H-006

Predicting Focus of Attention of Elderly Drivers

Onkar Krishna† Go Irie† Takahito Kawanishi† Kunio Kashino† Kiyoharu Aizawa‡

1. Introduction

Saliency models are frequently used for predicting focus of attention (FoA) while viewing a scene. Significant efforts have been made on developing these models. For example, [1] and [2] have proposed a data driven model for saliency prediction in free-viewing and car driving scenarios, respectively. However, existing models are mostly focused on and evaluated for adult's FoA, hence cannot be applied to the prediction of elderly's.

Motivated by these observations, we consider a problem of predicting elderly's FoA in this paper. A straight forward approach would be to train a prediction model, e.g., a deep convolutional neural network (CNN), on the FoA data of elderly participants. However, collecting a sufficient amount of training data of elderly's is challenging due to their physical or health conditions, hence, data-efficient approaches are needed.

In this paper, we consider a new framework to tackle this issue. Although the FoAs by adults and elderly generally look different, their tendencies can still be well characterized by the scene they watch. Based on this idea, we consider to leverage the knowledge of the FoA predictors constructed for the adults, e.g., [2], to predict the elderly's. To this end, we propose a new framework based on deep image translation. That is, given a scene, our method translates the saliency map of adults predicted by a state-of-the-art method to that of the elderly's, where the mapping is obtained by an encoder-decoder-type deep CNN.

Although a few recent attempts consider age-dependent saliency models [3, 4], our approach is based on image-to-image translation and is totally different from them. Evaluation experiments are done on the FoA dataset collected over elderly while driving on a car driving simulator. Results show that our model gives remarkable prediction accuracy.

2. Method

Given a sequence of video frames, our task is to predict which part of each frame an elderly person would pay attention while viewing the video sequence in a car driving task.

Following the previous paper [2], we consider to sequentially predict the FoA map of each frame. Specifically, given a sequence of k successive frames denoted by $\mathfrak{F}_n = \{f_{n-k+1}, f_2, \cdots, f_n\}$, we predict the FoA on f_n in the form of the probability of how likely each pixel of f_n is attended by an observer. We use $k = 16$ throughout this paper. We denote the predicted FoA of elderly and adults for f_n by e_n and a_n, respectively. The ground truth FoA of elderly observers is denoted by e_n^*.

2.1 Model

We approach to the task in the following two steps. We first predict the adult's FoA a_n, and then transform it to the elderly's FoA e_n. To this end, we propose a model that is designed to complete these two steps. The schematic overview of our model is illustrated in Fig. 1. In the first step, given an input sequence of a video, the adult's FoA is predicted by using a *predictor network* which is trained on a large fixation data of adult observers [2]. In the second step, the *translator network* is used to get the elderly's FoA map from the predicted FoA of adults. The translator network has a simple encoder-decoder architecture designed to have eight 4×4 convolution/deconvolution layers followed by the rectified linear activation (ReLU) function. We use another network, i.e., *discriminator network*, to facilitate the training of the translator network in an adversarial learning framework.

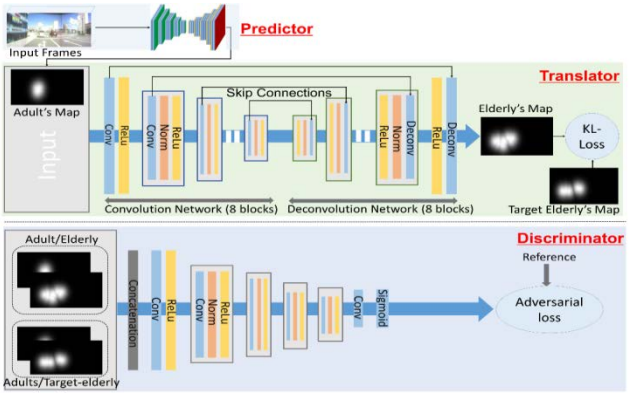

Fig. 1. Overview of our framework.

2.2 Training

Let us denote the translator and discriminator as T and D, respectively. Overall, we train these two networks in a unified training framework that solves the following optimization problem with respect to their parameters.

$$\min_D \max_T \mathcal{L}_{\mathrm{Adv}}(T, D) - \gamma \mathcal{L}_{\mathrm{KL}}(T) \qquad (1)$$

This objective is consisted of two major loss terms; content loss $\mathcal{L}_{\mathrm{KL}}$ and adversarial loss $\mathcal{L}_{\mathrm{Adv}}$.

