avoid the overfitting problem in training the networks with random initialization, a larger training dataset is required. To make the training set larger, in this section, the images of the HKU-IS and DUT-OMRON datasets, that were previously used only as test datasets, are also used for training, and the performance of the two architectures are evaluated only on images from the CSSD and PASCAL-S datasets. In addition, training of the two networks are stopped when they reach the same validation performances. This is due to the fact that even if the training continues, the overfitting problem cannot completely avoided and thus, we cannot obtain an appropriately trained network for both cases. This is worse for the network performing the standard convolution, since it has more parameters. The F_β and MAE values for images in the CSSD and PASCAL-S datasets obtained by the two versions of the network are given in Table 5.3. It is seen from this table that for images in the both datasets the proposed network performing the depthwise separable convolution provides better F_β and MAE values compared to the network performing the standard convolution.

Numbers of parameters and MAC operations for such the two networks are calculated and given in Table 5.4. It is seen that using the depthwise separable convolution for the same network results in 75.4% and 70.7% reductions in the numbers of parameters and MAC operations, respectively.

Table 5.3: Comparison of the two networks using the standard and depthwise separable convolutions when applied to the images of the CSSD and PASCAL-S datasets

Dataset	**CSSD**		**PASCAL-S**	
Performance Metric	F_β	**MAE**	F_β	**MAE**
Network with Standard Convolution	0.8830	0.0729	0.7340	0.1415
Proposed Network with Depthwise Separable Convolution	**0.8907**	**0.0617**	**0.7406**	**0.1355**

Table 5.4: Comparison of the number of parameters and the number of MACs between the two networks using the standard and depthwise separable convolutions

Method	**No. of Parameters** ($\times 10^6$)	**No. of MACs** ($\times 10^9$)
Network with Standard Convolution	12.22	2.152
Proposed Network with Depthwise Separable Convolution	3.01	0.630

5.4.3 Performance Comparison of the Proposed Deep Salient Object Detection Scheme with Algorithms 3.1 and 4.3

In this section, performance of the proposed deep salient object detection scheme is compared with that of the WT-based scheme proposed in Algorithm 3.1 Chapter 3, and the NSCT-based scheme proposed in Algorithm 4.3 Chapter 4.

The precision-recall curves obtained by the three proposed methods in this thesis for images of the four datasets are shown in Figures 5.6 and 5.7. Also, the F_β and MAE values obtained by these methods on images of the four datasets are given in Table 5.5. It is seen from Figures 5.6 and 5.7, and Table 5.5 that using the deep convolutional features results in a substantially improved performance compared to the other two schemes. However, it should be noted that the improved performance of the deep method is obtained at the expense of the computational cost required for the training process. Also, it is seen that the scheme using NSCT-based features provides a better performance compared to the WT-based scheme.

Table 5.5: F_β and MAE values obtained by applying the three WT-based, NSCT-based, and deep salient object detection methods proposed in Chapters 3, 4, and 5, respectively, on the images in the four datasets

Dataset	**HKU-IS**		**PASCAL-S**	
Performance Metric	F_β	**MAE**	F_β	**MAE**
Proposed Deep Method	**0.8694**	**0.0451**	**0.8050**	**0.1042**
Proposed NSCT-based Method	0.7354	0.1117	0.6976	0.1872
Proposed WT-based Method	0.6284	0.1458	0.5432	0.2170
Dataset	**DUT-OMRON**		**CSSD**	
Performance Metric	F_β	**MAE**	F_β	**MAE**
Proposed Deep Method	**0.7260**	**0.0681**	**0.9152**	**0.0527**
Proposed NSCT-based Method	0.6082	0.1218	0.8321	0.1053
Proposed WT-based Method	0.5274	0.1362	0.7376	0.1516

5.4.4 Performance Comparison with the State-of-the-art Schemes

In this section, the performance of the proposed method is compared with that of a number of state-of-the-art deep schemes for salient object detection, namely, Aggregating multi-level convolutional features method (AMULET) [51], deep hierarchical saliency method (DHS) [50], deep contrast learning method (DCL) [47], deep saliency method (DS) [46], recurrent fully convolutional network (RFCN) [49], encoded low level distance map method (ELD) [45], and multi-scale deep features method (MDF) [44]. It should be pointed out that, since the images in the DUT-OMRON dataset have been used in training the DHS method [50], it is evaluated on all the images except those from the DUT-OMRON dataset.

The saliency maps obtained using the proposed method as well as those from using the other methods for couple of images from each of the four datasets are shown in Figure 5.8. It is seen from this figure that the saliency maps obtained by the proposed method is more similar to the ground truth compared to those obtained by the other methods. For instance, in the saliency map generated by the proposed method for the image in the first row (Figure 5.8 (j)), the salient object is detected uniformly and with precise edges, whereas

the other methods either cannot detect the entire salient object (Figuers 5.8 (d) and (e)), or incorrectly detect some non-salient regions as the salient region (Figures 5.8 (c)-(d), and (f)-(i)). In the saliency maps obtained by the MDF and ELD methods (Figures 5.8 (c) and (d)), although the salient regions are detected uniformly, many non-salient regions are incorrectly detected as the salient regions as well. This can be due to the employment of image segments rather than image pixels in these two methods. As it is seen from the images in row 4 of Figure 5.8 (f), the DS method fails in differentiating the uniform non-salient region from the salient region. Also, in the saliency maps obtained by this method the boundary between the salient and non-salient region is not sharp (row 1 of Figure 5.8 (f)). As seen from the images in row 7 of Figure 5.8 (g), the DCL method incorrectly detects some of the uniform non-salient regions as salient ones. The AMULET, DHS, and RFCN schemes provide more accurate saliency maps compared to the other four schemes. However, they still fail in some images such as those in rows 5 and 7 of Figures 5.8 (e), (h) and (i).

Figures 5.9 and 5.10 illustrate the precision-recall curves averaged over all the images in each of the datasets obtained by using the various methods. It is seen from these figures that for the images in the HKU-IS, DUT-OMRON and CSSD datasets, the proposed method provides larger precision values for a wide range of recall values as compared to that provide by the other methods, whereas for the PASCAL-S dataset, each of the DS, AMULET, and RFCN schemes provides a slightly better performance.

In order to study the precision-recall performance of the various schemes, the F_β values corresponding to each point on the precision-recall curves shown in Figures 5.9 and 5.10 are calculated for images in each dataset. The maximum value for F_β resulting from the various methods for images in each of the four datasets are listed in Table 5.6. It is seen from this table that the proposed scheme provides the largest maximum value for F_β for all but the PASCAL-S dataset. For images in the PASCAL-S dataset, the DS method provides

Table 5.6: Maximum F_β values when applying the fixed thresholds, obtained by applying the proposed and other methods on images in the four datasets

Dataset	**HKU-IS**	**PASCAL-S**	**DUT-OMRON**	**CSSD**
Performance Metric	Max F_β	Max F_β	Max F_β	Max F_β
Proposed [85]	**0.8904**	0.7920	**0.7527**	**0.9176**
AMULET [51]	0.8845	0.8055	0.7154	0.9121
DHS [50]	0.8756	0.7943	-	0.9027
DCL [47]	0.8582	0.7712	0.6576	0.8938
DS [46]	0.8489	**0.8086**	0.7488	0.9004
RFCN [49]	0.8783	0.8037	0.7034	0.8982
ELD [45]	0.8178	0.7336	0.6778	0.8751
MDF [44]	0.8395	0.7045	0.6433	0.8473

the largest maximum value for F_β and the proposed scheme provides a value for this metric that is 2.1% lower than the largest maximum.

