Scanpath Prediction for Visual Attention using IOR-ROI LSTM

Zhenzhong Chen, Wanjie Sun
School of Remote Sensing and Information Engineering,
Wuhan University, Wuhan, P.R. China
{zzchen, sunwanjie}@whu.edu.cn

Abstract

Predicting scanpath when a certain stimulus is presented plays an important role in modeling visual attention and search. This paper presents a model that integrates convolutional neural network and long short-term memory (LSTM) to generate realistic scanpaths. The core part of the proposed model is a dual LSTM unit, *i.e.*, an inhibition of return LSTM (IOR-LSTM) and a region of interest LSTM (ROI-LSTM), capturing IOR dynamics and gaze shift behavior simultaneously. IOR-LSTM simulates the visual working memory to adaptively integrate and forget scene information. ROI-LSTM is responsible for predicting the next ROI given the inhibited image features. Experimental results indicate that the proposed architecture can achieve superior performance in predicting scanpaths.

1 Introduction

Selective visual attention mechanism of human vision system automatically extracts information from a subset of visual input by shifting our eyes to bring region of interest (ROI) onto the fovea where fine-grained analysis can be carried out. Most studies concerning selective visual attention are dealing with overt attention involving eye movements [Zhang and Lin, 2013]. Successive eye movements during scene exploration, called the visual scanpaths, are comprised of a series of saccades and fixations. Fixations indicate ROIs of observers and saccades represent rapid changes of gaze. Unlike saliency models that aim to measure the probability distribution of fixations with respect to the spatial layout of a stimulus, visual scanpath prediction models can record not only attention locations but also the order among fixations. Understanding and predicting scanpaths is a challenge yet important task, potentially contributing to advancements of vision research, visual art design, robot vision, virtual reality, which has received great research interests recently [Boccignone, 2016; Le Meur *et al.*, 2017].

Visual inhibition of return (IOR) facilitates foraging by discouraging re-examination of recently fixated locations and objects [Samuel and Kat, 2003; Wang and Klein, 2010]. Due to the limited capacity of the visual working memory (VWM), effects of IOR gradually fade over time, which allows the possibility to re-fixate previously attended regions. Such a process is fundamental to computational models of selective visual attention [Bays and Husain, 2012]. [Itti *et al.*, 1998] firstly implemented a bottom-up saliency model that outputs a sequence of fixations using a winner-take-all neural network and IOR. [Wang *et al.*, 2011] introduced a saccadic scanpath prediction model based on information maximization criteria, which incorporates the VWM to integrate fixated regions and to forget earlier ones at a constant rate. [Le Meur and Liu, 2015] proposed a new framework to predict visual scanpaths by modeling bottom-up saliency, oculomotor biases, and IOR. [Jiang *et al.*, 2016] presented a reinforcement learning approach, also employed IOR, with least-squares policy iteration to predict a sequence of human fixations, simulating the recorded eye-fixation examples.

A few studies have been conducted to integrate deep learning frameworks into scanpath generation. [Assens Reina *et al.*, 2017] introduced the saliency volume to generate scanpath on 360° images by sampling fixations from each temporal saliency slice in the saliency volume. But the accuracy of the generated scanpath heavily depends on the quality of saliency volume and sampling strategies. [Ngo and Manjunath, 2017] proposed a model that generates scanpath using both convolutional neural network (CNN) and recurrent neural network (RNN). High-level image features extracted by a pre-trained CNN are sent to RNN which predicts fixation transition probability from the current position to possible locations. This model utilized the standard LSTM [Hochreiter and Schmidhuber, 1997] to simulate the VWM, but the input to the LSTM is essentially a one-dimensional vector discarding relative spatial information which is indispensable to simulate human vision.

In this paper, we propose a bio-inspired and interpretable scanpath prediction architecture based on convolutional LSTM (ConvLSTM) [Shi *et al.*, 2015] that incorporates a novel IOR-LSTM along with ROI-LSTM to capture IOR dynamics and gaze shift behavior simultaneously. Results on the OSIE dataset [Xu *et al.*, 2014] and MIT dataset [Judd *et al.*, 2009] have verified the effectiveness of our model.

This work was supported in part by the National Natural Science Foundation of China under Grant 61771348 and 61471273.

