Sādhanā Vol. 42, No. 5, May 2017, pp. 671–685
DOI 10.1007/s12046-017-0627-7

© Indian Academy of Sciences

An efficient visual saliency detection model based on Ripplet transform

A DIANA ANDRUSHIA[1,*] and R THANGARAJAN[2]

[1]Karunya University, Coimbatore 641114, India
[2]Kongu Engineering College, Erode 638052, India
e-mail: andrushia@gmail.com; thangs_68@yahoo.com

MS received 5 December 2015; revised 19 August 2016; accepted 29 October 2016

Abstract. Even though there have been great advancements in computer vision tasks, the development of human visual attention models is still not well investigated. In day-to-day life, one can find ample applications of saliency detection in image and video processing. This paper presents an efficient visual saliency detection model based on Ripplet transform, which aims at detecting the salient region and achieving higher Receiver Operating Characteristics (ROC). Initially the feature maps are obtained from Ripplet transform in different scales and different directions of the image. The global and local saliency maps are computed based on the global probability density distribution and feature distribution of local areas, which are combined together to get the final saliency map. Ripplet-transform-based visual saliency detection is the novel approach carried out in this paper. Experimental results indicate that the proposed method based on Ripplet transformation can give excellent performance in terms of precision, recall, F measure and Mean Absolute Error (MAE), and is compared with 10 state-of-the-art methods on five benchmark datasets.

Keywords. Ripplet transform; visual saliency model; Receiver Operating Characteristics (ROC); visual performance.

1. Introduction

Visual attention plays a vital role in visual information processing of Human Visual System (HVS). HVS is able to collect important cues from the visual scenes, which are commonly referred to as visual attention mechanism. In order to model this mechanism, many computational saliency models had been brought forth in the past decades. As attention is an abstract concept, it is very hard to closely imitate this natural capability of human vision processing by a machine [1]. Many of these computational models find application in adaptive image compression, segmentation of *object-of-interest* in an image, automatic image thumbnailing, object detection and recognition, visual tracking, automatic creation of image collage, content-aware image resizing, non-photo realistic rendering, etc. [2]. Humans exhibit a strong and innate capability to process a visual scene and extract visual saliency features. As far as computational models are concerned, there are two types of models in the literature, namely the bottom-up and the top-down approaches. Bottom-up approaches work from the low-level visual features and move up to construct a saliency map and by virtue they are task independent. The top-down approaches are goal-driven and they use prior knowledge to achieve a goal such as target detection, scene classification, object recognition, etc. [3]. The top-down approach starts from a particular visual processing task.

In order to describe saliency, many of the bottom-up computational models rely on the contrast of the local features, local differences with the neighbouring areas and local complexity. The colour information plays an important role as one among the local features. The various colour channels of Lab, YCbCr and RGBYI are used to generate saliency maps. In this proposed method, RGBYI colour channel is employed because this colour channel mimics the visual cortex of human system by involving colour channel pairs such as yellow/blue, blue/yellow, red/green and green/red. It also includes colour difference channels and four broadly tuned colour channels (detailed in section 2.1).

A pioneering computational model of visual saliency was proposed by [4]. In their work the saliency map is obtained from low-level conspicuous visual features of colour, intensity and orientation. These maps are derived from the cross-scale combination of feature maps with centre surround difference. Ma and Zhang [5], Itti *et al* [4], Oliva *et al* [6], Harel *et al* [7] and Goferman *et al* [8] proposed saliency models based on local contrast information. Achanta *et al* [9], Cheng *et al* [10], Perazzi *et al* [11], Guo *et al* [12] and Hou and Zhang [13] focused on global

*For correspondence

features to construct saliency models. Based on this, it is inferred that the main tasks that are involved in visual attention modelling are as follows: (1) obtaining the feature maps for local and global features, (2) combining the various feature maps and (3) using top-down features if required for a particular application. In general, the visual saliency models can be constructed in spatial and frequency domains. The computational complexity of these models is very high and these models are not suitable for multi-scale spatial and frequency analyses.

In the recent years, many researchers have vested more interest to build computational visual models in the transform domain. Li *et al* [14, 15] proposed a saliency model based on amplitude and phase spectrum of Fourier transform and this model can process only the global information. Hence the local information is missed out. Guo *et al* [12], Hou and Zhang [13], Boris and Stiefelhagen [16] and Li *et al* [14, 15] proposed saliency models with the help of Fourier transform, which did not give satisfactory results for aperiodic signals and thus could not obtain local frequency components. Hence these models have restrictions in terms of image size. Fourier transform can give promising results for stationary signals only.

In order to overcome these difficulties many studies have recently shifted to Wavelet transform (WT) in the construction of visual computational model. WT is capable of providing multi-scale analysis in spatial and frequency domains. WT can represent singularities in a much better manner than what Fourier transform can. Moreover, WT can be applied for non-stationary signals also [17]. Murray *et al* [18] computed weight maps for each colour sub-bands using Wavelet coefficients. The local contrast and centre surround region energy ratio were considered to obtain the weight map. Saliency map was derived from the inverse Wavelet transform (IWT) of weight maps for each colour sub-band and this method was less sensitive to non-salient edges. Oakes and Abhayaratne [19] proposed a visual saliency detection method for video sequence. WT was used to obtain spatial and temporal saliency maps, which were combined together with spatial intensity and orientation contrasts information of the video sequence. Ma *et al* [20] proposed saliency model based on WT and entropy. Saliency maps of different colour channels are combined using the entropy and Wavelet coefficients. Similarly many researchers have investigated visual saliency detection using WT [3, 21–24].

Even though Wavelets have the capability to account for multi-scale spatial and frequency analyses, they are not apt at giving a compact representation of intermediate dimensional structures. Wavelets are very good to represent point singularities but when it comes to directional features they fall short [25]. The main reason is that Wavelets are generated by isotropic elements. The quantity of decomposition level should be very large when approximating a curve using WT [26]. The disadvantages of WT are overcome by using multi-directional and multi-scale transforms.

Candes [27] and Candes and Donoho [28] introduced an anisotropic geometric WT named Ridgelet transform, which is optimal to represent straight-line singularities and resolve the 1D singularity; 2D, i.e., line and curve, singularities are still not resolvable by Ridgelet transform, which is one of its main drawback. To analyse line and curve singularities, the multi-scale Ridgelet transform was proposed by Candes and Donoho [25] and it was called as Curvelet transform (CT). It has the capability to represent directional features and improved efficiency to represent curve-like edges [29]. Zhong *et al* [30] proposed a saliency detection method that is based on CT and 2D Gabor filter. It uses CT's edge, directional information and Garbor transform's spatial localization for the saliency map construction. This method also incorporates the influence of centre bias as a top-down cue.

The anisotropic property of CT can resolve 2D singularities and obey the parabolic scaling law $width = length^2$, which may not be optimal for all types of boundaries. In order to sharpen the scaling law, the Ripplet transform was proposed. The discretization of CT is a challenging task and it is needed in the pre-processing steps of computational models. Ripplet transform is a generalization of CT with two important parameters (support c and degree d). These parameters provide the anisotropy capability of representing singularities along arbitrarily shaped curves. In recent years many developments have been made in directional Wavelet systems to optimally represent the directional features of signals in higher dimensions.

The Ripplet transform is used in various applications such as image retrieval, image compression, image fusion, etc. Chowdhury *et al* [31] presented an image retrieval system for natural colour images. In this system, multi-scale geometric analysis of Ripplet transform is emphasized. Geng *et al* [32] investigated the high directionality of Ripplet transform property in multi-focus image fusion technique. Juliet *et al* [33] performed medical image compression with the help of Ripplet transform and it yielded compressed version of images with high quality. All these methods significantly produce good results in performance analysis.

As stated earlier, the saliency map of the existing models is generated by either considering the local features or the global ones and the correlation between global and local features is not exploited. The existing models are still in their infancy stage when it comes to handling some complex images with cluttered background, heterogeneous objects or multiple backgrounds. The transform domain approaches of visual saliency detection methods have larger computing speed and need lesser computational effort. Even though the existing WT-based saliency detection models give better results, they fail to represent the directional features in an effective manner. This is overcome by the multi-directional and multi-scale Ripplet transform in order to provoke the directional features in an exhaustive way, because Ripplet transform has some unique properties like multi-resolution, good localization, high directionality, general scaling and support, anisotropy, fast coefficient decay, etc. which are essential to describe images for saliency detection.

In this paper an efficient visual saliency detection method based on Ripplet transform is proposed. The main idea is that the salient regions are quite different from their background and the uncommon regions stand apart in terms of disparate properties of an image such as colour, orientation, texture and shape. Ripplet coefficients can effectively represent the local features of an image in multi-scale and multi-direction [31]. Hence the Ripplet transform is considered to represent the salient regions more effectively.

