Appearance-based Gaze Estimation Using Multi-task Neural Network

Mihan Liu¹, Hui Li²*, Huanhuan Dai

¹College of Computer Science, Sichuan University, Chengdu, Sichuan, 610065, China
²College of Computer Science, Sichuan University, Chengdu, Sichuan, 610065, China
³College of Computer Science, Sichuan University, Chengdu, Sichuan, 610065, China

*Corresponding author’s e-mail: lihuib@stu.scu.edu.cn

Abstract. In order to improve the gaze estimation accuracy through using the information in the eye image more adequately, an appearance-based gaze estimation neural network EyeNet has been proposed in this article. On the one hand, by decomposing the main task into two subtasks, this network could gain more information during the training stage. On the other hand, EyeNet could extract features in the image more efficiently by adding an auxiliary task which is detecting the iris center and eye ball center of the eye. Experiments show that by using these two tricks, the error decreased from 5.42 degree to 5.16 degree.

1. Introduction

Human’s eyes are the main organs of getting information from the outside world. By estimating the gaze direction of eyes, we can develop many applications which can be applied in many scenarios, such as computer-human interaction, VR, psychological research and driver monitor system, so gaze estimation has great research significance.

Gaze estimation methods can be divided into two categories as intrusive and non-intrusive based on the device it used to get the eye images. Non-intrusive gaze estimation methods don’t need subjects to wear device like glasses or helmet, thus non-intrusive gaze estimation is friendlier to subjects and has much more potential to be widely used in daily life.

In non-intrusive methods, gaze estimation can be categorized as feature-based, model-based and appearance-based[1]. With the development of deep learning[2] and massive gaze estimation datasets emerging, appearance-based gaze estimation made a huge progress. Researcher have made a lot study on appearance-based gaze estimation using neural networks. For example, Zhang[3] used LeNet-5 and VGG-16 as their base network and added the head pose as auxiliary information into the fully-connected layer to estimate the gaze direction. Zhang[4] argued that by giving different area with different weights, the information can be enhanced or depressed, so they put the whole face into the network and give different area with different weights and get better performance.

All current appearance-based methods regard gaze estimation as a simple regression task. They just send the image into the network and then get two numbers which stands for the pitch and yaw of the gaze direction. But getting the mapping from a raw eye picture to gaze direction is an ill problem due to the head pose variation, illumination changing and image resolution variation. We argue that simply regressing two numbers can’t fully utilize the information in eye images, thus may not gain the best performance in inference stage.
In order to extract features in eye images more efficiently and get better accuracy, we proposed an appearance-based gaze estimation network EyeNet. It has two main characteristics:

1. EyeNet divide the main task gaze estimation into two sub-tasks. Each sub-task is designed to predict one angle, yaw or pitch, and in every task we convert the simple regression task into a combination of regression and classification task. In this way, the network can learn more knowledge during the training stage and by adding the classification task the network can learn the relationship among the similar gaze direction pictures to get better performance.

2. By adding an auxiliary task, eye landmark detection, EyeNet can extract features more efficiently. Research in multi-task learning[5] shows that by adding an auxiliary task which is related to the main task, better performance of the main task can be achieved. We add an auxiliary task, eye landmarks localization, to the network and make the network focus on the vital features during the training stage to achieve better performance.

2. Network design
Here we will introduce the tricks we used in the EyeNet, the full picture of EyeNet and the loss function design of EyeNet.

2.1. Main task dividing
Gradient vanishing is a deadly problem for neural networks at training stage. In order to solve this problem many tricks have been proposed. ResNet[6] is one of the most successful way for solving the gradient vanishing problem. ResNet uses short-cuts in their network. The multi-branch way can make more information back propagate to the network to avoid the gradient vanishing problem, and as a result improve the network’s performance. In the similar way, we use two separate branches at the end of the network, each one for an angle, to predict gaze angles. The doubled output could bring doubled signal to the networks which could make the network learn more knowledge and get better performance.

