IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 27, 2019, accepted January 10, 2020, date of publication January 14, 2020, date of current version January 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2966628

Predicting Saccadic Eye Movements in Free Viewing of Webpages

CHEN XIA, (Member, IEEE), AND RONG QUAN
School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Chen Xia (cxia@nwpu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802314, in part by the China Postdoctoral Science Foundation under Grant 2017M623242, and in part by the Project Supported by the Natural Science Foundation of Shaanxi Province under Grant 2019JQ-296.

ABSTRACT Attention modeling for webpages has emerged as a new research direction in computer vision. Despite an amount of research effort, most studies have focused on estimating webpage saliency to reveal the static location of human fixations. Without temporal information, existing models cannot interpret the dynamic properties of the actual attention process in free-viewing webpages. To solve this problem, we propose a webpage-based saccadic model in this study to model dynamic visual search behaviors of humans when they view webpages. In the first stage, we utilize the support vector machine to learn the mapping from multilevel saliency features to an initial probability of being fixated. In the second stage, we combine the mechanisms of spatial bias and inhibition of return with the estimation of the initial probability to iteratively predict a sequence of successive fixations for each webpage image. Experimental results on a benchmark eye-tracking data set for webpages have demonstrated that the proposed model outperforms the state-of-the-art saccadic methods.

INDEX TERMS Saccadic scanpath, semantic hashing code, support vector machine, visual attention, web viewing.

I. INTRODUCTION

The webpage is a prominent platform for information communication on the Internet. With the prevalence of the Internet, webpages have played an important role in our daily life. According to the statistics on Internet Live Stats, there are over 1.5 billion websites on the world wide web. Moreover, the total number of Internet users has reached 4.5 billion in June 2019. Due to the ubiquitous webpage reading, it is necessary to investigate how humans deploy visual sources to acquire information in free-viewing webpages.

For the human visual system (HVS), there is an imbalance between input and computing resources. Embraced by overwhelming amounts of visual input, the HVS can still effectively process visual information to form a more accurate understanding of the external world [1]. The main reason for that is the selective visual attention. It serves as a mediating mechanism, which selects the most critical regions into detailed processing while limiting the influence of the rest areas [2]. Based on its important role in visual processing, modeling visual attention for webpages can contribute to revealing the internal working mechanisms of the HVS in the free viewing of webpages. On the other hand, predicting visual attention is beneficial to webpage design. A good webpage not only needs to meet functional needs, but also needs to have a reasonable layout. Visual attention prediction can help to improve the layout of webpages and thus achieve a better reading of Internet users.

Due to the significance of attention modeling for webpages, researchers have begun to predict where humans pay attention on webpages. Shen and Zhao [3] constructed the eye-tracking data set of webpages and used multiple kernel learning to learn saliency maps from recorded data. Zheng *et al.* [4] proposed an end-to-end learning framework to estimate task-driven webpage saliency. Despite multiple computational frameworks, the existing work usually aims at estimating two-dimensional saliency maps for webpages to encode the static saliency of distinct areas [3]–[6]. However, actual attention is a dynamic visual search process [7]. Even two regions with the same saliency value have different attended orders. As is shown in Fig. 1, a saliency map cannot explain the dynamic behaviors in attention, such as

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang.

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

FIGURE 1. Comparison between webpage saliency estimation and saccade prediction. (a) Webpages. (b) Saliency maps (fixation heat maps). (c) Saccadic scanpaths of different subjects. (d) Predicted saccadic scanpaths by the proposed model.

the temporal relationship between salient regions and the generation of saccadic scanpaths. To explore the dynamic attributes of human visual behaviors in free-viewing webpages, we propose a saccadic model for webpages, which can predict saccadic scanpaths according to the content of webpages.

The proposed model consists of two main stages. In the first stage, we calculate a saliency map to estimate the initial distribution of fixations for scanpath generation. Concretely, we first extract multilevel saliency features to represent each area from distinct perspectives. Then, we utilize the support vector machine (SVM) to learn the relationship between the extracted features and the initial probability of being fixated on webpages. In the second stage, we combine the mechanisms of spatial bias and inhibition of return (IoR) with saliency estimation to predict human saccadic scanpath on each webpage image. In summary, the main contributions are three-fold.

- We aim at investigating the dynamic properties of human attention in free-viewing webpages. Based on existing research on webpage saliency, it can establish a complete visual attention model for viewing webpages.
- We construct a more effective feature description for saliency estimation. We employ deep networks to extract deeper and more abstract representation from input. Then, based on learning from eye-tracking data, we can more accurately model the relationship between multilevel features and saliency.
- We model the spatial influence on scanpath generation, which can generate scanpaths more consistent with human eye movements. Experimental results have shown that spatial bias is beneficial to improve the performance on saccade prediction for webpages. By combing the influence of saliency, spatial bias, and IoR, the final model outperforms the state-of-the-art saccadic methods.

The remainder of the paper is organized as follows: Section II reviews related work on visual attention modeling. Section III describes the proposed saccadic model for webpages. The performances of models are evaluated in Section IV, and the conclusions are drawn in Section V.

II. RELATED WORK

A. VISUAL ATTENTION MODELS ON NATURAL IMAGES

Most existing visual attention models have originated from Treisman and Gelade's Feature Integration Theory (FIT) [8]. It suggests that different visual features are combined in parallel to affect the human attention process. Based on their study, Koch and Ullman [9] fused the influence of different features into a two-dimensional topographic map called ''saliency map'' to represent the conspicuity of each region in an image. Furthermore, Itti *et al.* [2] completely implemented the framework in [9] and built the classic bottom-up computational approach of attention. They proposed the center-surround (C-S) assumption and estimated saliency according to the feature contrast between the central pixel and the average of neighborhood pixels. Based on this milestone, predicting the saliency maps consistent with the eye movements of humans has become the primary task in attention modeling.

To obtain a more accurate saliency map, in the past two decades, a large number of models have emerged to improve Itti *et al.*'s model from distinct perspectives [2]. Firstly, the pixel-wise comparison has been replaced by patch-based feature difference which takes context information into account. For instance, Bruce and Tsotsos [10] determined saliency by comparing the independent components of central and surrounding patches. Borji [11] used the space-weighted feature dissimilarity between the center patch and other surrounding patches to represent local saliency. Han *et al.* [12] adopted the patches from image boundary to model background and calculated saliency by the reconstruction residual of a background-based network.

Secondly, the area for C-S comparison has expanded from local neighborhoods to nonlocal or global regions. As stated in previous research [13], the local comparison is insufficient when a region has low C-S contrast but the entire local region is globally rare. Therefore, models have begun to integrate more nonlocal information into C-S comparison. For instance, Xia *et al.* [14] described the C-S contrast based on an autoencoder network learned from each scene globally. Wang *et al.* [15] extended the range of context to a corpus of similar images to stress the regions deviating from traditional notions.

Thirdly, learning has been introduced in the calculation. On the one hand, methods have used learning as methodologies for building features in C-S comparison to enhance the generalization ability of models. For instance, Borji and Itti [13] learned V1-like features from natural scenes to compute local and global rarity. Vig *et al.* [16] searched for the optimal blend of features from a hierarchical model family. On the other hand, models have learned the relationship between features and saliency values directly. For instance, Shen and Zhao [17] adopted a linear classifier to learn the inference from extracted features to saliency. Wang and Shen [18] applied an end-to-end CNN to predict multilevel saliency from the input of an image.

Besides the research on saliency estimation for predicting human fixations, another branch of saliency calculation has emerged to detect object-level salient areas, which has achieved a wide range of applications. Cheng *et al*. [19] first segmented each scene into regions and computed saliency based on the comparison between regions. To use the information beyond the current image for saliency estimation, Wang *et al*. [20] calculated the saliency of each image by warping the annotations of similar scenes. For video salient object detection, Wang *et al*. [21] computed static and dynamic saliency by unitizing CNNs to learn the inference from input to labeled salient areas. In [22], Wang *et al*. calculated object-aware saliency to generate spatiotemporal saliency prior for video object segmentation. In [23], Wang *et al*. also applied a CNN to predict the attention bounding box for each image. Then, they integrated an aesthetics assessment based network to select from the attention-based candidate windows for photo cropping. In [24], Wang *et al*. estimated stereoscopic saliency based on disparity and edge, and used saliency to guide stereoscopic thumbnail generation.

