Learning to Predict Sequences of Human Visual Fixations

Ming Jiang, *Student Member, IEEE*, Xavier Boix, *Student Member, IEEE*, Gemma Roig, *Student Member, IEEE*, Juan Xu, Luc Van Gool, *Senior Member, IEEE*, and Qi Zhao, *Member, IEEE*

***Abstract*— Most state-of-the-art visual attention models estimate the probability distribution of fixating the eyes in a location of the image, the so-called saliency maps. Yet, these models do not predict the temporal sequence of eye fixations, which may be valuable for better predicting the human eye fixations, as well as for understanding the role of the different cues during visual exploration. In this paper, we present a method for predicting the sequence of human eye fixations, which is learned from the recorded human eye-tracking data. We use least-squares policy iteration (LSPI) to learn a visual exploration policy that mimics the recorded eye-fixation examples. The model uses a different set of parameters for the different stages of visual exploration that capture the importance of the cues during the scanpath. In a series of experiments, we demonstrate the effectiveness of using LSPI for combining multiple cues at different stages of the scanpath. The learned parameters suggest that the low-level and high-level cues (semantics) are similarly important at the first eye fixation of the scanpath, and the contribution of high-level cues keeps increasing during the visual exploration. Results show that our approach obtains the state-of-the-art performances on two challenging data sets: 1) OSIE data set and 2) MIT data set.**

***Index Terms*— Scanpath prediction, visual saliency prediction.**

I. Introduction

HUMANS and other primates shift their gaze to allocate processing resources to a subset of the visual input. Understanding and emulating the way that human observers free-view a natural scene have attracted much interest [1]–[3].

Manuscript received November 30, 2014; revised July 6, 2015 and October 23, 2015; accepted October 25, 2015. Date of publication January 7, 2016; date of current version May 16, 2016. This work was supported in part by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant R-263-000-B32-112, in part by the Defense Innovative Research Programme under Grant 9014100596, and in part by the ERC Advanced Grant VarCity. *(Corresponding author: Qi Zhao.)*

M. Jiang, J. Xu, and Q. Zhao are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: mjiang@nus.edu.sg; jxu@nus.edu.sg; eleqiz@nus.edu.sg).

X. Boix is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, also with the Computer Vision Laboratory, ETH Zurich, Zürich 8092, Switzerland, and also with the Laboratory for Computational and Statistical Learning, Center for Brains, Minds and Machines, Istituto Italiano di Tecnologia, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: xboix@mit.edu).

G. Roig is with the Computer Vision Laboratory, ETH Zurich, Zürich 8092, Switzerland, and also with the Laboratory for Computational and Statistical Learning, Center for Brains, Minds and Machines, Istituto Italiano di Tecnologia, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: gemmar@mit.edu).

L. Van Gool is with the Computer Vision Laboratory, ETH Zurich, Zürich 8092, Switzerland (e-mail: vangool@vision.ee.ethz.ch).

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/TNNLS.2015.2496306

Many computational models of visual attention aim at predicting the probability of fixating the eyes in a location of the visual scene, i.e., the saliency map. However, the problem of predicting the sequence of eye fixations remains considerably unexplored. The prediction of the sequence of eye fixations might be more involved than predicting the saliency map, since the temporal order of the fixations is discarded in a saliency map.

Computational saliency models for predicting the sequence of eye fixations are inspired by biological systems. Several psychological studies investigated the strategy underlying the temporal sequence of eye movements [4]–[6]. These studies introduced several hypotheses that may explain the strategies followed by the observers, but these models were not intended to predict the sequence of eye fixations. It was not until recently that the first computational models that predict the sequence of eye fixations in an unseen image were introduced [7], [8]. These methods introduced multiple cues extracted from the image that are useful for scanpath prediction. The results achieved by these methods are a remarkable feat.

In this paper, we introduce a model to combine different cues to predict the human visual scanpath. Our model dynamically combines the input cues by changing the contribution of each cue over the temporal sequence of the visual scanpath. This allows analyzing the temporal evolution of the importance of the different cues during visual exploration.

The parameters of our model are learnt from the examples of recorded human eye fixations, using least-squares policy iteration (LSPI). LSPI is a technique for reinforcement learning that we use to mimic the human visual scanpath. Reinforcement learning has been previously used to learn the models of visual attention to improve some computer vision and robotics tasks, such as object, action, and face recognition [9]–[11], visual search in surveillance [12], and autonomous navigation [13], [14], among others. We use similar techniques as some of these methods, but in our case, we aim at mimicking human eye fixations rather than using visual attention to improve a specific task.

In a series of experiments, namely, in the OSIE data set [15] and the MIT data set [16], we show that LSPI is able to effectively learn to combine cues for predicting the visual scanpath. The learned parameters suggest that the low-level and high-level cues are similarly important at the first eye fixation of the scanpath, while the weights of the high-level cues keep increasing with time. LSPI achieves better prediction of the

human visual scanpath than the heuristics based on generating scanpath predictions from saliency maps. In addition, we show that the cues we combine with LSPI outperform the current state-of-the-art methods [8].

II. Toward Visual Scanpath Prediction

In this section, we briefly revisit the literature on saliency models and the prediction of scanpaths in eye fixation, which is the main goal of this paper.

A. Visual Saliency Models

Early studies on computational saliency models are rooted in feature-integration theory [17]. Koch and Ullman [1] proposed a computational architecture based on this theory, and the first complete implementation and verification was done by Itti *et al.* [18]. In the last decade, a large number of computational models have been proposed following a similar framework [19], and also, a rich variety of theories and approaches have been introduced for the integration and design of features [20]–[24]. These models combine low-level features, such as color, intensity, and orientation, at numerous spatial scales. Several recent works show that semantic information can considerably boost the performance of these models [16], [25].

Another strand of research investigates the problem of finding the salient objects or regions in the image [26]–[29]. This should not be confused with the aforementioned works that predict the saliency map [30], i.e., the probability of the locations in the image where a human observer may fixate the eye. In this paper, in contrast to saliency map prediction or salient object detection, we aim at predicting the temporal sequence of eye fixations.

B. Visual Scanpath Prediction

Several visual saliency models generate sequences of eye fixations from saliency maps [18], [19]. These models cannot exploit the intrinsic temporal information of the visual scanpath because the saliency maps are static. Using temporal dynamics may lead to better accuracy prediction of the visual scanpath, and may reveal the influence of the different cues during visual exploration.

The pioneering work by Lee and Yu [4] introduced the principle of information maximization for understanding the visual scanpath. Inspired by this work, Renninger *et al.* [31] implemented a visual scanpath predictor for shape silhouettes. However, these models were not introduced for natural or realistic images.

It was not until recently that Wang *et al.* [7] introduced the first human visual scanpath predictor for natural images, which was based on the information maximization criteria. Later, Sun *et al.* [32] proposed to generate a scanpath by sequentially obtaining super-Gaussian component (SGC) and selecting fixations with maximum SGC responses. More recently, Liu *et al.* [8] improved the predictive accuracy in [7] by introducing semantics and transition probabilities in the model. This work can be seen as complementary to these previous methods of visual scanpath prediction. We focus on learning how to dynamically combine different input cues for visual scanpath prediction. Thus, the cues introduced in [7] and [8] could be integrated in the feature combination framework we propose. Yet, we show that the cues we use outperform the results of previous works.

