

# A novel hybrid approach for salient object detection using local and global saliency in frequency domain

Rinki Arya<sup>1</sup> · Navjot Singh<sup>1</sup> · R. K. Agrawal<sup>1</sup>

Received: 9 September 2014 / Revised: 17 April 2015 / Accepted: 15 June 2015 /

Published online: 15 July 2015

© Springer Science+Business Media New York 2015

**Abstract** In this paper, we introduce a fast and novel biologically plausible frequency domain approach to detect salient object which incorporates both local and global salient features. The proposed approach involves three phases. In the first phase, locally salient features are obtained as suggested in the research work of Bian and Zhang. Globally salient features are computed in the second phase using fast Walsh-Hadamard transform since it is computationally more efficient and faster than fast Fourier transform. Finally the saliency map is generated in terms of linear weighted combination of local and global saliency maps where the weights are determined using entropy measure. The performance is evaluated both qualitatively and quantitatively on two publicly available datasets and one new dataset derived from a publicly available dataset. Experiments show that the proposed model significantly outperforms other relevant existing state-of-the-art methods in both spatial and frequency domain. The proposed method is also computationally less expensive to detect salient object accurately.

**Keywords** Salient object detection · Local saliency · Global saliency · Fast Walsh-Hadamard transform (WHT) · Piecewise Frequency Domain Divisive Normalization (PFDN)

## 1 Introduction

While watching a play in a theatre, usually a sudden spotlight grabs the attention of the audience sitting in the dark. At any point of time during performance, a lot of visual stimuli like actors performing the play, their props and costumes, stage setup, etc. reach human eyes. In spite of this

---

✉ Rinki Arya  
rinki.arya89@gmail.com

Navjot Singh  
navjot.singh.09@gmail.com

R. K. Agrawal  
rkajnu@gmail.com

<sup>1</sup> School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi 110067, India

huge load of visual stimuli the spotlight guides the human gaze by identifying the areas of relevance in a scene. The mechanism of identifying the relevant regions in a given image or scene is called visual attention [6, 30]. These relevant regions or objects are termed as salient objects in the field of computer vision [6, 30]. Detection of these salient objects finds its real time applications in surveillance systems [21], remote sensing [36], image retrieval [3, 20] and object detection and recognition [35, 45]. It is helpful in automatic target detection [28, 30], robotics, image and video compression [28], automatic cropping/centering [41] to display objects on small portable screens [10], medical imaging [35], advertising a design [28], image collection browsing [40], image enhancement [18], video summarization [38] and many more.

Visual Attention is a cognitive process that helps humans and primates to rapidly select the highly relevant information from a scene [7]. This information is then further processed by high-level visual processes such as scene understanding and object detection. It is commonly believed that visual attention is guided by two components: (i) Bottom-up (BU) visual saliency, a data-driven and task independent component based on only low-level and image-based outliers and conspicuities, and (ii) a top-down (TD) component, a volitionally-controlled mechanism that guides attention and gaze in a task-dependent and goal-directed manner, orchestrating the sequential acquisition of information from the visual environment.

When information about specific search target, search task, and particular time or other constraints is not specified to an observer in advance then bottom-up (image-derived) information plays a predominant role in guiding attention toward potential interesting targets [29]. When attention is exploited by salient stimuli, it is considered to be bottom-up, memory-free, and reactive. It depends only on the instantaneous sensory input, without taking into account the internal state of the organism. It is driven by low-level stimulus in the scene. In some cases when backgrounds are highly cluttered, due to deficiency of top-down prior knowledge, bottom-up saliency algorithms usually respond to numerous unrelated low-level visual stimuli (false positives) and thus may miss the objects of interest (false negatives). Most of research works mostly focused on the bottom-up aspect of visual attention. Currently researchers started distinguishing the two very similar terms with the advancement of bottom-up approaches: fixation prediction and salient object detection [8, 11, 35]. The main objective of the fixation prediction models is to find the fixation points in a given scene. Fixation points are those points in the scene or image where human eyes focus if shown for a few seconds. These points are useful in eye movement prediction. The second category of models which are salient object detection models detects the most salient object in an image by drawing accurate silhouettes of the salient object. To draw accurate silhouettes, segmentation of the image into two regions, a salient object and background is needed. Both categories of models construct saliency maps which are useful for different purposes. The other way to guide and improve the attention is to use top-down, memory-dependent, or anticipatory mechanisms. Top-down attention is driven by cognitive factors such as knowledge, expectations and current goals [13]. The top-down methods are task-dependent and the human observation behavior is exploited to achieve specific goals. Top-down models are always integrated with the bottom-up models to generate saliency maps for localizing objects of interest. This bottom-up or top-down visual attention can be modelled in spatial domain and frequency domain to automatically generate the saliency map which encodes visual conspicuity stimulus. In general, spatial domain methods provide higher detection accuracy but take more computation time to obtain features. In literature, research works are suggested to determine features in frequency domain to reduce computation time.

Most of the models in frequency domain focus only on the local saliency while others focus only on the global saliency. However, in order to detect a salient object, both local as well as

global saliency information play a vital role. In literature [5, 12, 31], the research works have utilized fusion of the local and global features obtained in spatial domain to enhance detection accuracy but at the cost of higher computation time. To the best of our knowledge till date, there is no model proposed in literature to detect salient object which utilizes both local and global saliency information in frequency domain. So, in this paper we propose a novel and effective hybrid approach for salient object detection which utilizes both local and global saliency information in frequency domain to reduce computation time without degrading detection accuracy much. Local saliency is computed in terms of PFDN as suggested in the research work of Bian and Zhang [4] and global saliency is determined using fast Walsh - Hadamard transform (WHT) [15, 24, 44]. WHT is less computationally expensive as it takes only binary values  $\pm 1$  and requires only addition and subtraction operations. Finally the saliency map is generated in terms of linear weighted combination of local and global saliency where the weights are determined using entropy measure [26, 39, 43]. To check the efficacy of the proposed hybrid model, experiments are performed on two publicly available datasets and one new derived dataset and performance is compared with existing state-of-the-art methods in literature.

The contribution of this paper is threefold: 1) A fast frequency domain approach for salient object detection is proposed which allows the full use of local and global information for salient object detection unlike recent methods in frequency domain which model saliency either as a global phenomenon or local phenomenon; 2) We employ fast Walsh-Hadamard transform (WHT) to compute global saliency because of its simplicity, efficiency, and speed; 3) We derived object-contour based ground truth dataset to obtain exact shapes of salient objects. The performance is evaluated on this derived dataset to check how well our proposed method satisfies the accurate object shape for all the 5000 images of MSRA SOD image set B rather than using ground truth based on rectangle constraints.

The paper is organized as follows. Section 2 is the description and review of related state-of-the-art methods to detect salient object. In section 3, we present the proposed saliency model (HLGM) based on local and global saliency in frequency domain. The experimental setup and results are presented in section 4. Conclusion and future work are discussed in Section 5.

## 2 Related work

### 2.1 Bottom-up methods

Visual attention [6, 17] is achieved by either a fast bottom-up component or a slow task-dependent top-down component. Most of the researchers focus on computing bottom-up visual attention in spatial domain. Itti et al. [30] suggested a biologically plausible saliency detection approach which generates activation maps by employing the centre-surround operator through a number of scales and finally combines these activation maps into a saliency map based on the early primate visual system. Han et al. [25] proposed a model which uses Markov random field and region growing techniques in combination with the Itti et al.'s model [30] for salient objects segmentation in colour images. Bruce and Tsotsos [9] proposed a neurally plausible bottom-up salient object detection model which works on the principle of information maximization. Achanta et al. [1] proposed a frequency-tuned method for saliency detection. Achanta and Susstrunk [2] proposed a salient region detection approach using

maximum symmetric surround technique by assigning large bandwidth to the filter in the centre of images while small bandwidth at border.