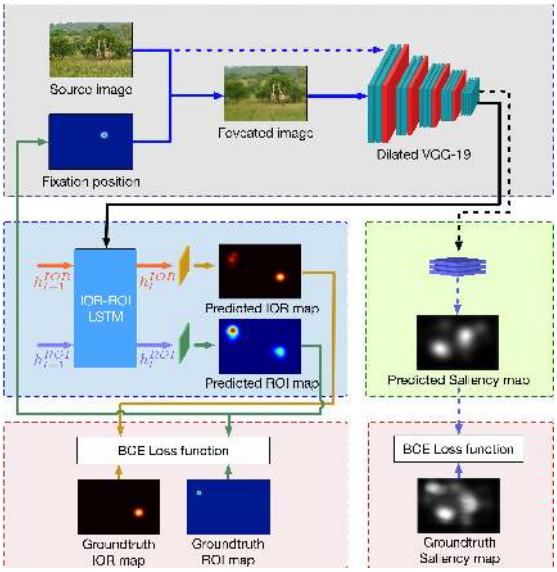


Figure 1: Architecture of the proposed scanpath prediction model, which includes three major components: an image feature extractor (light gray), the IOR-ROI LSTM module (light blue), and the saliency guidance network (light green).

2 Model Architecture

Figure 1 shows the overview of the proposed scanpath prediction architecture. There are three major components presented in this model: an image feature extractor, the IOR-ROI LSTM module and a saliency guidance network. The image feature extractor is responsible for synthesizing the foveated image and extracting visual feature maps using CNN. The IOR-ROI LSTM plays a critical role in predicting the next eminent ROI to which subjects tend to pay attention. The saliency guidance network is only employed during the training phase, forcing the feature extractor to encode more global saliency information. A new prediction loop is started by updating the foveated image using the latest fixation selected from the predicted ROI. In order to ensure that the IOR-ROI LSTM module can effectively learn visual IOR pattern, predicted IOR map is used to compute distance against its corresponding ground truth as an auxiliary training loss which will be jointly optimized with the ROI prediction loss.

2.1 Foveated Image and Image Feature Extractor

Thanks to the foveal system of human eyes, human can concentrate on a specific region with significant detail [Wandell, 1995]. The most area of the retina outside the foveal region is called peripheral vision which only occupies limited resources of the brain by perceiving visual information in a low-resolution manner. In our work, we employ the computational foveation model described in [Wang *et al.*, 2017] and implement it on the modern GPU to efficiently simulate the foveation process of the human vision system.

The image feature extractor can be any form of CNNs, *e.g.*, VGG-Net [Simonyan and Zisserman, 2014] and ResNet [He *et al.*, 2016]. We employ all convolutional and pooling layers of a pre-trained VGG-19 network to extract convolutional features. However, original VGG-19 is composed of five blocks where each of them is followed by a max pooling layer with the stride of 2×2, which considerably reduces spatial resolution of the resulting feature maps by a factor of 32. This could significantly degrade the quality of the feature maps. To circumvent this issue, we modify the original VGG-19 structure to produce feature maps with a downscale factor of 8, while maintaining the same receptive field of convolutional kernels. Last two max-pooling layers of the original VGG-19 is discarded and convolution operation in block 5 is replaced with dilated convolution [Yu and Koltun, 2016] which has kernel size of 3×3 and dilation rate of 2.

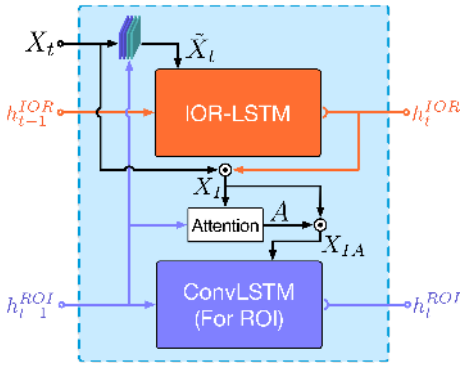


Figure 2: The IOR-ROI LSTM module consists of two parts: IOR-LSTM which is used to capture IOR pattern; attentive ROI-LSTM which is used to predict the next ROI.

2.2 IOR-ROI LSTM Module

IOR behavior is individual dependent and reacts differently to various visual features [Hotta *et al.*, 2010]. Thus, it is not bioplausible to mimic IOR using linear models which are usually integrated into computational visual attention and gaze shift models [Itti *et al.*, 1998; Wang *et al.*, 2011; Le Meur and Liu, 2015; Wang *et al.*, 2017]. To this end, we devise the IOR-ROI LSTM module that is capable of capturing IOR dynamics and gaze shift behavior of human observers simultaneously. As shown in Figure 2, the IOR-ROI LSTM module is designed to incorporate inhibited attentive mechanism in predicting ROI from which the next fixation is generated.

