Predicting Eye Fixations With Higher-Level Visual Features

Ming Liang, *Student Member, IEEE*, and Xiaolin Hu, *Senior Member, IEEE*

***Abstract*—Saliency map and object map are the two contrasting hypotheses for the mechanisms utilized by the visual system to guide eye fixations when humans are freely viewing natural images. Most computational studies define saliency as outliers of distributions of low-level features, and propose saliency as an important factor for predicting eye fixations. Psychophysical studies, however, suggest that high-level objects predict eye fixations more accurately and the early saliency only has a minor effect. But this view has been challenged by a study which shows opposite results, suggesting that the role of object-level features needs further investigations. In addition, little is known about the role of intermediate features between the low-level and the object-level features. In this paper, we construct two models based on mid-level and object-level features, respectively, and compare their performances against those based on low-level features. Quantitative evaluation on three benchmark natural image fixation data sets demonstrates that the mid-level model outperforms the state-of-the-art low-level models by a significant margin and the object-level model is inferior to most low-level models. Quantitative evaluation on a video fixation data set demonstrates that both the mid-level and object-level models outperform the state-of-the-art low-level models, and the latter performs better under three out of four standard metrics. When combined together the two proposed models achieve even higher performance. However, incorporating the best low-level model yields negligible improvements on all of the data sets. Taken together, these results indicate that higher level features may be more effective than low-level features for predicting eye fixations on natural images in the free viewing condition.**

***Index Terms*—Attention, saliency, feature, visual hierarchy.**

I. Introduction

WHEN our eyes are viewing natural scenes, the fixations are directed by attention to facilitate the processing of visual information. Attention can be driven by both bottom-up stimulus properties and top-down task dependent factors, which has received extensive investigations from psychology and neuroscience. Besides the direction of information flow, another important issue about attention is the representation of inputs, *i.e.* visual features, based on which attentional signals are computed. It is still unresolved what levels of features are involved in attention computation, and to what extent each of them can guide the attention, respectively. Two contrasting hypotheses on attention under the condition of free viewing are (low-level) saliency map and (high-level) object-based attention, which are briefly reviewed in what follows.

Most computational models follow the saliency map hypothesis [1], which defines saliency as outliers of the distribution of visual features on the image. Hence, the locations with distinct or rare features are assigned high saliency values and supposed to strongly attract attention. Consistent with psychological findings [2], [3], saliency models mostly operate on low-level feature channels [1], [4]–[6], such as orientation, color and intensity. These features resemble the stimuli to which neurons in early visual areas are sensitive. Psychophysical [7] and physiological [8] evidence suggests that the saliency map might be represented by neural responses in V1.

In contrast, the object-based hypothesis claims that objects may better predict fixations and directly guide attention [9], [10]. The major argument is that although saliency is shown to be correlated with fixations, it has only an indirect effect through the salient recognized objects. In a psychophysical experiment [9], an object map was obtained by weighting each pixel by its overlap counts with the recalled objects by human subjects. It was compared with the saliency map computed by Itti *et al.*'s model [1] in terms of the ability to predict eye fixations when the subjects are freely viewing daily life images. The object map outperformed the saliency map, and when the two maps were combined the performance was not significantly better than that of the object map alone. However, recently an analysis of the same data yielded conflicting results [11]. After appropriately addressing the spatial bias intrinsic in fixations, the object map did not perform better than Itti *et al.*'s model. In addition, the object map was outperformed by the state-of-the-art saliency models.

The visual system has a hierarchical organization, and it creates increasingly complex and invariant multi-level feature representations [12]–[14]. Previous experiments have indicated multiple candidate sites on cortex for attention computation including V1 [7], [8] and V4 [15], [16], which process

Manuscript received July 7, 2014; revised November 20, 2014; accepted January 8, 2015. Date of publication January 23, 2015; date of current version February 11, 2015. This work was supported in part by the National Basic Research Program (973 Program) of China under Grant 2013CB329403, in part by the National Natural Science Foundation of China under Grant 61273023, Grant 91120011, Grant 61332007, and Grant 91420201, and in part by the Tsinghua University Initiative Scientific Research Program under Grant 20121088071. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dimitrios Tzovaras.

M. Liang is with the State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, and also with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: liangm07@mails.tsinghua.edu.cn).

X. Hu is with the State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with the Center for Brain-Inspired Computing Research, Tsinghua University, Beijing 100084, China (e-mail: xlhu@tsinghua.edu.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/TIP.2015.2395713

1057-7149 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

different levels of features. A natural question is to what extent each feature level contributes to attention. In this study, we explored this issue by constructing two attention models based on higher-level visual features (higher than the low-level features used in typical saliency models [1]–[6]). The motivations for proposing these models are described as follows.

Low-level features and object-level features (throughout the paper we use "object-level" and "high-level" interchangeably) are two extremes of the hierarchical representations of visual information [17]. Neither hypothesis discussed above considers the effect of intermediate feature representations on attention, which we call *mid-level* features in this paper. For computational implementation of mid-level features, we follow the common definition in computer vision, that is, features built on low-level ones, having more semantic information but without direct description on high-level image structures. Unlike the simple low-level features which are universal building elements for all images, mid-level features are suggested to be more informative for object recognition [18]. This advantage attributes to their stronger ability for discriminating visual patterns [19], including selectivity to perceptually different patterns and invariance to perceptually irrelevant variations. Saliency by nature measures the differences between visual patterns. We hypothesize that the differences measured on mid-level features may better match the human perception of saliency than low-level features on natural images. A feature selection study [20] for saliency prediction supports this hypothesis but a systematic investigation is lacked. To further validate this hypothesis, we propose to use some mid-level features effective for object recognition to predict saliency in the conventional bottom-up saliency map framework. One should be aware that the saliency map is originally defined with respect to low-level features, but this notion is extended to mid-level features in this paper.

The object map in [9] was obtained based on the behavioral data of human subjects. There do exist some computational models that employ object-level features. A deep learning model [22] has been trained to detect saliency where the top layer features were used. These features were shown to correspond to semantic concepts. Another approach is to train a weighted sum of object detectors [23]–[25], but these detectors are not necessarily obtained in a hierarchical fashion. Our second model adopts this approach, with the motivation that there may be much space for improvement by incorporating more objects. Specifically, we train an object-level attention model using the high-level object bank (OB) [26] feature, which has 177 kinds of object detectors.

We did not attempt to devise new low-level saliency models by considering that through years of efforts many excellent models of this kind have been developed. These models were used as baselines for evaluating the two proposed higher-level models. We compared their performances in natural viewing conditions, and evaluated the roles of different feature levels.

The remaining content is organized as follows. Section II presents some related works. Section III and section IV describe the mid-level and high-level models, respectively. Section V presents the experimental results. Finally, section VI gives the discussion.

II. Related Work

Based on the feature integration theory [2] Koch and Ullman [27] first proposed the concept of saliency map, which was implemented by Itti *et al.* in a biologically plausible way [1]. Since then, many models have been proposed, and most of them are bottom-up models which do not need supervised training. They typically work in three steps. First, a set of feature maps over several channels are extracted from each location of the image. Second, saliency signals are activated on each feature map according to certain measures. Third, these signals are combined to form a master saliency map.

For the first step, low-level features are used by almost all saliency models. Gabor-filtered orientation, opponent color and intensity are used by Itti *et al.*'s model [1]. Similar features are used by GBVS [4] and AWS [6]. ICA bases which resemble the Gabor filters are used by AIM [5] and SUN [28]. Raw image patches are used in context aware saliency (CA) [29] and image manipulation saliency [30]. For models operating in the frequency domain [31], [32], the features are Fourier or discrete cosine transformation coefficients. Besides these linear features nonlinear features have also been used by some saliency models. Local steering kernels (LSKs) are used by SDSR [33]. The covariance matrix of a set of features, such as orientation, color, intensity, is used by covariance saliency [34]. In a recent feature selection study [20], some features including histograms of oriented gradients (HOG) [35] and histograms of colors are shown to be complementary to the low-level features for saliency prediction.