Recently, Mathe and Sminchisescu [33] introduced a scanpath predictor based on learning a model from the examples of human eye fixations. Note that our model is also learnt from examples, but the model we propose has several advantages over that in [33]. Our model uses a different set of weights for each stage of the visual scanpath, which can adjust the influence of each cue at the different stages of the prediction of the scanpath. In addition, our model directly relates the cues with the policy that predicts the scanpath, and allows analyzing the importance of each cue during the scanpath. This is not the case in Mathe and Sminchisescu model [33], since it learns the parameters of the reward rather than the policy, and it becomes involved to relate the importance of each parameter to the final prediction of the scanpath.

III. Human Visual Scanpath as a Markov Decision Process

In this section, we introduce a Markov decision process (MDP) to model the human visual scanpath. This will serve as the basis for our learning algorithm that we introduce in the following Section IV.

Formally, an MDP is represented with the tuple $(\mathcal{S}, \mathcal{A}, r, P)$ [34]. $\mathcal{S}$ is the set of possible states of the system, and a state $s \in \mathcal{S}$ encodes the current situation of the system. In particular, for visual scanpath prediction, the state represents all the information gathered through the visual exploration of the image, such as the locations of previous eye fixations, and features of the current visual exploration extracted from the image. $\mathcal{A}$ is the set of actions that the MDP can take at each stage. An action $a \in \mathcal{A}$ is the location in the image where the gaze will be fixed next. An MDP assumes that to deliver the next action, only the current state is needed.

After taking an action, the system receives a reward, which is denoted as the function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The reward $r(s, a)$ determines how valuable is to be in a certain state after taking an action. Note that the reward is unknown until the action is taken. Thus, at each time step, the MDP is in some state s, it decides taking an action a, the process responds by moving into a new state s', and the reward for being in state s' is evaluated. Since the state s' is uncertain until the action is executed, it is common to use the probability of transitioning to a destination state, s', from the current state, s, when action a has been taken. This probability is denoted as $P(s'|s, a)$.

The transition probability $P(s'|s, a)$ and the aforementioned reward $r(s, a)$ may be difficult to model for the case of eye fixations. We use a reinforcement learning algorithm that does not use a predefined model for $P(s'|s, a)$ and $r(s, a)$ that is introduced in the next section IV. Before that, we now review the common formulation for MDP assuming that we have a model for $P(s'|s, a)$ and $r(s, a)$. An MDP generates the next eye fixation using a policy that at each time step decides what action to take given the current state. Let $\pi : \mathcal{S} \to \mathcal{A}$ be the mapping between the current state and the action to be taken,

the so-called policy. The optimal policy maximizes at each time step the expected reward. This is expressed using the Bellman equation [35]

$$\begin{aligned}\pi(s) &= \arg\max_{a\in\mathcal{A}} Q^{\pi}(s,a) \\ &= \arg\max_{a\in\mathcal{A}} r(s,a) + \gamma\, \mathbb{E}_{s'\sim P}[Q^{\pi}(s',a)] \end{aligned} \tag{1}$$

where $\mathbb{E}_{s'\sim P}[Q(s',a)]$ is the expected value of $Q^{\pi}(s',a)$ over $s' \sim P(s'|a,s)$. Observe that the Bellman equation evaluates the reward of the current state, $r(s,a)$, and the expected reward in the following states after sequentially taking the actions following the policy. $Q^{\pi}(s,a)$ is the so-called value function, and note that it is defined recursively in order to evaluate the reward along the sequence of decision making. γ is the value of the weight of the importance of the expected future rewards, and it is used to alleviate the effects of the uncertainty on the state transitions, $P(s'|a,s)$.

The modeling of the transitioning probability, $P(s'|s,a)$, and the reward function, $r(s,a)$, is of crucial importance to be able to model the human visual scanpath as an MDP. However, true underlying reward and the transition probability of the human model remain unknown. As we mentioned previously, we take an alternative approach, in which we approximate the model by learning a value function that bypasses the definition of the transition probability and the reward function. This is LSPI [36], which allows learning a policy from the recorded human eye fixations that mimics the human behavior.

IV. Learning a Policy From Sequences of Fixations

In this section, we give a general overview of our algorithm. The main aim of our algorithm is to predict a sequence of locations in the image that mimic the sequence of human eye fixations while free viewing an image. During a training phase, the algorithm has access to several recorded sequences of eye-tracking results in natural images, which are used to learn a policy for scanpath prediction.

A common approach in the literature to learn a policy without specifying the reward function and the transition probability of the MDP is to approximate $Q^{\pi}(s,a)$ using a linear projection of a state-action descriptor [36]

$$Q^{\pi}(s,a) \approx \hat{Q}^{\pi}(s,a) = \mathbf{w}^T \phi(s,a) \tag{2}$$

where $\phi(s,a)$ is a vector of features extracted from the state-action pairs that we introduce in Section VI. $\mathbf{w}$ is a vector of parameters for the features in $\phi(s,a)$, which is learned during a training phase, and ^{T} is the transpose operator. Thus, the policy becomes

$$\pi(s) = \arg\max_{a\in\mathcal{A}} \hat{Q}^{\pi}(s,a) = \arg\max_{a\in\mathcal{A}} \mathbf{w}^T \phi(s,a). \tag{3}$$

This maximization can be solved by evaluating $\mathbf{w}^T\phi(s,a)$ among all actions and selecting the one with higher $\hat{Q}^{\pi}(s,a)$, since the number of actions is relatively small in practice. Thus, the execution of the MDP consists of iterating between solving (3) to take a new action and extracting $\phi(s,a)$ for the new state after taking the action. In this way, the definition of $\phi(s,a)$ and the learning of $\mathbf{w}$ determine the policy to generate the eye fixations with the MDP.

In Sections IV-A and IV-B, we first introduce the algorithm to learn the parameter $\mathbf{w}$, and then, we define the state-action mapping $\phi(s,a)$.

A. Least-Squares Policy Iteration

We use the algorithm of policy iteration to learn the policy [36], which has been shown to perform well in other problems with similar settings to ours [14]. LSPI is an iterative procedure that at each iteration i uses the current policy, π_i, to generate a new, improved policy π_{i+1}. At the beginning of the ith iteration of the system, actions are generated from the current policy, π_i, which generate new sequences of eye fixations that will be used for further improving the policy. Recall that the parameter of the policy is $\mathbf{w}$, and hence, we use $\mathbf{w}_i$ to denote the parameters at iteration i. The generated sequences are composed by the state-action mapping before executing the action, the state-action mapping after executing the action, and the received reward, which are defined as $\phi(s^n,a^n)$, $\phi(s'^n,\pi(s'^n))$, and $r(s^n,a^n)$, respectively. We use the superindex n to index the different sequences.