In this work, we find saliency map with two aspects: global as well as local. The feature maps are created by IRT on multi-level decomposition. Depending on each decomposition level, local variations are recorded in each feature map. The global saliency map is computed by considering the global distribution of local features, that is, the global probability density distribution. Another aspect called *entropy* is leveraged in the proposed method. Geometrically well-organized regions in an image have lower entropy than the disorganized ones [34]. We have used the concept of higher entropy and lower saliency for a disordered signal. The local saliency map is obtained by the idea of entropy and feature distribution. Thus the global and local saliency maps are computed based on the global probability density distribution, and feature distribution of local areas, which are combined together to get the final saliency map.

The proposed saliency detection is considered as a frequency domain analysis problem. For a fair evaluation, only frequency-based state-of-the-art models have been taken for comparison. The proposed method is compared with 10 state-of-the-art methods, namely, IT in Itti *et al* [4], SR in Hou and Zhang [13], SE in Murray *et al* [18], WT in Imamoglu *et al* [3], VS in Li *et al* [15], ST in Bao *et al* [26], WE in Ma *et al* [20], PQFT in Guo and Zhang. [35], HSC in Li *et al* [14] and DQCT in Boris and Stiefelhagen [16] on five benchmark datasets. Except IT [4] all the above state-of-the-art methods are transform domain approaches. IT [4] is the pioneer model of visual saliency detection. The performance of the proposed method is elicited in terms of ROC area, precision, recall, *F* measure and Mean Absolute Error (*MAE*). It can be shown that the proposed method can effectively handle the images with heterogeneous objects, homogenous objects, cluttered background and multiple objects. Both the subjective and objective evaluations reveal that the proposed Ripplet-transform-based saliency detection method has successfully achieved a higher saliency performance on five benchmark datasets compared to the 10 state-of-the-art saliency detection methods.

The proposed method significantly incorporates the local features with the global distribution, in order to obtain the local and global saliency maps. This is not accounted in many of the existing methods. Due to the multi-scale property of Ripplet transform, the saliency map of the proposed method incorporates both the low and high range of frequency components. Hence the saliency map of the proposed method has the same resolution as that of an input image, and the proposed method is applicable for images having different sizes of objects and uniformity. An added advantage is that the Ripplet decomposition of the proposed method continues up to the possible coarsest scale when stretched, in order to get the feature map. Each feature map can represent image details in multi-scale and multi-direction. The uncommon regions in an image are due to the variations in the local features. Since Ripplet transform is well suited to capture the orientation singularities, the Ripplet coefficients can effectively represent the local features of an image. Hence the salient regions are brought out from their surroundings. By this way, the true salient regions are obtained, which are independent of their sizes. The proposed method is experimented with the images involving multiple objects, complex background, single foreground object, two foreground objects, high contrast, etc., which supports the performance of proposed method.

The remaining part of the paper is organized as follows. Section 2 describes the proposed method of Ripplet-transform-based saliency detection, which includes the colour channel transformation (section 2.1), Ripplet transformation and feature map generation (section 2.2), global saliency map computation (section 2.3), local saliency map computation (section 2.4) and final saliency map computation (section 2.5). Section 3 presents the experimental results and quantifies the performance of the proposed method with the state-of-the-art methods. The conclusion of the paper is given in the final section.

2. Proposed model for saliency detection

Saliency is viewed as the unusual regions in the background of an image. The unusual regions are sourced by various properties of an image such as shape, colour, texture, orientation, etc. Any saliency detector could have the capability to find saliency under different conditions. Ripplet transform is very good to represent 2D singularities (edges, textures). Based on the time–frequency analysis of Ripplet transform [36], the Ripplet coefficients are more effective in representing the salient regions. Since the salient regions are quite different from their background, they can be brought out from their background using Ripplet Transform and the local image features can be represented in multi-scale and multi-directions by the Ripplet coefficients. The dissimilarities between unusual region and its background area are effectively represented by the Ripplet coefficients due to the vast orientation and texture information; also the Ripplet transform is very good to seize orientation [33] singularities. Hence the Ripplet transform is used to identify the salient regions in this proposed work.

An example is given in figure 1, where the original images are shown in the top row and the decomposed images in Ripplet domain are shown in the bottom row. In figure 1 the salient regions of the original images are elicited by colour (figure 1c), texture (figure 1d), orientation (figure 1e), shape (figure 1a) and density (figure 1b). It can be observed that the coefficients in the Ripplet domain are in agreement with the saliency depicted by the proper colour channel, scale and orientations accordingly as the case may be.

The block diagram of the proposed saliency detection model is illustrated in figure 2. A novel saliency detection model based on Ripplet transform is proposed. The Ripplet coefficients represent the images in multi-scale and multi-direction. The feature maps are generated by IRT on multi-level decomposition. The local and global saliency maps are obtained based on their contrasts and entropy. They are combined together to get the final saliency map.

2.1 *Colour channel transformation*

Images have different contrasts in the various colour channels. Appropriate colour channels must be used in order to identify the salient regions in an image. Hence the selection of colour channels plays an important role [37] in the detection of salient region in an image. The colour channels of Lab and YCbCr are mainly employed for saliency detection. Recently the colour channels of RGBYI were used [38] for saliency detection. In this proposed method an input image is converted into RGBYI colour channels. The channels of an input image are red r, green g and blue b. Four basic colour channels, namely R, G, B and Y, are computed using Eqs. (2) and (3). Intensity channel is obtained using Eq. (1). Human visual neurons are excited by one colour and inhibited by other colour [4]. The colour pairs of red/green, green/red, blue/yellow and yellow/blue exist in human visual cortex. Colour opponency also plays a major role in visual saliency detection [12]. Human retina contains a large number of photoreceptor cells called cone cells, which are responsive to the variation between green and red, and blue and yellow. The two colour difference channels RG and BY are computed using Eq. (4):

$$I = \frac{r+g+b}{3} \tag{1}$$

$$R = r - \frac{g+b}{2}, G = g - \frac{r+b}{2}, B = b - \frac{r+g}{2} \tag{2}$$

$$Y = \frac{r+g}{2} - \frac{|r-g|}{2} - b \tag{3}$$

$$RG = R - G \text{ and } BY = B - Y. \tag{4}$$

2.2 *Ripplet transformation and feature map generation*

The input colour image I^c is applied with an $n \times n$ 2D Gaussian low pass filter, in order to remove noise:

$$I = I^c * g_{n \times n} \tag{5}$$

where g is the $n \times n$ 2D Gaussian filter ($n = 3$ in this work); $*$ is the convolution operator; I is the noise-free image of I^c and the normalization is performed for each colour channel in the range of [0,255]. The parvocellular ganglion cells of human retina are more sensitive to orientation, colour and texture [39]. This property is imitated by the Ripplet transform [31]. The image sub-bands will be obtained by Discrete Ripplet Transform (DRT) for a number of levels. Four-level RT decomposition is performed ($N = 4$). This decomposition produces 11 sub-bands for each channel in the image. As there are nine channels, we get 99 sub-bands of an image:

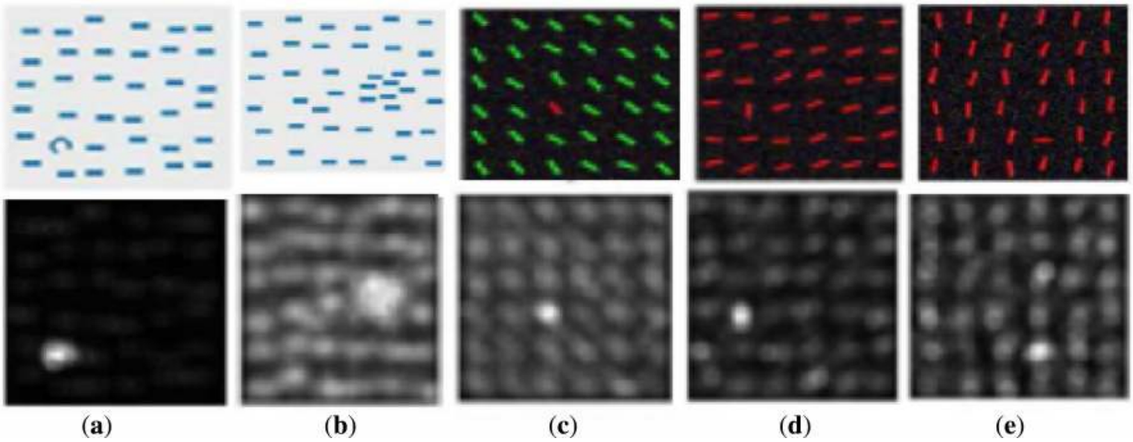

Figure 1. Original images (first row) vs. transformed images in Ripplet domain (second row).