For the second step, many approaches have been proposed which operate locally or globally on the feature maps. For local measures, the difference between two locations is weighted by their spatial distance. In Itti *et al.*'s model the consipicuity in each feature channel is calculated as the center surround contrast. GBVS [4] treats the image as a graph and the weight between two nodes is determined by their dissimilarity and distance. The saliency is then the equilibrium distribution of the Markov chain defined over the graph. To measure a multi-scale saliency, CAS [29] compares an image patch with other patches in both the same scale and neighboring scales. For global measures, the interaction between locations are independent of their spatial distance. AIM [5] employs an information theoretical approach and defines its saliency as the self-information. Inspired by the properties of neural responses, AWS [6] decorrelates the multi-scale feature vectors with PCA and uses their normalized amplitude to represent saliency. Another kind of global saliency measure is derived in the frequency domain, such as spectral residue saliency (SR) [31] and image signature (IS) [32].

For the third step, following the classical model of Itti *et al.* [1] many models [4]–[6] simply sum up the saliency on all feature maps. The weights are usually manually assigned for each feature channel. Besides linear combination, multiplication [34] and max [36] operations have also been used to integrate saliency information. Recently, a hierarchical structure was proposed to generate saliency maps at different scales and a Markovian method was used to integrate the saliency maps sequentially [37].

Different from the standard saliency framework, some attention models skip the second step. They extract and combine higher-level features (sometimes together with low-level saliency) to yield the attention map. Object-level features are employed by some models, in the form of object detectors [22], [24], [25]. These object-level features are not used for activating saliency, but are linearly combined in a supervised fashion to predict fixations. Recently, a set of neural networks with different number of layers were trained independently [38], and their top layer features were used to predict fixations. Then the networks were linearly combined. Some features in the model can be regarded as mid-level and some can be regarded as high-level, although they come from different networks.

This paper focuses on the first step, more specifically, the roles of mid-level and high-level visual features in attention modeling.

III. Mid-Level Attention Model

In this section, we describe the mid-level model. It has two feature channels, shape and color. The overall architecture is similar to traditional saliency models and the main difference lies in the extracted features, as described below.

A. Shape Channel

In image classification, the bag of words (BoW) [39] framework based on SIFT [21] descriptors is often used to obtain a mid-level visual representation. BoW assign the extracted descriptors to pre-trained visual words which correspond to typical local image structures. By pooling the visual words over the entire image, a histogram is obtained as the final representation. We use a modified BoW representation to obtain the shape features, in which a histogram is pooled from each local region.

The process of extracting a SIFT descriptor is briefly described as follows. Details can be found in [21]. First, the gradient orientation and magnitude at each sample point within a patch are computed. This step is similar to the Gabor filtering step in Itti *et al.*'s model [1]. Second, the patch is divided into 16 bins and the gradients in each bin are pooled into a histogram with 8 orientations. This operation introduces certain position invariance, which is in principle similar to the tuning property of complex cells. For computing the histogram, each gradient is weighted by a Gaussian window located at the descriptor center to avoid abrupt changes caused by small position shift, and the gradient is smoothly distributed to adjacent orientations. Third, the 16 histograms are concatenated to a vector of 128 dimensions. To be robust to illumination changes, thresholding and normalization operations are applied to the vector to yield the final descriptor. Figure 1 shows a SIFT descriptor located at the center of the big red square patch, where the 16 small squares are bins.

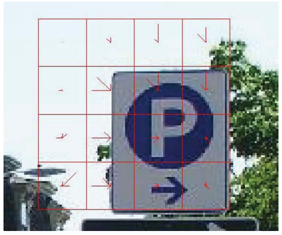

Fig. 1. A SIFT descriptor example. Each small red square is a bin, whose gradients are pooled into a histogram. The red lines in each bin denote the orientation information with line length denoting the magnitude. Best viewed in color.

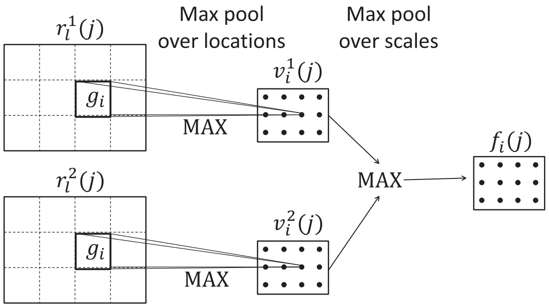

Fig. 2. Illustration of the pooling step in the shape channel. Two scales are shown.

A dictionary of visual words is trained for the BoW representation. A large set of SIFT descriptors are randomly extracted from an image dataset. The K-means clustering algorithm is then applied to the descriptors to learn K centroids of words $\mathbf{c}_j$ $(j = 1, \ldots, K)$, where each word $\mathbf{c}_j$ is a 128D vector. The dictionary is denoted by a matrix $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K)$.

For an input image, SIFT descriptors $\{\mathbf{d}_i\}$ are densely extracted from locations $\{i\}$. Each $\mathbf{d}_i$ is encoded by $\mathbf{C}$ and represented by $\mathbf{r}_i$ via locality-constrained linear coding (LLC) [40], which amounts to solving the following optimization problem:

$$\min_{\tilde{\mathbf{r}}_i} ||\tilde{\mathbf{C}}_i \tilde{\mathbf{r}}_i - \mathbf{d}_i||_2^2 \tag{1}$$

with the constraint $\mathbf{1}^T \tilde{\mathbf{r}}_i = 1$. Here $\tilde{\mathbf{C}}_i$ denotes the local dictionary formed by $\tilde{K}$ nearest neighbors (measured by Euclidean distance) of $\mathbf{d}_i$ among the K words. $\tilde{K} < 128$ is usually small. The local coordinate $\tilde{\mathbf{r}}_i$ is then converted to $\mathbf{r}_i$ by padding zeros in $\mathbf{r}_i$, which correspond to the $K - \tilde{K}$ unused visual words. LLC leads to a sparse code by using a small number of neighboring visual words for the reconstruction of each descriptor.

The image is then divided into non-overlapping grids $\{g_i\}$, where g_i denotes the grid cell centered at i. The visual words within each grid cell are max pooled to obtain a local region feature vector $\mathbf{v}_i$:

$$v_i(j) = \max_{l \in g_i} r_l(j) \tag{2}$$

where $v_i(j)$ is the jth feature of $\mathbf{v}_i$ and $r_l(j)$ is the jth element of $\mathbf{r}_l$ (see Figure 2).

The procedure described above deals with a single scale. To incorporate multi-scale information, several SIFT bin sizes are used. A single dictionary is trained for all scales. The SIFT extraction, LLC coding and max pooling steps are performed for each scale, followed by an extra pooling step over the scales (see Figure 2):

$$f_i(j) = \max_b v_i^b(j) \tag{3}$$

where b denotes the bin size.

To activate saliency, AWS [6] uses an approach by first removing the correlation between different features $f_i(j)$ and $f_i(k)$ and then normalizing each feature over the entire image. This approach can be adopted here. As LLC has led to a sparse code, in which the correlation between different features is largely removed, we directly normalize $f_i(j)$ to its standard score

$$\overline{f}_i(j) = (f_i(j) - \mu_j)/\sigma_j \tag{4}$$

where μ_j and σ_j denote the mean and standard deviation of $f_i(j)$ at all locations, respectively.

The saliency value at location i is defined as

$$s_i^{shape} = ||\overline{\mathbf{f}}_i||_2^2 \tag{5}$$

where $\overline{\mathbf{f}}_i = (\overline{f}_i(1), \overline{f}_i(2), \ldots, \overline{f}_i(K))^T$. The saliency map is then $S^{shape} = \{s_i^{shape}\}$.

B. Color Channel

Itti *et al.*'s model [1] uses opponent colors to encode color information, which correspond to the tuning properties of the V1 and V2 neurons. In the V4 area the neurons are tuned to hue while invariant to luminance changes [41]. We use color name features [42] to represent color information, as it is similar to higher-level color encoding in the cortex.