The fitting of a new policy at the ith iteration of the system, from a set of generated sequences, is done using the approximation of $Q^{\pi}(s,a)$ in (2) that we previously introduced. Then, we find the fixed point of the Bellman equation [37], [38], and the optimization problem becomes the solution of the following point process equation:

$$\begin{aligned}\mathbf{w}_{i+1} = \arg\min_{\mathbf{u}} \sum_{n}^{N} \|\mathbf{u}^T\phi(s^n,a^n) - r(s^n,a^n) \\ -\gamma\, \mathbf{w}_{i+1}^T \phi(s'^n,\pi(s'^n))\|^2 \end{aligned} \tag{4}$$

where N is the number of generated sequences. Note that (4) aims at approximating the Bellman equation through least squares, and finding that the optimal $\mathbf{u}$ is equal to $\mathbf{w}$ (fixed point of the Bellman equation). This can be solved in a closed form by solving a system of linear equations, which can be easily implemented in practice when the state-action mapping is low dimensional, that is our case [37], [38]. Thus, as shown in [36], the set of weights, $\mathbf{w}_{i+1}$, can be found by solving the system $\mathbf{A}\mathbf{w}_{i+1} = \mathbf{b}$ by iteratively updating the equations

$$\mathbf{b} \leftarrow \mathbf{b} + \frac{1}{N}\sum_{n}^{N} \phi(s^n,a^n) r(s^n,a^n) \tag{5}$$

$$\mathbf{A} \leftarrow \mathbf{A} + \frac{1}{N}\sum_{n}^{N} \phi(s^n,a^n)(\phi(s^n,a^n)^T - \gamma\, \phi(s'^n,\pi(s'^n))^T) \tag{6}$$

where $\mathbf{A}$ is a square matrix of the dimensionality of ϕ. It can be shown that by running LSPI until convergence, as the number of generated sequences increases, the algorithm learns the policy that maximizes the expected reward [36].

The reward we use to learn with LSPI is based on evaluating the performance of the eye fixations generated using the current policy. Recall that we have access to the exemplar trajectories of visual scanpaths that have been previously

recorded using an eye tracker. In order to evaluate how well the eye fixations generated mimic the recorded eye fixations from humans, we compute the distribution of eye fixation among all the subjects at the pixel level. The sequences are evaluated using the same criteria we use in Section V-B. Note that this is a proxy of the underlying reward of an MDP that models the human vision attention, since it enforces that the learned policy mimics the human visual scanpath.

B. State-Action Mapping $\phi(s, a)$

Since the MDP uses a discrete temporal basis, we divide the visual scanpath into different temporal consecutive stages. We use a total number of six stages, because from the experiments, we found that this value is neither too fine nor too coarse for the majority of the images, in order to capture different characteristics of the temporal evolution of the visual scanpath. In addition, we use a constant value of six stages for all the images for simplicity in the implementation.

In order to be able to learn the weights for combining the features, $\mathbf{w}$, and distinguishing among different stages in the sequence of eye fixations, we use a state-action mapping vector, $\phi(s, a)$, that differentiates among stages. To do so, we use $\phi'(s, a)$, which is the vector of the features computed in a particular stage of the sequence of eye fixations, and it is different for each stage. These features are data set dependent, and the concrete form of $\phi'(s, a)$ is introduced in Section VI. Given $\phi'(s, a)$, we define $\phi(s, a)$ using the indicator functions associated with each of the stages, and the final state-action mapping becomes the following vector:

$$\phi(s, a) = (\mathbf{I}[t = 1]\phi'(s, a), \ldots, \mathbf{I}[t = 6]\phi'(s, a))^T \quad (7)$$

where t indicates the stage of the sequence of eye fixations and $\mathbf{I}[\cdot]$ is the indicator function that takes value 1 when it is true and 0 otherwise. We can see by analyzing (7) that $\phi(s, a)$ is a sparse vector, since it is only different from 0 in the vector entries that corresponds to the current stage. In this way, when the policy is evaluated, i.e., $\mathbf{w}^T\phi(s, a)$, only the part of $\mathbf{w}$ corresponding to the stage t will have an effect, and the rest will be inhibited from the indicator function. This allows us to learn a different set of parameters for making decisions considering which stage of the visual exploration we are evaluating. Note that the learning of the different stages of the scanpath does not become independent among each other. This is because the order of the sequence is taken into account in the reward.

The features that we use for the state-action feature vector, $\phi'(s, a)$, are data set dependent. For this reason, in the following section V, we first describe the data sets, and then we introduce the state-action features.

V. Data Sets and Evaluation

In this section, we introduce the data sets and the evaluation criteria we use in the experiments.

A. Data Sets

We test our algorithm on two public eye-tracking data sets, namely, OSIE [15] and MIT data sets [16].

1) OSIE Data Set: The OSIE data set [15] contains 700 natural images and 12 semantic attributes (e.g., face, text, and motion) on 5551 outlined objects. Unlike previous eye-tracking data sets, a large portion of the images includes multiple dominant objects, making it a suitable data set for comparisons of the relative importance of the semantic cues. The ground-truth annotations of semantic objects are used as features. We randomly split the data set into 350 images for training and 350 for testing.

2) MIT Data Set: The MIT data set [16] is a widely used data set with 1003 natural images and object categories, such as face, pedestrian, and car, that can be detected with specifically trained object detectors. For the experiments on the MIT data set, we learn our model based on the features introduced by Judd *et al.* [16]. We randomly select 502 images for training and 501 images for testing.

B. Evaluation

We evaluate the prediction of visual scanpath using the evaluation metric proposed by Borji *et al.* [39]. This method first computes a mean-shift clustering for all human fixations, using the optimal bandwidth to maximize the interaction rate between the clusters. A unique character is assigned to each cluster center and the corresponding fixations, so each scanpath can be represented by a string. It measures the similarity between human subjects' scanpath and the model prediction with the Needleman–Wunsch string matching algorithm [40]. The matching scores for all the subjects are averaged to get the final evaluation score. For evaluation purposes, we provide an upper bound of the achievable performance by reporting the results of the human performance as the scanpath predictor. We compute the scanpath similarities between every two subjects, and average them to obtain the overall interobserver similarity.

In [39], the interaction rate is computed as the number of saccades between the clusters. This leads to a number of small clusters that may not represent the unit of attention in the scene (i.e., the objects). We improve the computation of the interaction rate as

$$I = \frac{N_b - N_w}{C} \quad (8)$$

where N_b and N_w represent the number of saccades between and within the clusters, respectively. The parameter C represents the number of clusters. With this improvement, the fixation clusters match the objects better than the original method.

In addition, differently from [39], instead of comparing the whole scanpath, we evaluate the models at multiple fixation stages. In particular, we compare the fixation sequences with lengths from 1 to 6 to see the change in the performance with different numbers of fixations, which allows a more explicit evaluation of the fixation order.

VI. Actions and Features

In this section, we introduce the state-action mapping $\phi'(s, a)$. First, we define the actions, and then, the features used to compute $\phi'(s, a)$ for each data set.

A. Actions

We propose to use superpixels rather than pixels as a base representation for eye-fixation prediction. Superpixels aim at segmenting the image into small segments that contain a maximum of one object inside. There are much more superpixels than possible objects, and superpixels are usually homogeneous in color or texture. We use superpixels extracted via energy-driven sampling superpixels [41] to segment the images into 300 superpixels, which is considered as one of the state-of-the-art methods for superpixel extraction.

We use superpixels instead of pixels to indicate eye fixation. This might be a natural choice, since when visual attention is allocated, it is not attracted by a single pixel, but a coherent region that represents an object or a part of an object [42]. In addition, using superpixels yields a computationally less expensive algorithm than using a pixel-based approach.

Thus, an action $a \in \mathcal{A}$ generates an eye fixation to a superpixel. The number of actions is equal to the number of superpixels. For evaluation purposes, we generate a final prediction at the pixel level using the centroid of the superpixel.

