Saliency Prediction with Scene Structural Guidance

Haoran Liang*, Ming Jiang†, Ronghua Liang* and Qi Zhao†

*College of Information Engineering, Zhejiang University of Technology, Hangzhou, China
†Department of Computer Science and Engineering, University of Minnesota, USA

Email: *{haoran, rhliang}@zjut.edu.cn, †mjiang@umn.edu, †qzhao@cs.umn.edu

Abstract—Previous works have suggested the role of scene information in directing gaze. The structure of a scene provides global contextual information that complements local object information in saliency prediction. In this study, we explore how scene envelopes such as openness, depth, and perspective affect visual attention in natural outdoor images. To facilitate this study, an eye tracking dataset is first built with 500 natural scene images and eye tracking data with 15 subjects free-viewing the images. We make observations on scene layout properties and propose a set of scene structural features relating to visual attention. We further integrate features from deep neural networks and use the set of complementary features for saliency prediction. Our features are independent of and can work together with many computational modules, and this work demonstrates the use of Multiple kernel learning (MKL) as an example to integrate the features at low- and high-levels. Experimental results demonstrate that our model outperforms existing methods and our scene structural features can improve the performance of other saliency models in outdoor scenes.

Keywords—visual saliency, eye-tracking dataset, scene envelop

I. INTRODUCTION

Humans have a tremendous ability to rapidly direct gaze and select the most relevant information from the visual world around us. Understanding and simulating this attentional mechanism has both scientific and economic impact, and are attracting increasing attention both in human vision and computational vision [1], [2]. Most commonly used methods in visual saliency are built to find regions and objects that stand out from their neighbors using both low-level features [3], [4] (i.e., color contrast, luminance contrast and orientation) and high-level features [5] (i.e., text, face and car), which are validated to be effective by either the biological or computational domains. On top of objects or semantic objects, this work focus on how structure in a scene affects human attention beyond salient objects itself.

This paper presents a focused study on the effect of scene structure on saliency in outdoor natural images. To understand where people look at with different scene structures, we first build a dataset that includes images of common outdoor scenes that are photographed in a standing position. Eye movement data were recorded when 15 subjects free-viewed the images. We make observations from the human fixation data and propose multiple scene level features. We classify the scene layout using intermediate layers of a deep convolutional neural network (CNN), followed by scene structural feature extraction. Multiple kernel learning was used for feature integration and saliency prediction. The main contributions of the paper are summarized as follows:

1. Structural factors relating to saliency in scenes are proposed and can be automatically detected by our simple method.
2. Scene structural information is used for the first time in a computational model to predict saliency. Multiple kernel learning (MKL) is used as the computational module for feature integration. Experimental results validate the effectiveness of the proposed features in this context.
3. A new eye tracking dataset is built for scene structural analysis and for scene saliency computation.

II. RELATED WORK

A. Saliency Model

The first saliency model was proposed by Koch and Ullman [2] and later implemented by Itti et al. [3], inspired by which, a number of algorithms have been developed to predict where humans look at in images along the same line [6], [7]. In these models, low-level features such as color, intensity, and orientation were extracted and feature channels were computed through center-surround filtering at multiple spatial scales, followed by a feature integration step using a linear mechanism to obtain the saliency map. On the basis of the effectiveness of color, intensity and orientation proved by many works in guiding visual search and attention-based computational model, various computational algorithms [8], [9] were proposed to infer saliency of different feature channels. The integration of multiple object detectors in general boosts prediction performance, especially scenes with the objects that have the detectors built and integrated [5]. With improvements of neural networks in the computer vision community and the

This work was done when Haoran Liang was a visiting student in the Zhao Lab.
application of deep architectures. Moreover, a large number of neural network-based saliency models have emerged in a short period of time. These models [10], [11] are trained to predict saliency in a single end-to-end manner, combining feature extraction, feature integration, and saliency value prediction.

