

A novel combination of second-order statistical features and segmentation using multi-layer superpixels for salient object detection

Rinki Arya¹ · Navjot Singh^{1,2} · R. K. Agrawal¹

Published online: 20 August 2016
© Springer Science+Business Media New York 2016

Abstract Salient object detection is one of the outstanding capabilities of the human visual system (HVS). The researcher community aims at developing a salient object detection model that matches the detection accuracy as well as computation time taken by the humans. These models can be developed in either spatial domain or frequency domain. Spatial domain models provide good detection accuracy at the cost of high computational time while frequency domain models offer fast computational speed to meet real-time requirements at the cost of poor detection accuracy. In order to induce a trade-off between computational time and accuracy, we propose a model which provides high detection accuracy without taking much of computation time. To detect the salient object with an accurate shape, we first segment the given image by utilizing a bipartite graph partitioning approach which aggregates multi-layer superpixels in a principled and effective manner. Second, the saliency of each segmented region is computed based on a hypercomplex Fourier transform (HFT) saliency map reconstructed using amplitude spectrum, filtered at an appropriate scale chosen using statistical features extracted from grey-level

co-occurrence matrix and original phase spectrum. Finally, a saliency map is generated by taking average of the HFT coefficients of each region in the segmented image and then using the average HFT intensity value of the entire image as a threshold to clearly separate salient object from the background. The performance of the proposed model is evaluated in terms of F -measure, area under curve (AUC), and computation time using six publicly available image datasets. Both qualitative and quantitative evaluations on six publicly available datasets demonstrate the robustness and efficiency of the proposed model against twenty popular state-of-the-art methods.

Keywords Salient object detection · Visual saliency · Gray-level Co-occurrence Matrix (GLCM) · Hypercomplex fourier transform (HFT) · Superpixel · Bipartite graph · Segmentation · Saliency map

1 Introduction

In the era of information technology, images are becoming a rich source of information pertaining to human daily life through social network sites like Facebook, Instagram, 9GAG, Twitter, etc. These images contain a lot of redundant and irrelevant information, which needs to be reduced while the relevant and informative visual data needs to be extracted. The process of extraction of this relevant and informative visual data (salient object) from an image is called salient object detection (SOD) [1–3]. Generally, it is believed that a salient object detection model should have at least the following three properties [3]: (i) it should detect the most salient object with the precise boundaries; (ii) it should generate the full resolution saliency map to correctly locate salient objects and preserve complete information

✉ Rinki Arya
rinki.arya89@gmail.com

Navjot Singh
navjot.singh.09@gmail.com

R. K. Agrawal
rkajnu@gmail.com

¹ School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi 110067, India

² National Institute of Technology, Srinagar, Pauri Garhwal, Uttarakhand 246174, India

of the original image; (iii) computational efficiency of the model should be high. SOD tries to imitate the human visual system (HVS) by concentrating on salient objects present in a complex scene. HVS has a remarkable capability of separating attentive and dominant objects from the background in a given scene even when these objects are witnessed for the first time. It exploits the visual attention mechanism [1–3] to detect salient objects. Detection of these salient objects plays a prominent role in several real time applications, such as image and video compression [4], image retargeting [5], image thumb nailing [6], image segmentation [7, 8], object recognition [9–11], content-aware image editing [12, 13], image classification/retrieval [14, 15], video surveillance systems [16], photo rearrangement [17], image quality assessment [18], remote sensing [19], automatic cropping/centering [20], to display objects on small portable screens [21], automatic target detection [22, 23], robotics [20, 21, 23], medical imaging [24], advertising a design [23], image collection browsing [25], image enhancement [26], and many more.

Visual attention [27, 28] can be achieved by using bottom-up and/or top-down approaches [1–3, 29]. Bottom-up approaches [30] are involuntary, stimulus-driven, and task independent. These approaches utilize low-level visual stimuli to extract features from an image to detect the salient object. The features can be extracted either at the local level or global level [2, 31–33]. The top-down approaches consider high-level information such as prior knowledge of the task, expectations, and emotions to detect the salient object [22, 34, 35]. They are slow and task dependant. Top-down approaches [36] work in combination with the bottom-up approaches for salient object detection. In this paper we reach our goal by using bottom-up mechanisms for salient region detection.

Salient object detection approaches [37–41] can be developed either in frequency domain or spatial domain to obtain the saliency map. Generally, frequency domain methods take low computation time to detect the salient object but do not provide good detection accuracy while spatial domain methods show better detection accuracy in comparison to frequency domain methods but need high computation time. To utilize the advantages of both spatial domain methods and frequency domain methods [42], we propose a salient object detection method which takes less computation time with good detection accuracy. The key idea of the proposed method is to detect a saliency map which can uniformly highlight the most salient object in the given image with accurate boundaries. The idea is implemented in three phases: In phase 1, we need a good segmentation algorithm to obtain an accurate shape of the object. But different segmentation algorithms produce different segmentation results under different parameters. Therefore, to fuse complementary information coming from existing segmentation

algorithms, we use an improved bipartite graph partitioning based segmentation algorithm which integrates a large number of superpixels generated from different segmentation algorithms under different parameters thereby giving better segmentation results. In phase 2, saliency value is assigned to each segmented region obtained from phase 1. To uniformly highlight the salient object in the segmented image, we take advantage of the hypercomplex image representation to combine multiple features (like colour and intensity) in order to get better performance rather than using a single feature. The saliency of each segmented region is obtained by reconstructing the image using the amplitude spectrum, filtered at a scale selected by minimizing our proposed average rank GLCM criterion and the original phase as saliency can be detected by convolving the amplitude spectrum of an image using a low pass Gaussian kernel of an appropriate scale [43]. To determine the optimal scale, we choose four statistical features namely, angular second moment, entropy, inverse difference moment, and contrast, which are complementary to each other. Together all these four features provide high discriminative power to distinguish two different spatial structures. Finally, we combine the spatial saliency information obtained from phase 2 and segmentation information from phase 1 to obtain the rough saliency map. We use HFT coefficients to locate the salient object in an image while superpixel segmentation is utilized to improve the object contours. Lastly, a final saliency map is obtained by clearly separating the foreground and background of the rough saliency map by using mean HFT intensity value of the entire image as a threshold. If the pixel saliency value in rough saliency map is greater than or equal to the threshold value then the pixel is considered to be salient, otherwise background. The main contributions of our paper are:

1. We propose a novel criterion based on statistical features extracted from GLCM to select an appropriate scale to filter amplitude spectrum of hypercomplex image Fourier transform to detect saliency in frequency domain. We have used aggregation of four statistical features namely, entropy, inverse difference moment (IDM), angular second moment (ASM), and contrast which are effective and complementary to each other. The main reason behind using GLCM is its capability to capture second order statistical properties of an image.
2. Although, frequency domain models work well to detect salient object in less computation time, but fail to capture the fine shape or structure of object. To obtain the accurate shape of the object, we utilize an existing bipartite graph partitioning based segmentation approach [43] which aggregates multi-layer superpixels in an effective way. To the best of authors' knowledge, this segmentation algorithm has not yet been used in

salient object detection field. The main reasons behind using this segmentation algorithm are (a) It provides good segmentation results with low computational complexity (b) It utilizes multi-layer superpixels, generated from Comaniciu & Meer's Mean Shift [44] and Felzenszwalb & Huttenlocher's [45] segmentation algorithms by varying their parameters, which helps in capturing the multi-scale visual patterns of an image in an effective manner.

3. To check the robustness and efficacy of our proposed model, the performance is assessed both quantitatively and qualitatively on six datasets. Experiments demonstrate that the proposed model considerably outperforms twenty existing state-of-the-art methods from both frequency and spatial domain.

The rest of the paper is organized as follows: Section 2 includes the relevant work. Section 3 describes our proposed method. Further, the experimental setup, dataset description and the obtained results are discussed in Section 4. Finally Section 5 includes conclusion and future research directions.

2 Related work

In 1998, Itti et al. proposed the first visual saliency model which created the first wave of interest in the area of visual saliency across multiple disciplines including computer vision, neuroscience, and cognitive psychology communities. They utilized feature integration theory to combine intensity, colour, and orientation feature maps into a saliency map [22]. In research work [46] Itti et al. model has been extended using region growing techniques. Later, in another research work [47] a computational model based on information maximization has been proposed to implement saliency using joint likelihood, independent component analysis (ICA), and Shannon's self-information. In the research work [48], a graph based visual saliency (GBVS) model has been suggested which utilize a Markov chain analysis to detect the salient object. Hou and Zhang [49] utilized spectral residual (SR) on the principle that redundancies are measured in terms of similarities. They suggested that the uneven regions account for the statistical singularities present in the spectrum. This research work is based on calculation of the spectral residual, which is computed by taking the difference between log spectrum of an input image and its smoothed version. Liu et al. [50, 51] extracted features at global, regional, and local level and used a supervised approach to partition the image into attention region and background region. In another research work [52] phase spectrum of Fourier transform (PFT) is singled out to be the most significant factor to locate the

