A Visual-Attention Model Using Earth Mover's Distance-Based Saliency Measurement and Nonlinear Feature Combination

Yuewei Lin, *Student Member*, *IEEE*, Yuan Yan Tang, *Fellow*, *IEEE*, Bin Fang, *Senior Member*, *IEEE*, Zhaowei Shang, Yonghui Huang, and Song Wang, *Senior Member*, *IEEE*

Abstract—This paper introduces a new computational visual-attention model for static and dynamic saliency maps. First, we use the Earth Mover's Distance (EMD) to measure the center-surround difference in the receptive field, instead of using the Difference-of-Gaussian filter that is widely used in many previous visual-attention models. Second, we propose to take two steps of biologically inspired nonlinear operations for combining different features: combining subsets of basic features into a set of super features using the L^m-norm and then combining the super features using the Winner-Take-All mechanism. Third, we extend the proposed model to construct dynamic saliency maps from videos by using EMD for computing the center-surround difference in the spatiotemporal receptive field. We evaluate the performance of the proposed model on both static image data and video data. Comparison results show that the proposed model outperforms several existing models under a unified evaluation setting.

Index Terms—Visual attention, saliency maps, dynamic saliency maps, earth mover's distance (EMD), spatiotemporal receptive field (STRF)

1 Introduction

Visual attention is an important mechanism in human vision: Despite the relatively large field of view, the human visual system processes only a tiny central region (the fovea) with great detail [18], [30], [44], [45], [17], [23], [33], [47]. This indicates that people usually focus on a small number of salient points (or locations) when they view a scene. Recently, developing computational models and algorithms to simulate the human visual attention has been attracting much interest in the computer vision society. An inclusion of a computational visual-attention model can substantially help address many challenging computer vision and image processing problems. For example, object detection and recognition can become much more efficient and more reliable by examining only the salient locations and ignoring large irrelevant background. Object tracking can also benefit from visual attention by examining only the spatiotemporally salient points. Because the neural mechanisms of the human vision system are still not fully known, it is a very challenging problem to build a comprehensive computational model that can well simulate the human visual attention mechanism. In the past decades, psychologists, neurobiologists, and computer scientists have all investigated visual attention from their own perspectives and benefit from the progress made in the other fields [10].

Previous research has shown that there are two kinds of visual attention mechanisms: bottom-up attention and top-down attention [7]. The bottom-up attention searches for the salient points based solely on the visual scene, i.e., image data, and therefore it is usually task irrelevant. Many computational visual-attention models developed in previous works are purely bottom-up [9] without assuming any specific prior knowledge on the objects and/or background. In specific applications where some prior knowledge is available or can be learned from training samples, people also include top-down mechanisms to improve the accuracy of the salient-point identification. For example, in [43] global scene configuration is used to guide the visual attention for localizing specific objects, such as people. It is worth mentioning that learning is not only used for top-down attention. Many pure bottom-up attention models also use learning to reveal some general knowledge that is applicable to different kinds of images in different applications [5], [14], [22], [29], [48]. In this paper, we focus on the pure bottom-up attention without using any task-relevant knowledge and without incorporating any learning components.

One of the most well-known bottom-up models for visual attention was developed by Itti et al. [16]. In Itti's model, an input image is first decomposed into the intensity, color, and orientation features in different image scales. A feature map is then generated by calculating the strength of each feature in each scale, where the feature strength at a point is defined by the center-surround difference at this point. In

- *Y. Lin and S. Wang are with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208. E-mail: ywlin.cq@gmail.com, songwang@cec.sc.edu.*
- *Y.Y. Tang is with the Department of Computer and Information Science, University of Macau, Macau and the College of Computer Science, Chongqing University, Chongqing 400030, China. E-mail: yytang@cqu.edu.cn.*
- *B. Fang, Z. Shang, and Y. Huang are with the College of Computer Science, Chongqing University, Chongqing 400030, China. E-mail: {fb, szw, hyh2009}@cqu.edu.cn.*

Manuscript received 23 Feb. 2011; revised 12 Nov. 2011; accepted 19 May 2012; published online 22 May 2012.
Recommended for acceptance by S. Avidan.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-02-0123.
Digital Object Identifier no. 10.1109/TPAMI.2012.119.

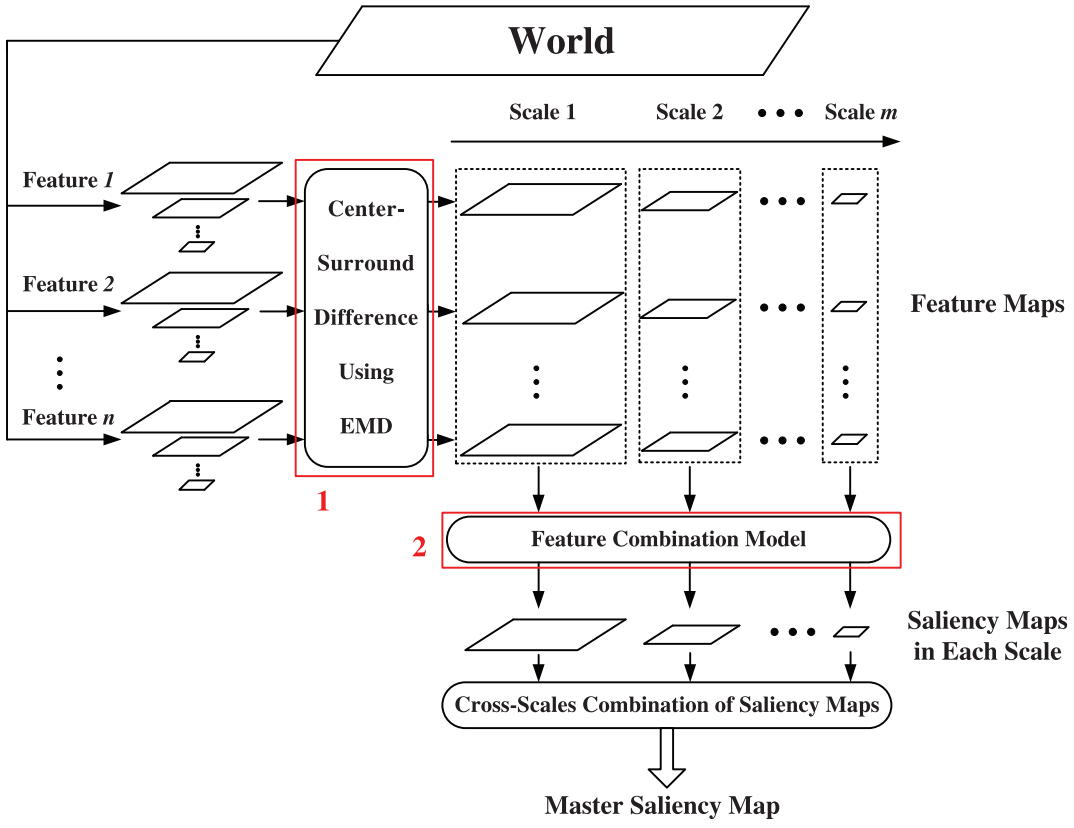


Fig. 1. The diagram of the proposed visual-attention model.

Itti's model, the center-surround difference is computed using a Difference-of-Gaussian (DoG) filter. After that, three conspicuity maps are constructed, one for each feature, by combining the feature strengths across multiple scales. Finally, these three conspicuity maps are linearly combined to produce a master saliency map. Parkhurst et al. [31] find that the saliency map produced by Itti's model shows better agreement with the human fixation points than that produced by chance. Recently, Itti's model has been extended to incorporate features other than intensity, color, and orientation. In [19], [38], dynamic saliency maps are generated from a video sequence by considering the motion feature. In [3], additional feature maps are constructed to reflect the symmetry and the object size in the image and then combined with other features to compute the master saliency map.

In this paper, we propose several new improvements over Itti's model. First, we propose using the Earth Mover's Distance (EMD) to measure the center-surround difference in the receptive field instead of using the DoG filter adopted in Itti's model. By comparing the histograms of the center and surround regions, EMD can provide a more robust measurement of their difference. Second, we propose using nonlinear operations, instead of the linear summation in Itti's model, for feature combination. More specifically, we propose to take two steps of biologically inspired nonlinear operations for combining the different features: First combining subsets of basic features into a set of *super features* using the L^m-norm and then combining all the super features using a Winner-Take-All (WTA) mechanism. Third, to construct the dynamic saliency maps from an input video, we extend the proposed model by computing the center-surround difference in the Spatiotemporal Receptive Field (STRF). These improvements are justified by an apple-to-apple performance comparison against several other existing visual-attention models in a unified experiment setting. The diagram of the proposed visual-attention model is illustrated in Fig. 1.

Recently, several new models have been developed for the bottom-up visual attention. In [4], [5], Bruce and Tsotsos measure the saliency using Shannon's self-information measure at each local image patch, where the feature of the patch is derived from an Independent Component Analysis (ICA) on a large number of patches in the image. In [48], Zhang et al. proposed a model of Saliency Detection using Natural Statistics (SUN), where a Bayesian inference is used to estimate the probability that there is a target at each location. Statistics on a large set of images are used to determine the priors in the Bayesian inference. In [12], Harel et al. described a Graph-Based Visual Saliency (GBVS) model where spectral graph analysis is used for computing the center-surround difference and its normalization. In [13], Hou and Zhang proposed using the spectral residual (SR) of an image as the saliency, where the spectral residual is defined by the log spectrum of an image and its smoothed version. In [14], Hou and Zhang further introduced a Dynamic Visual Attention (DVA) model by maximizing the entropy of the sampled visual features, where the entropy is measured by the incremental coding length. In [2], Avraham and Lindenbaum developed a validated stochastic model to estimate the probability that an image part is of interest and used this probability as saliency. In Section 5, we conduct

experiments to compare the performance of the proposed model with these models.