In the development of visual attention, saliency estimation has always been a research topic of great interest. With a large amount of research effort, the last decade has witnessed significant improvements in the prediction of saliency maps. However, a saliency map does not contain any dynamic information. Therefore, it cannot wholly account for actual human eye movements. To investigate the dynamic properties in visual attention, saccadic models have emerged to predict human saccadic scanpaths. The earliest research can be traced back to Itti *et al*.'s model [2], which employed winner-take-all (WTA) and IoR on each saliency map to generate a sequence of successive fixations. Besides, Lee and Yu [25] interpreted saccadic eye movements under the framework of information maximization. They iteratively selected the locations of the maximum complexity in responses for scanpath generation. Similarly, Wang *et al*. [26] directed the next fixation to the maximum on a residual perceptual information map measured by Site Entropy Rate.

In recent years, another tendency in saccade prediction is to regard saccade as a Markov process, with the next fixation determined by the maximum transition probabilities calculated based on the last fixation. Liu *et al*. [27] calculated the transition probabilities based on low-level saliency and the semantic content modeled by a Hidden Markov Model. Le Meur and Liu [28] combined saliency, oculomotor biases, and memory effect for estimating the transition probabilities. To further model the effect of stimuli and spatial location, Le Meur and Coutrot [29] extended the previous work [28] by training the model with distinct image categories and spatial locations. Also inspired by [28], Wu and Chen [30] introduced visual memory and combined it with oculomotor bias and IoR in the calculation of transition probabilities for gaze shifts.

B. VISUAL ATTENTION MODELS ON WEBPAGES

With the popularity of the Internet and the rapid development of big data, webpages have become one of the most important channels for humans to acquire information from the external environment. Due to the ubiquitous webpage data, it is necessary to understand how humans deploy their attention on webpages in free-viewing tasks. To this end, multiple methods have been proposed to investigate the attention process on webpages.

Shen and Zhao [3] pioneered in this direction and proposed an early saliency model for webpages. Firstly, they analyzed the features and mechanisms underlying webpage saliency. Then, they constructed the first eye-tracking data set of webpages by collecting the eye movements from 11 observers on 149 webpages. Finally, they learned from the data set via multiple kernel learning to obtain the model based on distinct features and positional bias. In [31], they extended the study of [3] by generating high-level representations from CNNs. Besides, Li *et al*. [5] improved Shen and Zhao's work [3] from two perspectives. For one thing, they introduced subband features to complement the features calculated in the space domain. For another, they detected object blobs in webpages to further enhance the performance of the model.

Besides webpage saliency under the free-viewing conditions, in recent years, researchers have begun to model task-driven attention on webpages to investigate the effect of targets on human eye movements. For instance, Zheng *et al*. [4] proposed an end-to-end learning framework to estimate the task-driven webpage saliency. They first constructed an eye-tracking data set with the stimuli from six categories. Then, they separately calculated task-specific and task-free fixation maps by learn from the data set. Finally, they integrated the effects of two components in the way of addition to derive the final saliency map.

In summary, the previous research on attention modeling for webpages has mainly focused on the estimation of webpage saliency [32]. However, these studies cannot interpret the temporal sequence of fixations, which is valuable for understanding actual human attention during visual exploration. Because of this, we present a saccadic model of webpages in this study to predict the shifts between fixations and to generate saccadic scanpaths on distinct webpage images.

III. SACCADIC MODEL FOR WEBPAGES

In this section, we describe the framework of the proposed saccadic model for webpages. The overall flowchart of the model is presented in Fig. 2. As is shown in the figure, the model consists of two main stages. In the first stage, we calculate a saliency map based on multilevel features to estimate the initial distribution of fixations. In the second stage, we combine the mechanisms of spatial bias and IoR with saliency estimation to iteratively predict fixations for saccadic scanpath generation. Besides, we have shown the overall algorithm in Algorithm 1 to provide more algorithmic details.

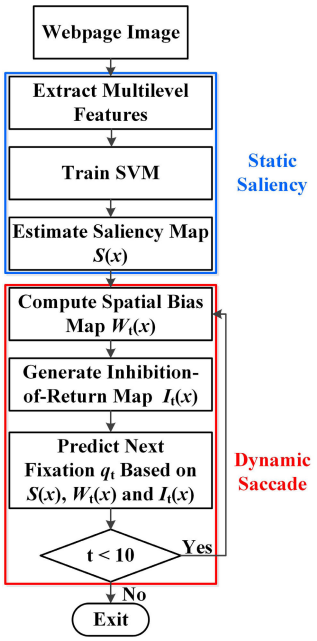


FIGURE 2. Framework of the proposed saccadic model for webpages. Given a webpage image, we predict the saccadic scanpath consisting of ten fixations according to the image content.

A. SALIENCY ESTIMATION WITH FEATURE FUSION

Results from perceptual research [28], [33] have indicated that saliency is an influential factor in guiding saccadic eye navigation. The salient regions with rare visual information usually present a high probability of being fixated. Therefore, we first compute the saliency map of each webpage image to estimate the initial probability of the gaze shift.

In saliency estimation, the core of our model is to learn the relationship between visual representation and saliency value from human eye-tracking data. We first construct the description of each region by extracting multilevel features. Then, we choose positive and negative samples separately from the most salient and nonsalient areas to train the parameters of mapping. Finally, based on the learned model, we calculate the saliency values of distinct locations to generate the saliency map of each test image.

To make a complete description of scenes, we use the features from different levels to represent each pixel. As is shown in Fig. 3, we first extract six low-level features to generate a physiologically plausible representation.

1) SUBBANDS OF THE STEERABLE PYRAMID

The pyramid subbands in four orientations and three scales.

2) 3-CHANNEL COLOR

The red, green, and blue color of each pixel.

3) PROBABILITY OF COLOR

The probability of the pixel's value in the corresponding color channel.

Algorithm 1 The Proposed Model of Predicting the Sequence of Saccadic Points

Input: Eye-tracking data D, Test image I, Number of saccadic points $n = 10$, time $t = 1$.

Output: A set of saccadic points $\{q_1, q_2, \cdots, q_n\}$.

Train SVM model

1: Select positive and negative samples according to the recorded data.
2: Extract multilevel features for each sample according to III-A.
3: Train SVM to obtain the parameters $\{\boldsymbol{W}_S, \boldsymbol{b}_S\}$ of the model.

Predict scanpath on a given image

1: Extract multilevel features $\boldsymbol{f}(\boldsymbol{x})$ for each pixel $\boldsymbol{x}$ according to III-A.
2: Compute the saliency value $\boldsymbol{S}(\boldsymbol{x}) = \boldsymbol{W}_S{}^T \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{b}_S$ by feeding the features $\boldsymbol{f}(\boldsymbol{x})$ into the trained model.
3: **while** $t \leq n$ **do**
4: Compute top-left bias map $\mu(\boldsymbol{x})$ according to (8).
5: Compute oculomotor bias map $\rho_t(\boldsymbol{x})$ according to (9).
6: Generate spatial bias $\boldsymbol{W}_t(\boldsymbol{x})$ by integrating $\mu(\boldsymbol{x})$ and $\rho_t(\boldsymbol{x})$ according to (10).
7: Compute IoR map $\boldsymbol{I}_t(\boldsymbol{x})$ according to (11).
8: Estimate integrated map $\boldsymbol{F}(\boldsymbol{x})$ as $\boldsymbol{S}(\boldsymbol{x})\boldsymbol{W}_t(\boldsymbol{x})\boldsymbol{I}_t(\boldsymbol{x})$.
9: Select the location of the maximum $\boldsymbol{F}(\boldsymbol{x})$ as q_t.
10: $t = t + 1$.
11: **end while**

4) ITTI MODEL FEATURES

The conspicuity under intensity, color, and orientation in Itti *et al.*'s model [2], which is computed by across-scale C-S contrast.