Color names map the raw RGB value to a 11D vector, each element denoting the probability of this pixel belonging to a certain basic color. The 11 basic color terms are defined according to their consistent use in English language and consensus among most speakers [43], including black, blue, brown, grey, green, orange, pink, purple, red, white and yellow. This mapping is learned from images on Internet [42]. For example, the red color training samples are obtained by typing 'red' and 'color' on Google, and cropping the red part of the returned images. Thus the training samples have some invariance to photometric changes.

The extraction of color name features involves two steps [44]. First, the RGB value of each pixel i is mapped to a 11D vector cn_i. Second, the image is divided into non-overlapping squared bins $\{P_i\}$ with the same size, with P_i centered at i. The color name vector $\mathbf{CN}_i$ is obtained by averaging all cn_l within P_i:

$$\mathbf{CN}_i = \sum_{l \in P_i} cn_l / N \tag{6}$$

where N denotes the number of pixels in each bin.

Inspired by SIFT, we concatenate the vectors of 4 × 4 neighboring bins to yield a 176D color feature $\mathbf{f}_i = (f_i(1), f_i(2), \ldots, f_i(176))^T$, where i denotes the center location of the 16 bins. This step enables the feature to represent more complex patterns.

To activate saliency, we adopt the same approach as that used in AWS [6]. First, principal component analysis (PCA) is used to decorrelate the color features

$$\mathbf{x}_i = \mathbf{U}^T \mathbf{f}_i \tag{7}$$

where the columns of $\mathbf{U}$ denote the principal components of the feature set over the input image. Then the saliency map $\mathbf{S}^{color} = \{s_i^{color}\}$ is computed based on the standard score of $\mathbf{x}_i$

$$\overline{x}_i(j) = (x_i(j) - \mu_j)/\sigma_j \tag{8}$$

where μ_j and σ_j denote the mean and standard deviation of $x_i(j)$ at all locations, respectively.

$$s_i^{color} = ||\overline{\mathbf{x}}_i||_2^2 \tag{9}$$

where $\overline{\mathbf{x}}_i = (\overline{x}_i(1), \overline{x}_i(2), \ldots, \overline{x}_i(176))^T$.

We tried the LLC technique to encode the color name features as in the shape channel but the result was not as good as that of PCA. A possible reason is that the distribution of color visual words is not sparse.

So far we have described the generation of color saliency map in a single scale. The extension to multiple scales is straightforward. We simply vary the bin size in equation (6) and repeat the feature extraction and the saliency activation steps, then sum up the resulting maps.

C. Channels Combination

The master map $S^{mid} = \{s_i^{mid}\}$ is obtained by normalizing and summing up S^{shape} and S^{color}:

$$S^{mid} = G * (\overline{S}^{shape} + \overline{S}^{color}) \tag{10}$$

where $\overline{S}^{shape}$ and $\overline{S}^{color}$ are normalized saliency maps with zero mean and unit standard deviation over the whole dataset. G is a Gaussian filter with its standard deviation being 0.04 times the longer dimension of S^{mid}, and $*$ denotes convolution. This setting is similar to that in many previous models [4], [32].

For convenience, this mid-level model is called histogram-based saliency (HBS) in what follows, as both shape and color channels involve histogram-based features.

IV. Object-Level Attention Model

A high-level object model is trained based on the object bank (OB) [26] feature. OB uses the responses of 177 object detectors [45], [46] to describe an image. These detectors operates in 12 scales. The detector response at location i and scale b is denoted as $\mathbf{r}_i^b$, which is a 177D vector. The training process is described below.

First, the multi-scale responses at each location i are max pooled over scales to obtain the feature map

$$f_i(j) = \max_b r_i^b(j) \tag{11}$$

where $f_i(j)$ and $r_i^b(j)$ denote the jth elements of $\mathbf{f}_i$ and $\mathbf{r}_i^b$, respectively.

Second, the attention map $\mathbf{S} = \{s_i\}$ is computed as

$$s_i = \mathbf{w}^T \mathbf{f}_i \tag{12}$$

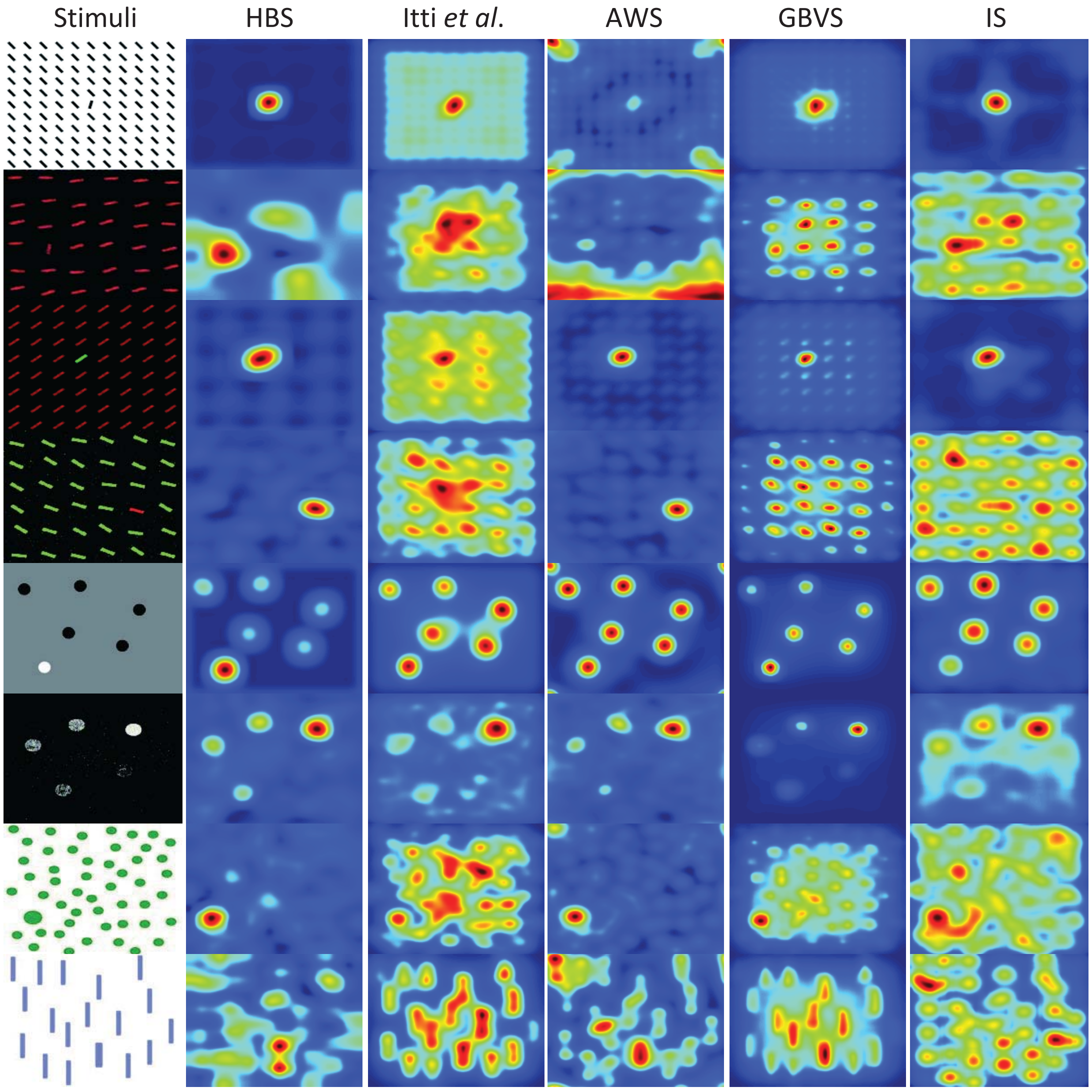

Fig. 3. Simple stimuli with the target pops out in orientation (row 1, 2), color (row 3, 4), intensity contrast (row 5, 6) and size (row 7, 8). From left to right, the columns denote the original image, saliency maps output by HBS, Itti *et al.*'s model, AWS, GBVS and IS. Best viewed in color.

where each element of $\mathbf{w}$ denotes the weight reflecting the priority of each object category during free viewing. To train $\mathbf{w}$, an support vector machine (SVM) based approach [24] is adopted. Images in a natural viewing dataset are split into a training set and a test set. The smoothed fixation maps are used as the ground truth. A set of pixels are randomly picked from the training images, with positive samples having high saliency and labels 1 while negative samples having low saliency and labels -1. For each sample the corresponding feature vector $\mathbf{f}$ is extracted. The features and labels are input into a linear SVM classifier to obtain $\mathbf{w}$, which are then used in the test phase. Because OB contains a large variety of generic objects and many of them may have little correlation with the fixation prediction task, l_1-norm SVM is used to select important features:

$$\min_{\mathbf{w}} \|\mathbf{w}\|_1 + c \sum_i \max\left(0, 1 - y_i \mathbf{w}^T \mathbf{f}_i\right)^2 \tag{13}$$

where c is a hyper-parameter. The l_1-norm penalty can drive the weights of useless features to zero [47].