This paper is organized as follows: In Section 2, we introduce EMD and use it to measure the center-surround difference, as indicated in box "1" in Fig. 1. In Section 3, we discuss the two steps of nonlinear operations for combining the basic features, as indicated in box "2" in Fig. 1. In Section 4, we extend the proposed model to construct dynamic saliency maps from a video. In Section 5, we evaluate the performance of the proposed model on standard datasets and compare its performance to several existing visual-attention models, followed by a brief conclusion in Section 6.

2 Center-Surround Difference Using EMD

In Itti's model, DoG filter is used to compute the center-surround difference. In particular, DoG filter is implemented by applying a Gaussian filter to the image in different scales and then computing their difference. In [11], Gao and Vasconcelos suggest the use of the histogram difference between the center and the surround as the center-surround difference. Specifically, they suggest the use of the KL divergence for this purpose. However, as a bin-by-bin dissimilarity measure, the KL divergence considers only the correspondence between the bins with the same index and does not consider the information across bins. It is also well known that the KL divergence is sensitive to the selection of the bin size [37]. In this paper, we propose to use EMD to compute the center-surround difference.

2.1 Earth Mover's Distance between Two Histograms

EMD was first introduced and used by Rubner et al. [36], [37] for measuring the color and texture difference, where the EMD is applied to the signatures of distributions rather than directly to the histograms. A histogram can be viewed as a special type of the signatures [28] and in this section, we briefly overview EMD between two normalized histograms (the total amount of a histogram is a constant, e.g., 1) with the same number of bins [28].

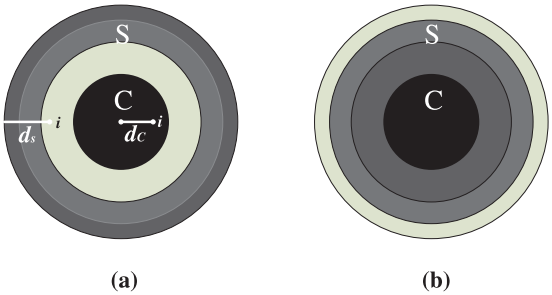
Let us consider EMD between two histograms $P = \{p_i, i = 1, 2, \ldots, n\}$ and $Q = \{q_j, j = 1, 2, \ldots, n\}$, where n is the number of bins. We introduce another all-zero n-bin histogram R and denote flow f_{ij} to be the amount that is moved from the bin i in P to the bin j in R. EMD between P and Q can then be defined as the minimum total flow (weighted by the moving distance of each flow) that is needed to make R to be identical to Q. Mathematically, EMD between P and Q can be written as

$$EMD(P, Q) = \min_{\{f_{ij}, i, j = 1, 2, \ldots, n\}} \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} d_{ij},$$

subject to

$$\sum_{j=1}^{n} f_{ij} = p_i, \quad \sum_{i=1}^{n} f_{ij} = q_j, \quad f_{ij} \geq 0, \quad i, j = 1, 2, \ldots, n,$$

where d_{ij} is the distance between the bins i and j. In this paper, we simply use the L^1 distance, i.e., $d_{ij} = |i - j|$.



(a) (b)

Fig. 2. An illustration of the motivation for constructing the weighted histograms. (a) and (b) Two center-surround regions with the same center-surround difference of 124.6667 when we use EMD on the unweighted intensity histograms. However, their center-surround differences are different (163.3668 for (a) and 87.3948 for (b)) when we use EMD on the proposed weighted histograms. In this illustrative example, the histograms are constructed using 256 bins.

2.2 EMD Based on Weighted Histogram

If we directly construct the histograms of the center and surround and then use the above EMD as the center-surround difference, the spatial information of the pixels is not considered. Intuitively, pixels near the border between the center and its surround are more important than the others for computing saliency. For example, let us consider two center-surround regions in Fig. 2. For both of them, the center C is a circular disk with radius 100 and intensity 0, and the surround S consists of three rings with outer radii 141, 173, and 200 pixels, respectively. Note that these three rings are of the same area for both Figs. 2a and 2b. The intensity of these three rings (from the outer ring to the inner ring) is 64, 90, and 220 in Fig. 2a and 220, 90, and 64 in Fig. 2b. Based on intensity histograms, it is easy to find that the center-surround differences in Figs. 2a and 2b are identical if we use EMD directly. However, perceptually the center-surround difference in Fig. 2a should be larger than that in Fig. 2b because there is clearly a larger intensity change across the center-surround border in Fig. 2a. In the following, we address this issue by introducing weighted histograms for both the center and the surround and then applying EMD to the weighted histograms.

First, we define a normalized weight $w(\cdot)$ for each pixel i in the center or the surround by

$$\begin{cases} w(i) = \dfrac{d_C(i)}{\sum_{j \in C} d_C(j)}, & \text{if } i \in \text{ center } C \\ w(i) = \dfrac{d_S(i)}{\sum_{j \in S} d_S(j)}, & \text{if } i \in \text{ surround } S, \end{cases}$$

where $d_C(i)$ denotes the euclidean distance from pixel i to the center of C and $d_S(i)$ denotes the shortest euclidean distance from pixel i to the outer boundary of S, as illustrated in Fig. 2a. Based on these weights, we construct normalized weighted histograms for the center and the surround, by using $w(i)$ as pixel-i's contribution to its histogram bin. By applying EMD to the weighted histograms of the center and the surround, we can achieve a center-surround difference that puts more weight on the pixels near their border. For example, the EMD-based center-surround difference is 163.3668 for the case shown in Fig. 2a and 87.3948 for the case shown in Fig. 2b by using the weighted histograms.

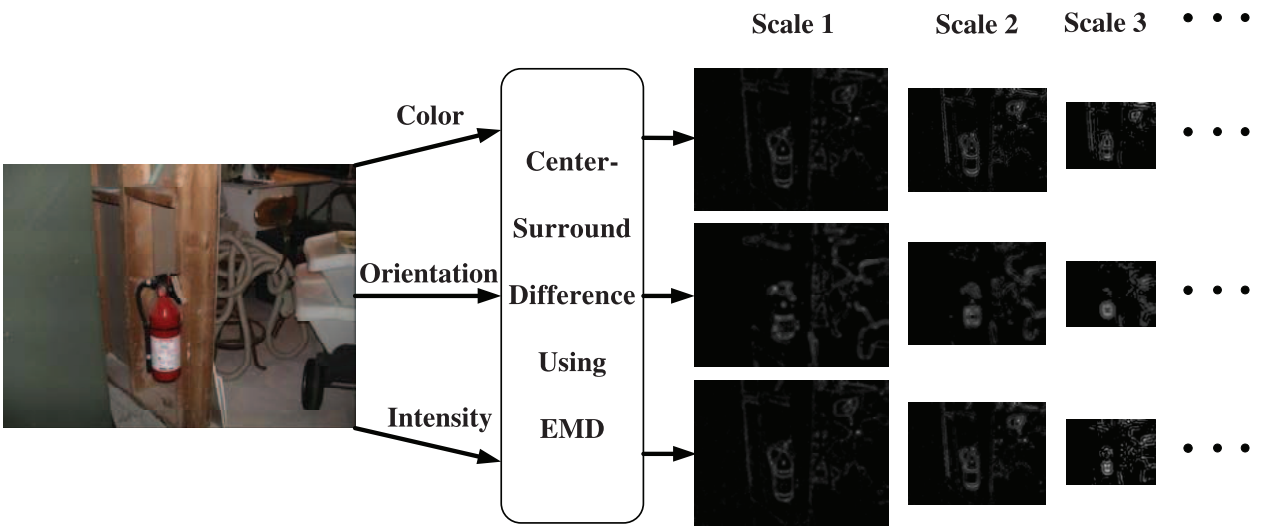


Fig. 3. An illustration of the pipeline of constructing the feature map for each feature in each scale.

Directly calculating EMD between two histograms is computationally expensive, with a complexity of $O(n^3)$ [28]. As mentioned above, we choose to use L^1-based bin distance d_{ij} and the normalized (weighted) histograms in this paper. With these choices, Levina and Bickel have proven that EMD equals the linear Wasserstein distance [25], which can be efficiently computed by

$$EMD(H_C, H_S) = \sum_{i=1}^{n} \left| \sum_{j=1}^{i} H_C(j) - \sum_{j=1}^{i} H_S(j) \right|, \qquad (1)$$

where H_C and H_S are the n-bin (normalized) weighted histogram for the center and the surround, respectively. Using (1), the EMD between two histograms can be computed with a linear complexity of $O(n)$.

As shown in Fig. 3, we construct weighted histograms based on different features (color, intensity, and orientation) and in different image scales. For each feature in each scale, we construct a feature map by calculating the center-surround difference at each pixel using the above-mentioned EMD, i.e.,

$$\mathbb{F}_l^f(x,y) = EMD\big(H_{C,l}^f(x,y), H_{S,l}^f(x,y)\big), \qquad (2)$$

where $\mathbb{F}_l^f$ denotes the feature map of feature f in scale l. $H_{C,l}^f(x,y)$ and $H_{S,l}^f(x,y)$ denote the weighted histograms of the center and the surround at pixel (x,y) in terms of feature f in scale l.

3 Feature Combination

Combining the feature maps for different features and from different scales is a very important component in visual attention [10]. In Itti's model [16], [17], [18], [46], feature maps for one feature in different scales are first linearly combined into a conspicuity map for this feature. Conspicuity maps for different features are then combined linearly into a master saliency map. In this paper, we first combine different features in each scale into a saliency map and then combine the saliency maps from different scales into a final master saliency map. When combining the saliency maps from different scales, we follow Itti's model by simply using the linear combination with equal contribution from different scales [16]. In this section, we focus on describing a biologically inspired model for combining the feature maps of different features into a saliency map in each scale.