Concretely, we resize the image to the resolution of $200 \times 200 \times 3$ and use Gaussian pyramid to subsample the image into 3 scales. At each scale l (where $l \in [0, 1, 2]$), we calculate the intensity map $I(l)$, four color maps $R(l)$, $G(l)$, $B(l)$, and $Y(l)$, and four orientation maps $O(l, \theta)$, $\theta \in \{0°, 45°, 90°, 135°\}$. Then, we implement the C-S operation $\ominus$ as the difference between fine and coarse scales. The C-S maps under intensity, color, and orientation are computed as (1), (2), and (3), respectively.

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)|, \tag{1}$$

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$$

$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|, \tag{2}$$

$$\mathcal{O}(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|, \tag{3}$$

where $[c, s] \in [0, 1], [0, 2], [1, 2]$. Finally, for each feature, we add the results of different scales to obtain the conspicuity under intensity $\bar{\mathcal{I}}$, color $\bar{\mathcal{C}}$, and orientation $\bar{\mathcal{O}}$, respectively.

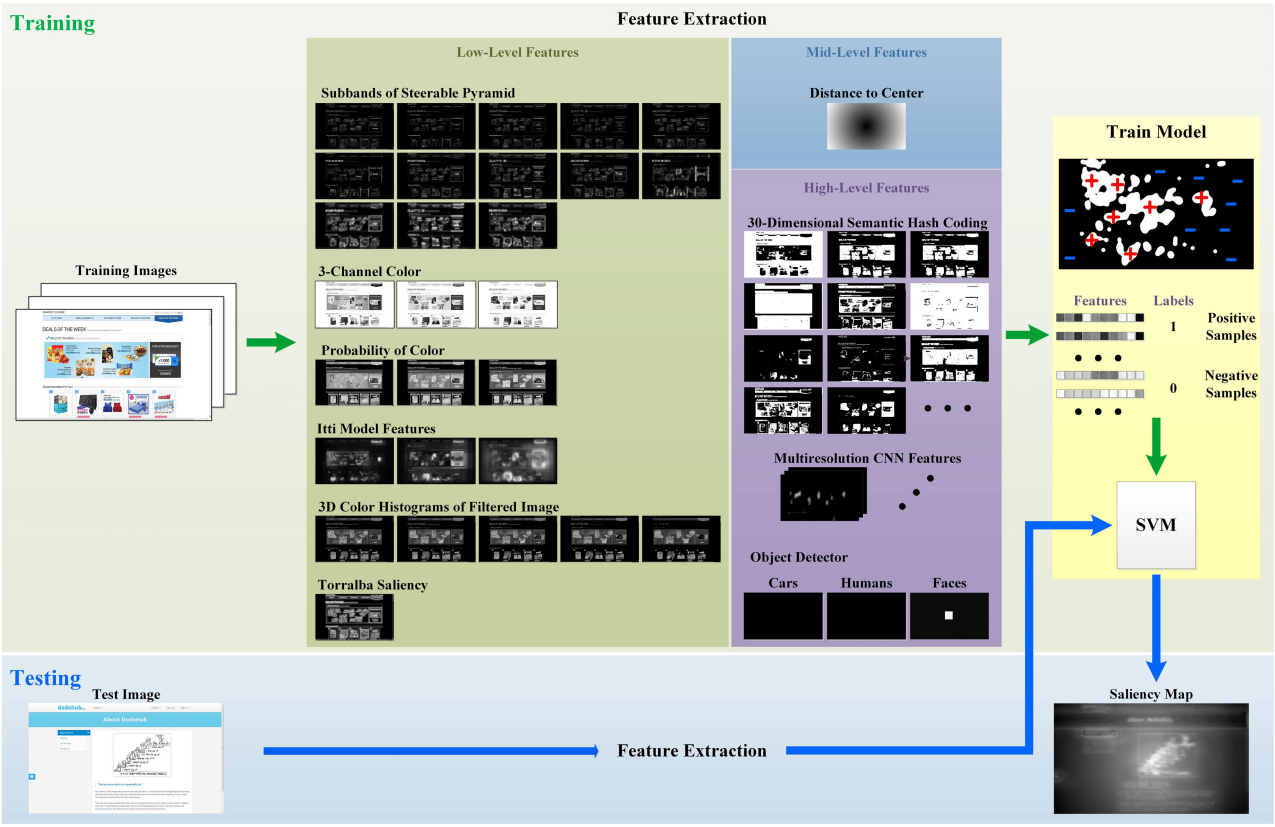


FIGURE 3. Block diagram of saliency estimation for webpages. We first extract a set of visual features from training images. Then, we choose positive and negative samples separately from the most salient (top 20% of the human heatmap) and nonsalient areas (bottom 30% of the human heatmap) to train the parameters of SVM. Finally, we use the learned model to predict the saliency map of a test image.

5) 3D COLOR HISTOGRAMS OF FILTERED IMAGE

The probability of each color according to 3D color histograms of the image filtered with a median filter under five scales.

6) TORRALBA SALIENCY

The saliency value calculated by [34], which defines saliency as the difference between the target velocity and the average of distractors.

7) DISTANCE TO CENTER

For mid-level features, we compute the distance to the center of images because important figures tend to be placed in the center of webpages [35].

8) SEMANTIC HASHING CODE

Semantic hashing, as an effective means to find a compressed representation of high-dimensional input data, has been successively applied in multiple fields. The semantic hashing is in fact to convert raw data from images to short binary code. The code can organize the data into a memory space where nearby addresses store the pointers of similar semantic objects [36]. To obtain binary code, an autoencoder network as Fig. 4 is used for feature learning. As can be seen from the figure, the network is a reconstruction network, which has symmetrical encoder and decoder. The input first passes through the encoder with the gradually decreasing number of units to obtain short binary code. Then, the decoder generates the output based on the binary code in the central layer.

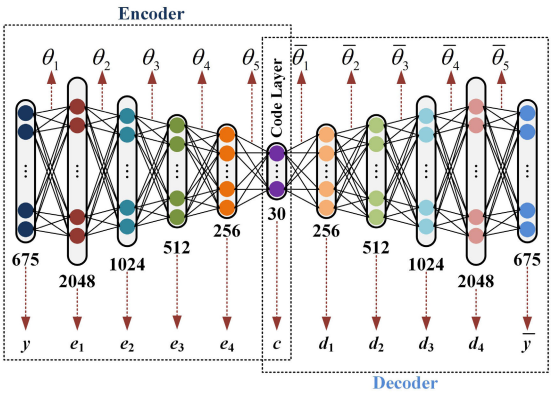


FIGURE 4. Autoencoder for extracting semantic hashing code. The network consists of symmetrical encoder and decoder.

For high-level features, we combine the methodologies of semantic hashing, multiresolution CNN, and object detection for feature extraction.

Concretely, as is shown in Fig. 4, given an input vector $\boldsymbol{y}$, we can derive the output of the first hidden layer in the

encoder as:

$$\boldsymbol{e}_1 = \text{sigmoid}(\boldsymbol{W}_1\boldsymbol{y} + \boldsymbol{b}_1), \tag{4}$$

where sigmoid is the sigmoid function, and $\boldsymbol{\theta}_1 = \{\boldsymbol{W}_1, \boldsymbol{b}_1\}$ is the parameter of the first hidden layer in the encoder, with weight of $\boldsymbol{W}_1$ and bias of $\boldsymbol{b}_1$. Then, $\boldsymbol{e}_1$ iteratively passes through the following hidden layers in the encoder to generate a vector as the output of the fifth hidden layer. To generate binary code $\boldsymbol{c}$, the values in the vector are rounded to 1 or 0 in the forward pass while this rounding is ignored in the back-propagation. In this way, noise can be introduced in the network to make the model more robust [36]. Next, in the decoding process, code $\boldsymbol{c}$ first passes through the first hidden layer in the decoder as:

$$\boldsymbol{d}_1 = \text{sigmoid}(\bar{\boldsymbol{W}}_1\boldsymbol{c} + \bar{\boldsymbol{b}}_1), \tag{5}$$

where $\bar{\boldsymbol{\theta}}_1 = \{\bar{\boldsymbol{W}}_1, \bar{\boldsymbol{b}}_1\}$ is the parameter of the first hidden layer in the decoder. After several iterations, the network can generate an output of $\bar{\boldsymbol{y}}$ at the end of the decoder. Based on $\bar{\boldsymbol{y}}$, a loss function can be defined as:

$$L(\boldsymbol{y}, \bar{\boldsymbol{y}}) = -\sum\nolimits_i \left(y_i \log \bar{y}_i + (1 - y_i) \log (1 - \bar{y}_i)\right), \tag{6}$$

where i is the index of elements in the vectors. Finally, for the training set T of all training samples, the total loss function is calculated as:

$$L_{AE}(\boldsymbol{\theta}) = \sum_{\boldsymbol{y} \in T} L(\boldsymbol{y}, \bar{\boldsymbol{y}}). \tag{7}$$

Due to the restricted encoder and the loss function of minimizing the reconstruction error, the central layer can work as a filter to adaptively extract a compressed representation of the input.