To be self-contained, basic knowledge of ConvLSTM is first introduced. In ConvLSTM, fully connected structures in standard LSTM are substituted with convolutional layers which can process spatial features while reducing parameters. Like standard LSTM, ConvLSTM is also driven by three self-parameterized controlling gates, *i.e.*, an input gate i_t, a forget gate f_t, and an output gate o_t. Gates controlling and states transition of ConvLSTM can be described as follows:

$$\begin{aligned} i_t &= \sigma(W_i * x_t + U_i * h_{t-1} + b_i) \\ f_t &= \sigma(W_f * x_t + U_f * h_{t-1} + b_f) \\ o_t &= \sigma(W_o * x_t + U_o * h_{t-1} + b_o) \\ g_t &= \tanh(W_g * x_t + U_g * h_{t-1} + b_g) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{1}$$

where x_t is the input at time step t; h_{t-1} and h_t are the hidden states of the ConvLSTM at time step $t-1$ and t, respectively; g_t is candidate content to be written into the cell state c_t; all

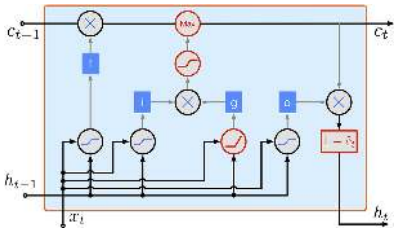


Figure 3: Inner structure of the IOR-LSTM. It is a modified version of original ConvLSTM, better simulating working mechanism of the VWM.

Ws, Us, and bs are learned weights and biases. The $*$ and $\odot$ represent convolution operation and Hadamard product, respectively.

1) IOR-LSTM: IOR-LSTM is in charge of learning to capture IOR dynamics when generating scanpaths. It should be fed with information from currently and previously attended regions. Specifically, its input is combined by:

$$\tilde{X}_t = f(W_{IOR} * [X_t, h_{t-1}^{ROI}]) \tag{2}$$

where $X_t \in \mathbb{R}^{W' \times H' \times C}$ represents image features, containing hints of currently fixated location due to the foveation process. W', H' and C denotes width, height and dimension of the extracted feature maps, respectively. $h_{t-1}^{ROI} \in \mathbb{R}^{W' \times H' \times D}$ is the hidden state of the ROI-LSTM which encodes attention history. D is the dimension of state maps. The fused representation of current and previous attention $\tilde{X}_t \in \mathbb{R}^{W' \times H' \times D}$ is obtained by convolving the concatenation of X_t and h_{t-1}^{ROI} through a non-linear activation function $f(\cdot)$.

The IOR-LSTM is modified from the original ConvLSTM to better simulate operating mechanism of the VWM where recently attended objects are retained, and earlier attended regions are gradually decayed. As shown in Figure 3, the learned forget gate selectively removes earlier information from previous cell state c_{t-1} by multiplying it with an adaptive forgetting factor in $(0, 1)^{W' \times H' \times D}$. Currently observed information is first transformed into g_t and then modulated by the input gate which determines what and how much information should be passed in. Different from the original ConvLSTM that updates memory through an addition operation, we simulate the iterative process of the VWM in the mental representation by updating cell state with the element-wise maximal values between decayed memory and information to be written as suggested by [Wang *et al.*, 2011], which can be described mathematically as:

$$c_t = \max(f_t \odot c_{t-1}, \lceil i_t \odot g_t \rceil) \tag{3}$$

where $\lceil \cdot \rceil$ is used to truncate values to the range $[0, 1]$, which essentially simulates the saturation of memory neurons. Besides, the original non-linear activation of candidate memory g_t is replaced with $ReLU(\cdot)$ to only retain positive activation.

In the IOR-LSTM, each channel of c_t indicates the spatial VWM usage of one type of features. Hence, values of the cell state c_t can be considered as the inhibition strength of each feature map. The more memory is occupied, the intenser that spatial location should be inhibited. To guide the ROI-LSTM to explore regions where its corresponding VWM has no or shallow footprint, hidden state of the IOR-LSTM is updated as:

$$h_t^{IOR} = 1 - o_t \odot c_t \tag{4}$$

As expressed by Eqn. 4, unexplored regions will have higher values than attended areas. The feature-wise inhibition masks h_t^{IOR} produced by IOR-LSTM are multiplied with feature maps of the foveated image X_t element-wisely, which forms the inhibited feature maps X_I that can be considered as the context describing where should not be focused during the next attention.

