Learning Saliency by MRF and Differential Threshold

Guokang Zhu, Qi Wang, Yuan Yuan, *Senior Member, IEEE*, and Pingkun Yan, *Senior Member, IEEE*

***Abstract*—Saliency detection has been an attractive topic in recent years. The reliable detection of saliency can help a lot of useful processing without prior knowledge about the scene, such as content-aware image compression, segmentation, etc. Although many efforts have been spent in this subject, the feature expression and model construction are far from perfect. The obtained saliency maps are therefore not satisfying enough. In order to overcome these challenges, this paper presents a new psychologic visual feature based on *differential threshold* and applies it in a supervised Markov-random-field framework. Experiments on two public data sets and an image retargeting application demonstrate the effectiveness, robustness, and practicability of the proposed method.**

***Index Terms*—Computer vision, differential threshold, machine learning, Markov random field (MRF), saliency detection, visual attention.**

I. Introduction

THE HUMAN visual system is remarkably effective at finding particular objects from a scene. For instance, when looking at images in the first row of Fig. 1, people are usually attracted by some specific objects within them (i.e., strawberries, a leaf, a flag, a person, and a flower, respectively). This ability of the human visual system to identify the salient regions in the visual field can enable one to "withdraw from some things in order to deal effectively with others" [1], [2], i.e., allocate the limited perceptual processing resources in an efficient way. It is believed that visual processing involves two stages: a preattentive stage that processes all the information available in parallel and, then, an attentive stage, in which the partial information inside of the attentional spotlight is glued together in serial for further processing [3]–[5]. In this paper, the preattentive functionality of the human visual system is imitated by a computer vision technique—saliency detection.

Saliency detection aims at providing the computational identification of scene elements that are notable to human observers. The detected result is presented as a grayscale image. The lighter the pixel is, the more salient it might be. Typical examples of saliency detection results are demonstrated in Fig. 1. Implementing the functionality of human visual attention by means of saliency detection is considered to be an important component in computer vision because of a wide range of applications, such as adaptive content delivery [6], video summarization [7], image quality assessment [8], [9], content-aware image compression and scaling [10]–[12], image segmentation [13], [14], object detection [2], and object recognition [15]–[17]. However, computer vision methods are still far from satisfying compared with biological systems. Therefore, researchers have never stopped making efforts to provide a more effective and efficient method for automatic saliency detection.

A. Related Work

Classical saliency detection methods choose to employ a "low-level" approach to calculate contrasts of image regions with respect to their surroundings, by selecting one or more low-level features such as color, intensity, and orientation [18]. The produced saliency results are topographically arranged maps, which integrate the normalized information from one or more feature maps to represent the visual saliency of a specific scene. According to the techniques that they used, these methods can broadly be categorized into three groups: biologically inspired, fully computational, and a combination of them.

Biologically inspired methods are based on the imitation of the selective mechanism of the human visual system. Itti *et al.* [19] introduce a groundbreaking saliency model, which is inspired by the biologically plausible architecture of the human visual system proposed by Koch and Ullman [21]. They first extract multiscale features from three complementary channels using a *difference of Gaussian* (DoG) approach. Then, the across-scale combination and normalization are employed to fuse the obtained features to an integrated saliency map. Based on this work, Walther *et al.* [22] propose to recognize salient objects in images by combining the saliency detection method of [19] with a hierarchical recognition model, while Frintrop *et al.* [23] capture saliency by computing center-surround differences through square filters and employ integral images to speed up the computation process. Recently, Garcia-Diaz *et al.* [24], [25] propose to take into account the perception role of *nonclassical receptive fields*. They start from the multiscale decomposition on the features of color and local orientation and use the

Manuscript received February 24, 2012; revised July 20, 2012 and December 3, 2012; accepted December 28, 2012. Date of publication February 21, 2013; date of current version November 18, 2013. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB707104, by the National Natural Science Foundation of China under Grants 61172142, 61172143, 61105012, 61125106, and 91120302, by the Natural Science Foundation Research Project of Shaanxi Province under Grant 2012JM8024, and by the 50th China Postdoctoral Science Foundation under Grant 2011M501487. This paper was recommended by Editor F. Hoffmann.

The authors are with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China (e-mail: zhuguokang@opt.ac.cn; crabwq@opt.ac.cn; yuany@opt.ac.cn; pingkun.yan@opt.ac.cn).

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/TSMCB.2013.2238927

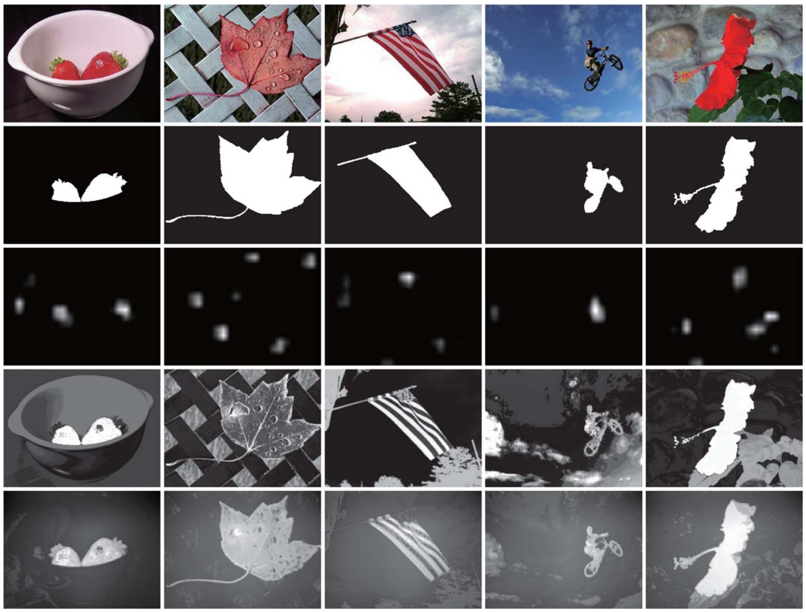


Fig. 1. Saliency detection. From top to bottom, each row respectively represents the original images, the ground truths, and the saliency maps calculated by IT [19], HC [20], and the proposed method.

statistical distance between each feature and the center of the distribution to produce the saliency map.

Different from biologically inspired methods, the fully computational methods calculate saliency maps directly by contrast analysis [18]. For example, Achanta *et al.* [26] evaluate saliency as the Euclidean distance between the average feature vectors of the inner subregion of a sliding window and its neighborhood region. Later, in order to preserve the boundaries of salient objects in the saliency map, Achanta *et al.* [18] present a frequency-tuned algorithm, which can retain more frequency content from the examined image than previous techniques. Gao *et al.* [27]–[29] model saliency detection as a classification problem. They measure the saliency value of a location as the *Kullback–Leibler* divergence between the histogram of a series of DoG and Gabor filter responses at the point and its surrounding region. Seo and Milanfar [30] propose to use *local steering kernels* (LSKs) as features, which measure saliency in terms of the amount of gradient contrast between the examined location and its surrounding region. Differently, Hou and Zhang [31] present a spectral residual method independent of low-level image features and prior knowledge. They start from a thorough statistics and experimental analysis of the log Fourier spectrum of natural images. Then, they propose to detect saliency by calculating the contrast between the original and the locally averaged log Fourier spectrum of the examined image.

