A Saliency Detection Model Using Low-Level Features Based on Wavelet Transform

Nevrez İmamoğlu, Weisi Lin, *Senior Member, IEEE*, and Yuming Fang

***Abstract*—Researchers have been taking advantage of visual attention in various image processing applications such as image retargeting, video coding, etc. Recently, many saliency detection algorithms have been proposed by extracting features in spatial or transform domains. In this paper, a novel saliency detection model is introduced by utilizing low-level features obtained from the wavelet transform domain. Firstly, wavelet transform is employed to create the multi-scale feature maps which can represent different features from edge to texture. Then, we propose a computational model for the saliency map from these features. The proposed model aims to modulate local contrast at a location with its global saliency computed based on the likelihood of the features, and the proposed model considers local center-surround differences and global contrast in the final saliency map. Experimental evaluation depicts the promising results from the proposed model by outperforming the relevant state of the art saliency detection models.**

***Index Terms*—Feature map, saliency detection, saliency map, visual attention, wavelet transform.**

I. Introduction

VISUAL attention is one of the primary features of the Human Visual System (HVS) to derive important and compact information from the natural scenes [1], [2]. Since the surrounding environment includes an excessive amount of information, visual attention mechanism enables a reduction of the redundant data which benefits perception during the selective attention process [1]–[4]. Many studies have tried to build computational models to simulate this mechanism [5], [6].

There are two types of the visual attention mechanism (and therefore two types of computational modeling as well): bottom-up and top-down approaches. The bottom-up approach is stimulus-driven, mostly obtained from early features, and task independent [7], [8]. However, the top-down approach, which is goal-driven, consists of high-level data processing and prior knowledge to support the tasks such as object recognition, scene classification, target detection, identification of the contextual information, etc. [9], [10]. Both of the computational models aim at generating saliency maps to detect the salient regions for images. We are to explore bottom-up visual attention modeling in this work.

Manuscript received August 19, 2011; revised December 12, 2011 and March 23, 2012; accepted June 03, 2012. Date of publication October 16, 2012; date of current version December 12, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Feng Wu.

The authors are with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: nimamoglu@ntu.edu.sg; wslin@ntu.edu.sg; fa0001ng@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2012.2225034

One of the earliest computational models of bottom-up visual attention was proposed by Itti *et al.* [10], [11]. The algorithm obtains the saliency map based on the intensity, color, and orientation conspicuity maps [11]. These conspicuity maps are attained by across-scale addition of feature maps, while the feature maps capture the center-surround differences between various Gaussian pyramid and oriented pyramid scales [11]. Since the saliency map is computed in coarser scales, local information loss is unavoidable in this algorithm. To detect global conspicuity differences between feature maps, Itti *et al.* [10], [11] gave two different types of normalization for the feature maps: 1) global non-linear normalization (being simple and suitable for real-time applications but not biologically plausible); 2) iterative filtering with the difference of Gaussian (biologically plausible but computationally expensive). However, these normalization and iterative filtering procedures are not sufficiently good to relate global feature differences with the correlation among feature dimensions, since each map is normalized regardless of the statistical information of other maps even though these two normalization schemes provide some global reasoning. Another way of combining the feature maps was proposed by Oliva *et al.* [12]. In [12], the feature maps are generated from the spatial decomposition of the color sub-bands and Gaussian distribution of the local features is computed for the saliency map. Saliency map computation based on a probability density function avoids the need of normalization and pooling operations that are necessary in [10]. However, it emphasizes global contrast more than the local contrast even it is based upon the local features.

Some studies try to build visual attention models by taking color contrast information into account in various ways [13], [14]. Achanta *et al.* [14] firstly obtained the Gaussian-filtered image and then used its arithmetic mean vector to derive the saliency map instead of the spatial decomposition. The CIE Lab color space is used for each image location to form a feature vector, and then, the absolute difference between the Gaussian-blurred image and the arithmetic mean vector is calculated to derive the saliency map [14]. The saliency map is processed by the mean-shift segmentation algorithm to extract the regions of interest in the input image for better performance. This approach is mostly good for images with large and homogeneous objects with clear boundaries [14]. Hence, it has limitations on applications because of its dependency on object size and uniformity.

Recent studies have tried to obtain the saliency map for images in the transform domain [15]–[17]. The Fourier Transform (FT) can be expressed with the polar form as two different components: phase and amplitude spectrums. Stated by Oliva *et al.* [15], the former carries information concerning the form and the position of local image structure; the latter holds the information of the global composition of the image that relates to the

overall scene layout [15]. Thus, the eccentricity or irregularities on spectral information can lead to attention regions of the scene [16]. Hou *et al.* first defined the saliency residual with log amplitude spectrum [16], and the saliency map is derived by applying the inverse FT on an exponential function which combines spectral residual and phase spectrum information [16]. Guo *et al.* [17] identified the input image frame in the spatio-temporal domain with four features: motion, intensity, and two chromatic colors. These four features are used as the quaternion data for an input scene. In this model, the quaternion FT (QFT) is firstly calculated, and then, the amplitude spectrum values are set to a constant value for each QFT feature so that the saliency map is constituted from the inverse QFT of the phase spectrum data [17]. In [16] and [17], the FT is the key for the final saliency map. Therefore, these models result in a saliency map in which the global irregularities of the scene can be more dominant than the local irregularities. Another disadvantage of this model is the high down-sampling requirement for images, which would yield spatial information loss. Moreover, it is known that FT may encounter difficulties and lead unsatisfactory results with non-stationary or a-periodic signals [18], [19], since such signals are better to be analyzed locally rather than globally. Hence, schemes in [16] and [17] have limitations regarding the image size requirement and saliency information used.

Recently, Wavelet transform (WT) has begun to attract researchers' effort in visual attention modeling [20], [21]. The advantage of the WT is its ability to provide multi-scale spatial and frequency analysis at the same time [18]. Tian *et al.* [20] proposed a WT-based salient detector for image retrieval which depends on local and global variations of wavelet coefficients at multi-scale levels. The idea is to account for both global and local features in the WT space: the points with higher global variation based on the absolute wavelet coefficients at courser scales are selected, and these selected points are tracked along the finer scales to detect the salient points [20]. The salient points will be the sum of the wavelet coefficients at tracked points along the multi-scale WT [20]. However, this algorithm is only able to give salient points rather than salient regions compared to the models [11], [14] and [16]. Therefore, even though salient points tracked at the finest resolution of the WT can represent the image for image retrieval [20], it is hard to compare this algorithm for attention region or object detection as a computational model of saliency detection. Murray *et al.* [21] derived weight maps from the high-pass wavelet coefficients of each level based on the WT decomposition. Two important aspects for the derivation of weight maps are the consideration of the human sensitivity to local contrast, and the energy ratio of central and surrounding regions to mimic the center-surround effect [21]. In [21], the saliency map is obtained by the inverse WT (IWT) of weight maps for each color sub-band [21]. Although WT representation is better than FT for images by providing more local analysis, there is lack of accounting for global contrast because the computation is based on the local differences in [21]. Hence, the local contrast is more dominant than the global contrast in the saliency model of [21].