2) ROI-LSTM: In determining the next ROI, we use the soft attention model described in [Cornia *et al.*, 2017] that convolves X_I along with hidden state of the ROI-LSTM h_{t-1}^{ROI} to output a 3D tensor summarizing the hidden representations focusing on the input X_I. The hidden representations are then projected into a single channel output using 1×1 convolution filters. This is followed by a $softmax(\cdot)$ function to normalize the input across spatial pixels. The whole process can be expressed as:

$$A = softmax(K * \tanh(W_A * X_I + U_A * h_{t-1}^{ROI} + b_A)) \tag{5}$$

The attention map A is then applied to each channel of the inhibited image feature maps X_I through the element-wise product to produce the inhibited attentive image features:

$$X_{IA} = A \odot X_I \tag{6}$$

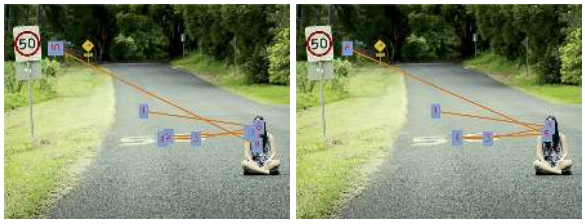
Finally, the ROI-LSTM, with original ConvLSTM structure driven by Eqn. 1, takes X_{IA} as input and updates its hidden state h_t^{ROI} which will be used to generate the ROI distribution map from which the point with the maximum value is selected as the next fixation.

2.3 Saliency Guidance Network

Visual saliency is one type of visual attention representations and saliency map demonstrates rich information about to what extent a specific region grabs observers' visual attention. Therefore, encoding saliency information into image features provides the IOR-ROI LSTM with better cues on where it should fixate. This is achieved by partially adopting readout network from the DeepGaze II [Kümmerer *et al.*, 2017] which takes the original image features as input and outputs the predicted saliency map. Thus, the ground truth saliency map of the input image can be used to guide the training of the high-level image feature extractor. The readout network, consisting of four 1×1 convolution layers followed by ReLU activation functions, transforms image features across various channels into the final saliency map. We train the saliency guidance network to encourage last convolutional features to output higher activation corresponding to more salient regions of the original image.

2.4 Loss Function

During the training phase, outputs produced by the IOR-ROI LSTM at time t are the IOR distribution $IOR_t \in [0, 1]^{W \times H}$ and the next ROI distribution $ROI_t \in [0, 1]^{W \times H}$. W and H represent width and height of the input image. These two outputs are functions of the hidden state h_t^{IOR} and h_t^{ROI}, respectively. They are modeled as one 1×1 convolutional layer followed by Sigmoid non-linearity. Besides, another output



(a) Raw fixations (b) Refined fixations

Figure 4: Example scanpath from the OSIE dataset.

during training phase is the predicted saliency map S_t from the saliency guidance network. To jointly optimize the entire model, we define the loss function at time step t as:

$$L_t(IOR_t, ROI_t, S_t) = \alpha\mathcal{L}(IOR_t, IOR_t^*) + \beta\mathcal{L}(ROI_t, ROI_t^*) + \gamma\mathcal{L}(S_t, S_t^*) \quad (7)$$

where IOR_t^*, ROI_t^* and S_t^* denote ground truth of the IOR, ROI and saliency map, respectively; $\mathcal{L}(\cdot)$ is the binary cross entropy function, while α, β and γ are three scalar factors setting to balance the three loss functions.

3 Experiments

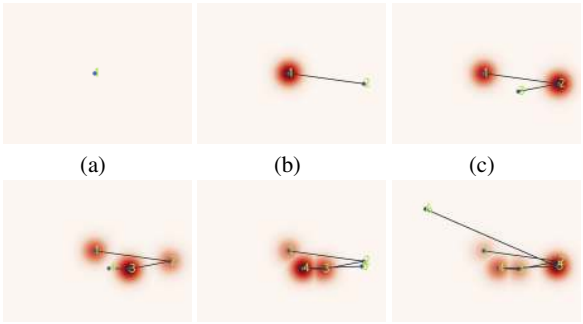
3.1 Datasets

We trained and evaluated our proposed model on the widely used public OSIE and MIT eye-tracking datasets consisting of natural images along with eye-tracking data from different participants. In the OSIE dataset, there are 700 images with 800×600 pixels. Eye-tracking data was acquired from 15 participants for each image. The MIT dataset is composed of 1003 images with resolution ranging from 405 to 1024. Eye-tracking data was also recorded from 15 subjects for each image.

We observed a common phenomenon in both OSIE and MIT datasets that most scanpaths have consecutive fixations located within very close distance, *i.e.*, those fixations fall into the same ROI. Assuming that ROI corresponds to valid receptive field of the fovea, *i.e.*, 5° of central field of vision [Wandell, 1995], 94.85% scanpaths in the OSIE dataset and 81.36% scanpaths in the MIT dataset have successive fixations belonging to the same ROI. To be consistent with the IOR theory, it is appropriate to combine successive fixations within the same ROI into one fixation, the merge process is iteratively executed until no consecutive fixations are within a given ROI. According to the experiment setup collecting the OSIE and MIT dataset, 1° of visual angle contains 24 pixels in the OSIE images [Xu *et al.*, 2014] and 35 pixels in the MIT images [Judd *et al.*, 2009], thus radius of ROI for the OSIE and MIT dataset are 60 pixels and 88 pixels, respectively. Figure 4(a) shows an example scanpath in the OSIE dataset. All $(n+1)$-th fixation located within a radius of 60 pixels around the n-th fixation should be merged into one fixation. Figure 4(b) shows the refined scanpath in which the fourth and fifth fixation are merging results of (4, 5) fixations and (6, 7, 8, 9) fixations shown in Figure 4(a).