B. Scene Layout and Spatial Envelope Saliency

Scene layout estimation from a single image has been an active topic of research over the past 30 years. Torralba and Oliva [12] showed that visual features of natural images are strongly scale-dependent, which allows a simple model based on the output magnitudes of a bank of localized multi-scale oriented filters to determine the absolute depth range of a given scene image. A recent work by M. R. Greene [13] has shown that global properties describing the three-dimensional layout of a scene, such as the dominant depth, openness, or perspective of an environment, can be perceived at first glance, and can influence scene categorization. After that, Ross and Oliva [14] made further progress to bridge visual perception and computer models by investigating the usefulness of global feature in representing scene layout properties. The most relevant works to the proposed method are those that focused on investigating how task instructions affect gaze patterns. Torralba et al. [15] learned target locations using a set of scene prototypes in a cluster-weighted regression model. The locations were then combined with the bottom-up features to form a final map representing salient regions. Similarly, Krista et al. [16] extended Torralba’s model by adding detectors for people search. Both works showed that the structure of scene affected the way people look, though they were in a task-specific context and required learning of the particular search targets for spatial priors.

III. OBSERVATIONS AND STRUCTURAL FEATURES

A. Dataset Collection

We collected a total of 500 natural images, representing a variety of common outdoor environments from the SUN2012 database [17] containing approximately 12,000 scene images. We randomly selected instances from 65 common categories in daily life and removed indoor scenes as well as those not photographed from a standing position. The images were all scaled to 256 × 256. Examples of the images are illustrated in Figure 2. The numbers of each scene are uneven as it is in the SUN2012 dataset, and we did not add images outside SUN2012 to make the number balanced since all the images there are with elaborate annotations. We kept them consistent to facilitate further research.

Fifteen students (8 male and 7 female, between the ages of 18 and 25) with corrected or uncorrected normal eyesight freely viewed the full set of images. These images were presented on a 22-inch LCD monitor (placed 57cm from the subjects), and eye movements of the subjects were recorded using an Eyelink 1000 (SR Research, Os-goode, Canada) eye tracker, at a sample rate of 1000Hz. The screen resolution was set to 1680 × 1050, and the images were scaled to occupy the full screen when presented on the display. Therefore, the visual angle of the stimuli was about 38.8° × 29.1°, and each degree of visual angle contained about 26 pixels in the image.

B. Observations from Eye Tracking Data

The typical appearances of different scene layouts are shown in Figure 3(a). We categorize the content of image into three parts for a better understanding and ease of description, see in Figure 3(b). Vertical refers to the region excluding sky and ground. We analyze the eye fixation data collected from 15 subjects by visualizing their fixation maps. Each fixation map was generated by convolving a 2D Gaussian filter on fixation points gathered from all the subjects in the dataset on one particular image. We report key observations from eye fixation maps, which are listed below.

Convex parts in open scenes attract attention. The convex parts usually appear in open and far scenes. From Figure 1, 4, we can see that most fixations fall into the vertical regions (i.e., regions between the ground and sky) of a scene, in which we observe that gazes often land on objects that are close to the boundary curve between sky and the vertical region, such as houses on the ground, or cars in the street. The most attractive regions often appear to stand out among all the visual references, stretching into the sky, which is denoted as convex parts here. Zooming in further the fixations, the top-middle parts of these objects particularly attract attention, such as the mountaintop, skyscraper, and chimney (Figure 4).

Vanishing points attract attention. Vanishing points usually appear in perpendicular scenes. We observe that vanishing points attract attention strongly and rapidly (see in Figure 4c). Bias toward vanishing points leads to a stronger spatial bias in fixations than center bias [18]. This is because viewing from a standing position, the intensity of visual references increases when it goes deeper and farther along the way where most
objects rest on. It can be seen that even when nothing appears at the vanishing point, people still have the tendency to figure out what is going to show up.