The average precision, recall and F_β values obtained using the proposed and the other methods when applying the adaptive threshold are calculated and shown in Figures 5.11 and 5.12. It is seen from these figures that the proposed scheme provides the largest F_β value for images in the HKU-IS, DUT-OMRON and CSSD datasets and the largest precision value for images in the DUT-OMRON and CSSD datasets.

The F_β and MAE values obtained using the proposed and all the other methods are given in Table 5.7. It is seen from this table that the proposed salient object detection method provides the best values for F_β and MAE when applied to the images in the HKU-IS, DUT-OMRON and CSSD datasets. Only for the images in the PASCAL-S dataset, the RFCN method provides the best F_β and MAE values and the proposed method is the second best providing a value for F_β that is 1.02% lower than that provided by the RFCN and a value for MAE that is 0.3% larger than the best MAE value.

In order to evaluate the computational complexity of the proposed method, number of parameters and that of MAC operations to generate a single saliency map for an image of size $224 \times 224 \times 3$ are computed and given in Table 5.8. In calculating the number of

Table 5.7: F_β and MAE values obtained by applying the proposed method and the other methods on images in the four datasets. The best and second best results are shown in red and blue fonts, respectively

Dataset	**HKU-IS**		**PASCAL-S**		**DUT-OMRON**		**CSSD**	
Metric	F_β	**MAE**	F_β	**MAE**	F_β	**MAE**	F_β	**MAE**
Proposed [85]	**0.8694**	**0.0451**	**0.8050**	**0.1042**	**0.7260**	**0.0681**	**0.9152**	**0.0527**
AMULET [51]	0.8525	**0.0484**	0.7952	0.1046	0.6740	0.0928	0.8983	**0.0590**
DHS [50]	0.8664	0.0496	0.8004	0.1065	–	–	**0.8988**	0.0626
DCL [47]	0.8452	0.0516	0.7729	0.1143	0.6693	0.0922	0.8919	0.0619
DS [46]	0.8040	0.0644	0.7748	0.1165	0.6697	**0.0818**	0.8629	0.0718
RFCN [49]	**0.8695**	0.0531	**0.8133**	**0.1039**	**0.7004**	0.0890	0.8945	0.0715
ELD [45]	0.7882	0.0723	0.7486	0.1249	0.6408	0.0965	0.8558	0.0713
MDF [44]	0.8141	0.0744	0.7615	0.1403	0.6820	0.0895	0.8702	0.0788

parameters and the number of MAC operations for the methods compared, only the network part of the methods are considered. More specifically, the computational complexity involved with the oversegmentation steps in the DCL, DS, RFCN, ELD, and MDF schemes is ignored. It is seen from this table that the proposed method is significantly lower in computational complexity, both in terms of the number of parameters and the number of MAC operations, than the other methods are. More specifically, the proposed network has only 3.01 millions parameters whereas the other methods have the number of parameters ranging from 28.37 to 134.67 millions parameters. Also, the proposed network requires only 0.63 billion MAC operations to generate a saliency map for an image of size $224 \times 224 \times 3$, whereas the other methods need between 7.23 to 123.15 billions operations to generate a saliency map for the same image.

Overall by considering the computational complexity and different performance metrics for the various schemes when applied to the images of the four datasets, the proposed scheme outperforms the other ones, whereas the AMULET and DHS methods stand as the second and third best methods, respectively. It should be noted that, the AMULET and DHS schemes are computationally more expensive compared to the proposed method. In the AMULET method, the number of MAC operations and the number of parameters are,

Table 5.8: Comparison of the Number of parameters and the number of MACs in the proposed and the other schemes

Method	**No. of Parameters** ($\times 10^6$)	**No. of MACs** ($\times 10^9$)
Proposed [85]	3.01	0.63
AMULET [51]	33.16	27.35
DHS [50]	93.9	15.81
DCL [47]	66.25	123.15
DS [46]	134.27	62.67
RFCN [49]	134.67	62.88
ELD [45]	28.37	15.39
MDF [44]	60.97	7.23

respectively, 43.4 and 11.02 times more than that of the proposed method. Compared to the DHS scheme, in the proposed method the number of MAC operations and the number of parameters are reduced by factors of 25.1 and 31.2, respectively.

5.5 Summary

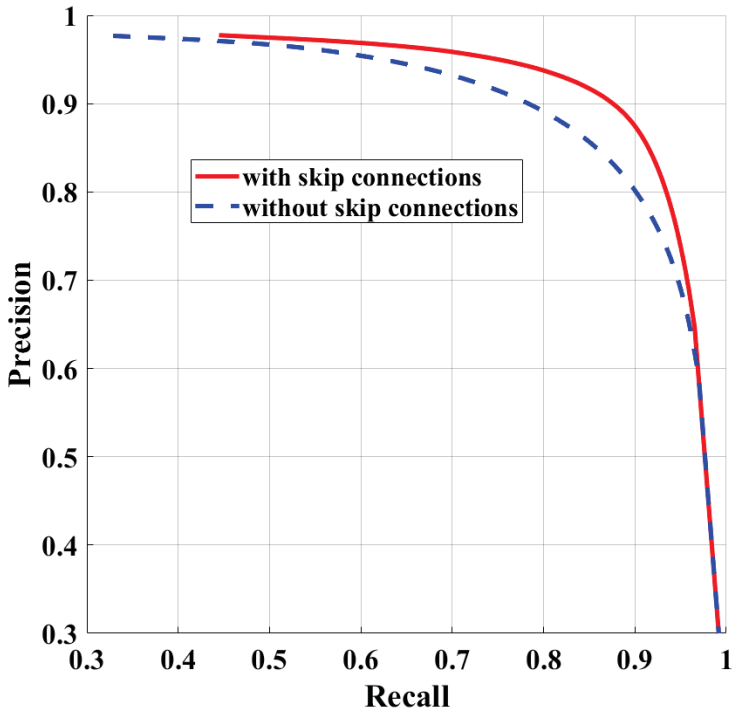
Deep convolutional neural networks have been shown to be effective in extracting multi-level and multi-resolution image features required in salient object detection. Combining the features extracted at different layers of the network by employing skip connections between the layers can improve the performance of the network. However, the existing deep salient object detection schemes suffer from high computational complexity in view of their employing the standard convolution for feature extraction. This problem gets even worse when the number of channels in the network layers is increased by employing the skip connections.