3.2 Inhibition and ROI Map Generation

It is important to retrieve ground truth IOR to supervise our model to learn reasonable and interpretable IOR behavior.



(a) (b) (c)

(d) (e) (f)

Figure 5: Ground truth inhibition map for each state of the scanpath shown in Figure 4(b).

However, datasets we used do not provide any IOR information. Therefore, we need to define what re-fixation is and how IOR attenuates only with the given fixations' location and order information. We consider re-fixation as an action that the n-th fixation falls into ROI centered at the $(n-t)$-th fixation. The $(n-t)$-th fixation is called re-fixated point of the n-th fixation. When re-fixation happened, inhibition effects exerted around the re-fixated point should fade completely. Hence, inhibition decay rate of the re-fixated point depends on the number of time steps that re-fixation happened since the re-fixated point has been visited for the first time. For other locations, inhibition effects decline to a given threshold until the end of the scanpath. Initial inhibition strength around the i-th attended location $\mathbf{p}_i$ is computed as:

$$\mathbf{\Phi}_i(\mathbf{x}) = \exp(-\frac{||\mathbf{x}-\mathbf{p}_i||_2^2}{\sigma^2}) \quad (8)$$

where $\mathbf{x}$ represents location in the inhibition map and σ controls the spread of the inhibition which is set to 2.5° of visual angle in our experiments. The ground truth inhibition map is a combination of previous decayed inhibition map and the latest inhibition map via the $\max$ operation:

$$\mathbf{\Phi}^*(\mathbf{x}) = \max_{i=1}^{k} Decay(\mathbf{\Phi}_i(\mathbf{x})) \quad (9)$$

Figure 5 illustrates the synthesized ground truth inhibition maps with respect to the scanpath shown in Figure 4(b). When predicting the first fixation, Figure 5(a), there is no inhibition effect presented. But when predicting the second fixation, Figure 5(b), regions around the first fixation is inhibited. It is worth noting that the second fixation is the re-fixated point of the fifth fixation, thus, inhibition effects around the second fixation disappeared after two time steps when the fifth fixation shown in Figure 5(e) is going to be predicted. Inhibition effects applied on other area decay step-by-step but will not die out.

Each ground truth fixation associates with one unique ground truth ROI map. ROI map is modeled as Gaussian blob, computed with the similar formula as Eqn. 8, centered at the ground truth fixation with the σ setting to 2.5° of visual angle which is the common valid fovea radius.

3.3 Implementation Details

We randomly splitted the OSIE and MIT datasets into 80% training data and 20% test data. We fully trained the IOR-ROI LSTM and saliency guidance part while only fine-tuning

Models	DTW↓	MultiMatch↑				ScanMatch↑
		Vector	Direction	Length	Position	
Inter-subject	1344 (154.71)	0.867 (0.008)	0.708 (0.018)	0.89 (0.008)	0.836 (0.01)	0.451 (0.017)
Itti's model (Ground truth saliency)	2254 (75.36)	0.83 (0.011)	0.639 (0.019)	0.877 (0.012)	0.71 (0.013)	0.263 (0.018)
SGC	2157 (73.4)	0.828 (0.005)	0.634 (0.013)	0.869 (0.003)	0.717 (0.011)	0.261 (0.018)
SaltiNet	2062 (52.26)	0.838 (0.005)	0.655 (0.014)	0.869 (0.004)	0.725 (0.007)	0.279 (0.013)
Wang's model	2089 (47.96)	0.832 (0.016)	0.718 (0.022)	0.834 (0.019)	0.761 (0.02)	0.283 (0.016)
Le Meur's model (Ground truth saliency)	1533 (94.78)	0.849 (0.006)	0.651 (0.012)	0.872 (0.007)	0.765 (0.012)	0.36 (0.021)
Single ConvLSTM	1468 (102.67)	0.863 (0.011)	0.658 (0.024)	0.893 (0.012)	0.797 (0.013)	0.37 (0.023)
Dual ConvLSTM	1449 (92.26)	0.864 (0.01)	0.679 (0.022)	0.892 (0.011)	0.801 (0.014)	0.387 (0.022)
IOR-ROI LSTM	1453 (97.14)	0.867 (0.01)	0.678 (0.023)	0.89 (0.011)	0.803 (0.013)	0.399 (0.021)
IOR-ROI LSTM + Saliency guidance	**1431** (82.78)	**0.868** (0.007)	**0.691** (0.017)	**0.896** (0.009)	**0.806** (0.015)	**0.413** (0.025)

Table 1: Results (with standard deviation in parentheses) of 3 metrics for different models in the OSIE dataset.