To train the network, we sample $d \times d$ unlabeled patches from the data set of webpages [3]. Then, we transform each patch into the form of a vector to feed into the network. With the training samples, we adopt a two-stage learning process to train the parameters of the network [37]. Firstly, we initialize the network layer by layer to obtain a set of sensible initial parameters for the network. We take the encoder as a stack of Restricted Boltzmann Machines (RBMs) to obtain initial values by contrastive divergence and use its transposes to initialize the decoder [14]. Based on the initialization, we use backpropagation to fine-tune the parameters globally according to (7). Finally, with the trained network, we can extract the 30-dimensional binary code of semantic hashing by inputting the patch of the pixel into the network.

9) MULTIRESOLUTION CNN FEATURES

Deep learning has played an important part in the progress of saliency. CNN, which is inspired by the function of visual cells, can capture semantic features of data in a hierarchical way [18], [38]. Therefore, in this study, we also extract CNN features to complement the hierarchically semantic information of images. Concretely, we adopt the trained network in [38] for feature learning. Given a pixel $\boldsymbol{x}$, we first extract $79 \times 79 \times 3$ central patch $\boldsymbol{c}(\boldsymbol{x})$ and $158 \times 158 \times 3$ surrounding patch $\boldsymbol{s}(\boldsymbol{x})$ from the center of $\boldsymbol{x}$. Then, we resize the surrounding patch into $79 \times 79 \times 3$ $\bar{\boldsymbol{s}}(\boldsymbol{x})$ as the size of the input in the second stream. Next, we feed $\boldsymbol{c}(\boldsymbol{x})$ and $\bar{\boldsymbol{s}}(\boldsymbol{x})$ into the input of the first and second streams, respectively. Finally, we combine the output in the third convolutional layers of two streams as CNN features.

10) OBJECT DETECTION

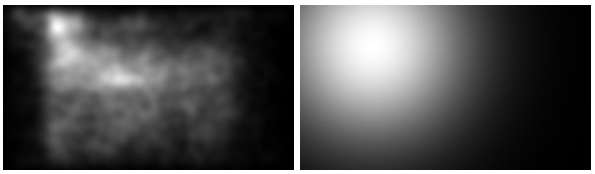
The analyses on the eye movements of webpages [3] have shown that humans tend to pay more attention to the regions of faces and persons. Therefore, we generate binary maps of objects to obtain a robust description of objects under different scenarios. Firstly, we derive the map of cars and persons by car and person detectors implemented by Felzenszwalb *et al.*'s model [39]. Then, we obtain the maps of faces by Viola and Jone's face detector [40].

Based on feature extraction, the next step is to construct the relationship between multilevel features and saliency value for estimating the saliency of test images. To solve this problem, we use SVM to learn from human eye-tracking data on webpages. Firstly, we randomly choose half of the data set for training, with the rest as test images. Then, for each training image, we sample ten positive points from the top 20% of the fixation density map and ten negative points from the bottom 30%. As a result, for each training sample, we can extract the pair of multilevel features and the corresponding label (1 for positive samples and 0 for negative samples, respectively). Finally, we use all pairs of sampled data to train the SVM. In testing, we adopt the method similar to regression for estimating the continuously changing saliency value $\boldsymbol{S}(\boldsymbol{x})$ of each pixel $\boldsymbol{x}$. Specifically, the saliency is calculated by ${\boldsymbol{W}_S}^T\boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{b}_S$ where $\boldsymbol{f}(\boldsymbol{x})$ refers to the features and $\{\boldsymbol{W}_S, \boldsymbol{b}_S\}$ denotes the parameters of the SVM.

B. SACCADIC SCANPATH GENERATION

Saliency estimation can provide an initial probability of being fixated for each pixel to guide the selection of fixation at each moment. However, besides saliency, there are other factors influencing saccade. One critical factor is spatial bias. It means that human eye movements exhibit certain spatial biases on webpages because of the unique properties of webpage images. Therefore, in this subsection, we first combine the spatial influence from distinct perspectives to model the spatial bias for the determination of saccade.

The first is top-left bias. As is shown in Fig. 5, there is a tendency to make saccade towards the top-left regions. Furthermore, previous research has demonstrated that the top-left bias mainly presents in the initial fixations on scanpaths [17]. The main reason is that the upper left corner of images usually contains important information of scenes, such as the headlines. Observers tend to pay more attention to the top-left regions at the initial moment when they try to obtain a general understanding of the scene. To model the spatial influence of top-left bias at initial stages, we generate a top-left bias map to introduce the traction towards top-left regions in the saccade



(a) (b)

FIGURE 5. Top-left bias of fixations on webpages. (a) Fixation density map generated by convolving all fixations in the data set [17] with a 2D Gaussian function. (b) Map of top-left bias.

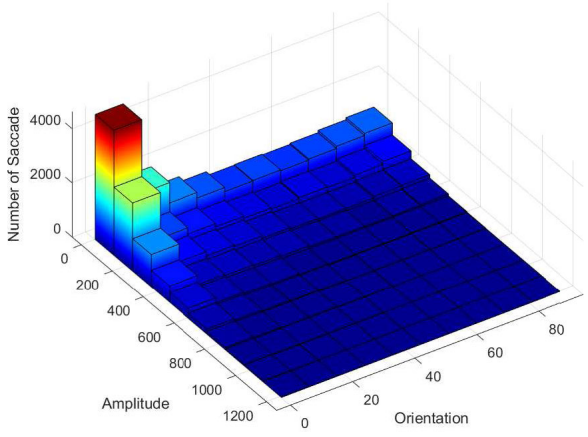


FIGURE 6. Oculomotor bias of saccade on webpages. Saccade amplitude is the spatial distance between two adjacent fixations. Saccade orientation is the absolute value of the angle with the X-axis. 0° is a horizontal saccade while 90° is a vertical saccade.

prediction of the first four fixations, which is calculated as:

$$\mu(\boldsymbol{x}) \propto 1 - \bar{d}(\boldsymbol{x}, \boldsymbol{x}_{tp}), \tag{8}$$

where $\boldsymbol{x}_{tp} = [h/4, w/4]$ refers to the pixel at the top-left region, with h and w as the height and width of an image, respectively. Besides, $\bar{d}(\boldsymbol{x}, \boldsymbol{x}_{tp})$ denotes the normalized distance from the pixel $\boldsymbol{x}_{tp}$. Consequently, the generated top-left bias map is shown in Fig. 5b.

The second is the oculomotor bias represented by saccade amplitude and saccade orientation. To investigate the oculomotor bias in free-viewing webpages, we analyze all the eye movement data of the eye-tracking database [3]. As is shown in Fig. 6, we plot the two-dimensional histogram of saccade amplitude and saccade orientation. Concretely, for saccade amplitude, we calculate the spatial distance between every two successive fixations. For saccade orientation, we compute the absolute value of the angle between the X-axis and each gaze shift. Therefore, a value close to 0° indicates that the saccade tends to be horizontal. Otherwise, a value close to 90° refers to a vertical saccade.

