FOVEATED NEURAL NETWORK: GAZE PREDICTION ON EGOCENTRIC VIDEOS

Mengmi Zhang[⋆†] *Keng Teck Ma*[⋆] *Joo Hwee Lim*[⋆] *Qi Zhao*[†,‡]

[⋆] Institute for Infocomm Research, Astar, Singapore
[†] National University of Singapore, Singapore
[‡] University of Minnesota, USA

ABSTRACT

A novel deep convolution neural network is proposed to predict gaze on current frames in egocentric videos. Inspired by human visual system, we introduce a fovea module responsible for sharp central vision and name our model as Foveated Neural Network (FNN). The retina-like visual inputs from the region of interest on the previous frame are analysed and encoded. The fusion of the hidden representations of the previous frame and the feature maps of the current frame guides the gaze prediction on the current frame. In order to simulate motion, we also include the dense optical flow between these adjacent frames as additional input. Experimental results show that FNN outperforms the state-of-the-art algorithms in the publicly available egocentric dataset. The analysis of FNN demonstrates that the hidden representations of the foveated visual input from the previous frame as well as the motion information between adjacent frames are efficient in improving gaze prediction performance in egocentric videos.

Index Terms— Visual Attention, Saliency, Egocentric Videos, Gaze, Fovea

1. INTRODUCTION

One important property of human perception is that we focus selectively on parts of the visual world at one time and allocate processing resources on the primary region of interests in high resolution. This property enables us to reduce the complexity of the scene and ignore the distraction from irrelevant features. In line with this fundamental role in human perception, attentional modeling has been extensively studied in computer vision. In this paper, we are interested in predicting gaze locations (where humans look) on current frames in egocentric videos. Different from normal videos, egocentric videos are recorded from a first person perspective and involve complex motions due to head movements.

Previous works for estimating human visual attention are based on saliency detection [1]. There are both bottom-up [2, 3, 3, 4, 5, 6, 7] and top-down streams [8, 9, 10]. The pioneering saliency prediction models adopted feature integration theory [11] where fusion of low level features, such as color, contrast, and intensity, attract human visual attention.

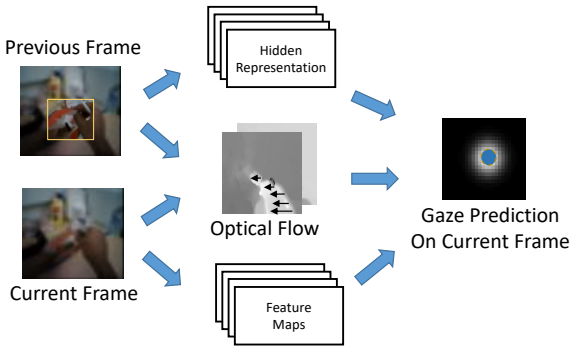


Fig. 1. Flowchart of gaze prediction in egocentric videos. Given the previous frame, the region of interest (yellow rectangle) is foveated and encoded into hidden representations. The optical flow is computed between the previous and current frames to simulate motion. The feature maps are extracted from the current frame. The hidden representations of the foveated visual inputs, the feature maps of the current frame, and the optical flow between these adjacent frames can then be used for predicating the gaze location (blue dot).

Huang et al. had greatly improved saliency prediction on images by leveraging on rich pools of semantic regions or objects in the scene from deep convolution neural network (2D-CNN)[12]; but the temporal information is discarded. Bazzani et al. extended the saliency prediction on static images in the spatial domain to the temporal saliency prediction in normal videos using Long Short Term Memory (LSTM)[13]. Meur and Coutrot's model incorporate systematic bias in semantic visual category for scanpath prediction [14]. There is another recurrent visual attention model where Mnih et al.'s algorithm predicted sequence of fixations on images [15]. One of the most related works, Li et al. directly addressed gaze prediction problem on egocentric videos where they predefined egocentric cues [16], *e.g.* hand poses.

We propose a novel deep neural network for gaze prediction on current frames on egocentric videos. Inspired by the foveal system of human eyes, we introduce a foveated mechanism to process visual inputs and name our model as Foveated Neural Network (FNN). To avoid accumulating errors by feeding the predicted gaze back to the model using recurrent

neural network, we use a feed-forward 2D-CNN where only the previous and current frames are required. The flowchart is shown in Figure 1.

2. PROPOSED MODEL

We first introduce an overview of our model, named as Foveated Neural Network (FNN), followed by a detailed analysis of each module in FNN. We provide training and implementation details in the end.