In this paper, we propose a novel saliency detection model based on high-pass coefficients of the wavelet decomposition after eliminating some high-frequency components of the image. The idea is to create the feature maps by IWT on the multi-level decomposition. Each feature map contains local variations in multi-scale resolution; in other words, they represent band-pass local information with different frequency bandwidths. The advantage of the proposed model is that we create more detailed feature maps (edge to texture) by applying IWT on various decomposition levels. This helps to observe the irregularities with different bandwidths. Then, two saliency maps are created: local and global saliency maps. Here, we create these two saliency maps for two reasons: i) to avoid the normalization of each feature map separately which is not efficient for considering the statistical relation among the feature maps for the saliency in a global perspective; ii) to incorporate local and global saliency as two different maps to make sure of taking both local and global contrast into consideration sufficiently. Finally, the local and global maps are combined to yield the final saliency map.

As described above, the existing visual attention models are proposed in different domains (spatial, FT, and WT) [11]–[21]. For comparison purpose in this work, the most relevant models to the proposed method are [11] and [14] in the spatial domain, [16] in the FT domain, and [21] in the WT domain. These existing models generate saliency maps while considering either local features or global ones more than the other, or they are not good enough to relate global features from the local ones as stated earlier. On the contrary, the proposed saliency detection model includes both local and global saliency information by integrating local feature differences with the global distribution of these features. The global saliency is based on the likelihood of local features for a given location as in [12]. It also takes the statistical relation among the feature maps into account. In the proposed model, the saliency map includes both narrow and wide range of frequency components due to the multi-scale derivation. Thus, the proposed model can be used for images with different object sizes and extents of uniformity. It also generates the saliency map with the same resolution of the input image. Another advantage is that the wavelet decomposition in the proposed model continues until the possible coarsest scale is reached, in order to attain the feature maps. By this way, we can obtain salient regions independent from their sizes or uniformity due to the features with high contrast from edge to texture since the salient points can be defined as the feature variations of the location with its surroundings based on the WT coefficients [20]. Experimental results demonstrate that the proposed algorithm produces better performance with respect to the relevant existing models.

The rest of this paper is organized as follows: Section II first introduces the concept of wavelet decomposition, and then presents the proposed model in detail, with proper analysis and justification; experimental results and evaluation for the proposed model are given in Section III; the final section gives conclusions for the paper.

II. The Proposed Saliency Detection Model

A. Wavelet Analysis

Even though wavelets were firstly introduced in the early 20th century by Alfred Haar, most of the developments in this area have been progressed since the late 20th century [18]. Recently, the use of wavelets in signal analysis has been

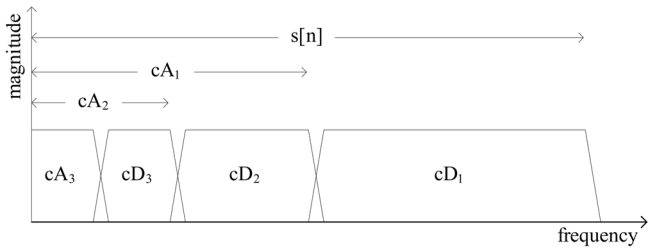


Fig. 1. Illustration of frequency components for a 3-level wavelet decomposition [19]: s[n] is the 1-D signal which covers all the frequency components; cA_1, cA_2, and cA_3 are the approximation data for low-pass outputs to depict the frequency content for each level of decomposition; cD_1, cD_2, and cD_3 are the details of the signal s[n] as the high-pass outputs to depict the frequency content for each level of the decomposition.

rapidly increasing in many engineering applications such as signal de-noising, compression, enhancement, video coding, and pattern classification, etc.

The signal analysis for frequency components can be achieved by FT in a global context, but it is not possible to make time-frequency analysis with FT (i.e., local frequency components cannot be obtained [18], [19], [22], [23]); the short-time-Fourier transform (STFT) can be utilized to perform local frequency analysis, since it yields the frequency information of a given time (so the technique is referred as time-frequency analysis) [18], [19], [22], [23]. It should be noted that STFT can be applied by taking the spatial interval instead of the time interval while dealing with the image since there is not any time information for still images. However, there are some limitations with this technique. With STFT, there is a constant resolution with spatial and frequency analysis, so the success of the application depends on the selection of the spatial interval [18], [19], [22], [23]. In addition, trying several spatial intervals on the image increases the computation time for the application.

The multi-scale wavelet analysis is able to perform better local frequency analysis since it examines the signal at different bands and bandwidths [18], [19], [22], [23]. Wavelet analysis is a process of applying multi-resolution filter-banks to the input signal [18], [23]. One property of the orthogonal wavelet filter-banks is that the approximation and detailed signals are obtained from the two frequency bands: low-pass and high-pass, respectively [18], [19], [23]. For the three-level wavelet decomposition, the frequency components can be simply expressed as 1-D data, as shown in Fig. 1 [19].

The philosophy behind the saliency generation is to create features and feature maps which represent the contrast or center-surround difference, taking both local and global factors into account. For example, in [14], the band-pass filter is stated to obtain the contrast values in the image for the saliency information, so this information can be demonstrated by the band-pass output of the image as in [11], [14]. Hence, among the band-pass regions, the visually dominant ones are most likely to be the salient regions in the scene. The wavelet decomposition has the advantage in extracting oriented details (horizontal, vertical and diagonal) in the multi-scale perspective, and enables high spatial resolution with higher frequency components and low spatial resolution with lower frequency components without information loss in details during the decomposition process [18], [23]. Therefore, the WT with selected decomposition levels can provide feature maps in the band-pass regions without low-frequency content after reconstruction by neglecting approximation signals. In Fig. 1, it can be seen that without the approximation data, for lower levels of IWTs, information with smaller bandwidths and higher-frequency components will be included in the feature map. However, for higher levels of IWTs, larger bandwidths with more frequency components will be included in the feature map after the reconstruction process. This means that reconstruction with details from fine to coarse scales will generate feature maps representing the edge to texture differences.

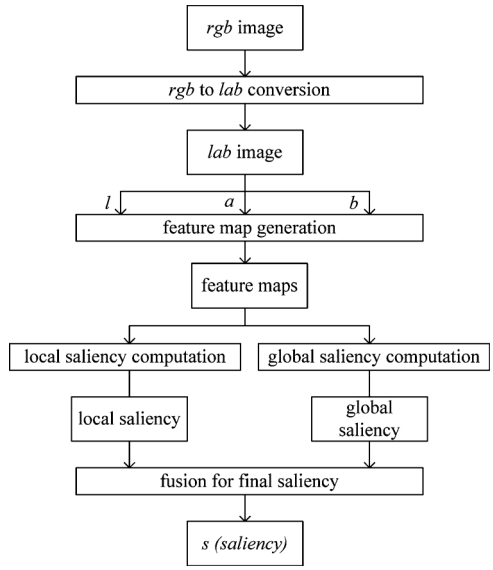


Fig. 2. The framework of the proposed saliency detection model (l: intensity channel, a: red-green color channel, and b: blue-yellow color channel).