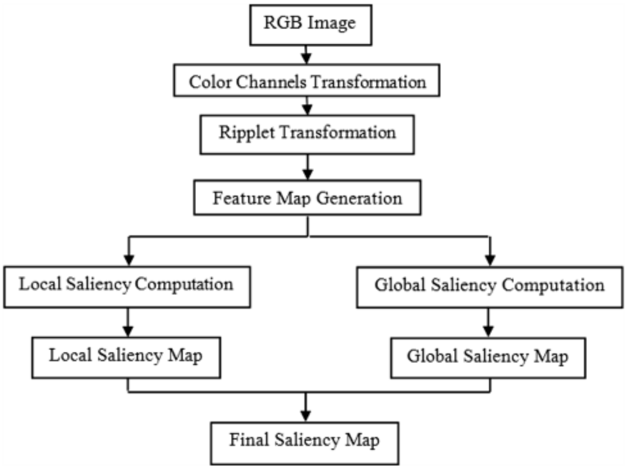

Figure 2. Proposed saliency detection model.

$$R^c_{lk} = DRT_N(I). \tag{6}$$

The channels of noise-free image I are represented by *c; c* ∈{r,g,b,I,R,G,B,RG,BY}; R^c_{lk} is the Ripplet coefficient representing the details of the image at *l* scale and *k* direction (in this work, *k* is taken as 8 for every level). The Ripplet coefficients represent the images in multi-scale and multi-direction. Using the Inverse Ripplet Transform we obtain the feature map $f^c_s(x, y)$, which is generated for *s*-level decomposition for each image sub-band, i.e. $s \in (1,\ldots,N)$. A scaling factor η of 10^3 is introduced to limit the feature map, since there are a large number of feature values in the following equation:

$$f^c_s(x,y) = \left\{IRT\left[R^c_{lk}\right]\right\}^2/\eta. \tag{7}$$

2.3 *Global saliency computation*

After the computation of feature maps, the next stage is to determine the saliency maps. In order to get the global saliency map, the global distribution of local features needs to be considered. Feature vector of the location (*x*,*y*) is represented by *f*(*x*,*y*) with the length of 9 × *N* and it is given in Eq. (8). Hence there are 36 features for each location. The global saliency map is constructed using the method discussed in Imamoglu *et al* [3]. The likelihood of the features in the feature maps is defined by a Probability Density Function (PDF) with normal distribution.

$$\text{Feature vector}\, f(x,y) = [f^c(x,y),\ldots,f^c_N(x,y)]^T. \tag{8}$$

The Gaussian PDF in multi-dimension is given by

$$p(f(x,y)) = \frac{1}{2\pi^{n/2}\left|\sum\right|^{1/2}} \times e^{(-1/2(f(x,y)-\mu)^T\sum^{-1}(f(x,y)-\mu))} \tag{9}$$

where $\sum = E\left[(f(x,y)-\mu)(f(x,y)-\mu)^T\right]$ is an $n \times n$ covariance matrix ($n = 9 \times N$), $\left|\sum\right|$ is the determinant of the covariance matrix, μ is the mean vector of each feature map ($\mu = E[f]$) and, T is the transpose operation.

The covariance matrix and mean of each feature map are features of global statistics. Hence the PDF defined in Eq. (9) can be taken as a global outcome. In order to achieve smooth results, a 2D Gaussian low pass filter ($I_{k\times k}$) is utilized ($k = 5$ in this work):

$$S_{global}(x,y) = \left(\log(p(f(x,y))^{-1})\right)^{0.5} * I_{k\times k} \tag{10}$$

where $S_{global}(x,y)$ is the global saliency map. It includes the statistical relation of feature map in order to provide important information that is not well highlighted by local contrast. Due to the structure of the scene, the result from Eq. (10) gives the saliency map for small salient region and may lose some local information.

The examples of global saliency map are given in figure 3b. It shows how the local features' distribution is balanced with local contrast for a salient region. Figure 3b and c exhibits, respectively, the global and local saliency map, where the salient regions are nearly curbed because of the smooth background. There are some cases where the local contrast is suppressed too much by global saliency (see row 1 of figure 4b and c). For many natural images it is not reasonable to assume that their backgrounds are smooth. In some cases, the global saliency may give potential salient regions that are considered as less salient locally, and similarly in other cases the locally salient region may not manifest as globally salient region as demonstrated in figure 4.

In figure 4c it is observed that the lower part of the wing and head of the bird have high local saliency. While considering the global distribution of local features (see figure 4b), wings are given more prominance rather than the head. Hence it is understood that different saliency maps can be used to adjust the saliency efficiently for local features. For a better saliency map, it is mandatory to add the global information of the local features.

2.4 *Local saliency computation*

Entropy is an important parameter to find the amount of information in an image or scene. It helps to get the local saliency map. The local area of an image usually has salient regions or salient objects [1]. Geometrically well-organized regions in an image have lower entropy than disorganized ones [34]. We have used the concept of higher entropy and lower saliency for a disordered signal. For an image, if the local areas are very much nearer to each other, then the entropies of the local areas become larger. But none of them are deemed salient. If some of the areas are extremely brought out from their backgrounds and their saliencies are

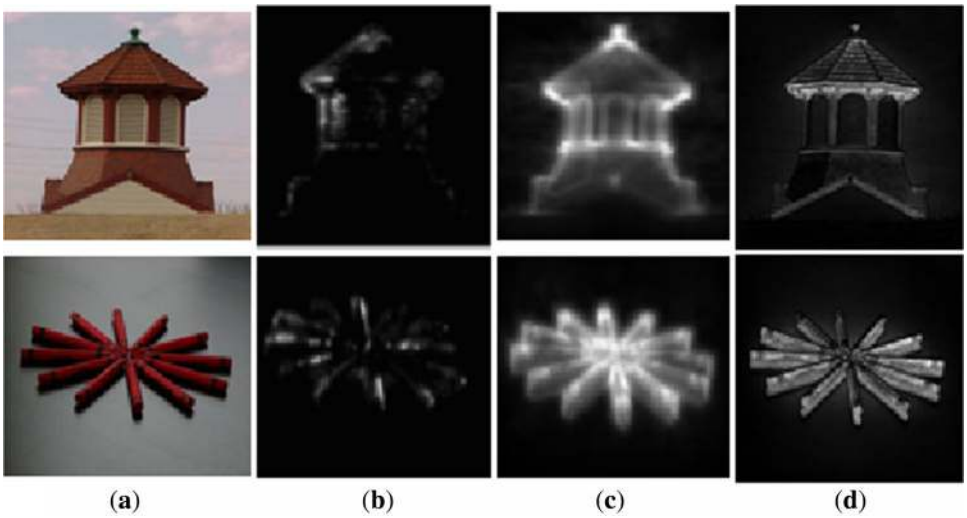

Figure 3. Examples of global and local saliency maps for the images with smooth background: (**a**) colour images, (**b**) global saliency map, (**c**) local saliency map and (**d**) final saliency map.

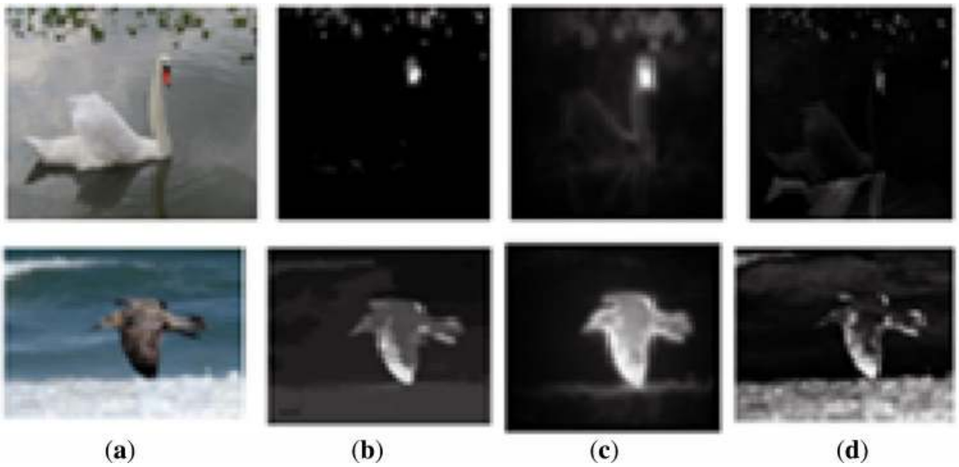

Figure 4. Examples of global and local saliency maps for the images without smooth background: (**a**) colour images, (**b**) global saliency map, (**c**) local saliency map and (**d**) final saliency map.

higher, their entropies are lower. For example, assume that there is a pretty white dove walking around the green grass. For this instance, the pretty white dove is viewed as saliency. The global and local contrasts are considered to detect the global and local saliency maps, which are narrated in the earlier part of this paper. These contrasts are nothing but the texture information. If the pretty white dove has very soft feathers and the green grass has rugged surface, then the green grass may have higher chance of being viewed as salient. Hence, finding the colour differences and contrast in the local area of an image tend to add more importance in the saliency detection process.

The energy in an image is reflected by its entropy, but the distribution of energy cannot be known. In order to obtain the local saliency map more effectively, feature distribution and entropy are used in the proposed method. The examples of proposed local saliency map are given in figures 3c and 4c. It should be noted that there are quite a many variations in some regions between local saliency map (figures 3c and 4c) and global saliency map (figures 3b and 4b).

