GAZEATTENTIONNET: GAZE ESTIMATION WITH ATTENTIONS

Haoxian Huang[1] *Luqian Ren*[1] *Zhuo Yang*[1*] *Yinwei Zhan*[1] *Qieshi Zhang*[2] *Jujian Lv*[3]

[1]School of Computer Science and Technology, Guangdong University of Technology
[2]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3]School of Computer Science, Guangdong Polytechnic Normal University

ABSTRACT

Predicting gaze point on mobile devices without calibration in unconstrained environments has great significance on human computer interaction. Appearance-based gaze estimation methods have been improved due to the recent advance in convolutional neural network (CNN) models and the availability of large-scale datasets. CNN models have limitations on extracting the global information of features and ignore the important information of local features. In this paper, we propose a novel structure named GazeAttentionNet. To improve the accuracy of gaze estimation, we use the global and local attention modules to utilize both global and local features. Firstly, we use MobileNetV2 and the self-attention layers as the global attention module to extract global features. Secondly, we add the local attention module containing the spatial attention to extract local features. With GazeAttentionNet, we achieve an excellent result on the GazeCapture dataset. The average errors of mobile phones and tablets are 1.67 cm and 2.37 cm.

Index Terms— gaze estimation, self-attention, spatial attention

1. INTRODUCTION

Gaze is an important part of humans in the visual world. We can acquire information from our surroundings while expressing our potential thoughts through the gaze. Recently, gaze estimation has been widely used in many scientific research fields, including human computer interaction [1], assisted driving [2] and psychology [3]. Also, with the development of hardware, gaze estimation researches recently focus on utilizing commodity hardware like webcams or the front-facing cameras available in mobile phones and tablet devices. The images captured by these devices can be directly used as inputs of the appearance-based gaze estimation models.

The methods of gaze estimation can be divided into model-based and appearance-based approaches [4]. The model-based approaches are firstly used to address the gaze estimation problem by using geometry features of eyes. Recently, the availability of large-scale datasets and novel deep learning technologies makes appearance-based methods possible to performance well on gaze estimation tasks. With the outstanding ability of feature extraction, CNN models have been generally used in computer vision tasks including gaze estimation. For example, Krafka et al. [5] proposed the iTracker model which uses a CNN model as the basic structure to extract features of the left eye, right eye, and face images respectively. These features are combined with the face grid input by using fully connected layers to estimate the gaze point on mobile devices. But most CNN models can not have great use of global and local features which result in limitations to achieve better performance on gaze estimation.

Our contributions are as follows:

(1) We combine MobileNetV2 and the self-attention layers as the global attention module.

(2) We propose the GazeAttentionNet structure by introducing both global and local attentions.

(3) We can achieve an excellent result on the GazeCapture dataset with an average error of 1.67 cm and 2.37 cm on mobile phones and tablets respectively.

2. RELATED WORKS

2.1. Model-based Approaches

Model-based methods explore the characteristics of human eyes to identify a set of distinctive features around the eyes. The pupil and corneal reflections are common features used for eye localization. Then model-based methods utilize these visual features to fit a geometric 3D eye model to perform gaze estimation. Model-based methods can be subdivided into corneal reflection and shape-based methods, depending on whether they rely on external light sources to detect eye features. Liu et al. [6] proposed a 3D gaze estimation method which works based on iris features using a single camera and a single light source. Model-based gaze estimation approaches tend to suffer from low image quality and variable lighting conditions.

Corresponding author is Zhuo YANG with email yangzhuo@gdut.edu.cn, dr.yangzhuo@qq.com.

2.2. Appearance-based Approaches

Appearance-based approaches directly use the images of eyes or faces as inputs of the model which maps the image data to the gaze vector or gaze point. These approaches need a large amount of data to train the models to predict the gaze point accurately. Krafka et al. [5] introduced a large-scale dataset of mobile gaze estimation named GazeCapture which contains data from over 1450 people and consists of almost 2.5 million frames. MPIIGaze [7] is a dataset for unconstrained 3D gaze direction estimation containing a large number of images from different participants. These images are collected from the daily lives of the participants within several months, which means the background of these images is different and the light condition also varies. With these large-scale appearance-based datasets, researchers try to improve the accuracy of gaze estimation and make it suitable to apply to mobile devices. In [8] Jha et al. proposed an approach that converted the regression problem into a classification problem, predicting the probability at the output instead of a single direction to improve the accuracy of the gaze estimation task. In [9] Chang et al. proposed a highly efficient high-frame-rate eye-tracking method suitable for the performance-constrained mobile environment.