In this chapter a salient object detection scheme has been developed by proposing a deep convolutional neural network in which features are extracted by performing lightweight depthwise separable convolution operations. The proposed network has been designed to have an encoder part and a decoder part. The encoder layers extract features at increasingly

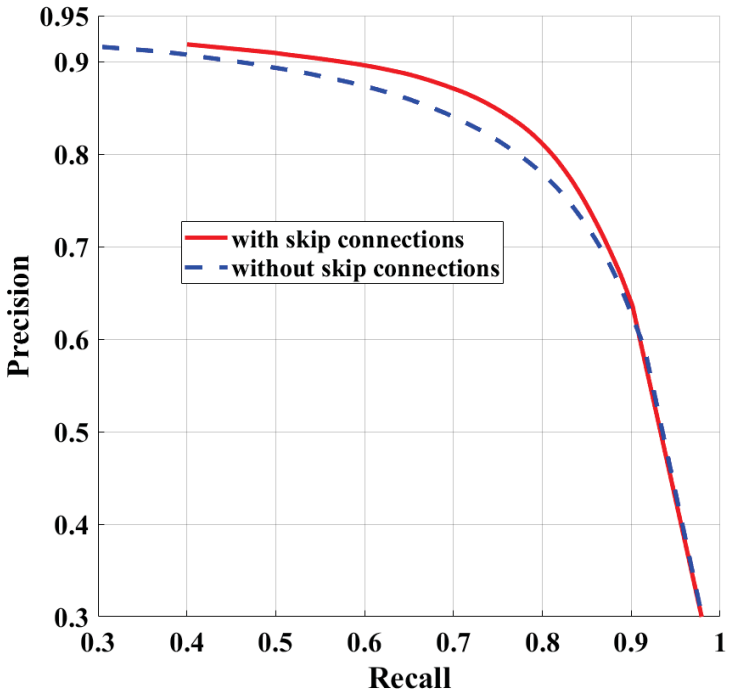
lower resolutions through the strided and normal depthwise separable convolution operations. The decoder layers, on other hand, extract features that are even of higher levels again by employing depthwise separable convolution. The spatial resolutions of the decoder feature maps are progressively increased to finally that of the input image by using the transposed convolutions. In the design of the proposed network a skip connection is established between an encoder layer and a judiciously chosen decoder layer that essentially aims to control the sparsity problem in the feature maps of the decoder resulting from raising the spatial resolution of its feature maps and to curtail the gradient vanishing problem of the network. In order to avoid the computational cost involved with the adjustment of the spatial resolution of the feature maps from the encoder and decoder before the concatenation of the maps, each pair of the layers participating in concatenation are chosen to have the same resolution. However, increasing the number of channels in a concatenating decoder layer affects the network's complexity. In the proposed network, this slight increase in the complexity has been made affordable in view of the drastic reduction of the network complexity by exploiting the depthwise separable convolution operation throughout the network.

The proposed method has been evaluated on images from different datasets, and compared with a number of existing deep salient object detection methods. Experimental results have shown that the performance of the proposed method is generally superior to that of the other methods by providing higher precision, recall and F_β values and lower MAE values, while at the same time having a significantly smaller computational complexity. In comparison to the other existing networks, in the proposed network the number of parameters is reduced by a factor that ranges from 9.43 to 44.7, and the number of MAC operations is reduced by a factor ranging from 11.5 to 195.5. Specifically, in the proposed network the number of parameters and the number of MAC operations are reduced by factors of 11.02 and 43.4, respectively, compared to the second best performing network (the AMULET

method). Finally, it needs to be emphasized that the proposed network has succeeded in outperforming the other state-of-the-art deep salient object detection networks in spite of its much lower complexity in terms of the number of parameters and the number of MAC operations. This superiority in the performance of the proposed network can be attributed to, among the characteristics of the proposed scheme, the use of the depthwise separable convolution that has allowed in some instances the passage of negatively valued features as well in training the network.

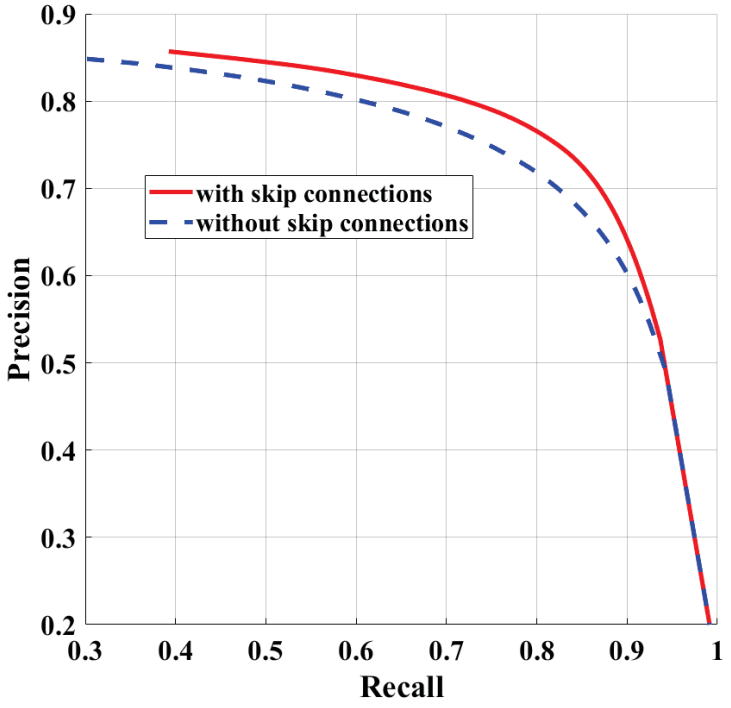


(a)

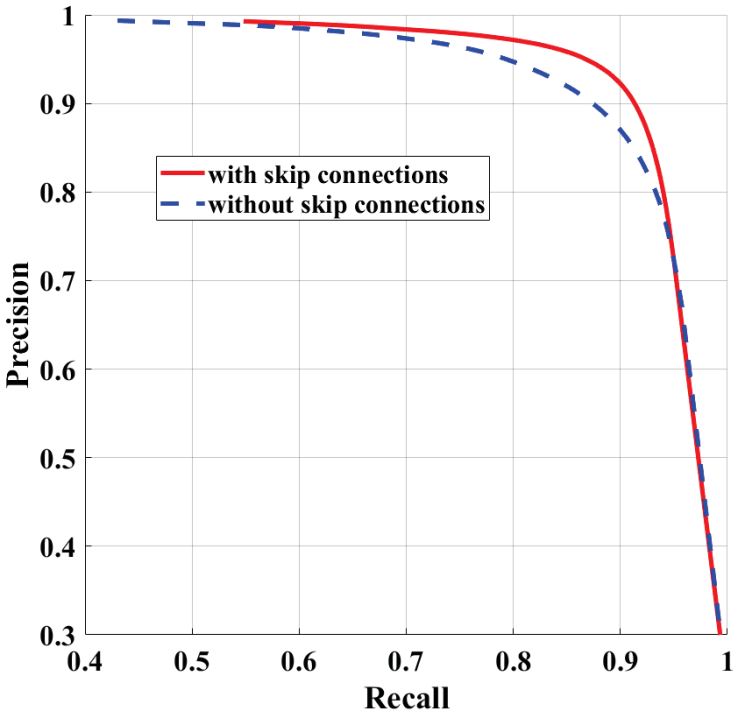


(b)

Figure 5.4: Precision-recall curves obtained the proposed salient object detection network with and without the skip connections between the encoder and decoder layers for images in (a) HKU-IS and (b) PASCAL-S datasets.