As can be seen from Fig. 6, observers tend to generate short and horizontal saccade when free-viewing webpages. From another perspective, the oculomotor bias has demonstrated that the determination of the current fixation is influenced by the position of the last fixation [28]. Therefore, we define an oculomotor bias map that depends on the last fixation to enhance the probability of short and horizontal saccade. To be specific, to generate the oculomotor bias map, we use a two-dimensional Gaussian function, which is calculated as:

$$\rho_t(\boldsymbol{x}) \propto \exp\left[-\left(\frac{(x - x_{t-1})^2}{2\sigma_x^2} + \frac{(y - y_{t-1})^2}{2\sigma_y^2} \right) \right], \tag{9}$$

where $[x, y]$ and $[x_{t-1}, y_{t-1}]$ denote the two-dimensional coordinates of the pixel $\boldsymbol{x}$ and the last fixation $\boldsymbol{q}_{t-1}$, respectively. Besides, we set asymmetric standard deviations with $\sigma_x = \min[w, h]/3$ and $\sigma_y = \sigma_x \times 4$ to encourage horizontal saccade. Finally, based on the top-left bias and the oculomotor bias, the map of spatial bias can be computed as:

$$\boldsymbol{W}_t(\boldsymbol{x}) = \begin{cases} \dfrac{\mu(\boldsymbol{x})}{\sum_{\boldsymbol{x}} \mu(\boldsymbol{x})} \dfrac{\rho_t(\boldsymbol{x})}{\sum_{\boldsymbol{x}} \rho_t(\boldsymbol{x})} & \text{if } t \leqslant 4 \\ \dfrac{\rho_t(\boldsymbol{x})}{\sum_{\boldsymbol{x}} \rho_t(\boldsymbol{x})} & \text{otherwise} \end{cases} \tag{10}$$

Besides spatial bias, IoR mechanism is another important factor in dynamic attention [2], [28]. Its essence is to prevent the following shifts returning to previously attended regions in a period. Previous research has shown that each saccade takes 30-70 ms while the inhibition of each local region lasts approximately 500-900 ms [2]. It means that in the generation of ten fixations, each previously fixated region would be inhibited to return. To model the IoR mechanism on saccade, we calculate the IoR map $\boldsymbol{I}_t(\boldsymbol{x})$ at stage t as:

$$\boldsymbol{I}_t(\boldsymbol{x}) = \begin{cases} 0 & \text{if } d(\boldsymbol{x}, \boldsymbol{q}_{t-1}) \leqslant r \\ \boldsymbol{I}_{t-1}(\boldsymbol{x}) & \text{otherwise} \end{cases} \tag{11}$$

where $d(\boldsymbol{x}, \boldsymbol{q}_{t-1})$ is the spatial distance between pixel $\boldsymbol{x}$ and the last fixation $\boldsymbol{q}_{t-1}$. Since we focus on modeling the saccade in a free-viewing process, similar to [2], we set the local region of each fixation as a simple disk with radius of $r = \min[w, h]/12$.

After modeling the multiple factors, we integrate the influence of saliency, spatial bias, and IoR into an integrated map $\boldsymbol{F}(\boldsymbol{x})$ as:

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{S}(\boldsymbol{x}) \boldsymbol{W}_t(\boldsymbol{x}) \boldsymbol{I}_t(\boldsymbol{x}). \tag{12}$$

Then, we take the maximum on the map $F(\boldsymbol{x})$ as the current fixation $\boldsymbol{q}_t$, which is expressed as:

$$\boldsymbol{q}_t = \arg\max_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{x}). \tag{13}$$

Based on the selection of the current fixation $\boldsymbol{q}_t$, we update the spatial bias map and IoR map in the following calculation. After several iterations of prediction and updating, we can finally generate a sequence of fixations and construct a saccadic scanpath on each webpage image. An example of the generation of successive fixations based on the integration of multiple factors can be found in Fig. 7.

IV. EXPERIMENTAL RESULTS

In this section, firstly, we describe the comparison models and implementation details in IV-A. Then, we introduce three

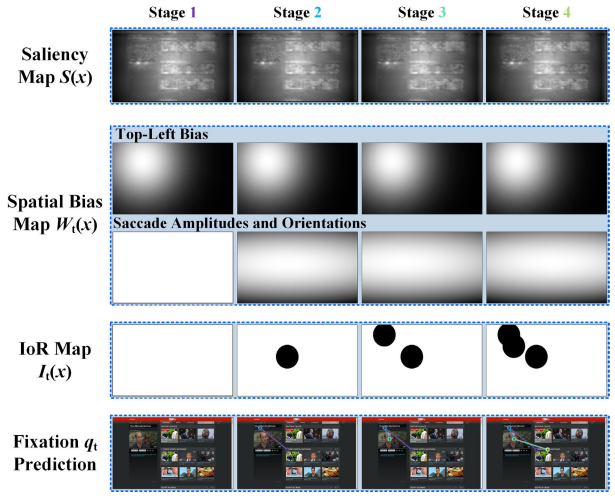


FIGURE 7. Generation of a sequence of fixations by integrating the influence of saliency, spatial bias, and IoR.

benchmark evaluation metrics for saccadic scanpath prediction in IV-B. Finally, we evaluate different combinations of saccadic factors and compare our predicted scanpaths on webpages with the state-of-the-art models in IV-C.

A. EXPERIMENTAL SETTINGS

1) DATA SET

In this study, we adopted the eye-tracking data set of webpages proposed by Shen and Zhao [3] for the evaluation of saccadic models. It includes 149 webpage images of three categories, namely pictorial, text, and mixed. Based on the images, they recorded the eye movements of 11 subjects with age from 21 to 25. Subjects were positioned at a distance of 0.6m from a screen of 1360 $\times$ 768 pixels. Besides, each scene was presented for 5s in the collection of eye-tracking data.

2) COMPUTATIONAL SACCADIC MODELS

To verify the validity of the proposed saccadic model, we compared the prediction of the proposed model with the results of the state-of-the-art saccadic models. Firstly, we compared the proposed model with Itti *et al.*'s attention model [2], which applied WTA and IoR on saliency maps for scanpath generation. Secondly, we compared the proposed model with Boccignone and Ferraro's model [41] of Constrained Levy Exploration (CLE). Thirdly, we compared the proposed model with Le Meur and Liu's saccadic model [28], which combined saliency estimation and oculomotor bias for scanpath prediction. Fourthly, we compared the proposed model with Xia *et al.*'s iterative representation learning (IRL) model [7] to output scanpaths based on a self-taught learning framework.

Moreover, human performance should be one of the benchmarks for [7]. Therefore, we also evaluated the inter-observer performance. Concretely, given an evaluation metric, we first evaluated each human scanpath by taking other human scanpaths on the same image as the ground truth. Then, we averaged the results across all human scanpaths and all images to measure the inter-observer performance "IO".

3) IMPLEMENTATION DETAILS

To obtain the semantic hashing code, we first extracted 200,000 15 $\times$ 15 unlabeled patches from webpages to train the autoencoder in Fig. 4, with the dimension of each input vector as 675 $=$ 15 $\times$ 15 $\times$ 3. Then, we adopted the network with the layer of size 675-2048-1024-512-256-30 in the encoder. For the network, the code units in the central layer are binary neurons, while the others are logistic neurons. In the two-stage training, we utilized 100 mini-batches with 50 epochs and 200 with 100 epochs, respectively. Besides, we updated the weights after each mini-batch with a learning rate of 0.001 and 0.1 for the central RBM and other RBMs in the encoder, respectively. To train the SVM model, we took half of the data set as training images and the other half as test images.

B. EVALUATION METRICS

Unlike saliency estimation, there has been a lack of consensus about the evaluation metrics for scanpath prediction [7]. Therefore, to fairly compare models, we adopted multiple saccadic metrics to evaluate the performance of models from different perspectives.

1) TIME-DELAY EMBEDDING (TDE)

TDE is a piece-based metric. It was first introduced to the evaluation of saccadic scanpath prediction by Wang *et al.* in a work of scanpath prediction on natural images [26]. Its calculation consists of three main steps. Firstly, we divided the predicted scanpath and all human scanpaths into saccadic pieces of k fixations. For instance, $\boldsymbol{C}_p^k(t) = (\boldsymbol{c}_p(t), \cdots, \boldsymbol{c}_p(t+k-1))$ is a k-dimensional saccadic piece with $\boldsymbol{c}_p(t)$ denoting the t-th predicted fixation. As a result, we can obtain the sets of predicted scanpath ($\boldsymbol{S}_p^k$) and human data ($\boldsymbol{S}_h^k$) by dividing the predicted scanpath and all human scanpaths into k-dimensional saccadic pieces, respectively.