Our proposed feature combination model is inspired by the primary visual cortex (V1) model proposed by Li and Koene [26], [24]. V1 is the simplest, earliest, and best studied visual area in the brain. The V1 model is a biologically based model that describes how the V1 neural responses can create a saliency map. Specifically, in the V1 model, saliency at a specific location is determined by the V1 cell with the greatest firing rate by following a Winner-Take-All mechanism. Additionally, some V1 cells only respond to a single feature and the others may be tuned to more than one feature. The latter are usually called feature conjunctive cells [26]. For example, there are CO cells that can respond to both color and orientation [24].

As illustrated in Fig. 4 [24], there are two major differences between the V1 model and Itti's model for feature combination. First, the feature combination is linear in Itti's model while it is nonlinear in the V1 model. Second, basic features (e.g., color, intensity, orientation) are directly combined in Itti's model, while according to the V1 model, some features may be associated to reflect the feature-conjunctive cells (e.g., CO and MO in Fig. 4) before they are combined with other features to generate the saliency map. Recently, people have found problems of using the linear feature-combination model. Poirier et al [32] pointed out: "... incremental changes in homogeneity had a greater effect on saliency when homogeneity was high than when it was low. This effect was observed both within and between dimensions. A purely additive combination (e.g., [17]) can therefore be ruled out, and models assuming such a combination rule would need to be updated to account for the current results." In [35], Riesenhuber and Poggio also found that MAX-like mechanisms at some stages of the circuitry seem to be more compatible with neurophysiological data than the linear summation mechanism with equal

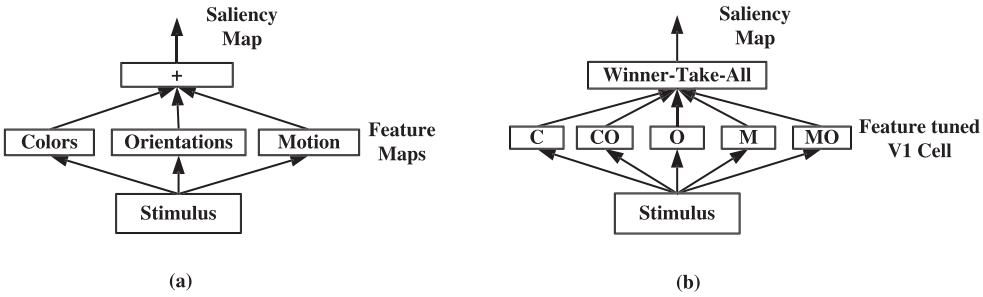


Fig. 4. An illustration of the differences between (a) the linear feature-combination used in Itti's model and (b) the nonlinear feature combination in the V1 model. This figure was adapted from [24].

weights. In [26], Li also pointed out that neither the neural mechanisms nor the exact underlying cortical areas responsible for the feature and saliency maps have been clearly specified in the linear summation model.

In this paper, we follow the V1 model to develop a model for nonlinear feature combination, as shown in Fig. 5. Given the set of N basic features (e.g., color, orientation, intensity),

$$\mathcal{F} = \{f_1, f_2, \ldots, f_N\},$$

we construct a set of $\tilde{N}$ *super features*,

$$\tilde{\mathcal{F}} = \{\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_{\tilde{N}}\},$$

where each super feature $\tilde{f}_i$ represents a subset of features in $\mathcal{F}$. If $\tilde{f}_i$ contains more than one basic feature, it models the response of a feature conjunctive cell, such as CO and MO.

In [42], To et al. used the L^m-norm[1] for combining the perception of the complex and suprathreshold visual elements in naturalistic visual images. Specifically, the L^m-norm [39], [42] over n real numbers a_i, $i = 1, 2, \ldots, n$, is defined by

$$\left(\sum_{i=1}^{n} a_i^m\right)^{1/m},$$

where m is a preset summating exponent. Following this model, in this paper we use the L^m-norm to construct superfeature maps $\mathbb{F}_l^{\tilde{f}}$, $\tilde{f} \in \tilde{\mathcal{F}}$, from the involved basic feature maps by

$$\mathbb{F}_l^{\tilde{f}}(x,y) = \left[\sum_{f \in \tilde{f}} \left[\mathbb{F}_l^f(x,y)\right]^m\right]^{1/m}. \quad (3)$$

We then use the WTA mechanism for combining the super features, i.e.,

$$\mathbb{S}_l(x,y) = \max_{\tilde{f} \in \tilde{F}} \mathbb{F}_l^{\tilde{f}}(x,y), \quad (4)$$

where $\mathbb{S}_l$ is the derived saliency map in scale l.

Note that if we only construct one super feature that involves all the basic features and use the L^m-norm with exponent $m = 1$, the above nonlinear feature combination model is degenerated to the linear feature combination model. In this paper, we consider three basic features of color, intensity, and orientation as in most previous visual-attention models and as suggested in [42], set the summating exponent $m = 2.8$ in constructing super features. We construct two super features:

$$\tilde{f}_1 = \{\text{Color}, \text{Orientation}\},$$
$$\tilde{f}_2 = \{\text{Intensity}\}.$$

The super feature $\tilde{f}_1$ reflects the CO tuned cells in the V1 model. We choose intensity itself as a separate super feature without associating it to other basic features because there is no evidence of any intensity-tuned cells [10]. In Fig. 6, we show the different saliency maps generated from a sample image when using the L^m-norm or linear summation for constructing super features. Note that, compared to the feature combination of directly taking the maximum over these three basic features, our construction of super features puts a relative lower weight on the intensity feature because the L^m-norm of the color and orientation features is always larger than or equal to the maximum of these two features. In the later experiments, we show that these two super features lead to better visual attention performance than the other possible ways of superfeature construction. Fig. 7 summarizes the proposed nonlinear feature combination in a scale.

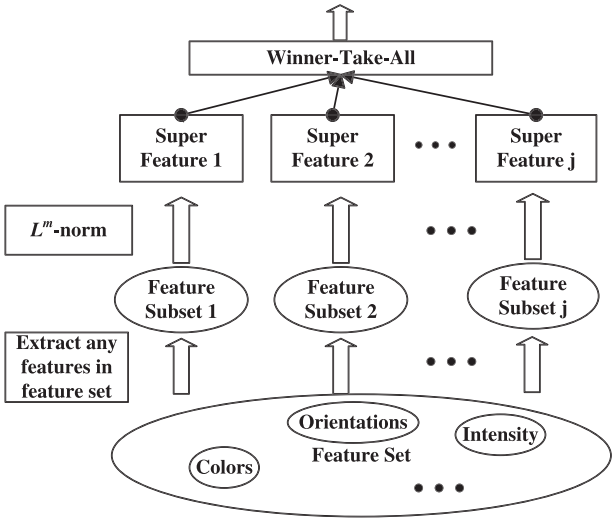


Fig. 5. An illustration of the proposed feature combination model.

1. In [42], this nonlinear operator was called Minkowski summation. However, in mathematics, Minkowski summation usually indicates the dilation of two sets in geometry. To avoid confusion, in this paper we call it L^m-norm instead of Minkowski summation.

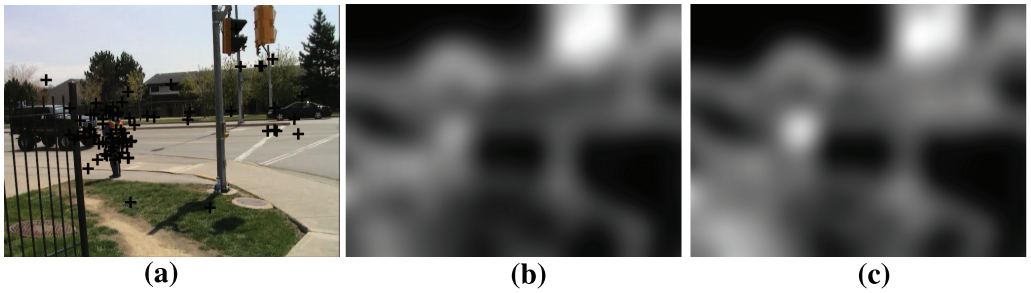


Fig. 6. An example to illustrate the use of the L^m-norm for constructing super features. (a) An input image. (b) The resulting saliency map when using the linear summation to construct the super feature $\{\text{Color}, \text{Orientation}\}$. (c) The resulting saliency map when using the proposed L^m-norm ($m = 2.8$) for constructing the super feature $\{\text{Color}, \text{Orientation}\}$.

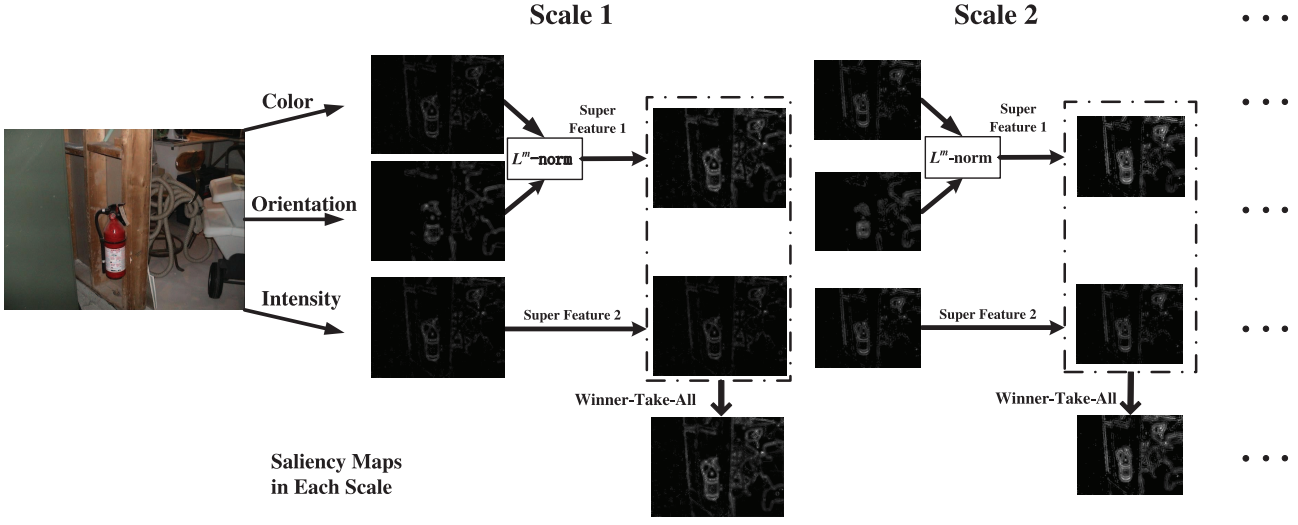


Fig. 7. An illustration of the proposed nonlinear feature combination model on the three basic features of color, intensity, and orientation.