salient regions. The research work [53] used pulsed discrete cosine transform (PCT) to imitate lateral surround inhibition behaviour of neurons. In research work [13] a frequency tuned (FT) saliency detection approach based on *Lab* colour space has been suggested which computes Euclidean distance between pixel vector in a Gaussian filtered image and the average vector for the input image. Bian and Zhang [54, 55] proposed a saliency detection approach that integrates the speed of frequency domain models with the topology of biologically based methods under the assistance of frequency domain divisive normalization (FDN). But, this model takes global surround into consideration. In research work [56], FDN model has been extended into piecewise frequency domain divisive normalization (PFDN) in order to relax the global surround constraint by dividing the image into overlapping local patches and conducting FDN on every patch in order to provide better biological plausibility. In another research work [57], maximum symmetric surround technique has been implemented to allocate small bandwidth to the border filter and large bandwidth to the centre filter. Guo and Zhang [58] suggested an approach to compute the saliency map by exploiting quaternion representation of an image by extending PFT model. In research work [59], a context-aware saliency detection method has been suggested. This is a top-down approach which makes use of both global and local saliency by exploiting some visual organization rules. This method is computationally expensive and takes 124 seconds per image to detect salient object. Shen and Wu [60] incorporated the concept of low rank matrix (LRMR) to detect the most salient image region. In this model they make use of some features extracted from the given image like colour, orientation, mean-shift segmentation and low-rank matrix. The shape information of the objects was clear, but the model is computationally expensive. In research work [61], amplitude spectrum of quaternion Fourier transform (AQFT) and the human visual sensitivity has been proposed to detect image saliency. In research work [62], colour, intensity, and texture features are extracted from the discrete cosine transform (DCT) coefficients from JPEG bit-stream to compute the DCT block differences. In 2013, Li et al. [63] proposed that saliency in an image can be detected by convolving image amplitude spectrum with a low-pass Gaussian kernel of an appropriate scale. To detect the saliency, spectrum scale space (SSS) is obtained by convolving the amplitude spectrum of an image with low-pass Gaussian kernel of different scales. Thereafter, an optimal scale with the smallest entropy value is chosen to generate the best saliency map. They employed the hypercomplex Fourier transform (HFT) in order to combine multidimensional feature maps (like colour, intensity) for spectrum scale-space analysis. Li et al. [64] designed a saliency detector by utilizing both supervised and unsupervised learning processes. They stated

that phase of intermediate frequencies is the key of image saliency. In research work [65], a salient region detection approach based on global saliency and local saliency in the frequency domain by using fast Walsh-Hadamard transform (FWHT) and PFDN respectively has been suggested. Jiang et al. [66] suggested a supervised learning based salient object detection approach (DRFI) which involves learning of a random forest repressor which helps in integrating a high-dimensional regional feature descriptor to predict region saliency score and determining the most discriminative features automatically. The experimental results, shown by Jiang et al. in their paper, are effective but its computational speed is much higher. It takes around 24 hours for training and 28 seconds for testing. Zou and Komodakis [67] suggested a salient object detection method based on a hierarchy-associated rich feature (HARF) construction framework. This framework makes a hierarchy of basic features obtained from multi-level regions and incorporates multi-layered deep learning features also to characterize the context of the whole object/background. The experimental results, shown by Zou and Komodakis in their paper, are good but its computational speed is much higher. It takes around 30 hours for training and 37 seconds. Sun et al. [68] used the concept of absorption of Markov chain to detect salient object in the given image. In research work [38], global contrast and spatial coherence have been proposed for salient object detection. Perazzi et al. [69] suggested the concept of saliency filters for saliency detection. Yan et al. [70] contributed to the salient object detection area by giving the concept of Hierarchical saliency. Liu et al. [71] suggested an approach for salient object detection based on saliency tree. Zhu et al. [72] suggested a background

measure and an optimization framework for salient object detection. Li et al. [73] contributed to the field by proposing a dense and sparse reconstruction based salient region detection method. In research work [74], fusion of visual saliency and generic objectness has been suggested for salient object detection. Jiang et al. [75] considered context and shape prior for salient object detection for better shape information but this method is computationally expensive. In research work [76], graph-based manifold ranking has been proposed for salient region detection. Margolin et al. [77] suggested a saliency framework based on the distinctness of patches in the given image. In research work [27], a high-dimensional colour transform (HDCT) has been proposed for salient object detection. This research work maps RGB colour space feature vector, from a low dimension to a high-dimensional by finding the optimal linear combination of colour coefficients.

3 The proposed model

In this section, we present an efficient approach to detect salient object in an image. To better capture intrinsic structural information and improve computational efficiency, we first segment the given image by utilizing a bipartite graph partitioning approach which aggregates multi-layer superpixels in an effective and principled manner. Second, the saliency of each segmented region is computed based on a hypercomplex Fourier transform (HFT) saliency map reconstructed using amplitude spectrum, filtered at an appropriate scale chosen using statistical features extracted from grey — level co-occurrence matrix and original phase spectrum.

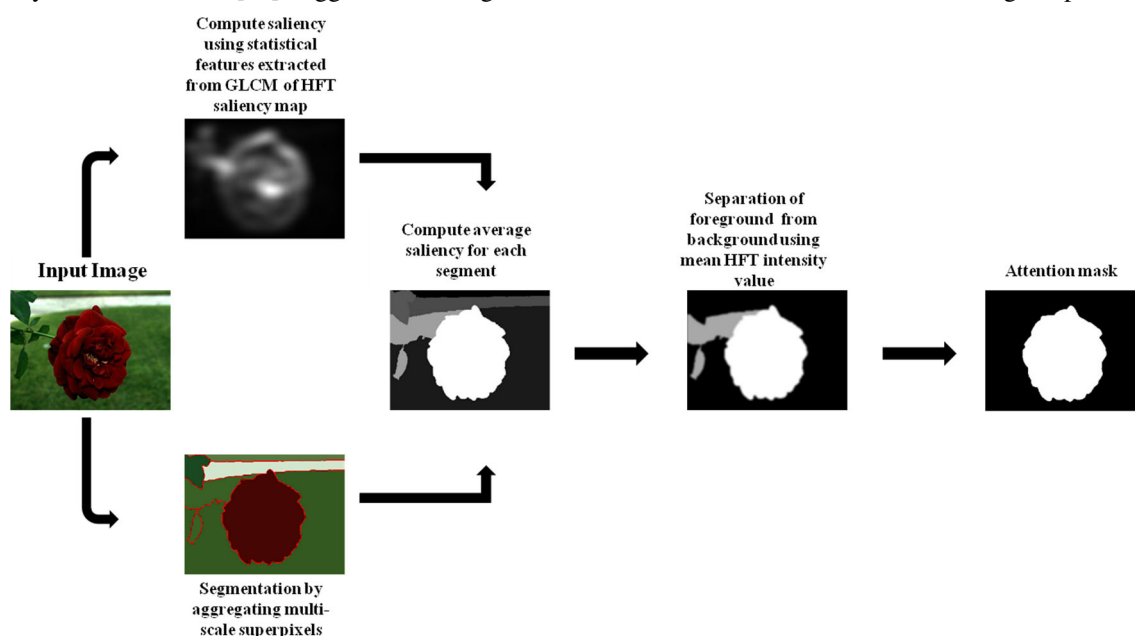


Fig. 1 Outline of detecting salient object using our proposed model (BHGT)

Thereafter, a rough saliency map is generated by taking average of the HFT coefficients of each region in the segmented image. Finally, the mean HFT intensity value of the entire image is used as a threshold to separate salient object from the background in the final saliency map. If the pixel saliency value is greater than or equal to the threshold value than the pixel is considered to be salient, otherwise if the pixel saliency value is less than the threshold then it is assigned a value of 0. The outline of the proposed model (BHGT) is illustrated in Fig. 1.

3.1 Segmentation of an image by aggregating superpixels using a bipartite graph

In this section, we aim at partitioning the entire image into multiple segments to obtain accurate shape information of the most salient object. The approach of first segmenting the given image and then choosing the most salient object was not used by research community in the past due to the lack of highly accurate segmentation algorithm with low computational complexity [3]. To overcome this challenge, recently superpixels are becoming increasingly popular in the field of salient object detection. Superpixel, a group of pixels with similar characteristics, is quite accurate in segmenting objects and is very fast to compute. The choice of a good segmentation algorithm considerably affects the performance of our proposed model. In literature, several algorithms such as Minimum Spanning Tree [45], Normalized Cut [78], and Mean Shift [44] have been used for image segmentation purpose. Recently Li et al. [43] proposed an improved image segmentation algorithm by taking advantage of different and complementary information coming from various popular segmentation algorithms [44, 45, 78]. In order to fuse this complementary information, Li et al. collected a variety of superpixels generated by different segmentation algorithms with varying parameters. Superpixels generated in this way help in capturing diverse and multi-scale visual patterns in the input image. To effectively aggregate these multi-layer superpixels, Li et al. proposed a bipartite graph partitioning based segmentation framework which is constructed over both pixels and superpixels. These pixels and superpixels work as the vertices of the bipartite graph and edges between these vertices are established on the basis of superpixel cues and smoothness cues. To enforce superpixel cues, a pixel is connected to the superpixel if pixel is a part of that superpixel while smoothness cues are enforced by connecting each superpixel to itself and its nearest neighbour in the feature space among its spatially adjacent superpixels. This bipartite graph segmentation framework can be computationally customized to an unbalanced bipartite graph structure which can be solved using a linear-time spectral algorithm. Development of this spectral algorithm to solve an unbalanced bipartite graph

structure makes Li et al. algorithm computationally very fast and efficient. For a given image \mathbf{I} and a set of superpixels \mathcal{S} , an unbalanced and undirected bipartite graph with two sets of vertices χ and γ can be represented as $\mathbf{G} = \{\chi, \gamma, \mathbf{B}\}$, where $\chi = \mathbf{I} \cup \mathcal{S} = \{x_i\}_{i=1}^{N_\chi}$ and $\gamma = \mathcal{S} = \{y_j\}_{j=1}^{N_\gamma}$ with N_χ and N_γ as the number of nodes in χ and γ respectively.