More recently, Rahtu *et al.* [32] propose a saliency measure based on a Bayesian framework, which calculates the local contrast of a set of low-level features in a sliding window. Goferman *et al.* [33] propose a context-aware saliency model, which aims at extracting an image subregion representing the scene and is based on the contrast between each image patch and the corresponding k most similar patches in the image. Liu *et al.* [34], [35] formulate the problem of saliency detection as an image segmentation task. In their method, novel features such as center-surround histogram, multiscale contrast, and color spatial distribution are employed to extract the prominent regions through *conditional random field* (CRF) learning. Cheng *et al.* [20] propose a regional saliency extraction method simultaneously evaluating the global contrast and spatial coherence. Wang *et al.* [36] define saliency as an anomaly relative to a given context and detect salient regions in the image associated with a large dictionary of images through *k-nearest-neighbor* retrieval.

The third category of methods is partly inspired by biological models and partly dependent on the techniques of fully computational methods. For instance, Harel *et al.* [37] design a graph-based method, which first forms activation maps by using some certain features (e.g., by default, color, intensity, and orientation maps are computed) and then combines them in a manner that highlights conspicuity. Bruce and Tsotsos [38] describe a biologically plausible model of saliency detection based on the maximum information sampled from a scene and calculate the *probability density function* based on a Gaussian *kernel density estimate* in a neural circuit. Zhang *et al.* [39] provide a Bayesian framework for the saliency task. They consider saliency as the probability of a target to be outstanding based on the analysis of Shannon's self-information and mutual information. Other than these, Judd *et al.* [40] train a saliency detection model by *support vector machine* (SVM), which utilizes multilevel image features to tackle the eye-tracking data.

B. *Limitations of Existing Methods*

Although various methods for saliency detection have been presented in the past few years and a laudable performance for human attentional spotlight prediction has been achieved in some circumstances, there are still several limitations for these methods.

The first limitation is the **integration model**. Although there are large number of cues, such as color, texture, shape, depth,

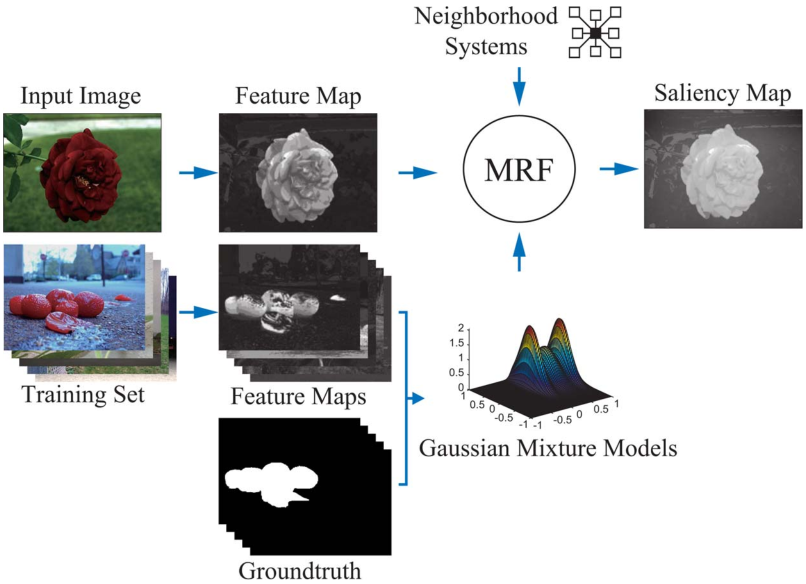


Fig. 2. Summary of the proposed method. The salient object detection problem is modeled by an MRF, where a group of biologically inspired salient features is incorporated in the detection procedure through MRF learning.

shadow, and motion, which have been considered as influential factors on visual attention in existing works, most of the mainstream methods [18], [20], [30], [32], [39], [41] integrate them only based on direct contrast calculation, which is not able to include priors. However, the ability to incorporate the prior information is important in many computer vision tasks [42].

Recently, learning-based methods have become very popular and seem to be generating promising results. However, there are still remaining problems. For example, Judd *et al.* [40] employ an SVM classifier for saliency detection. The training set employed in their work is composed of images with ground-truth fixation points instead of regions. Sometimes, a few detected fixation points are adequate for further applications. However, in most cases, the desired exports are salient regions.

The second limitation is the **feature description**. Many descriptions of the saliency features [30], [32], [36], [43], [44] based on the aforementioned cues are only applicable to scene-specific database or limited by the strict conditions of use. For example, Seo and Milanfar [30] propose to use LSK as the saliency feature descriptor based on the covariance matrices within the local windows. It is obvious that this definition will highlight regions with higher local complexity. However, saliency cannot always be equated with local complexity. For example, Fig. 1 shows that the much less complex regions containing the flag or the flower appear to be significantly more salient. Wang *et al.* [36] describe saliency feature based on searching through the enormous online image database. Their method therefore greatly limits its promotion potential by the harsh conditions in practice.

In fact, there are many proverbial principles of the human vision system, which can play a significant role in breaking through the bottleneck of constructing the more effective and compact feature descriptions. Nevertheless, there is no substantial progress in regard to ingeniously introducing these principles into computer vision.

C. Overview of the Proposed Method

The presented method, named *differential threshold and Markov random field (MRF)-based visual saliency* (DTMBVS), formulates salient object detection as a *maximum a posteriori probability* (MAP) estimation problem, which can be solved by finding the optimal binary labels that discriminate the salient regions from the background of the scene through MRF learning. Fig. 2 shows the flowchart. The main contribution of this paper is a tractable method suitable for saliency detection. This is motivated by the need for overcoming the limitations of existing methods and takes advantage of two components.

First, the saliency detection problem is modeled by an *MRF*. MRF and its variants (such as CRF [45], [46] and DRF [47]) have achieved many successes in computer vision. The primary advantages of MRF are the regularization, which has the ability to form fields with locally coherent labels, and the strong robustness to noise. In this paper, two biologically inspired features are incorporated in the detection procedure through MRF learning.

The closest to DTMBVS is the method proposed by Liu *et al.* [35] which extracts a prominent region through CRF learning. The proposed method differs from the one in [35] mainly in two aspects. First, the proposed method models the posterior probability as the product of two likelihoods [see (1)], each of which is then individually formulated as an MRF energy representation, while Liu's method models the posterior probability directly as an MRF, which actually is a linear combination of features. The proposed model takes advantage of the simplicity to be understood and implemented in practice. Second, the employed features are much different. Liu's method utilizes the multiscale contrast, center-surround histogram, and color spatial distribution, while the proposed method mainly uses a psychologically inspired color feature (see Section II-B).