Content Loss. The content loss requires the translator network T to output the ground truth FoA e_n^* in per pixel bases. The loss function used here is KL-divergence which allows us to directly compare the two probability distributions denote by $e_{n,i}$ and $e_{n,i}^*$. Specifically it is given in the following form.

$$\mathcal{L}_{\mathrm{KL}}(T) = \sum_n \sum_i e_{n,i}^* \left(\log\left(e_{n,i}^*\right) - \log\left(e_{n,i}\right)\right) \qquad (2)$$

Adversarial Loss. The content loss defined above only aims at approximating the conditional distribution of e_n, i.e., $p(e_n; \mathfrak{F}_n)$, which may not fully capture the underlying correlation between e_n and a_n. To model the joint distribution of $p(e_n, a_n; \mathfrak{F}_n)$, we

† NTT Communication Science Laboratories
‡ The University of Tokyo

Fig. 2. Qualitative results of the predicted FoA maps.

introduce an additional loss based on adversarial learning. This is specifically formulated as follows.

$$\mathcal{L}_{\mathrm{Adv}}(T,D) = \mathbb{E}_{(a_n,e_n^*)\sim p(a_n,e_n^*;\,\Im_n)}\left[\log\left(D(a_n,e_n^*)\right)\right] + \mathbb{E}_{a_n\sim p(a_n;\,\Im_n),e_n\sim p(\,e_n|a_n)}\left[1-\log\left(D(a_n,e_n)\right)\right] \quad (3)$$

3. Experiments

We empirically demonstrate the effectiveness of our framework on a car driving scenario, which is one of the most important applications of FoA estimation.

3.1 Setup

Dataset. Since there is no existing dataset that is suitable for our case, we collected our own dataset by using a car driving simulator. The detail of our dataset construction process is described as follows. 18 observers belonging to two different age groups, adults and elderly, were recruited. The adult and elderly observers had mean age of 26 and 75 years, respectively. All the observers had normal or corrected to the normal vision. Each participant was asked to use the driving simulator and to safely drive a car to reach a certain destination. The simulator shows a video sequence to each participant consisting of 10,000 frames of road environment. An eye-tracking system called 'Smart-Eye' is used for recording the eye-gaze movement of each participant while driving in real time. The recorded fixation maps are overlaid frame-by-frame on each video after dynamic time warping (DTW) based frame alignments. We obtained 9,652 continuous fixation maps correspond to the 9,652 frames of the video stimuli for both age groups.

Training Details. We trained our translator and discriminator networks from scratch during 100 epochs by using Adam with the learning rate of 0.0002 and momentum of 0.5. The weight for the content loss (KL-divergence term) γ is fixed to 100.

3.2 Quantitative Results

We first evaluate our model and compare it with several existing models [2,5-7] in terms of three standard metrics for FoA prediction, including Pearson's Correlation Coefficient (CC), Similarity (SIM), and KL-Divergence. The results are shown in Table 1. We can see that our model outperforms all the baselines. Compared to the bottom-up unsupervised approaches [5-7], the performance gain of our method is very high. The reason for the severe failure of these models is that they are based on bottom-up features, which is suitable in predicting FoA during free-viewing. The performance of the top-down approach [2] is the most competitive among all the baselines. However, our model still outperforms it.

Table 1. Prediction accuracy of our model in comparison with baseline over the test dataset.

Algorithm	CC ↑	SIM ↑	KL-div. ↓
Ours	**0.7717**	**0.6832**	**2.18**
DrEyeVe[2]	0.6386	0.5324	4.06
[2] (fine-tuned)	0.6575	0.5535	-
Wang15 [5]	0.1305	0.2232	5.60
Wang [6]	0.0901	0.2595	4.90
ML-net[7]	0.1478	0.2940	-

3.3 Qualitative Results

We next show some qualitative results in Fig. 2. These results show the excellent ability of our deep image translation network in predicting elderly's FoA in a driving scenario. The first example (first row) shows that while taking the right turn our model accurately mimic the elderly's attention at the corner of the pedestrian crossing, whereas the other models fails to do so.

4. Conclusion

We attempted to predict the elderly driver's FoA based on image-to-image translation. We have empirically proved that our model outperformed several existing methods.

Acknowledgement. We would like to thank Prof. Kimihiko Nakano of The University of Tokyo for availing the driving simulator and assisting with the data collection.

References

[1] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model," IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 5142–5154, 2018.
[2] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara, "Predicting the driver's focus of attention: dr(eye)ve project," IEEE transactions on pattern analysis and machine intelligence, 2018.
[3] Onkar Krishna, and Kiyoharu Aizawa. "Age-adapted saliency model with depth bias." In Proceedings of the ACM Symposium on Applied Perception, p. 5. ACM, 2017.
[4] Onkar Krishna, Andrea Helo, Pia R¨am¨a, and Kiyoharu Aizawa, "Gaze distribution analysis and saliency prediction across age groups," PloS one, vol. 13, no. 2, pp. e0193149, 2018.
[5] Wenguan Wang, Jianbing Shen, and Fatih Porikli, "Saliency-aware geodesic video object segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3395–3402.
[6] Wenguan Wang, Jianbing Shen, and Ling Shao, "Consistent vide saliency using local gradient flow optimization and global refinement," IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 4185–4196, 2015.
[7] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "A Deep Multi-Level Network for Saliency Prediction," in International Conference on Pattern Recognition (ICPR), 2016.