**Low-level feature plays an important role.** Observations show quite a strong correlation between low-level feature maps and object semantics in scene-centric images, thus suggesting leveraging these maps in this work. For example, visual references with edges in each orientation (e.g., text and sign) usually have large responses in orientation maps.

### IV. Computational Model

Based on observations from eye tracking data, we report in this section a model that combines both low- and scene-level features for eye fixation prediction. In this section, we show the details of the proposed features and the saliency model.

#### A. Structural Features

To incorporate scene layout properties in our model, we specify the three main geometric classes, namely ground, vertical and sky (see in Figure 3 (b)), which have relatively predictable spatial distributions and can be reasonably detected by previous works. Here we use geometry context [19] to make a rough segmentation of the input images. In practice, we replace the superpixel algorithm with simple linear iterative clustering (SLIC) [20] to achieve improvement in speed as well as in memory efficiency. The confidence maps for ground, sky and vertical region are then converted to binary maps by taking the most likely label for each superpixel. In other words, if one superpixel has the highest probability representing the sky, it is set to 1 in the sky map and 0 in ground and vertical maps.

**Ground horizon.** Similar to the horizon feature applied in [5], the horizontal line is particularly important in outdoor scenes with a standing view. In this work, we simply calculate the vertical projection curve \(G(x)\) (\(x\) denotes the horizontal coordinate) for the ground map and select the maximum value \(H_l = \max(G(x))\) as the height of the horizontal line in images. Note that in some images the horizontal line is falsely located in sky region due to the absence of vertical content or the fluctuation of boundaries, we simply constrain the horizontal line to be under the sky region before looking for it.

**Convex part.** Starting from the horizontal line, we calculate above the upper bound of horizon line the vertical projection curve \(C(x)\) in the vertical map and smooth it using mean filter. We then select the top \(n(n \leq 3)\) convex parts that have the local maximum value by \(C'(x) = 0\) and \(C''(x) < 0\), where \(C'(x)\) and \(C''(x)\) are the first and second derivatives of \(C(x)\) respectively. Finally, the vertical coordinates of the convex points are set to \(\mu H_v\), where \(H_v = \{H_{v1}, ..., H_{vn}\}\) and \(H_{ci}\) denotes the height of the \(i\)th convex part (i.e. approximately the red region of the color bar in Figure 5) and \(\mu\) is a scalar which is empirically set to 0.8.

**Vanishing point.** The vanishing point normally lies where the ground stretches, therefore it can be accurately conjectured if the ground is segmented correctly. Due to the convex shape of the ground, the pixel selected for deciding horizontal line approximates the vanishing point in most images. A typical example is shown at the bottom in Figure 5. Most failures are related to the instability of flatness within the ground region. In order to make it robust in the detection of the vanishing point, we take into consideration the local regions within the horizon part of the images. In particular, we densely divide the horizon region into \(k\) patches (Figure 5c) and use the contextual features \(F_v\) from the geometric context to represent each patch. The contextual features are the average confidence for the patch itself, average confidences for the patch windows above and below the patch, and the above-center and below-center differences for each of the eight geometric classes, yielding a 40-dimensional feature vector. We select some typical examples with accurately located vanishing point to be prototypes. The calculated feature \(F_v\) is compared to prototype’s feature \(F_p\) at the same location as \(d = ||F_v - F_p||_2\). Finally, the patch with minimum \(d\) is selected as a candidate position for the vanishing point.

#### B. Multi-scale Low-level Features

**Low-level features:** The classic low-level features proposed by Itti-Koch [3] include multi-scale color, intensity and orientation maps. We in this work use the same color and intensity maps with the implementation of Itti et al. and the feature maps shown on the left side in Figure 6 show the effectiveness of the two simple and biologically plausible features in representing objects in low spatial frequency. In addition, the Gabor filters

---

**Fig. 4.** Three main factors including (a) the horizontal line, (b) the convex parts and (c) the vanishing point are shown in the yellow boxes. These factors encode the scene structure and complement bottom-up features that highlight regions with contrast.