The feature distribution is obtained first, which represents the probability of saliency of a given feature map $f(x, y)$ in a given colour image I^c. The concept of pixelwise similarity and information theroy is utilized, in order to obtain the feature distribution. Enhancement of the probabilty of a feature appearing in the image is given by a similarity function. Y_p^k is the similarity function of k^{th} feature at position p, given as follows:

$$Y_p^k = \frac{1}{M \times N} \sum_{i \in I^c i \neq p} Q(b|f_p^k - f_i^k| + 1) \qquad (11)$$

$$Q(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{12}$$

where $M \times N$ is the size of the input image and b is a constant ($b = -0.01$ in this work). Smaller contrast between f_p^k and f_i^k yields higher value for $Q.Y_p^k$ can be observed as appearance probability [1] of k^{th} feature at location p, enriched by all the comparable features at other locations of the image I^c. The proposed method can quantify the saliency in multi-scale, because the apperance probability is determined on multi-level. In order to obtain the entropy $\varepsilon_N(x, y)$, the $(k+1) \times (k+1)$ neighbourhood (k is taken as 5 in this work) is used as local area for every location (x, y). At the location $(x, y)$the saliency value is determined by

$$\varepsilon_s^c(x, y) = -\sum_{i=1}^{2k+1} Y_p^k(x_i y_i) \times \log_{10} Y_p^k(x_i y_i) \tag{13}$$

$$\varepsilon_s(x, y) = \sum_c N(\varepsilon_s^c(x, y)). \tag{14}$$

The normalization operator is $N(.)$; The local saliency map is obtained by incorporating $\varepsilon_s^c(x, y)$ for every location (x, y) as

$$S_{local} = \left(\sum_{s=1}^{N} \varepsilon_s(x, y) \right) * g_{n \times n}. \tag{15}$$

S_{local} is the local saliency map.

2.5 *Final saliency map computation*

The local saliency map (Eq. (15)) and global saliency map (Eq. (10)) are combined together [3] in order to obtain the final saliency map:

$$S_{final}(x, y) = M\left(S_{local}(x, y) \times e^{S_{\text{global}}(x, y)} \right). \tag{16}$$

In order to decline the amplification effect modulation, M is introduced on the saliency map, where $M\,(.) = (.)^{\ln\sqrt{2}}/\sqrt{2}$. Goferman *et al* [8] stated that the saliency values around the salient points boost the enhancement of the saliency map performance. The focus of attention points have more impact than those far away from the attention. In this method, points with saliency values greater than 0.7 are assigned as focus of attention points.

$$s(x, y) = S_{final}(x', y')(1 - d_{FOA}(x, y)). \tag{17}$$

At the point (x, y) the saliency value is represented by $s(x, y)$. Let (x', y') be the most salient point. The saliency value of (x', y') is represented by $S_{final}(x', y')$; $d_{FOA}(x, y)$ is the distance between location (x, y) and its nearest FOA at the location (x', y'). In the focused region the saliency values are high, which will be reflected in the final saliency map also.

3. Experimental results

3.1 *Set-up*

The experimental work is carried out for different datasets that are widely used in the saliency detection literature.

[1]MSRA 5000: This datasets includes 5,000 [40] natural images along with their corresponding ground truth. It contains single salient objects of medium sizes per image with a simple background. The ground truth are given by human-labeled rectangles around salient regions. Various images are presented in these dataset, namely natural scenes, animals, indoor, outdoor, etc. MSRA 1000: This dataset is a subset of MSRA-5000 dataset. It includes 1,000 natural images with a new set of ground truth [9].

[2]SED: This dataset contains [41] two sub-datasets, namely SED1 and SED2, which consist of images under different conditions such as single and multiple foregrounds. SED1 dataset comprises100 images with a single salient object. SED2 dataset consists of 100 images with two salient objects. All these 200 images are given with pixel-wise ground truth annotations for the salient objects.

[3]SOD: This dataset includes 300 images [42] from Berkeley Segmentation Dataset (BSD). The boundaries of the salient objects are given for all the 300 images. Binary ground truth of each image is also provided. This dataset contains images with multiple objects in complex background. It is one of the challenging dataset for saliency detection. Detailed characteristics of these benchmark datasets are given in table 1.

Table 1. List of benchmark datasets used.

Name	Size (no of images)	Characteristics
MSRA 5000 [40]	5000	Single object, simple background, high contrast
MSRA 1000 (subset of MSRA 5000) [9]	1000	Single object, simple background, high contrast
SED1 [41]	100	Single foreground object
SED2 [41]	100	Two foreground objects
SOD [42]	300	Multiple objects, complex background

[1] http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salientobject.html

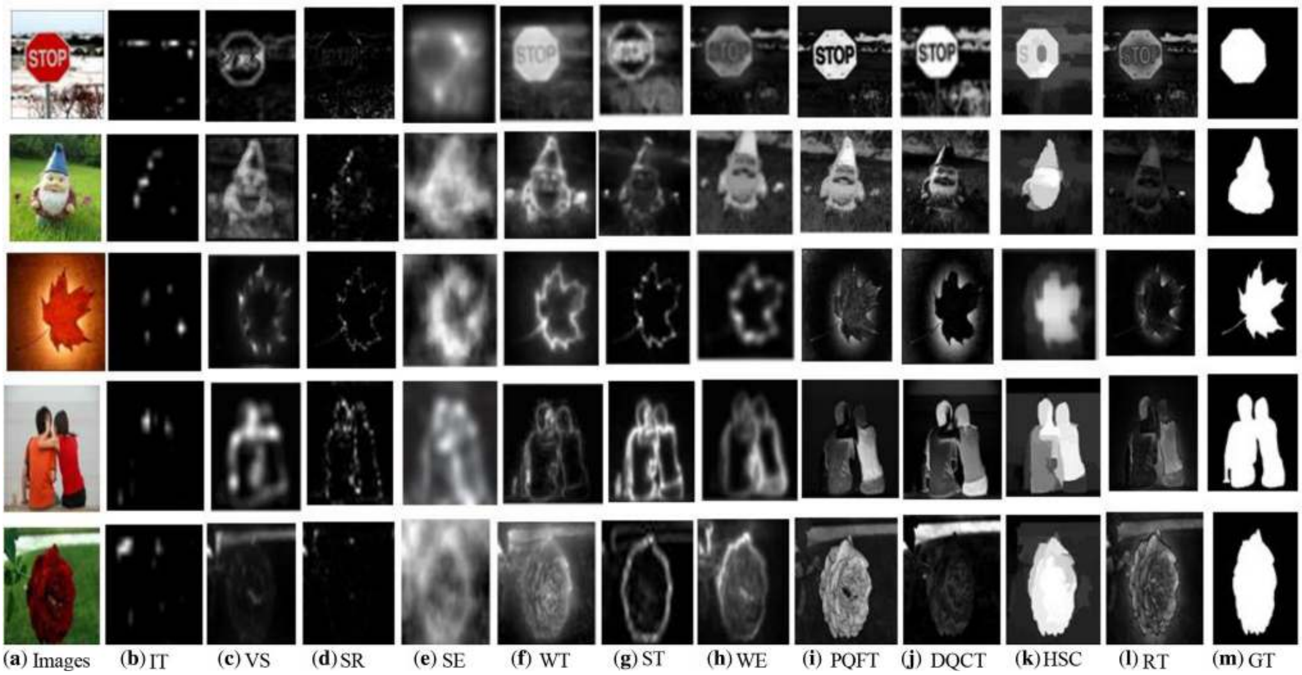

Figure 5. Examples of saliency maps over MSRA-1000 dataset: (**a**) input images, (**b**)–(**k**) saliency maps of other existing methods, (**l**) proposed RT method and (**m**) ground truth.

3.2 *Performance analysis*

The performance of the proposed approach is evaluated by the benchmark datasets arrayed in table 1 and the results are compared with 10 state-of-the-art saliency models (figures 5–8). As discussed earlier, existing models were proposed in different domains. The selected existing methods are IT in Itti *et al* [4], VS in Li *et al* [15], SR in Hou and Zhang [13], SE in Murray *et al* [18], WT in Imamoglu *et al* [3], ST in Bao *et al* [26], WE in Ma *et al* [20], PQFT in Guo and Zhang [35], HSC in Li *et al* [14] and DQCT in Boris and Stiefelhagen [16]. Detailed descriptions of these methods are discussed in the earlier section of this manuscript. The rationale for choosing these models is discussed here.

IT is the pioneer computational model of visual saliency detection. The other nine models are proposed in various transform domains. The VS, SR, PQFT and HSC are Fourier-based models while SE, WT and WE are wavelet-based models. DQCT is Fourier- and cosine-based model and ST is a shearlet-based model. In order to do a fair comparison, all saliency maps are obtained by the state-of-the-art models and they are normalized in the same range [0,255] of original images.