The spatial domain models are generally complex and highly computational which limits their usage in real time applications. To overcome these limitations, researchers employed frequency domain techniques for salient object detection. Recently Hou and Zhang [27] used spectral residual of Fourier transform to detect the salient objects. Guo et al. [22] pointed out that phase spectrum of Fourier transform (PFT) is the most important key to determine the position of salient objects rather than amplitude spectrum and proposed a saliency detection model based on the phase spectrum of the Fourier transform. Guo and Zhang [23] extended PFT model to compute the multi-resolution spatiotemporal saliency map which uses quaternion representation of the image. Yu et al. [46] proposed a salient object detection model which is based on the concept of pulsed discrete cosine transform. Bian and Zhang [4] suggested a frequency domain divisive normalization (FDN) approach for saliency detection using contourlet transform and frequency divisive normalization. FDN exhibits biological plausibility as it utilizes the concept of initial feature extraction and cortical surround inhibition. Bian and Zhang [4] extended FDN by decomposing an image into overlapping local patches and then conducting piecewise FDN (PF DN) on these patches. Recently Fang et al. [14] utilized amplitude spectrum of Quaternion Fourier Transform (AQFT) of different patches of an image to detect salient object. More recently, Li et al. [34] built saliency detection model based on the hypercomplex Fourier transform (HFT). In this method Gaussian functions of different variances are used to filter the log amplitude spectrum while keeping the phase spectrum.

## 2.2 Top-down methods

Top-down approaches are integrated with the bottom-up approaches in order to detect the salient locations. Zhang et al. [47] proposed a Bayesian framework based approach to classify a pixel into salient object or background object by taking position, area and intensity saliencies and a maximum saliency difference technique into consideration. Goferman et al. [19] proposed a context-aware saliency detection algorithm by exploiting four principles: local low level, global, visual organization rules and high-level factors. Liu et al. [37] proposed a supervised approach by incorporating a set of features to depict a salient object at the local, regional and global level. The proposed method consists of two phases. In the first phase the multi-scale contrast, center-surround histogram and color spatial distribution features are extracted from the image. Then in the second phase, conditional random field is employed to linearly combine these features into a saliency map.

## 3 Hybrid approach based on local and global saliency maps

In this paper, we propose a hybrid approach (HLGM) that utilizes benefits of both local and global saliency information in frequency domain for salient object detection. The proposed HLGM model involves three phases. In the first phase, local saliency information is computed using PF DN as suggested in the research work of Bian and Zhang [4]; in the second phase, global saliency information is computed using fast Walsh-Hadamard transform [15, 24, 44]; and finally, the hybrid saliency map is determined in terms of weighted linear combination of the local and global saliency maps where the weights are determined using entropy measure. The extracted hybrid saliency map is used to produce segmentation mask around region of

interest. This segmentation mask around region of interest is also called the attention mask. We will describe the attention mask more in a later section 3.3. Brief details of these three phases are discussed below.

### 3.1 Local saliency computation

Most of the models working in frequency domain are global in nature except some models which compute local saliency information over image patches such as the PFDN model, which shows biologically plausibility in frequency domain [4]. In spectral domain, each feature map can be seen as a sub-band in the frequency spectrum and saliency map is generated through all scales and orientations [4]. So in the first phase of HLGGM, we have utilized PFDN to compute local saliency map. In this phase, the input image is first transformed into LAB colour space and for a given colour channel, the image is decomposed into a number of scales by employing a Laplacian pyramid. Then for a given colour channel and a scale, the input image is partitioned into  $P$  overlapping local patches. For a given colour channel and a scale, Fourier transform of a patch ( $\mathbf{I}_p$ ) is computed as

$$R_p(k) = F_k(\mathbf{I}_p) \quad \forall p = 1, 2, \dots, P \quad (1)$$

where  $F$  represents Fourier transform,  $R_p(k)$  is the Fourier transform coefficient at frequency component  $k$  for patch  $p$ . These Fourier coefficients are grouped using the decomposition scheme shown in Fig. 1 as suggested in the research work of Bian and Zhang [4]. This decomposition scheme separates the amplitude spectrum of the input image into feature maps in four scales with 16, 8, 4, and 1 orientations from the highest scale to the lowest, which result in total 29 sub-bands corresponding to the 29 feature maps.

Then normalization term for  $i^{\text{th}}$  sub-band,  $E_i$  is calculated as

$$E_i = \sqrt{w \sum_{k \in i} \frac{\|R_p(k)\|^2}{N} + \sigma^2} \quad \forall i = 1, 2, \dots, 29 \quad (2)$$

where  $N$  is the number of pixels.  $w$  and  $\sigma^2$  represent constants which are set to 1. After computing the normalization term for each sub-band,  $E_i$  the divisive normalization coefficient of sub-band  $i$  in the frequency domain can be obtained by

$$\hat{R}_p(k)_{k \in i} = \frac{R_p(k)_{k \in i}}{E_i} \quad \forall i = 1, 2, \dots, 29 \quad (3)$$

The normalization in above equation suppresses frequency band with high energy concentration which helps in obtaining a desired saliency map. The combined divisive normalization coefficient for  $p^{\text{th}}$  patch is given by

$$\hat{\mathbf{R}}_p = \cup_i \left( \hat{R}_p(k)_{k \in i} \right) \quad \forall i = 1, 2, \dots, 29 \quad (4)$$

The saliency maps  $\mathbf{SM}_p$  of  $p^{\text{th}}$  patch is obtained [4] as

$$\mathbf{SM}_p = \mathbf{W} \left| F^{-1} \left\{ \hat{\mathbf{R}}_p \right\} \right|^2 \quad (5)$$

where  $\mathbf{W}$  denotes the windowing function for edge effects removal.

For each scale and a given colour channel, the saliency maps for all patches are combined by taking the maximum value at each pixel location. The obtained saliency maps from all scales are resized to the dimension of original image and the final local saliency map (**LSM**) is obtained in terms of spatial maximum across all scales and color channels. Finally the saliency map is smoothed using Gaussian filter and values of **LSM** are normalized to [0, 1].

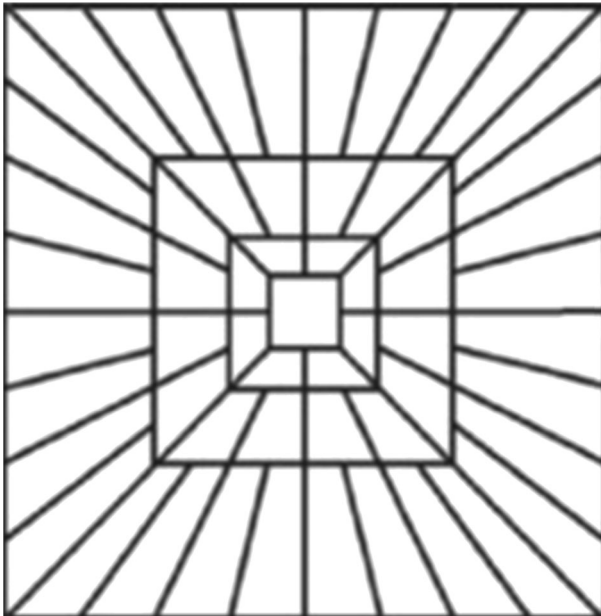
### 3.2 Global saliency computation

To reduce the computation time to capture global saliency of the image in the proposed HLGGM model, we have used fast Walsh-Hadamard transform (WHT) [15, 16, 24, 44] instead of fast Fourier transform (FFT). The elements of basis vectors of WHT take only +1 and -1 values while FFT requires complex multiplications. The computations involved in WHT are very simple because when an image is projected onto the basis images, all that is needed to do is to multiply each pixel by ±1. So the WHT is computationally more efficient than FFT [26, 42]. The WHT [26] coefficients of the image **I** of size  $N \times N$  where  $N=2^n$  is computed as