4 Constructing Dynamic Saliency Maps from Video

Videos provide information to construct dynamic saliency maps over time. In this section, we extend the computation of the center-surround difference from a single static image to a sequence of image frames to construct a dynamic saliency map. Using the center-surround difference for saliency map reflects the function of the Receptive Field (RF) in neurophysiology [8]: Classical RF has a roughly circular, center-surround organization [15], as shown in Fig. 8. There are two primary configurations: One is shown in Fig. 8a, where the RF center is responsive to bright stimuli and its surround is responsive to dark stimuli, and the other one is shown in Fig. 8b, where the RF center is responsive to dark stimuli and its surround is responsive to bright stimuli.

While the RFs in spatial coordinates are widely used, the RF organizations are not actually static. When examined in the space-time domain, the RFs of most cells in the geniculo-cortical pathway exhibit striking dynamics [6]. Recent measurement techniques have made it possible to plot the full spatiotemporal RF (STRF) of the neurons, which include specific excitatory and inhibitory subregions that vary over time [1], [6], [8]. As shown in Fig. 9a, an X-T (spatial x-axis over the temporal t-axis) plot summarizes how the 1D spatial organization of the RF changes over time. This X-T plot typically exhibits a center-surround organization in space and a biphasic structure in time. Panel A in Fig. 9a shows the temporal response curves obtained by slicing through the X-T data at the center of the RF, whereas panels B and C in Fig. 9a show the spatial RF profiles determined at two different times ($t = 60$ ms and $t = 25$ ms, respectively). Fig. 9b shows the approximate construction by thresholding the X-T plot shown in Fig. 9a. We use this approximated construction of the STRF profile for defining the center and the surround in the space-time domain. Note that the regions in the top left and top right of Fig. 9a are not reflected in Fig. 9b because the corresponding excitatory is low (see Panel B) and is ignored after the thresholding.

From Fig. 9b, we can see that, in the STRF, a surround is made up of two parts: a spatial surround and a temporal surround. The difference between the center and the spatial surround reflects the static saliency and the difference between the center and the temporal surround reflects the motion saliency. Combining both of them, we can derive the dynamic saliency at each location in the space-time

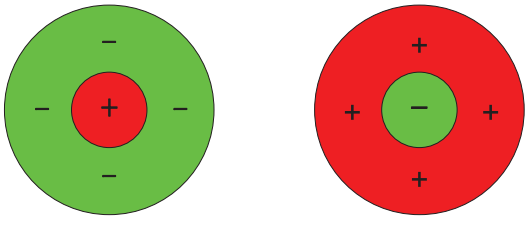


Fig. 8. An illustration of the spatial RF structure of neurons.

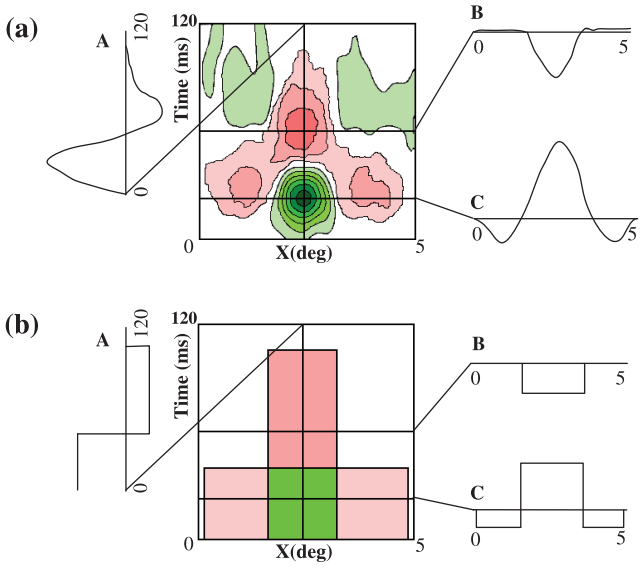


Fig. 9. An illustration of the STRF organization. (a) A sample STRF profile (X-T plot), adapted from [6]. (b) Approximate construction of the STRF profile (X-T plot).

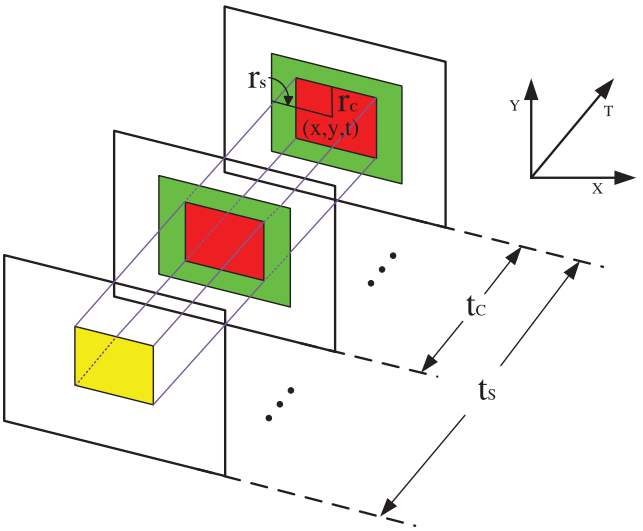


Fig. 10. An illustration of the center and the surround used in the proposed method for deriving a dynamic saliency map.

domain. Specifically, at a spatiotemporal location (x, y, t) we define the center C and the surround S in the STRF as

$$\begin{aligned} C(x,y,t) &= \{(x',y',t') | \max(|x-x'|, |y-y'|) < r_C, \\ &\quad 0 < t - t' < t_C\}, \\ S(x,y,t) &= \{(x',y',t') | r_C < \max(|x-x'|, |y-y'|) < r_S, \\ &\quad 0 < t - t' < t_C\} \\ &\quad \cup \{(x',y',t') | \max(|x-x'|, |y-y'|) < r_C, \\ &\quad t_C < t - t' < t_S\}, \end{aligned}$$

where r_C and r_S define the center and the surround spatially, and t_C and t_S define the center and the surround temporally, as illustrated in Fig. 10.

At each spatiotemporal location, we construct the feature maps by using EMD based on the weighted histograms as described in Section 2 (see (1) and (2)). For the feature selection, superfeature construction, and feature combination, we use the same methods as described in Section 3 (see (3) and (4)).

5 Experiments

As in many previous works, we use Bruce's data [4], [5] for evaluating the visual-attention performance on static images and Itti's data [20], [21] for evaluating the visual-attention performance on videos. Bruce's data consist of 120 color images, on each of which a set of human-eye tracking fixations and a human density map (blurred fixations) are provided as the ground truth for evaluation. Itti's data consist of 50 original video clips and 50 MTV video clips that are constructed by dividing and reassembling original video clips. On each of these video clips, a set of fixation points is provided as the ground truth for evaluation.

For the proposed model, we use three basic features: color, intensity, and orientation. For the calculation of these features, we follow Itti's model: two color features $r - g$ and $b - \min(r, g)$, one intensity feature $\frac{r+g+b}{3}$, and four orientation features constructed by applying Gabor filters along four orientations $\{0°, 45°, 90°, 135°\}$, with r, g, and b being the original RGB color values. In constructing the super features, we set the summating exponent $m = 2.8$. Weighted histograms are constructed with 16 bins for computing the EMD-based center-surround difference in all our experiments. In handling static images, i.e., Bruce's data, we set the boundary of the center to be a 3×3-pixel square and the outer boundary of the surround to be 7×7-pixel square. For each static image, we also construct an image pyramid with five scales and then use the four coarser scales (i.e., excluding the original scale) for computing saliency maps. In handling videos, i.e., Itti's data, we set $r_C = 1$ pixel, $r_S = 3$ pixels, $t_C = 1$ frame, and $t_S = 3$ frames. For each video clip, we also construct a spatial pyramid with five scales and then use the three coarsest scales for computing the dynamic saliency maps.

On Bruce's data, we compare the performance of the proposed model with several other existing visual-attention models, including Itti's model (using the implementation at http://www.klab.caltech.edu/~harel/share/gbvs.php), the GBVS model [12], the model of Attention based on Information Maximization (AIM) [5], the Spectral Residual model [13], the DVA model proposed in [14], the SUN model [48], and the Esaliency model [2]. On Itti's data, we compare the performance of the proposed model with the performance of Itti's model [16], the Variance model [34], and the Surprise model [20], [21].