For convenience, this attention model is called OB in what follows.

V. Experiments

A. Parameter Settings

The VLFeat [48] library was used to extract SIFT descriptors. Four bin sizes $\{4, 6, 8, 10\}$ were used for both the shape and color channels, the same as the default settings for PHOW descriptors in VLFeat. To inhibit the responses in regions with very low contrast, any SIFT descriptor with its l_2-norm smaller than $20/b$ was thresholded to be zero where b is the side length of the bin. This is because SIFT descriptors with larger bins tend to have lower contrast. The visual word number K was 1024 (larger K has led to similar or slightly better results on different datasets but significantly increased the computing time). We found that the performance of the models were not sensitive with K. Five nearest neighbors were used in LLC (equation (1)), which is the default setting in the reference [40].

In equation (13) c was set to 0.01. 10-fold cross validation over the dataset was carried out. In each trial, the weights for the images in the test fold were learned on the training folds. After cross validation, every image was assigned a set of test weights.

B. Synthetic Patterns

To gain some heuristic insight on the difference between the mid-level and low-level features, we qualitatively evaluated HBS over 26 synthetic images, which were collected from the supplement materials of [49]. The dictionary in the shape channel was trained from these images. We did not test the object model here because these images did not contain any high-level objects. The images were classified into three categories. Several typical low-level models including Itti *et al.*'s model [1], AWS [6], GBVS [4] and IS [32] were used for comparison. The results are shown in Figure 3, 4 and 5.

The first category contains simple stimuli in which the targets pop out in a basic feature dimension, such as orientation, color, intensity and size (see Figure 3). Stimuli 2, 4, and 6 have more inhomogeneous background than 1, 3 and 5. HBS showed visually satisfactory results consistently for all stimuli. For stimuli 4, only HBS and AWS found the target. For stimulus 5, only HBS and GBVS detected the highest saliency at the single white spot.

The second category contains more complex stimuli (see Figure 4). HBS successfully detected all targets but other models did not perform so well. For instance, in stimuli 6 the target is a regular pentagon and the distractors are all circles. All low-level models failed to find this target.

The third category contains three pairs of asymmetric stimuli (see Figure 5). HBS again successfully detected all targets. Other models generally did not perform well, especially for the last 4 stimuli.

These results demonstrate the superiority of HBS over the low-level models in detecting distinct patterns. This superiority attributes to the stronger representation power of mid-level features, which enables better discrimination of distinctness and suppression of irrelevant variations.

C. Free Viewing Fixations on Natural Images

1) Datasets: Three public human fixation datasets were used in our experiments. The first is the Toronto dataset [5]. It consists of 120 indoor and outdoor images with 681 × 511 pixels. Each image has a resolution of 681 × 511 and was viewed by 20 subjects for 4 seconds. This dataset contains less scene categories and many images have no particular objects of interest. The second is the MIT dataset [24]. It consists of 1003 natural indoor and outdoor images. The longest image dimension is 1024 pixels and the other dimension ranges from 405 to 1024 pixels. Each image was viewed by 15 subjects for 3 seconds. The images contains many meaningful daily life objects such as faces and texts, which rarely appears in the other two datasets. The third is the Kootstra dataset [50]. It consists of 100 outdoor images, which are separated into five categories, including animals, automan, flowers, buildings and nature. Each image has a resolution of 1024 × 768 and was viewed by 31 subjects for 5 seconds. This dataset is the most difficult one as the fixation consistency among subjects is the lowest.

For each dataset, all images were used to train the dictionary of SIFT visual words. 1000 SIFT descriptors were extracted for each bin size and each image.

2) Evaluation Metrics: Several metrics have been proposed in the literature to quantify the ability of saliency models for predicting human fixations. Three of them are adopted here: Area Under the ROC Curve (AUC) [51], Normalized Scanpath Saliency (NSS) [52] and Linear Correlation Coefficient (CC) [53]. AUC is the most widely used metric in saliency model evaluations. Given a threshold th, pixels with saliency value higher than th are classified as fixated (positive samples) while others as non-fixated (negative samples). Using human fixation data as the ground truth, for each th we calculate a true positive rate TPR and a false positive rate FPR, which can be represented by a point in the 2D space. By varying th a curve can be obtained, which is called the ROC curve. AUC measures how well a saliency map predicts the human fixation data on an image. A perfect prediction corresponds to an AUC of 1 and a random guess corresponds to an AUC of 0.5 (chance level). NSS is the average value at human fixation locations of a saliency map, which has been normalized to have zero mean and unit variance. A higher NSS denotes a better performance, and NSS $= 0$ corresponds to a random guess. CC is defined as $\frac{cov(F,S)}{\sqrt{var(F)var(S)}}$, which measures the linear relationship between the fixation map F and a saliency map S. $CC = 1$ denotes a perfect positive linear relationship and $CC = 0$ denotes no linear relationship.

Unfortunately, all of these three metrics suffer from the so-called center bias [54]. Center bias describes the tendency of humans to watch the center of an image regardless of its content, due to the photographer's bias and subjects' viewing strategy [55]. The fixation data of all public datasets shows a high center bias so that a trivial Gaussian center model can achieve high evaluation scores [49]. Some saliency models contain implicit [4] or explicit [24] center prior while other models do not, making the direct comparison unfair.

There are two approaches to deal with this problem. First, use an evaluation metric which is free from the center bias. Shuffled AUC [28] is such a metric. In shuffled AUC, the negative samples are all fixation points across the dataset, except those in the current image. In this way the positive samples and negative samples have the same spatial distribution, and the influence of center bias is eliminated.

Second, add a center prior to each model and evaluate the models in the usual way. In fact, metrics other than shuffled AUC can provide supplementary information about the models. In our experiments, a center map was combined with the saliency map or attention map obtained by any model as follows:

$$S_c = w_1(\text{center map}) + w_2(\text{model map}) \tag{14}$$

where w_1 and w_2 denote combination weights. For fair comparison, the l_2-norm SVM was used to train the optimal weights for each model. The combination was performed