$$WH(u, v) = \frac{1}{N} \sum_{r=0}^{N-1} \sum_{c=0}^{N-1} I(r, c) (-1)^{\sum_{i=0}^{n-1} [b_i(r)p_i(u) + b_i(c)p_i(v)]} \quad 0 \leq u, v \leq N-1 \quad (6)$$

where  $b_i(r)$  is the  $i^{th}$  bit in the binary representation of  $r$ .  $(u, v)$  is the index in frequency domain and  $(r, c)$  is the index in spatial domain. In addition,  $p_i(u)$  is found as follows:

$$\begin{aligned} p_0(u) &= b_{n-1}(u) \\ p_i(u) &= b_{n-i}(u) + b_{n-i-1}(u) \quad \text{for } i = 1 \dots n-1 \end{aligned}$$



**Fig. 1** 2D-multiscale directional filter bank of contourlet transform

These WHT coefficients correspond to the frequencies ranging from lowest to highest from the origin with a lot of mid-range frequencies as shown in Fig. 2, where origin is at the top left corner. LF, MF, and HF refer to the Low Frequencies, Middle Frequencies, and High Frequencies respectively in the image.

We pick all the high frequencies coefficients as shown in Fig. 2, which can be selected by the following mask **M**.

$$M(u, v) = \begin{cases} 0, & \text{if } u \leq \frac{9N}{10} \text{ and } v \leq \frac{9N}{10} \\ 1, & \text{otherwise} \end{cases}$$

The selected WHT coefficients are given by

$$WH'(u, v) = WH(u, v) * M(u, v) \tag{7}$$

The global saliency map (**GSM**) is computed as

$$GSM(r, c) = \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} WH'(u, v) (-1)^{u+v} \sum_{i=0}^{n-1} [b_i(r) p_i(u) + b_i(c) p_i(v)] \quad \text{for } 0 \leq r, c \leq N-1 \tag{8}$$

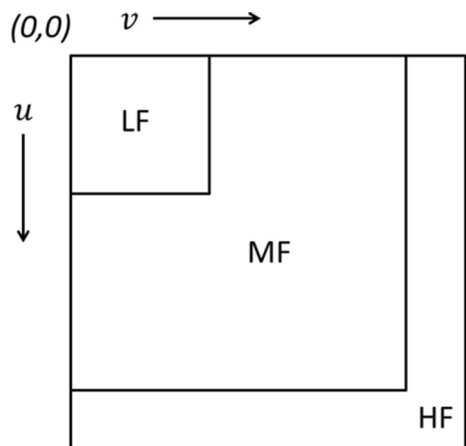
The values of **GSM** are normalized to [0,1]. Finally the obtained saliency map **GSM** is resized to the dimension of original image.

Being a real, symmetric and orthogonal transform, the WHT transformation matrix **H** has the following properties [26, 39, 43]:

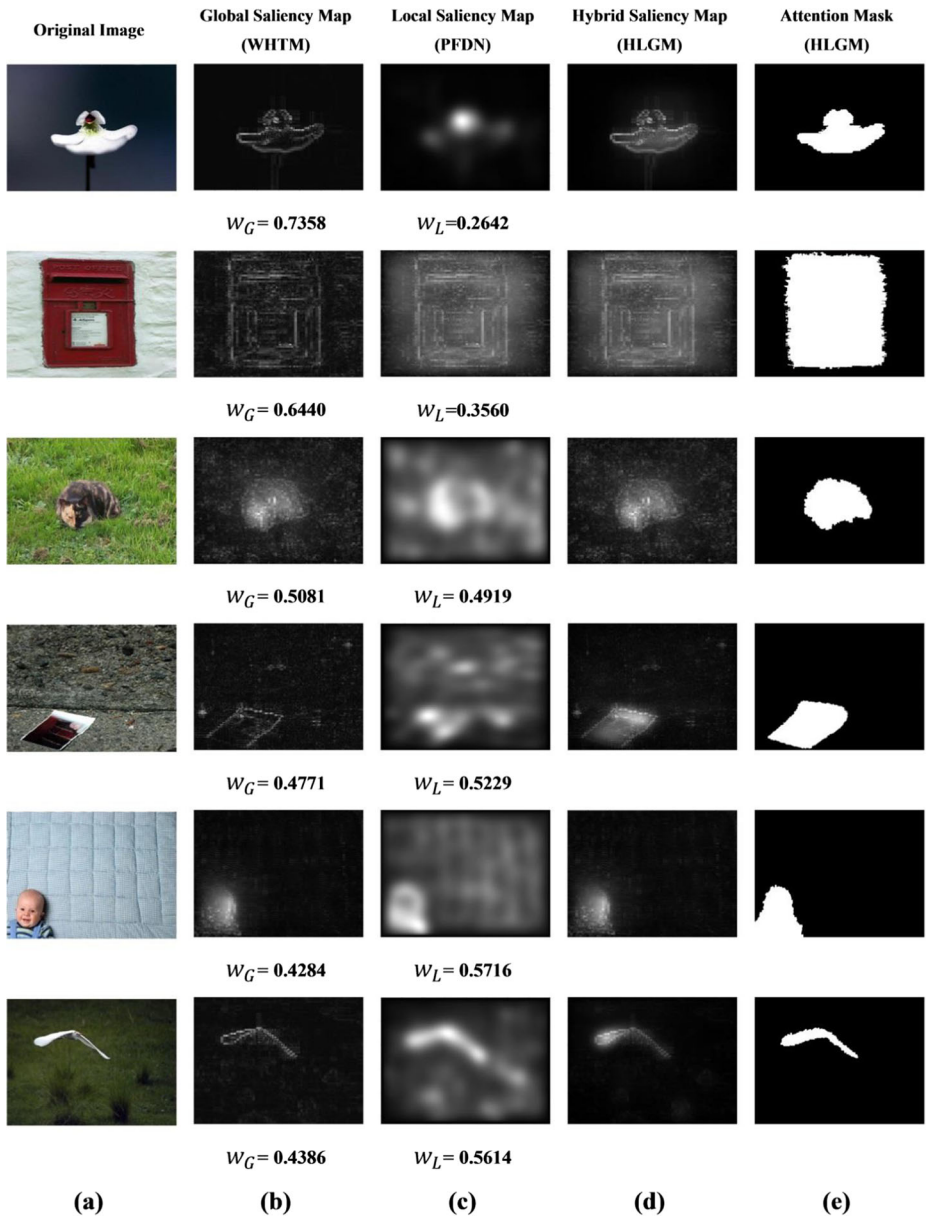
$$H = H^* = H^T = H^{-1}$$

The most attractive aspect of the WHT is that it involves only addition and subtraction computations, with no multiplication operation. Since multiplication is a time consuming operation, using WHT saves a significant amount of computation time [15, 24, 26, 39, 43, 44]. In this manuscript, this global approach based on WHT to compute global saliency, is abbreviated as WHTM (Walsh-Hadamard Transform Method).