Then, for each saccadic piece $\boldsymbol{C}_p^k(t) \in \boldsymbol{S}_p^k$, we calculated its distance $d_k(t)$ to the set of human scanpaths $\boldsymbol{S}_h^k$ as:

$$d_k(t) = \min_{i,\tau} \left\| \boldsymbol{C}_p^k(t) - \boldsymbol{C}_{h_i}^k(\tau) \right\|_2 / k, \boldsymbol{C}_{h_i}^k(\tau) \in \boldsymbol{S}_h^k, \quad (14)$$

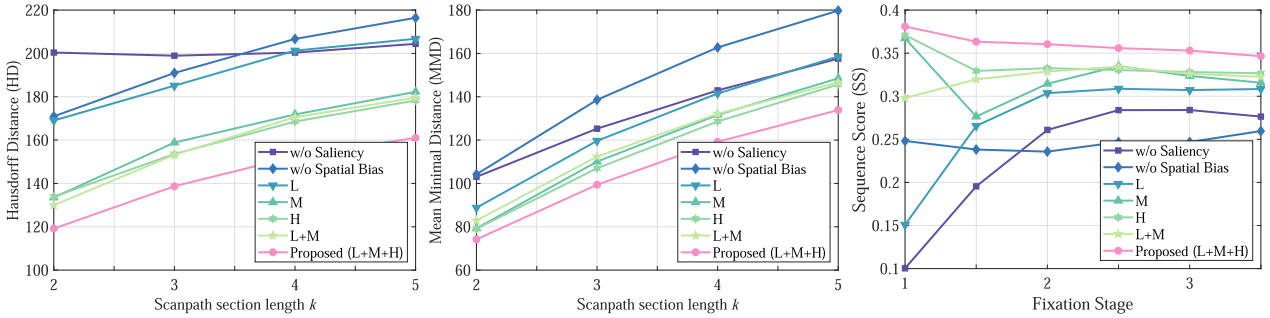
where $\boldsymbol{C}_{h_i}^k(\tau)$ refers to the saccadic piece extracted from the time τ of i-th observer's scanpath.

Finally, we adopted Hausdorff distance (HD) and mean minimal distance (MMD) to measure the total similarity between the pair of predicted and actual human scanpaths. Concretely, HD was defined as the maximum among the distances of all the pieces on the predicted scanpath:

$$D_k^1 = \max_t d_k(t). \quad (15)$$

Besides, MMD was defined as the mean minimal distance:

$$D_k^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} d_k(t), \quad (16)$$



(a) HD of distinct factors and features (b) MMD of distinct factors and features (c) SS of distinct factors and features

FIGURE 8. Comparison on different combinations of factors and features under TDE (HD and MMD) and SS. HD and MMD are divergence measures (should be minimized), while SS is a similarity measure (should be maximized).

where n_k denotes the total number of k-dimensional saccadic pieces included in the prediction. It should be noted that in the experiment we varied k from 2 to 5 as [26] to investigate the performance of models with the outputted saccadic scanpath of distinct lengths.

2) SEQUENCE SCORE (SS)

SS is a string-based metric. It was proposed by Borji *et al*. [42] to convert the scanpaths into strings for comparison. Its calculation consists of three main steps. Firstly, we utilized mean-shift to cluster the fixations, which can better integrate global information of the scene into evaluation than classifying fixations according to grids. Then, we assigned a set of characters to each cluster and converted the scanpaths into strings by determining the cluster of each fixation. Finally, we adopted the Needleman-Wunsch algorithm to calculate the similarity between each pair of predicted and human scanpaths. Based on the pair-wise similarity calculation, we can derive the total measurement of SS by averaging the results across all human scanpaths on each image and all images in the data set. Besides, similar to TDE [26], we varied the length of scanpath for comparison from 1 to 6 by truncating subsequent fixations to evaluate the model at different stages of scanpath generation.

3) MULTIMATCH (MM)

MM is a vector-based metric. It was proposed by Jarodzka and Holmqvist [43] and has been widely used in the recent challenges of saccadic scanpath prediction [44]. Its calculation also includes three main steps. Firstly, we encoded the scanpaths into geometrical vectors. Each saccade can present the shift between two fixations. By taking the fixations as the points in the two-dimensional coordinate system, saccade can be regarded as vectors for dissimilarity comparison. Concretely, for two scanpaths $\boldsymbol{s}_1 = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_N)$ and $\boldsymbol{s}_2 = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_M)$, we computed the distance between the vectors of saccades $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$. In the calculation of distance, the following measures can be used: Position (the distance in the locations), Amplitude (the distance in saccade amplitude), Direction (the distance in saccade orientation), and Shape (the distance in vector representation). Then, we utilized dynamic programming to align scanpaths based on their shapes temporally. Concretely, we computed a cost matrix with each entry representing the distance between pair-wise vectors. After that, we adopted a dynamic programming approach to calculate the minimal accumulating cost in the cost matrix. The final dissimilarity between two scanpaths was reflected by the last entry in the accumulating cost matrix. By traversing all pairs of predicted scanpath and human scanpath in the data set, we can derive the average score under the metric MM.

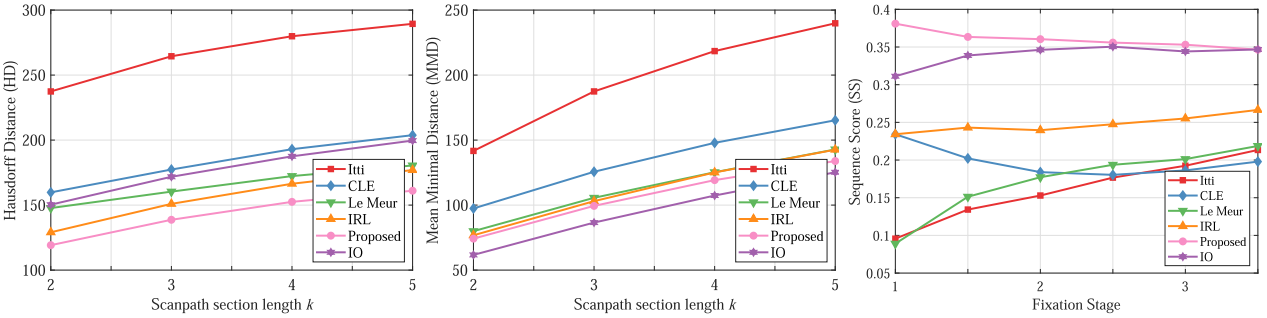
TABLE 1. Comparison on different combinations of factors and features under MM. The best scores are in Bold Face font.

Algorithm	Position	Length	Direction	Shape
w/o Saliency	0.213	0.060	0.130	0.380
w/o Spatial Bias	0.193	0.086	0.187	0.414
L	0.198	0.069	0.143	0.384
M	0.175	0.070	0.130	0.360
H	0.172	0.059	0.123	0.341
L+M	0.174	0.068	0.124	0.356
Proposed (L+M+H)	**0.151**	**0.059**	**0.121**	**0.314**

C. COMPARISON OF SCANPATH PREDICTION

In this subsection, we first evaluated different combinations of saccadic factors based on the metrics in IV-B. Firstly, we constructed the model without the estimation of saliency "w/o Saliency" to predict saccadic scanpaths based on random initial positions and the modeling of spatial bias and IoR. Then, we generated the model "w/o Spatial Bias" to iteratively output fixations based on saliency estimation and IoR. Finally, we compared these models with the proposed model ("Proposed"), which fuses the factors of saliency, spatial bias, and IoR to generate scanpaths on webpages.