Second, a new *differential threshold*-based visual feature is introduced for feature extraction (see Fig. 3). The *differential*

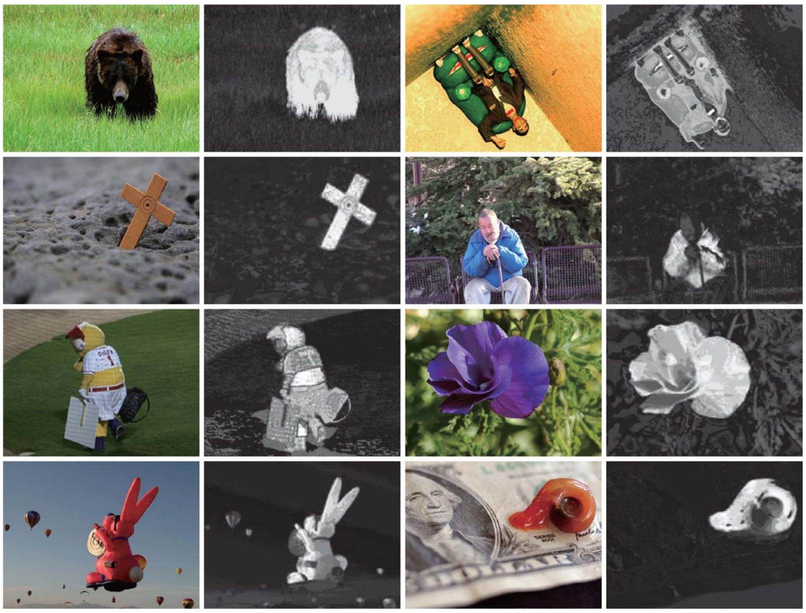


Fig. 3. Examples of *differential threshold*-based visual feature maps.

threshold refers to the minimal differences that can be discriminated by the human visual system between two homogeneous physical stimuli, which can be quantified by the concept of *just noticeable difference* (JND) [48]. Ernest Heinrich Weber, an experimental psychologist, discovered that the JND between two physical stimuli is not an absolute amount but a relative one associated with the intensity of the preceding stimulus [48]. Inspired by this principle, the proposed visual feature employs one JND as a unit to compartmentalize each individual color channel. Then, the color statistics of the obtained JNDs are utilized to define contrast for each pixel.

The rest of this paper is organized as follows. Section II introduces the proposed framework and the *differential threshold-based* visual feature for saliency detection. Section III presents the extensive experiments conducted to prove the effectiveness of the proposed method. Section IV demonstrates a content-aware image retargeting application, and the conclusion follows in Section V.

II. Model Description

In this section, saliency detection is modeled by an MRF. At the same time, two biologically inspired salient features are incorporated in the detection procedure through MRF learning.

A. MRF Structured Field of Saliency Detection

This section specifies a general saliency detection method, which aims to estimate the probability (i.e., between 0 and 1) of each pixel to be salient in a visual scene according to the image features.

In our method, the input image X with pixels $\{x_i\}$ is described by two kinds of features, $F = \{f_i\}$ and $S = \{s_i\}$, where f_i and s_i are the proposed *differential threshold*-based visual feature and the relative position of x_i, respectively. This method will provide X a binary mask L to classify each pixel x_i with a label $l_i \in \{1, 0\}$, which indicates whether this pixel belongs to the salient region.

The prior information is represented by the notation $\mathcal{G}$, which takes into account the statistical properties of the selected features, as well as the supervisory labels. Concretely, $\mathcal{G}$ is introduced to represent the distributions of the features in the salient region (denoted as $G_{F,1}$ and $G_{S,1}$) and background (denoted as $G_{F,0}$ and $G_{S,0}$) and the distribution of the proportion R of the salient pixels in the entire image (denoted as G_R). Each distribution is described by a *Gaussian mixture model* (GMM).

Since $\mathcal{G}$ can be learned from the training set, the saliency detection problem can thus be translated to a MAP problem

$$p(L|\mathcal{G}, X) \propto p(X, L|\mathcal{G}) = p(X|L, \mathcal{G}) \cdot p(L|\mathcal{G}). \tag{1}$$

Insofar, the pixel coordinate and the *differential threshold*-based visual feature are assumed to independently affect the saliency detection. Therefore, the probability $p(X|L, \mathcal{G})$ is made of two distinct parts

$$p(X|L, \mathcal{G}) = p(F, S|L, \mathcal{G}) = p(F|L, \mathcal{G}) \cdot p(S|L, \mathcal{G}). \tag{2}$$

Then, consider the probability $p(L|\mathcal{G})$ of the mask L given all the related parameters $\mathcal{G}$. We assume that the mask will form a label field with interactions between neighboring pixels, and the labels in the mask obey the distribution described by G_R. Therefore, this probability is assumed to combine two independent models

$$p(L|\mathcal{G}) = p(R_L|G_R) \cdot R_L \cdot p_{corr}(L). \tag{3}$$

The first part $p(R_L|G_R) \cdot R_L$ constrains the mask L with the GMM description of R. R_L is the proportion of pixels labeled as 1 by L in the image. The second part $p_{corr}(L)$ encodes neighbor correlations imposed by the MRF, which regularizes fields with locally coherent labels. This field is defined on a grid (8-connectivity).

Then, the conditional probability $p(L|\mathcal{G}, X)$ can be rewritten using an energy function E, $p(L|\mathcal{G}, X) \propto \exp(-E)$, which makes the solving of the MRF easier

$$E = U_1 + U_2 + \Sigma_{i,j \in \mathbb{N}} V_{i,j} \tag{4}$$

where $\mathbb{N}$ represents couples of graph neighbors in the pixel grid. The sum over $V_{i,j}$ represents the interaction potential, which is defined as

$$V_{i,j} = \begin{cases} -\beta, & l_j = l_i, x_j \in \mathcal{N}_i \\ +\beta, & l_j \neq l_i, x_j \in \mathcal{N}_i \end{cases} \tag{5}$$

where $\mathcal{N}_i$ denotes the neighbors of x_i and the constant β is experimentally chosen to be 0.5.

U_1 is the unary potential of the observations, while U_2 is the likelihood energy of the label distribution. These two terms are determined by

$$U_1 = -\log\left[p(F|L, \mathcal{G}) \cdot p(S|L, \mathcal{G})\right] \tag{6}$$

$$U_2 = -\log\left[p(R_L|G_R) \cdot R_L\right]. \tag{7}$$

After obtaining the component items of E, the MAP problem can be transformed to finding an optimal binary mask L to end the iterating process

$$L^* = \arg\max_L p(L|\mathcal{G}, X) \propto \arg\min_L E. \tag{8}$$

Once L^* has been obtained, there is a direct way to define saliency value $S(x_i)$ as the probability of x_i to be labeled with 1 while the others are satisfied with L^*, i.e.,

$$\begin{aligned} S(x_i) &= p(l_i = 1|\mathcal{G}, X) \\ &= p(f_i|G_{F,1}).p(s_i|G_{S,1}).p(R_{L^*}|G_R).R_{L^*}.e^{-\Sigma_{j\in\mathcal{N}_i} V_{i,j}}. \end{aligned} \tag{9}$$

Finally, it should be noted that each GMM is simply estimated by a recursive *expectation-maximization* algorithm, with each mixture made of three components, and the MAP is approximately estimated by an *iterated conditional mode* algorithm.

B. Differential Threshold-Based Visual Feature

Inspired by the discovery made by Ernest Heinrich Weber [48] that the perceptual difference between two homogeneous physical stimuli is not an absolute amount but a relative one associated with the intensity of the preceding stimulus, this paper proposes a *differential threshold*-based contrast model to define the saliency feature for each pixel in the image.