**Fig. 2** Range of Low Frequencies (LF), Middle Frequencies (MF) and High Frequencies (HF)







**Fig. 3** **a** Original image **b** Global saliency map and its corresponding global weight **c** Local saliency map and its corresponding local weight **d** Hybrid saliency map using HLGM approach **e** Attention mask generated from hybrid saliency map

### 3.3 Composition saliency computation

It is possible that for some images global saliency information plays a vital role to detect a salient object properly while for others local saliency information is important. Both of these saliency information need to be combined in such a way that the dominant saliency



information gets higher weight. To capture the composite information in the proposed HLGGM model, we used a linear weighted combination of the local and global saliency information to compute the final saliency map (**FSM**) which is obtained as

$$\mathbf{FSM} = w_L \times \mathbf{LSM} + w_G \times \mathbf{GSM} \quad (9)$$

where  $w_L$  and  $w_G$  represent the weights to be assigned to the local and global saliency maps. A desired saliency map should highlight the salient objects while suppressing the objects which are non-salient. In such case the histogram of saliency values will not be uniformly distributed over all bins and the corresponding entropy will be small. It can be easily observed from a few experiments that the saliency map with the minimum entropy value clearly separates the salient region from the background. The lower is the entropy of saliency map, pixels of salient object are minimally scattered. Hence, higher weight is assigned to a saliency map with lower entropy and lower weight is assigned to a saliency map with higher entropy to obtain better saliency map. To choose the weights which satisfy this criterion, the weights  $w_L$  and  $w_G$  can be assigned inversely proportional to entropy of local saliency map and entropy of global saliency map respectively. For some images, local weights and global weights corresponding to local saliency maps and global saliency maps respectively are shown in column (b) and (c) of Fig. 3. Entropy of the local and global saliency maps are computed as

$$E_L = -\sum_{i=1}^b p_i(\mathbf{LSM}) \log_2 p_i(\mathbf{LSM}) \quad \text{and} \quad E_G = -\sum_{i=1}^b p_i(\mathbf{GSM}) \log_2 p_i(\mathbf{GSM}) \quad (10)$$

where  $b$  represents the number of bins,  $p_i(\cdot)$  indicates the probability of pixels belonging to the  $i^{\text{th}}$  bin in the histogram. In the experiments, the number of bins used to compute the hybrid saliency map has been set to 16 bins. The weights  $w_L$  and  $w_G$  are computed as

$$w_L = \begin{cases} \frac{E_G}{E_L + E_G} & \text{if } E_L \neq 0 \text{ and } E_G \neq 0 \\ 1 & \text{if } E_G = 0 \text{ and } E_L \neq 0 \end{cases} \quad (11)$$

$$w_G = \begin{cases} \frac{E_L}{E_L + E_G} & \text{if } E_L \neq 0 \text{ and } E_G \neq 0 \\ 1 & \text{if } E_L = 0 \text{ and } E_G \neq 0 \end{cases} \quad (12)$$

The saliency map **FSM** is normalized between  $[0, 1]$ . The normalized saliency value of pixel  $p$  is computed as

$$\mathbf{FSM}_N(p) = \frac{\mathbf{FSM}(p) - m}{M - m} \quad (13)$$

where  $m = \min_{p \in P} \mathbf{FSM}(p)$  and  $M = \max_{p \in P} \mathbf{FSM}(p)$  represent the minimum and maximum values of the saliency map respectively, and  $P$  indicates the set of all pixels in the image. A threshold is required to classify a pixel  $p$  into an attention pixel or a background pixel. For this purpose, generally a fixed threshold is selected which is half of the maximum saliency value. But a fixed threshold may not be suitable for all saliency maps. In our experiments, we used an adaptive threshold  $\tau$  that is dependent on the saliency map. The adaptive threshold  $\tau$  is

calculated in two steps. In the first step, a Canny edge operator is applied on the normalized saliency map  $\mathbf{FSM}_N$  to generate the object's silhouette. The edge information  $E$  for every pixel  $p$  is given as

$$E(p) = \text{edge}(\mathbf{FSM}_N(p)) = \begin{cases} 1, & p \text{ is an edge pixel} \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

In the second step, the average of the saliency values present at the object's silhouette is used as a threshold  $\tau$  to classify a pixel  $p$  into an attention pixel or a background pixel. The threshold  $\tau$  is computed as

$$\tau = \frac{\sum_{p \in P} E(p) \cdot \mathbf{FSM}_N(p)}{\sum_{p \in P} E(p)} \quad (15)$$

A binary threshold map  $\mathbf{T}$  from the grayscale saliency map  $\mathbf{FSM}_N$  is generated as

$$\mathbf{T}(p) = \begin{cases} 1, & \mathbf{FSM}_N(p) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

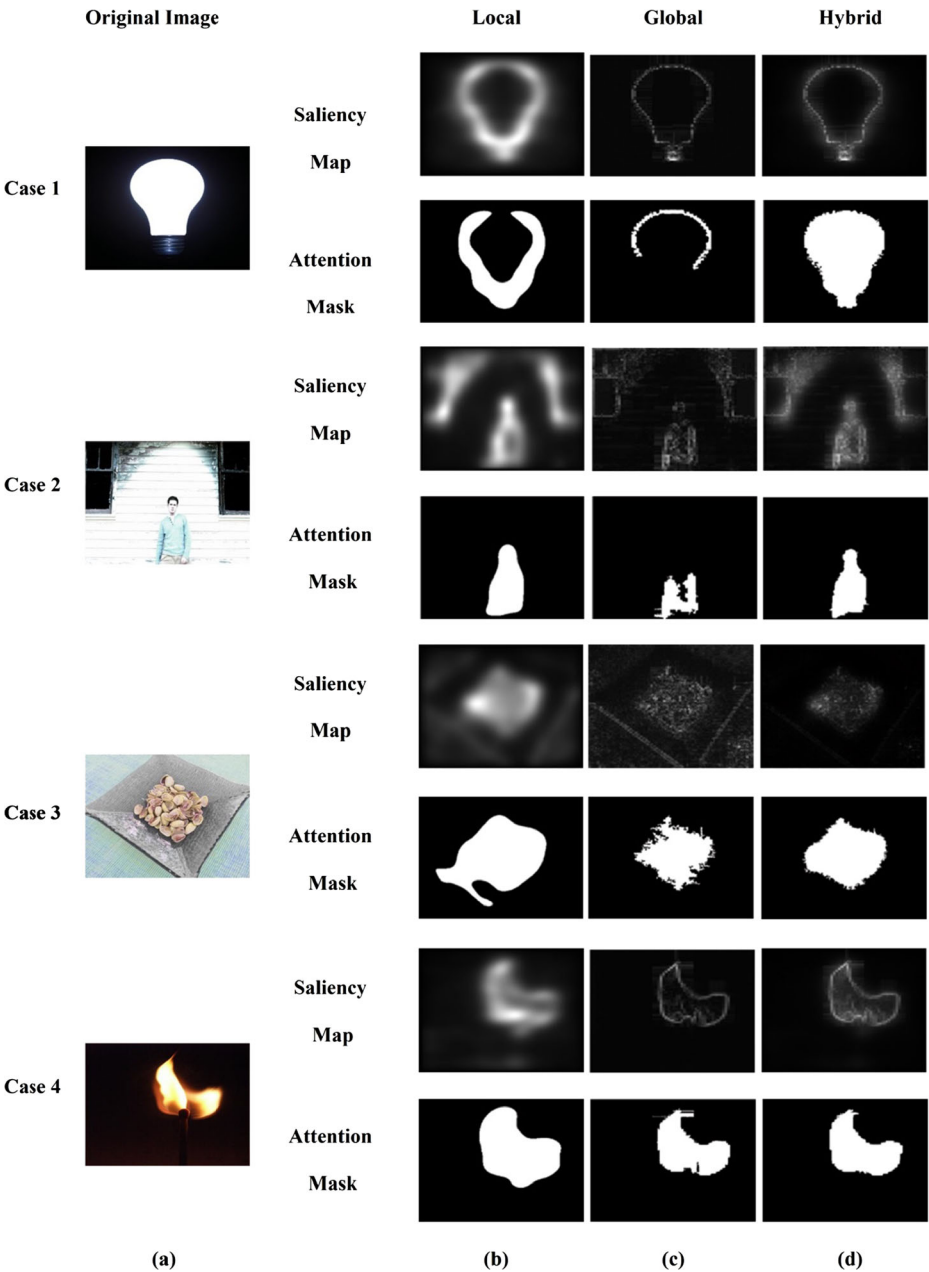
where the values of  $\mathbf{T}$  corresponding to 1 represent attention pixels and 0 indicate background pixels. From Eq. (16), a threshold map  $\mathbf{T}$  is generated which contains several objects. These objects contain several holes. By holes we here mean a set of background pixels that cannot be reached by filling in the background from the edge of the object. First we fill the holes and obtain connected components in the threshold map  $\mathbf{T}$ . Then connected component labelling is done by identifying the connected components in threshold map  $\mathbf{T}$  and assigning each connected component a unique label using 8-connected neighbourhood. Then after discarding background, all the remaining connected components are sorted according to their area. Finally the connected component with the largest area is chosen as an attention mask corresponding to the saliency map.