2.1. Architecture Overview

We formulate the gaze prediction problem on the current frame of egocentric videos as: given the previous frame and the current frame, FNN outputs the saliency map for the current frame. Hence, the spatial coordinate with the maximum probability on the saliency map is the predicted gaze location. The overview of FNN is presented in Figure 2. FNN divides into three modules: *Pre-process Module (PP)*, *Fovea Module (F)*, and *Re-alignment and Post-process Module (RP)*.

In *PP*, based on the current frame of low resolution, FNN extracts the feature maps useful for gaze prediction and estimates the region of interest (ROI) on the current frame. The center of ROI will be used in the next iteration. In *F*, given ROI on the previous frame, *F* simulates the human fovea and outputs the feature maps extracted from the retina-like image patches centered over ROI. They are of different resolution and cover different sizes of the receptive field. The patch covering the large receptive field is of low resolution while the one covering the small receptive field is of high resolution. In *RP*, the extracted feature maps from the patches are re-aligned based on the center of ROI and concatenated with the feature maps extracted from the current frame. The combined feature maps are post-processed and output the refined saliency map and hence, the predicted gaze location on the current frame.

We define an egocentric frame I of low resolution and high resolution using superscript l and h respectively. The subscript denotes time t. A saliency map is defined as a probability distribution of gaze locations; thus, the spatial coordinate of the maximum probability in the saliency map is the predicted gaze location f^r. Similarly, we use the estimated saliency map obtained from the low-resolution frame to propose ROI centered at f^c. We use superscript r as the refined gaze location (the output of FNN) and superscript c as the center of the proposed ROI.

2.2. Fovea Module

Given an egocentric high-resolution frame I^h_{t-1} and the center of ROI f^c_{t-1} at time $t-1$, ROI is attended in a foveated manner. In order to simulate the attentional processing in the retina, we use the same approach as [15]. Instead of assessing the frame in high resolution across all pixels, F extracts the retina-like representation focused on f^c_{t-1}, *i.e.*different image patches of limited bandwidths centered at f^c_{t-1}. In our case, we use three bandwidths: $H \times H$, $\frac{H}{2} \times \frac{H}{2}$ and $\frac{H}{4} \times \frac{H}{4}$; however, not limited to three, F can be generalized to more than three depending on the applications. When the receptive field centered at f^c_{t-1} exceeds the frame boundary, we use zero padding to fill in the empty areas. These multiple resolution patches are then scaled to the same size $\frac{H}{4} \times \frac{H}{4}$. This is to simulate the fovea where the patch covering small receptive field (Patch1) is of high resolution whereas the patch covering large receptive field (Patch3) is downsampled to be of low resolution. Thus, it enables F to allocate the small amount of processing power (the same number of parameters in 2D-ConvNetPatch) on the large area of the frame in low resolution (Patch3) and vice versa.

As shown in [12], convolution layers of high levels in 2D convolution neural network (2D-ConvNet) trained for object recognition are effective in predicting saliency. We use the pre-trained 2D-ConvNet on ImageNet for feature extraction. The feature maps from these multiple resolution patches are extracted using branches of 2D-ConvNetPatch. The branches have the same architecture and share the same network parameters. The outputs of F are feature maps denoted as $FP1_{t-1}$, $FP2_{t-1}$ and $FP3_{t-1}$ respectively. Each of their feature maps are of size $\frac{H}{16} \times \frac{H}{16}$.

2.3. Pre-process Module

Before assessing to ROI of the current frame in high resolution, I^l_t (size $\frac{H}{2} \times \frac{H}{2}$) is perceived in low resolution at time t. *PP* uses 2D-ConvNetPreprocess for encoding features of I^l_t and 2D-ConvNetCoarse for proposing the ROI. As egocentric videos involve head motions, we compute the dense optical flow OF_t between I^l_t and I^l_{t-1} from [17] and use it to implicitly represent motions between adjacent frames. 2D-ConvNetPreprocess takes five channels as inputs: RGB channels from I^l_t and OF_t in horizontal and vertical axis. We denote the output from 2D-ConvNetPreprocess as feature maps FP_t with each feature map of size $\frac{H}{16} \times \frac{H}{16}$. FP_t and $FP1_{t-1}$, $FP2_{t-1}$, $FP3_{t-1}$ from F are of the same size and they will be used for predicting gaze location on the current frame. Based on FP_t extracted from I^l_t, 2D-ConvNetCoarse proposes one ROI where the model may be interested in focusing attention on. The ROI is represented using the center of ROI denoted as f^c_t. f^c_t is obtained by taking the spatial coordinate of the maximum on the saliency map estimated from I^l_t in low resolution. It will be used in F in the next iteration (time $t+1$) where FNN predicts the next gaze location on frame I^l_{t+1}.