B. Overview of the Proposed Model

The proposed model is to create the feature maps by increasing the bandwidths or the frequency components from higher to lower values. As shown in Fig. 2, it integrates two different maps referred as local and global saliency maps. Both maps are attained by features of the same level, based upon the wavelet coefficients.

C. Feature Map Generation

The first step of the proposed model illustrated in Fig. 2 starts with the computation of feature maps. First of all, instead of using *rgb* color space for saliency detection, an image is converted to the *CIE Lab* color space (CIE illumination *D65* model is selected for conversion as the *white-point* parameter in *Matlab*® *rgb* to CIE *Lab* converter). The conversion is needed due to the fact that the Lab color space is uniform and similar to the human perception, with a luminance and two-chromatic channels (*RG* and *BY*) [9]. To remove noise, we apply an $m \times m$ 2D Gaussian low-pass filter to the input color image $\mathbf{g}^c$:

$$\mathbf{g}'^c = \mathbf{g}^c * \mathbf{l}_{m \times m} \tag{1}$$

where $\mathbf{l}$ is the $m \times m$ 2-D filter; $\mathbf{g}'^c$ is noise-removed version of g^c; $*$ denotes the convolution operation. For noise reduction, a small filter size is selected ($m = 3$ in this work) to filter very

high frequency noise [14]. Then, each channel is normalized to the range of $[0, 255]$.

The sub-bands of the image will be formed by WT for a number of levels, as to be shown in (2). Daubechies wavelets (Daub.5) are chosen since its filter size is appropriate for pixel neighborhoods, computation time, and the overall result.

$$[\mathbf{A}_N^c, \mathbf{H}_s^c, \mathbf{V}_s^c, \mathbf{D}_s^c] = WT_N\left(\mathbf{g}'^c\right) \tag{2}$$

where N is the maximum number of the scaling for WT decomposition process, i.e., the resolution index $s \in \{1, \ldots, N\}$ and the N^{th} level corresponds to the coarsest resolution; c is the channels of $\mathbf{g}'^c$ as $c \in \{L, a, b\}$; $\mathbf{A}_N^c$ (to represent scaling coefficients) is the approximation output at the coarsest resolution for each channel; $\mathbf{H}_s^c$, $\mathbf{V}_s^c$ and $\mathbf{D}_s^c$ are the wavelet coefficients of horizontal, vertical and diagonal details for the given c and s, respectively.

The wavelet coefficients representing the details of the image at various scales are used to create several feature maps with increasing frequency bandwidths. The feature maps can be calculated by IWT. Since we already apply the Gaussian filter, we can create feature maps from the details of WT by neglecting approximation data during the IWT process. Hence, several feature maps can be obtained while representing the contrast from edge to texture. The approximation data of the selected decomposition level s is not used during IWT operation as in (3) below, to detect the global saliency:

$$f_s^c(x, y) = \frac{\left(IWT_s\left(\mathbf{H}_s^c, \mathbf{V}_s^c, \mathbf{D}_s^c\right)\right)^2}{\eta} \tag{3}$$

where $f_s^c(x, y)$ is the feature map generated for the s^{th} level decomposition for each image sub-band c, η is the scaling factor (since the range of the *Lab* input image for each channel is $[0, 255]$, there is a large range for feature values in (3); therefore, an appropriate value of η is the scaling factor to limit the feature maps, and $\eta = 10^4$ in (3) where this scaling is necessary to avoid the huge variation in the covariance matrix among the feature maps during the computation of global saliency map in (4)); $IWT_s(\cdot)$ is the reconstruction function referring to the IWT of $\mathbf{H}_s^c$, $\mathbf{V}_s^c$ and $\mathbf{D}_s^c$ by neglecting the $\mathbf{A}_N^c$. Thus, for $s \in \{1, \ldots, N\}$ and $c \in \{L, a, b\}$, Equation (3) creates $3 \times N$ feature maps for an input color image, and each feature map's resolution is equal to the size of the input image. In Fig. 3, five sample color images are given at the top of the figure, and, the feature maps of L, a, and b channels for each sample image are demonstrated respectively. For each color channel at each row of the Fig. 3, there are 8 feature maps from left to right representing the reconstruction results from wavelet coefficients of the 1st-level decomposition to the 8th-level decomposition for the given input images (Fig. 3). It should be noted that 8-level reconstruction consists of details from the coarsest scale (the 8th level) to the finest scale (the 1st level) wavelet coefficients, the 7-level reconstruction consists of details from the 7th level to the finest scale (the 1st level) wavelet coefficients, and so on.

D. Global Distribution of Features

After obtaining the feature maps, the next step is to calculate the global distribution of the local features to obtain the global saliency map. From $f_s^c(x, y)$ in (3), a location (x, y) can be represented as a feature vector $\mathbf{f}(x, y)$ with a length of $3 \times N$ (3 channels L, a and b, and N-level wavelet-based features for each channel) from all feature maps.

Examples have been given in Fig. 3, where each example image has 8 feature maps for each channel, and thus there are 24 features for each location (x, y). Regarding the feature maps, the likelihood of the features at a given location can be defined by the probability density function (PDF) with a normal distribution [12], [24]. Therefore, the Gaussian PDF in multi-dimensional space can be written as [24], [12]:

$$p\left(\mathbf{f}(x, y)\right) = \frac{1}{(2\pi)^{n/2}\left|\Sigma\right|^{1/2}} \times e^{\left(-1/2(\mathbf{f}(x,y)-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{f}(x,y)-\boldsymbol{\mu})\right)} \tag{4}$$

with

$$\Sigma = E\left[\left(\mathbf{f}(x, y) - \boldsymbol{\mu}\right)\left(\mathbf{f}(x, y) - \boldsymbol{\mu}\right)^T\right] \tag{5}$$

where $\boldsymbol{\mu}$ is the mean vector containing the mean of each feature map, i.e., $\boldsymbol{\mu} = E[\mathbf{f}]$; T is the transpose operation; Σ in (5) is the $n \times n$ covariance matrix; $n = 3 \times N$, the number of the feature vector referring to the dimension of the feature space including 3 color channels and N feature maps for each color channel, and $|\Sigma|$ is the determinant of the covariance matrix [24].