Figure 4 depicts the local, global and final saliency maps and their corresponding attention masks on certain images for comparison. **Case 1** corresponds to the case where local saliency as well as global saliency does not give good result individually. **Case 2** corresponds to the case where local saliency performs better than global saliency. **Case 3** corresponds to the case where global saliency is better than local saliency. **Case 4** corresponds to the case where both the local saliency and global saliency show good and almost comparable results. But in all above mentioned four cases our proposed hybrid approach HLGGM renders better performance both in terms of saliency maps and attention masks.

## 4 Experimental setup and results

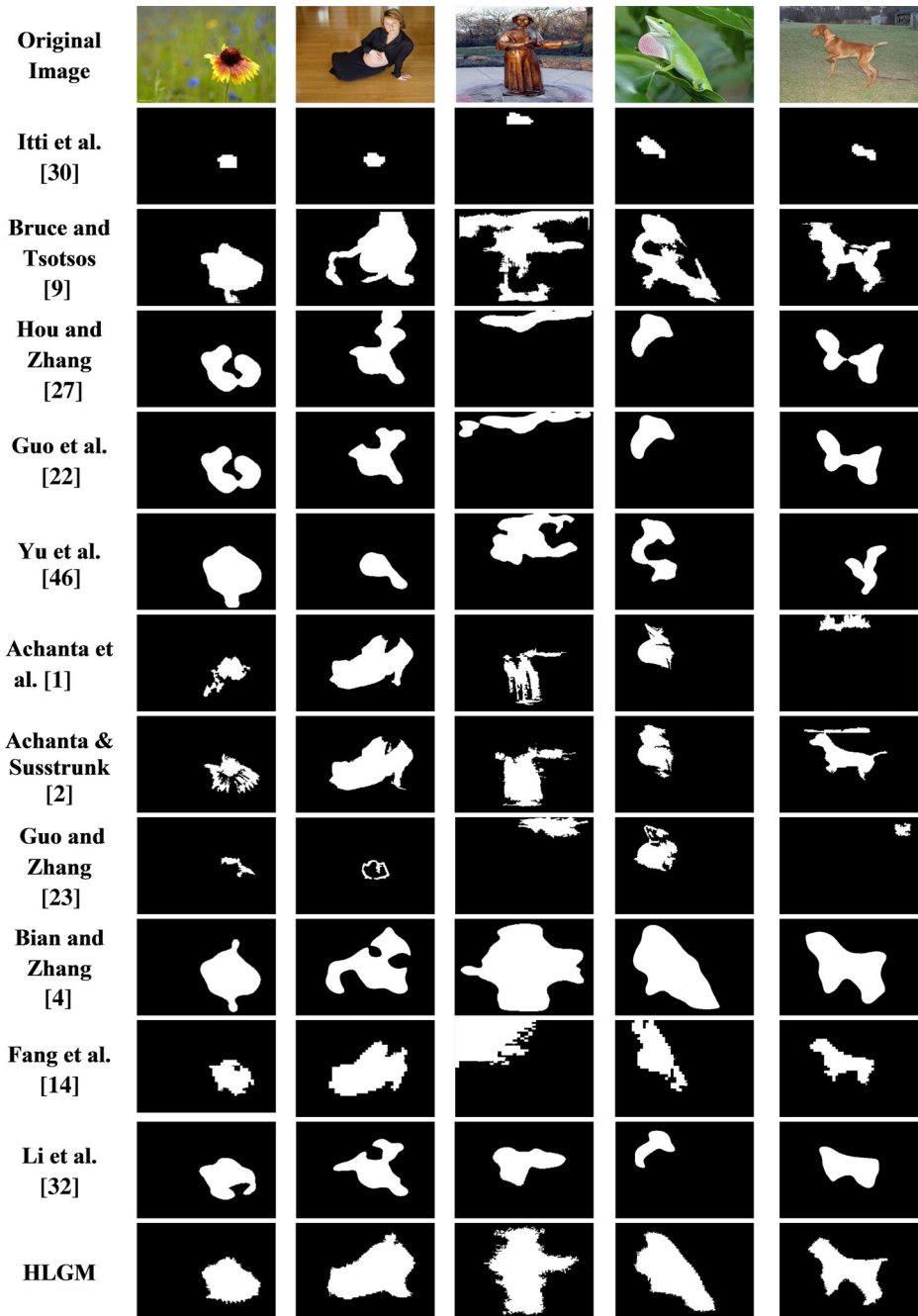
To check the efficacy of the proposed HLGGM model, the performance is evaluated both qualitatively and quantitatively, and is compared with the existing approaches [1, 2, 4, 9, 14, 22, 23, 27, 30, 34, 46]. The performance of the HLGGM and eleven other state-of-the-art models is examined using two popular and publicly available datasets, and one new ground truth based dataset. The first one is Microsoft Research Asia Salient Object Database<sup>1</sup> (MSRA SOD)

<sup>1</sup> [http://www.research.microsoft.com/enus/um/people/jiansun/salientobject/salient\\_object.htm](http://www.research.microsoft.com/enus/um/people/jiansun/salientobject/salient_object.htm)



**Fig. 4** **a** Original image. **b** Local saliency maps and corresponding attention masks. **c** Global saliency maps and corresponding attention masks. **d** Final saliency maps and corresponding attention masks by the proposed model

image set B. It contains 5000 high quality colour images of various object categories and scene types in 10 subfolders with their ground truth manually labelled by nine users. The result is in the form of a rectangle that is bounding the salient object. The



**Fig. 5** Qualitative evaluation of the HLGM model and eleven other state-of-the-art models

**Table 1** Quantitative comparison for MSRA SOD image set B

| Model Name             | Precision    | Recall       | F -measure   |
|------------------------|--------------|--------------|--------------|
| Local Approach (PFDN)  | 0.593        | <b>0.944</b> | 0.728        |
| Global Approach (WHTM) | <b>0.714</b> | 0.676        | 0.695        |
| Hybrid Approach (HLGM) | 0.686        | 0.877        | <b>0.770</b> |

second one is Binary Masks,<sup>2</sup> containing 1000 images selected from 5000 images of MSRA SOD image set B. These images are manually segmented and the result is in the form of a binary mask. Achanta et al. [1] suggested that the bounding box based ground truth is inaccurate as it may contain various objects in a single box. In order to overcome this problem, they suggested an object-contour based ground truth dataset. But they had chosen only 1000 images out of 5000 images. We derived a new ground truth based dataset from a publicly available dataset called the SAA\_GroundTruth.<sup>3</sup> It contains all the 5000 images of MSRA SOD image set B which are manually segmented in such a way that the result matches the majority of bounding boxes as suggested in MSRA SOD image set B. All the images are of size 400×300 or 300×400 having intensity values in [0, 255]. For both the qualitative evaluation and quantitative evaluation, all the experiments regarding our proposed approach and other state-of-the-art models are carried out using Windows 7 environment over Intel(R) Xeon(R) processor with a speed of 2.27 GHz and 4GB RAM.