For every color channel, it is compartmentalized by a sequence of intervals called JNDs. Each JND is determined by a set of ethological and psychological experiments conducted by the authors. First, 15 participants are chosen to be subjects in the experiments. Then, the JNDs of each color channel are determined by increasing the value in this channel based on the preceding JNDs until a perceivable change happens to the subjects. The increment with a chance of 50% of being reported to be different from the proceeding color by the subjects is considered as the new JND in the examined channel. In this process, the other two color channels are fixed to the corresponding average. This procedure is repeated for each color channel until the color space is quantized to a limited number of color prototypes.

TABLE I
UPPER BOUNDS OF EACH COLOR INTERVAL DIVIDED BY THE OBTAINED JNDs IN THE RGB COLOR SPACE

	The upper bounds of each bin									
R	30	51	78	101	124	159	184	221	255	
G	28	51	80	95	124	141	172	193	228	255
B	34	61	88	125	152	183	208	235	255	

After implementing this procedure in the RGB color space, the R, G, and B channels are finally divided into 9, 10, and 9 JNDs with unequal intervals, respectively. For each interval, its corresponding upper bound is presented in Table I. Then, the color statistics of the rerepresented image are used to calculate the visual feature for each pixel. To be specific, the feature value of a pixel x_i is determined by

$$f(x_i) = f(I_i) = \sum_{j=1}^{n} p_j D(I_i, I_j) \tag{10}$$

where I_i is the color of pixel x_i, n is the total number of colors presented in the image, and p_j is the frequency of color I_i in the image. $D(I_i, I_j)$ is the color distance between I_i and I_j. Employing different color spaces or color distance formulas often leads to completely different values for $D(I_i, I_j)$. In our experiments, best results were obtained by measuring the distance with a new color distance formula in the *CIELAB* color space. More details are specified in Sections III-C and III-D.

By using the *differential threshold* to quantize the color space, we can reduce the number of colors to $n = 810$. Perhaps the closest to our feature are the features of [49] and [20]. These two features are based on the color statistics of images similar to (10), which is designed with a computational complexity of n^2. Differently, the feature of [49] reduces n^2 by utilizing only gray-level information of images, i.e., $n^2 = 256^2$. This feature has the obvious disadvantage that a lot of useful color information is ignored. Cheng *et al.* [20] propose to quantize each individual color channel of the RGB color space to 12 equidifferent values, which reduces n to $12^3 = 1728$. However, using rigid division to split color channels with equal intervals does not have a theoretical basis, and thus, it is unclear how effective they are. In contrast, employing a full color space and a nonrigid division is consistently more appropriate with human visual characteristics.

Although the color contrast can be computed efficiently by a more compact representation of the color information contained in images, this representation can also introduce some artifacts that confound some similar colors with the excessive different values. In order to overcome this negative effect, a smoothing procedure is employed to refine the feature values. This procedure is implemented by replacing the feature value

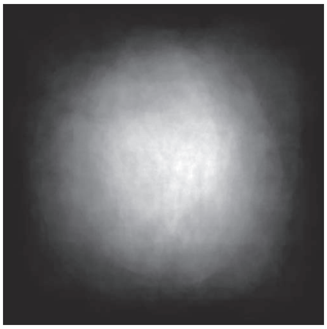


Fig. 4. Statistical map illustrates the saliency distribution for 200 normalized training images. Each pixel in the map indicates the possibility of being salient for a normalized image with fixed size. The high intensity indicates the high possibility of a location being salient.

of each color with the weighted feature values of similar colors [measured by (19)]

$$f(I_i) = \sum_{I_j \in \mathbb{M}_i} w(i,j) f(I_j) \tag{11}$$

where $\mathbb{M}_i$ denotes the m nearest colors of I_i. The weight $\{w(i,j)\}_j$ depends on the difference between the color I_i and I_j and satisfies the normalization constraints $0 \leq w(i,j) \leq 1$ and $\sum_j w(i,j) = 1$. More specifically

$$w(i,j) = \frac{1}{Z(i)} e^{-D(I_i,I_j)^2/h^2} \tag{12}$$

where $Z(i) = \sum_j e^{-D(I_i,I_j)^2/h^2}$ is the normalizing factor. The parameter h controls the decay of the weights as a function of the color differences. m and h are experimentally fixed to $n/4$ and 4, respectively, in all our experiments.

C. Center Bias

The center bias is also taken into account due to the fact that humans naturally put their attentional spotlight near the center of the image at the first glance, and photographers actually have a bias to make objects of their interest close to the center of the scene. This rule is true for the 200 training images as illustrated in Fig. 4. For these reasons, the proposed method additionally includes a feature which indicates the Euclidean distance to the center of the image for each pixel.

III. Experiments

A. Image Data Set

In order to evaluate the performances of DTMBVS under different settings, as well as to compare this method with state-of-the-art saliency detection methods, two publicly available data sets are employed. The first data set containing 1000 images is manually constructed by Achanta *et al.* [18] and has achieved great popularity in saliency detection [20]. Each of the selected image in this data set contains one or several salient objects and has an accurate object-contour-based ground truth. The experiments randomly select 200 images and their associated ground truths as the training set and use the remaining 800 as the testing set. Then, the trained model is further tested on the second data set, MSRA-B [34], [35], which contains 5000 well-labeled images, much larger than the first one.

B. Evaluation Measure

In the experiments for quantitative evaluation, the criterion called *precision-recall* curve is chosen to capture the tradeoff between accuracy and sensitivity. This parametric curve is sketched by varying the threshold used to binarize the saliency map. *Precision* measures the rate of the positive detected salient region to the whole detected region, while *recall* measures the rate of the positive detected salient region to the ground truth. Moreover, *F-measure* [50], which is a weighted harmonic mean of *precision* and *recall*, is also taken to provide a single index. More specifically, given the image with pixels $X = \{x_i\}$ and binary ground truth $G = \{g_i\}$, for any detected binary saliency mask $L = \{l_i\}$, these three indexes are defined as

$$precision = \sum_i g_i l_i / \sum_i l_i \tag{13}$$

$$recall = \sum_i g_i l_i / \sum_i g_i \tag{14}$$

$$F_\alpha = \frac{precision \times recall}{(1-\alpha) \times precision + \alpha \times recall} \tag{15}$$

where α is set to 0.5 according to Martin *et al.* [50].

C. Color Space Selection

A color space is a model where the independent components of color are precisely defined, by which one can quantify, generate, and visualize colors. Different color spaces are suitable for different applications. Therefore, it is necessary to experimentally determine the most appropriate color space for a particular task. In this section, six mostly employed color spaces, *CMYK*, *HSV*, *RGB*, *XYZ*, *YCbCr*, and *CIELAB*, are evaluated on the data set constructed by Achanta *et al.* to select the most suitable one for the proposed method.

The final results are shown in Fig. 5(a). For each color space, the corresponding best distance metric is employed (the discussion of distance metric is presented later in Section III-D). It is manifest that adopting *CMYK*, *HSV*, and *CIELAB* color spaces always obtains better results than using the others, and the *CIELAB* color space achieves better performance than *CMYK* and *HSV* most of the time according to the *precision-recall* curves. Moreover, the *F-measure* bars also show that the proposed method based on the*CIELAB* color space outperforms the other five in this perspective. According to these analyses, it is reasonable to believe that the *CIELAB* color space is more suitable for the proposed method. Therefore, the following experiments are all conducted on this color space.