Using the PDF in (4), the global saliency map can be computed as (6) below. As can be seen in (6), the result is filtered with a $k \times k$ 2-D Gaussian low-pass filter to obtain a smooth map where $k = 5$, which is a commonly used filter size in many saliency applications as in [9], [10], [11], [12].

$$s_G(x, y) = \left(\log\left(p\left(\mathbf{f}(x, y)\right)^{-1}\right)\right)^{1/2} * \mathbf{1}_{k \times k} \tag{6}$$

where s_G includes the information both locally and globally since it is computed from the local features in (3). However, it can be seen as a global saliency map because its effect on global distribution on the saliency map is much higher, and may become dominant (i.e., overestimated) due to the content or structure of the scene. Also, it includes the statistical relation among the feature maps so it may give some important information which cannot be detected well enough by the local contrast. Moreover, the result from (6) may generate a saliency map with small salient regions, and thus causes some loss in local saliency information. It can be seen that the global saliency map examples in Fig. 4(b) are quite similar to the local saliency map (given in (7)) examples in Fig. 4(c). This is the result when the distribution of the local features for the salient regions is balanced with the local contrast for the given features. However, there are also cases where global saliency can suppress the local contrast too much (see Figs. 5(b) and 5(c)). On the other hand, it may yield important salient regions which are less salient locally or the locally salient regions may not be as attractive as globally salient regions. As can be seen in Fig. 5(g), the wing

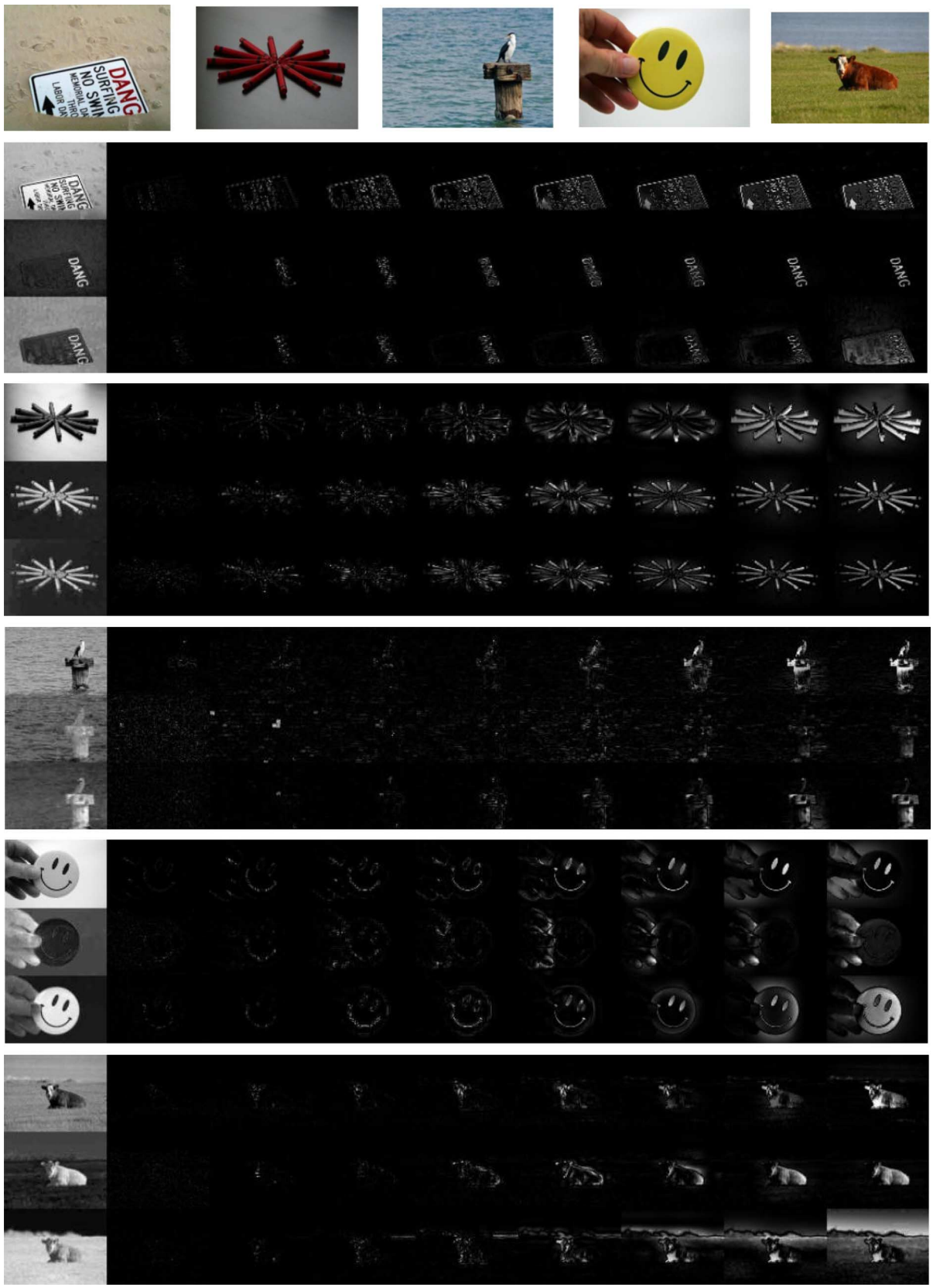
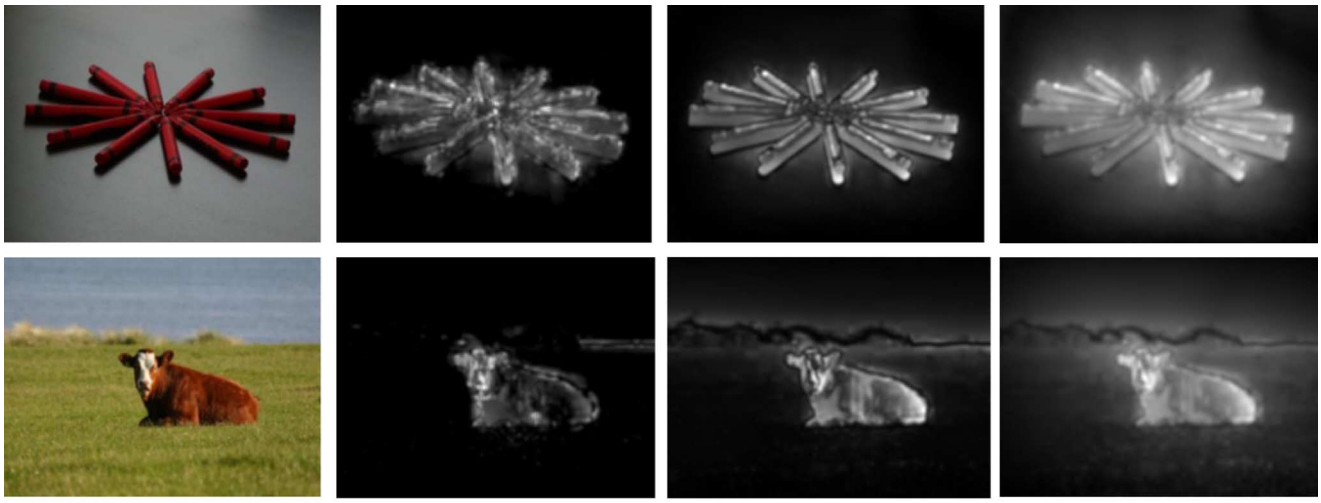


Fig. 3. Images and their feature maps from the CIE Lab color space as luminance and two chromatic channels respectively (RG/BY).

and the head of the animal have high local saliency; however, the global distribution of the local features gives more attention to the head rather than the wings (Fig.5(f)). Therefore, examples in Fig. 5 show that different saliency maps can be beneficial to adjust the saliency for local features or to alleviate overestimation for global information for the better saliency map.