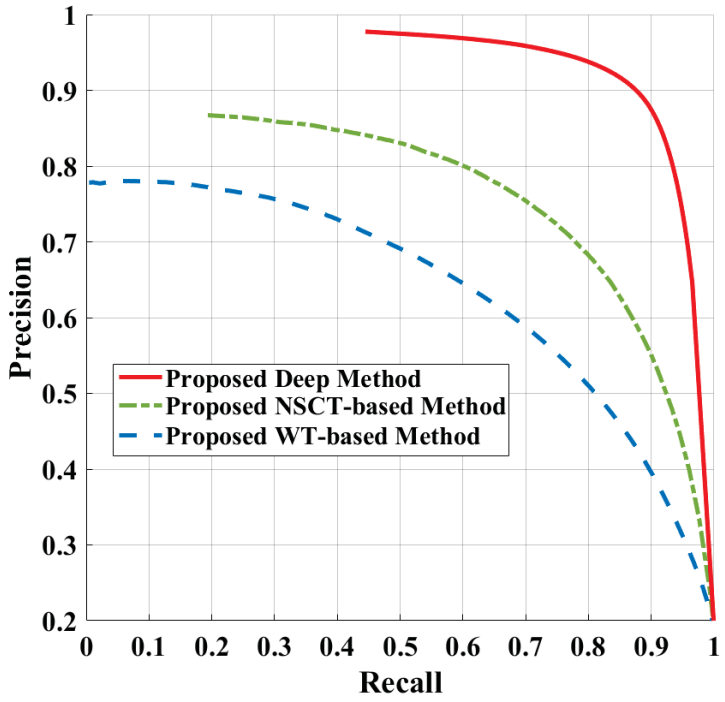


(a)

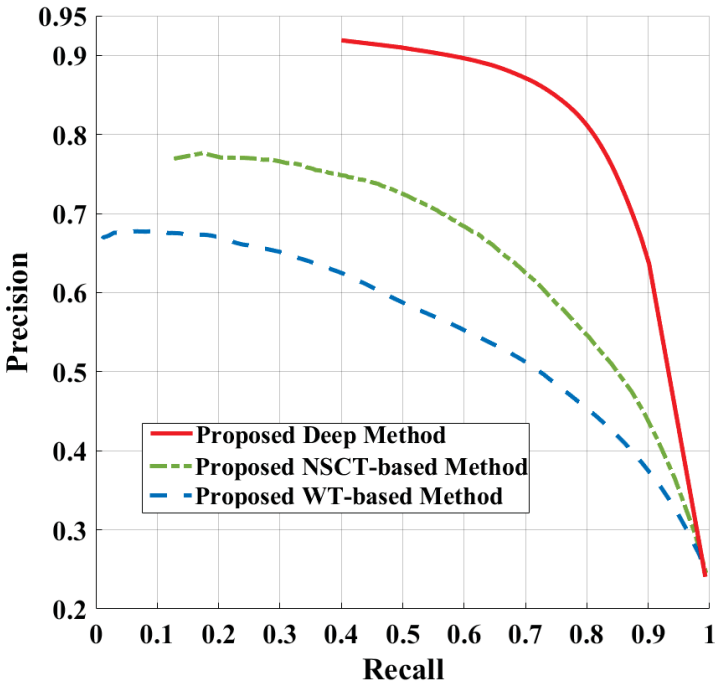


(b)

Figure 5.5: Precision-recall curves obtained the proposed salient object detection network with and without the skip connections between the encoder and decoder layers for images in (a) DUT-OMRON and (b) CSSD datasets.

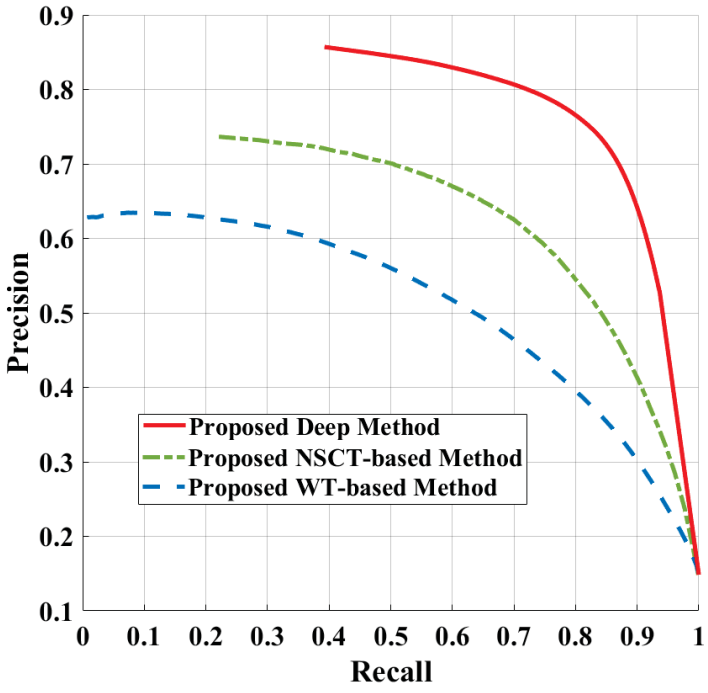


(a)

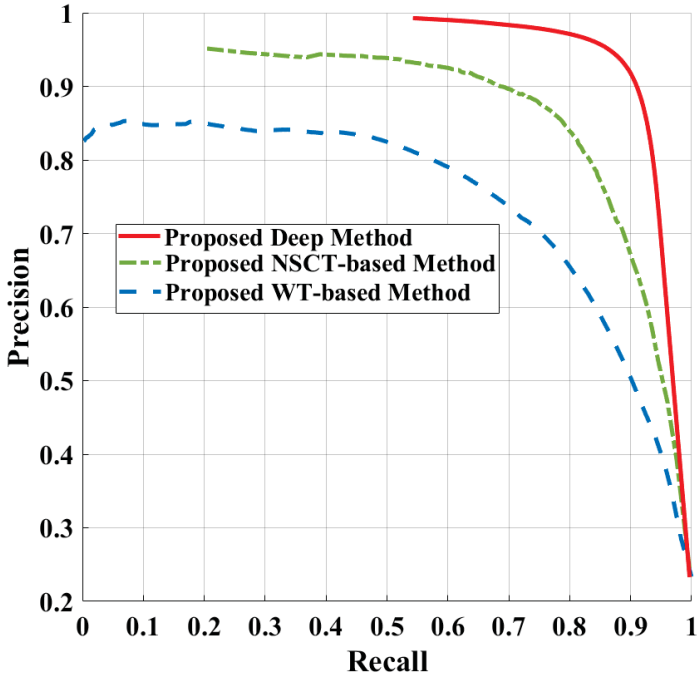


(b)

Figure 5.6: Precision-recall curves obtained by applying the three WT-based, NSCT-based, and deep salient object detection methods proposed in Chapters 3, 4, and 5, respectively, on the images in (a) HKU-IS and (b) PASCAL-S datasets.



(a)



(b)

Figure 5.7: Precision-recall curves obtained by applying the three WT-based, NSCT-based, and deep salient object detection methods proposed in Chapters 3, 4, and 5, respectively, on the images in (a) DUT-OMRON and (b) CSSD datasets.