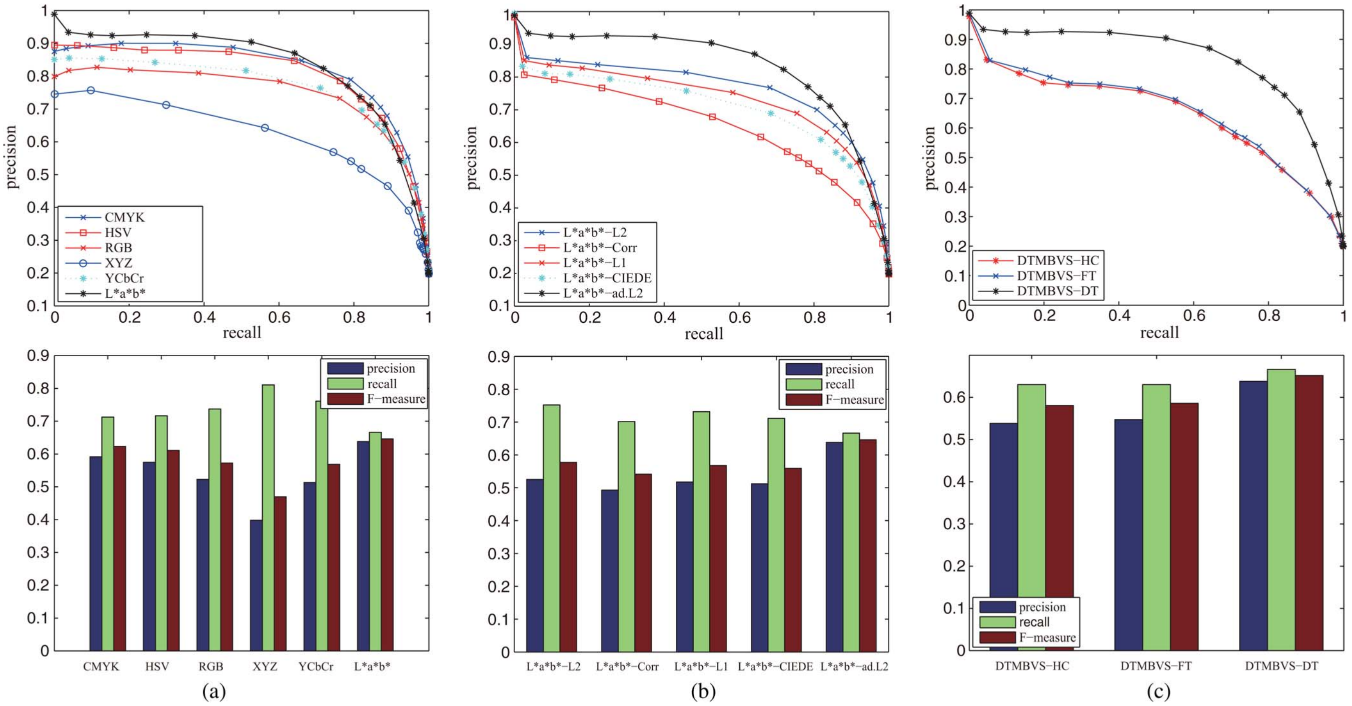


Fig. 5. *Precision-recall curves* and *averaged precision, recall, and F-measure bars* of saliency detection results by the proposed method using different settings. (a) Different color spaces. (b) Different color difference measurements in *CIELAB* color space. "$L^*a^*b^* - \mathrm{L2}$," "$L^*a^*b^* - \mathrm{Corr}$," "$L^*a^*b^* - \mathrm{L1}$," "$L^*a^*b^* - \mathrm{CIEDE}$," and "$L^*a^*b^* - \mathrm{ad.L2}$" are corresponding to (17), (18), (16), *CIEDE2000*, and (19), respectively. (c) Different color representations. "DTMBVS-HC," "DTMBVS-FT," and "DTMBVS-DT" are corresponding to the DTMBVS based on the color representation employed in [18] and [20] and the *differential threshold*-based color representation, respectively. All the experiments are based on the data set constructed by Achanta *et al.*

D. Color Difference Measurement

Measuring the difference or distance between two colors is an important aspect for color analysis. There are three commonly used formulas in practical applications [51], [52]. In general applications, the color difference of two points, I_1 and I_2, in a full color space is often directly derived from (16) or (17)

$$D(I_1, I_2) = \|I_1 - I_2\|_1 \tag{16}$$

$$D(I_1, I_2) = \|I_1 - I_2\|_2. \tag{17}$$

The third commonly used formula is based on the concept of correlation coefficient

$$D(I_1, I_2) = 1 - r(I_1, I_2) \tag{18}$$

where $r(I_1, I_2)$ is the correlation coefficient between I_1 and I_2.

Aside from these three universal formulas, there are also others designed for specific color spaces. In these formulas, the *CIEDE2000* color difference formula has been considered as the best choice for the *CIELAB* color metric [53].

It is notable that the aforementioned formulas give the same weight to each color channel. However, human observers are not with the equivalent sensibility to the hue, saturation, and lightness changes [54]. Hence, here, the authors propose other color difference formulas for the *CIELAB* (l^*, a^*, b^*) color space to match the characteristics of human perception

$$D(I_1, I_2) = \sqrt{\phi\,(l_1^* - l_2^*)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2} \tag{19}$$

$$\phi = \begin{cases} 1, & \text{if } \sigma_{L^*} \geq \max(\sigma_{a^*}, \sigma_{b^*}) \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

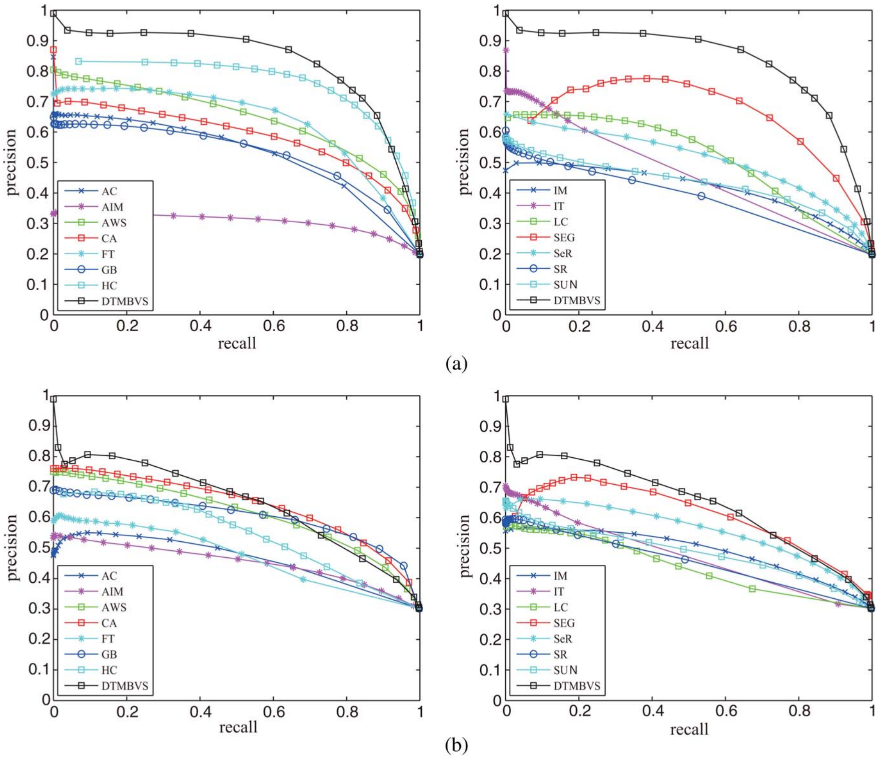
where σ_{L^*}, σ_{a^*}, and σ_{b^*} are the variances of the corresponding color channels.