Models	DTW↓	MultiMatch↑				ScanMatch↑
		Vector	Direction	Length	Position	
Inter-subject	1072 (300.07)	0.873 (0.004)	0.663 (0.017)	0.889 (0.004)	0.839 (0.008)	0.421 (0.024)
Itti's model (Ground truth saliency)	1539 (209.77)	0.849 (0.006)	0.6 (0.014)	0.894 (0.006)	0.753 (0.011)	0.307 (0.009)
SGC	1769 (212.56)	0.831 (0.004)	0.603 (0.019)	0.858 (0.008)	0.712 (0.007)	0.258 (0.012)
SaltiNet	1957 (129.04)	0.844 (0.003)	0.637 (0.026)	0.865 (0.008)	0.734 (0.003)	0.276 (0.026)
Wang's model	1865 (179.94)	0.809 (0.006)	0.669 (0.028)	0.784 (0.021)	0.731 (0.007)	0.244 (0.015)
Le Meur's model (Ground truth saliency)	1319 (194.52)	0.863 (0.006)	0.631 (0.018)	0.886 (0.005)	0.789 (0.011)	0.349 (0.029)
Single ConvLSTM	1307 (73.58)	**0.884** (0.011)	0.59 (0.033)	0.89 (0.022)	0.826 (0.012)	0.363 (0.03)
Dual ConvLSTM	1298 (161.84)	0.874 (0.014)	0.676 (0.027)	0.898 (0.018)	0.823 (0.017)	0.378 (0.029)
IOR-ROI LSTM	1264 (178.26)	0.876 (0.12)	**0.687** (0.029)	0.902 (0.013)	**0.829** (0.017)	0.385 (0.028)
IOR-ROI LSTM + Saliency guidance	**1231** (141.22)	0.875 (0.006)	0.683 (0.023)	**0.904** (0.01)	0.827 (0.007)	**0.39** (0.031)

Table 2: Results (with standard deviation in parentheses) of 3 metrics for different models in the MIT dataset.

dilated layers of the pre-trained VGG-19 network. Furthermore, training images and fixation coordinates were randomly flipped horizontally to generate more training data. The model was trained using the Adam optimizer along with step learning rate decay strategy which decays learning rate at a constant rate of 0.9. Initial learning rate for the IOR-ROI LSTM and saliency guidance part is 1×10^{-4} while fine-tuning the dilated VGG-19 with an initial learning rate of 1×10^{-5}.

3.4 Evaluation Metric

There are a few scanpath evaluation metrics quantifying different aspects of predicted scanpath performance [Anderson *et al.*, 2015]. The experiments employed three evaluation metrics. We first adopted the Dynamic Time Warp algorithm (DTW) used in [Le Meur and Liu, 2015], which can measure the similarity between two sequences with different lengths. The second evaluation metric, MultiMatch, introduced in [Dewhurst *et al.*, 2012] consists of five separate measures that assess the similarity between two scanpaths with respect to shape, direction, length, position, and duration. We also evaluated scanpath similarity using ScanMatch [Cristino *et al.*, 2010] which is able to take spatial, temporal and sequential similarities into account simultaneously. Similar to the compared benchmark [Le Meur and Liu, 2015], we only evaluate the performance on the spatial distribution aspects, where the fixation duration prediction is out of the scope of our paper.

3.5 Experimental Results

We compared our results against the state-of-the-art scanpath prediction models, *i.e.*, Itti's model [Itti *et al.*, 1998], SGC [Sun *et al.*, 2012], SaltiNet [Assens Reina *et al.*, 2017], Wang's model [Wang *et al.*, 2011] and Le Meur's model [Le Meur and Liu, 2015]. To generate scanpath for a specific image using Itti's and Le Meur's model, we need to provide them with a saliency map corresponding to the image

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

Figure 6: Comparison of generated scanpaths with ground truth human eye-tracking scanpaths. The first row shows the ground truth scanpaths; The second row presents the generated scanpaths.

and, hence, quality of the generated scanpath can be significantly affected by the accuracy of the input saliency map. Thus, in comparison experiments, we fed Itti's and Le Meur's model with the test image and its ground truth saliency map. Moreover, to validate that the superiority of our model does come from incorporating learned IOR dynamics, a performance comparison between scanpath prediction model using just single ConvLSTM and the model with the IOR-ROI LSTM was also conducted. To better demonstrate the effectiveness of the designed IOR-LSTM, we compared the performance of the dual ConvLSTM model, by replacing the IOR-LSTM with original ConvLSTM, with that of the IOR-ROI LSTM. Furthermore, since inter-subject scanpaths similarities evaluate different natural and real human scanpaths. Thus, inter-subject scanpath similarities can be used as a reference quantifying the quality of generated scanpaths. The less performance gap between the generated scanpaths and that of the real human scanpaths, the more realistic the generated scanpaths are.