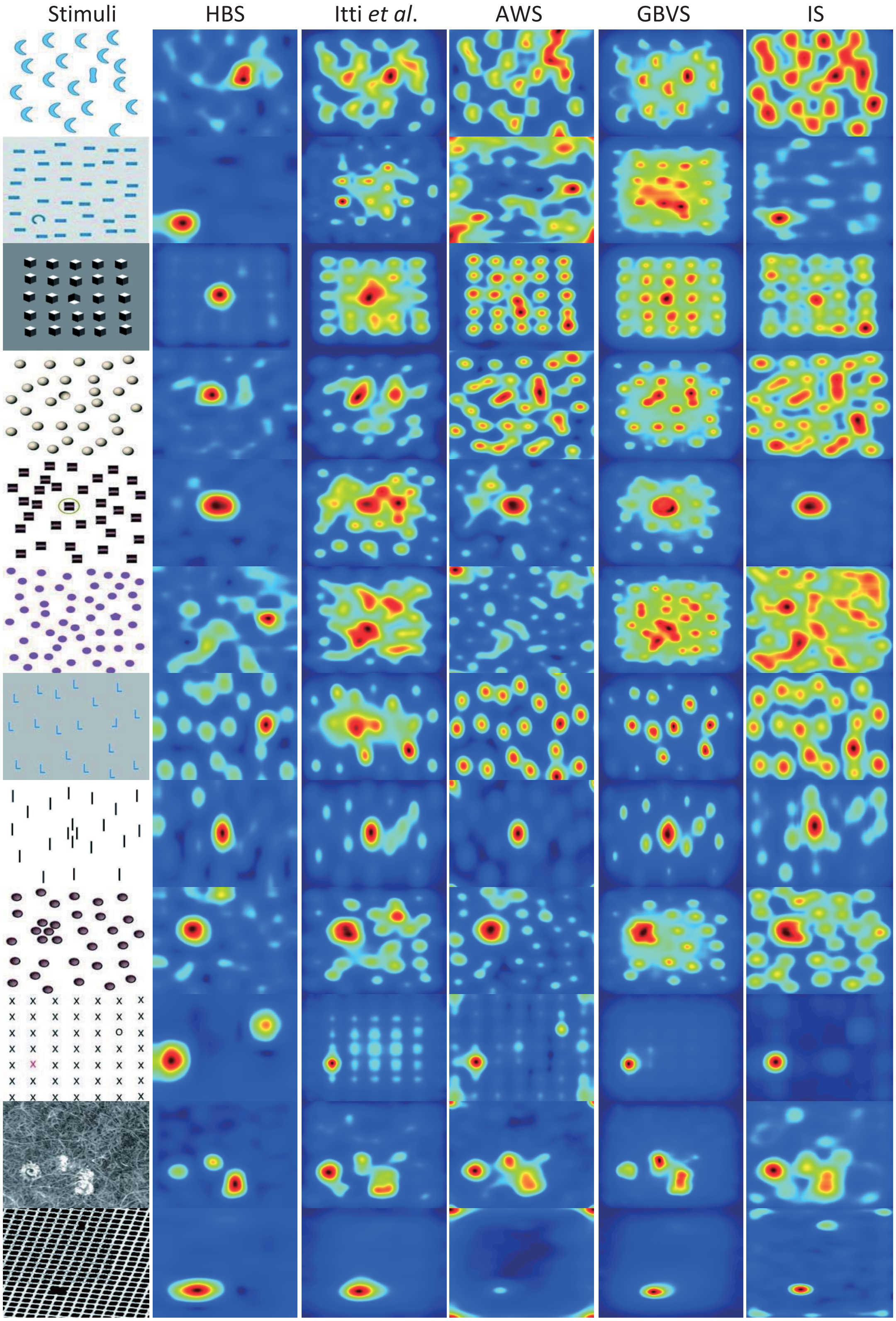

Fig. 4. Complex stimuli. From left to right, the columns denote the original image, saliency maps output by HBS, Itti *et al.*'s model, AWS, GBVS and IS. Best viewed in color.

by a 10-fold cross validation. In each fold, the average of smoothed fixation maps of the training images was used as the center model.

Therefore, for each model, besides the shuffled AUC, we also report AUC, NSS and CC for the combination of the model and the center model.

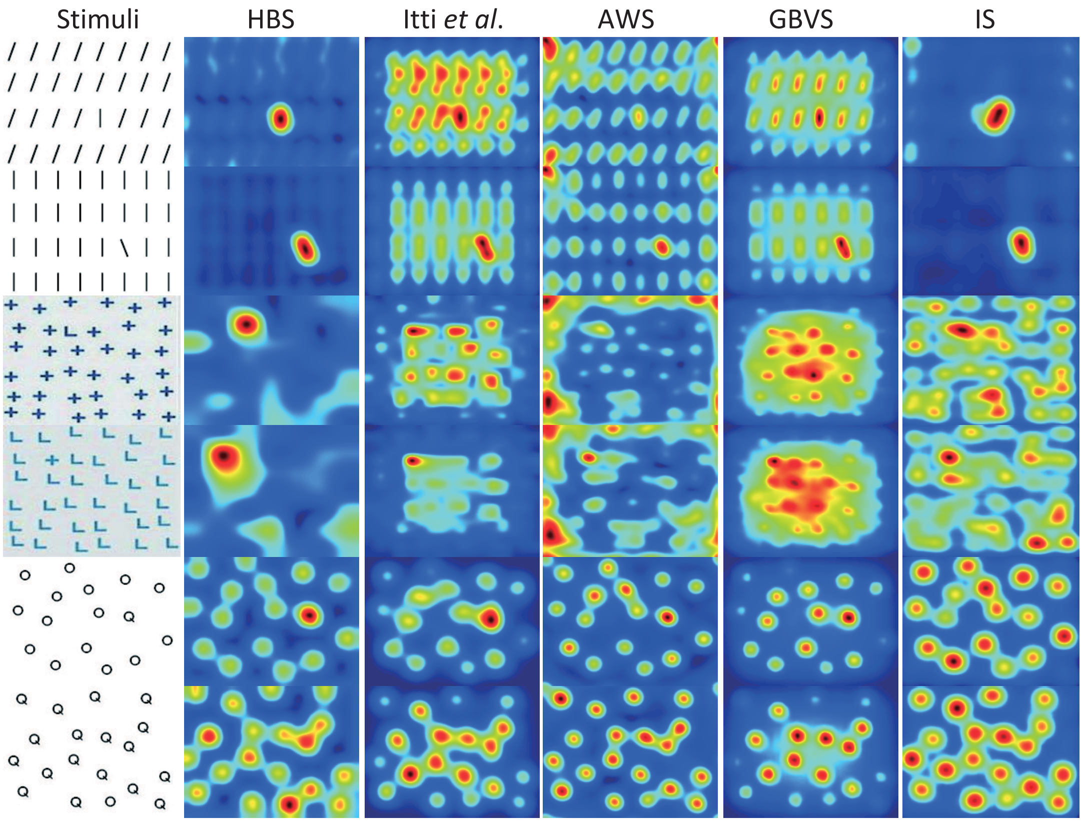

Fig. 5. Asymmetry stimuli. From left to right, the columns denote the original image, saliency maps output by HBS, Itti's model, AWS, GBVS and IS. Best viewed in color.

TABLE I
EVALUATION RESULTS OVER THE TORONTO DATASET[1]

Model Type	Model	Shuffled AUC	SEM	AUC	SEM	NSS	SEM	CC	SEM
Low-level	AWS [6]	**0.723**	0.0075	**0.850**	0.0050	**1.664**	0.0440	**0.626**	0.0106
	SDSR [33]	0.713	0.0082	0.842	0.0053	1.575	0.0413	0.605	0.0115
	IS [32]	0.711	0.0080	0.846	0.0050	1.584	0.0410	0.610	0.0112
	ICL [56]	0.701	0.0076	0.830	0.0053	1.507	0.0440	0.578	0.0135
	COV [34]	0.701	0.0081	0.836	0.0049	1.475	0.0394	0.566	0.0119
	CA [29]	0.696	0.0073	0.838	0.0056	1.534	0.0408	0.588	0.0115
	SR [31]	0.689	0.0090	0.834	0.0052	1.486	0.0407	0.574	0.0119
	AIM [5]	0.674	0.0075	0.838	0.0047	1.460	0.0374	0.560	0.0115
	Itti *et al.* [1]	0.662	0.0092	0.838	0.0052	1.495	0.0402	0.576	0.0115
	SUN [28]	0.651	0.0084	0.819	0.0054	1.382	0.0424	0.539	0.0140
	GBVS [4]	0.644	0.0088	0.836	0.0052	1.522	0.0392	0.586	0.0106
Mid-level	HBS	**0.749**	0.0068	**0.863**	0.0047	**1.760**	0.0463	**0.661**	0.0104
High-level	OB	0.647	0.0069	0.834	0.0047	1.375	0.0303	0.538	0.0094
Spatial Bias	Center	0.499	0.0000	0.807	0.0000	1.312	0.0000	0.518	0.0000
Combination	eDN [38]	—	—	0.841	0.0049	1.333	0.0270	0.531	0.0090
	AWS + HBS + OB	**0.747**	0.0065	**0.868**	0.0043	**1.761**	0.0424	**0.664**	0.0091
	HBS + OB	**0.747**	0.0065	**0.868**	0.0043	**1.760**	0.0429	**0.663**	0.0093

3) Models for Comparison: Our models were compared against 11 low-level saliency models. All of their results were obtained with the original implementations downloaded from the authors' websites. For Itti *et al.*'s model, the version from the GBVS package [4] was used. For SDSR [33] we used the global saliency measure which performed better in our experiments. For covariance saliency model (COV) [34], we used the version parametrized by SigmaPoints which performed better in our experiments.