#### 4.1 Qualitative evaluation

The qualitative evaluation of the proposed model and eleven other state-of-the-art models on five images can be seen in Fig. 5. We have chosen these five images from the test data set which contain objects differing in shape, size, position, type etc. The following observations regarding the attention masks can be drawn from Fig. 5:

- Itti et al. [30] worked at the local level and neglected the global details, hence gave disappointing results.
- Bruce and Tsotsos [9] gave better saliency results than Itti et al. [30] by utilizing the information maximization approach.
- Hou and Zhang [27] lacked the shape information of the objects.
- Guo et al. [22] gave unsatisfactory results with lacked shape details.
- Yu et al. [46] failed to give satisfactory results with deteriorated shapes.
- Achanta et al. [1] gave clear results for some images but it deteriorated for others.
- Achanta and Susstrunk [2] gave better results than its previous work but included some extra information.
- Guo and Zhang [23] failed to notice the shape information of the object.

<sup>2</sup> [http://ivrgwww.epfl.ch/supplementary\\_material/RK\\_CVPR09/GroundTruth/binarymasks.zip](http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/GroundTruth/binarymasks.zip)

<sup>3</sup> E-mail at "rinki.arya89 @ gmail.com or navjot.singh.09@gmail.com"

**Table 2** Quantitative comparison for Binary Masks

| Model Name             | Precision    | Recall       | F -measure   |
|------------------------|--------------|--------------|--------------|
| Local Approach (PFDN)  | 0.508        | <b>0.887</b> | 0.646        |
| Global Approach (WHTM) | <b>0.709</b> | 0.517        | 0.598        |
| Hybrid Approach (HLGM) | 0.657        | 0.734        | <b>0.693</b> |

- Bian and Zhang [4] gave better results than all the above mentioned models in terms of saliency detection. The shape information needs to be enhanced and it also contained unnecessary details.
- Fang et al. [14] missed finer shape details. It was able to localise the objects properly but with deteriorated shapes.
- Li et al. [34] either missed some portion of the object or gave extra information of the object that was not required.
- The proposed HLGM model gave better shape information and clear boundaries of the object.

## 4.2 Quantitative evaluation

The quantitative evaluation of the proposed model and eleven other state-of-the-art models is done in terms of precision, recall, F -measure, and computation time. Using the ground truth  $\mathbf{G}$  and the detection result  $\mathbf{R}$ , precision, recall, F -measure are calculated as

$$Precision = \frac{TP}{TP + FP} ; \quad Recall = \frac{TP}{TP + FN} ; \quad F_{\alpha} = \frac{(1 + \alpha) \times Precision \times Recall}{\alpha \times Precision + Recall}$$

where  $TP = \sum_{\mathbf{G}(x,y)=1} \mathbf{R}(x,y) ; \quad FP = \sum_{\mathbf{G}(x,y)=0} \mathbf{R}(x,y) ; \quad FN = \sum_{\mathbf{R}(x,y)=0} \mathbf{G}(x,y)$

(17)

where 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.

While computing value of F -measure, we have chosen  $\alpha=1$  to give equal weightage to both precision and recall.

Tables 6, 7 and 8 show the quantitative performance evaluation of the proposed method in comparison to the other state-of-the-art methods on MSRA SOD image set B, Binary Masks and SAA\_Ground Truth respectively. The computation time taken by the models can be

**Table 3** Quantitative comparison for SAA Ground Truth

| Model Name             | Precision    | Recall       | F -measure   |
|------------------------|--------------|--------------|--------------|
| Local Approach (PFDN)  | 0.495        | <b>0.857</b> | 0.628        |
| Global Approach (WHTM) | <b>0.669</b> | 0.498        | 0.571        |
| Hybrid Approach (HLGM) | 0.624        | 0.669        | <b>0.646</b> |

**Table 4** Average weight with standard deviation for local and global approaches

| Model Name             | Average weight±Standard deviation |
|------------------------|-----------------------------------|
| Local Approach (PFDN)  | 0.468±0.1619                      |
| Global Approach (WHTM) | 0.532±0.1619                      |

observed from Table 9. To compare the proposed approach with the existing state-of-the-art methods in terms of computation time, all experiments are carried out using Windows 7 environment over Intel(R) Xeon(R) processor with a speed of 2.27 GHz and 4GB RAM to avoid biases. The best results are shown in bold.

To highlight the contributions of local (PFDN) and global approach (WHTM) to our proposed approach (HLGM), the quantitative results of the local approach (PFDN), global approach (WHTM) and hybrid approach (HLGM) are shown in Tables 1, 2 and 3.

It can be observed that for all the three datasets, the performance of the proposed method (HLGM) is better than both local approach (PFDN) and global approach (WHTM) in terms of F-measure. It can be also noted that for all the three datasets, the performance of global approach (WHTM) is better in comparison to local approach (PFDN) in terms of precision only. On the other hand, for all the three datasets, the performance of local approach (PFDN) is better in comparison to global approach (WHTM) in terms of Recall and F-measure. Experimental results suggest that the performance of the hybrid approach (HLGM) is better due to combination of both local (PFDN) and global approach (WHTM) and the way they are combined.

For each image in the MSRA SOD B image set B containing 5000 images, we calculate a local and a global weight corresponding to local and global saliency map respectively. In this way, 5000 local weights and 5000 global weights are computed corresponding to 5000 local and 5000 global saliency maps respectively. Now an average local weight is calculated by taking the average of all 5000 local weights. Similarly an average global weight of all 5000 global weights is calculated. Average local weight with standard deviation for local approach (PFDN) and average global weight with standard deviation for global approach (WHTM) are shown in Table 4.

The computation time taken by local method, global method and the proposed method can be observed from Table 5. It can also be clearly observed from Table 5 that Global approach (WHTM) is fast enough to compute global saliency in real time.