2.4. Re-alignment and Post-process Module

After obtaining $FP1_{t-1}$, $FP2_{t-1}$ and $FP3_{t-1}$, *RP* realigns these feature maps based on f^c_{t-1}. The realignment process includes the following steps: 1). scale $FP1_{t-1}$, $FP2_{t-1}$ and

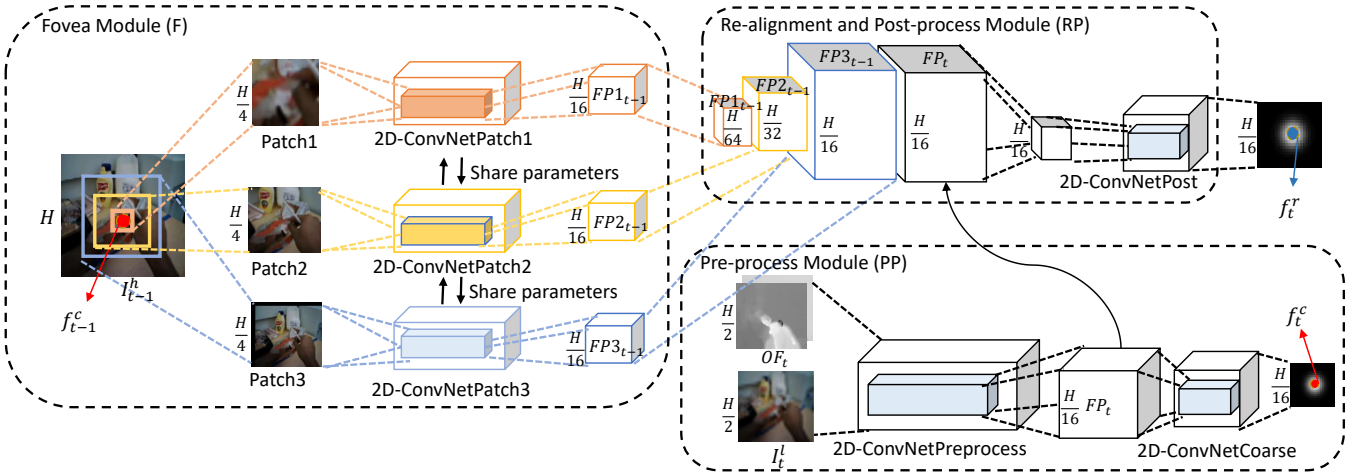


Fig. 2. Architecture of our model for Gaze Prediction on Current Frame. There are three modules: *Fovea Module (F)*, *Pre-process Module (PP)* and *Re-alignment and Post-process Module (RP)*. In *PP*, the inputs to 2D-ConvNetPreprocess are 5 channels: RGB channels from I_t^l and the optical flow OF_t in horizontal and vertical axis. Its outputs are the feature maps FP_t. 2D-ConvNetCoarse outputs the estimated region of interest centered at f_t^c (red dot). f_t^c will then be used in the next iteration (time $t+1$). In *F*, given the high-resolution frame I_{t-1}^h and its region of interest centered at f_{t-1}^c, it extracts Patch1, 2, 3 with the different receptive coverage. These 3 patches are scaled to be of the same size. In *RP*, the outputs from PP (feature maps $FP1$, $FP2$ and $FP3$) are realigned based on f_{t-1}^c. The concatenated feature maps from $FP1_{t-1}$, $FP2_{t-1}$, and $FP3_{t-1}$ together with FP_t are the inputs to 2D-ConvNetPost for estimating the refined saliency map at time t. The coordinate with the maximum probability in the saliency map is the refined gaze location f_t^r at time t (blue dot).

$FP3_{t-1}$ to $\frac{H}{64} \times \frac{H}{64}$, $\frac{H}{32} \times \frac{H}{32}$ and $\frac{H}{16} \times \frac{H}{16}$ respectively; 2). add in zero paddings to each of the four sides of each feature map by $\frac{3H}{128}$ in $FP1_{t-1}$ and $\frac{H}{64}$ in $FP2_{t-1}$; 3). shift the concatenated feature maps back to f_{t-1}^c with respect to I_{t-1}^h. The realignment process is used for consolidating all the feature maps across multiple resolution patches to the same spatial location with respect to I_{t-1}^h.