A recent model eDN [38] was also compared, which integrates multi-level features from neural networks in a supervised manner. We calculated the scores on the MIT and Toronto datasets based on the saliency maps provided by the authors. The model has an explicit center prior, which has suppressed the contribution of features and led to very low shuffled AUC scores. To be fair we only report its AUC, NSS and CC scores. In addition, the saliency maps were combined with our center model according to equation (14), which has improved the scores.

4) Results: Tables I, II and III show the comparison results over the Toronto, MIT and Kootstra datasets, respectively. The average scores over all images and their standard error of the mean (SEM) are reported. AWS [6] performed the best among

[1] The bold numbers indicate higher than or equal to the best scores of all low-level models.

TABLE II

EVALUATION RESULTS OVER THE MIT DATASET

Model Type	Model	Shuffled AUC	SEM	AUC	SEM	NSS	SEM	CC	SEM
Low-level	AWS [6]	**0.701**	0.0040	**0.852**	0.0021	**1.592**	0.0188	**0.488**	0.0045
	COV [34]	0.678	0.0038	0.841	0.0020	1.449	0.0164	0.453	0.0045
	IS [32]	0.675	0.0042	0.844	0.0022	1.484	0.0166	0.465	0.0045
	CA [29]	0.675	0.0040	0.842	0.0022	1.474	0.0164	0.461	0.0044
	ICL [56]	0.671	0.0043	0.835	0.0022	1.458	0.0179	0.455	0.0048
	SDSR [33]	0.658	0.0044	0.837	0.0024	1.446	0.0167	0.456	0.0047
	AIM [5]	0.666	0.0035	0.842	0.0020	1.436	0.0157	0.450	0.0043
	SR [31]	0.660	0.0042	0.835	0.0023	1.451	0.0174	0.454	0.0047
	Itti *et al.* [1]	0.650	0.0041	0.841	0.0021	1.449	0.0158	0.455	0.0043
	GBVS [4]	0.642	0.0043	0.843	0.0022	1.471	0.0159	0.459	0.0041
	SUN [28]	0.638	0.0038	0.830	0.0022	1.398	0.0174	0.439	0.0050
Mid-level	HBS	**0.728**	0.0037	**0.865**	0.0020	**1.672**	0.0185	**0.512**	0.0042
High-level	OB	0.669	0.0033	**0.853**	0.0019	1.411	0.0123	0.437	0.0032
Spatial Bias	Center	0.510	0.0000	0.818	0.0000	1.325	0.0000	0.422	0.0000
Combination	eDN [38]	—	—	**0.854**	0.0019	1.359	0.0108	0.430	0.0031
	AWS + HBS + OB	**0.736**	0.0035	**0.873**	0.0019	**1.696**	0.0173	**0.517**	0.0038
	HBS + OB	**0.736**	0.0035	**0.873**	0.0019	**1.693**	0.0173	**0.516**	0.0038

TABLE III

EVALUATION RESULTS OVER THE KOOTSTRA DATASET

Model Type	Model	Shuffled AUC	SEM	AUC	SEM	NSS	SEM	CC	SEM
Low-level	AWS [6]	**0.626**	0.0071	**0.700**	0.0055	**0.808**	0.0349	**0.576**	0.0148
	IS [32]	0.603	0.0080	0.689	0.0055	0.730	0.0295	0.548	0.0145
	CA [29]	0.603	0.0063	0.686	0.0054	0.734	0.0317	0.542	0.0153
	SDSR [33]	0.602	0.0084	0.688	0.0057	0.731	0.0312	0.542	0.0147
	COV [34]	0.601	0.0069	0.681	0.0054	0.703	0.0320	0.523	0.0162
	ICL [56]	0.598	0.0074	0.680	0.0056	0.720	0.0340	0.533	0.0166
	AIM [5]	0.590	0.0060	0.685	0.0054	0.712	0.0303	0.527	0.0152
	SR [31]	0.588	0.0076	0.680	0.0056	0.700	0.0331	0.525	0.0168
	Itti *et al.* [1]	0.580	0.0074	0.682	0.0055	0.708	0.0313	0.531	0.0154
	GBVS [4]	0.557	0.0063	0.678	0.0054	0.693	0.0283	0.527	0.0146
	SUN [28]	0.556	0.0076	0.666	0.0059	0.657	0.0358	0.497	0.0191
Mid-level	HBS	**0.643**	0.0070	**0.708**	0.0059	**0.874**	0.0380	**0.611**	0.0144
High-level	OB	0.540	0.0056	0.671	0.0052	0.659	0.0278	0.505	0.0145
Spatial Bias	Center	0.495	0.0000	0.660	0.0000	0.643	0.0000	0.490	0.0000
Combination	AWS + HBS + OB	**0.637**	0.0070	**0.711**	0.0056	**0.876**	0.0350	**0.618**	0.0129
	HBS + OB	**0.634**	0.0070	**0.710**	0.0058	**0.873**	0.0356	**0.615**	0.0131

low-level saliency models under all four metrics, and was used as a baseline. Another baseline was the Gaussian center model, which had a chance level performance for shuffled AUC and set a lower bound for AUC, NSS and CC. Note that the reported scores in the tables may not be exactly the same as, though very close to, those in other papers due to some differences in implementation details. For example, there may be differences in the point density of ROC curves, smoothed fixation map used in CC metric, or the center models accounting for the center bias. But these factors have a similar influence on all models and the relative ranks may not change.

HBS outperformed the other models under all metrics over the three datasets, which indicates that the mid-level features better predict the fixations than low-level features. On the other hand, the high-level OB model was surpassed by most low-level models over the Toronto and Kootstra datasets. This is possibly due to the lack of meaningful objects in images of these two datasets. This hypothesis is partly supported by the fact that OB achieved relatively better performance over the MIT dataset, which contains many daily life objects. OB had similar AUC score with AWS, but was outperformed by AWS under the other three metrics. When compared with some other low-level models such as SDSR and AIM, OB had competitive performance. It achieved higher shuffled AUC and AUC but lower NSS and CC. These results indicate that objects do not have stronger prediction power than low-level and mid-level features; this is true at least for the proposed OB model.

With three kinds of models characterized by low-level, mid-level and high-level features, an interesting question arises: if we combine the models together, what is the performance of the resulting model? We then linearly combined AWS (because it performed best among low-level models over all datasets in our experiments) and the proposed HBS and OB models. The linear weights were obtained using the same approach as that for combining the center map and model map (see [24]). Over the Toronto and Kootstra datasets where the OB model was less effective, the combined model exhibited very close performance to HBS under any metric (see Tables I and III). Over the MIT dataset, the combined model showed clear advantages over HBS under the shuffled AUC and AUC metrics (see Table II).