The across affinity matrix $\mathbf{B} = (b_{ij})_{N_\chi \times N_\gamma}$ is defined as follows:

$$b_{ij} = \begin{cases} \alpha, & x_i \in y_j, x_i \in \mathbf{I}, y_j \in \mathcal{S} \\ e^{-\beta d_{ij}}, & \text{if } x_i \sim y_j, x_i \in \mathcal{S}, y_j \in \mathcal{S} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $N_\chi = |\mathbf{I}| + |\mathcal{S}|$ and $N_\gamma = |\mathcal{S}|$ and $d_{i,j}$ signifies the distance between the features of superpixels x_i and y_j . α and β are set to greater than 0 to balance superpixel and smoothness cues. \sim signifies a certain neighbourhood between superpixels. Using bipartite graph \mathbf{G} , input image \mathbf{I} is segmented into k segments by accumulating same label nodes into a segment with the help of spectral clustering algorithm. To segment an image into k groups, k bottom eigenvectors of generalized eigenvalue problem are computed as:

$$\mathbf{L}\mathbf{f} = \lambda\mathbf{D}\mathbf{f} \quad (2)$$

where \mathbf{L} and \mathbf{D} represent graph Laplacian and degree matrix respectively. \mathbf{D} is calculated as: $\mathbf{D} = \text{diag}(\mathbf{B}\mathbf{1})$. This eigenvalue problem could be solved by applying Lanczos method to the normalized affinity matrix $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ by treating \mathbf{G} as a sparse graph. But it takes $O(k(N_\chi + N_\gamma)^{3/2})$ running time [43]. This eigenvalue problem could also be solved by SVD but it also takes $O(k(N_\chi + N_\gamma)^{3/2})$ running time as explained in [43]. As it can be observed from across affinity matrix \mathbf{B} that the number of columns in \mathbf{B} are much larger than the number of rows $N_\chi = N_\gamma + |\mathbf{I}|$, and $\mathbf{I} \gg N_\gamma$ so we have $N_\chi \gg N_\gamma$. This large variation between the number of rows and number of columns clearly demonstrates the unbalanced structure of the bipartite graph. To exploit the unbalanced structure, Li et al. proposed a transfer cut method to compute bottom k eigenvectors in reduced time as it transforms the eigenvalue problem in to the following:

$$\mathbf{L}_\gamma \mathbf{v} = \eta \mathbf{D}_\gamma \mathbf{v} \quad (3)$$

where $\mathbf{L}_\gamma = \mathbf{D}_\gamma - \mathbf{W}_\gamma$, $\mathbf{D}_\gamma = \text{diag}(\mathbf{B}^T \mathbf{1})$, and $\mathbf{W}_\gamma = \mathbf{B}^T \mathbf{D}_\chi^{-1} \mathbf{B}$. \mathbf{L}_γ is Laplacian of the bipartite graph $\mathbf{G}_\gamma = \{\gamma, \mathbf{W}_\gamma\}$ as $\mathbf{D}_\gamma = \text{diag}(\mathbf{W}_\gamma \mathbf{1})$. In this way the task of partitioning graph \mathbf{G} into k groups takes $O(2k(1 + d_\chi)N_\chi + kN_\gamma^{\frac{3}{2}})$ time where d_χ is the average number of edges connected to each node in χ . Our work belongs to salient object detection, for which a comprehensive discussion about segmentation approaches is beyond the scope of this paper. We refer readers to research article proposed by Li et al. [43] for a detailed discussion of this segmentation approach. To choose the most salient region among these k segmented regions R_p , $p = 1, \dots, k$, the saliency value of each region,

R_p , needs to be computed. To find out the saliency value of each region, we utilize statistical features obtained from grey-level co-occurrence matrix (GLCM) of hypercomplex Fourier transform (HFT) saliency map, which is discussed in Section 3.2.

3.2 Saliency using statistical features extracted from GLCM of HFT saliency map

We have utilized a frequency domain approach to compute saliency of all the segmented regions obtained from Section 3.1 because of their fast computational speed. In general, frequency domain based visual attention models belong to one of the two categories: 1) models which generate a saliency map by processing the spectrum of each colour channel separately and then fusing the individual saliency maps into the final saliency map, and 2) models which merge colour channel images into a quaternion image and then use the hypercomplex Fourier transform (HFT) to obtain the quaternion spectrum for processing. The models, belonging to the latter category show better performance in comparison to the models belonging to the first category [79]. In this paper, we focussed on the latter category of frequency models to compute saliency map by utilizing quaternion or hypercomplex representation to combine multiple features (like colour and intensity) in order to get better performance [63, 79] rather than using a single feature as in [49]. In 1996, Castlman [80] came up with the properties of amplitude spectrum and phase spectrum of discrete Fourier transform. He pointed out that amplitude spectrum represents how much of each frequency components exists within the image and phase spectrum signifies where each of the frequency components resides within the image representing information regarding local properties of the image. In 2007, Hou and Zhang suggested a frequency domain approach which utilize spectral residual (SR) of the amplitude spectrum to calculate saliency in an image while keeping phase spectrum for computational purpose only. Although SR model provided good results however it doesn't make use of the phase spectrum. In 2008, Guo et al. singled out phase spectrum of an image's Fourier transform as the key to locate salient objects in an image and proposed the phase spectrum of Fourier transform (PFT) model [52, 58] for grey-scale image by taking inverse Fourier transform (IFT) of the phase spectrum alone by keeping amplitude spectrum constant. The PFT model is computationally very fast but it only takes grey-scale images as an input without considering other low-level features e.g. colour, orientation, motion etc. Motivated by this, Guo and Zhang [58] extended PFT approach to phase spectrum of quaternion Fourier transform (PQFT) [58] by taking advantage of the quaternion representation of the image. Quaternion representation of an image represents each pixel

by a quaternion of low level features in a whole unit to achieve high accuracy without losing any information [79]. These low level features may include colour, intensity, orientation, motion etc. Recently, Li et al. [63] pointed out that different filter scales can capture different types of salient regions e.g. a small-scale kernel detects the large salient regions efficiently while very small-scale may not suppress the repeated pattern satisfactorily. Similarly a large-scale kernel works best to detect small salient regions but a very large-scale only highlight the salient object boundaries. Therefore, it is important to choose an appropriate scale for Gaussian kernel to suppress repeated patterns sufficiently. To realize this idea, Li et al. [63] proposed a hypercomplex Fourier transform (HFT) model by smoothing the log amplitude spectrum with Gaussian functions of different variances while keeping the phase spectrum constant. They used hypercomplex representation of an image by fusing multiple low-level features like intensity, colour etc. into a hypercomplex matrix.

3.2.1 Saliency map calculation using hypercomplex Fourier Transform(HFT)

Quaternion Fourier transform was first applied to colour images in the research work of [81] by utilizing a discrete version of Ell's transform [82, 83]. Later, Pei et al. implemented quaternion Fourier transform (QFT) [84] by considering the transform mentioned in [85]. In 2007, Ell and Sangwine [86] proposed QFT of colour images which was utilized by Guo et al. [52, 58] to compute saliency map for colour images in saliency domain for the first time. Inspired by the research work of [52, 58, 86], Li et al. also used the concept of hypercomplex matrix to combine multiple feature maps as follows:

$$\mathbf{f} = w_1 \mathbf{f}_1 + w_2 \mathbf{f}_2 \mathbf{i} + w_3 \mathbf{f}_3 \mathbf{j} + w_4 \mathbf{f}_4 \mathbf{k} \quad (4)$$

where w_m denote the m^{th} weight, $w_1 = 0$, $w_2 = 0.5$, $w_3 = w_4 = 0.25$ and \mathbf{f}_m represents the m^{th} feature map. $\mathbf{f}_2 = \mathbf{I}_s = (\mathbf{r} + \mathbf{g} + \mathbf{b})/3$, $\mathbf{f}_3 = \mathbf{RG} = \mathbf{R} - \mathbf{G}$, and $\mathbf{f}_4 = \mathbf{B} - \mathbf{Y}$ [87, 88], where \mathbf{r} , \mathbf{g} , \mathbf{b} represent the red, green and blue channels of a given input colour image and $\mathbf{R} = \mathbf{r} - (\mathbf{g} + \mathbf{b})/2$, $\mathbf{G} = \mathbf{g} - (\mathbf{r} + \mathbf{b})/2$, $\mathbf{B} = \mathbf{b} - (\mathbf{r} + \mathbf{g})/2$ and $\mathbf{Y} = ((\mathbf{r} + \mathbf{g}))/2 - |\mathbf{r} - \mathbf{g}|/2 - \mathbf{b}$. Here \mathbf{RG} corresponds Red/green, green/red colour pair and \mathbf{BY} corresponds to the blue/yellow, yellow/blue colour pair for implementing the "colour opponent-component" system which exists in human visual cortex [88]. In general, quaternion has 4 degree of freedom which represents 4 independent feature channels. In our case, as we have w_1 to be zero and other three to be non-zero components, it has only 3 degree of freedom. $F_H[u, v]$ can be denoted in the polar form as:

$$F_H[u, v] = \|F_H[u, v]\| e^{i\mu\Phi(u, v)} \quad (5)$$

$F_H [u, v]$ is the frequency domain representation of $f (n, m)$. The hypercomplex transform and the inverse transformation of the image can be represented as:

$$F_H [u, v] = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\mu 2\pi ((\frac{mv}{M}) + (\frac{nu}{N}))} f (n, m) \quad (6)$$

$$f (n, m) = \frac{1}{\sqrt{MN}} \sum_{v=0}^{M-1} \sum_{u=0}^{N-1} e^{\mu 2\pi ((\frac{mv}{M}) + (\frac{nu}{N}))} F_H [u, v] \quad (7)$$

Its phase spectrum $P (u, v)$, amplitude spectrum $A (u, v)$, and eigen-axis spectrum $E (u, v)$ are defined as:

$$A (u, v) = \|F_H [u, v]\| \quad (8)$$

$$P (u, v) = \Phi (u, v) = \tan^{-1} \frac{\|V (F (u, v))\|}{S (F (u, v))} \quad (9)$$

$$E (u, v) = \mu (u, v) = \frac{V (F (u, v))}{\|V (F (u, v))\|} \quad (10)$$

To choose the best scale, Li et al. generated a sequence of smoothed amplitude spectrum called Spectrum Scale-Space (SSS) $\Lambda = \{\Lambda_b\}$ by adopting Gaussian filter smoothing of amplitude spectrum $A (u, v)$ at different scales by using (11).

$$\Lambda (u, v; b) = (g (., .; b) * A) (u, v) \quad (11)$$

where b signifies scale parameter, $b = 1, \dots, B$. $B = \lceil \log_2 \min \{Height, Width\} \rceil + 1$, where $Height$ $Width$ denote the height and width of the input image respectively. The Gaussian filter $g (u, v; b)$ is defined as:







$$g (u, v; b) = \frac{1}{\sqrt{2\pi 2^{b-1} t_0}} e^{-(u^2+v^2)/2^{2b-1} t_0^2} \quad (12)$$

where t_0 is experimentally set to 0.5. Different saliency maps corresponding to different scales are generated by performing the inverse transform for a given amplitude spectrum Λ_b , while keeping the phase spectrum $P (u, v)$ and eigenaxis spectrum $E (u, v)$ constant by using the (13):

$$S_b = g * \left\| F_H^{-1} \left\{ \Lambda_b (u, v) e^{EP(u,v)} \right\} \right\|^2 \quad (13)$$

where g represents a fixed scale Gaussian kernel. Thus a series of saliency maps $\{S_b\}$ is obtained. To choose the best saliency map S from $\{S_b\}$, we propose a novel criterion based on statistical features obtained from GLCM, which is discussed in (Fig. 2).