In 2D-ConvNetPost, we use one 2D convolution layer to fuse the consolidated information on the previous frame together with FP_t from the current frame. The fused information is post-processed by another two fully connected layers before generating the final predicted saliency map of size $\frac{H}{16} \times \frac{H}{16}$. The coordinate with the maximum probability in the saliency map is the predicted gaze location f_t^r on I_t^l.

2.5. Training and Implementation Details

We train FNN in stochastic gradient descent with learning rate 0.01 and batch size 1. The fixation map (the ground truth) is defined as the binary map with human gaze locations. As a common practice, we put an isotropic gaussian mask over the binary map and normalize it to be $[0, 1]$. Same as [12], we minimize Kullback-Leibler divergence (KLD) loss between the predicted saliency map and the fixation map. All the weights from 2D-ConvNet in FNN are pre-loaded using VGG-16 trained on ImageNet [18]. The parameter H is set to be 1200. All the numbers of feature channels for $FP1_{t-1}$, $FP2_{t-1}$, $FP3_{t-1}$ and FP_t are 512. The input frames to FNN are normalized to $[0, 1]$ with mean and standard deviation. We implement the proposed algorithm in Torch.

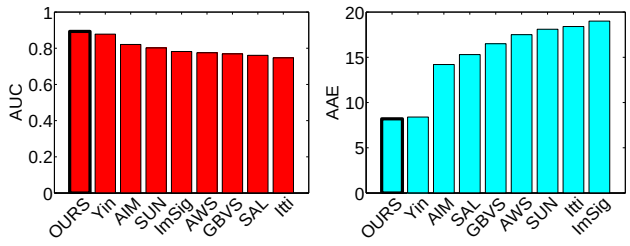
We evaluate FNN using the publicly available egocentric dataset, GTEA [9]. It contains 17 video sequences in total with each video lasting for 4 minutes on average. 14 human subjects are asked to prepare for meals in a kitchen at their own wishes while wearing the eye-tracking devices. For fair comparison, we choose videos 1, 4, 6-22 as training and validation sets while the rest are used for testing same as [16].

3. EXPERIMENT

We compare FNN with the state-of-the-art using standard evaluation metrics on one publicly available dataset. In the following subsections, we introduce the evaluation metrics and comparative methods. In the end of the section, we present the results and the detailed analysis.

3.1. Metrics

We used two standard evaluation metrics to measure the performance of gaze prediction: Area Under the Curve (AUC) [19] and Average Angular Error (AAE) [16]. AUC is commonly used in the saliency prediction literature. It measures the consistency between a predicted saliency map and a fixation map of human gazes. AAE is used in the gaze tracking literature and measures the error between the predicted and the human gaze locations in an angular distance.



(a) AUC Evaluation (b) AAE Evaluation

Fig. 3. Results on GTEA Dataset. Evaluation of Gaze Prediction using Area Under the Curve (AUC) in (a) and using Average Angular Error (AAE) in (b). The comparative methods are introduced in Section 3.2

	AAE	AUC
SALICON (SAL)	16.5	0.76
SAL + 2 Fully Connected Layers (FC)	10.6	0.80
SAL + FC + OpticalFlow (OF)	8.33	0.88
SAL+ FC + OF + FoveaOnPreviousFrame	**8.15**	**0.89**

Table 1. Evaluation of Ablated Models. From top to bottom, only one component is added into the previous model at one time. They are evaluated using Average Angular Error (AAE) and using Area Under the Curve (AUC). Number denoted in bold is the best.

3.2. Comparative Methods

We compare our method with the state-of-the-art saliency prediction algorithms: Graph-based Visual Saliency (GBVS) [3], Saliency Using Natural Statistics (SUN) [4], Adaptive Whitening Saliency (AWS) [5], Attention based on Information Maximization (AIM) [6], Itti's Model (Itti) [20], Image Signature Saliency (ImSigLab) [7] and SALICON [12]. In particular, SALICON is a 2D-ConvNet with the current frame as the only input. We fine-tune SALICON on the training set and evaluate its predicted saliency maps in the test set.

In addition, we include [16] as it directly addresses the gaze prediction problem on egocentric videos by using Hidden-Markov model for the temporal dynamics.

3.3. Results and Analysis

The results in AUC and AAE are presented in Figure 3. FNN outperforms the state-of-the-art algorithms on gaze prediction on current frames in egocentric videos in both AAE and AUC.

Compared with saliency prediction algorithms, FNN yields a significant boost in gaze prediction performance. Though SALICON learns the semantic features useful for gaze prediction, it fails to take temporal information into account. See the ablation study in Table 1 (Row 3).