(a) (b) (c) (d)

Fig. 4. Examples for similar local and global saliency responses. (a) Color images. (b) Global saliency maps obtained by (6). (c) Local saliency maps obtained by (7). (d) Final saliency maps.

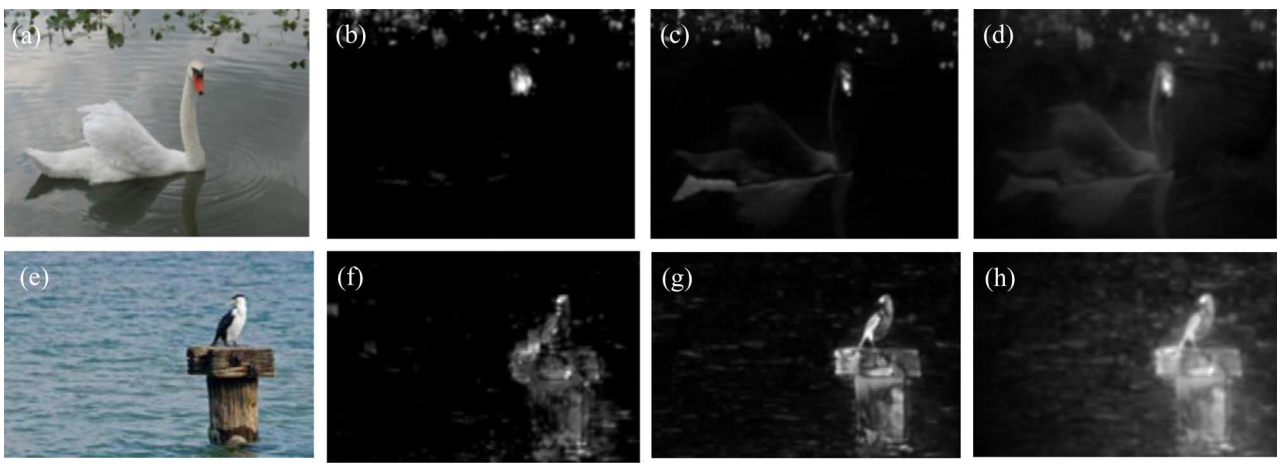


Fig. 5. Examples for different local and global responses. ((a), (e)) Color images. ((b), (f)) Global saliency maps obtained by (6). ((c), (g)) local saliency maps obtained by (7). ((d), (h)) final saliency maps.

E. Linear Combination of the Local Features

In this work, local saliency is created by fusing the feature maps at each level linearly without any normalization operation in [11], as the formula to be given in (7) below. Hence, this new map will be computed based on the local features computed in (3) in which the maximum value between channels of the input image is taken into account at each level. Figs. 4(c), 5(c), and 5(g) demonstrate the local saliency maps of the proposed model. It can be seen that there are differences in some regions between global saliency maps (Figs. 5(b), 5(f)) and local saliency maps (Figs. 5(c), 5(g)). The performance evaluation for local and global saliency maps is also given in Fig. 6 in Section III to show their differences.

The feature maps obtained in (3) are used for calculating the local saliency map as:

$$s_L(x,y) = \left(\sum_{s=1}^{N} \arg\max \left(f_s^L(x,y), f_s^a(x,y), f_s^b(x,y) \right) \right) * \mathbf{1}_{k\times k} \quad (7)$$

where $f_s^L(x,y)$, $f_s^a(x,y)$, and $f_s^b(x,y)$ are the feature maps at scale s for L, a and b channels respectively; $s_L(x,y)$ is the local saliency map.

F. Combination of the Global and Local Saliency Maps

Based on (6) and (7), we can create the global and local saliency maps. The final saliency is the result of combining these two maps. The integration is performed to modulate the local saliency map with its corresponding global saliency map defined as:

$$s'(x,y) = M\left(s'_L(x,y) \times e^{s'_G(x,y)} \right) * \mathbf{1}_{k\times k} \quad (8)$$

where $s'(x,y)$ is the final saliency map, $s'_L(x,y)$ and $s'_G(x,y)$ are the local and global saliency maps linearly scaled to the range $[0,1]$. Since the modulation is applied by the multiplication of local saliency and the exponential value of the global saliency, $M(.)$ as $M(\cdot) = (\cdot)^{\ln\sqrt{2}}/\sqrt{2}$, is used as the non-linear normalization function to diminish the effect of amplification on the map. The possible values of the output in (8) will be between 0 and 1 due to the parameter selection. In (8), we obtain

a saliency map which considers local features at a location with its respective global feature distribution (shown in Figs. 4(d), 5(d) and 5(h)). Therefore, the global relation between local feature maps is established without the need of any complex feature map normalization process for enhancing each feature map as in [10].

In addition, we enhance the final saliency map with a similar fashion in [25]. As stated by Goferman *et al.* [25], based on Gestalt laws [26], saliency values to describe the regions of interest can be reevaluated around the regions that are the most salient points of the scene. The idea is: the locations around the focus of attention (FoA) have to be more attentive than those away from the FoA [25]. Therefore, saliency values around the most salient points are increased to enhance the performance of the saliency map as [25]:

$$s(x,y) = s'(x',y')\left(1 - d_{cFoA}(x,y)\right) \tag{9}$$

where $s(x,y)$ is the saliency value at point (x,y), $s'(x',y')$ is the salient value of the most salient points at the location (x',y') extracted from the saliency map in (8) with a threshold of 0.8 as in [25], $d_{cFoA}(x,y)$ is the distance between the location (x,y) and its closest FoA at the location (x',y'). Obviously, the saliency values around the salient regions will increase in the final saliency map; on the other hand, the saliency values of the points that are distant to the attention regions will decrease or remain largely unchanged. The proposed final saliency has better performance than the local and global saliency maps as shown in Fig. 6. The analysis and evaluation are given in Section III for the proposed algorithm.

III. Experimental Results

In this work, the Microsoft public database [27] including 5000 color images is used to evaluate the performance of the proposed model quantitatively. Besides the color images, there are also ground-truths for images in the database: the human-labeled attention regions highlighted with a bounding box created by 9 subjects [27]. These bounding boxes represent the attention regions, i.e., the object/region of interest in scenes perceived by the subjects. As a result of averaging the human responses, the performance evaluation of a saliency detection model can be achieved quantitatively by checking the consistency between the human-labeled ground-truth and the saliency map from any computational model.