Fig. 2 Binary images with different spatial structures, but with the same histogram

| DIFFERENT SPATIAL STRUCTURES | CLASSICAL ENTROPY VALUE | AVERAGE ASM VALUE | AVERAGE IDM VALUE | AVERAGE CONTRAST VALUE | AVERAGE ENTROPY VALUE |
|---|-------------------------|-------------------|-------------------|------------------------|-----------------------|
|  | 1 | 0.491 | 0.992 | 0.436 | 0.741 |
|  | 1 | 0.483 | 0.985 | 0.865 | 0.779 |
|  | 1 | 0.465 | 0.967 | 1.821 | 0.847 |
|  | 1 | 0.443 | 0.946 | 3.020 | 0.918 |
|  | 1 | 0.410 | 0.909 | 5.064 | 1.00 |
|  | 1 | 0.274 | 0.582 | 2.339 | 1.34 |

3.2.2 Finding the proper scale using GLCM criterion

As we have already discussed that we need to choose a specific scale in the sequence $\{S_b\}$ to select the best saliency map. To choose the best saliency map, Li et al. [63] gave the idea of computing the conventional entropy on the smoothed image filtered by a Gaussian kernel of an appropriate scale to capture the spatial structure of the saliency map. And it is a challenge in itself to choose an appropriate scale for a low-pass Gaussian kernel as mentioned in Li et al. research work. Although, Li et al. method considers spatial structure of the saliency map to a limited extent, however their criteria is not able to choose the best saliency map as shown in Fig. 3. In order to overcome this problem, we propose a novel criterion to determine the optimal scale o_p based on statistical features extracted from GLCM of HFT saliency map to capture the spatial geometric information which is ignored when entropy is computed directly on either HFT saliency map or filtered HFT saliency map [63]. As shown

$$P_{\Delta x, \Delta y}(ij) = \sum_{p=1}^m \sum_{q=1}^n \begin{cases} 1, & \text{if } I(p, q) = i \text{ and } I(p + \Delta x, q + \Delta y) = j \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where the offset $(\Delta x, \Delta y)$ depends on the angular relationship θ and distance d between the neighbouring pixels. We have used eight directions in this paper with distance $d = 1, 2$ and angular relationship $\theta = \{0^0, 45^0, 90^0, 135^0\}$.

To determine the optimal scale, we choose four statistical features namely, Angular second moment, Entropy, Inverse difference moment, and Contrast, which are complementary to each other. These four statistical features are defined as:

- a) **Angular second moment (ASM):** ASM is a measure of homogeneity of an image, which takes high values when the image is homogeneous. It is also called energy/uniformity.

$$ASM = \sum_{i,j} P_{d,\theta}(i, j)^2 \tag{15}$$

- b) **Inverse difference moment (IDM):** IDM measures the local homogeneity of an image and assigns a relatively higher value for homogeneous images. It is also called homogeneity.

$$IDM = \sum_{i,j} \frac{P_{d,\theta}(i, j)}{1 + |i - j|} \tag{16}$$

- c) **Contrast:** Local variations in an image can be efficiently captured using contrast. Large amount of variation in an image corresponds to high contrast value.

$$Contrast = \sum_{i,j} (i - j)^2 P_{d,\theta}(i, j) \tag{17}$$

in Fig. 2, binary images with different spatial structures and same probability distributions of intensity have the same classical entropy. In such situation, entropy is not able to distinguish two different geometrical shapes which also play an important role in detecting a salient object. It can also be seen from Fig. 3 that Li et al. improved entropy criterion is also not able to choose the best saliency map in spite of considering spatial structure to some extent. So we utilize statistical features generated from GLCM to capture the spatial structure of an image. It can be seen from Fig. 2 that for each statistical feature, the average value of eight scores corresponding to eight offsets is not same for different spatial structures while conventional entropy is same for all structures. Grey-level co-occurrence matrices are originally defined by Haralick in 1973 [89]. It can be briefly described as the matrix \mathbf{P} of relative frequencies with two neighbouring pixels separated by an offset $(\Delta x, \Delta y)$ in the image \mathbf{I} of size $m \times n$, one with gray level i and the other with gray level j

- d) **Entropy:** Randomness of the image texture (intensity distribution) can be easily captured by entropy. A homogeneous image will result in a lower entropy value, while an inhomogeneous (heterogeneous) region will result in a higher entropy value.

$$Entropy = - \sum_{i,j} P_{d,\theta}(i, j) \log P_{d,\theta}(i, j) \tag{18}$$

Together all these four features provide high discriminative power to distinguish 8 different kind of saliency maps of an image. Second order statistic based features were built from GLCM matrix with $d = 1, 2$ and $\theta = \{0^0, 45^0, 90^0, 135^0\}$. We have eight scores for each feature corresponding to each saliency map. In case of entropy and contrast, ranks are assigned to all the saliency maps in ascending order, i.e. the lower score is given the numerically lower rank. While in case of homogeneity and energy, ranks are assigned to all the saliency maps in descending order i.e. the higher score is given the numerically lower rank. If two or more saliency maps have the equal scores for a particular statistical feature then we average the ranks for the tied values. For each saliency map, we take the average of total ranks assigned to that saliency map. In this way we have in total 8 ranks corresponding to 8 saliency maps. We employ the average rank of a saliency map as the criterion to determine the optimal scale

$$b_p = \{R(S_j)\} \tag{19}$$


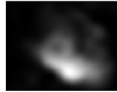
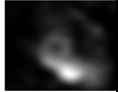
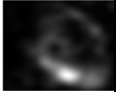
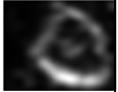
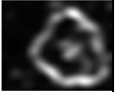
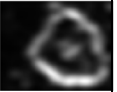
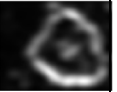
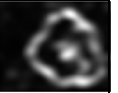

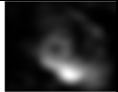



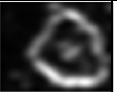




























































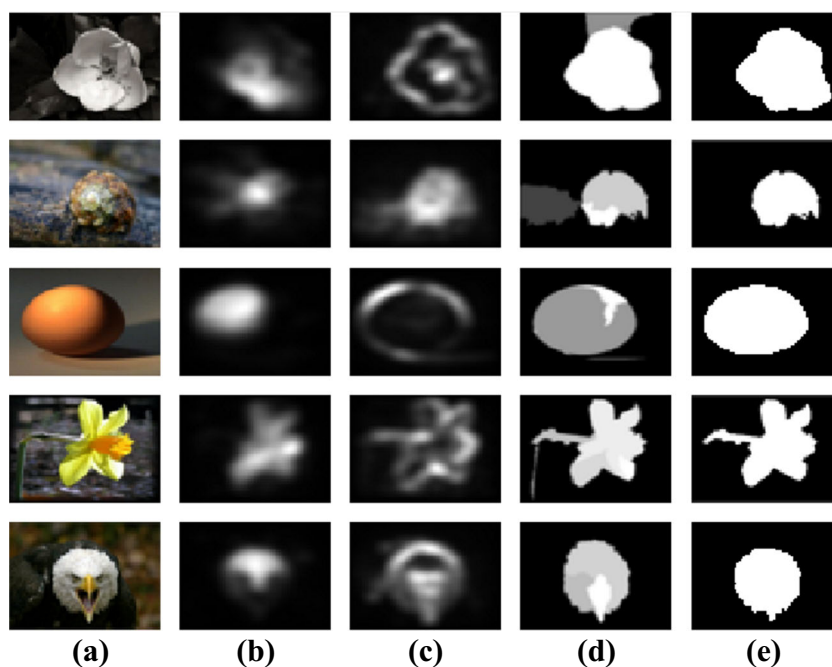
| Image | Saliency Maps at different scales | | | | | | | |
|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |
| | Scale 1 | Scale 2 | Scale 3 | Scale 4 | Scale 5 | Scale 6 | Scale 7 | Scale 8 |
| Best scale selection using Li et al. (2013) |  |  |  |  |  |  |  |  |
| | 11.6968 | 12.1669 | 13.9718 | 15.7597 | 15.2519 | 13.8142 | 13.1855 | 13.1215 |
| Best scale selection using our proposed criterion |  |  |  |  |  |  |  |  |
| | 215 | 206 | 224 | 167 | 119 | 101 | 77 | 43 |
|  |  |  |  |  |  |  |  |  |
| | Scale 1 | Scale 2 | Scale 3 | Scale 4 | Scale 5 | Scale 6 | Scale 7 | Scale 8 |
| Best scale selection using Li et al. (2013) |  |  |  |  |  |  |  |  |
| | 13.0428 | 11.7005 | 9.9992 | 10.1096 | 11.8584 | 12.9724 | 13.8158 | 14.2973 |
| Best scale selection using our proposed criterion |  |  |  |  |  |  |  |  |
| | 256 | 224 | 180 | 168 | 129 | 95 | 68 | 32 |
|  |  |  |  |  |  |  |  |  |
| | Scale 1 | Scale 2 | Scale 3 | Scale 4 | Scale 5 | Scale 6 | Scale 7 | Scale 8 |
| Best scale selection using Li et al. (2013) |  |  |  |  |  |  |  |  |
| | 13.2977 | 12.4085 | 13.0981 | 14.6748 | 15.3098 | 15.3474 | 15.5848 | 15.681 |
| Best scale selection using our proposed criterion |  |  |  |  |  |  |  |  |
| | 222 | 146 | 177 | 227 | 160 | 106 | 57 | 57 |

Fig. 3 Average GLCM rank using our proposed method and entropy value using Li et al. [63] criterion corresponding to eight different saliency maps

Fig. 4 **a** original image **b** Best saliency maps chosen using entropy criterion **c** Best saliency map chosen using GLCM criterion **d** Final saliency map using our proposed method **e** Ground truth



where $R(S_j) = R_{contrast}(S_j) + R_{homogeneity}(S_j) + R_{energy}(S_j) + R_{entropy}(S_j)$

$R_{homogeneity}(S_j)$ is the rank of the saliency map S_j corresponding to *homogeneity* score, $R_{contrast}(S_j)$ is the rank of the saliency map S_j corresponding to *contrast* score, $R_{energy}(S_j)$ is the rank of the saliency map S_j corresponding to *energy* score, and $R_{entropy}(S_j)$ is the rank of the saliency map S_j corresponding to *entropy* score.