One may wonder how much AWS has contributed to the overall performance in the combined model. We then tested the combination without AWS. The resulting model showed only negligible difference from that including AWS, suggesting that the higher-level features are enough for this prediction task (see Tables I, II and III). In comparison with eDN which

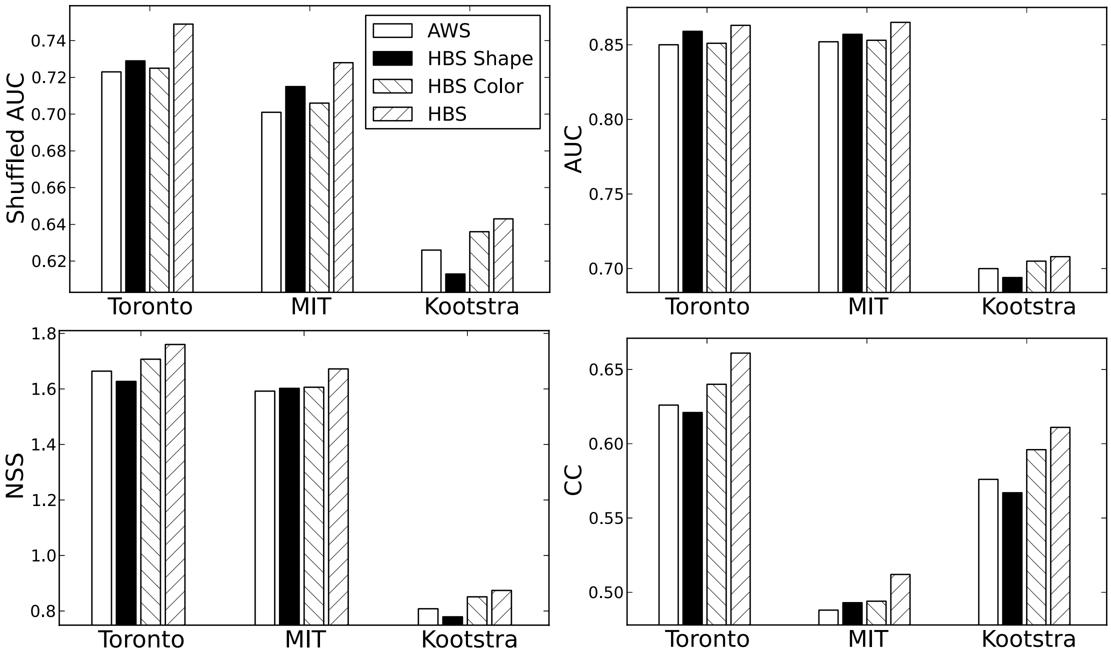

Fig. 6. The scores of HBS shape and color channels, respectively. The scores of AWS and HBS are given for comparison.

TABLE IV
EVALUATION RESULTS OVER THE ASCMN DATASET

Model Type	Model	Shuffled AUC	SEM	AUC	SEM	NSS	SEM	CC	SEM
Low-level	AWS [6]	**0.669**	0.0013	0.790	0.0010	**1.019**	0.0057	0.343	0.0017
	IS [32]	0.663	0.0013	0.790	0.0010	0.999	0.0056	0.339	0.0017
	CA [29]	0.660	0.0014	0.789	0.0010	0.995	0.0055	0.338	0.0017
	SDSR [33]	0.665	0.0013	0.791	0.0010	1.014	0.0056	0.341	0.0017
	ICL [56]	0.664	0.0013	0.787	0.0010	0.988	0.0056	0.335	0.0017
	AIM [5]	0.650	0.0012	0.789	0.0010	0.981	0.0053	0.334	0.0016
	SR [31]	0.666	0.0013	0.788	0.0010	1.003	0.0058	0.340	0.0018
	Itti *et al.* [1]	0.663	0.0012	**0.794**	0.0009	1.018	0.0054	**0.344**	0.0016
	GBVS [4]	0.637	0.0013	0.791	0.0009	1.005	0.0051	0.341	0.0015
	SUN [28]	0.626	0.0013	0.777	0.0010	0.933	0.0060	0.321	0.0018
Mid-level	HBS	**0.675**	0.0013	**0.794**	0.0009	**1.029**	0.0055	**0.346**	0.0016
High-level	OB	0.665	0.0011	**0.800**	0.0009	**1.039**	0.0049	**0.356**	0.0014
Spatial Bias	Center	0.546	0.0000	0.775	0.0000	0.921	0.0000	0.317	0.0000
Combination	AWS + HBS + OB	**0.692**	0.0011	**0.809**	0.0009	**1.088**	0.0048	**0.369**	0.0014
	HBS + OB	**0.691**	0.0011	**0.809**	0.0009	**1.084**	0.0048	**0.368**	0.0014

combines deep network features of three levels, our model combined by HBS and OB achieved better results.

HBS has two feature channels, shape and color. We evaluated them separately to reveal the contribution of each channel. See Figure 6. The scores of AWS and HBS are also shown in the figure for comparison. These two channels exhibited similar performances and were both competitive to AWS. Interestingly, the color channel alone outperformed AWS under all metrics.

D. Free Viewing Fixations on Videos

We then evaluated the models for predicting fixations on videos in the free viewing condition, using the ASCMN dataset [57]. It consists of 24 videos in five categories: abnormal motion, surveillance, crowd motion, moving camera and sudden salient motion. The videos were viewed by 13 subjects, and the gazes were recorded at a video-based frame rate. The same evaluation metrics were used as before. Although motion features should be useful for this task, we focused on the roles of static features only as none of the aforementioned models contains motion features. The evaluation results are shown in Table IV. Note that in [57] the fixation heat maps rather than fixations were used as the ground truth and the saliency maps were preprocessed before evaluation. Therefore the scores of some methods such as SUN in [57] are different from those in Table IV.

Among the low-level models, AWS achieved the highest shuffled AUC and NSS, while Itti *et al.*'s model achieved the highest AUC and CC. HBS again outperformed all low-level models. OB achieved a shuffled AUC higher than most of the low-level models except AWS, and the highest scores under the other three metrics. Notably, OB performed even better than HBS under AUC, NSS and CC.

Again we combined different models and tested the performance. The combination of HBS and OB achieved higher performance than each of them alone. Adding AWS further improved the performance but the improvement was negligible.

In summary, the higher-level models performed better than low-level models in predicting saliency in videos.

OB exhibited better performance on videos than on static images, possibly because the salient regions in videos are often related with semantic objects such as people and cars.

VI. Discussion

In this study we explored the hypothesis that higher-level features contribute to the guidance of eye fixations in the free viewing condition. To achieve this goal, we proposed two attention models HBS and OB, respectively, based on higher-level features, and analyzed their roles in predicting eye fixations in free viewing experiments by comparing with low-level saliency models. In the natural image experiments, HBS outperformed the state-of-the-art models over all three benchmark datasets and OB was effective over a semantically rich dataset, although surpassed by some existing low-level models. In the video experiment, HBS outperformed all low-level models and OB achieved the best scores under three of four evaluation metrics. Furthermore, in both experiments the combination of the two models achieved even higher performance. However, incorporating the best low-level model did not yield noticeable improvement. These results, together with [20], suggest that, in addition to low-level features studied extensively in the literature, higher-level features also contribute to the guidance of attention in the absence of explicit tasks, and this contribution might surpass that of low-level features.

Attention is classified as bottom-up saliency and top-down attention, in which the information flow is upstream or downstream along the visual pathway, respectively. A natural hypothesis is that all layers along the pathway contribute to the guidance of attention. Besides the evidence supporting V1's role in saliency computation [7], [8], previous studies have shown that certain mechanisms in higher-level cortex areas are also involved in attention. The reaction time in visual search experiments is influenced by the 3D depth perception [58], which is possibly processed beyond V1. The figure-ground perception is able to guide the attention [59]. Physiologically it may correspond to the stereoscopic perception [60] and border-ownership [61] in higher cortex areas. There are also physiological evidences [15], [16] that the activity in V4 represents a saliency map. However, previous models on mapping visual features to eye fixations emphasize low-level features. This study demonstrates that mid-level and object-level features are also important, and even more effective than low-level features, though we are unable to assign them to specific stages on the visual pathway due to their abstract modeling procedures.