The following can be observed from Tables 6, 7 and 8:

**Table 5** Computational time taken by models

| Model Name             | Time (in sec) per image |
|------------------------|-------------------------|
| Local Approach (PFDN)  | 0.25                    |
| Global Approach (WHTM) | 0.008                   |
| Hybrid Approach (HLGM) | 0.26                    |



**Table 6** Quantitative comparison for MSRA SOD image set B

| Model Name                | Precision    | Recall       | F –measure   |
|---------------------------|--------------|--------------|--------------|
| Guo and Zhang [23]        | 0.815        | 0.245        | 0.377        |
| Hou and Zhang [27]        | 0.761        | 0.526        | 0.622        |
| Guo et al. [22]           | 0.764        | 0.528        | 0.625        |
| Achanta et al. [1]        | 0.717        | 0.575        | 0.638        |
| Itti et al. [30]          | 0.672        | 0.614        | 0.642        |
| Yu et al.[46]             | 0.797        | 0.579        | 0.671        |
| Fang et al. [14]          | 0.671        | 0.675        | 0.673        |
| Li et al. [34]            | <b>0.892</b> | 0.580        | 0.702        |
| Bian and Zhang [4]        | 0.593        | <b>0.944</b> | 0.728        |
| Achanta and Susstrunk [2] | 0.786        | 0.704        | 0.743        |
| Bruce and Tsotsos [9]     | 0.728        | 0.762        | 0.745        |
| HLGM                      | 0.686        | 0.877        | <b>0.770</b> |

- In terms of precision, Li et al. [34] outperforms the rest of the state-of-the-art methods for MSRA and Binary-Mask datasets. Achanta and Susstrunk [2] shows the highest precision for SAA dataset.
- In terms of recall, Bian and Zhang [4] model furnishes the best performance for all three datasets. But it gives the lowest precision for MSRA and SAA and also shows poor precision for Binary-Mask dataset.
- The performance of model proposed by Guo and Zhang [23] is not good in terms of recall and F-measure for all the three datasets.

In terms of F -measure, the proposed method outperforms other state-of-the-art methods for all the three datasets which can also be observed in Fig. 6. This signifies that the proposed

**Table 7** Quantitative comparison for Binary Masks

| Model Name                | Precision    | Recall       | F -measure   |
|---------------------------|--------------|--------------|--------------|
| Guo and Zhang [23]        | 0.694        | 0.117        | 0.201        |
| Guo et al. [22]           | 0.578        | 0.379        | 0.458        |
| Hou and Zhang [27]        | 0.502        | 0.440        | 0.469        |
| Yu et al.[46]             | 0.626        | 0.403        | 0.489        |
| Achanta et al. [1]        | 0.599        | 0.606        | 0.603        |
| Itti et al. [30]          | 0.550        | 0.695        | 0.614        |
| Li et al. [34]            | <b>0.776</b> | 0.535        | 0.633        |
| Fang et al. [14]          | 0.638        | 0.636        | 0.637        |
| Bian and Zhang [4]        | 0.508        | <b>0.887</b> | 0.646        |
| Achanta and Susstrunk [2] | 0.635        | 0.670        | 0.652        |
| Bruce and Tsotsos [9]     | 0.535        | 0.859        | 0.659        |
| HLGM                      | 0.657        | 0.734        | <b>0.693</b> |

**Table 8** Quantitative comparison for SAA Ground Truth

| Model Name                | Precision    | Recall       | F -measure   |
|---------------------------|--------------|--------------|--------------|
| Guo and Zhang [23]        | 0.677        | 0.114        | 0.195        |
| Guo et al. [22]           | 0.583        | 0.371        | 0.454        |
| Hou and Zhang [27]        | 0.588        | 0.372        | 0.456        |
| Yu et al.[46]             | 0.617        | 0.382        | 0.472        |
| Fang et al. [14]          | 0.593        | 0.544        | 0.568        |
| Itti et al. [30]          | 0.545        | 0.609        | 0.575        |
| Li et al. [34]            | 0.755        | 0.490        | 0.594        |
| Bruce and Tsotsos [9]     | 0.527        | 0.777        | 0.628        |
| Achanta et al. [1]        | 0.800        | 0.517        | 0.628        |
| Bian and Zhang [4]        | 0.495        | <b>0.857</b> | 0.628        |
| Achanta and Susstrunk [2] | <b>0.801</b> | 0.524        | 0.634        |
| HLGM                      | 0.624        | 0.669        | <b>0.646</b> |

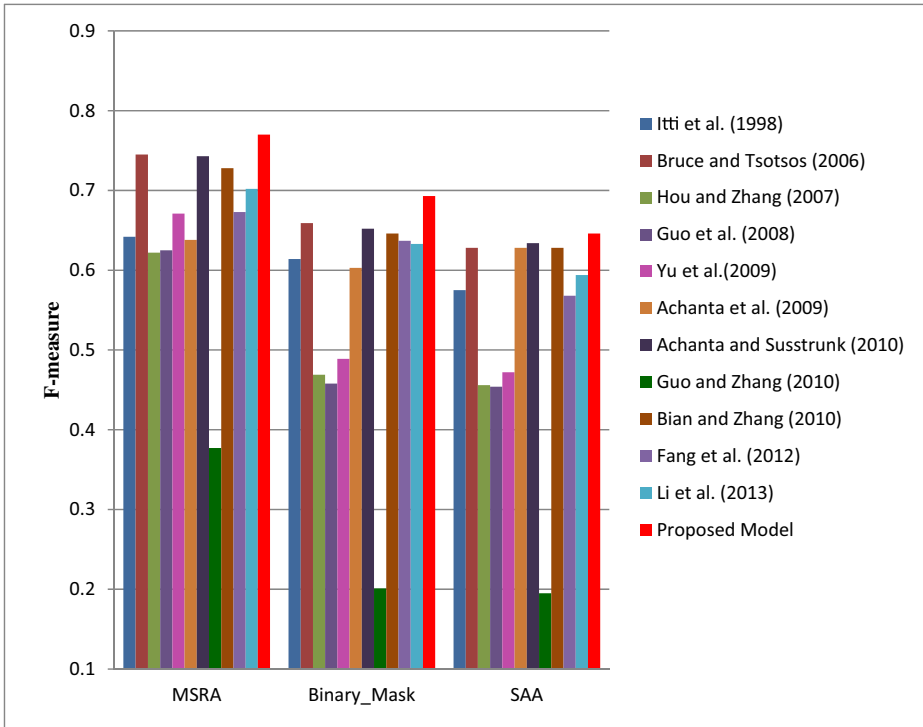
method provides good performance both in terms of precision and recall whereas other models give high precision value but low recall value and vice-versa.

The following can be observed from Table 9:

- The model suggested by Guo et al. [22] takes the least computational time. But this method does not perform well in terms of precision, recall and F-measure.
- Bruce and Tsotsos [9] model takes maximum time.
- Hou and Zhang [27], Yu et al. [46], Achanta et al. [1], Bian and Zhang [4], Guo and Zhang [23], and Guo et al. [22] models take less computation time than the proposed model but their performance is not as good in terms of F-measure.
- The computation time taken by the proposed model is considerably less in comparison to Itti et al. [30], Bruce and Tsotsos [9], Achanta and Susstrunk [2],

**Table 9** Computational time for the proposed model and different state-of-the-art models

| Model Name                | Time (in sec) per image |
|---------------------------|-------------------------|
| Itti et al. [30]          | 1.70                    |
| Bruce and Tsotsos [9]     | 50.8                    |
| Hou and Zhang [27]        | 0.02                    |
| Guo et al. [22]           | 0.018                   |
| Yu et al.[46]             | 0.06                    |
| Achanta et al. [1]        | 0.17                    |
| Achanta and Susstrunk [2] | 0.31                    |
| Guo and Zhang [23]        | 0.18                    |
| Bian and Zhang [4]        | 0.25                    |
| Fang et al. [14]          | 21.9                    |
| Li et al. [34]            | 0.69                    |
| HLGM                      | 0.26                    |



**Fig. 6** Comparison of F-measure values for state-of-the art methods and proposed method

Fang et al. [14] and Li et al. [34] which can be helpful in detecting salient object with higher detection accuracy in real-time.