When comparing scanpaths from various subjects on the same stimulus, it can be observed that they can be either similar or very different. Due to the explicit supervision of IOR behavior during the training phase, the IOR-ROI LSTM tends

(a) 1→2 (b) 2→3 (c) 3→4 (d) 4→5 (e) 5→6 (f) 6→7 (g) 7→8

Figure 7: Visualization of the IOR dynamics and attention shift behavior for each fixation prediction step when generating a scanpath with 8 fixations. The first row shows IOR effects when estimating the next fixation; The second row shows heat maps of attention in which the region with the highest value is selected as the next fixation position.

to be confused if we provide it with scanpaths of the same image from different observers with very different IOR behavior. Therefore, we trained the IOR-ROI LSTM for each subject separately to model a consistent IOR dynamics and ROI shift behavior. We randomly selected 10 subjects from the OSIE and MIT dataset for training, and evaluation scores of each subject-specific model were averaged as the final result.

For each model, 10 scanpaths were generated for a given image. The length of each scanpath was determined as mean scanpath length of the training dataset, *i.e.*, 8 fixations for scanpaths in both the OSIE and MIT datasets. All three evaluation metrics were computed by comparing each predicted scanpath to all subjects' ground truth scanpath and final results were the average of the 10 scores of each evaluation metric. To measure inter-subject performance, we first evaluated the performance of each subject by treating scanpaths of other subjects as the ground truth. Then, the average value of all subjects was used as inter-subject performance.

Tables 1 and 2 report performance of different models on the OSIE and MIT datasets in terms of the mentioned evaluation metrics, respectively. From Table 1, we can see that the model combined the IOR-ROI LSTM with saliency guidance network outperforms all the other models, except the inter-subject reference model, in all three evaluation metrics. However, as presented in Table 2, MultiMatch vector measure of the IOR-ROI LSTM based model does not perform very well when compared with that of the single ConvLSTM model. This may be caused by the inconsistency of image content type and image scale in the MIT dataset, which is also confirmed by [Wang *et al.*, 2017]. The overall performance of scanpath prediction using just a single ConvLSTM is higher than that of Le Meur's saccadic scanpath prediction model, while the IOR-ROI LSTM model can significantly boost scanpath generation performance, which clearly indicates that performance improvement does come from the advantage of explicitly modeling of IOR dynamics and attention shift. Furthermore, the incorporating of saliency information brings another performance gain, especially in the ScanMatch. The dual ConvLSTM can achieve similar performance in separate measures of the MultiMatch when compared with that of the IOR-ROI LSTM model. But it can be obviously observed that the IOR-ROI LSTM model outperforms the dual ConvLSTM model in the ScanMatch, which suggests that our proposed IOR-LSTM, combined with the ConvLSTM for ROI prediction, can better capture sequential characteristics of scanpaths in the test dataset. It is also worth noting that our proposed model can achieve comparable results in the MultiMatch vector, direction, and length similarity metrics when compared with that of the inter-subject, which means that the proposed model is able to well capture the overall shape, saccade direction and saccade amplitude characteristics of scanpaths in the test dataset.

In addition to qualitative evaluation, Figure 6 illustrates the visual comparison of scanpaths generated by our proposed model with the ground truth human eye-tracking scanpaths. As is shown in the figure, fixations of the generated scanpath almost cover the same locations as fixations of the ground truth scanpath do. We can also observe that the generated scanpaths have high shape similarity with the ground truth scanpaths, which manifests the effectiveness of our proposed scanpath prediction model and also gives a well visual explain on why the proposed model can obtain comparable performance in MultiMatch vector, direction and length measures shown in Tables 1 and 2.