**Fig. 5.** The illustration for obtaining convex point and vanishing point for different types of scene.
have been found to be particularly appropriate for texture representation and discrimination. We set up Gabor filters with 4 orientations (i.e., $0^\circ$, $45^\circ$, $90^\circ$, $135^\circ$) to the input images at 5 scales, yielding 20 orientation feature maps.

The usefulness of surroundedness has been shown in boolean map saliency (BMS) [21], one of the state-of-the-art models in predicting eye fixation, and it uses color features only. BMS makes use of simple image processing operations to leverage the topological structural cue, showing the superiority of the center-surround color map. The maps in Figure 6 illustrate that although BMS fails to represent the scene structural features (i.e., the vanishing point), the surrounded and contrastive regions are emphasized with less redundancy than feature maps from Itti et al. Hence the map obtained by BMS is selected as an additional feature map at the original scale.

**Structure bias:** Since the convex point and vanishing point attract attention strongly and rapidly, we take into account this bias by using a distance-to-scene (DTS) map which is quite similar to the distance-to-center (DTC) map used in [5]. The value of a pixel in DTS map is defined as following:

$$L(x, y) = 1 - \frac{\sqrt{(x - s_x)^2 + (y - s_y)^2}}{\sqrt{(s_x)^2 + (s_y)^2}}$$  \hspace{1cm} (1)

where $x$ and $y$ are the indexes of pixel, $s_x$ and $s_y$ are the indexes of the nearest convex point or vanishing point to the pixel.

**C. Learning a Saliency Model with Multiple Kernels**

In the experiment, we train a classifier on our image set using 10-fold cross validation. Respectively, 450 images are used for training and 50 images for testing each time. We randomly select 5 pixels respectively from the top 20% and bottom 40% regions, yielding a training set of 4500 samples (2250 positive samples and 2250 negative samples). Each sample is represented as a feature vector which is generated by concatenating the values from every feature map on one particular pixel, i.e., the 25 feature maps in our experiment lead to a 25-dimensional feature vector for each sample. Both training and testing samples are normalized to have a zero mean and unit variance. Instead of a linear SVM model, we use multiple kernel learning (MKL) [22] to integrate different features from all levels. By combining multiple kernels of support vector machines (SVMs) instead of one, MKL is able to remove the assumptions of kernel functions and eliminate the burdens of manual parameter tuning in kernel functions of SVMs.

We use the simpleMKL algorithm [22] in our model to solve this MKL problem and the probability of eye fixations on each point ($f(x)$). The final saliency map $S$, can then be obtained by

$$S = \{\max (f(x), 0) * g) \circ L$$  \hspace{1cm} (2)

where $g$ is a gaussian mask that is used to smooth the saliency map, $*$ is the convolution operator, $L$ refers to the DTS map and $\circ$ denotes the Hadamard product operator.

**V. EXPERIMENTAL RESULTS**

This section reports experiments and results to validate the proposed model in predicting eye fixations in scenes. For a fair comparison and a comprehensive assessment, the fixation prediction results of all the models were measured by three similarity metrics and all the evaluation scores presented in this section are obtained as the highest score by varying the smooth parameter (standard deviation of a Gaussian mask) from 1% to 5% of the image width in a step of 0.05%. An illustration of the training and prediction procedure are shown in Figure 7.

**A. Experimental settings**

For each image, each of the three layout properties namely openness, depth, and perspective was rated on a continuous 1-6 scale [14], e.g., to rate the degree of depth, “1” represents near and “6” represents far. The score determines which structural feature exists. In this paper, the deep scene features from the higher level layer of deep convolutional neural networks (CNet) are used as generic features to determine which attributes exist inside training images. In our implementation, we use the Caffe [23] package to generate the 4096-dimensional feature vector as descriptor from the 7-th fully-connected layer in the pre-trained Places205-CNN [24] model. After that, we simply employ support vector regression (SVR) to perform regression on descriptors to generate the ratings for scene properties. This procedure brings not only simplicity by avoiding parameter selection in GIST or CWM, but also boosted accuracy. The scene attribute decides which structure map to compute and should be identified before the saliency prediction. For the
defined features above, the convex part is only calculated if the scene is considered to be open or far. Similarly, the vanishing point appears in perpendicular scenes only.