The biggest difference between the proposed formula and the previously discussed four equations is that the proposed formula is based on a heuristic perceptual fact that lightness has less effect than the color opponent in the visual attention process, unless the lightness variation is large enough. This principle is also improved in our experiments. Fig. 5(b) demonstrates the comparative results. It is manifest that the best performance is achieved by employing the proposed formula.

E. Effects of Differential Threshold Representation

The proposed DTMBVS contains a *differential threshold*-based color representation, which aims to imitate the human visual system better while having high computational efficiency. In this section, two relevant color quantization methods [18], [20] are employed to substitute the *differential threshold*-based color representation. Then, their results are evaluated to verify the relative advantage of *differential threshold*.

Fig. 5(c) plots the results of the DTMBVS method based on different color quantizations. All these curves and bars altogether can prove that the *differential threshold*-based color representation has a clear advantage compared with other existing methods. Therefore, it is reasonable to believe that the proposed color representation is more suitable for the proposed method.



(a)

(b)

Fig. 6. *Precision-recall curves* of the proposed DTMBVS and state-of-the-art saliency detection methods on (a) data set constructed by Achanta *et al.* and (b) MSRA-B data set.

F. Comparison With Other Methods

The results of the proposed method are compared with 14 state-of-the-art saliency detection methods. They are respectively AC [26], AIM [38], AWS [24], [25], CA [33], FT [18], GB [37], HC [20], IM [41], IT [19], LC [49], SEG [32], SeR [30], SR [31], and SUN [39]. These 14 methods are selected according to four certain principles following [18] and [20]: recency (CA, HC, IM, and AWS are proposed during the last two years), high citation frequency (AIM, GB, IT, and SR have been cited over 200 times), variety (LC and HC are global contrast based; AC, FT, SeR, and SUN are local contrast based; IT and AWS are biologically inspired; and SEG and SR are fully computational), and relation to the proposed DTMBVS (HC and LC). The code for LC is from Cheng *et al.* [20]. For the other 13 selected methods, their codes are downloaded from the authors' home pages. Every method is used to compute saliency maps for all the testing images. Then, the obtained results are compared with the labeled ground truth for quantitative evaluation.

Fig. 6(a) illustrates the results on the data set constructed by Achanta *et al.* The *precision-recall* curves show that the proposed method clearly outperforms AC, AIM, AWS, CA, FT, GB, IM, IT, LC, SEG, SeR, SR, and SUN. The proposed method can locate salient regions with much more accuracy than these 13 existing methods, i.e., yield higher *precision* with the same *recall* rate over the 800-image testing set and vice versa. The proposed method outperforms HC most of the time, except the disadvantage of lower *precision* rates when the tasks place more emphasis on achieving extremely high *recall* rates. However, in practical applications, simply emphasizing the extremely high *recall* rate is not a satisfying choice. A balance between the *precision* and the *recall* rate must be more appropriate [50]. According to this principle, the moderate *precision* and *recall* rates should be referred to. In this case, the proposed method is the best choice in most applications.

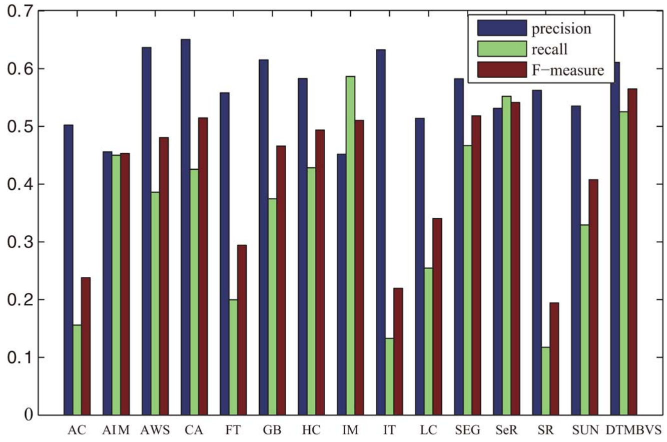


Fig. 7. *Averaged precision, recall, and F-measure bars* of the proposed DTMBVS and state-of-the-art saliency detection methods on MSRA-B data set.

In order to further validate the effectiveness and robustness of the proposed DTMBVS, the MSRA-B image set is also employed in this comparative experiment. The *precision-recall* results are presented in Fig. 6(b). It is obvious that DTMBVS outperforms AC, AIM, FT, HC, IM, IT, LC, SEG, SeR, SR, and SUN on this image set. However, as can be seen in Fig. 6(b), the *precision-recall curves* cannot provide discriminative clues

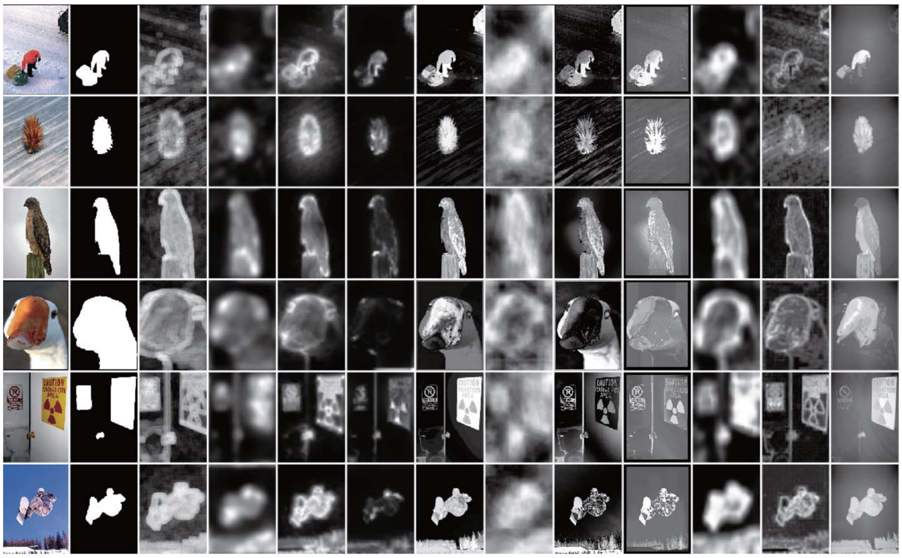


Fig. 8. Saliency detection results. From left to right, each column respectively represents the original images, the ground-truth labels, the saliency maps calculated by AIM [38], AWS [24], [25], CA [33], GB [37], HC [20], IM [41], LC [49], SEG [32], SeR [30], and SUN [39], and the saliency maps calculated by the proposed method.

for AWS, CA, and GB. Therefore, the *F-measure bars* should be taken to provide more discriminative information. The comparative results are presented in Fig. 7, which shows that the proposed DTMBVS clearly dominates other competitors in the *F-measure* indicator.