B. Features

We use a total of 19 different features for the OSIE data set, and 35 for the MIT data set. These features range from low-level cues to semantics, which we describe as follows.

1) Low-Level Features: We use the common bottom-up saliency features for both the data sets. They are generated based on three biologically plausible feature channels, namely, color, intensity, and orientation [18]. For each channel, normalized center-surround differences are computed at multiple scales and integrated linearly. In the proposed framework, which is based on superpixels, the pixelwise feature maps are averaged within each superpixel.

In the MIT data set, in addition to the aforementioned features, the local energy of steerable pyramid filters, RGB colors, probabilities, and histograms, and the Torralba saliency map are provided [43].

2) Semantic Features: It has been shown that certain object categories attract attention more strongly and rapidly than others [15], [16], [25]. For OSIE data set, we use the list of objects of interest and attributes available with the data set, which are shown to be relevant for saliency map prediction [15]. We use the 12 attributes provided in the data set, which include face, text, motion, and extra, and one additional feature to represent the annotated objects in the OSIE data set without semantic features. For MIT data set, we use the four features provided with the data set, which include the response of detectors for face, car, pedestrian, and horizon.

One scenario where discrepancy often happens between the saliency models and the human behaviors is that low-level features tend to highlight object contours with local contrast. In comparison, humans look more at objects and, particularly, at their center regions. Studies have also shown a strong correlation between eye fixations and objectness [44]–[47]. Therefore, in our experiments, the ground-truth object segmentation [15] and the detected objects [16] allow a more accurate prediction of human eye fixations. For each object, we estimate its center as the centroid of the corresponding image region, and then, place a Gaussian blob at the center of each object. In both the data sets, we do this by setting the object center pixels to 1 in the feature map and others to 0, and blurring the map with a Gaussian kernel ($\sigma = 2°$).

3) Center Bias: The center bias is an important cue for the prediction of saliency maps [16]. This may be due, for instance, to the photographer bias, the straight-ahead position, the tendency to center the eyeball within its orbit, and the tendency to look at the screen center due to strategic advantages. For the OSIE data set, we model it with a nontime-varying 2-D Gaussian distribution at the screen center. We set σ of the Gaussian distribution following the standard procedure in the literature [48]. For the MIT data set, to be consistent with Judd *et al.* [16], we use the same distance-to-center channel in their saliency model. We model the center bias differently in order to make a fair comparison between our model and these two methods. To generate the feature for each superpixel, we do the same procedure as in the low-level features. It yields a 1-D feature that indicates the distance to the center. We use two different procedures for each data set in order to reproduce the standard procedure used in the literature for each data set.

4) Spatial Distribution of Eye-Fixation Shifts: For both the data sets, OSIE and MIT, we include as a feature a prior probability of shifting the eye from a superpixel to another. We use the prior model introduced in [8], which is a 2-D Cauchy distribution. Let $(x, y)_t$ be the position of the predicted eye fixation at stage t. We define a 1-D feature that evaluates the prior probability that the next eye-fixation shifts to the location (x, y), i.e., the centroid of a superpixel, and it is defined as $(\|(x, y - (x, y)_t\|_2^2 + 60^2)^{-(3/2)}$ [8].

5) Indicator of Visited: We include another 1-D feature to indicate when the superpixels have received a fixation during the sequence of eye fixations. This feature is equal to 1 when a superpixel closer to two hops in the neighborhood has been selected in a fixation, and 0 otherwise. This feature may be useful to avoid generating new eye fixations in the superpixels already visited. We use it for the OSIE and MIT data sets.

VII. Experiments

In this section, we report the results of our method in the OSIE data set [15] and the MIT data set [16], explained in Section V. We first discuss the learning algorithm and the learned parameters, and finally, report the results compared with several baselines and the state-of-the-art methods.

A. Learning

Recall that we use LSPI to learn the parameters (Section IV). We set $\gamma = 0.6$, since it achieves the highest accuracy in a twofold cross-validation on the training set (in the following, we also report the results for $\gamma = 0$). At each iteration of LSPI, we generate a number of 15 fixations for each image in the training set. The 15 fixations are generated with the policy, extracting the fixation that have highest value function. In this way, we are able to generate more training sequences, and we found that 15 yields a good compromise between the computational cost and the accuracy. We found

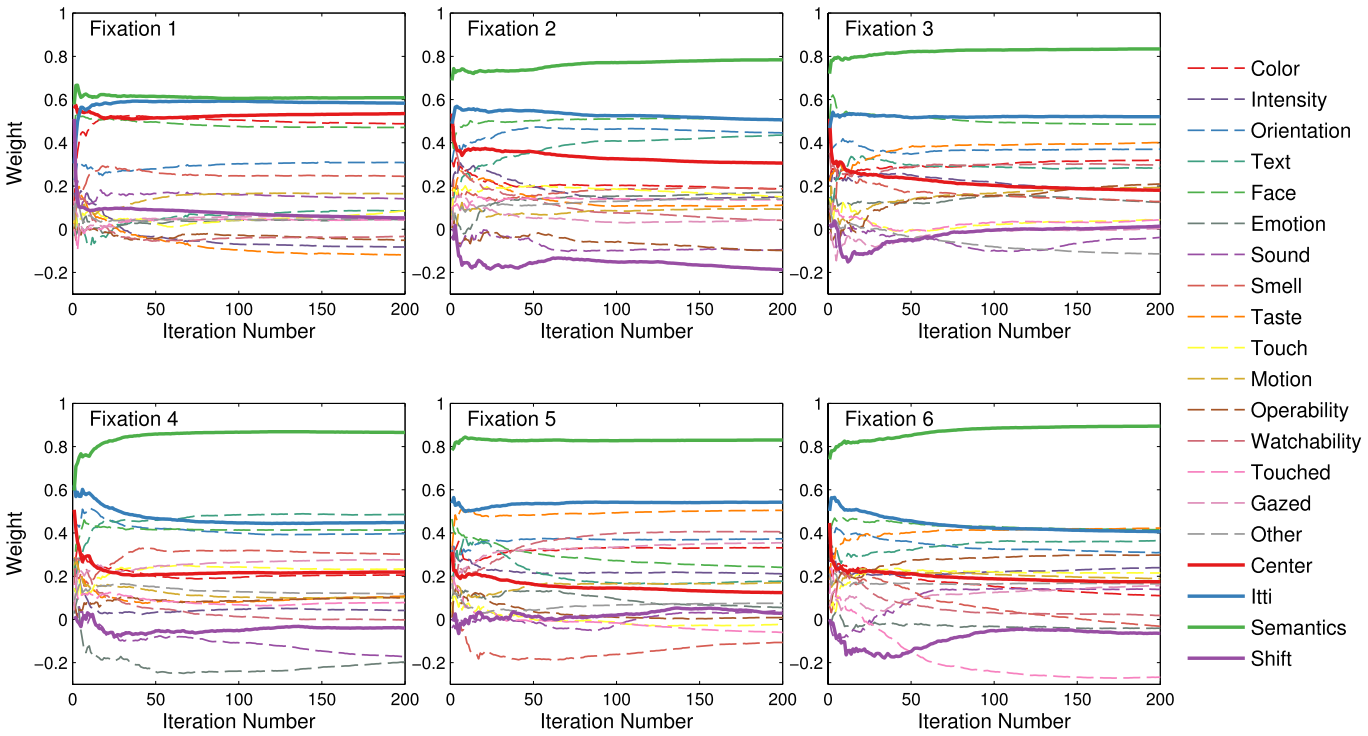


Fig. 1. Convergence of LSPI on OSIE data set. Parameter weights while learning with LSPI, for each of the six fixations in their temporal order in the scanpath. We group the cues into four groups, and plot with thicker lines the total weight of the cues in the groups.