Prior to the model comparisons, we have evaluated the performance of local saliency, global saliency, and final saliency in our model to demonstrate how they affect the overall performance. The quantitative performance for the database is evaluated based on overall *precision P*, *recall R*, and *F-Measure F* F_α, as defined below respectively [27]:

$$P = \frac{\sum_x \sum_y \left(t(x,y) \times s(x,y)\right)}{\sum_x \sum_y s(x,y)} \tag{10}$$

$$R = \frac{\sum_x \sum_y \left(t(x,y) \times s(x,y)\right)}{\sum_x \sum_y t(x,y)} \tag{11}$$

$$F_\alpha = \frac{(1+\alpha) \times P \times R}{\alpha \times P + R} \tag{12}$$

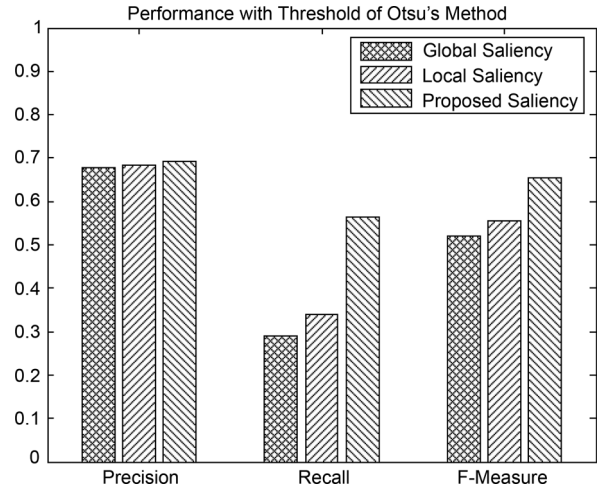


Fig. 6. Overall performance evaluation for the local, global and final saliency maps of the proposed model.

where $t(x,y)$ is the ground-truth map, $s(x,y)$ is the saliency map from the computational model, and α in (12) is a positive parameter to decide the relative importance of the precision over the recall in evaluating the precision (a greater value for α indicates the higher importance of recall over precision; α is chosen as 0.3 in this work).

For the experimental results, P is related to the saliency detection performance of the computational model; R is the ratio of salient regions from correct detection and ground-truth; *F-Measure* is a performance measure as being the harmonic mean of P and R [27]. In the evaluation, the generated saliency map is converted to a binary image with an appropriate threshold (as to be discussed next) for performance comparison, and $t(x,y)$ and $s(x,y)$ in (10) and (11) are the binary maps to calculate P, R, and *F-Measure* values.

For this analysis, Otsu's automatic threshold algorithm [28] is selected for the binary map generation since it makes the test less independent than the user defined threshold values. P, R and *F-measure* results for the proposed local, global and final saliencies are given in Fig. 6. It can be seen that overall performance of the local saliency is better than global saliency for the 5000-image database, while their *precision* performances are very similar; but *recall* and *F-measure* values of the local saliency are better than those of the global saliency. From Fig. 6, the final saliency map results in better performance than local or global saliency, regarding P, R, and *F-measure* results. It can be concluded that global saliency map and local saliency map may carry important information of different regions for some images as in Figs. 5(f) and 5(g). This demonstrates the advantage of combining the local and global saliency for the proposed model to create the final saliency map.

As mentioned in Section I, we have selected the works by Itti *et al.* [11], Achanta *et al.* [14], Hou *et al.* [16], and Murray *et al.* [21] for comparison. The reason is that each work above is unique due to their technical approaches, and they represent spatial-based models, FT-based models, and WT-based models, respectively. Since our intention is to compare the saliency detection approaches, we have used the saliency computation part

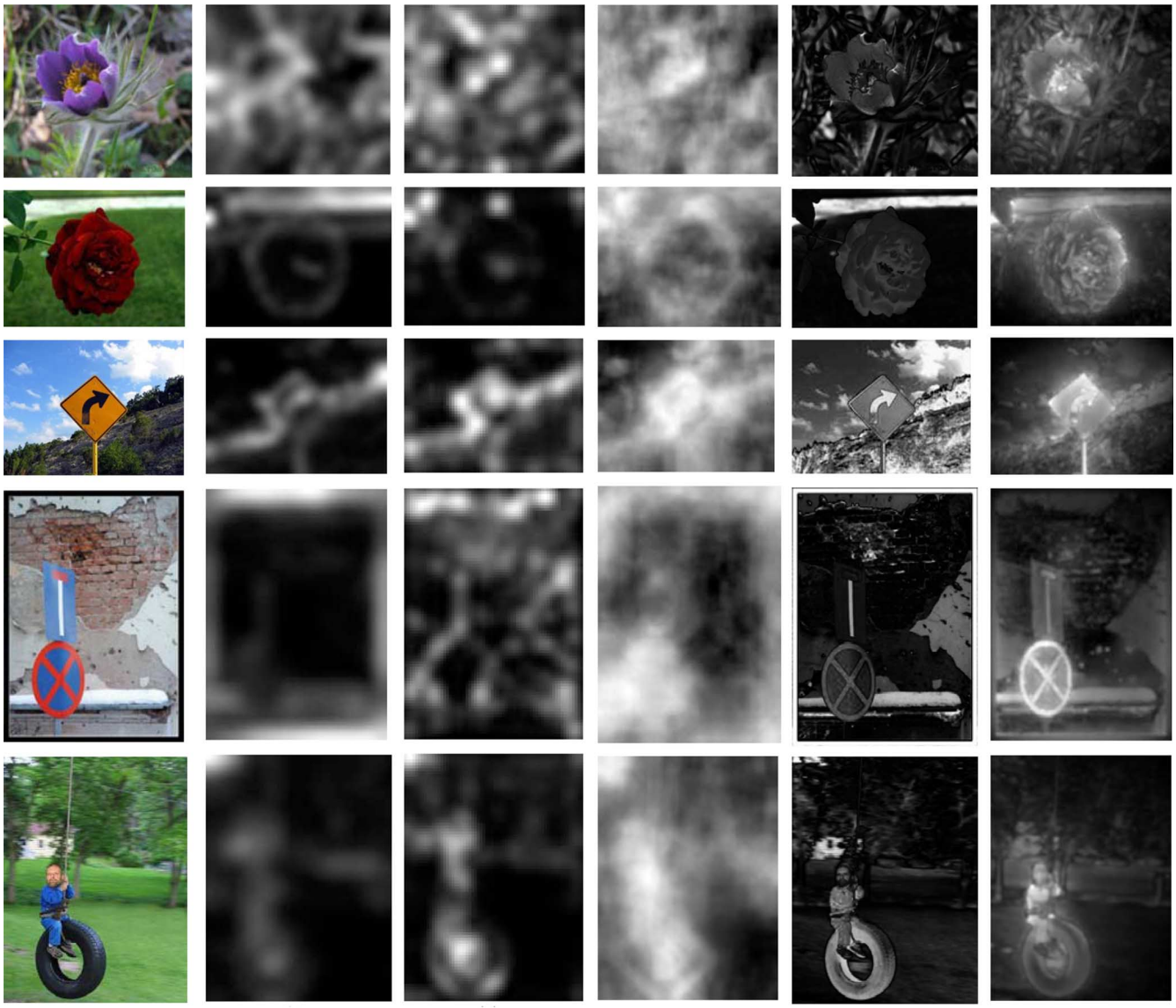


Fig. 7. Performance evaluation. (a) Color images. (b) Saliency maps of Itti's model [11]. (c) Saliency maps of Hou's model [16]. (d) Saliency maps of Murray's model [21]. (e) Saliency maps of Achanta's model [14]. (f) Saliency maps of the proposed model.

of [14] by neglecting the segmentation part, because all the selected models (including the proposed model) are saliency map generation based on low-level features without segmentation. In Fig. 7, some examples are given for comparison with models in [11], [14], [16], [21], and the proposed model.