It has been shown that increased level of features can improve the object recognition accuracy. For example, SIFT features are more useful than gradients orientation [21], BoW features are more useful than SIFT features [40], and even higher level features are more useful than BoW features [62]. Our results show a similar trend for attention from low-level to mid-level features. But we cannot assert this trend along the hierarchy further. In fact, we cannot claim relative contributions of mid-level features and high-level features. One reason is that the proposed mid-level model (HBS) and high-level model (OB) have exhibited different performances on different datasets. In fact, over the three static image datasets HBS performed better and over the video dataset OB performed better. But a more important reason is that they compute attention in unsupervised and supervised manners, respectively, which obscures the contributions of features.

A potential solution to the above problem is to utilize deep learning models and train all levels of features in an unsupervised or supervised manner. Vig *et al.* [38] tested different levels of features for fixation prediction, but these features were actually top-layer features from different networks. The model was outperformed by the combination of HBS and OB. A possible reason is that it did not use saliency activation but directly mapped the features to fixations in a supervised manner. Saliency activation might be important for this task, especially for lower-level features. But the exact reason is unknown and needs further investigation.

References

[1] 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. 20, no. 11, pp. 1254–1259, Nov. 1998.

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

[3] J. M. Wolfe, "Guided Search 2.0: A revised model of visual search," *Psychon. Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.

[4] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006.

[5] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006.

[6] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Decorrelation and distinctiveness provide with human-like saliency," in *Proc. 11th Int. Conf. Adv. Concepts Intell. Vis. Syst.*, 2009, pp. 343–354.

[7] L. Zhaoping, "Attention capture by eye of origin singletons even without awareness—A hallmark of a bottom-up saliency map in the primary visual cortex," *J. Vis.*, vol. 8, no. 5, 2008, Art. ID 1.

[8] X. Zhang, L. Zhaoping, T. Zhou, and F. Fang, "Neural activities in V1 create a bottom-up saliency map," *Neuron*, vol. 73, no. 1, pp. 183–192, 2012.

[9] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vis.*, vol. 8, no. 14, 2008, Art. ID 18.

[10] A. Nuthmann and J. M. Henderson, "Object-based attentional selection in scene viewing," *J. Vis.*, vol. 10, no. 8, 2010, Art. ID 20.

[11] A. Borji, D. N. Sihite, and L. Itti, "Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser *et al.*'s data," *J. Vis.*, vol. 13, no. 10, 2013, Art. ID 18.

[12] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[13] M. P. Young, "Objective analysis of the topological organization of the primate cortical visual system," *Nature*, vol. 358, no. 6382, pp. 152–155, 1992.

[14] L. G. Ungerleider and J. V. Haxby, "'What' and 'where' in the human brain," *Current Opinion Neurobiol.*, vol. 4, no. 2, pp. 157–165, 1994.

[15] J. A. Mazer and J. L. Gallant, "Goal-related activity in V4 during free viewing visual search: Evidence for a ventral stream visual salience map," *Neuron*, vol. 40, no. 6, pp. 1241–1250, 2003.

[16] N. P. Bichot, A. F. Rossi, and R. Desimone, "Parallel and serial neural mechanisms for visual search in macaque area V4," *Science*, vol. 308, no. 5721, pp. 529–534, 2005.

[17] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.

[18] S. Ullman, M. Vidal-Naquet, and E. Sali, "Visual features of intermediate complexity and their use in classification," *Nature Neurosci.*, vol. 5, no. 7, pp. 682–687, 2002.

[19] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[20] M. Liang and X. Hu, "Feature selection in supervised saliency prediction," *IEEE Trans. Cybern.*, doi: 10.1109/TCYB.2014.2338893.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[22] C. Shen, S. Mingli, and Q. Zhao, "Learning high-level concepts by training a deep network on eye fixations," in *Proc. Deep Learn. Unsupervised Feature Learn. Workshop, Conjunct. NIPS*, 2012.

[23] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates Inc., 2007.

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

[25] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 438–445.

[26] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates Inc., 2010, pp. 1378–1386.

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

[28] 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, 2008, Art. ID 32.

[29] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[30] R. Margolin, L. Zelnik-Manor, and A. Tal, "Saliency for image manipulation," *Vis. Comput.*, vol. 29, no. 5, pp. 381–392, 2012.

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

[32] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.

[33] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, 2009, Art. ID 15.

[34] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.*, vol. 13, no. 4, 2013, Art. ID 11.

[35] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. CVPR*, vol. 1. Jun. 2005, pp. 886–893.

[36] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 478–485.

[37] T. Shi, M. Liang, and X. Hu, "A reverse hierarchy model for predicting eye fixations," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 2822–2829.

[38] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. CVPR*, 2014, pp. 2798–2805.

[39] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. Int. Workshop Multimedia Inf. Retr.*, 2007, pp. 197–206.

[40] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 3360–3367.

[41] B. R. Conway, S. Moeller, and D. Y. Tsao, "Specialized color modules in macaque extrastriate cortex," *Neuron*, vol. 56, no. 3, pp. 560–573, 2007.

[42] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1523, Jul. 2009.

[43] B. Berlin, *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA, USA: Univ. California Press, 1991.

[44] F. Shahbaz Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez, "Color attributes for object detection," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 3306–3313.

[45] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[46] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 577–584, 2005.

[47] J. Zhu, S. Rosset, T. J. Hastie, and R. Tibshirani, "1-norm support vector machines," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2003.

[48] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1469–1472.

[49] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2013.

[50] G. Kootstra, B. de Boer, and L. R. B. Schomaker, "Predicting eye fixations on complex visual stimuli using local symmetry," *Cognit. Comput.*, vol. 3, no. 1, pp. 223–240, 2011.

[51] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*, vol. 1. New York, NY, USA: Wiley, 1966.

[52] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 18, pp. 2397–2416, 2005.

[53] T. Jost, N. Ouerhani, R. von Wartburg, R. Müri, and H. Hügli, "Assessing the contribution of color in visual attention," *Comput. Vis. Image Understand.*, vol. 100, nos. 1–2, pp. 107–123, 2005.

[54] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, 2007, Art. ID 4.

[55] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. Vis.*, vol. 9, no. 7, 2009, Art. ID 4.

[56] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates Inc., 2008, pp. 681–688.

[57] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, and T. Dutoit, "Dynamic saliency models and human attention: A comparative study on videos," in *Proc. 11th ACCV*, vol. 7726. 2013, pp. 586–598.

[58] L. Zhaoping, N. Guyader, and A. Lewis, "Relative contributions of 2D and 3D cues in a texture segmentation task, implications for the roles of striate and extrastriate cortex in attentional selection," *J. Vis.*, vol. 9, no. 11, 2009, Art. ID 20.

[59] F. T. Qiu, T. Sugihara, and R. von der Heydt, "Figure-ground mechanisms provide structure for selective attention," *Nature Neurosci.*, vol. 10, no. 11, pp. 1492–1499, 2007.

[60] F. T. Qiu and R. von der Heydt, "Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules," *Neuron*, vol. 47, no. 1, pp. 155–166, 2005.

[61] H. Zhou, H. S. Friedman, and R. von der Heydt, "Coding of border ownership in monkey visual cortex," *J. Neurosci.*, vol. 20, no. 17, pp. 6594–6611, 2000.

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates Inc., 2012, pp. 1097–1105.

Ming Liang (S'13) received the B.E. degree in computer science and technology from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2007, and the M.E. degree in computer science and technology from Tsinghua University, Beijing, in 2010, where he is currently pursuing the Ph.D. degree.

His current research interests include neural networks and their applications in computer vision.

Xiaolin Hu (S'01–M'08–SM'13) received the B.E. and M.E. degrees in automotive engineering from the Wuhan University of Technology, Wuhan, China, in 2001 and 2004, respectively, and the Ph.D. degree in automation and computer-aided engineering from The Chinese University of Hong Kong, Hong Kong, in 2007. He is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include artificial neural networks, computer vision, and computational neuroscience. He is also an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.