In order to validate the performance of the proposed RT model in saliency detection, Receiver Operating Characteristic (ROC) curve is adopted as one of the objective evaluation metric. The saliency maps are combined with salient regions and non-salient regions. As stated in Imamoglu *et al* [3] the percentage of target points in the GT falling into the salient points of the saliency map is known as True Positive Rate (TPR), while the percentage of the background points falling into the salient points is called False Positive Rate (FPR). The ROC curve is drawn between TPR and FPR. Each point on ROC represents different trade-offs between false positives and false negatives. Figure 9 shows the ROC curve for the proposed and existing methods for the five datasets. This demonstrates the efficacy of the proposed RT model for improving the saliency detection performance. For further analysis, the area under ROC curve (AUC) is also calculated on five datasets. Figure 10 clearly demonstrates the better performance of the proposed RT method in terms of larger ROC area. The proposed RT model thus proves to have effectively contributed in saliency measurement and has yielded improved performance on all the five datasets.

3.3 *Subjective evaluation*

In order to perform subjective comparison, the saliency maps of the proposed method and 10 existing methods on the five benchmark datasets are portrayed in figures 5–8. It can be observed that if the images have simple background and homogenous objects, then the high-quality saliency maps are produced by various models (such as rows 3 and 4 of figure 5). The proposed RT model can highlight the

[2] http://www.wisdom.weizmann.ac.il/~vision/Seg_Evaluation_DB/dl.html

[3] http://elderlab.yorku.ca/SOD/index.html

salient region even though the image background is cluttered (rows 2 and 4 in figure 7, rows 1 and 3 in figure 8 and rows 4 and 5 in figure 6) and images having heterogeneous objects (for example automobiles in figures 7 and 8, and a person in figures 6 and 8). Hence it can be inferred that the images with complex background can be effectively handled by the proposed RT model.Multiple objects in an image (such as rows 1, 2 and 4 in figure 8 and last row in figure 7) can also be effectively highlighted by the proposed RT method. Where many saliency detection models have failed to detect the multiple objects in an image, the proposed method proves its applicability in such scenarios also. The proposed RT model comparatively performs well for large-scale (row 2 of figure 7, last row of figure 5 and rows 1 and 5 of figure 6) and small-scale salient objects (last row of figure 7). Comparing with existing methods, the proposed RT method effectively handles the images with a wide spectrum of features like simple background, homogenous objects, heterogeneous objects, cluttered background and multiple objects.

3.4 *Objective evaluation*

The performance of the proposed method is objectively evaluated by ROC curves, which have been discussed in section 3.2. In addition to this, the parameters precision (P), recall (R) and F measure provide the objective evaluation. These parameters are given as follows:

$$P = \frac{\sum_x \sum_y ((GT(x,y) \times SM(x,y))}{\sum_x \sum_y SM(x,y)} \qquad (18)$$

$$R = \frac{\sum_x \sum_y ((GT(x,y) \times SM(x,y))}{\sum_x \sum_y GT(x,y)} \qquad (19)$$

$$F_\alpha = \frac{(1+\alpha)PR}{\alpha P + R} \qquad (20)$$

where $GT(x,y)$ is the ground truth and $SM(x,y)$ is the saliency map of the proposed method; α is taken as 0.3 in this work [40].

Precision P relates the percentage of salient pixels correctly assigned. Recall R relates the correct detection of salient regions in accordance with ground truth. F measure F_α is the harmonic mean of P and R. These performance measures have been calculated for both the proposed model and the existing models. The Otsu automatic threshold algorithm [43] and mean value of saliency map are used to obtain binary images of the proposed saliency map.

The ROC curves and PR charts are given in figures 9 and 11, respectively. It is evident from ROC curves and PR charts in figures 9 and 11 that the proposed RT method yields superior performance compared with those of extant

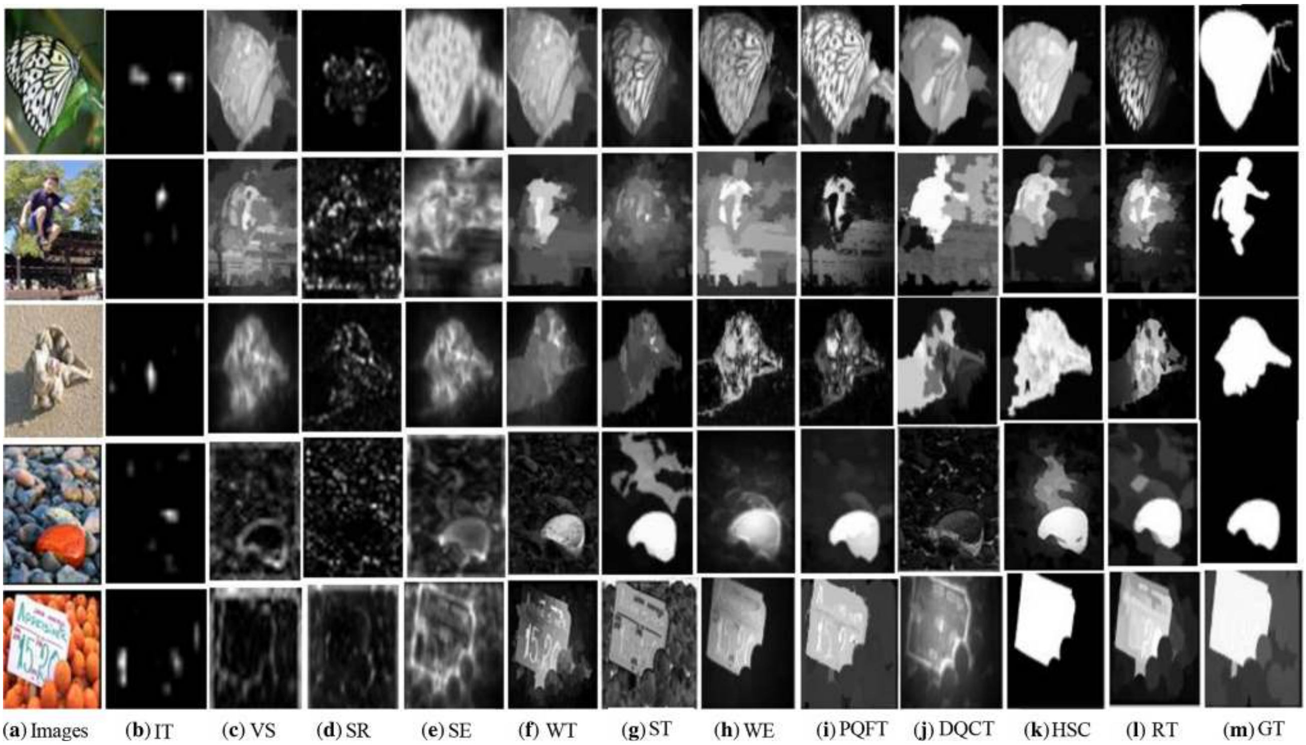

Figure 6. Examples of saliency maps over MSRA-5000 dataset: (**a**) input images, (**b**)–(**k**) saliency maps of other existing methods, (**l**) proposed RT method and (**m**) ground truth (GT).

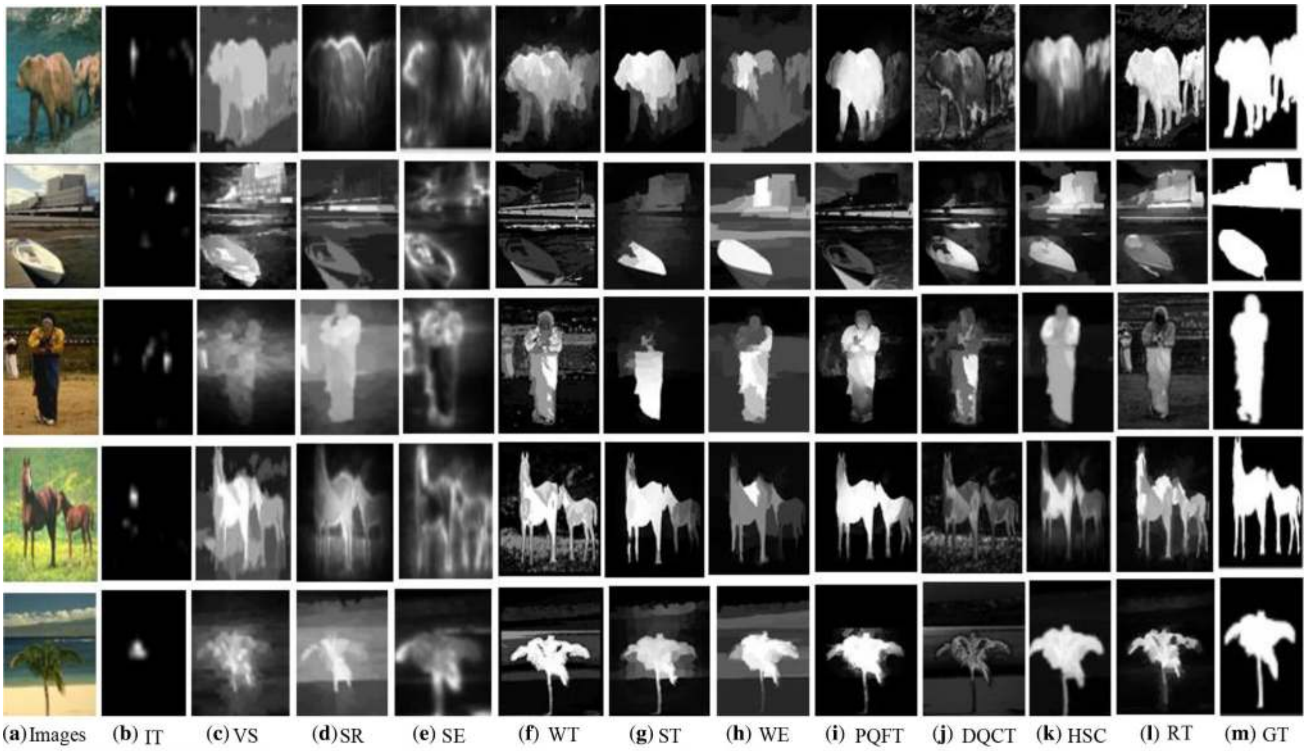

Figure 7. Examples of saliency maps over SED 1 (row 1–3) and SED 2 (row 4–6) datasets: (**a**) input images, (**b**)–(**k**) saliency maps of other existing methods, (**l**) proposed RT method and (**m**) ground truth.