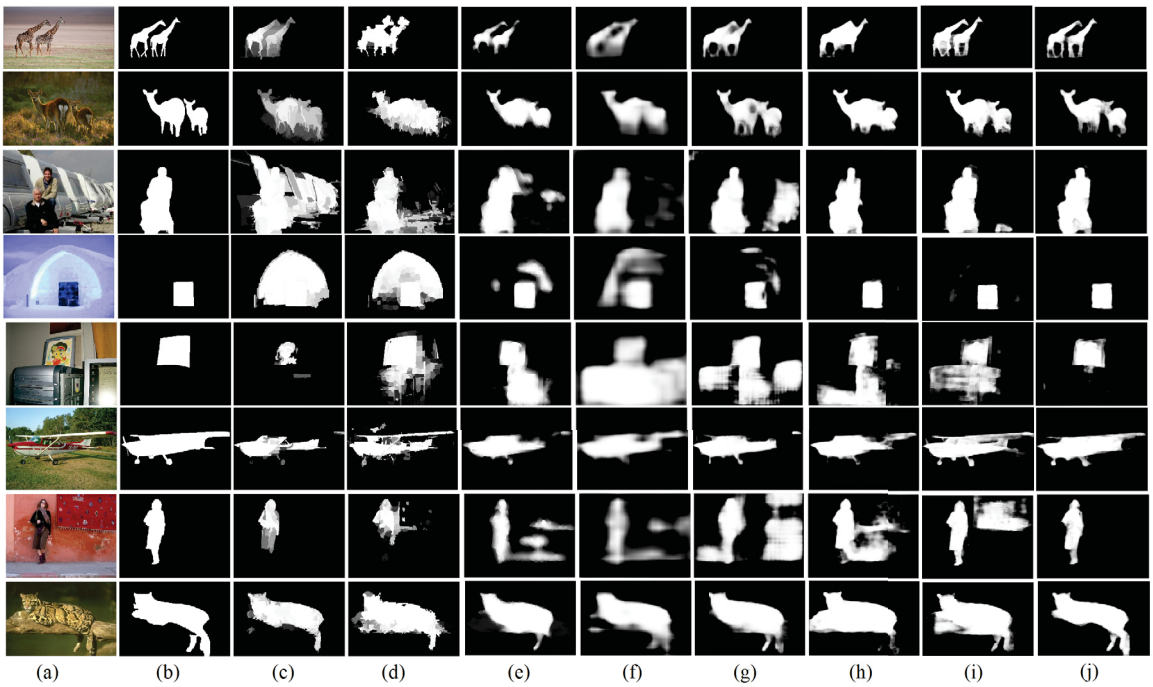
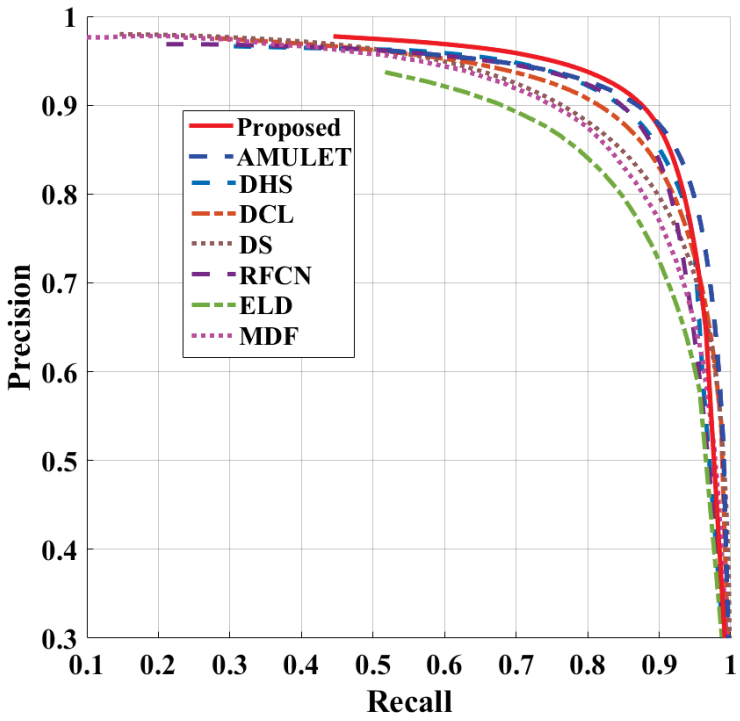
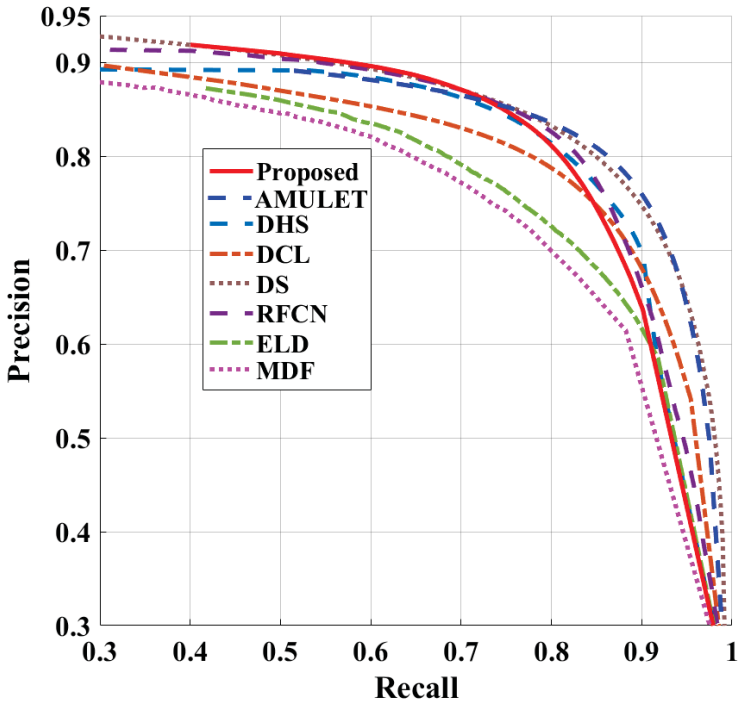


Figure 5.8: Saliency maps obtained by applying the proposed method and the other methods on images from the datasets in [31, 44, 75, 76]. (a) Original image. (b) Ground truth. (c) MDF [44]. (d) ELD [45]. (e) RFCN [49]. (f) DS [46]. (g) DCL [47]. (h) DHS [50]. (i) AMULET [51]. (j) Proposed method [85] .

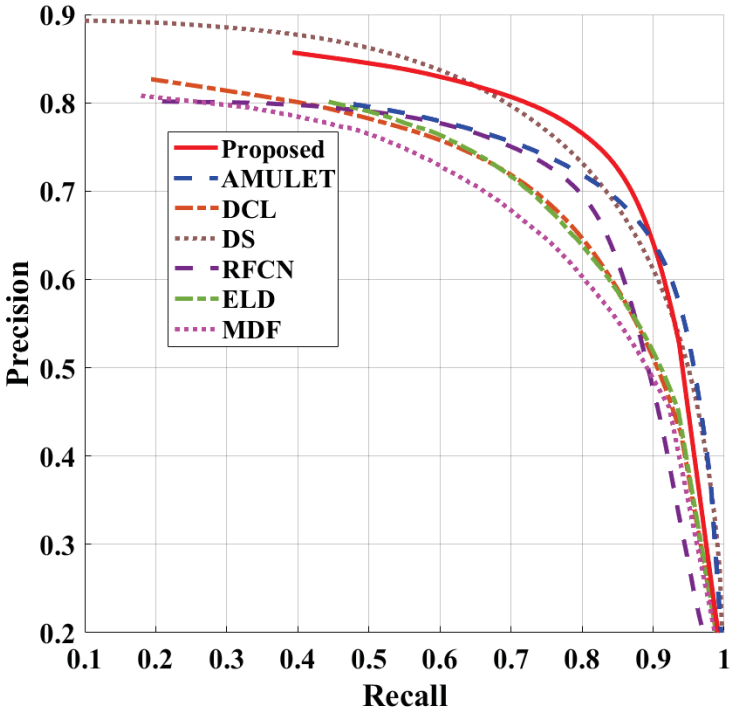


(a)