The comparison under TDE and SS is shown in Fig. 8, and the results under MM are shown in Table 1. Firstly, the comprehensive comparison across different combinations can show the importance of integrating the influence of saliency and spatial bias in scanpath prediction. Secondly, the comparison under "Position" in Table 1 indicates that saliency estimation helps to provide a more accurate



(a) HD on the data set [3] (b) MMD on the data set [3] (c) SS on the data set [3]

FIGURE 9. Comparison under TDE (HD and MMD) and SS on the eye-tracking data set of webpage [3]. "IO" refers to the inter-observer performance.

prediction for the position of the fixations. Similarly, the significant improvement made by the proposed model at initial stages (e.g., fixation stage 1 under SS in Fig. 8c) can also demonstrate the effectiveness of estimating the initial distribution of fixations by saliency estimation. In other words, static saliency and dynamic saccade are not independent, but coherent mechanisms which interpret the visual behaviors. Thirdly, by comparing the evaluation under "w/o Spatial Bias" and "Proposed", we can conclude that modeling spatial bias is an effective means to obtain scanpaths more consistent with human data.

On the other hand, we discussed an ablation analysis to investigate the influence of different levels of features. The results under distinct combinations of features are shown in Fig. 8 and Table 1. In the comparison, "L", "M", and "H" refer to the models using low-level, mid-level, and high-level features, respectively. As can be observed from the results, the model "H" outperforms models "L" and "M", which illustrates the contribution of high-level features. Also, the comparison between "L+M" and the proposed model can show the important role of high-level cues in scanpath prediction. The final model "L+M+H" that integrates multiple levels of features can comprehensively describe the input, thus outperforming other combinations.

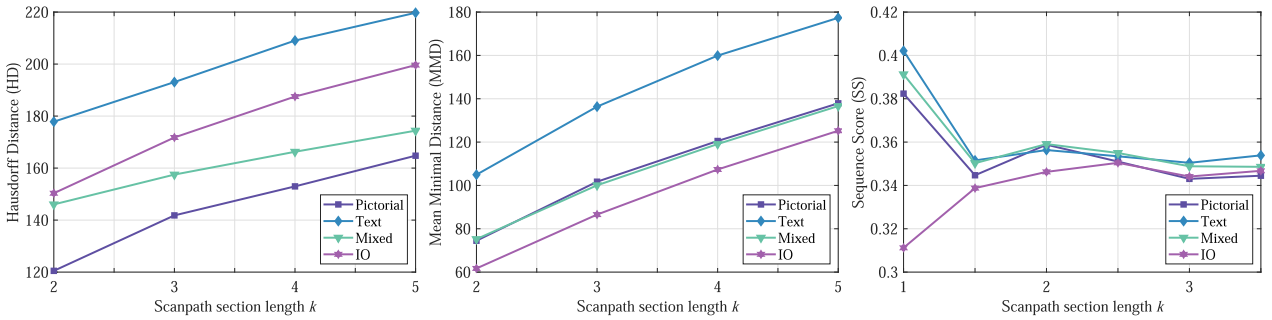
In the second part, we compared the performance of the state-of-the-art models on the data set of webpage [3]. The evaluation under TDE is shown in Figs. 9a and 9b. As can be seen from the results of HD and MMD, the rankings of models are consistent under different values of k. Moreover, the proposed saccadic model can obtain a more accurate prediction of human scanpaths on webpages than other predictive models. It should be noted that the proposed model surpasses the "IO" model under HD. It is mainly because of the limited number of observers for comparison (11 subjects). The upper bound can only be derived by calculating the similarity with the data from an unlimited number of observers [7].

The evaluation under SS is shown in Fig. 9c. By observing the figure, we can draw some conclusions as follows. Firstly, the proposed model built by integrating the learning-based saliency and factors from saccade can outperform other algorithms at different stages of scanpath generation. Secondly, in the first three stages, the proposed model can achieve a better prediction of scanpaths than "IO" model, which indicates the importance of modeling top-left bias. In contrast, with the length of scanpaths increasing, the consistency among human data is gradually enhanced. Therefore, the predictive models still have room for improvement by better exploring the mechanisms in long-term eye movements. Thirdly, the advantages of the proposed model over the state-of-the-art ones for natural scenes (e.g., "IRL" [7] and "Le Meur" [28]) further demonstrate that it is necessary to model human saccadic behaviors on webpages. Fourthly, an overall comparison with the SS scores under natural images [7] can show that human eye movements on webpages is less consistent than those on natural images. It is because that webpage images usually include multiple figures or objects. There are more possible saccade orders to extract information.

TABLE 2. Performance of MM on the test data set [3]. Four properties were used for evaluation. The best scores are in Bold Face font.

Algorithm	Position	Length	Direction	Shape
Itti [2]	0.260	0.154	0.161	0.506
CLE [41]	0.222	0.065	0.183	0.449
Le Meur [28]	0.211	0.062	0.149	0.400
IRL [7]	0.193	0.060	0.156	0.395
Proposed-*Pictorial*	0.150	**0.057**	0.121	0.310
Proposed-*Text*	0.151	0.059	0.126	0.312
Proposed-*Mixed*	**0.149**	0.058	0.123	0.312
Proposed	0.151	0.059	0.124	0.314
IO	0.156	0.062	**0.119**	**0.296**

The evaluation under MM is shown in Table 2. It can be observed from the table that the proposed model has advantages over other models. On the one hand, the proposed model can achieve a more accurate prediction of the fixated positions by adopting the feature-fusing learning-based saliency to estimate the initial prohibition of being fixated. On the other hand, by integrating the mechanisms of spatial bias and IoR, the proposed model can also surpass the others in the properties of "Length" and "Direction". As a result, the predicted scanpaths of the proposed model are more consistent with human scanpaths.



(a) HD on the data set [3] (b) MMD on the data set [3] (c) SS on the data set [3]

FIGURE 10. The evaluation under TDE (HD and MMD) and SS on the subsets "Pictorial", "Text", and "Mixed" in data set [3].

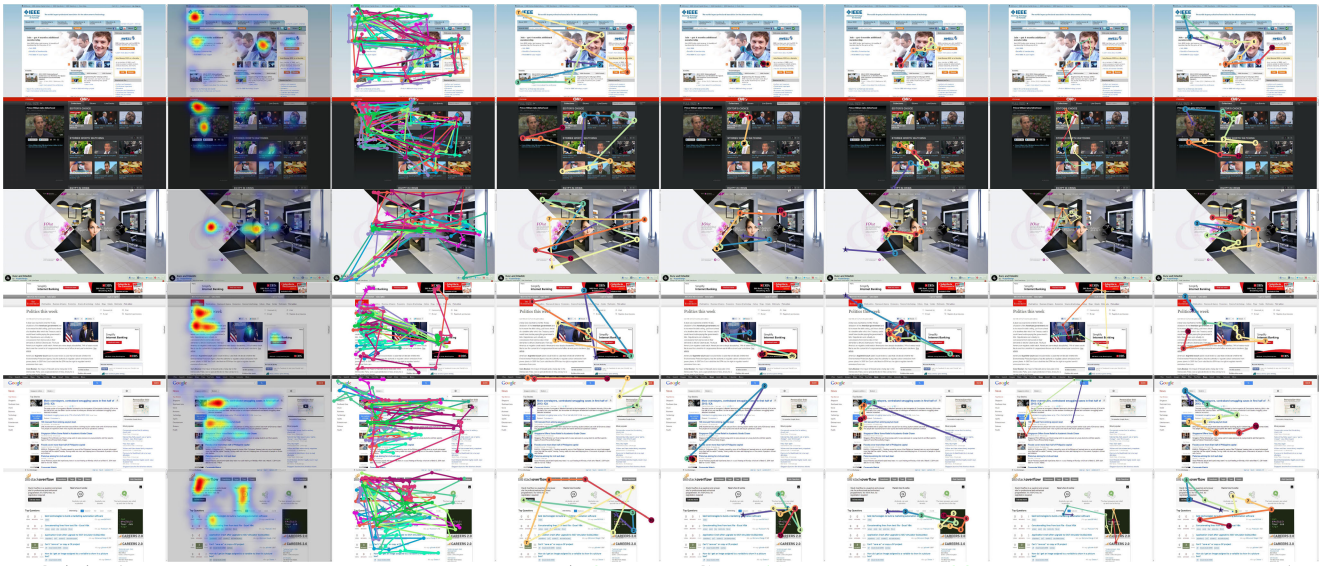


FIGURE 11. Visual comparison of the scanpaths from distinct models. "Heatmap" refers to the distribution of all human fixations. "Human Scanpaths" shows all human scanpaths on each webpage image. For both human and predicted scanpaths, pentagrams are the starting points.