## 5 Conclusion and future work

In this paper, we proposed a novel and fast biologically plausible frequency domain approach for salient object detection. The proposed approach determined salient object by considering both local saliency and global saliency. The proposed approach involved three phases. In the first phase, locally salient features were generated using the research work of Bian and Zhang. In the second phase, globally salient features were computed using fast Walsh-Hadamard transform. Finally, the saliency map was obtained in terms of linear weighted combination of local and global saliency where the weights were calculated using entropy measure. The performance of the proposed model was evaluated in terms of precision, recall, F-measure and computation time using two publicly available image datasets and one new dataset. Experiments on variety of images showed that the proposed approach outperformed Bian and Zhang model which considers only local saliency and other existing state-of-the-art methods in terms of F-measure. The proposed approach was found to be less computationally expensive to detect salient object accurately.

There are many possible remaining issues for further investigation such as partial occlusion, intra-class variation, viewpoint variation, background clutter, and

articulation. In future non-linear combination of features can be planned to evaluate the performance. We also plan to extend our work to detect any number of salient objects or no salient object at all.

**Acknowledgments** The first author expresses her gratitude to the University Grant Commission (UGC), India for the obtained financial support in performing this research work.

## References

- Achanta R, Hemamiz S, Estraday F, Susstrunk S (2009) Frequency-tuned salient region detection. *IEEE Conf Comput Vis Pattern Recognit (CVPR)* 1597–1604
- Achanta R, Susstrunk S (2010) Saliency detection using maximum symmetric surround. *IEEE Int Conf Image Process (ICIP)* 2653–2656
- Amit Y (2002) 2D Target detection and recognition, models, algorithms and networks. MIT Press, Cambridge, MA
- Bian P, Zhang LM (2010) Visual saliency: a biologically plausible contourlet-like frequency domain approach. *Cogn Neurodyn* 4(3):189–198
- Borji A, Itti L (2012) Exploiting local and global patch rarities for saliency detection. *IEEE Conf Comput Vis Pattern Recognit (CVPR)* 478–485
- Borji A, Itti L (2013) State-of-the-art in visual attention modeling. *IEEE Trans Pattern Anal Mach Intell* 35(1):185–207
- Borji A, Sihite DN, Itti L (2012) Salient object detection: a benchmark. *Proc Eur Conf Comput Vis Lecture Notes Comput Sci* 414–429
- Borji A, Sihite DN, Itti L (2012) Salient object detection: a benchmark. *Eur Conf Comput Vis* 414–429
- Bruce NDB, Tsotsos JK (2006) Saliency based on information maximization. *Adv Neural Inf Process Syst* 18: 155–162
- Chen L, Xie X, Fan X, Ma W, Shang H, Zhou H (2002) A visual attention model for adapting images on small displays. Technical Report, Microsoft Research Redmond
- Cheng M-M, Mitra NJ, Huang X, Torr PHS, Hu S-M (2011) Salient object detection and segmentation. *IEEE Trans Pattern Anal Mach Intell. Technical Report, TPAMI-2011-10-0753*
- Cheung Y, Peng Q (2012) Salient region detection using local and global saliency. *21st Int Conf Pattern Recognit (ICPR)* 210–213
- Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev* 3:201–215
- Fang Y, Lin W, Lee B-S et al (2012) Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum. *IEEE Trans Multimed* 14(1):187–198
- Fine NJ (1949) On the Walsh functions. *Trans Am Math Soc* 65:372–414
- Fine NJ (1950) The generalized Walsh functions. *Trans Am Math Soc* 69:66–77
- Frintrop S, Rome E, Christensen HI (2010) Computational visual attention systems and their cognitive foundation: a survey. *ACM Trans Appl Percept* 7(1):1–46
- Gasparini F, Corchs S, Schettini R (2007) Low quality image enhancement using visual attention. *Opt Eng* 46(4):40502–040504
- Goferman S, Zelnik-Manor L, Tal A (2012) Context-aware saliency detection. *IEEE Trans Pattern Anal Mach Intell* 34:1915–1926
- Gonzalez RC, Woods RE (2002) Digital image processing. Prentice-Hall, Upper Saddle River
- Graefe R, Efenberger W (1996) A novel approach for the detection of vehicles on freeways by real time vision. In *Intelligent Vehicles* 363–368
- Guo CL, Ma Q, Zhang LM (2008) Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. *IEEE Conf Comput Vis Pattern Recognit (CVPR)* 1–8
- Guo CL, Zhang LM (2010) A novel multiresolution spatio temporal saliency detection model and its applications in image and video compression. *IEEE Trans Image Process* 19(1):185–198
- Hadamard J (1893) Resolution d'une question relative aux determinants. *Bull Sci Math* 17:240–246
- Han J, Ngan KN, Li MJ, Zhang HJ (2006) Unsupervised extraction of visual attention objects in color images. *IEEE Trans Circuit Syst Video Technol* 16:141–145
- Hassan M, Osman I, Yahia M (2007) Walsh-hadamard transform for facial feature extraction in face recognition. *Int J Comput Inf Sci Eng* 1(5)

27. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. *IEEE Conf Comput Vis Pattern Recognit (CVPR)* 1–8
28. Itti L (2000) Models of bottom up and top down visual attention. Dissertation, California Institute of Technology, Pasadena
29. Itti L (2005) Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Vis Cogn* 12:1093–1123
30. Itti L, Koch C, Niebur E (1998) A model of saliency based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Mach Intell* 20:1254–1259
31. 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
32. Kanan C, Cottrell G (2010) Robust classification of objects, faces, and flowers using natural image. *IEEE Conf Comput Vis Pattern Recognit (CVPR)* 2472–2479
33. Karssemeijer N (2006) Detection of stellate distortions in mammograms. *IEEE Trans Med Imaging* 15:611–619
34. 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
35. Li Y, Hou X, Koch C, Rehg JM, Yuille AL (2014) The secrets of salient object segmentation. *IEEE Conf Comput Vis Pattern Recognit* 280–287
36. Li Z, Itti L (2011) Saliency and gist features for target detection in satellite images. *IEEE Trans Image Process* 20:2017–2029
37. Liu T, Yuan Z, Sun-Wang J, Zheng N, Tang X, Shum HY (2011) Learning to detect a salient object. *IEEE Trans Pattern Anal Mach Intell* 33:353–366
38. Ma Y, Hua X, Lu L, Zhang H (2005) A generic framework of user attention model and its application in video summarization. *IEEE Trans Multimed* 7(5):907–919
39. Pratt WK, Kane J, Andrews JC (1969) Hadamard transform image coding. *Proc IEEE* 57(1):58–68
40. Rother C, Bordeaux L, Hamadi Y, Blake A (2006) Autocollage. *ACM Trans Graph* 25:847–852
41. Santella A, Agrawala M, Decarlo D, Salesin D, Cohen M (2006) Gaze based interaction for semi-automatic photo cropping. *SIGCHI Conf Hum Factors Comput Syst* 771–780
42. Scott E (1999) *Computer vision and image processing*. Prentice Hall, Upper Saddle River
43. Seberry J, Wysocki BJ, Wysocki TA (2005) On some applications of Hadamard matrices. *Metrika* 62:221–239
44. Walsh JL (1923) A closed set of orthogonal functions. *Am J Math* 55:5–24
45. Walther D, Koch C (2006) Modeling attention to salient proto-objects. *Neural Netw* 19:1395–1407
46. Yu Y, Wang B, Zhan LM (2009) Pulse discrete cosine transform for saliency-based visual attention. *The 8th Int Conf Dev Learn (ICDL)* 1–6
47. Zhang W, Wu QMJ, Wang G, Yin H (2010) An adaptive computational model for salient object detection. *IEEE Trans Multimed* 12:300–315



**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. Currently he is pursuing 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.