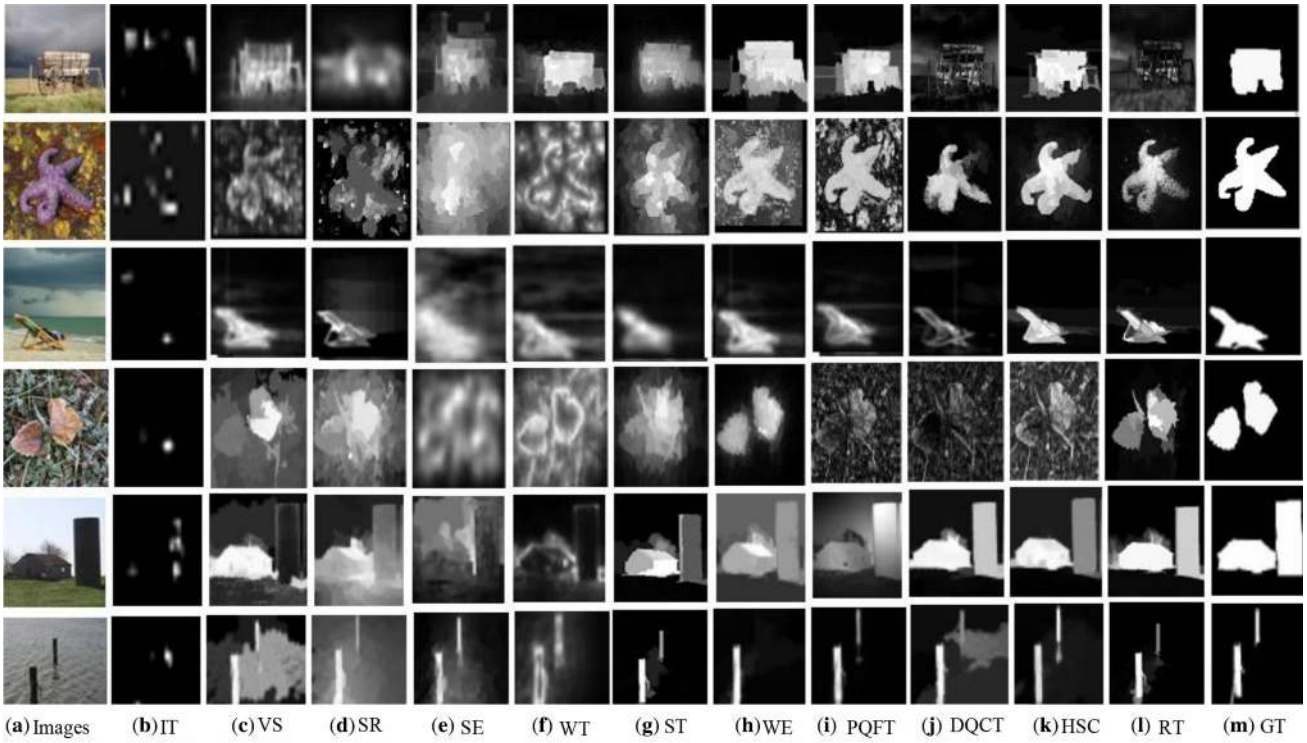

Figure 8. Examples of saliency maps over SOD dataset: (**a**) input images, (**b**)–(**k**) saliency maps of other existing methods, (**l**) proposed RT method and (**m**) ground truth.

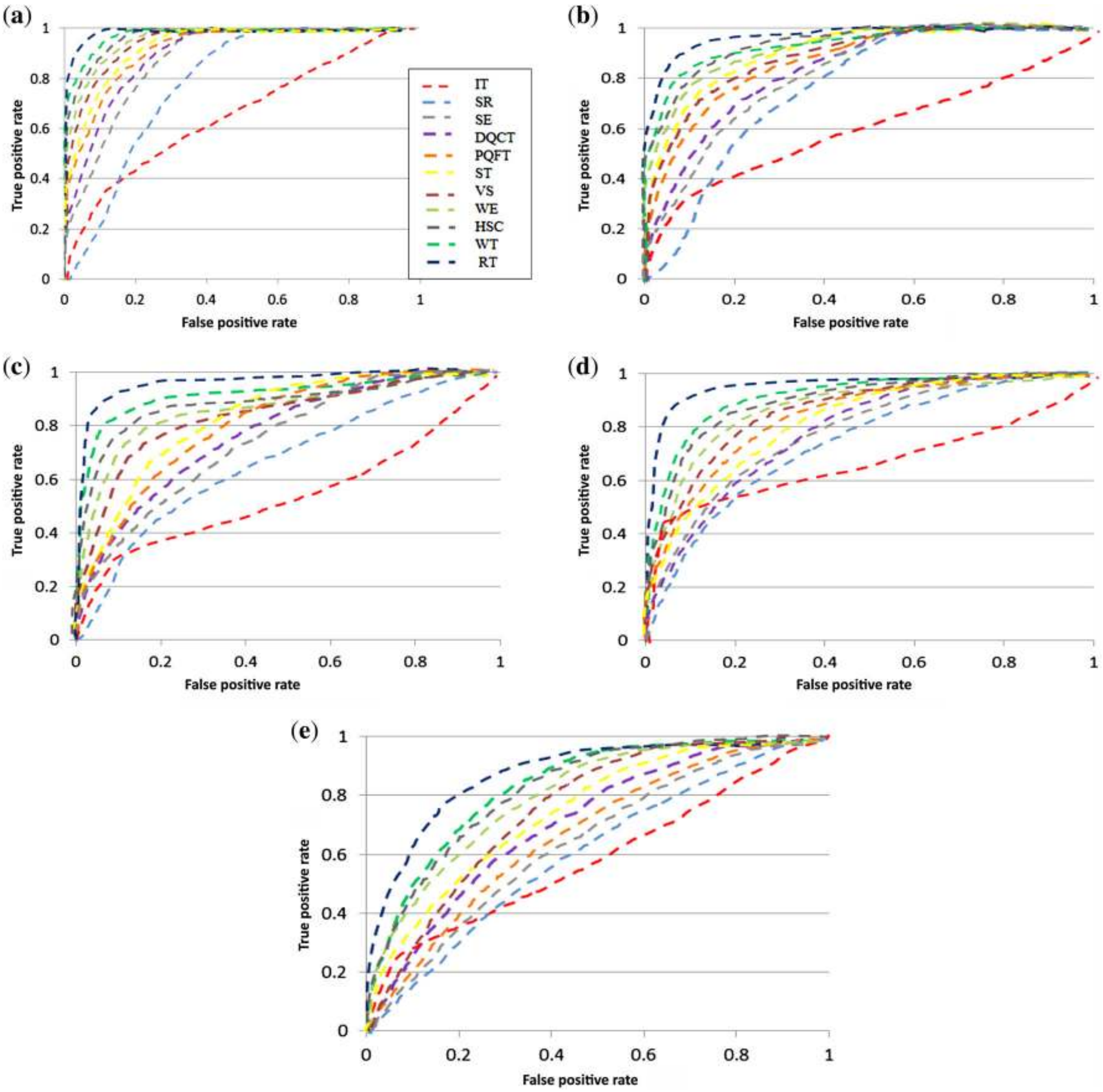

Figure 9. ROC of the proposed RT method and other existing methods: (**a**) MSRA 1000, (**b**) MSRA 5000, (**c**) SED 1, (**d**) SED 2 and (**e**) SOD.

methods. Figure 10 shows the ROC area of all five benchmark datasets, wherein the proposed method is seen to consistenly outperform all the state-of-the-art-methods on all the five datasets. MSRA-1000 dataset consistently scores high ROC area for all the methods and SOD dataset consistently scores low ROC area. Hence, the SOD dataset remains challenging for saliency detection.