Gaze estimation has been regarded as a regression problem for a long time. These days researchers like Ruiz [7] found that the gap between classification and regression might not be so big. They proposed a network to estimate human head pose in a combination way and get an awesome result. Gaze estimation and head pose estimation have some common features. For example, they both are finding a mapping from image to numbers. So, we apply this approach in our EyeNet. For every output branch of EyeNet, the loss of each predicted angle are the combination of regression loss and classification loss. The classification loss could let network to learn the relationships between similar angle pictures and the regression loss could make sure that the output angle of network is accurate.

2.2. Auxiliary task
In Sec. 2.1, we said that gaze estimation is similar to head pose estimation, but they also have many differences. First, eye images are usually smaller than face images. More importantly, gaze direction could be inferred through the features like iris center and eyeball center. For example, Park [8] proposed GazeMap which is a pictorial representation of 3D gaze. Their work proved that by adding some auxiliary information like iris margin and eyeball margin, neural network could concentrate on the important features during training stage and the performance of neural network could be improved in this way.

Rajeev [9] proposed a multi-task neural network, HyperFace. They use one neural network to predict face landmarks, gender, head pose from one image at the same time. Their work showed that every single sub-task’s performance has been improved due to that every sub-task cooperates with each other and let the network learning more information in the image than single-task neural network.

Inspired by their excellent works, the second trick that EyeNet used is adding an auxiliary task, eye landmarks detection. Specifically speaking, the auxiliary task is detecting iris center and eyeball center, because of that the ray starts from the eyeball center and ends with iris center can roughly indicates the gaze direction in the 3D world.
There are two types of landmark detection methods, regressing coordinates directly or predicting a heatmap which stands for the landmarks. In our work, we choose the latter. Heatmaps in landmark detection are grayscale images, and the value of pixels which are nearer to the landmarks is greater than farther ones. Figure 1 shows the heatmaps we used in EyeNet.

![Eye landmarks and corresponding heatmaps.](image)

2.3. Neural network architecture

Figure 2 shows the structure of EyeNet. The backbone of EyeNet is hourglass module that proposed by Newell [10]. They used it in human pose estimation because it can learn the relation among human’s body parts. Here we use 3 hourglass modules because the size of eye crop is relatively smaller than human body image.

We can see from figure 2 that there are two outputs path of EyeNet. The top path is the auxiliary task we mentioned at sec2.2. It outputs two heatmaps, one for iris center and one for eyeball center.

The bottom path is the main task, gaze estimation. It splits into two sub-paths that each path stands for one angle, pitch or yaw. We take the upper path, path of yaw, as an example. The output of hourglass modules will go through a full-connected layers and outputs a probability distribution, and then we calculate the expectation of the probability distribution to get prediction of yaw.

2.4. Loss Functions

As shown in figure 2, there are three outputs in our network and each one has its own loss. The total loss is the combination of all the losses.

![Network architecture](image)

The total loss $l_{total}$ contains two big parts, gaze estimation loss $l_{gaze}$ and landmark detection loss $l_{landmarks}$, as in equation (1). Equation (2) gives the computation of $l_{landmarks}$. In equation (2), $h$ and $\hat{h}$ stand for heatmap outputted from EyeNet and ground truth heatmap respectively, and subscript $i$ and...
$j$ stands for horizontal and vertical coordinates respectively. Gaze estimation loss is composed of two parts, $l_{\text{pitch}}$ and $l_{\text{yaw}}$. The computation of each angle is demonstrated as equation (4). In equation (4), $y$ and $\hat{y}$ stand for predicted angle and ground truth angle respectively, C and R stand for classification loss function and regression loss function respectively. We used cross-entropy loss for classification and mean square loss for regression in this paper.

\[
l_{\text{total}} = l_{\text{gaze}} + l_{\text{landmarks}} \tag{1}
\]
\[
l_{\text{landmarks}} = \left( \sum_{i,j} (h_{ij} - \hat{h}_{ij})^2 \right)^{1/2} \tag{2}
\]
\[
l_{\text{gaze}} = l_{\text{pitch}} + l_{\text{yaw}} \tag{3}
\]
\[
l_{\text{angle}} = R(y, \hat{y}) + C(y, \hat{y}) \tag{4}
\]

3. Experiments and results

3.1. Training

We use MPIIGaze[3] as our training and testing datasets. The MPIIGaze dataset has 213659 images which are captured among 15 subjects during their daily life. The variation of illumination, resolution and head pose is huge, thus estimating gaze direction on this dataset is a big challenge.