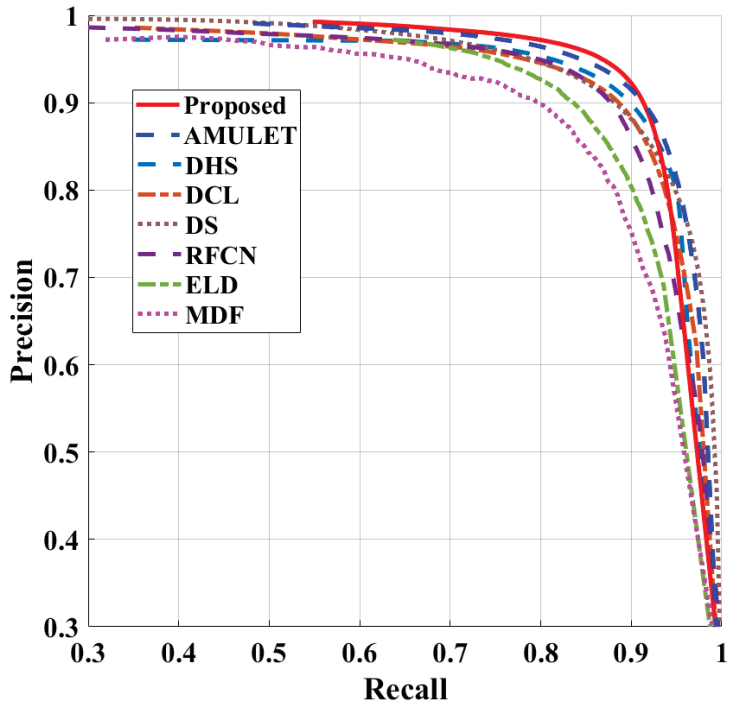


(b)

Figure 5.9: Precision-recall curves obtained by applying the proposed and the other methods for images in (a) HKU-IS, and (b) PASCAL-S datasets.

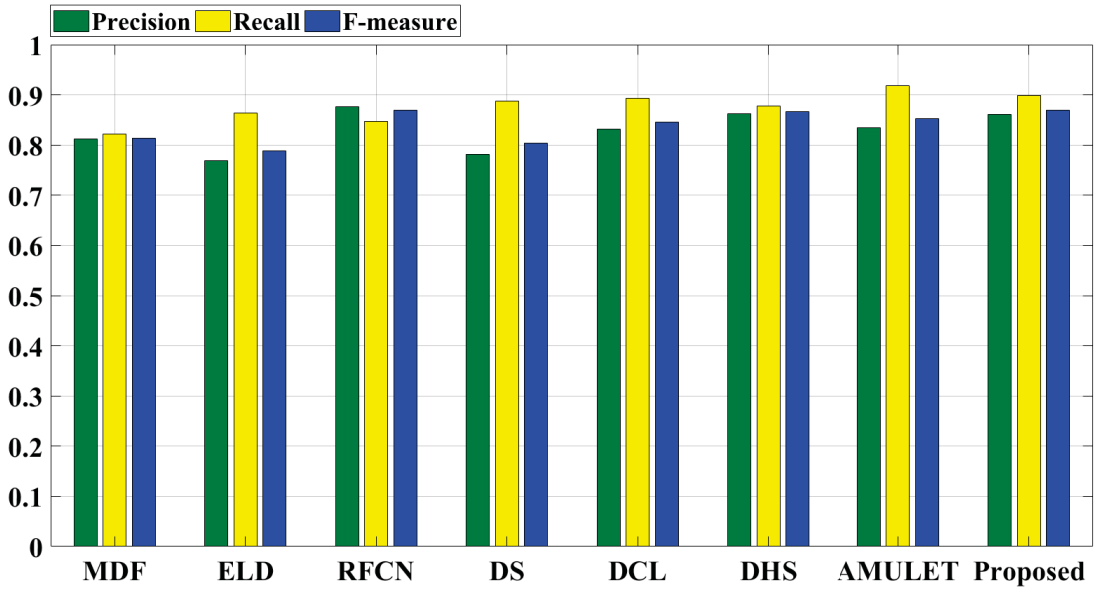


(a)

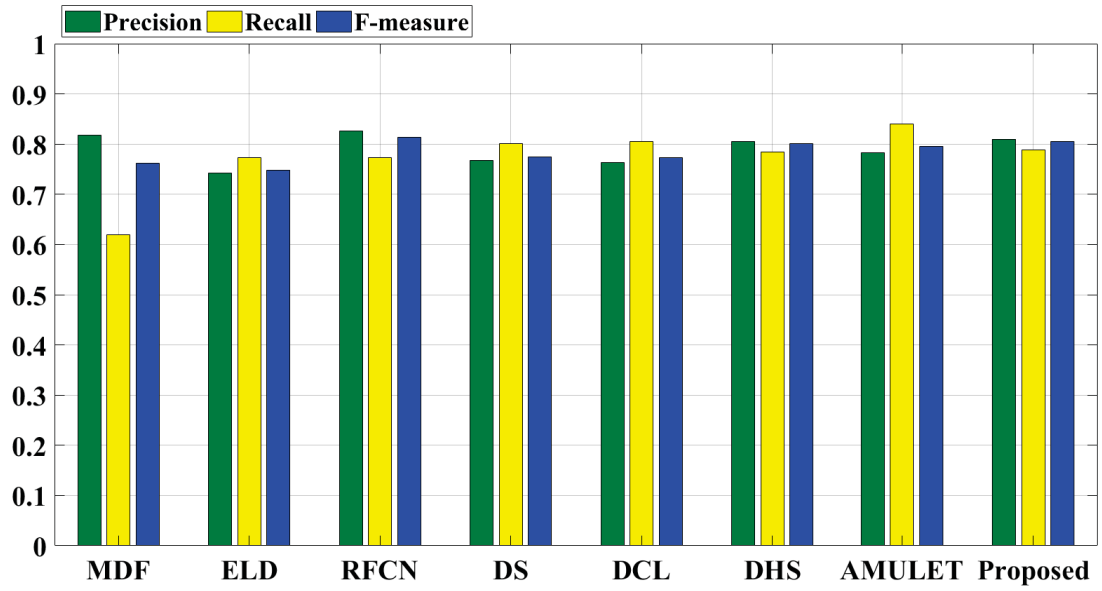


(b)

Figure 5.10: Precision-recall curves obtained by applying the proposed and the other methods for images in (a) DUT-OMRON, and (b) CSSD datasets.