Many previous works reported the performance on Bruce's data using the ROC curves and their Area Under Curve (AUCs): The master saliency map is first thresholded to a binary map, which is then compared to the ground truth for determining the true-positive rate (TPR) and the false-positive rate (FPR). By varying the threshold for the master saliency map, we construct an ROC curve. However, we found that the AUCs reported on the previous works may not be directly comparable because of the following two reasons:

TABLE 1
AUCs of the Proposed Model and the Comparison Models on Bruce's Data under Their Own Settings

Model	AUCs Reported in Previous Works	Ground Truth We Used	AUCs We Obtained	With Center Bias?
AIM [4], [5]	0.7810 [5]	Density Map, $T_d = 0.50$	0.7800	No
DVA [14]	0.7928 [14]	Density Map, $T_d = 0.15$	0.7911	Yes
SUN [48]	0.6682 [48]	Fixation Points	0.6664	No
Itti's Model [16]	0.6146 [48]	Fixation Points	0.7973	Yes
Itti's Model [16]	None	Density Map, $T_d = 0.0$	0.8020	Yes
Itti's Model [16]	None	Density Map, $T_d = 0.3$	0.8651	Yes
GBVS [12]	None	Fixation Points	0.8253	Yes
SR [13]	None	Fixation Points	0.6952	No
Esaliency [2]	None	Fixation Points	0.7118	Yes
Proposed Model	None	Fixation Points	0.7403	No
Proposed with Bias	None	Fixation Points	0.8365	Yes

TABLE 2
AUCs of the Proposed Model and Other Comparison Models on Bruce's Data under Four Unified Settings Where the Center Bias Is Removed

Model	Setting 1	Setting 2	Setting 3	Setting 4
Itti's Model [16]	0.6530±0.0092	0.5734±0.0052	0.6608±0.0094	0.7055±0.0117
GBVS [12]	0.6587±0.0087	0.5514±0.0051	0.6408±0.0091	0.6979±0.0122
SR [13]	0.6759±0.0089	0.5712±0.0055	0.6690±0.0089	0.7072±0.0113
DVA [14]	0.6867±0.0074	0.5732±0.0054	0.6715±0.0079	0.7257±0.0100
AIM [4], [5]	0.6788±0.0089	0.5866±0.0054	0.6645±0.0094	0.7129±0.0116
SUN [48]	0.6669±0.0103	0.5690±0.0064	0.6343±0.0102	0.6793±0.0130
Esaliency [2]	0.6550±0.0083	0.5669±0.0048	0.6411±0.0088	0.6847±0.0115
Proposed Model	**0.6959±0.0091**	**0.5881±0.0055**	**0.6809±0.0090**	**0.7275±0.0113**

Setting 1 uses the fixation points as the ground truth and Settings 2, 3, and 4 use the human density map as the ground truth, with thresholds $T_d = 0$, 0.1, and 0.2, respectively. For each setting and each model, we try different blurring factors to the obtained saliency maps to achieve the best average AUC over Bruce's data.

1. In some previous works, AUCs are produced by directly using the fixation points as the ground truth (e.g., [48]), while in other works they are produced by using the human density map as the ground truth (e.g., [14]). In addition, when using the human density map as the ground truth, we first need to select a threshold to make it a binary map. We found that different thresholds T_d may have to be used for generating the binary map and producing AUCs reported in previous works.
2. Center bias has been known to be a serious issue in visual attention: The regions near the image center are more likely to be salient than the regions near the image perimeter [40], [41]. As listed in Table 1, some previous models take advantage of the center bias and some do not [48].

Table 1 summarizes the specific settings we found that can obtain the previously reported AUCs on Bruce's data for several comparison models, including the AIM model [4], [5], the DVA model [14], and the SUN model [48]. For the DVA

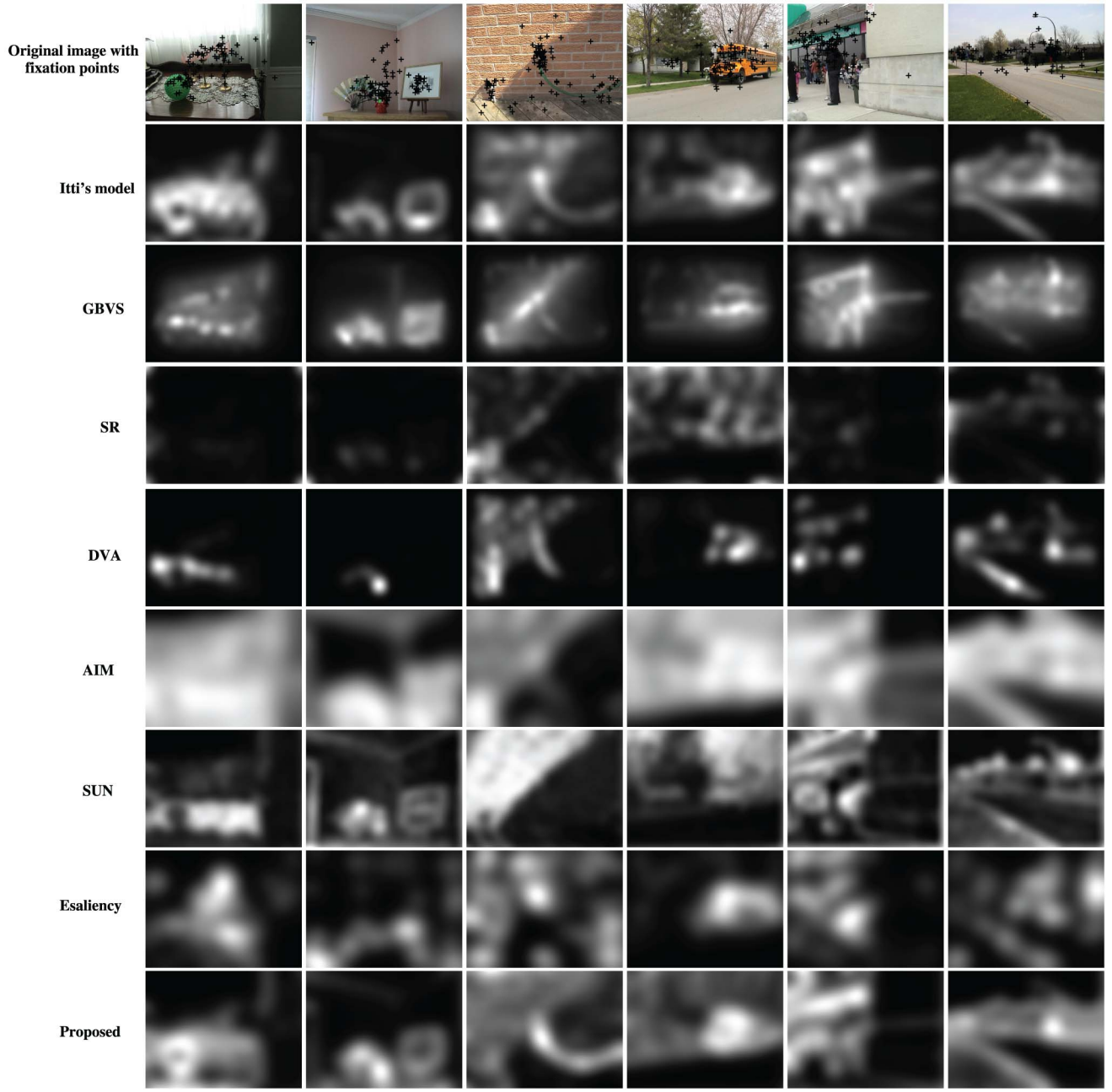


Fig. 11. Master saliency maps generated by the proposed model and the comparison models on a set of images. From the top row to the bottom row are the original image with fixation points (in dark crosses) and saliency maps produced by Itti's model [16], the GBVS model [12], the SR model [13], the DVA model [14], the AIM model [4], [5], the SUN model [48], the Esaliency model [2], and the proposed model, respectively.

model [14], the AUC reported in the previous work is obtained by using the human density map as the ground truth, but the threshold T_d used for generating this AUC is not given. We test different T_ds and find that $T_d = 0.15$ can lead to an AUC of 0.7911, which is very close to the one reported in the previous work [14]. For the AIM model [5], we could not achieve the previously reported AUC of 0.7810 by using fixation points as the ground truth. However, we find that by using the human density map as the ground truth with $T_d = 0.15$ we can get a very similar AUC of 0.7800 for the AIM model. We cannot get exactly the same AUCs reported in the previous works because of the possible different implementation details such as the density of the points for generating the ROC curve. In Table 1, we also give the AUCs of Itti's model under three different settings. We cannot tune the settings to get the AUC of Itti's model reported in [48] because we are using a different implementation, which implicitly introduces a center bias. For the SR model [13], the GBVS model [12], and the Esaliency model [2], there are no previously reported AUCs on Bruce's data. We simply use the fixation points as the ground truth and include the resulting AUCs in Table 1. In this table, we also report the AUC of the proposed model under two different settings: "Proposed Model" indicates the model as described above without any additional processing and "Proposed with Bias" indicates the altered proposed model where a center bias is introduced by multiplying the resulting master saliency map by a Gaussian kernel

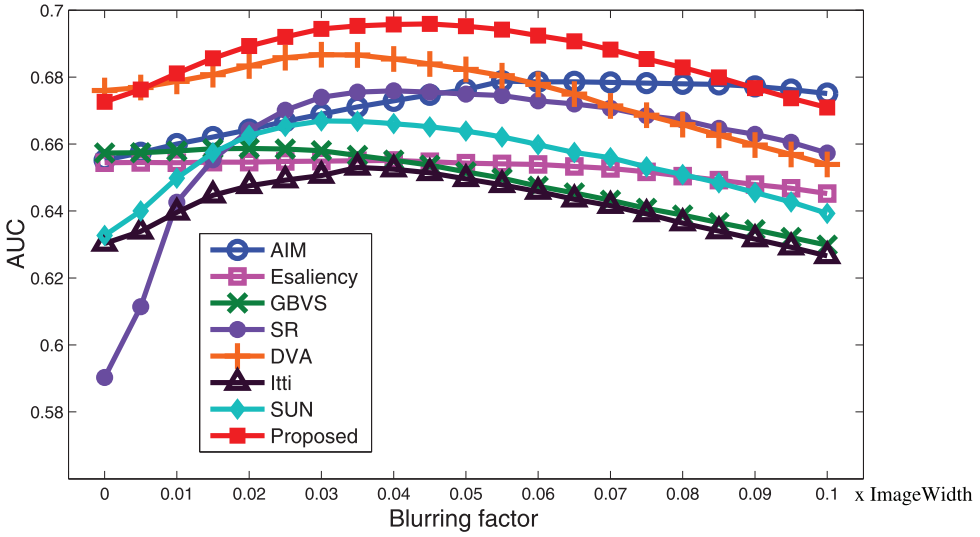


Fig. 12. The average AUCs over Bruce's data for each model when using different blurring factors to smooth the master saliency map. All these AUCs are obtained under unified Setting 1, i.e., directly using the fixation points as the ground truth and removing the center bias using an additional processing developed in [48].

$$G(x,y) = \exp\left(\frac{-(x-x_0)^2}{2\sigma_x^2} + \frac{-(y-y_0)^2}{2\sigma_y^2}\right),$$

where (x_0, y_0) is the coordinate of the image center and σ_x and σ_y are set as the one-fifth of the image width and height, respectively. From Table 1, we can clearly see that different evaluation settings can lead to completely different AUCs, even for the same model. For example, when using the human density map to evaluate Itti's model, different T_ds result in different AUCs. In addition, introducing a center bias (either explicitly or implicitly) can substantially change the resulting AUCs. For example, the AUC of the proposed model increases from 0.7403 to 0.8365 by introducing a center bias. Therefore, it is not meaningful to compare the performance of different models by examining their AUCs from different settings.