We choose the saliency map with the lowest average rank. In Fig. 3, we have shown eight saliency maps with their average GLCM rank and entropy value calculated using Li et al. [63] criterion. For both the proposed method and Li et al. [63] method, the best saliency map is chosen corresponding to the best score, shown in red. It can be clearly observed from Fig. 3 that our proposed method choose better saliency map in comparison to Li et al. [63] method. The proposed method shows better result by taking second order statistical properties of the saliency maps into consideration using GLCM.

3.3 Generation of final saliency map

In this paper, we combined the spatial saliency information obtained from Section 3.2 and segmentation in Section 3.1 to obtain the rough saliency map with accurate object silhouettes. We have used HFT coefficients to locate the salient object in an image while superpixel segmentation is utilized to improve the object contours. Given a segmented region $R_p, p = 1, \dots, k$, where k is the number of segmented regions, the average intensity of each region R_p , is computed based on the corresponding HFT coefficients of

the region in the saliency map obtained in Section 3.2. Each pixel $x \in R_p$, is assigned with the average intensity value v

$$v = \sum_{i=1}^{|R_p|} v_i / |R_p|, \quad \forall x \in R_p, p = 1, \dots, k \quad (20)$$

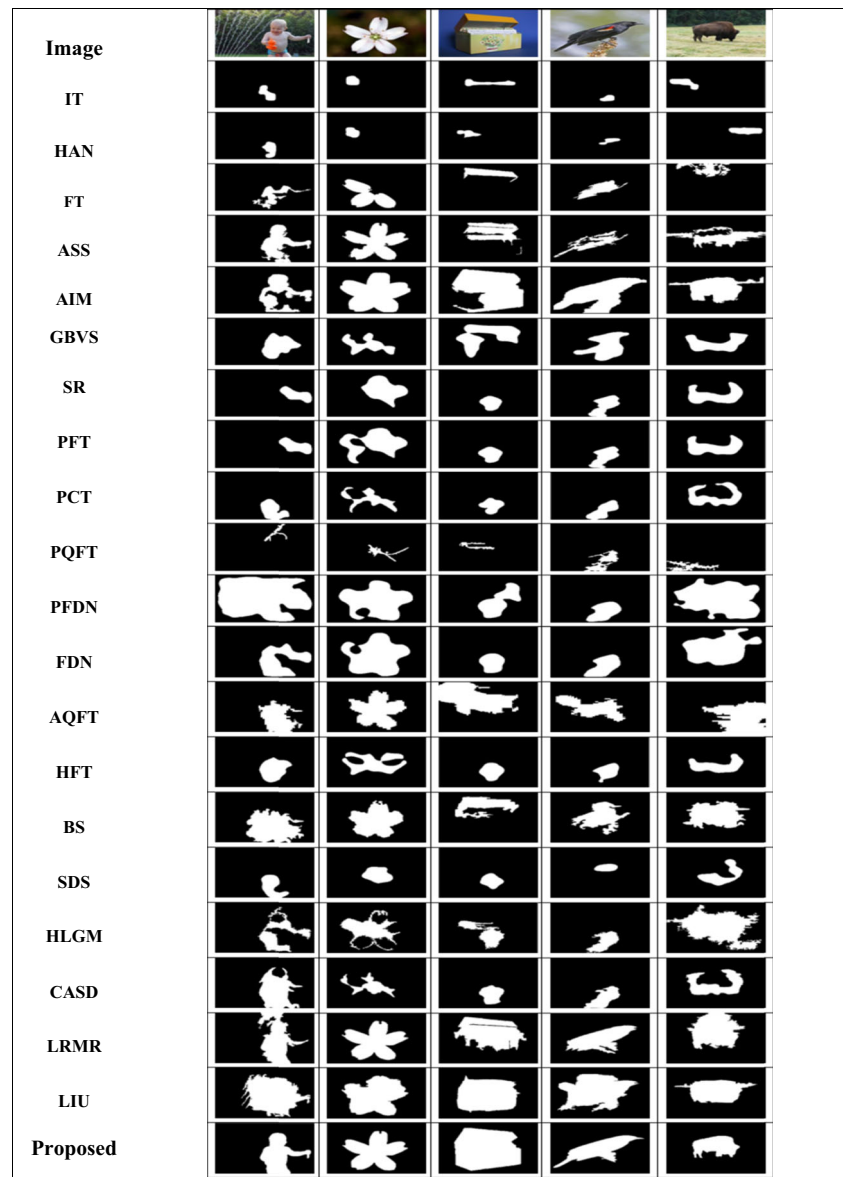
where v_i is the intensity value of the i^{th} pixel x_i . $|R_p|$ is the cardinality of pixels in R_p . Lastly, a final saliency map is obtained by clearly separating the foreground and background of the rough saliency map by using mean HFT intensity value of the entire image as a threshold. If the pixel saliency value is greater than or equal to the threshold value than the pixel is considered to be salient, otherwise if the pixel saliency value is less than the threshold then it is assigned a value of 0. It can be observed in the saliency maps that the boundary of the salient object is improved by incorporating segmentation results.

Original image, best saliency map chosen using Li et al. [63] criterion, best saliency map selected using GLCM criterion, final saliency map using the proposed method, and ground truth are shown in columns a, b, c, d, and e respectively of Fig. 4. It can be easily seen that the best saliency maps chosen by our proposed GLCM based criterion capture more information than the best saliency maps chosen using Li et al. [63] criterion.

4 Experimental setup and results

Experiments are carried out on six datasets to check the robustness and efficacy of our suggested approach (BHGT). BHGT is compared both qualitatively and quantitatively

Fig. 5 Qualitative comparison of the proposed model with existing twenty state-of-the-art models



against twenty popular state-of-the-art models such as IT [22], FT [13], ASS [57], AIM [47], GBVS [48], SR [49], PFT [52], PCT [53], PQFT [58], PFDN [56], FDN [54, 55], AQFT [61], HFT [63], BS [62], SDS [90], CASD [59], LRMR [60], HAN [46], LIU [50, 51] and HLGM [65]. All the experiments are carried out using Windows 7 environment over Intel (R) Xeon (R) processor with a speed of 2.27 GHz and 4GB RAM. The six datasets used for experiments are described below:

A Salient Object Datasets

- a) **MSRA SOD**: Microsoft Research Asia Salient Object Database¹ (MSRA SOD) image set B contains 5000 images, along with ground truth of each

image in the form of a rectangle bounding the salient object.

- b) **ASD**: Achanta Saliency Database (ASD²) contains 1000 images chosen from 5000 images of MSRA-B dataset, along with ground truth of each image in the form of a binary mask. In research work [13], object-contour based ground truth dataset is preferred over rectangle based ground truth as a rectangle may include many objects.
- c) **SAA_GT**: We derived a new ground truth based dataset called the SAA_GT³ [65] which contains all the 5000 images of MSRA-B dataset, along with

¹<http://www.research.microsoft.com/enus/um/people/jiansun/salientobject/salient.object.htm>.

²http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/GroundTruth/binarymasks.zip.

³E-mail at “rinki.arya89@gmail.com” or “navjot.singh.09@gmail.com”.

Table 1 Quantitative comparison for MSRA-B, ASD, and SAA_GT datasets

| Model | MSRA-B | | ASD | | SAA_GT | |
|-------|--------------|--------------|--------------|--------------|--------------|--------------|
| | F-measure | AUC | F-measure | AUC | F-measure | AUC |
| IT | 0.642 | 0.663 | 0.614 | 0.529 | 0.575 | 0.590 |
| AIM | 0.745 | 0.705 | 0.659 | 0.631 | 0.628 | 0.673 |
| SR | 0.622 | 0.659 | 0.469 | 0.505 | 0.456 | 0.581 |
| GBVS | 0.742 | 0.698 | 0.650 | 0.579 | 0.634 | 0.636 |
| PFT | 0.625 | 0.706 | 0.458 | 0.731 | 0.454 | 0.723 |
| PCT | 0.671 | 0.737 | 0.489 | 0.777 | 0.472 | 0.746 |
| FT | 0.638 | 0.669 | 0.603 | 0.625 | 0.628 | 0.648 |
| FDN | 0.763 | 0.782 | 0.678 | 0.826 | 0.643 | 0.795 |
| PQFT | 0.377 | 0.540 | 0.201 | 0.538 | 0.195 | 0.530 |
| PFDN | 0.728 | 0.830 | 0.646 | 0.856 | 0.628 | 0.833 |
| AQFT | 0.673 | 0.728 | 0.637 | 0.807 | 0.568 | 0.764 |
| ASS | 0.743 | 0.698 | 0.652 | 0.630 | 0.634 | 0.664 |
| BS | 0.775 | 0.849 | 0.752 | 0.853 | 0.699 | 0.834 |
| HFT | 0.702 | 0.858 | 0.633 | 0.827 | 0.594 | 0.873 |
| SDS | 0.516 | 0.734 | 0.443 | 0.779 | 0.400 | 0.753 |
| HLGM | 0.770 | 0.860 | 0.693 | 0.848 | 0.646 | 0.818 |
| CASD | 0.737 | 0.776 | 0.737 | 0.705 | 0.702 | 0.741 |
| LRMR | 0.792 | 0.783 | 0.799 | 0.713 | 0.752 | 0.753 |
| HAN | 0.718 | 0.675 | 0.668 | 0.546 | 0.605 | 0.601 |
| LIU | 0.767 | 0.802 | 0.795 | 0.733 | 0.824 | 0.767 |
| BHGT | 0.818 | 0.873 | 0.830 | 0.892 | 0.837 | 0.878 |

ground truth of each image in the form of a binary mask.