According to Perazzi *et al* [11], Mean Absolute Error (*MAE*) may be taken as the test criterion that gives a more balanced comparison between continuous saliency map and the binary ground truth. If the whole salient regions are uniformly highlighted by the saliency map then *MAE* will be very small.

The *MAE* is defined as follows:

$$MAE = \frac{1}{N \times M} \sum_{x=1}^{N} \sum_{y=1}^{M} |SM(x,y) - GT(x,y)| \qquad (21)$$

where N and M are, respectively, the width and the height of the respective saliency map and ground truth image. *MAE* yields better estimation of the dissimilarities between the saliency map (prior to thresholding) and binary ground truth. Figure 12 indicates that the proposed method significantly outperforms the other state-of-the-art-approaches in terms of *MAE* measure.

Three test criteria were taken for the objective evaluation of the proposed method. The overall performance of the proposed RT model is good under both subjective and objective analysis and the outcomes are reliable. The

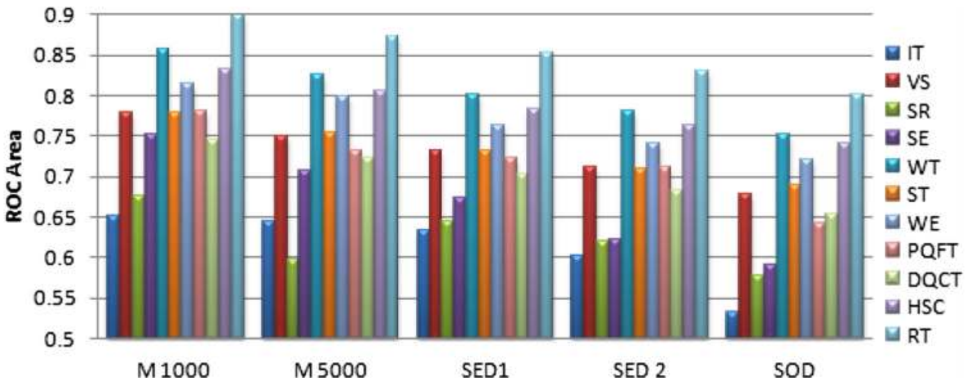

Figure 10. AUC of the proposed RT method and other existing methods.

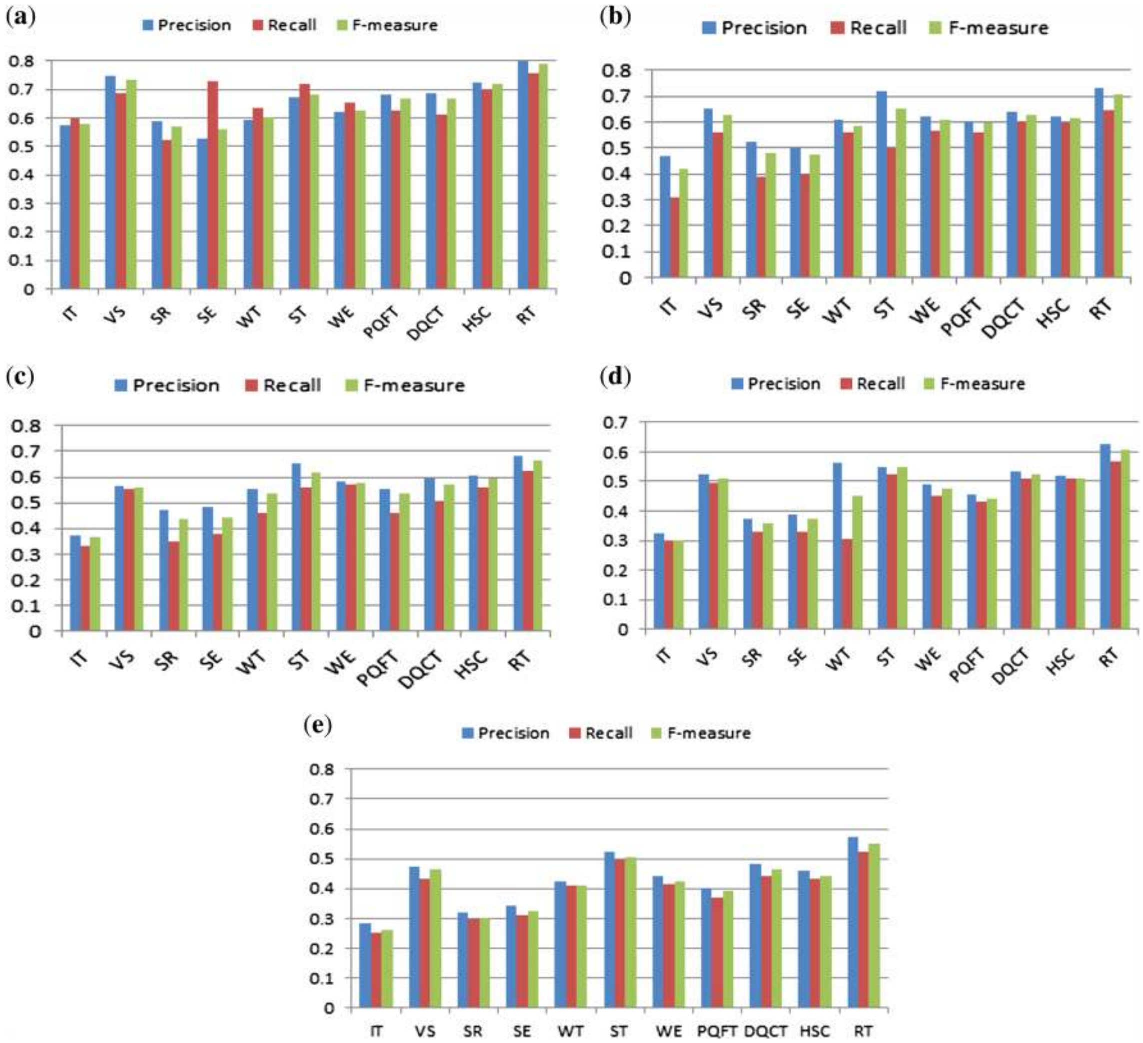

Figure 11. Precision, recall and *F* measure of the proposed RT method and other existing methods: (**a**) MSRA 1000, (**b**) MSRA 5000, (**c**) SED 1, (**d**) SED 2 and (**e**) SOD.

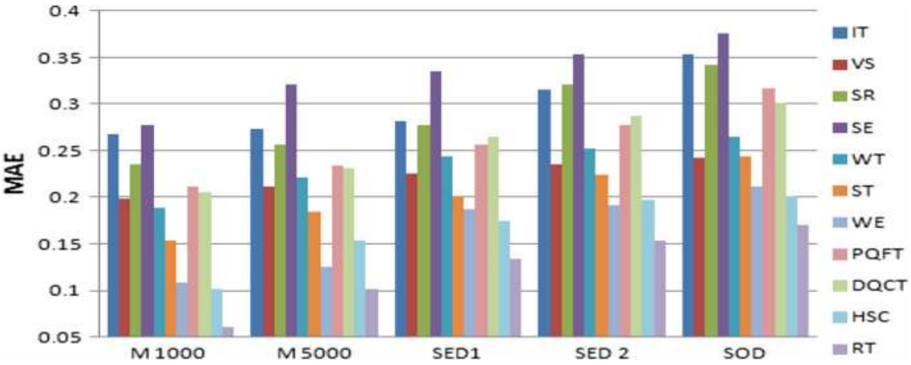

Figure 12. *MAE* of the proposed RT method and state-of-the-art-methods on the five datasets.

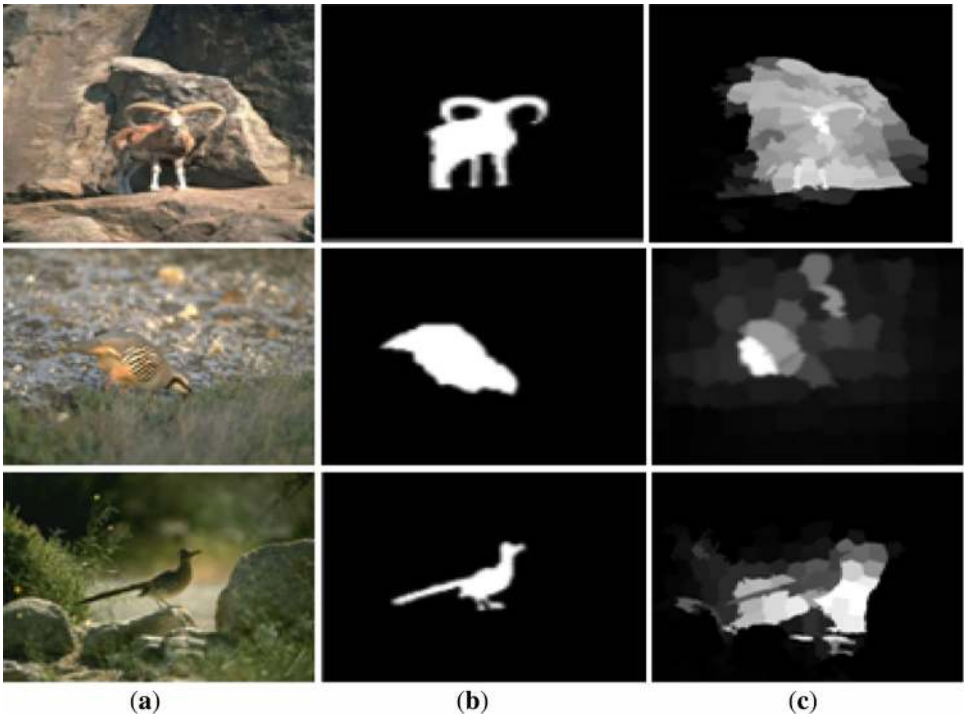

Figure 13. Some examples of failure cases: (**a**) image, (**b**) ground truth and (**c**) proposed RT method.

accuracy or efficiency of the proposed RT model purely depends on the applications one may choose.