To conduct an apple-to-apple comparison, we propose a unified setting by directly using the fixation points as the ground truth and removing the center bias using an additional processing developed in [48]. In particular, when computing the AUC on one image we take the ground-truth fixation points on this image as positive samples and the ground-truth fixation points on all the other images, i.e., the remaining 119 images excluding the evaluated image in Bruce's data, as negative samples. In addition, in many previous works, the obtained master saliency maps are usually blurred by a Gaussian filter when they are evaluated against the ground-truth fixation points or human density maps. In Table 1, all the reported AUCs are obtained by using the default settings in respective software packages. Therefore, we used their default blurring factors, i.e., the

TABLE 3
AUCs of a Variety of Altered Models to Justify Individual Components of the Proposed Model

Models	AUC (blurring factor optimized for each model)	AUC (blurring factor optimized for Proposed Model)
Itti's Model	0.6530±0.0092	0.6514±0.0093
Nonlinear	0.6766±0.0094	0.6766±0.0094
KL + Linear	0.6848±0.0089	0.6848±0.0089
KL + Nonlinear	0.6926±0.0089	0.6919±0.0090
EMD	0.6888±0.0094	0.6888±0.0094
EMD+Nonlinear+MAX	0.6918±0.0095	0.6918±0.0095
EMD + Nonlinear without L^m-norm	0.6898±0.0093	0.6898±0.0093
Proposed Model	**0.6959±0.0091**	**0.6959±0.0091**

TABLE 4
AUCs of the Altered Models by Using Different Super Features

	Super Features	AUC (blurring factor optimized for each case)	AUC (blurring factor optimized for Proposed Model)
Case 1	{Color, Intensity}, {Orientation}	0.6916±0.0095	0.6916±0.0095
Case 2	{Intensity, Orientation}, {Color}	0.6842±0.0098	0.6842±0.0098
Case 3	{Color}, {Intensity}, {Orientation}	0.6857±0.0097	0.6852±0.0095
Case 4	{Color, Intensity, Orientation}	0.6862±0.0095	0.6862±0.0095
Case 5	{Color, Orientation}, {Intensity}	**0.6959±0.0091**	**0.6959±0.0091**
Case 6	{Orientation from Color}, {Intensity}	0.6877±0.0096	0.6857±0.0095

standard deviation of the Gaussian filter. By using a different blurring factor for smoothing the obtained saliency map, we may get a different AUC. To achieve a fairer comparison, for each compared model, we exhaustively try all possible blurring factors in the range of 0 to 10 percent of the image width, and pick the optimal blurring factor that leads to the best average AUCs over Bruce's data for this model. The column "Setting 1" of Table 2 shows such best average AUCs of the proposed model and other comparison models under this unified setting. Fig. 11 shows the master saliency maps generated by the proposed model and these comparison models under this setting. We can see that, on the first image, the proposed model can better recognize the saliency of the ball than most other models. On the third image, the proposed model can better recognize the saliency of the water pipe than most other models.

For comparison, the columns "Setting 2," "Setting 3," and "Setting 4" of this table show the best average AUCs of these models under other three unified settings where the center bias is removed but the human density map is used as the ground truth, with thresholds $T_d = 0$, 0.1, and 0.2, respectively. As in the previous works, we also include their standard errors. We can see that the AUCs usually increase with the increase of T_d. Therefore, we suggest the direct use of the unified Setting 1, i.e., using the fixation points as the ground truth, for performance evaluation. Fig. 12 shows the average AUCs over Bruce's data for each model, when using different blurring factors in Setting 1. We can see that, with the increase of the blurring factor, the average AUC increases initially but drops later for each model.

We also conduct experiments to justify each newly developed component of the proposed model under the above-mentioned unified setting, i.e., Setting 1 in Table 2. Starting from Itti's model, we construct a set of altered models, each of which only incorporates a subset of our newly developed components and the resulting AUCs are summarized in Table 3. In Table 3, "Nonlinear" indicates the altered model where we only use the proposed nonlinear feature combination but not the EMD-based center surround difference in Itti's model. "KL+Linear" indicates the altered model in which we use KL-divergence for computing the center-surround difference and the linear summation for feature combination. "KL+Nonlinear" indicates the altered model in which we use KL-divergence for computing the center-surround difference and the proposed nonlinear operations for feature combination. "EMD" indicates the altered model where we only use EMD for the center-surround difference, but not the proposed nonlinear feature combination. "EMD+Nonlinear+MAX" indicates the altered model where all the components are the same as the above proposed model except that the MAX operator instead of linear combination is used for combining salient maps from different scales. "EMD+Nonlinear without L^m-norm" indicates the altered model where all the components are the same as the above proposed model except that a linear summation instead of the L^m-norm is used for superfeature

TABLE 5
Performance of the Proposed Model and Three Comparison Models on Itti's Data

Model	Original Video Clips		MTV Video Clips	
	KL-Divergence	AUC	KL-Divergence	AUC
Variance [34]	0.120±0.004	0.624±0.002	0.087±0.003	0.605±0.002
Itti's Model [18]	0.179±0.006	0.663±0.003	0.143±0.005	0.647±0.003
Surprise [20], [21]	0.213±0.006	0.676±0.003	0.196±0.006	0.665±0.003
Proposed Model	**0.279±0.007**	**0.699±0.003**	**0.273±0.008**	**0.698±0.003**

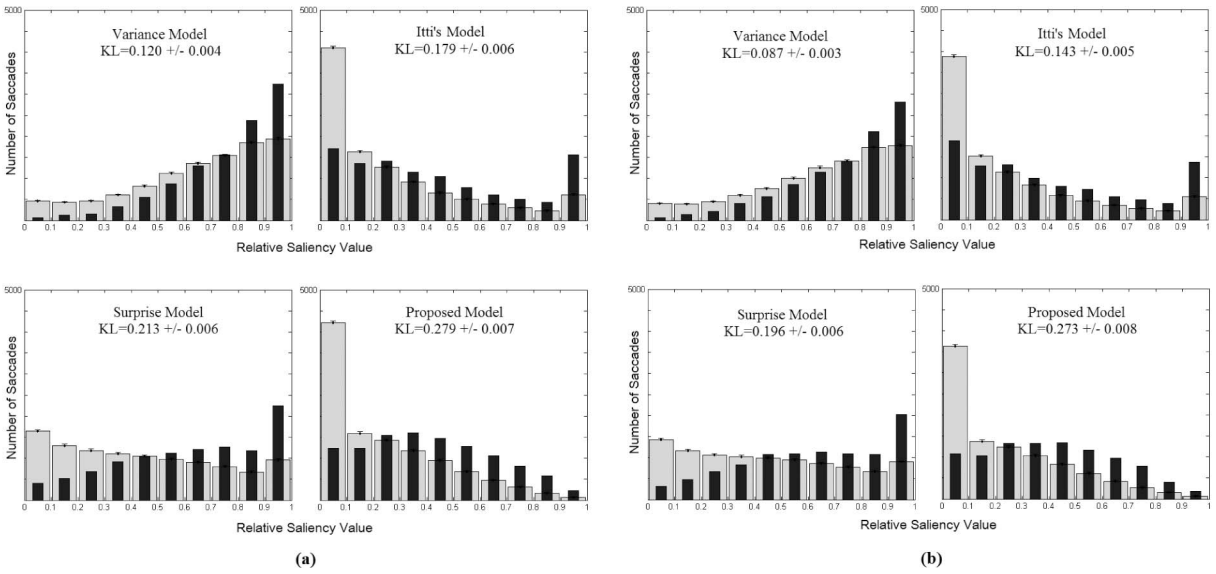


Fig. 13. Histograms constructed for evaluating the performance using the KL-divergence. Dark bins indicate the histogram for fixation points and gray bins indicate the histogram for randomly selected points. (a) Histograms constructed for the original video clips. (b) Histograms constructed for the MTV video clips. In both (a) and (b), four such histogram pairs are shown for the Variance model, Itti's model, the Surprise model, and the proposed model, respectively.

construction. For clarity, we also include the AUCs of Itti's model and the proposed model in this table. In Table 3, the column "AUC (blurring factor optimized for each model)" shows the best average AUC when each model or altered model uses its own optimal blurring factor. The column "AUC (blurring factor optimized for Proposed Model)" shows the average AUCs when all the models and altered models use a same blurring factor that leads to the best average AUC for the proposed model. By comparing to the performance of Itti's model, we can see that the introduction of each new component leads to an improved performance. By comparing to the performance of the proposed model, we can see that the integration of all these new components lead to a further improved performance. By comparing the performance of "EMD+Nonlinear+MAX" and the performance of the proposed model, we can see that the use of a MAX operator cannot produce a better performance than the linear summation for combining the saliency maps from difference scales.