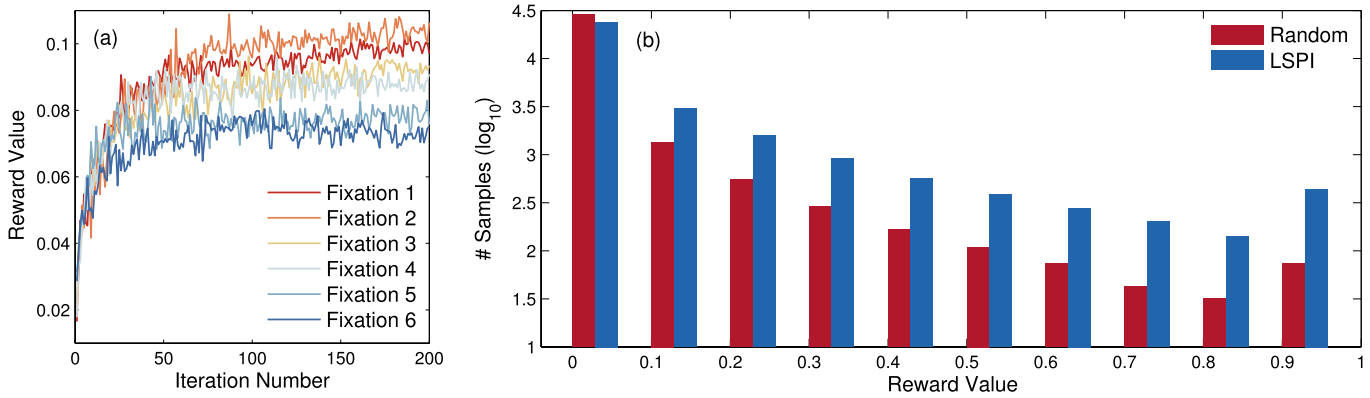


Fig. 2. Statistics of the reward during learning on OSIE data set. (a) Average and (b) distribution of the reward during learning with LSPI.

that the indicator of visit feature has an undesirable impact in the learning algorithm, because it has a strong tendency to take large negative values (it forces to avoid generating new eye fixations in the superpixels already visited). To significantly reduce the convergence time and avoid stability problems with the values of the weights, we fix the weight of the indicator visit cue to $-\infty$, and do not learn it. Note that it is still part of $\phi(s, a)$. We report the analysis of the learning on OSIE data set, but similar conclusions could be extracted with the MIT data set.

In Fig. 1, we show the progression of the parameter values during learning. Each plot corresponds to the parameters of one stage from the six different stages in which we divided the visual exploration according to their temporal order (from the first fixation to the sixth). We can observe that after 100 iterations, the LSPI converges.

In Fig. 2(a), we show the average of the reward during learning. We can see that the policy gets higher rewards, until about the 100th iteration, when it converges. We can also observe that the received reward is worse at the final stages of the visual scanpath than at the beginning. As we show in the following, the same tendency is observed when evaluating the performance of the prediction of the scanpath with LSPI, as shown in Fig. 5. This gives us some reassurance that the learning objective (maximization of the reward) is impaired with the final performance. In Fig. 2(b), we illustrate the distribution of the reward in the last iteration of the learning, i.e., the reward for the learned policy that we use in the rest of the experiments. We can see that the rewards obtained with the policy are much higher than with randomly selecting superpixels, especially for higher rewards (note that it is in logarithm scale).

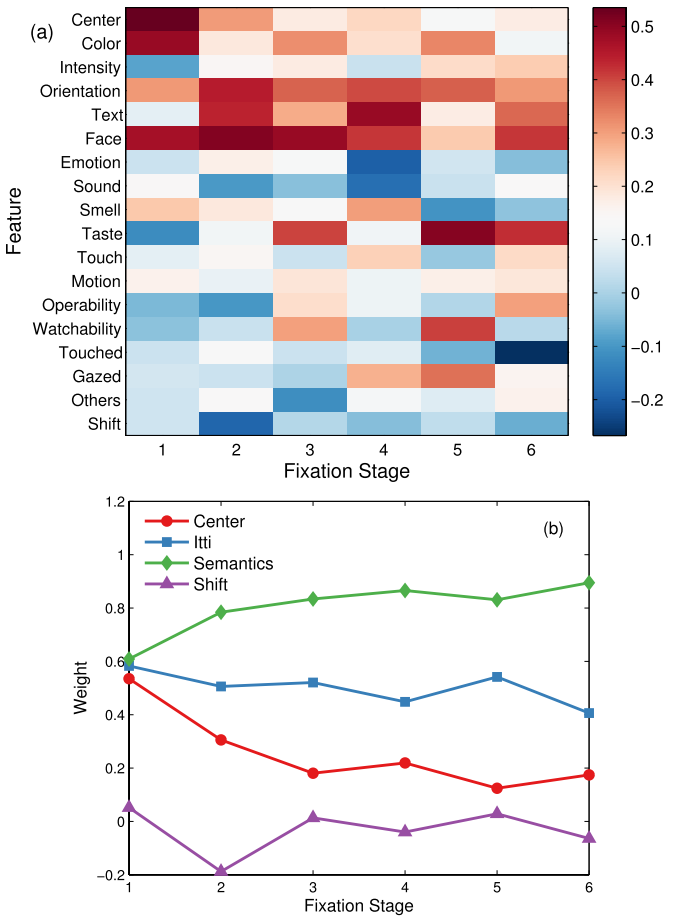


Fig. 3. Learned parameters on OSIE data set. (a) Weights of the different cues for each of the six fixations in which we divided the visual exploration. (b) We group the low-level cues by Itti-Koch saliency into one group and all the semantics into another group, and show the total weight of the cues in each group of cues. Note that the cues of the image center and the prior of the eye fixation shifts are plotted separately.

B. Analysis of the Learned Parameters

In Fig. 3(a), we show the learned weights for the different cues of the OSIE data set. We can directly compare them because all the features take values between 0 and 1. We observe that the weight of the center channel is high at first, and it decreases monotonically with time, suggesting that subjects tend to look at the center of the image at the beginning of the visual scanpath, and the center bias decreases during the visual exploration [48], [49]. Faces consistently attract attention strongly, from the very first fixation, while text becomes dominant from the second fixation suggesting the important role of the two semantic cues in gaze deployment [25]. At the later stages of the scanpath, other semantics, such as taste, watchability, gazed, and operability, also become relevant [50]–[52].