We trained 15 models, one for one subject. We use cross-subject training procedure which is that using one person for testing and using the rest persons for training. Using cross-subject training could show the generalization ability of the proposed network.

All of these models are trained using Adam[11] as optimizer. The hyper-parameters are as follows: learning rate 0.0002, batch size 32, weight decay 0.0001 and epochs 10. The image send into the network are all of size 156 * 96.

3.2. Metrics

In this paper, error is determined as the Euclidean distance between the predicted angles $(p, y)$ and ground truth angles $(\hat{p}, \hat{y})$ as given by equation (5). The error of single model is the average of all samples tested on it and the final error is the average error of all 15 models.

\[
\text{error} = \left((p - \hat{p})^2 + (y - \hat{y})^2\right)^{1/2} \tag{5}
\]

3.3. Results

In order to prove that the tricks we applied are effective, we compared the performance among three conditions: hourglass-only, hourglass + main task dividing, hourglass + main task dividing + auxiliary task. From the result showed in table 1, we can see that the two tricks that we proposed can both improve the performance of the network.

| Model                              | Error (degree) |
|------------------------------------|----------------|
| Hourglass only                     | 5.42           |
| Hourglass + main task dividing     | 5.27           |
| Hourglass + main task dividing + Auxiliary task | 5.16 |

We also compared the performance with some methods that others proposed at table 2.

| Method   | Error (degree) |
|----------|----------------|
| Wood[12] | 9.58           |
| Nie[13]  | 7.1            |
| Zhang[3] | 5.5            |
| Ours     | 5.16           |
4. Conclusion
In this paper, we improved the performance of gaze estimation using neural networks in a multi-task way. Our experiments show that both of the tricks we applied are effective. It proved that by adding more signals at the end of neural networks, the network will gain better predicting accuracy at testing stage. And our experiments also show that the positions of iris center and eyeball center are important information in gaze estimation. With good using of these features, great improvement can make in gaze estimation.

References
[1] Park, S., Zhang, X., Bulling, A., Hilliges, O. (2018). Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications. Warsaw.
[2] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521: 436-444.
[3] Zhang, X., Sugano, Y., Fritz, M., Bulling, A. (2017). MpiiGaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41: 162-175.
[4] Zhang, X., Sugano, Y., Fritz, M., Bulling, A. (2017). It’s written all over your face: Full-face appearance-based gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Hawaii. pp. 51-60.
[5] Zhang Y, Yang Q. (2017) An overview of multi-task learning. National Science Review, 5: 30-43.
[6] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas. pp. 770-778.
[7] Ruiz, N., Chong, E., Rehg, J. M. (2018). Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 2074-2083.
[8] Park, S., Spurr, A., Hilliges, O. (2018). Deep pictorial gaze estimation. In Proceedings of the European Conference on Computer Vision (ECCV). Munich. pp. 721-738.
[9] Ranjan, R., Patel, V. M., Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41, 121-135.
[10] Newell, A., Yang, K., Deng, J. (2016.). Stacked hourglass networks for human pose estimation. In: European conference on computer vision. Amsterdam. pp. 483-499.
[11] Kingma, D. P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
[12] Wood, E., Baltrušaitis, T., Morency, L. P., Robinson, P., & Bulling, A. (2016). Learning an appearance-based gaze estimator from one million synthesised images. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications pp. 131-138.
[13] Nie, S., Zheng, M., Ji, Q. (2018). The deep regression bayesian network and its applications: Probabilistic deep learning for computer vision. IEEE Signal Processing Magazine, 35, 101-111.