Several visual comparisons are also presented in Fig. 8 for qualitative evaluation. Only the top ten of the 14 aforementioned methods, AIM, AWS, CA, GB, HC, IM, LC, SEG, SeR, and SUN, are selected according to the *F-measure* indicator to compare with the proposed one. As can be seen from Fig. 8, the competitive ten methods tend to produce internally incongruous or morphological changed salient regions, while the proposed DTMBVS is prone to generate much more consistent results.

The salient region detected by the proposed method is visibly distinguished with the background. From this result, it can be inferred that the selected features (i.e., the *differential threshold*-based visual feature and the spatial constraint) can be more distinguishable than others and the model constructed with the MRF is more appropriate. Therefore, the proposed method has the ability to locate the truly salient regions in a greater probability for each image.

G. Robustness to Noise

Generally, in consideration of the actual conditions, a well-defined model is the one that not only achieves satisfying results for the testing data but also can resist a significant level of noise. In order to evaluate the robustness of all these saliency detection methods, a significant level of white Gaussian noise, which keeps the $SNR = 20$ dB, is added to each testing image. The second column of Fig. 9 shows an example image with added noise. This image is disturbed by numerous white spots. However, the object-background content can still be distinguished easily by human eyes.

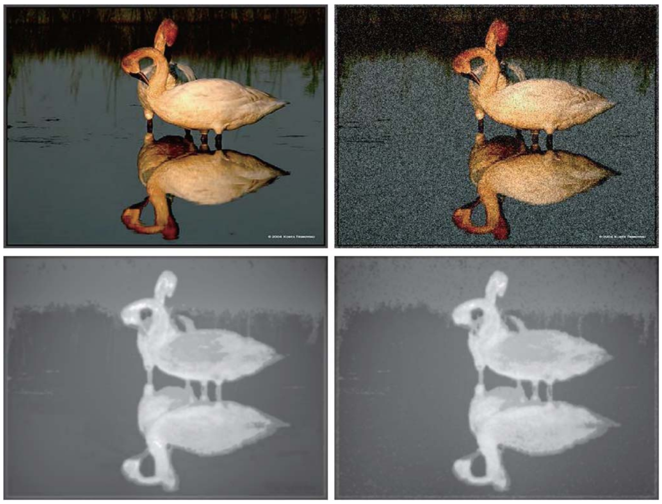


Fig. 9. First row presents an original image and the corresponding noisy image. The second row presents the saliency maps calculated by the proposed method.

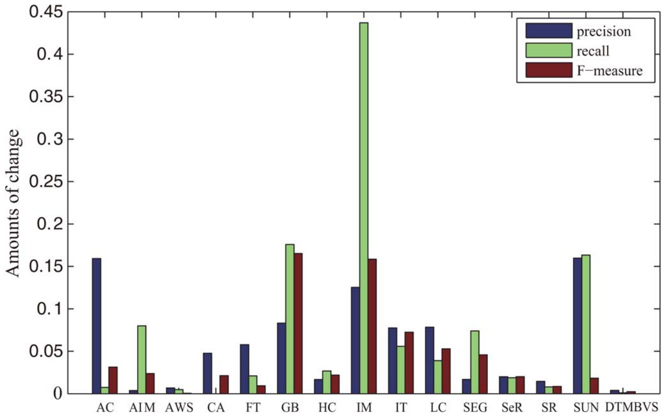


Fig. 10. Performance comparison after adding noise to images. The bars are used to indicate the amounts of changes in the average values of *precision*, *recall*, and *F-measure*. The SNR after adding noise is kept constant at 20 dB.

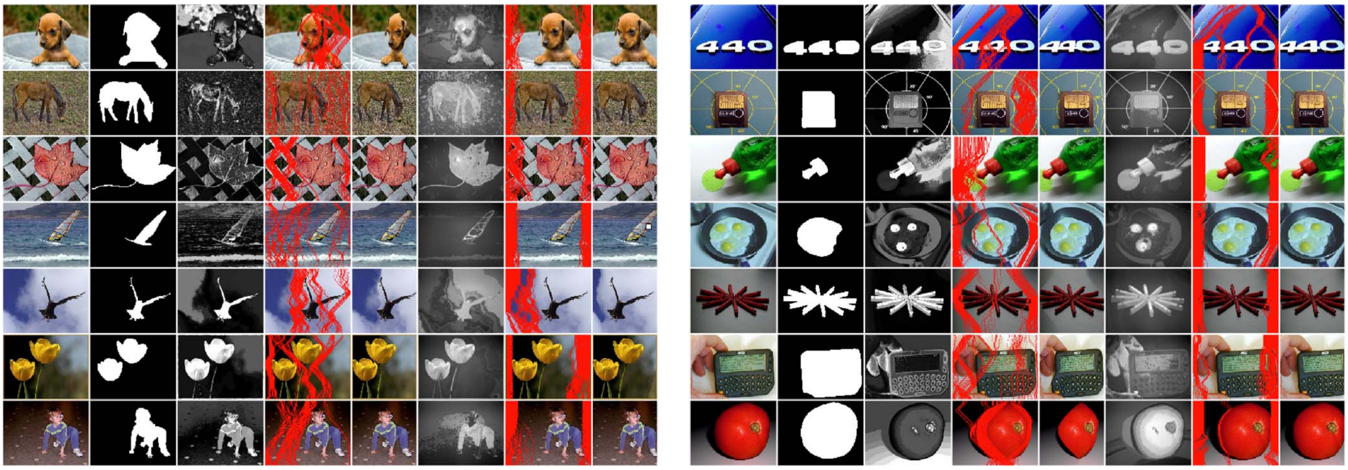


Fig. 11. Image retargeting. From left to right, each column respectively represents the original images, ground-truth saliency labels, saliency maps calculated by HC [20], carved seams by HC, resized image by HC, saliency maps calculated by the proposed DTMBVS, carved seams by DTMBVS, and the resized image by DTMBVS.

Then, the amounts of changes in the average values of *precision*, *recall*, and *F-measure* are compared to show the influence of the white Gaussian noise on different methods. Fig. 10 demonstrates the results on the data set constructed by Achanta *et al.* All these bars altogether show that the proposed DTMBVS performs a remarkable robustness to white Gaussian noise compared with other existing methods.

The reasons for the robustness of the proposed method are mainly the following: 1) The smoothing procedure in feature representation is effective at reducing the negative effects of the added white Gaussian noise, and 2) the employed model, which can strengthen the spatial constraint by MRF, is suitable for resisting this kind of interference.

IV. Content-Aware Image Retargeting Application

In some degree, practicability is an important evaluation aspect for a method. From this view, an excellent saliency detection method should enable many practical applications, which will achieve better results if human visual attention characteristics have been taken into account. Among these applications, a typical one named content-aware image retargeting is selected for further evaluation, which aims at flexibly resizing images by removing/expanding the noninformative regions.

Seam carving is a popular technique for image retargeting [55]. A seam is defined as a connected path of pixels going from the top (left) of an image to the bottom (right). By repeatedly removing or inserting seams, the image can be retargeted to the expected size. To obtain satisfying results, the removed or inserted seams should ensure that the salient regions in the image should not be disturbed.