Appearance-based approaches have the potential to work on low-quality images captured by the webcams or front-facing cameras on phones or tablets. Given the success of previous appearance-based gaze estimation researches and the availability of huge labeled datasets, in this work, we focus on this kind of method. It's well known that the Transformer models, which mainly consist of self-attention layers have excellent performance on processing the NLP tasks with the ability to utilize the global information of sentences. In this paper, we combine the MobileNetV2 and the self-attention layers to acquire the global information of features, and we add the spatial attention to utilize the local features of inputs to estimate the gaze point more accurately.

3. GAZEATTENTIONNET

3.1. Global Attention Module

The structure of the global attention module is shown in Fig. 1. The global attention module is made up of MobileNetV2 and the self-attention layers. MobileNetV2 is used to extract a compact feature representation of the input images. And the self-attention layers are used to extract the global information of inputs by computing a weighted sum of features maps. The self-attention layers are organized based on the structure of Transformer's encoder [10].

Backbone. MobileNetV2 is used to process the input images into feature maps. It contains the inverted residual module with a linear bottleneck. This module takes a low-dimensional compressed representation as an input, which is first expanded to high dimension and filtered with

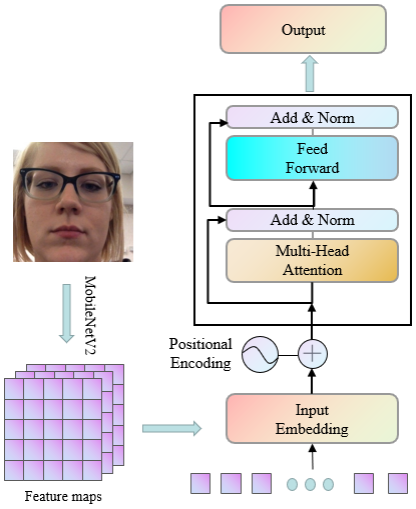

Fig. 1. The structure of the global attention module consists of MobileNetV2 and the self-attention layers. The MobileNetV2 extracts the image's features forming a 1280×7×7 feature map. The feature map are embedded with a positional encoder as the input of the self-attention layers.

a lightweight depthwise convolution. It can alleviate the problem of feature degradation, reduce the amount of computation required for convolution and make use of the input's information better at the same time.

Self-attention Layers. The self-attention mechanism is an attention mechanism relating different positions of a single sequence to compute a representation of the whole sequence. Given a feature matrix X, the feature is projected into queries Q, keys K, and the values V. We calculate the scaled dot product of vectors Q, K, V and apply the softmax function to obtain the weights on values. The operation of the self-attention mechanism can be summarized as [10]:

$$Attention(Q, K, V) = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_m}})V \qquad (1)$$

where $\frac{1}{\sqrt{d_m}}$ is the scaling factor, controling the fluctuation of dot product.

In this paper, the self-attention layers are organized based on the structure of Transformer's encoder composed of a stack of six identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a position-wise fully connected feed-forward network. And there is a residual connection around each of the two sub-layers, followed by layer normalization. Also, the fixed positional encodings are added to the input of each attention layer.

3.2. Local Attention Module

Attention mechanism has been widely used in computer vision tasks, such as image classification [11] and image captioning [12]. The key idea is to reweight features using a score map to emphasize important features and suppress less useful ones [13]. To increase the contrast of the different regions contributing to gaze estimation, we introduce a spatial attention to GazeAttentionNet.

Spatial Attention. In this paper, the spatial attention includes three convolution layers with filter size 1×1 followed by a linear unit layer. The input of the spatial attention is an activation tensor U of size 256×28×28 after the process of four convolution layers. The output is a 28×28 spatial weight matrix W. Then, the weight matrix W is resized to 256 through a fully connected layer and a Sigmod function.