In this work, the similarity metrics we use include shuffled Area Under Curve (sAUC), Normalized Scanpath Saliency (NSS) and Linear Correlation Coefficient (CC), whose codes and descriptions are available online [25].

**B. Performance Evaluation**

Among the comparative models in our experiment, except for the Judd model [5] that is with object detectors, all the other methods are purely bottom-up including Boolean Map Saliency (BMS) [21], the Adaptive Whitening Saliency (AWS) [26], the Attention based on Information Maximization (AIM) model [27], the Image Signature (SIG) model [28], the Graph Based Visual Saliency (GBVS) model [29], the Saliency Using Natural Statistics (SUN) model [30], and the Itti et al.’s model [3].

We further explore the advantage of the MKL-based framework and the contributions from the structural feature maps. Specifically, we add the result obtained by the SVM-based framework (denoted as OUR-SVM) as well as the model without structural feature maps (denoted as OUR-NOS). From the curves in Figure 8 that is calculated following Borji’s implementation [25], we observe that the proposed model outperforms all the other saliency models with all the three similarity metrics. We also visualize human fixation maps and saliency maps generated from different saliency models mentioned above on sample images selected from our scene dataset in Figure 9.

From the saliency maps, it can be seen that the scene layout properties do not stand out as a specific entity for people to gaze to. Instead, it is more of a contextual guidance for people to decide where to look at with a certain level of top-down knowledge. For instance, in the bottom row in Figure 9, other models consider the car on the left to be the most salient object, while people prefer the inconspicuous car with smaller size near the vanishing point which gets emphasized in our model that produces a better saliency map.

The MKL is able to combine different features by choosing suitable kernels to produce a result (in sAUC) that is 2.26% better than that obtained from the SVM-based framework. We also find that the evaluation scores decrease significantly if we remove the structural feature maps, showing the effectiveness of structural features in guiding attention in scenes.

**C. Integrating Scene Features with Existing Models**

We further test the proposed structural features on other existing eye tracking dataset which also contain a variety of natural outdoor scene, *i.e.* the Toronto [31] and MIT [5] datasets.
Since the proposed method only handles outdoor scenes while more than half of the MIT dataset is indoor or object focused cases, all the images from MIT dataset are fed into PlacesCNN [24] at first so that the outdoor cases can be automatically picked out. Particularly, an image will be chosen only when score obtained from SVR is larger than \( \theta \), forming a subset of 212 images. In our experiment, \( \theta \) is set to 4.1, 4.5, 3.8 for depth, openness, perspective respectively.

The proposed scene structural features can also be used with existing models and the effectiveness is validated with two well-referred models, i.e., Judd and BMS. Table I shows model performance in Toronto and MIT datasets. Models with a “4S” are those where the scene structural features are incorporated into the comparative models. In particular, in Judd’s model we replace the DTC maps with our DTS maps, and in BMS, the saliency maps are pixel-wise multiplied with our DTS maps. The results demonstrate that the structural features can be used independently to effectively improve the performance of general saliency models. The superior performance of the proposed model, and the considerable improvement after adding structural bias into the saliency maps suggest that scene layout property plays an important role in directing gaze.

VI. DISCUSSION AND CONCLUSION

This work studies the role of scene structures in visual attention. We first build an eye tracking dataset on scene images with different layout properties namely openness, depth, and perspective. A number of structural features are proposed based on observations on human fixation data. The layout is determined by the intermediate layers of a deep CNN and features are later extracted by our proposed method. A computational model that integrates multiple level features is developed using multiple kernel learning. Our method could be used together with other models including deep learning models.

REFERENCES