- d) **SOD**: SOD⁴ a collection of 500 images of Berkeley Segmentation Dataset (BSD), where salient object boundaries are annotated by seven subjects. The images are more complex than the first three datasets making it difficult for the models to produce convincing results.
- e) **SEDI**: SEDI⁵ denoted as “single-object database” is a collection of 100 images, containing only one object, which are annotated by 3 subjects.
- f) **SED2**: SED2⁶ denoted as “two-object database” is also a collection of 100 images, containing two objects, which are annotated by 3 subjects.

All the images are of size 400 × 300 or 300 × 400 having intensity values in [0, 255].

Evaluation measures ROC score (or AUC (area under the ROC curve)), F- Measure, and computation time are adopted to measure the performance of the proposed model (BHGT) and other twenty state-of-the-art models on six

datasets. The outcome of the salient object detection procedure is a saliency map. A suitable threshold [65, 91] is applied on the saliency map to generate the attention mask **R**. The obtained attention mask is used to compute the precision, recall, and F-measure. Using the ground truth **G** and the detection result **R**, F- Measure is calculated as

$$F_{\beta} = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{21}$$

where β is chosen to be 1. Precision and recall are computed as:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

TP (true positives) is the number of salient pixels that are detected as salient pixels. FP (false positives) is the number of background pixels that are detected as salient pixels. FN (false negatives) is the number of salient pixels that are detected as background pixels.

$$TP = \sum_{G(x,y)=1} R(x,y)$$

$$FP = \sum_{G(x,y)=0} R(x,y) \quad FN = \sum_{R(x,y)=0} G(x,y) \tag{22}$$

Some of the models are better in terms of precision while others excel in terms of recall. A model is considered to

Table 2 Quantitative comparison for SOD, SED1, and SED2 datasets

| Model | SOD | | SED1 | | SED2 | |
|-------|--------------|--------------|--------------|--------------|--------------|--------------|
| | F-measure | AUC | F-measure | AUC | F-measure | AUC |
| IT | 0.252 | 0.550 | 0.248 | 0.623 | 0.413 | 0.601 |
| AIM | 0.500 | 0.796 | 0.657 | 0.880 | 0.597 | 0.861 |
| SR | 0.395 | 0.732 | 0.454 | 0.780 | 0.476 | 0.796 |
| GBVS | 0.550 | 0.813 | 0.642 | 0.868 | 0.570 | 0.821 |
| PFT | 0.366 | 0.714 | 0.422 | 0.753 | 0.467 | 0.794 |
| PCT | 0.413 | 0.736 | 0.475 | 0.807 | 0.473 | 0.781 |
| FT | 0.402 | 0.595 | 0.471 | 0.650 | 0.649 | 0.676 |
| FDN | 0.511 | 0.775 | 0.647 | 0.838 | 0.605 | 0.795 |
| PQFT | 0.397 | 0.729 | 0.439 | 0.781 | 0.474 | 0.771 |
| PFDN | 0.513 | 0.801 | 0.693 | 0.873 | 0.565 | 0.767 |
| AQFT | 0.446 | 0.702 | 0.516 | 0.733 | 0.589 | 0.729 |
| ASS | 0.470 | 0.790 | 0.582 | 0.840 | 0.663 | 0.797 |
| BS | 0.560 | 0.799 | 0.656 | 0.829 | 0.593 | 0.766 |
| HFT | 0.508 | 0.803 | 0.581 | 0.879 | 0.521 | 0.731 |
| SDS | 0.368 | 0.725 | 0.472 | 0.817 | 0.337 | 0.683 |
| HLGM | 0.511 | 0.773 | 0.619 | 0.829 | 0.653 | 0.838 |
| CASD | 0.505 | 0.791 | 0.566 | 0.833 | 0.555 | 0.813 |
| LRMR | 0.564 | 0.794 | 0.710 | 0.860 | 0.676 | 0.814 |
| HAN | 0.320 | 0.584 | 0.412 | 0.654 | 0.512 | 0.648 |
| LIU | 0.538 | 0.796 | 0.681 | 0.868 | 0.561 | 0.812 |
| BHGT | 0.587 | 0.815 | 0.794 | 0.893 | 0.837 | 0.853 |

⁴<http://elderlab.yorku.ca/~vida/SOD/>.

⁵http://www.wisdom.weizmann.ac.il/~vision/Seg_Evaluation_DB.

⁶http://www.wisdom.weizmann.ac.il/~vision/Seg_Evaluation_DB.

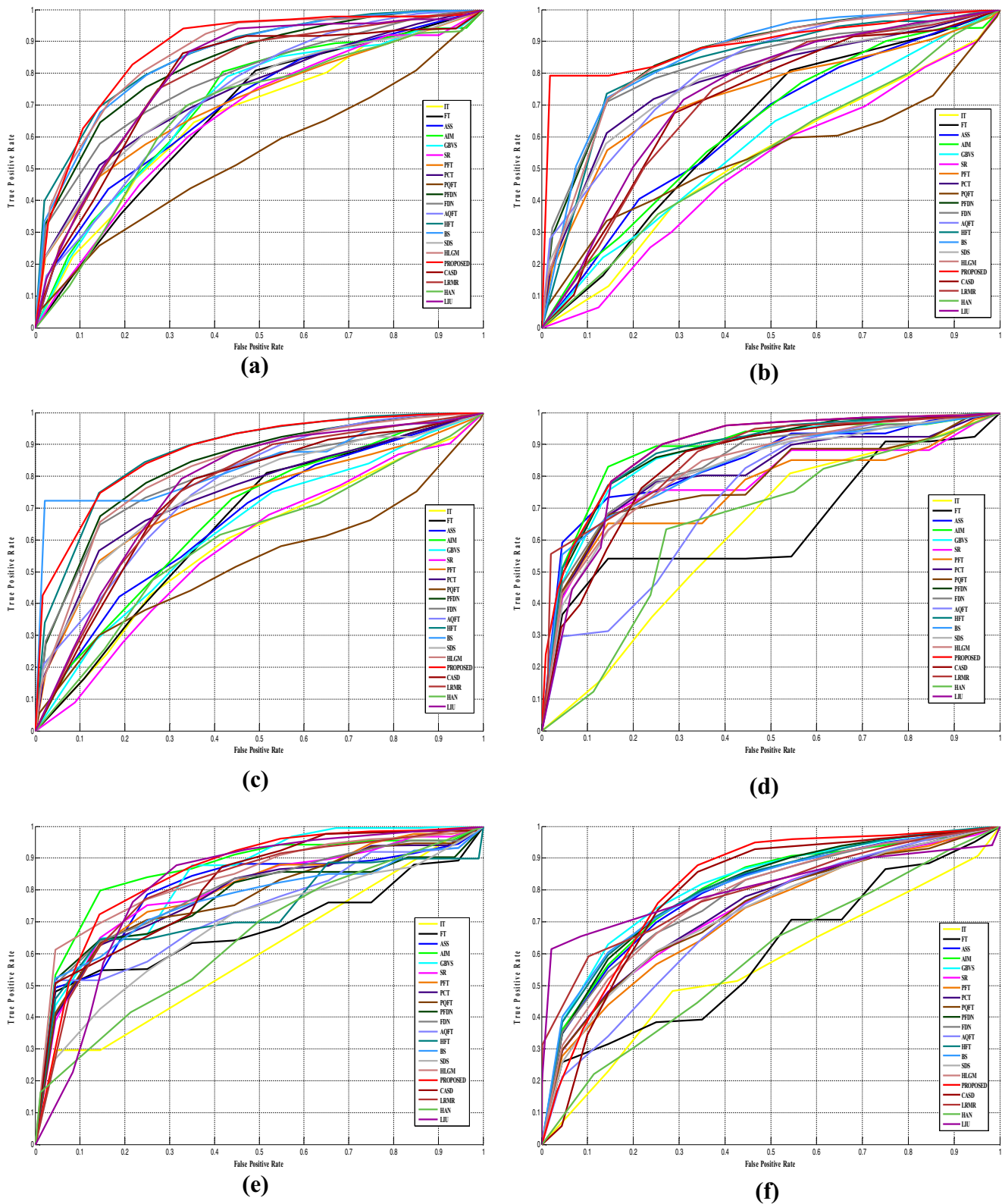


Fig. 6 ROC for the three datasets **a** MSRA-B **b** ASD **c** SAA_GT **d** SED1 **e** SED2 **f** SOD

be good if it shows better performance in terms of both precision and recall. F-measure is the weighted harmonic

mean of precision and recall. AUC is computed by drawing a receiver operator characteristic (ROC) curve. ROC curve

is plotted between the true positive rate (TPR) and the false positive rate (FPR). TPR and FPR are given by

$$TPR = \frac{TP}{\sum_{(x,y)} \mathbf{G}(x, y)}$$

$$FPR = \frac{FP}{W \times H - \sum_{(x,y)} \mathbf{G}(x, y)}$$

where W and H represents the width and height of the image respectively. The saliency maps corresponding to the proposed model as well as state-of-the-art models are first normalized between $[0, 255]$. Then 256 thresholds are chosen one by one and the values of TPR and FPR are computed and the ROC curve is plotted and finally area under the curve (AUC) is calculated.