3.5 *Failure analysis*

As discussed in the previous sub-sections the proposed RT model outperforms the state-of the art models in terms of three popular criteria, i.e. ROC, Precision–Recall–*F* measure and *MAE*. However, there exist some images that are still challenging for saliency detection task in the proposed RT method as well as other existing methods. If the image contains a salient object whose colour matches with that of its background then the saliency is wrongly highlighted. Figure 13 shows counter-examples of the proposed RT model. In these images, the salient objects and the background are confusingly blended with each other. Such cases are not handled effectively by the proposed method and as well as by the state-of-the-art methods. These problems will be worked out in the near future.

4. Conclusions

To conclude this paper, a novel saliency detection model using Ripplet transform has been presented. The main focus is on detection of salient regions in the Ripplet domain for achieving higher Receiver Operating Characteristics (ROC). The novelty of the work lies in using the Ripplet transform in the detection of saliency in a visual scene. The

Ripplet transform can represent singularities along arbitrarily shaped curves and thereby it yields feature maps with different scales and in different directions. This characteristic of Ripplet transform is leveraged in the computation of global saliency map and local saliency map based on global probability density distribution and feature distribution of local areas. Experimental results convincingly show that the proposed Ripplet-transform-based visual saliency detection model has better performance compared with the 10 state-of-the-art models on five benchmark datasets. As a future development, the depth from the focus cue can be utilized to enhance the visual quality of the saliency map.

References

[1] Lin R J and Lin W S 2014 A Computational visual saliency model based on statistics and machine learning. *J. Vis.* 14(9): 1–18

[2] Huang J B and Ahuja N 2012 Saliency detection via divergence analysis: a unified perspective. In: *Proceedings of the International Conference on Pattern Recognition, Tsukuba*, pp 2748–2751

[3] Imamoglu N, Lin W S and Fang Y M 2013 A saliency detection model using low-level features based on wavelet transform. *IEEE Trans. Multimedia* 15(1): 96–105

[4] Itti L, Koch C and Niebur E 1998 A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20(11): 1254–1259

[5] Ma Y F and Zhang H J 2003 Contrast-based image attention analysis by using fuzzy growing. In: *Proceedings of the ACM International Conference on Multimedia*, Berkeley, pp 374–381

[6] Oliva A, Torralba A, Castelhano M S and Henderson J M 2003 Topdown control of visual attention in object detection. In: *Proceedings of the IEEE International Conference on Image Processing*, pp 253–256

[7] Harel J, Koch C and Perona P 2006 Graph-based visual saliency. In: *Proceedings of Advances in Neural Information Processing Systems 19*, Cambridge, USA: MIT Press, pp. 545–552

[8] Goferman S, Zelnik Manor L and Tal A 2010 Context-aware saliency detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 2376–2383

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

[10] Cheng M M, Zhang G X, Mitra N J, Huang X and Hu S M 2015 Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 37(3): 569–582

[11] Perazzi F, Krahenbuhl P, Pritch Y and Hornung A 2012 Saliency filters: contrast based filtering for salient region detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 733–740

[12] Guo C, Ma Q and Zhang L 2008 Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 1–8

[13] Hou X and Zhang L 2007 Saliency detection: a spectral residual approach. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 1–8

[14] Li C, Xue J, Zheng N, Lan X and Tian Z 2013 Spatio-temporal saliency perception via hypercomplex frequency spectral contrast. *Sensors* 13: 3409–3431

[15] Li J, Levine M D, An X, Xu X and He H 2013 Visual saliency based on scale-space analysis in the frequency domain. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(4): 996–1010

[16] Boris S and Stiefelhagen R 2012 Quaternion-based spectral saliency detection for eye fixation prediction. In: *Proceedings of the European Conference on Computer Vision*, pp 116–129

[17] Merry R J E 2005 *Wavelet theory and application – a literature study.* The Netherlands: Eindhoven University of Technology

[18] Murray N, Vanrell M, Otazu X and Parraga C A 2011 Saliency estimation using a non-parametric low-level vision model. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 433–440

[19] Oakes M and Abhayaratne C 2012 Visual saliency estimation for video. In: *Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services*, Dublin, pp 1–4

[20] Ma X, Xie X, Lam K M and Zhong Y 2015 Efficient saliency analysis based on wavelet transform and entropy theory. *J. Vis. Commun. Image R.* 30: 201–207

[21] Laurent C, Laurent N, Maurizot M and Dorval T 2006 In depth analysis and evaluation of saliency-based color image indexing methods using wavelet salient features. *Multimedia Tools Appl.* 31(1): 73–94

[22] Li J, Tian Y, Huang T and Gao W 2010 Probabilistic multi-task learning for visual saliency estimation in video. *Int. J. Comput. Vis.* 90(2): 150–165

[23] Li Z Q, Fang T and Huo H 2010 A saliency model based on wavelet transform and visual attention. *Sci. China Inf. Sci.* 53(4): 738–751

[24] Tian Q, Sebe N, Lew M S, Loupias E and Huang T S 2001 Image retrieval using wavelet-based salient points. *Electron. Imaging* 10(4): 835–849

[25] Candes E and Donoho D 2004 New tight frames of curvelets and optimal representations of objects with piecewise singularities. *Commun. Pure Appl. Math.* 57(2): 219–266

[26] Bao L, Lu J, Li Y and Shi Y 2014 A saliency detection model using shearlet transform. *Multimedia Tools Appl.* 74(11): 4045–4058

[27] Candes E J 1998 *Ridgelets: theory and applications*. Dissertation, Department of Statistics, Stanford University, Stanford

[28] Candes E J and Donoho D 1999 Ridgelets: a key to higher-dimensional intermittency. *Philos. Trans. Math. Phys. Eng. Sci.* 357(1760): 2495–2509

[29] Ma J W and Plonka G 2010 The curvelet transform. *IEEE Signal Process. Mag.* 27(2): 118–133

[30] Zhong S H, Liu Y, Shao L and Wu G 2011 Unsupervised saliency detection based on 2D Gabor and Curvelets transforms. In: *Proceedings of the International Conference on Internet Multimedia Computing and Service (ICIMCS'11)*, pp 146–149

[31] Chowdhury M, Das S and Kundu M K 2013 A Ripplet transform based statistical framework for natural color image retrieval. In: *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, pp 492–502

[32] Geng P, Huang M, Liu S, Feng J and Bao P 2014 Multifocus image fusion method of Ripplet transform based on cycle spinning. *Multimedia Tools Appl.* 75(17): 10583–10593

[33] Juliet S, Rajsingh E B and Ezra K 2016 A novel medical image compression using Ripplet transform. *J. Real-Time Image Proc.* 11(2): 401–412

[34] Duncan K and Sarkar S 2012 Relational entropy-based saliency detection in images and videos. In: *Proceedings of the IEEE International Conference on Image Processing*, pp 1093–1096

[35] Guo C and 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

[36] Xu J, Yang L and Wu D O 2010 Ripplet—a new transform for image processing. *J. Vis. Commun. Image R.* 21(7): 627–639

[37] Luo W, Li H, Liu G and Ngan K N 2012 Global salient information maximization for saliency detection. *Signal Process. Image Commun.* 27(3): 238–248

[38] Borji A 2012 Boosting bottom-up and top-down visual features for saliency estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 438–445

[39] Kaplan E and Shapley R M 1986 The primate retina contains two types of ganglion cells, with high and low contrast sensitivity. *Proc. Natl. Acad. Sci.* 83(8): 2755–2757

[40] Liu T, Sun J, Zheng N N, Tang X and Shum H Y 2011 Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(2): 353–367

[41] Alpert S, Galun M, Basri R and Brandt A 2007 Image segmentation by probabilistic bottom-up aggregation and cue integration. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 1–8

[42] Movahedi V and Elder J H 2010 Design and perceptual validation of performance measures for salient object segmentation. In: *Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp 49–56

[43] Gonzalez R C, Woods R E and Eddins S L 2004 *Digital signal processing using Matlab.* Englewood Cliffs, NJ: Prentice Hall