We use four convolution layers and the spatial attention as the local attention module to compute the spatial weight matrix. The structure of the local attention model is shown in Fig. 2. The global attention's outputs V fuse with the local attention module's outputs W by the element-wise multiplication. The process can be summarized as:

$$F = V \odot W \tag{2}$$

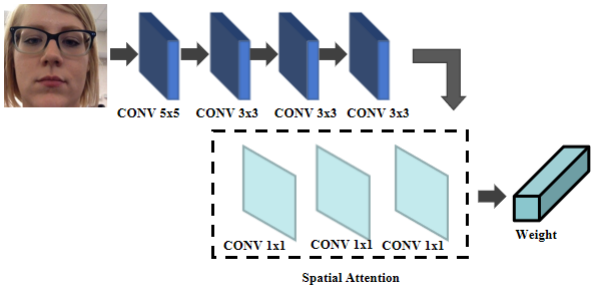

Fig. 2. The followings of the basic convolution layers in the first row are the ReLU function and Maxpool layer. In the spatial attention, the ReLU function is added after the first two convolution layers separately. A Sigmoid function is added after a fully connected layer which is used to compress the size of the weight matrix.

3.3. Feature Fusion

As shown in Fig. 3, the proposed GazeAttentionNet structure consists of three components, including the global attention module, local attention module, and fully connected layers. A two-branch structure is designed to extract the global and local information of the eyes and face images respectively. These feature maps will be fused after the process of these two attention modules.

The global attention module is used to acquire the main information of the images processed into a feature map with a size of 1280×7×7 by MobileNetV2. Then, the feature map will be fed into the self-attention layers with fixed positional encoders. The outputs of the global attention module are fused with the local attention module's outputs by the element-wise multiplication forming feature maps of the left eye, right eye, and face.

Lastly, we use the fully connected layers to combine the feature maps of the eyes and face with the face gird input to estimate the gaze point.

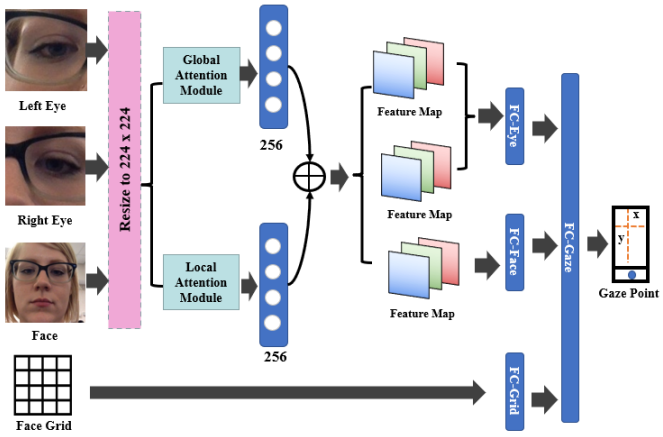

Fig. 3. The GazeAttentionNet contains the global attention module, local attention module, and fully connected layers. The global and local attention module is the first branch and second branch separately in Fig. 3. The details of these two attention modules are shown in Fig. 1 and Fig. 2.

4. EXPERIMENTS

4.1. Data Preparation

In this paper, we adopt GazeCapture [5] dataset to test the performance of our GazeAttentionNet structure. The GazeCapture dataset is a large-scale dataset containing almost 2.5 million frames. All of the frames are collected by the front-facing cameras of mobile devices. Besides, during the process of collecting data, participants are asked to rotate the devices to change the position of the camera including putting it on the top, bottom, left, and right. These operations can make the dataset's data more diversified, which is beneficial to test model's robustness. For training, each of the samples is treated independently while for testing, we average the predictions of the samples to obtain the prediction on the original test sample.

4.2. Training Details

We implement our experiments and design by using PyTorch (version=1.9.0) and Python (version=3.8). To compare the performance of different models, the hyperparameters are set

as the same in overall experiments. The optimizer we use is the SGD optimizer with a learning rate of 0.0001, a momentum of 0.9, and a weight decay of 0.0005 throughout the training procedure. The batch size of the experiments is set to 32.

4.3. Preliminary Experiments

Before testing our GazeAttentionNet on the GazeCapture dataset, we test CNN models such as AlexNet, ResNet18, VGG11, GoogleNet, and MobileNetV2. The structures of these CNN models are different. For example, the AlexNet consists of convolution layers, pooling layers, and fully connected layers using a linear combination while the ResNet18 has a residual structure to reuse the features of former layers. All of them are the famous CNN models used in image classification with excellent performance.

These CNN models are used to replace the convolution layers of the iTracker model as the basic structure. The eyes, face images, and face grid are used as inputs to test these models' performance on the GazeCapture dataset. The results of these experiments are shown in the top half of Table 1. We can find out that MobileNetV2 achieves an average error with the value of 1.75 cm and 2.67 cm on mobile phones and tablets, which is lower than other CNN models. So we adopt MobileNetV2 as the backbone of our global attention module.