4.1 Qualitative evaluation

In Fig. 5, we have shown qualitative comparison of our proposed model (BHGT) with twenty state-of-the-art models on five images. These five images are selected from the test data set containing objects with different shape, size, position, type etc. Figure 5 clearly shows that shape information and boundaries are well captured using our proposed model (BHGT) as compared to other state-of-the-art models.

4.2 Quantitative evaluation

The performance of the proposed model (BHGT) and other twenty popular state-of-the-art models is evaluated in terms of F-measure, area under curve (AUC), and computation time. Table 1 shows the quantitative performance of the proposed method in comparison to twenty state-of-the-art methods for MSRA-B, ASD, and SAA_GT datasets. Table 2 shows the quantitative performance of the proposed method in comparison to twenty state-of-the-art methods on SOD, SED1, and SED2 datasets. The performance of the proposed model (BHGT) against twenty state-of-the-art models can also be seen in the form of ROC curves shown in Fig. 6a, b, c, d, e, and f for MSRA-B, ASD, SAA_GT, SED1, SED2, SOD datasets respectively. It can be clearly seen from the Fig. 6 that the ROC curve of proposed model BHGT, shown in red colour, covers the maximum area under the ROC curve for all the six datasets and hence gives the highest AUC value for all the datasets.

The best results are shown in bold. It can be noted from Tables 1, 2 that our proposed method outperforms all twenty state-of-the-art methods for all the six datasets in terms of

Table 3 Average time per image

| Method Name | Time (in sec) per image |
|-------------|-------------------------|
| IT | 1.70 |
| AIM | 50.8 |
| GBVS | 59.8 |
| SR | 0.02 |
| PFT | 0.018 |
| PCT | 0.06 |
| FT | 0.17 |
| ASS | 0.31 |
| PQFT | 0.18 |
| PFDN | 0.25 |
| FDN | 0.29 |
| AQFT | 21.9 |
| HFT | 0.69 |
| HLGM | 0.26 |
| SDS | 0.17 |
| CASD | 124.0 |
| LRMR | 71.9 |
| HAN | 12.4 |
| LIU | 25.7 |
| BS | 7.57 |
| BHGT | 5.98 |

F-measure and AUC. Higher values for both precision and recall are obtained for our proposed method while other models show high precision value but low recall value or vice-versa.

Following conclusions can be derived from Table 3:

1. Guo et al. model [52] requires the least computational time. But this method does not show good results for F-measure and AUC.
2. Goferman et al. [59] model requires maximum amount of time.
3. Hou and Zhang [49], Itti et al. [22], Guo et al. [52], Yu et al. [53], Achanta et al. [13], Bian and Zhang [56], Guo and Zhang [58], Li et al. [63], Arya et al. [65], Li et al. [64], and Achanta and Susstrunk [57] models need low computation time as compared to the proposed model but these methods perform poor in terms of F-measure and AUC in comparison to our proposed model.
4. Our proposed model takes less computation time as compared to Bruce and Tsotsos [47], Fang et al. [61], Goferman et al. [59], Shen et al. [60], Han et al. [92], Liu et al. [50] and Harel et al. [48] models.

5 Conclusion and future work

In this paper, we propose a bottom-up salient object detection model (BHGT) which takes the advantage of both spatial domain and frequency domain. The key idea of the proposed method is to detect a saliency map which can uniformly highlight the most salient object in the given image with accurate shape information. To accurately capture shape information, we need a good segmentation algorithm. But different segmentation algorithms produce different segmentation results under different parameters. Therefore, to fuse complementary information coming from existing segmentation algorithms, we use an improved bipartite graph partitioning based segmentation algorithm which integrates a large number of superpixels generated from different existing segmentation algorithms under different parameters thereby giving better segmentation results. To uniformly highlight the salient object in the segmented image, we take advantage of the hypercomplex image representation to combine multiple features in order to get better performance. The saliency of each segmented region is obtained by reconstructing the image using the original phase and the amplitude spectrum, filtered at a scale selected by minimizing our proposed average rank GLCM criterion. To determine the optimal scale, we choose four statistical features namely, angular second moment, entropy, inverse difference moment, and contrast, which are complementary to each other. Together all these four features provide high discriminative power to distinguish two different spatial structures by capturing spatial relationships of the image pixels. Finally a saliency map is obtained by taking average of the HFT coefficients of each region in the segmented image and taking the mean HFT intensity value of the entire image as a threshold to clearly separate salient object from the background in the final saliency map. The performance of the proposed model (BHGT) is evaluated in terms of F-measure, AUC and computation time on six publicly available image datasets. Experimental results demonstrated that the BHGT outperformed the existing state-of-the-art methods in terms of both qualitatively and quantitatively (F -measure and AUC) on all the six datasets. In the proposed model, we are choosing only one saliency map corresponding to the optimal scale but other abandoned maps may also contain meaningful saliency information. In our future work, we intend to incorporate the meaningful saliency information from different abandoned saliency maps in the determination of final saliency map. Work also needs to be extended to detect any number of salient objects or no salient object at all. It would also be interesting to include the top-down information in future to improve the performance.

Acknowledgments The first author expresses her sincere and reverential gratitude to University Grant Commission (UGC), India, for the obtained financial support during this research.

References

1. Borji A, Itti L (2013) State-of-the-art in visual attention modeling. *IEEE Trans Pattern Anal Mach Intell* 35(1):185–207
2. Borji A, Cheng M-M, Jiang H, Li J (2015) Salient object detection: a benchmark. *IEEE Trans Image Process* 24(12):5706–5722
3. Borji A, Cheng Ming-Ming, Jiang H, Li J (2014) Salient object detection: a survey. arXiv:1411.5878
4. Hadizadeh H, Bajic IV (2014) Saliency-aware video compression. *IEEE Trans Image Process* 23(1):19–33
5. Wang Y-S, Tai C-L, Sorkine O, Lee T-Y (2008) Optimized scale-and-stretch for image resizing. *ACM Trans Graph (TOG)* 27(5):118–118
6. Marchesotti L, Cifarelli C, Csurka G (2009) A framework for visual saliency detection with applications to image thumbnailing. In: *IEEE 12th international conference on computer vision*, pp 2232–2239
7. Boykov YY, Jolly Marie-Pierre (2001) Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In: *Eighth IEEE international conference on computer vision*, vol 1, pp 105–112
8. Chul B, Ko B, Nam J-Y (2006) Object-of-interest image segmentation based on human attention and semantic region clustering. *J Opt Soc Am A (JOSA A)* 23(10):2462–2470
9. Gopalakrishnan V, Hu Y, Rajan D (2010) Random walks on graphs for salient object detection in images. *IEEE Trans Image Process* 19(12):3232–3242
10. Ueli, Walther D, Koch C, Perona PR (2004) Is bottom-up attention useful for object recognition? In: *IEEE computer society conference on computer vision and pattern recognition*, vol 2, pp II–37
11. Navalpakkam V, Itti L (2006) An integrated model of top-down and bottom-up attention for optimizing detection speed. In: *IEEE computer society conference on computer vision and pattern recognition*, vol 2, pp 2049–2056
12. Achanta R, Ssstrunk S (2009) Saliency detection for content-aware image resizing. In: *IEEE international conference on image processing (ICIP)*, pp 1005–1008
13. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: *IEEE conference on computer vision and pattern recognition*, pp 1597–1604
14. Alpert S, Galun M, Brandt A, Basri R (2012) Image segmentation by probabilistic bottom-up aggregation and cue integration. *IEEE Trans Pattern Anal Mach Intell* 34(2):315–327
15. Jian MW, Dong JY, Ma J (2011) Image retrieval using wavelet-based salient regions. *The Imaging Science Journal* 59(4):219–231
16. Huang K, Tao D, Yuan Y, Li X, Tan T (2011) Biologically inspired features for scene classification in video surveillance. *IEEE Trans Syst Man Cybern Part B Cybern* 41(1):307–313
17. Park J, Lee J-Y, Tai Y-W, Kweon IS (2012) Modeling photo composition and its application to photo re-arrangement. In: *IEEE international conference on image processing (ICIP)*, pp 2741–2744
18. Ninassi A, Le Meur O, Le Callet P, Barba D (2007) Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In: *IEEE international conference on image processing*, vol 2, pp II–169