In Fig. 3(b), we bring together the cues of the OSIE data set into several groups, and plot the total weight of the cues in the group. We can observe that the semantics are, in general, the most important group in deciding where to look at, followed by low-level image features. While semantics and low-level cues contribute comparably to the first fixation, the importance of semantics keeps increasing as time goes, as more top down factors come into play [53]. Center bias is

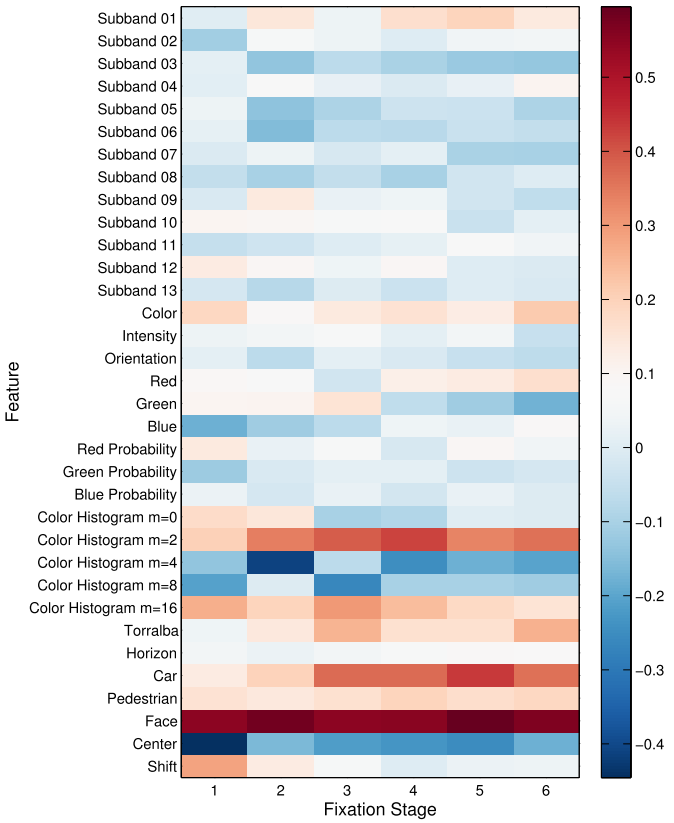


Fig. 4. Learned parameters on MIT data set. Weights of the different cues for each of the six fixations in which we divided the visual exploration.

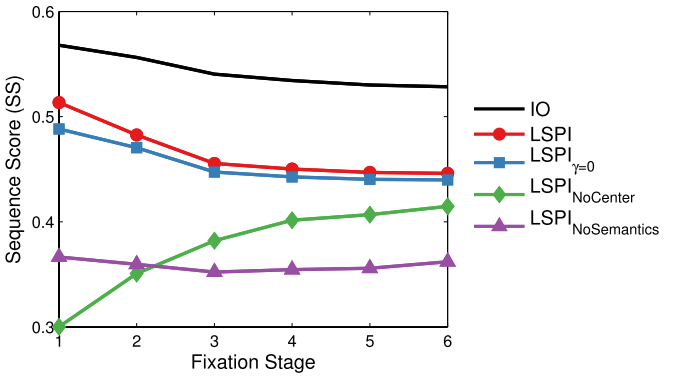


Fig. 5. Evaluation of LSPI with different feature sets and parameters on OSIE data set. LSPI$_{\gamma=0}$ is with a greedy policy, LSPI$_{\text{NoCenter}}$ is taking the center-bias channel out, and LSPI$_{\text{NoSemantics}}$ is taking the semantic features out. IO indicates the interobserver performance.

the strongest at the beginning of the scanpath, as humans start their visual explorations from the center of the image (by both the experimental setup and the strategic advantages to look at the center [48], [49]), and it clearly decreases with time.

The important roles of the semantic features and center bias are also found in the MIT data set, as illustrated in Fig. 4. Note that the center bias has negative weights because it is the distance to center, and it enforces fixations close to the center.

C. Impact of the Components of LSPI

We first evaluate several baselines based on LSPI to analyze the contribution of the different components in our model. We can observe in Fig. 5 that taking into account

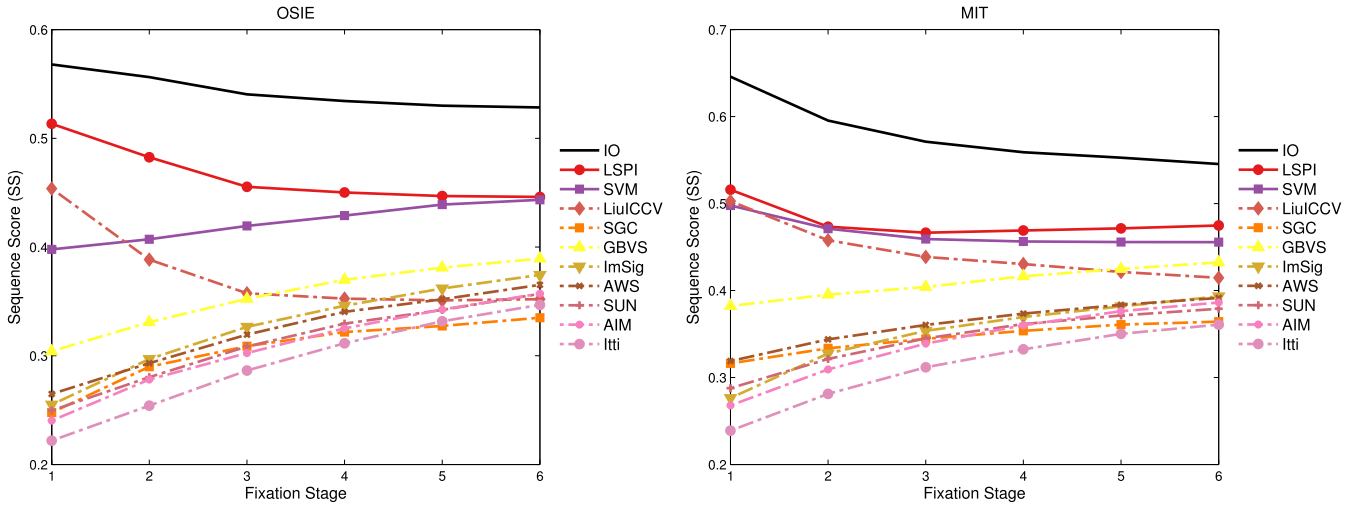


Fig. 6. Evaluation of LSPI and baseline models. The LSPI model is compared with the state-of-the-art LiuICCV [8], SGC [32], and the winner-take-all heuristic [18] from several saliency maps, such as support vector machine (SVM) [15], [16], graph-based visual saliency (GBVS) [20], image signature (ImSig) [23], adaptive whitening saliency (AWS) [54], saliency using natural statistics (SUN) [21], attention based on information maximization (AIM) [22], and Itti [18]. The SVM is in solid line, because it uses the same low- and semantic-level features and center bias as LSPI. IO indicates the interobserver performance.

Fig. 7. Qualitative evaluation of LSPI and baselines on OSIE data set. Human ground truth and predicted visual scanpaths of the LSPI and the state-of-the-art saliency models.

the reward in the future actions does improve over a policy learned with a greedy reward ($\gamma = 0$). Since there was a significant uncertainty in the human eye-movement behavior, subjects attended to salient regions in different temporal orders. Therefore, using $\gamma = 0.6$ is shown to alleviate the effects of the uncertainty. To show the contribution in the

Fig. 8. Qualitative evaluation of LSPI and baselines on MIT data set. Human ground truth and predicted visual scanpaths of the LSPI and the state-of-the-art saliency models.

performance of the different cues, we take the semantic-based cues and the image center out from $\phi(s, a)$ (in the plot, indicated as $\text{LSPI}_{\text{NoSemantics}}$ and $\text{LSPI}_{\text{NoCenter}}$, respectively). We can observe that the performance without including the semantics and the center cues is significantly lower. This finding agrees with the previous studies that emphasize the roles of the semantic features [15], [16], [25] and the center bias [16], [48], [49] in saliency prediction. In particular, as can be seen, without the center-bias channel, the performance is significantly decreased at the early fixations, which coincides with our analysis of the parameter weights in Section VII-B.