In addition, we also compared the results under each subset. The subsets are ''Pictorial'', ''Text'', and ''Mixed''. The example images of the three subsets can be seen from Fig. 1. ''Pictorial'' includes 50 images, each of which has a dominant picture with less text. ''Text'' includes 50 images with informative text. ''Mixed'' includes 49 images with each of thumbnail pictures and text. The results in Fig. 10 and Table 2 have shown that the proposed model performs best on the subset of ''Pictorial'' but performs worst on ''Text''. It has implied that modeling text representation is important, although texts can be popped out by low-level features, such as intensity and orientation. Therefore, determining how to extract text-based features to improve the saccade prediction on text images will be a future research direction.

Besides quantitative comparison, we also provided a visual comparison of generated scanpaths in Fig. 11. In the figure, the second column ''Heatmap'' is the ground truth of saliency estimation, which can show the distribution of human fixations on each scene. The third column ''Human Scanpaths'' displays all human scanpaths on each image with different colors. As can be observed from the figure, the proposed scanpaths are more similar to human data than other saccadic algorithms. In addition, the comparison between the models with (the proposed method and ''Le Meur'') and without (''Itti'' and ''CLE'') the modeling of oculomotor bias can demonstrate the importance of learning the bias of saccade amplitude and saccade orientation in scanpath generation.

V. CONCLUSION

Despite a large amount of research effort in the last two decades, there are still some limitations in visual attention modeling. For one thing, previous research has usually focused on saliency estimation of natural scenes. The calculation of webpage saliency has still been a challenge. For another, the study on dynamic attention mechanisms is a bit limited [7].

To address these problems, we have proposed a saccadic model for webpages to investigate human dynamic eye movements in free-viewing webpages. In the first stage, we have estimated the initial distribution of fixations based on multilevel saliency learning. In the second stage, we have combined the factors of saliency, top-left bias, oculomotor bias, and IoR to predict fixations for scanpath generation iteratively. Qualitative and quantitative experimental results have demonstrated the advantages of the proposed model over other state-of-the-art methods.

For future work, we will extend this study from the following perspectives. Firstly, we will explore the task-driven dynamic visual attention on webpages. Based on attention modeling in free-viewing conditions, we will investigate the influence of tasks on the generation of saccadic scanpaths. Secondly, besides the dissimilarity of scanpaths presented by targets, we will also focus on the difference in scanpaths presented by distinct groups of subjects to solve the classification problems in multiple applications, such as disease identification and age recognition [45]. Thirdly, we will integrate the methodologies of deep learning into the estimation of saliency and the modeling of saccadic properties. On the one hand, we will build a large-scale eye-tracking data set of webpages to complement the lack of large-scale webpage-based data set. On the other hand, we will take advantage of the learning ability of the deep models to reveal deeper dynamic attention mechanisms from human eye movements.

REFERENCES

[1] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.
[2] 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.
[3] C. Shen and Q. Zhao, "Webpage saliency," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Sep. 2014, pp. 33–46.
[4] Q. Zheng, J. Jiao, Y. Cao, and R. W. Lau, "Task-driven webpage saliency," in *Proc. 15th Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 287–302.
[5] J. Li, L. Su, B. Wu, J. Pang, C. Wang, Z. Wu, and Q. Huang, "Webpage saliency prediction with multi-features fusion," in *Proc. 23rd IEEE Int. Conf. Image Process.*, Phoenix, AZ, USA, Sep. 2016, pp. 674–678.
[6] Y. Li and Y. Zhang, "Webpage saliency prediction with two-stage generative adversarial networks," 2018, *arXiv:1805.11374*. [Online]. Available: https://arxiv.org/abs/1805.11374
[7] C. Xia, J. Han, F. Qi, and G. Shi, "Predicting human saccadic scanpaths based on iterative representation learning," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3502–3515, Jul. 2019.
[8] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.
[9] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.
[10] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2005, pp. 155–162.
[11] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 438–445.
[12] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.
[13] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 478–485.
[14] C. Xia, F. Qi, and G. Shi, "Bottom–up visual saliency estimation with deep autoencoder-based sparse reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1227–1440, Jun. 2016.
[15] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley, "Image saliency: From intrinsic to extrinsic context," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 417–424.
[16] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. 27th IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2798–2805.
[17] C. Shen and Q. Zhao, "Learning to predict eye fixations for semantic contents using multi-layer sparse network," *Neurocomputing*, vol. 138, no. 11, pp. 61–68, Aug. 2014.
[18] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May. 2018.
[19] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 409–416.
[20] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5025–5034, Nov. 2016.
[21] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.
[22] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.
[23] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.
[24] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Trans. Visual. Comput. Graph.*, vol. 23, no. 8, pp. 2014–2027, Aug. 2017.
[25] T. S. Lee and S. X. Yu, "An information-theoretic framework for understanding saccadic eye movements," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, Dec. 1999, pp. 834–840.
[26] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating human saccadic scanpaths on natural images," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 441–448.
[27] H. Liu, D. Xu, Q. Huang, W. Li, M. Xu, and S. Lin, "Semantically-based human scanpath estimation with hmms," in *Proc. 14th IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3232–3239.
[28] O. Le Meur and Z. Liu, "Saccadic model of eye movements for free-viewing condition," *Vis. Res.*, vol. 116, pp. 152–164, Nov. 2015.
[29] O. Le Meur and A. Coutrot, "Introducing context-dependent and spatially-variant viewing biases in saccadic models," *Vis. Res.*, vol. 121, pp. 72–84, Apr. 2016.
[30] Y. Wu and Z. Chen, "Saliency map generation based on saccade target theory," in *Proc. 18th IEEE Int. Conf. Multimedia Expo*, Hong Kong, Jul. 2017, pp. 529–534.
[31] C. Shen, X. Huang, and Q. Zhao, "Predicting eye fixations on webpage with an ensemble of early features and high-level representations from deep network," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2084–2093, Nov. 2015.
[32] Y. Zheng, F. Zhou, L. Li, X. Bai, and C. Sun, "Mutual guidance-based saliency propagation for infrared pedestrian images," *IEEE Access*, vol. 7, pp. 113355–113371, 2019.
[33] M. Spratling, "Predictive coding as a model of the V1 saliency map hypothesis," *Neural Netw.*, vol. 26, pp. 7–28, Feb. 2012.
[34] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[35] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep. 2009, pp. 2106–2113.
[36] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval," in *Proc. 19th Eur. Symp. Artif. Neural Netw.*, Bruges, Belgium, Apr. 2011.
[37] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[38] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 392–404, Feb. 2018.
[39] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2009.
[40] P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 4, no. 4, pp. 34–47, 2001.
[41] G. Boccignone and M. Ferraro, "Modelling gaze shift as a constrained random walk," *Phys. A, Stat. Mech. Appl.*, vol. 331, nos. 1–2, pp. 207–218, Jan. 2004.
[42] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *Proc. 14th IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 921–928.
[43] H. Jarodzka and K. Holmqvist, "A vector-based, multidimensional scanpath similarity measure," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Austin, TX, USA, Mar. 2010, pp. 211–218.
[44] J. Gutiérrez, E. David, Y. Rai, and P. Le Callet, "Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still images," *Signal Process., Image Commun.*, vol. 69, pp. 35–42, Nov. 2018.
[45] O. Le Meur, A. Coutrot, Z. Liu, P. Rämä, A. Le Roch, and A. Helo, "Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4777–4789, Oct. 2017.

CHEN XIA received the B.Eng. and Ph.D. degrees from Xidian University, Xi'an, China, in 2010 and 2017, respectively. She is currently an Assistant Professor with the School of Automation, Northwestern Polytechnical University, Xi'an. Her research interests include computer vision, deep learning, saliency estimation, and saccadic scanpath prediction.

RONG QUAN received the B.S. degree in information engineering from Northwestern Polytechnical University, in 2013, where she is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with the School of Automation. Her research interests include computer vision, machine learning, deep reinforcement learning, object co-segmentation, visual saliency detection, and human scanpath prediction.