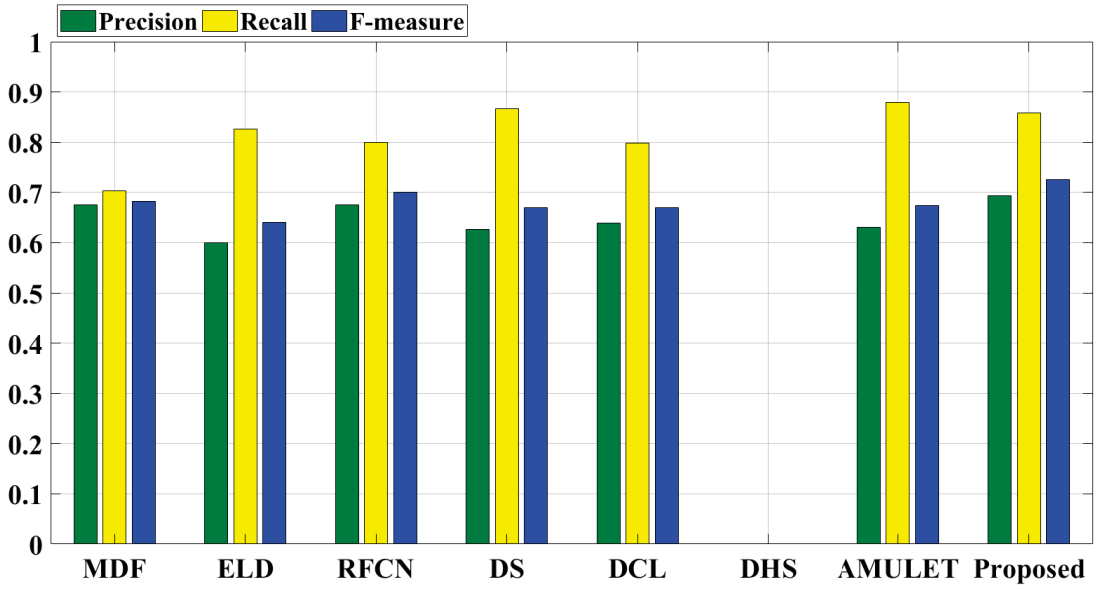


(a)

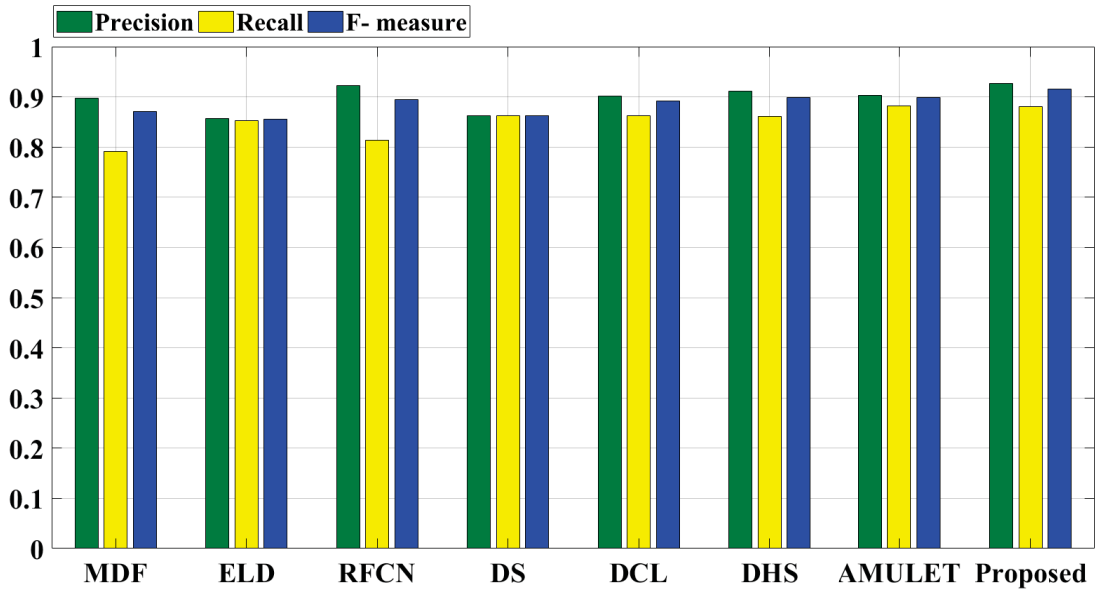


(b)

Figure 5.11: Precision, recall, and F_β obtained by applying the proposed and other salient object detection methods on images in (a) HKU (b) PASCAL-S datasets.



(a)



(b)

Figure 5.12: Precision, recall, and F_β obtained by applying the proposed and other salient object detection methods on images in (a) DUT-OMRON and (b) CSSD datasets.

Chapter 6

Conclusion

6.1 Concluding Remarks

Salient object detection is an active research topic due to its various applications in image processing and computer vision tasks. This thesis has been concerned with the problem of salient object detection in images by devising schemes for multi-resolution multi-level features extraction. A number of algorithms have been developed for extracting features in the wavelet and non-subsampled contourlet transform domains as well as in deep convolutional domain and for their efficient utilization in salient object detection.

In the first part of this thesis, multi-resolution frequency domain feature extraction algorithms have been used to detect the salient objects in images. In view of the capabilities of the WT in extracting multi-resolution image features, a salient object detection method has been proposed by making use of the WT-based features. The multi-resolution features of the image have been extracted using the wavelet coefficients of the three color channels. To combine the features of different color channels, the extracted features have then been fused based on the concept of entropy and border avoidance criterion. In order to evaluate the performance of the proposed method, experiments have been conducted on images from a number of datasets, and it has been shown that the proposed scheme is superior to

that of the other frequency domain schemes in being able to detect the salient object precisely. In addition, to circumvent lack of directional selectivity in the features extracted by the WT, another salient object detection scheme has also been proposed based on image features extracted by the NSCT. The local and global features have been extracted from the local variations of the low-pass coefficients and the distribution of the directional subband coefficients, respectively. Experimental results have shown the superiority of the proposed method to that of the more recent salient object detection schemes.

The second part of this thesis has been concerned with developing a low complexity salient object detection scheme utilizing deep convolutional features. In this part, a deep salient object detection network with an encoder-decoder architecture has been developed where the features are extracted by performing the depthwise separable convolution. Performing the lightweight depthwise separabale convolution in the layers of the proposed network has been resulted in an extremely low complexity compared to the existing deep salient object detection schemes. Such a low complexity has made the use of skip connections affordable in order to improve the performance of the network. In addition, the decomposition of convolution into two separate parts of depthwise and poitwise convolutions, has provided an opportunity to place the non-linear activation unit in between the two parts and thus has made the passage of some negative-valued features as well through the network possible. Experimental results on images from different datasets have shown that the proposed network generally outperforms the other deep salient object detection schemes in terms of precision, recall, F_β and MAE values, while at the same time having a drastically smaller computational complexity.

6.2 Scope for Future Work

There are a number of additional studies that can be undertaken along the research work presented in this thesis. Some of the possible studies are as follows:

- Image features could be extracted using a different image representation such as a weighted graph representing the relations among the pixels or regions of an image.

- In this thesis the depthwise separabale convolution has been used in a deep network with an encoder-decoder architecture. Other deep architectures for salient object detection can be investigated by making use of the lightweight depthwise separable convolution. Moreover, the deep salient object detection network presented in this thesis is developed in a supervised deep learning approach. Developing salient object detection networks using a deep reinforcement learning approach could also be investigated.

- The proposed salient object detection schemes for images can be extended to videos in order to detect the most visually informative region of video scenes.

References

[1] B. C. Ko and J.-Y. Nam, "Object-of-interest image segmentation based on human attention and semantic region clustering," *Journal of the Optical Society of America A*, vol. 23, no. 10, pp. 2462–2470, 2006.

[2] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2009, pp. 817–824.

[3] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, Oct 2004.

[4] X. Y. Stella and D. A. Lisin, "Image compression based on visual saliency at individual scales," in *Proc. International Symposium on Visual Computing*, Las Vegas, NV, USA, 2009, pp. 157–166.

[5] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, Jan 2010.

[6] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, Washington, DC, USA, June 2004, pp. II–II.