Furthermore, we conduct experiments to justify the proposed biologically inspired super features. Specifically, we alter the proposed model by using different super features and evaluate the performance, also under the above-mentioned unified setting, i.e., Setting 1 in Table 2. Given three basic features of intensity, color, and orientation, there are, in total, five different cases to construct the super features, as shown by Case 1 through Case 5 in Table 4. Note that the super features shown in Case 5 are the ones we use for the proposed model. In Case 6, we construct the orientation feature from color instead of from intensity. More specifically, we apply the same Gabor filters along the four directions $\{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ to the two color feature maps and then take the average of filtering results along each direction to construct the orientation feature, which is then combined with the intensity feature by using a WTA mechanism. As before, the column "AUC (blurring factor optimized for each case)" shows the best average AUCs when each case uses its own optimal blurring factor. The column "AUC (blurring factor optimized for Proposed Model)" shows the average AUCs when all the cases use a same blurring factor that leads to the best average AUC for the proposed model. We can see that the use of the proposed super features produces a better performance than using the other super features. Note that in this experiment we always use the proposed nonlinear operations for feature combination.

In evaluating the proposed model on Itti's data, we use two measures. First, just like the evaluation on Bruce's data, we compute AUCs by using the fixation points as the ground truth. Second, by following [20], [21], we construct two histograms: one for the fixation points and the other for a set of randomly selected points, in terms of the saliency computed by a visual-attention model, and then calculate the KL-divergence between these two histograms as a performance measure. The larger the KL-divergence, the better the performance. Table 5 shows the performance of the proposed model and three comparison models, on 50 original clips and 50 MTV clips, respectively. For AUCs, we show the average and the standard error over the respective clips. For KL-divergence, by following [21] we repeat the random sampling 100 times to get the average and standard error. Fig. 13 shows the histogram pairs constructed for computing the KL divergence in this performance evaluation [21]. Note that here Itti's model for constructing the dynamic saliency maps uses the additional motion and flicker features [20], [21]. We can see that the proposed model produces better performance than the comparison models in terms of both AUC and the KL divergence. Fig. 14 shows the selected frames of the generated dynamic saliency maps. On the first video, we find that the proposed model can better recognize the high saliency of the two walking persons.

We run all our experiments in a Thinkpad laptop with a dual-core 2.70 GHz Intel i7-2620 CPU and 4.00 GB memory. The proposed model is implemented using Matlab. Average

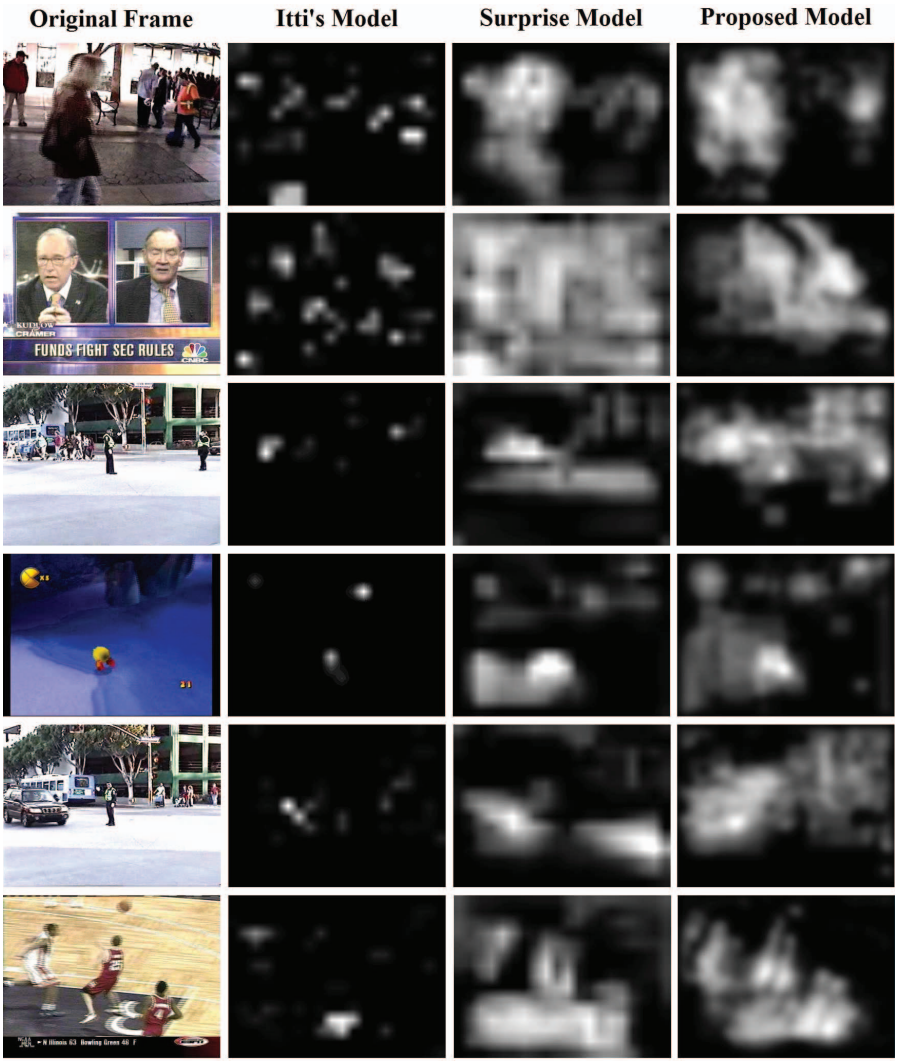


Fig. 14. Selected frames of the dynamic saliency maps generated by the proposed model and the comparison models. From the left column to the right column are the original frame and saliency maps produced by Itti's model [16], the Surprise model [20], [21], and the proposed model, respectively.

running time is 12.53 seconds for processing a static image in Bruce's data and 7.59 seconds per video frame for processing video clips in Itti's data. The most time-consuming step is the construction of the histograms for computing EMD-based center-surround difference. Since this histogram construction and the computing of the EMD-based center-surround difference are local operations, we expect they can be substantially speeded up by using a GPU implementation. It takes less time to process a video frame in Itti's data than process a static image in Bruce's data because, as mentioned above, we take coarser scales when processing videos.

6 Conclusion

In this paper, we developed a novel computational model for visual attention. We first used the weighted histograms and EMD for computing the center-surround difference. We also developed a two-step nonlinear operation to combine different basic features by following the findings in the neurobiology discipline. We finally extended this model to process videos for constructing dynamic saliency maps, where the major step is to use the weighted histograms and EMD for computing the spatiotemporal center-surround difference. For performance evaluation, we investigated different evaluation settings used in the previous works and described a unified setting that can provide fairer apple-to-apple comparison. We conducted experiments on both Bruce's dataset, which consists of static images, and Itti's dataset, which consists of video clips, and found that the proposed model produces a better performance than many existing models.

Acknowledgments

This work was supported, in part, by the National Natural Science Foundations of China (NSFC-61173130, NSFC-61173129, and NSFC-90820306), and the Natural Science Foundation of Chongqing, China (CSTC-2009AB5002 and CSTC-2010BB2217), and SRG010-FST11-TYY, MYRG187(Y1-L3)-FST11-TYY, and MYRG205(Y1-L4)-FST11-TYY. This work was also supported, in part, by the US National Science Foundation (NSF) (IIS-0951754 and IIS-1017199), the US Air Force Office of Scientific Research (FA9550-11-1-0327), and the US Army Research Laboratory under

Cooperative Agreement Number W911NF-10-2-0060 (US Defense Advanced Research Projects Agency (DARPA) Mind's Eye Program). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Army Research Laboratory or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein. Part of this work was conducted while Y. Lin was with Chongqing University. A preliminary version of this work has appeared in a conference proceeding [27].

REFERENCES

[1] E.A. Allen and R.D. Freeman, "Dynamic Spatial Processing Originates in Early Visual Pathways," *J. Neuroscience,* vol. 26, no. 45, pp. 11763-11774, 2006.

[2] T. Avraham and M. Lindenbaum, "Esaliency (Extended Saliency): Meaningful Attention Using Stochastic Image Modeling," *IEEE Trans. Pattern Analysis Machine and Intelligence,* vol. 32, no. 4, pp. 693-708, Apr. 2010.

[3] M.Z. Aziz and B. Mertsching, "Fast and Robust Generation of Feature Maps for Region-Based Visual Attention," *IEEE Trans. Image Processing,* vol. 17, no. 5, pp. 633-644, May 2008.

[4] N.D.B. Bruce and J.K. Tsotsos, "Saliency Based on Information Maximization," *Proc. Advances in Neural Information Processing Systems,* pp. 155-162, 2006.

[5] N.D.B. Bruce and J.K. Tsotsos, "Saliency, Attention, and Visual Search: An Information Theoretic Approach," *J. Vision,* vol. 9, no. 3, pp. 1-24, 2009.

[6] D. Cai, G.C. Deangelis, and R.D. Freeman, "Spatiotemporal Receptive Field Organization in the Lateral Geniculate Nucleus of Cats and Kittens," *J. Neurophysiology,* vol. 78, pp. 1045-1061, 1997.

[7] M. Corbetta and G.L. Shulman, "Control of Goal-Directed and Stimulus-Driven Attention in the Brain," *Nature Rev. Neuroscience,* vol. 3, no. 3, pp. 201-215, 2002.

[8] G.C. DeAngelis, I. Ohzawa, and R.D. Freeman, "Receptive-Field Dynamics in the Central Visual Pathways," *Trends in Neurosciences,* vol. 18, no. 10, pp. 451-458, 1995.

[9] B.A. Draper and A. Lionelle, "Evaluation of Selective Attention under Similarity Transformations," *Computer Vision and Image Understanding,* vol. 100, nos. 1/2, pp. 152-171, 2005.

[10] S. Frintrop, E. Rome, and H.I. Christensen, "Computational Visual Attention Systems and Their Cognitive Foundations: A Survey," *ACM Trans. Applied Perception,* vol. 7, no. 1, article 6, 2010.

[11] D. Gao and N. Vasconcelos, "Bottom-Up Saliency Is a Discriminant Process," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 185-190, 2007.

[12] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *Proc. Neural Information Processing Systems,* 2006.