Though Li's work [16] uses the hidden markov model for temporal dynamics, FNN performs better with an improvement of 2.4% ($(8.33 - 8.18)/8.33 = 2.4\%$) in AAE due to the enriched pool of semantic feature representations in the network and the fovea module on the previous frame.

To further explore the effect of individual components introduced in FNN, we conduct an ablation study and report the results in Table 1. We build up FNN based on SALICON and we add in one component at a time. SALICON is a feedforward 2D-ConvNet with the last few fully connected layers removed. We added in 2 fully connected layers in the end which boosts up the performance to a significant extent in terms of AAE (Row 2). Compared with SALICON containing only convolution and pooling operations within a local receptive field, we hypothesize that the added 2 fully connected layers fuse all the information across space and increase the capacity of saliency representations.

To study the effect of the foreground and background motions, we add in the dense optical flow between the current frame and the previous frame as inputs to the network (Row 3). The first convolution layer has two additional input channels. The results improve by 2 in AAE and 0.08 in AUC. It suggests that the motion estimation between adjacent frames is an important egocentric cue for gaze prediction.

We present the result of FNN (Row 4). Compared with the one in Row 3, we add in the fovea module and fuse its feature maps with the one-stream network. Result shows an improvement of 0.18 in AAE and 0.01 in AUC. It explains that the integration of the foveated information on the previous frame is useful for predicting gaze on the current frame.

According to [16], there exists a strong center bias for gaze distributions on current frames in egocentric videos since the large gaze shift often gets compensated by the head motions. Hence, we use sAUC to evaluate FNN and compare it with the center bias. We create the artificial center as the predicted gaze location and we put an isotopical gaussian mask over the center for sAUC evaluation. We report sAUC results in GTEA: FNN (0.65) and center bias (0.5). It confirms that FNN predicts gaze locations more than center bias.

4. CONCLUSION

We present a novel foveated neural network for gaze prediction on egocentric videos. Evaluation results on the publicly available dataset demonstrate that FNN outperforms the state-of-the-art methods. The integration process of proposing, attending and analysing ROI on the previous frame as well as the feature extraction from the current frame helps gaze prediction performance. We also incorporate head movement to FNN by introducing the dense optical flow as the additional feature inputs. We will extend FNN to more than two adjacent frames by introducing a memory module in the near future.

5. ACKNOWLEDGEMENTS

This work was supported by the Reverse Engineering Visual Intelligence for cognitive Enhancement (REVIVE) programme funded by the Joint Council Office of A*STAR, Singapore.

6. REFERENCES

[1] Christof Koch and Shimon Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of intelligence*, pp. 115–141. Springer, 1987.

[2] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI*, , no. 11, pp. 1254–1259, 1998.

[3] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[4] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, pp. 32–32, 2008.

[5] Antón Garcia-Diaz, Xosé R Fdez-Vidal, Xosé M Pardo, and Raquel Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012.

[6] Neil Bruce and John Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2005, pp. 155–162.

[7] Xiaodi Hou, Jonathan Harel, and Christof Koch, "Image signature: Highlighting sparse salient regions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[8] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.," *Psychological review*, vol. 113, no. 4, pp. 766, 2006.

[9] Alireza Fathi, Yin Li, and James M Rehg, "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision*. Springer, 2012, pp. 314–327.

[10] Ali Borji, Dicky N Sihite, and Laurent Itti, "Probabilistic learning of task-specific visual attention," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 470–477.

[11] Anne M Treisman and Garry Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[12] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *IEEE ICCV*, 2015, pp. 262–270.

[13] Loris Bazzani, Hugo Larochelle, and Lorenzo Torresani, "Recurrent mixture density network for spatiotemporal visual attention," *arXiv preprint arXiv:1603.08199*, 2016.

[14] Olivier Le Meur and Antoine Coutrot, "Introducing context-dependent and spatially-variant viewing biases in saccadic models," *Vision research*, vol. 121, pp. 72–84, 2016.

[15] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., "Recurrent models of visual attention," in *NIPS*, 2014, pp. 2204–2212.

[16] Yin Li, Alireza Fathi, and James M Rehg, "Learning to predict gaze in egocentric video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3216–3223.

[17] Thomas Brox, Christoph Bregler, and Jitendra Malik, "Large displacement optical flow," in *CVPR*. IEEE, 2009, pp. 41–48.

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

[19] Ali Borji, Hamed R Tavakoli, Dicky N Sihite, and Laurent Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *IEEE ICCV*. IEEE, 2013, pp. 921–928.

[20] Laurent Itti and Christof Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10, pp. 1489–1506, 2000.