D. Comparison of Scanpath Prediction From Saliency Maps

We evaluate several baselines that generate visual scanpaths directly from the saliency maps with the winner-take-all heuristic, as proposed in [18] and [19]. These are GBVS [20], ImSig [23], AWS [54], SUN [21], AIM [22], and Itti [18]. Fig. 6 shows that LSPI is able to outperform these models on both the data sets. These results coincide with what was observed in previous works for scanpath prediction [7], [8], [33] that modeling the temporal information yields a better performance than the models that discard the temporal information using a saliency map.

We also evaluate a baseline based on linear SVM models and winner-take-all [15], [16], which integrate the same features as in the LSPI model, but it first computes a saliency map, and from that it generates the sequence of eye fixations. The SVM is trained using the features extracted from all superpixels in the training images. Results in Fig. 6 show that the SVM achieves a lower prediction accuracy than the LSPI. Since the same cues are used for the LSPI and the SVM, this experiments show that the temporal dynamics taken into account in LSPI are useful for the visual scanpath prediction. Note that the difference in performance between the LSPI and the SVM is higher at the beginning of the scanpath than at the end. This is because at the first stages of the visual exploration, our model learns that it is more effective to use the low-level features, and this cannot be modeled by the SVM baseline, since it uses the same weights for all the stages.

E. Comparison With the State-of-the-Art Models

We compare our results with the state-of-the-art scanpath prediction models LiuICCV [8] and SGC [32]. We use

their codes to predict the sequences of six eye fixations. Fig. 6 shows that our method outperforms the state-of-the-art scanpath prediction models, and the learning-based LiuICCV model yields a better performance than the feedforward SGC model without semantic features. Our method obtains higher performance compared with LiuICCV at the later stage of the visual scanpath than at the beginning. This is because in the later stages, semantic features are more useful for the prediction, and our model uses a much richer set of semantic features than LiuICCV. This method uses an unsupervised learning to extract the semantic classes, while we use the output of object detectors learn specifically to detect the semantic classes that have been shown to be useful for saliency prediction.

In Fig. 6, we also provide the results of the human performance as a scanpath predictor (denoted as IO). We can observe that there is room for improvement, since there is a significant gap between our model and the human performance.

F. Computational Cost

The models were implemented in MATLAB 2013b, running on a Dell Optiplex 990 with Intel i7-2600 CPU @ 3.40 GHz and 16-GB RAM. The computational cost at the testing time is dominated by the feature extraction (more than 1 s/image). Evaluating the policy in LSPI consists of a dot product between the feature vector and the set of learnt parameters, which for all the superpixels in an image takes $\sim$9.8 ms. For the methods based on generating the saliency map, the computational cost is higher than directly predicting the eye fixation (e.g., 251.2 ms for the SVM baseline), since extracting the eye fixations from the saliency map requires evaluation inhibition of return, which has a higher computational cost than the dot products done in LSPI policy evaluation.

At training time, the computational cost of learning the model with LSPI is significative, because it requires generating a large amount of sequences of eye fixations. For the MIT data set, every iteration of the algorithm takes $\sim$12.6 ms/image and 6.3 s in total. The sequences can be generated in parallel for each image, and then the policy can be updated. The total training time is of 21 min. The computational cost of learning the SVM baseline is 3.4 s, which is much lower than the LSPI, because it does not require to generate multiple sequences of eye fixations.

G. Qualitative Results

Finally, in Figs. 7 and 8, we depict the examples of the visual scanpaths predicted by the LSPI and the baselines, including the human scanpath with the highest interobserver similarity. As shown in the Fig. 7 and Fig. 8, fixations from both the LSPI and SVM based methods generally land on semantically meaningful objects, while other models tend to focus on the regions with distinct low-level features or object contours. Compared with the SVM and the state-of-the-art models that fixate mostly in a focused region, the LSPI model generates better scanpaths covering a wide range of objects, showing the effect of the temporal-variant weights of features.

VIII. Conclusion

We introduced a model to predict the sequences of human eye fixations while free viewing natural images. Our model is learned with LSPI from recorded human eye fixations in natural images, and allows the integration of multiple cues by considering the different stages of visual exploration. The experimental results show that taking the temporal dynamics for integrating multiple cues into account results in better visual scanpath prediction. In addition, the results obtained by our model outperform the state-of-the-art results in the automatic prediction of the sequences of eye fixations.

References

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

[2] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annu. Rev. Neurosci.*, vol. 23, no. 1, pp. 315–341, 2000.

[3] J. Han *et al.*, "Video abstraction based on fMRI-driven visual attention model," *Inf. Sci.*, vol. 281, pp. 781–796, Oct. 2014.

[4] T. S. Lee and S. X. Yu, "An information-theoretic framework for understanding saccadic eye movements," in *Advances in Neural Information Processing Systems*, vol. 12, S. A. Solla, T. K. Leen, and K. Müller, Eds. Cambridge, MA, USA: MIT Press, 2000, pp. 834–840.

[5] J. Najemnik and W. S. Geisler, "Optimal eye movement strategies in visual search," *Nature*, vol. 434, no. 7031, pp. 387–391, 2005.

[6] B. T. Sullivan, L. Johnson, C. A. Rothkopf, D. Ballard, and M. Hayhoe, "The role of uncertainty and reward on eye movements in a virtual driving task," *J. Vis.*, vol. 12, no. 13, p. 19, 2012.

[7] 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.

[8] 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.

[9] L. Paletta and G. Fritz, "Reinforcement learning for decision making in sequential visual attention," in *Attention in Cognitive Systems. Theories and Systems From an Interdisciplinary Viewpoint* (Lecture Notes in Computer Science), vol. 4840, L. Paletta and E. Rome, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 293–306.

[10] B. Goodrich and I. Arel, "Reinforcement learning based visual attention with application to face detection," in *Proc. 25th IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Providence, RI, USA, Jun. 2012, pp. 19–24.

[11] T. Darrell and A. Pentland, "Active gesture recognition using learned visual attention," in *Advances in Neural Information Processing Systems*, vol. 8, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA, USA: MIT Press, 1996, pp. 858–864.

[12] S. Minut and S. Mahadevan, "A reinforcement learning model of selective visual attention," in *Proc. 5th Int. Conf. Auto. Agents*, Montreal, QC, Canada, May 2001, pp. 457–464.

[13] A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Learning sequential visual attention control through dynamic state space discretization," in *Proc. 26th IEEE Int. Conf. Robot. Autom.*, Kobe, Japan, May 2009, pp. 2258–2263.

[14] C. Kwok and D. Fox, "Reinforcement learning for sensing strategies," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, vol. 4. Sendai, Japan, Sep. 2004, pp. 3158–3163.

[15] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, p. 28, 2014.