Firstly, we evaluate the performance of the proposed model based on the P, R, and *F-measure* measures. Since the threshold effects the performance evaluation, we have selected two different methods to obtain the threshold value; using Otsu's automatic threshold model [28] and mean value of the saliency map. Figs. 8 and 9 show the overall performance of the existing models and our proposed algorithm based on the threshold values from Otsu's method and averaging the saliency map respectively. Experimental results prove the reliability of the proposed model quantitatively according to the overall performance from 5000 images. For both threshold selection methods, our model outperforms others based on the precision and F-Measure values (see Figs. 8 and 9). Regarding the Otsu's threshold method, the recall value of the work [21] is better than our model; however, its precision and F-measure values are relatively low due to the high recall value which causes more irrelevant salient regions occur (Fig. 7(d)). On the other hand, the precision value of [21] is the lowest among the compared models, and our model has the best precision and F-Measure values for this case too even though the recall value of our model is similar to that of [21] based on the threshold with mean value.

Secondly, we also evaluate the performance of the proposed model based on the Receiver Operating Characteristic (ROC) [25]. The saliency maps include two different regions: salient and non-salient regions. Percentage of the salient regions from the ground-truth intersecting with the salient region from the saliency map is called True Positive Rate (TPR). Percentage of the non-salient regions from the ground truth intersecting with

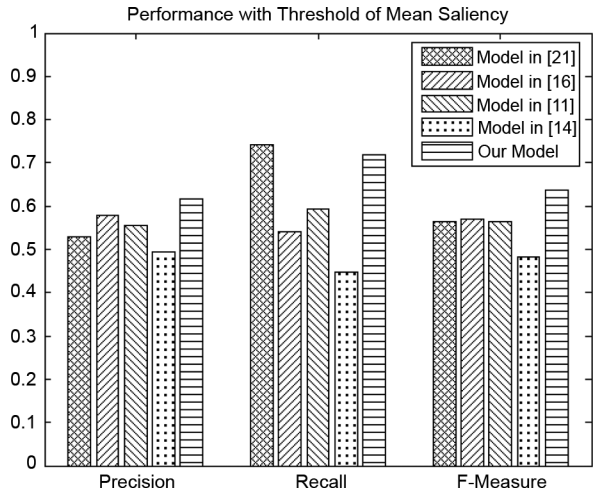


Fig. 8. Performance comparison between the proposed model and other existing ones.

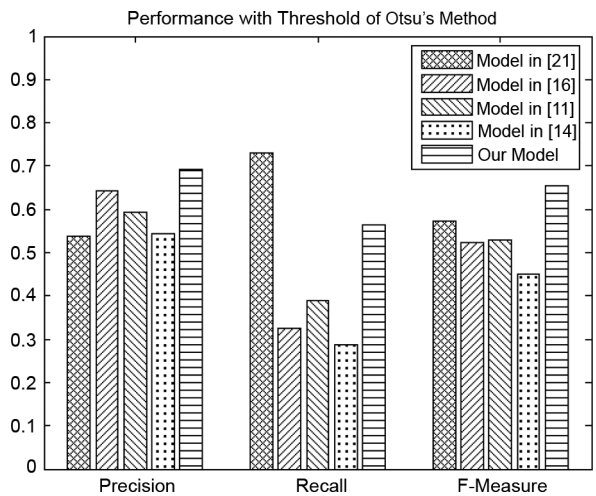


Fig. 9. Performance comparison between the proposed model and other existing ones.

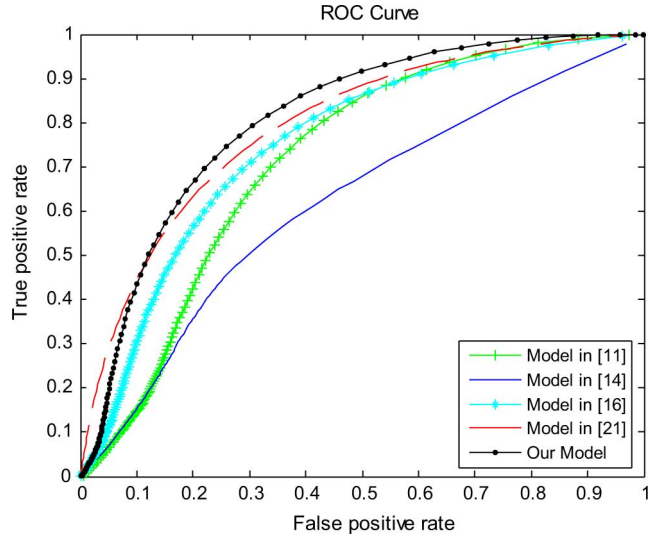


Fig. 10. ROC curves for different saliency detection models.

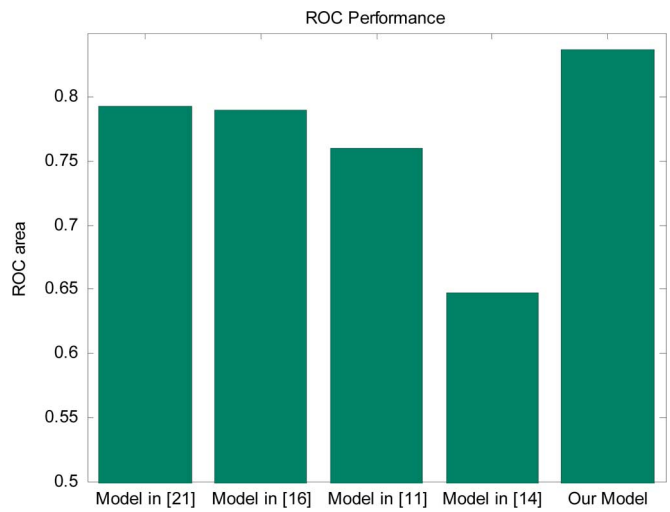


Fig. 11. ROC area results from different saliency detection models.

the salient region from the saliency map is called False Positive Rate (FPR). From these data, the ROC curves for the saliency detection models are shown in Fig. 10.

A larger ROC area means better performance for a saliency detection model. Fig. 11 shows the ROC areas for the compared models and the proposed model. The saliency model [14] without the segmentation integration has the smallest ROC area compared to other models. The models [16] and [21] have similar results by yielding better performance than those in [14] and [11]. The proposed algorithm has the largest ROC area among the compared models.

In sum, the proposed algorithm is first tested based on the precision, recall and F-measure values. Then, the ROC area test is selected as another performance measurement to compare the proposed model with others. Regarding these two test criteria, the overall performance of the proposed model is reliable and yields better results with respect to the relevant existing works.

IV. Conclusions

In this paper, a novel bottom-up computational model of visual attention is proposed to obtain the saliency map for images based on wavelet coefficients. Various feature maps are generated by IWT with the band-pass regions of the image in various scales. The proposed model derives feature maps of band-pass filtered regions from the input image with increasing frequency bandwidths. It can be seen as an adaptation of the center-surround structure of the HVS (human visual system) since feature maps includes components from the edge to the texture based on the multi-level wavelet decomposition. Using these features, the local and global saliency maps are introduced to form the final saliency map.