[7] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM International Conference on Multimedia*, Berkeley, CA, USA, 2003, pp. 374–381.

[8] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 10, 2007.

[9] Y. Fang, Z. Chen, W. Lin, and C. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3888–3901, Sept 2012.

[10] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. ACM International Conference on Multimedia*, Juan-les-Pins, France, 2002, pp. 533–542.

[11] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, Oct 2005.

[12] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, June 2012, pp. 23–30.

[13] J. Han, E. J. Pauwels, and P. De Zeeuw, "Fast saliency-aware multi-modality image fusion," *Neurocomputing*, vol. 111, pp. 70–80, 2013.

[14] A. Abdulmunem, Y.-K. Lai, and X. Sun, "Saliency guided local and global descriptors for effective action recognition," *Computational Visual Media*, vol. 2, no. 1, pp. 97–106, 2016.

[15] J. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 862–875, April 2015.

[16] Y. Yuan, J. Wang, B. Li, and M. Q.-H. Meng, "Saliency based ulcer detection for wireless capsule endoscopy diagnosis," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 2046–2057, 2015.

[17] Z. Çamlica, H. R. Tizhoosh, and F. Khalvati, "Medical image classification via svm using lbp features from saliency-based folded data," in *Proc. IEEE International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL, USA, Dec 2015, pp. 128–132.

[18] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[19] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 996–1010, 2013.

[20] 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, no. 11, pp. 1254–1259, Nov 1998.

[21] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Advances in Neural Information Processing Systems*, Canada, 2007, pp. 545–552.

[22] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, Jan 2013.

[23] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[24] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, USA, 2007, pp. 1–8.

[25] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, USA, June 2008, pp. 1–8.

[26] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Proc. IEEE International Conference on Computer Vision Systems (ICCV)*, Santorini, Greece, 2008, pp. 66–75.

[27] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009, pp. 1597–1604.

[28] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

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

[30] N. İmamoğlu, W. Lin, and Y. Fang, "A saliency detection model using low-level features based on wavelet transform," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 96–105, 2013.

[31] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA, June 2013, pp. 3166–3173.

[32] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, June 2014, pp. 2814–2821.

[33] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec 2015, pp. 1404–1412.

[34] M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, March 2015.

[35] W. Tu, S. He, Q. Yang, and S. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 2334–2342.

[36] D. Liu, F. Chang, and C. Liu, "Salient object detection fusing global and local information based on nonsubsampled contourlet transform," *Journal of the Optical Society of America A*, vol. 33, no. 8, pp. 1430–1441, 2016.

[37] R. Arya, N. Singh, and R. Agrawal, "A novel hybrid approach for salient object detection using local and global saliency in frequency domain," *Multimedia Tools and Applications*, vol. 75, no. 14, pp. 8267–8287, 2016.

[38] M. Rezaei Abkenar, H. Sadreazami, and M. O. Ahmad, "Graph-based salient object detection using background and foreground connectivity cues," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, Sapporo, Japan, May 2019, pp. 1–5.

[39] D. Lee Fugal, *Conceptual Wavelets in Digital Signal Processing*. San Diego, California: Space & Signals Technical Pub, 2009.

[40] R. Merry and M. Steinbuch, "Wavelet theory and applications," *Eindhoven university of technology, Department of mechanical engineering, Control systems technology group, literature study*, 2005.

[41] L. Wang, H. Lu, X. Ruan, and M. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015, pp. 3183–3192.

[42] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015, pp. 1265–1274.

[43] S. He, R. W. Lau, W. Liu, Z. Huang, and Q. Yang, "Supercnn: A superpixelwise convolutional neural network for salient object detection," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 330–344, 2015.

[44] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 5455–5463.

[45] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 660–668.

[46] X. Li, L. Zhao, L. Wei, M. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, Aug 2016.

[47] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 478–487.

[48] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, Honolulu, HI, USA, 2017, p. 2.

[49] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. European Conference on Computer Vision (ECCV)*, Amsterdam, the Netherlands, 2016, pp. 825–841.

[50] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 678–686.

[51] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Venice, Italy, 2017, pp. 202–211.

[52] A. Borji, M. Cheng, H. Jiang, and J. Li, "Salient object detection: A survey," *CoRR*, vol. abs/1411.5878, 2014. [Online]. Available: http://arxiv.org/abs/1411.5878

[53] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. NY, USA: Pearson/Prentice Hall, 2018.

[54] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, July 1989.

[55] S. Mallat, *A Wavelet Tour of Signal Processing*, 3rd ed. Orlando, FL, USA: Academic Press, Inc., 2008.

[56] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*. Prentice Hall Englewood Cliffs, 1995.

[57] I. Daubechies, *Ten Lectures on Wavelets*. Siam, 1992.

[58] A. L. Da Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: theory, design, and applications," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3089–3101, 2006.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009, pp. 248–255.

[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[62] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015, pp. 1–9.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 770–778.

[64] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 3431–3440.

[65] Y. Liu, M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017, pp. 5872–5881.

[66] R. Girshick, "Fast r-cnn," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec 2015, pp. 1440–1448.

[67] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017, pp. 1951–1959.

[68] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?" in *Proc. European Conference on Computer Vision (ECCV)*, Amsterdam, the Netherlands, 2016, pp. 443–457.

[69] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[70] M. Rezaei Abkenar and M. O. Ahmad, "Superpixel-based salient region detection using the wavelet transform," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, Montreal, QC, Canada, May 2016, pp. 2719–2722.

[71] M. Rezaei Abkenar and M. O. Ahmad, "Salient region detection using efficient wavelet-based textural feature maps," *Multimedia Tools and Applications*, vol. 77, no. 13, pp. 16 291–16 317, Jul 2018.

[72] K. Koffka, *Principles of Gestalt Psychology*. Routledge, 2013.

[73] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Bombay, India, 1998, pp. 839–846.

[74] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[75] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, June 2014, pp. 280–287.

[76] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA, 2013, pp. 1155–1162.

[77] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3169–3176.

[78] M. Rezaei Abkenar and M. O. Ahmad, "Quaternion-based salient region detection using scale space analysis," in *Proc. Signal Processing and Intelligent Systems Conference (SPIS)*, Tehran, Iran, 2015, pp. 78–82.

[79] M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2091–2106, 2005.

[80] M. Rezaei Abkenar, H. Sadreazami, and M. O. Ahmad, "Patch-based salient region detection using statistical modeling in the non-subsampled contourlet domain," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, Baltimore, MD, USA, May 2017, pp. 1–4.

[81] M. Rezaei Abkenar, H. Sadreazami, and M. O. Ahmad, "Salient region detection using feature extraction in the non-subsampled contourlet domain," *IET Image Processing*, vol. 12, no. 12, pp. 2275–2282, 2018.

[82] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, vol. 1, Sep. 2003, pp. I–253.

[83] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, Nov 2012.

[84] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and e cient algorithms," *IEEE transactions on image processing*, vol. 2, no. 2, pp. 176–201, 1993.

[85] M. Rezaei Abkenar and M. O. Ahmad, "Lcnet: A low complexity deep salient object detection network using depthwise separable convolution operation," under review for journal publication.

[86] L. Sifre and S. Mallat, "Rigid-motion scattering for image classification," *Ph. D. thesis*, 2014.

[87] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[88] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 4510–4520.

[89] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[90] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[91] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[92] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[93] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.