Generally, seam carving is implemented by finding the path with the minimum cumulative energy and removing it from the image [55]. In this paper, the detected saliency maps are used for energy function definition. Through retargeting images to the 75% width of the original ones and judging the results subjectively, the practicability of the employed saliency detection method can be evaluated directly. The HC [20], which has the best performance among the competitive methods according to the *precision-recall* curves in Fig. 6, is chosen to conduct the comparative experiment.

Fig. 11 presents the intermediate process and final results. It is manifest that the DTMBVS maps can help to produce more eye-pleasing images than HC maps. The DTMBVS can generate saliency maps with higher and more consistent saliency values in target regions than HC. In this case, the seams through these regions are accurately avoided, and the resized results therefore have less distortions.

V. Conclusion

Humans can efficiently fix their attentional spotlight to the areas of interest even when no more cues other than color are employed. This gives us a clue that simple image features and saliency models might be competent for a good performance. In this paper, a supervised method for saliency detection has been presented. The proposed method mainly incorporates the single biologically inspired saliency feature, *differential threshold*-based color feature, to predict the possibilities of each pixel being salient through MRF learning. Its performance has been evaluated on two public image data sets. Experiments indicate that the proposed method outperforms other 14 state-of-the-art saliency detection methods in terms of both effectiveness and robustness. An application of seam carving is also involved, which intuitively exemplifies the usefulness of the proposed method.

Although various saliency detection methods have been proposed, the performance of these methods is still far from satisfying compared with the human visual system, particularly when tackling images of complex scenes. This is largely because many valuable cognitive principles of visual attention in the human visual system have not yet been considered. It is reasonable to believe that further introducing these principles in saliency detection will be beneficial for improving the state of the art.

REFERENCES

[1] W. James, *The Principles of Psychology*, vol. 1. New York: Henry Holt, 1890.

[2] Y. Yu, G. K. I. Mann, and R. G. Gosine, "An object-based visual attention model for robotic applications," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 5, pp. 1398–1412, Oct. 2010.

[3] M. I. Posner, J. A. Walker, F. J. Friedrich, and R. D. Rafal, "Effects of parietal injury on covert orienting attention," *J. Neurosci.*, vol. 4, no. 7, pp. 1863–1874, Jul. 1984.

[4] A. Belardinelli, F. Pirri, and A. Carbone, "Bottom-up gaze shifts and fixations learning by imitation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 2, pp. 256–271, Apr. 2007.

[5] A. M. Treisman and S. Sato, "Conjunction search revisited," *J. Exp. Psychol.: Human Percept. Perform.*, vol. 16, no. 3, pp. 459–478, Aug. 1990.

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

[7] S. Marat, M. Guironnet, and D. Pellerin, "Video summarization using a visual attentional model," in *Proc. Eur. Signal Process. Conf.*, 2007, pp. 1784–1788.

[8] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.

[9] J. You, A. Perkis, M. M. Hannuksela, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 561–564.

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

[11] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, 2012, to be published.

[12] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 2232–2239.

[13] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.

[14] C. Jung and C. Kim, "A unified spectral-domain approach for saliency detection and its application to automatic object segmentation," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1272–1283, Mar. 2012.

[15] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. II-37–II-44.

[16] L. Yang, N. Zheng, J. Yang, M. Chen, and H. Chen, "A biased sampling strategy for object categorization," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1141–1148.

[17] D. Walthera, U. Rutishausera, C. Kocha, and P. Peronaa, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Comput. Vis. Image Understand.*, vol. 100, no. 1/2, pp. 41–63, Oct. 2005.

[18] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.

[19] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 11, pp. 1254–1259, Nov. 1998.

[20] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 409–416.

[21] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[22] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition—A gentle way," in *Proc. Biol. Motiv. Comput. Vis.*, 2002, pp. 472–479.

[23] S. Frintrop, M. Klodt, and E. Rome, "A real-time visual attention system using integral images," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2007, pp. 1–10.

[24] A. Garcia-Diaz, V. Leborán, X. Fdez-Vidal, and X. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *J. Vis.*, vol. 12, no. 6, pp. 1–22, Jun. 2012.

[25] A. Garcia-Diaz, X. Fdez-Vidal, X. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image Vis. Comput.*, vol. 30, no. 1, pp. 51–64, Jan. 2012.

[26] R. Achanta, F. J. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Proc. Comput. Vis. Syst.*, 2008, pp. 66–75.

[27] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 481–488.

[28] D. Gao and N. Vasconcelos, "Integrated learning of saliency, complex features, and object detectors from cluttered scenes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 282–287.

[29] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *J. Vis.*, vol. 8, no. 7, pp. 1–18, Jun. 2008.

[30] H. J. Seo and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 45–52.

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

[32] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 366–379.

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

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

[35] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[36] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley, "Image saliency: From intrinsic to extrinsic context," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 417–424.

[37] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[38] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 155–162.

[39] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 1–20, Dec. 2008.

[40] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 2106–2113.

[41] 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. Vis. Pattern Recognit.*, 2011, pp. 433–440.

[42] S. Baluja, "Using expectation to guide processing: A study of three real-world applications," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 1–7.

[43] Q. Wang, P. Yana, Y. Yuana, and X. Li, "Multi-spectral saliency detection," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 34–41, Jan. 2013.

[44] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Visual saliency by selective contrast," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 7, pp. 1150–1155, Jul. 2013.

[45] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Texton-boost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 1–15.

[46] J. Verbeek and B. Triggs, "Scene segmentation with CRFs learned from partially labeled images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1553–1560.

[47] J. Shotton, A. Blake, and R. Cipolla, "Contour-based learning for object detection," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2005, pp. 503–510.

[48] E. Weber, H. E. Ross, D. J. Murray, and E. H. Weber, *On the Tactile Senses*, 2nd ed. London, U.K.: Psychology Press, 1996.

[49] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[50] D. R. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.

[51] A. Gijsenij, T. Gevers, and M. P. Lucassen, "A perceptual comparison of distance measures for color constancy algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 208–221.

[52] J. Hao, "Human pose estimation using consistent max-covering," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1357–1364.

[53] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Res. Appl.*, vol. 30, no. 1, pp. 21–30, Feb. 2005.

[54] H. Y. Lee, H. K. Lee, and Y. H. Ha, "Spatial color descriptor for image retrieval and video segmentation," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 358–367, Sep. 2003.

[55] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 1–10, Jul. 2007.

Guokang Zhu is currently working toward the Ph.D. degree in the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

His research interests include computer vision and machine learning.

Qi Wang received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Postdoctoral Researcher with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include computer vision and pattern recognition.

Yuan Yuan (M'05–SM'09) is a Researcher (Full Professor) with the Chinese Academy of Sciences, Xi'an, China, and her main research interests include visual information processing and image/video content analysis.

Pingkun Yan (S'04–M'06–SM'10) received the B.Eng. degree in electronics engineering and information science from the University of Science and Technology of China, Hefei, China, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore.

He is a Full Professor with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include computer vision, pattern recognition, machine learning, and their applications in medical imaging.