The final saliency map represents both the local contrast of each location on the scene and the global distribution of the features as an *amplifier* for local saliency values. The local saliency map is calculated based on the linear combination of each level's maximum value in the feature maps within L, a and b channels, while the global saliency map is computed based on the

normal distribution of the local features. The final saliency detection from the combination of the local and global information has performed well on the public database with 5000 images and the associated human-labeled ground truth. Extensive experimental evaluation confirms that the proposed model performs better than the relevant existing models under different test conditions.

REFERENCES

[1] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[2] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, pp. 219–227, 1985.

[3] J. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search," *J. Exp. Psychol.: Human Percept. Perform.*, vol. 15, no. 3, pp. 419–433, 1989.

[4] J. Wolfe, "Guided search 2.0: A revised model of guided search," *Psychonomic Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.

[5] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, "Modelling visual attention via selective tuning," *Artif. Intell.*, vol. 78, no. 1–2, pp. 507–545, Oct. 1995.

[6] E. Niebur and C. Koch, "Computational Architectures for Attention," in *The Attentive Brain*, R. Parasuraman, Ed. Cambridge, MA: MIT Press, 1998, pp. 163–186.

[7] J. M. Wolfe, S. J. Butcher, and M. Hyle, "Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons," *J. Exp. Psychol. Human Percept Perform.*, vol. 29, pp. 483–502, 2003.

[8] O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model the bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[9] S. Frintrop, "VOCUS: A visual attention system for object detection and goal directed search," Ph.D. dissertation, Rheinische Friedrich-Wilhelms-Universiẗat Bonn, Bonn, Germany, 2005.

[10] L. Itti, "Models of bottom-up and top-down visual attention," Ph.D. dissertation, Dept. Computat. Neur. Syst., California Inst. Technol, Pasadena, 2000.

[11] L. Itti, C. Koch, and E. Niebur, "Model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[12] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proc. IEEE Int. Conf. Image Processing*, 2003, vol. 1, pp. 253–256.

[13] Y. F. Ma and H. J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 374–381.

[14] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. Comput. Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[16] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. Comput. Vision and Pattern Recognition*, 2007, pp. 1–8.

[17] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Int. Conf. Comput. Vision and Pattern Recognition*, 2008, pp. 1–8.

[18] R. J. E. Merry, *Wavelet Theory and Application: A Literature Study, DCT 2005.53*. Eindhoven, The Netherlands: Eindhoven Univ. Technol., 2005.

[19] D. L. Fugal, *Conceptual Wavelets in Digital Signal Processing: An In-depth Practical Approach for the Non-Matematician*. San Diego, CA: Space & Signals Technical Publishing, 2009, pp. 1–78.

[20] Q. Tian, N. Sebe, M. S. Lew, E. Loupias, and T. S. Huang, "Image retrieval using wavelet-based salient points," *J. Electron. Imag.*, vol. 10, 4, pp. 835–849, 2001.

[21] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. IEEE Int. Conf. Comput. Vision and Pattern Recognition*, 2011.

[22] Y. Kocyigit and M. Korurek, "EMG signal classification using wavelet transform and fuzzy logic classifier," *ITU dergisi/d mühendislik*, vol. 4, no. 3, 2005.

[23] John and L. Semmlow, *Biosignal and Biomedical Image Processing: MATLAB-Based Applications*. New York: Marcel Decker, 2004.

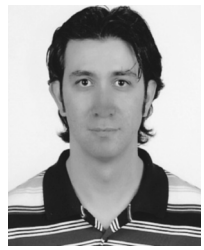
[24] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. London, U.K.: Academic/Elsevier, 2009, pp. 20–24.

[25] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Int. Conf. Comput. Vision and Pattern Recognition*, 2010, pp. 2376–2383.

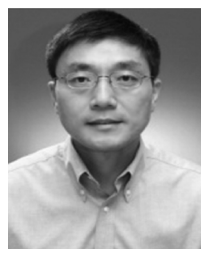
[26] K. Koffka, *Principles of Gestalt Psychology*. London, U.K.: Routledge & Kegan Pul, 1955.

[27] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Int. Conf. Comput. Vision and Pattern Recognition*, 2007, pp. 1–8.

[28] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Signal Procesing Using Matlab®*. Englewood Cliffs, NJ: Prentice Hall, 2004.



Nevrez İmamoğlu received the B.E. degree in Computer Engineering with double major in Electronics & Communication Engineering from Çankaya University, Ankara, Turkey, in 2007, and the M.S. degree in Electrical and Electronics Engineering from TOBB University of Technology and Economics, Ankara, Turkey, in 2010. He worked as a research associate for 1 year during 2010–2011 at School of Computer Engineering, Nanyang Technological University, Singapore. He is currently pursuing the Ph.D. degree in Medical System Engineering from the Division of Artificial Systems Science, Chiba University, Chiba, Japan. His research interests include image/video processing, computer vision, pattern recognition, and intelligent systems.



Weisi Lin (M'92–SM'98) received his B.Sc. and M.Sc. from Zhongshan University, China, and the Ph.D. degree in computer vision from King's College London, U.K. He served as the Lab Head, Visual Processing, and the Acting Department Manager, Media Processing, for the Institute for Infocomm Research. Currently, he is an Associate Professor in the School of Computer Engineering, Nanyang Technological University, Singapore. His areas of expertise include image processing, perceptual modeling, video compression, multimedia communication and computer vision. He has published over 190 refereed papers in international journals and conferences. He is currently on the editorial boards of IEEE TRANSACTIONS ON MULTIMEDIA, IEEE SIGNAL PROCESSING LETTERS and *Journal of Visual Communication and Image Representation*.

Dr. Lin is a Chartered Engineer (U.K.), a fellow of the Institution of Engineering Technology, and an Honorary Fellow, Singapore Institute of Engineering Technologists. He organized special sessions in IEEE International Conference on Multimedia and Expo (ICME 2006, 2012), IEEE International Workshop on Multimedia Analysis and Processing (2007), IEEE International Symposium on Circuits and Systems (ISCAS 2010), Pacific-Rim Conference on Multimedia (PCM 2009), SPIE Visual Communications and Image Processing (VCIP 2010), Asia Pacific Signal and Information Processing Association (APSIPA 2011, 2012), and MobiMedia 2011. He gave invited/keynote/panelist/tutorial talks in International Workshop on Video Processing and Quality Metrics (2006), IEEE International Conference on Computer Communications and Networks (2007), SPIE VCIP 2010, PCM 2007, PCM 2009, IEEE ISCAS 2008, IEEE ICME 2009, APSIPA 2010, and IEEE International Conference on Image Processing (2010). He has been elected as a Distinguished Lecturer of APSIPA (2012–13).

Yuming Fang received the B.E. degree in Software Engineering from Sichuan University in 2006 and the M.S. degree in Communication and Information System from Beijing University of Technology in 2009. He is currently pursuing the Ph.D. degree from the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include image/video processing and computer vision.