We also visualized the IOR dynamics and attention shift behavior for each fixation prediction step in Figure 7. As shown in the first row of Figure 7(a), when generating the second fixation, surrounding region of the first fixation is strongly inhibited and its corresponding region in the ROI map, shown in the second row of Figure 7(a), has relatively low value, then another salient object is going to be attended. Another thing should be noted is that as the scanpath proceeds, IOR effects on previously fixated location decay adaptively and ROI value of that region can gradually recover to a high level, which finally results in the re-fixation. This process is illustrated in Figure 7(c). When predicting the fourth fixation, inhibition in the first fixation declines below a threshold and ROI of that area becomes the highest. Figure 7(d)(f)(g) also demonstrate the occurrence of re-fixation.

4 Conclusion

In this paper, we present a model to predict scanpath. In the proposed model, a novel IOR-ROI LSTM module can simultaneously model IOR dynamics and attention shift behavior. Compared with existing computational scanpath estimation approaches which model IOR in a linear manner, our IOR-LSTM, simulating the VWM, learns IOR dynamics based on visual features extracted by CNNs. Experimen-

tal results show that our proposed scanpath prediction model outperforms state-of-the-art computational saccadic scanpath generation approach. Explicitly modeling IOR in scanpath estimation can effectively boost fidelity of the generated scanpath.

References

[Anderson *et al.*, 2015] Nicola C. Anderson, Fraser Anderson, Alan Kingstone, and Walter F. Bischof. A comparison of scanpath comparison methods. *Behav. Res. Methods*, 47(4):1377–1392, 2015.

[Assens Reina *et al.*, 2017] Marc Assens Reina, Xavier Giró-i Nieto, Kevin McGuinness, and Noel E. O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *ICCV Workshop on EPIC*, 2017.

[Bays and Husain, 2012] Paul M. Bays and Masud Husain. Active inhibition and memory promote exploration and search of natural scenes. *J. Vis.*, 12(8):8, 2012.

[Boccignone, 2016] Giuseppe Boccignone. A probabilistic tour of visual attention and gaze shift computational models. *arXiv preprint arXiv:1607.01232*, 2016.

[Cornia *et al.*, 2017] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. In *ICVSS*, 2017.

[Cristino *et al.*, 2010] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D. Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behav. Res. Methods*, 42(3):692–700, 2010.

[Dewhurst *et al.*, 2012] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behav.Res. Methods*, 44(4):1079–1100, 2012.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[Hotta *et al.*, 2010] Shinji Hotta, Shigeyuki Oba, and Shin Ishii. Visual attention model involving feature-based inhibition of return. *Artif. Life Robotics*, 15(2):129–132, 2010.

[Itti *et al.*, 1998] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[Jiang *et al.*, 2016] Ming Jiang, Xavier Boix, Gemma Roig, Juan Xu, Luc Van Gool, and Qi Zhao. Learning to predict sequences of human visual fixations. *IEEE Trans. Neural Netw. Learn. Syst.*, 27(6):1241–1252, 2016.

[Judd *et al.*, 2009] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[Kümmerer *et al.*, 2017] Matthias Kümmerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *ICCV*, 2017.

[Le Meur and Liu, 2015] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision Res.*, 116:152–164, 2015.

[Le Meur *et al.*, 2017] Olivier Le Meur, Antoine Coutrot, Zhi Liu, Pia Rämä, Adrien Le Roch, and Andrea Helo. Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood. *IEEE Trans. Image Processing*, 26(10):4777–4789, 2017.

[Ngo and Manjunath, 2017] Thuyen Ngo and B.S. Manjunath. Saccade gaze prediction using a recurrent neural network. In *ICIP*, 2017.

[Samuel and Kat, 2003] Arthur G. Samuel and Donna Kat. Inhibition of return: A graphical meta-analysis of its time course and an empirical test of its temporal and spatial properties. *Psychon Bull. Rev.*, 10(4):897–906, 2003.

[Shi *et al.*, 2015] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sun *et al.*, 2012] Xiaoshuai Sun, Hongxun Yao, and Rongrong Ji. What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In *CVPR*, 2012.

[Wandell, 1995] Brian A. Wandell. *Foundations of Vision*. Sinauer Associates, 1995.

[Wang and Klein, 2010] Zhiguo Wang and Raymond M. Klein. Searching for inhibition of return in visual search: A review. *Vision Res.*, 50(2):220–228, 2010.

[Wang *et al.*, 2011] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. Simulating human saccadic scanpaths on natural images. In *CVPR*, 2011.

[Wang *et al.*, 2017] Yixiu Wang, Bin Wang, Xiaofeng Wu, and Liming Zhang. Scanpath estimation based on foveated image saliency. *Cogn. Process.*, 18(1):87–95, 2017.

[Xu *et al.*, 2014] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *J. Vis.*, 14(1):28, 2014.

[Yu and Koltun, 2016] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[Zhang and Lin, 2013] Liming Zhang and Weisi Lin. *Selective visual attention: computational models and applications*. John Wiley & Sons, 2013.