[13] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2007.

[14] X. Hou and L. Zhang, "Dynamic Visual Attention: Searching for Coding Length Increments," *Proc. Advances in Neural Information Processing Systems,* pp. 681-688, 2008.

[15] D.H. Hubel and T.N. Wiesel, "Receptive-Field Dynamics in the Central Visual Pathways," *J. Physiology,* vol. 160, no. 1, pp. 106-154, 1962.

[16] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis Machine and Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[17] L. Itti and C. Koch, "A Saliency Based Search Mechanism for Overt and Covert Shifts of Visual Attention," *Vision Research,* vol. 40, nos. 10-12, pp. 1489-1506, 2000.

[18] L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nature Rev. Neuroscience,* vol. 2, no. 3, pp. 194-203, 2001.

[19] L. Itti, "Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention," *IEEE Trans. Image Processing,* vol. 13, no. 10, pp. 1304-1318, Oct. 2004.

[20] L. Itti and P. Baldi, "Bayesian Surprise Attracts Human Attention," *Proc. Advances in Neural Information Processing Systems,* vol. 18, pp. 547-554, 2006.

[21] L. Itti and P. Baldi, "Bayesian Surprise Attracts Human Attention," *Vision Research,* vol. 49, no. 10, pp. 1295-1306, 2009.

[22] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," *Proc. 12th IEEE Int'l Conf. Computer Vision,* 2009.

[23] E.I. Knudsen, "Fundamental Components of Attention," *Ann. Rev. of Neuroscience,* vol. 30, pp. 57-78, 2007.

[24] A.R. Koene and Z. Li, "Feature-Specific Interactions in Salience from Combined Feature Contrasts: Evidence for a Bottom-Up Saliency Map in V1," *J. Vision,* vol. 7, no. 7, pp. 1-14, 2007.

[25] E. Levina and P. Bickel, "The Earth Mover's Distance Is the Mallows Distance: Some Insights from Statistics," *Proc. Eighth IEEE Int'l Conf. Computer Vision,* vol. 2, pp. 251-256, 2001.

[26] Z. Li, "A Saliency Map in Primary Visual Cortex," *Trends in Cognition Science,* vol. 6, no. 1, pp. 9-16, 2002.

[27] Y. Lin, B. Fang, and Y.Y. Tang, "A Computational Model for Saliency Maps by Using Local Entropy," *Proc. 24th AAAI Conf. Artificial Intelligence,* pp. 967-973, 2010.

[28] H. Ling and K. Okada, "An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 5, pp. 840-853, May 2007.

[29] 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 Analysis and Machine Intelligence,* vol. 33, no. 2, pp. 353-367, Feb. 2011.

[30] S.P. Liversedge and J.M. Findlay, "Saccadic Eye Movements and Cognition," *Trends in Cognition Science,* vol. 4, no. 1, pp. 6-14, 2000.

[31] D. Parkhurst, K. Law, and E. Niebur, "Modeling the Role of Salience in the Allocation of Overt Visual Attention," *Vision Research,* vol. 42, pp. 107-123, 2002.

[32] F. Poirier, F. Gosselin, and M. Arguin, "Perceptive Fields of Saliency," *J. Vision,* vol. 8, no. 15, pp. 1-19, 2008.

[33] U. Rajashekar, I. van der Linde, A.C. Bovik, and L.K. Cormack, "GAFFE: A Gaze-Attentive Fixation Finding Engine," *IEEE Trans. Image Processing,* vol. 17, no. 4, pp. 564-573, Apr. 2008.

[34] P. Reinagel and A.M. Zador, "Natural Scene Statistics at the Centre of Gaze," *Network,* vol. 10, pp. 341-350, 1999.

[35] M. Riesenhuber and T. Poggio, "Hierarchical Models of Object Recognition in Cortex," *Nature Neuroscience,* vol. 2, no. 11, pp. 1019-1025, Nov. 1999.

[36] Y. Rubner, C. Tomasi, and L.J. Guibas, "A Metric for Distributions with Applications to Image Databases," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 59-66, Jan. 1998.

[37] Y. Rubner, C. Tomasi, and L.J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Int'l J. Computer Vision,* vol. 40, no. 2, pp. 99-121, 2000.

[38] F. Shic and B. Scassellati, "A Behavioral Analysis of Computational Models of Visual Attention," *Int'l J. Computer Vision,* vol. 73, no. 2, pp. 159-177, 2007.

[39] R.N. Shepard, "Attention and the Metric Structure of Stimulus Space," *J. Math. Psychology,* vol. 1, pp. 54-87, 1964.

[40] B. Tatler, R. Baddeley, and I. Gilchrist, "Visual Correlates of Fixation Selection: Effects of Scale and Time," *Vision Research,* vol. 45, no. 5, pp. 643-659, 2005.

[41] B.W. Tatler, "The Central Fixation Bias in Scene Viewing: Selecting an Optimal Viewing Position Independently of Motor Biases and Image Feature Distributions," *J. Vision,* vol. 7, no. 4, pp. 1-17, 2007.

[42] M. To, P.G. Lovell, T. Troscianko, and D.J. Tolhurst, "Summation of Perceptual Cues in Natural Visual Scenes," *Proc. Royal Soc. B,* vol. 275, no. 1649, pp. 2299-2308, 2008.

[43] A. Torralba, A. Oliva, M.S. Castelhano, and J.M. Henderson, "Contextual Guidance of Eye Movements and Attention in Real-World Scenes: The Role of Global Features on Object Search," *Psychology Rev.,* vol. 113, no. 4, pp. 766-786, Oct. 2006.

[44] S. Treue, "Visual Attention: The Where, What, How and Why of Saliency," *Current Opinion Neurobiology,* vol. 13, no. 4, pp. 428-432, Aug. 2003.

[45] J.K. Tsotsos, "Analyzing Vision at the Complexity Level," *Behavioral and Brain Sciences* vol. 13, no. 3, pp. 423-445, 1990.

[46] D. Walther and C. Koch, "Modeling Attention to Salient Proto-Objects," *Neural Networks,* vol. 19, no. 9, pp. 1395-1407, 2006.

[47] J.M. Wolfe and T.S. Horowitz, "What Attributes Guide the Deployment of Visual Attention and How Do They Do It?" *Nature Rev. Neuroscience*, vol. 5, no. 6, pp. 495-501, June 2004.
[48] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," *J. Vision*, vol. 8, no. 7, pp. 1-20, 2008.

Yuewei Lin received the BS degree in optical information science and technology from Sichuan University, Chengdu, China, and the ME degree in optical engineering from Chongqing University, Chongqing, China. He is currently working toward the PhD degree in the Department of Computer Science and Engineering at the University of South Carolina. His current research interests include computer vision and image/video processing. He is a student member of the IEEE.

Yuan Yan Tang received the BS degree in electrical and computer engineering from Chongqing University, Chongqing, China, the MS degree in electrical engineering from the Beijing University of Post and Telecommunications, Beijing, China, and the PhD degree in computer science from Concordia University, Montreal, Quebec, Canada. He is a chair professor in the Faculty of Science and Technology at the University of Macau (UM). Before joining UM, he served as a chair professor in the Department of Computer Science at Hong Kong Baptist University and dean of the College of Computer Science at Chongqing University, China. He is a chair of the Technical Committee on Pattern Recognition of the IEEE Systems, Man, and Cybernetics Society (IEEE SMC) for his great contributions to wavelet analysis, pattern recognition, and document analysis. Recently, he was elected as one of the executive directors of the Chinese Association of Automation Council. With all his distinguished achievement, he is also the founder and editor-in-chief of the *International Journal of Wavelets, Multi-resolution, and Information Processing (IJWMIP)*, and an associate editor of the *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, and the *International Journal on Frontiers of Computer Science (IJFCS)*. He has been presented with numerous awards such as the First Class of Natural Science Award of Technology Development Centre, Ministry of Education of the People's Republic of China in November 2005 and the Outstanding Contribution Award by the IEEE Systems, Man, and Cybernetics Society in 2007. He has published more than 300 papers, books, and book chapters. He is a fellow of the IEEE and of the Pattern Recognition Society (IAPR).

Bin Fang received the BS degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, the MS degree in electrical engineering from Sichuan University, Chengdu, China, and the PhD degree in electrical engineering from the University of Hong Kong, Hong Kong. He is currently a professor with the College of Computer Science, Chongqing University, Chongqing, China. His research interests include computer vision, pattern recognition, information processing, biometrics applications, and document analysis. He has published more than 120 technical papers and is an associate editor of the *International Journal of Pattern Recognition* and *Artificial Intelligence*. He has been the program chair and a committee member for many international conferences. He is a senior member of the IEEE.

Zhaowei Shang received the BS degree in computer science from the Northwest Normal University, Lanzhou, China, in 1991, the MS degree from the Northwest Polytechnical University, Xi'an, China, in 1999, and the PhD degree in computer engineering form Xi'an Jiaotong University, Xi'an, in 2005. He is currently an associate professor with the Department of Computer Science, Chongqing University, Chongqing, China. His research interests include computer vision, pattern recognition, image processing, and wavelet analysis.

Yonghui Huang received the BE degree in computer engineering from Chongqing University, Chongqing, China, in 2009. He is currently working toward the master's degree in the Department of Computer Science, Chongqing University. His research interests include machine learning, pattern recognition, and image annotation.

Song Wang received the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC) in 2002. From 1998 to 2002, he also worked as a research assistant in the Image Formation and Processing Group at the Beckman Institute of UIUC. In 2002, he joined the Department of Computer Science and Engineering at the University of South Carolina, where he is currently an associate professor. His research interests include computer vision, medical image processing, and machine learning. He is currently serving as the publicity/web portal chair of the Technical Committee on Pattern Analysis and Machine Intelligence of the IEEE Computer Society, and as an associate editor of *Pattern Recognition Letters*. He is a senior member of the IEEE and a member of the IEEE Computer Society.