Table 1. Results of different models and the GazeAttentionNet on the GazeCapture dataset.

Model	Mobile Phone Error(cm)	Tablet Error(cm)
iTracker [5]	2.04	3.32
SAGE [14]	1.78	2.72
TAT [15]	1.77	2.66
AlexNet	1.85	2.83
ResNet18	1.82	2.72
VGG11	1.82	2.72
GoogleNet	1.79	2.73
MobileNetV2	1.75	2.67
GazeAttentionNet	**1.67**	**2.37**

4.4. GazeAttetionNet Experiments

4.4.1. Experiments Set Up

To evaluate the proposed GazeAttentionNet's performance, we perform the following two experiments in the same experiment environment. We have tested different CNN models' performance on the GazeCapture dataset in the preliminary experiments and the results are shown in Table1. Inspired by these, we combine MobileNetV2 and the self-attention layers as the global attention module. Besides, we test its performance on the GazeCapture dataset by replacing the convolution layers of the iTracker model as same as the preliminary experiments and using the same inputs.

Furthermore, we add the local attention module to acquire the spatial weight of the input images and fuse it with the outputs of the global attention module forming the GazeAttentionNet. We test GazeAttentionNet's performance with the same experimental process.

4.4.2. Experiments Results

The average errors of the experiment which uses the global attention module as the basic structure to extract features are 1.68 cm and 2.51 cm on mobile phones and tablets separately. Compared to MobileNetV2's results shown in Table 1, it proves that the self-attention layers have a positive impact on the model. And GazeAttentionNet experiment's results are shown in the last row in Table 1 with better performance compared to other models including the recent works [5, 14, 15] on GazeCapture dataset with the lowest errors.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel and effective structure for gaze estimation named GazeAttentionNet, which utilizes both global and local features to estimate the gaze point on mobile devices. The global attention module which contains the self-attention layers can acquire the global features of inputs. The local attention module can get a spatial weight of the inputs focusing on the important parts of inputs. With the GazeAttentionNet, we achieve an outstanding result on the GazeCapture dataset with average errors of 1.67 cm and 2.37 cm on mobile phones and tablets respectively. To promote appearance-based gaze estimation on mobile devices, we will improve the model efficiency in the future.

6. ACKNOWLEDGEMENTS

This research is supported by National Natural Science Foundation of China (Nos.61907009, U1813205, U1913202) and Science and Technology Planning Project of Guangdong Province (No.2019B010150002).

7. REFERENCES

[1] Zhimin Wang, Haofei Wang, Huangyue Yu, and Feng Lu, "Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 5, pp. 524–534, 2021.

[2] Simone Dari, Nikolay Kadrileev, and Eyke Hüllermeier, "A neural network-based driver gaze classification sys-

tem with vehicle signals," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[3] Bikun Yang, Jinshi Cui, Yuqiang Tong, Li Wang, and Hongbin Zha, "Recognition of infants' gaze behaviors and emotions," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3204–3209.

[4] Dan Witzner Hansen and Qiang Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

[5] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba, "Eye tracking for everyone," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2176–2184.

[6] Jiahui Liu, Jiannan Chi, Wenxue Hu, and Zhiliang Wang, "3D model-based gaze tracking via iris features with a single camera and a single light source," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 2, pp. 75–86, 2021.

[7] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 162–175, 2019.

[8] Sumit Jha and Carlos Busso, "Estimation of gaze region using two dimensional probabilistic maps constructed using convolutional neural networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3792–3796.

[9] Yuhu Chang, Changyang He, Yingying Zhao, Tun Lu, and Ning Gu, "A high-frame-rate eye-tracking framework for mobile devices," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1445–1449.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[11] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: Convolutional block attention module," in *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, Eds., Cham, 2018, pp. 3–19, Springer International Publishing.

[12] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6298–6306.

[13] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] Junfeng He, Khoi Pham, Nachiappan Valliappan, Pingmei Xu, Chase Roberts, Dmitry Lagun, and Vidhya Navalpakkam, "On-device few-shot personalization for real-time gaze estimation," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1149–1158.

[15] Tianchu Guo, Yongchao Liu, Hui Zhang, Xiabing Liu, Youngjun Kwak, Byung In Yoo, Jae-Joon Han, and Changkyu Choi, "A generalized and robust method towards practical gaze estimation on smart phone," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1131–1139.