[16] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 2106–2113.

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

[18] 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.

[19] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.

[20] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, vol. 19, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. Cambridge, MA, USA: MIT Press, 2007, pp. 545–552.

[21] 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, p. 32, 2008.

[22] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, p. 5, 2009.

[23] 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.

[24] Y. Zhang, J. Han, and L. Guo, "Saliency detection by combining spatial and spectral information," *Opt. Lett.*, vol. 38, no. 11, pp. 1987–1989, 2013.

[25] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *J. Vis.*, vol. 9, no. 12, p. 10, 2009.

[26] R. Valenti, N. Sebe, and T. Gevers, "Image saliency by isocentric curvedness and color," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 2185–2192.

[27] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[28] Y. Lu, W. Zhang, H. Lu, and X. Xue, "Salient object detection using concavity context," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 233–240.

[29] L. Xu, H. Li, L. Zeng, and K. N. Ngan, "Saliency detection using joint spatial-color constraint and multi-scale segmentation," *J. Vis. Commun. Image Represent.*, vol. 24, no. 4, pp. 465–476, 2013.

[30] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. 27th IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 280–287.

[31] L. W. Renninger, J. M. Coughlan, P. Verghese, and J. Malik, "An information maximization model of eye movements," in *Advances in Neural Information Processing Systems*, vol. 17, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA, USA: MIT Press, 2005, pp. 1121–1128.

[32] X. Sun, H. Yao, and R. Ji, "What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1552–1559.

[33] S. Mathe and C. Sminchisescu, "Action from still image dataset and inverse optimal control to learn task specific visual scanpaths," in *Advances in Neural Information Processing Systems*, vol. 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran & Associates Inc., 2013, pp. 1923–1931.

[34] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 2014.

[35] R. Bellman, "A Markovian decision process," *Indiana Univ. Math. J.*, vol. 6, no. 4, pp. 679–684, Apr. 1957.

[36] M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *J. Mach. Learn. Res.*, vol. 4, pp. 1107–1149, Dec. 2003.

[37] J. A. Boyan, "Least-squares temporal difference learning," in *Proc. 16th Int. Conf. Mach. Learn.*, Bled, Slovenia, Jun. 1999, pp. 49–56.

[38] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Mach. Learn.*, vol. 22, nos. 1–3, pp. 33–57, 1996.

[39] 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, VIC, Australia, Dec. 2013, pp. 921–928.

[40] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Molecular Biol.*, vol. 48, no. 3, pp. 443–453, 1970.

[41] M. Van den Bergh, X. Boix, G. Roig, and L. Van Gool, "SEEDS: Superpixels extracted via energy-driven sampling," *Int. J. Comput. Vis.*, vol. 111, no. 3, pp. 298–314, 2015.

[42] G. Kanizsa, *Organization in Vision: Essays on Gestalt Perception*. New York, NY, USA: Praeger Pub., 1979.

[43] 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.

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

[45] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[46] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2009–2021, Dec. 2013.

[47] J. Han, L. Sun, X. Hu, J. Han, and L. Shao, "Spatial and temporal visual attention prediction in videos using eye movement data," *Neurocomputing*, vol. 145, pp. 140–153, Dec. 2014.

[48] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, p. 9, 2011.

[49] 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, p. 4, 2007.

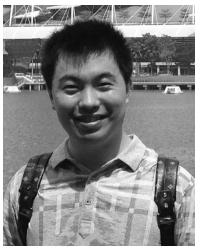
[50] K. J. Friston, G. Tononi, G. N. Reeke, Jr., O. Sporns, and G. M. Edelman, "Value-dependent selection in the brain: Simulation in a synthetic neural model," *Neuroscience*, vol. 59, no. 2, pp. 229–243, 1994.

[51] P. J. Whalen *et al.*, "Human amygdala responsivity to masked fearful eye whites," *Science*, vol. 306, no. 5704, p. 2061, 2004.

[52] M. S. Beauchamp, K. E. Lee, J. V. Haxby, and A. Martin, "FMRI responses to video and point-light displays of moving humans and manipulable objects," *J. Cognit. Neurosci.*, vol. 15, no. 7, pp. 991–1001, 2003.

[53] S. K. Mannan, C. Kennard, and M. Husain, "The role of visual salience in directing eye movements in visual object agnosia," *Current Biol.*, vol. 19, no. 6, pp. R247–R248, 2009.

[54] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *J. Vis.*, vol. 12, no. 6, p. 17, 2012.



Ming Jiang (S'13) received the B.Eng. and M.Eng. degrees from Zhejiang University, Hangzhou, China, in 2004 and 2008, respectively. He is currently pursuing the Ph.D. degree with the National University of Singapore, Singapore.

His current research interests include computer vision, visual cognition, and computational neuroscience.

Xavier Boix (S'10) received the Ph.D. degree in computer vision from ETH Zurich, Zürich, Switzerland, in 2014.

He is currently a Post-Doctoral Researcher with the National University of Singapore, Singapore. He is also a Research Affiliate with the Brain and Cognitive Science Department, Massachusetts Institute of Technology, Cambridge, MA, USA.

Dr. Boix received the Intel Doctoral Student Award from ETH Zurich.

Gemma Roig (S'10) received the Ph.D. degree in computer vision from ETH Zurich, Zürich, Switzerland, in 2014.

She is currently a Post-Doctoral Fellow with the Brain and Cognitive Science Department, Massachusetts Institute of Technology, Cambridge, MA, USA.

Juan Xu received the M.E. degree from the National University of Singapore, Singapore, in 2014.

She is currently a Research Associate with Dr. Q. Zhao at the National University of Singapore. Her current research interests include visual attention and the visual search ability in children with autism spectrum disorder.

Luc Van Gool (S'79–M'81–SM'10) received the Dr.-Ing. degree in electromechanical engineering from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1981.

He is currently a Professor with the Katholieke Universiteit Leuven and ETH Zurich, Zurich, Switzerland, where he leads computer vision research and teaches computer vision. He has authored over 200 papers in this field. His current research interests include 3-D reconstruction and modeling, object recognition, tracking, and gesture analysis.

Prof. Van Gool has been a Program Committee Member of several major computer vision conferences. He received several best paper awards. He is a Co-Founder of ten spin-off companies.

Qi Zhao (SM'04–M'10) received the M.Sc. and Ph.D. degrees in computer engineering from the University of California at Santa Cruz, Santa Cruz, CA, USA, in 2007 and 2009, respectively.

She was a Post-Doctoral Researcher with the Department of Computation and Neural Systems and the Division of Biology at the California Institute of Technology, Pasadena, CA, USA, from 2009 to 2011. She is currently an Assistant Professor with the Electrical and Computer Engineering Department, National University of Singapore (NUS), Singapore, where she is the Principal Investigator with the Visual Information Processing Laboratory and involved in computational vision and cognitive neuroscience. She also holds an appointment with the Ophthalmology Department and the Interactive and Digital Media Institute at NUS. She has authored over 30 journal and conference papers in top computer vision, cognitive neuroscience, and machine learning venues, and has edited a book titled *Computational and Cognitive Neuroscience of Vision* (Springer) that provides a systematic and comprehensive overview of vision from various perspectives, ranging from neuroscience to cognition, and from computational principles to engineering developments. Her current research interests include computational vision, machine learning, computational cognition, and neuroscience.