19. Li Z, Itti L (2011) Saliency and gist features for target detection in satellite images. *IEEE Trans Image Process* 20(7):2017–2029
20. Santella A, Agrawala M, DeCarlo D, Salesin D, Cohen M (2006) Gaze-based interaction for semi-automatic photo cropping. In: SIGCHI conference on human factors in computing systems, pp 771–780
21. Chen L-Q et al. (2003) A visual attention model for adapting images on small displays. *Multimedia Systems* 9(4):353–364
22. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Mach Intell* 20(11):1254–1259
23. Itti L (2000) Models of bottom-up and top-down visual attention. California Institute of Technology, Doctoral dissertation
24. Karssemeijer N, te Brake GM (1996) Detection of stellate distortions in mammograms. *IEEE Trans Med Imaging* 15(5):611–619
25. Rother C, Bordeaux L, Hamadi Y, Blake A (2006) Autocollage. *ACM Trans Graph (TOG)* 25(3):847–852
26. Gasparini F, Corchs S, Schettini R (2007) Low-quality image enhancement using visual attention. *Opt Eng* 46(4):040502–040502
27. Kim J, Han D, Tai Y-W, Kim J (2016) Salient region detection via high-dimensional color transform and local spatial support. *IEEE Trans Image Process* 25(1):9–23
28. Zhu L, Klein DA, Frintrop S, Cao Z, Cremers AB (2014) A multisize superpixel approach for salient object detection based on multivariate normal distribution estimation. *IEEE Trans Image Process* 23(12):5094–5107
29. Itti L, Koch C (2001) Computational modelling of visual attention. *Nat Rev Neurosci* 2(3):194–203
30. Borji A (2012) Boosting bottom-up and top-down visual features for saliency estimation. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, pp 438–445
31. Cheung Y-M, Peng Q (2012) Salient region detection using local and global saliency. In: *21st international conference on pattern recognition (ICPR)*, pp 210–213
32. Jia C, Hou F, Duan L (2013) Visual saliency based on local and global features in the spatial domain. *Int J Comput Sci* 10(3):713–719
33. Borji A, Itti L (2012) Exploiting local and global patch rarities for saliency detection. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, pp 478–485
34. Zhang L, Yang L, Luo T (2016) Unified saliency detection model using color and texture features. *Plos one* 11(2):e0149328
35. Itti L (2005) Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Vis Cogn* 12(6):1093–1123
36. Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* 3(3):201–215
37. Wang K, Lin L, Lu J, Li C, Shi K (2015) PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence. *IEEE Trans Image Process* 24(10):3019–3033
38. Cheng M, Mitra NJ, Huang X, Torr PHS, Hu S (2015) Global contrast based salient region detection. *IEEE Trans Pattern Anal Mach Intell* 37(3):569–582
39. Naqvi SS, Browne WN, Hollitt C (2016) Salient object detection via spectral matting. *Pattern Recogn* 51:209–224
40. Huang X, Su Y, Liu Y (2016) Iteratively parsing contour fragments for object detection. *Neurocomputing* 175:585–598
41. Huo L, Jiao L, Wang S, Yang S (2016) Object-level saliency detection with color attributes. *Pattern Recogn* 49:162–173
42. Levine MD, An X, He H (2011) Saliency detection based on frequency and spatial domain analysis. In: *British machine vision conference (BMVC)*, pp 86.1–86.11
43. Li Z, Wu X-M, Chang S-F (2012) Segmentation using superpixels: a bipartite graph partitioning approach. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, pp 789–796
44. Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. *IEEE Trans Pattern Anal Mach Intell* 24(5):603–619
45. Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. *Int J Comput Vis* 59(2):167–181
46. Han J, Ngan KN, Li M, Zhang H-J (2006) Unsupervised extraction of visual attention objects in color images. *IEEE Trans Circuits Syst Video Technol* 16(1):141–145
47. Bruce N, Tsotsos J (2005) Saliency based on information maximization. *Adv Neural Inf Proces Syst*:155–162
48. Harel J, Koch C, Perona P (2006) Graph-based visual saliency. In: *Advances in neural information processing systems*, pp 545–552
49. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: *IEEE conference on computer vision and pattern recognition*, pp 1–8
50. Liu T et al (2011) Learning to detect a salient object. *IEEE Trans Pattern Anal Mach Intell* 33(2):353–367
51. Liu T et al (2007) Learning to detect a salient object. In: *IEEE conference on computer vision and pattern recognition*, pp 1–8
52. Guo C, Qi Ma, Zhang L (2008) Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, pp 1–8
53. Yu Y, Wang B, Zhang L (2009) Pulse discrete cosine transform for saliency-based visual attention. In: *IEEE 8th international conference on development and learning*, pp 1–6
54. Bian P, Zhang L (2008) Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In: *International conference on neural information processing*, pp 251–258
55. Bian P, Zhang L (2010) Visual saliency: a biologically plausible contourlet-like frequency domain approach. *Cogn Neurodyn* 4(3):189–198
56. Bian P, Zhang L (2010) Piecewise frequency domain visual saliency detection. In: *IEEE third international conference on information and computing (ICIC)*, vol 3, pp 269–272
57. Achanta R, Susstrunk S (2010) Saliency detection using maximum symmetric surround. In: *17th IEEE international conference on image processing (ICIP)*, pp 2653–2656
58. Guo C, Zhang L (2010) A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans Image Process* 19(1):185–198
59. Goferman S, Zelnik-Manor L, Tal A (2012) Context-aware saliency detection. *IEEE Trans Pattern Anal Mach Intell* 34(10):1915–1926
60. Shen X, Ying Wu (2012) A unified approach to salient object detection via low rank matrix recovery. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, pp 853–860

61. Fang Y et al. (2012) Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum. *IEEE Trans Multimedia* 14(1):187–198
62. Fang Y, Chen Z, Lin W, Lin C-W (2012) Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Trans Image Process* 21(9):3888–3901
63. Li J, Levine MD, An X, Xu X, He H (2013) Visual saliency based on scale-space analysis in the frequency domain. *IEEE Trans Pattern Anal Mach Intell* 35(4):996–1010
64. Li J, Duan L-Y, Chen X, Huang T, Tian Y (2015) Finding the secret of image saliency in the frequency domain. *IEEE Trans Pattern Anal Mach Intell* 37(12):2428–2440
65. Arya R, Singh N, Agrawal RK (2015) A novel hybrid approach for salient object detection using local and global saliency in frequency domain. *Multimedia Tools and Applications*: 1–21
66. Huaizu, Yuan, Zejian, Cheng, Jiang M-M, Gong Y, Zheng N, Wang J (2014) Salient object detection: a discriminative regional feature integration approach. arXiv:1410.5926
67. Zou W, Komodakis N (2015) HARK: Hierarchy-associated rich features for salient object detection. In: *IEEE international conference on computer vision*, pp 406–414
68. Sun J, Lu H, Liu X (2015) Saliency region detection based on Markov absorption probabilities. *IEEE Trans Image Process* 24(5):1639–1649
69. Perazzi F, Philipp, Krahenbuhl, Yael, Pritch, Hornung A (2012) Saliency filters: contrast based filtering for salient region detection. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, pp 733–740
70. Yan Q, Xu L, Jianping, Jia JS (2013) Hierarchical saliency detection. In: *IEEE conference on computer vision and pattern recognition*, pp 1155–1162
71. Zhi, Zou, Wenbin, Le Meur, Liu O (2014) Saliency tree: a novel saliency detection framework. *IEEE Trans Image Process* 23(5):1937–1952
72. Zhu W, Liang S, Wei Y, Sun J (2014) Saliency optimization from robust background detection. In: *IEEE conference on computer vision and pattern recognition*, pp 2814–2821
73. Li X, Lu H, Zhang L, Ruan X, Yang M-H (2013) Saliency detection via dense and sparse reconstruction. In: *IEEE international conference on computer vision*, pp 2976–2983
74. Chang K-Y, Liu T-L, Chen H-T, Lai S-H (2011) Fusing generic objectness and visual saliency for salient object detection. In: *IEEE international conference on computer vision (ICCV)*, pp 914–921
75. Jiang H et al (2011) Automatic salient object segmentation based on context and shape prior. In: *BMVC*, vol 6, p 9
76. Yang C, Zhang L, Lu H, Ruan X, Yang M-H (2013) Saliency detection via graph-based manifold ranking. In: *IEEE conference on computer vision and pattern recognition*, pp 3166–3173
77. Margolin R, Tal A, Zelnik-Manor L (2013) What makes a patch distinct? In: *IEEE conference on computer vision and pattern recognition*, pp 1139–1146
78. Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Transaction on Pattern and Machine Intelligence* 22(8):888–905
79. Zhang L, Lin W (2013) *Selective visual attention: computational models and applications*. Wiley
80. Castleman Kenneth R (1996) *Digital image processing*. Prentice Hall Press, Upper Saddle River
81. Sangwine SJ (1996) Fourier transforms of colour images using quaternion or hypercomplex, numbers. *Electron Lett* 32(21):1979–1980
82. Ell TA (1992) *Hypercomplex spectral transformations*. University of Minnesota
83. Ell TA (1993) Quaternion-fourier transforms for analysis of two-dimensional linear time-invariant partial differential systems. In: *32nd IEEE conference on decision and control*, pp 1830–1841
84. Pei S-C, Ding J-J, Chang J-H (2001) Efficient implementation of quaternion Fourier transform, convolution, and correlation by 2-D complex FFT. *IEEE Trans Signal Process* 49(11):2783–2797
85. Sangwine SJ, Ell TA (2000) The discrete Fourier transform of a colour image. *Image Processing II Mathematical Methods, Algorithms and Applications*:430–441
86. Ell TA, Sangwine SJ (2007) Hypercomplex Fourier transforms of color images. *IEEE Trans Image Process* 16(1):22–35
87. Itti L, Baldi PF (2005) Bayesian surprise attracts human attention. *Adv Neural Inf Proces Syst*:547–554
88. Engel S, Zhang X, Wandell B (1997) Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature* 388(6637):68–71
89. Haralick RM, Shanmugam K, Dinstein IH (1973) Textural features for image classification. *IEEE Trans Syst Man Cybern*:610–621
90. Li J, Duan L-Y, Chen X, Huang T, Tian Y (2015) Finding the secret of image saliency in the frequency domain. *IEEE Trans Pattern Anal Mach Intell* 37(12):2428–2440
91. Singh N, Arya R, Agrawal RK (2014) A novel approach to combine features for salient object detection using constrained particle swarm optimization. *Pattern Recogn* 47(4):1731–1739
92. Han J, Ngan KN, Li M, Zhang H-J (2006) Unsupervised extraction of visual attention objects in color images. *IEEE Trans Circuits Syst Video Technol* 16(1):141–145



Ms. Rinki Arya received M.Tech (Computer Science and Technology) from Jawaharlal Nehru University, New Delhi. Currently she is pursuing Ph.D. (Computer Vision and Pattern Recognition) from Jawaharlal Nehru University, New Delhi. Her current research areas are computer vision, image processing, object detection, pattern recognition, and feature extraction.



Mr. Navjot Singh obtained M.Tech (Computer Science and Technology) from Jawaharlal Nehru University, New Delhi and Ph.D. (Computer Vision and Pattern Recognition) from Jawaharlal Nehru University, New Delhi. His current research areas are computer vision, image processing, object detection, pattern recognition, and feature extraction and classification.



Prof. R. K. Agrawal obtained MTech. (Computer Application) from Indian Institute of Technology Delhi, New Delhi and PhD (Computational Physics) from University of Delhi, Delhi. Presently, he is working as a Professor at School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. His current research areas are: Classification, feature extraction and selection for pattern recognition problems in domains of image processing, security, and bioinformatics.