Gaze Estimation in Driver Monitor System

Abstract

Gaze estimation has been one of the hottest topics in the field of computer vision in recent years. It is challenging due to high cost of acquiring data and overfitting caused by small training data. To alleviate such problems, we present a flexible gaze acquisition system which can be applied to Driver Monitor System (DMS). Unlike the complex acquisition systems in the laboratory, our acquisition system only relies on a binocular IR camera and a laser rangefinder with a checkerboard calibration board which is not limited by the venue and suitable for rapid cockpit modeling. In addition, we proposed a multi-task gaze estimation method with Semi-Supervised training strategy based on image style conversion. Experiments show that the proposed method can effectively improve the performance of the model training on small training data.

1 Introduction

Gaze reflects the driver’s concentration direction when driving (9). The application of gaze estimation is mainly to detect the distraction state by tracking the driver’s gaze direction or the position of the gaze point through the DMS camera in the cabin. Gaze estimation can also be applied to automatic screen brightening, HUD, and multi-mode interaction combined with intelligent voice.

Gaze estimation remains a challenging task due to the ground truth of gaze angle cannot be completed by manual annotation (4, 5, 7, 11, 12, 3, 1). firstly, we could not adopt the existing multi-eye gaze acquisition scheme (8) in the laboratory due to the flexibility of scene construction and the limitation of binocular camera only. Secondly, the head movement is difficult to predict, and the generalization ability of the algorithm is needed to offset the error caused by the change of head pose; finally, gaze estimation algorithm needs to be robust to the influence of the eye occlusion, illumination and other environments.

Therefore, we built a high-precision acquisition system for the gaze angle and gaze point in the cabin. At the same time, the cabin is divided into left front windshield, right front windshield, left rearview mirror, right rearview mirror, central control screen, and other six areas. The deep learning model is used to predict the driver’s gaze Angle and gaze point areas, and the driver’s distraction state is cooperatively judged from two different dimensions.

2 Gaze Acquisition System

The principle of gaze acquisition system mainly converts the starting point and the gaze point of the gaze to the camera coordinate system through the transformation of the multi-eye camera and the spatial coordinate system while collecting image data, and constructs a 3D vector from the starting point to the gaze point. The pitc and YAW angles can be calculated in the camera coordinate system. The starting point of the gaze is the key point of the eye pupil, which can be calculated by building a binocular camera for distance measurement. However, in the actual process, the binocular ranging
method based on triangulation should try to keep the two cameras of the same model, and the two imaging planes are coplanar. Otherwise, it must be corrected by stereo rectification, a process that produces unmeasured errors. By the binocular calibration of the camera can obtain the rotation $R$ and translation $T$ matrices from the main camera to the other camera.

$$Z_{c1} \mu_1 = \begin{bmatrix} f_{x1} & 0 & c_{x1} & 0 \\ 0 & f_{y1} & c_{y1} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_{c1} \\ Y_{c1} \\ Z_{c1} \\ 1 \end{bmatrix} (1)$$

$$Z_{c2} \mu_2 = \begin{bmatrix} f_{x2} & 0 & c_{x2} & 0 \\ 0 & f_{y2} & c_{y2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_{c2} \\ Y_{c2} \\ Z_{c2} \\ 1 \end{bmatrix} (2)$$

$$\begin{bmatrix} X_{c1} \\ Y_{c1} \\ Z_{c1} \\ 1 \end{bmatrix} = \begin{bmatrix} R \quad T \end{bmatrix} \begin{bmatrix} X_{c2} \\ Y_{c2} \\ Z_{c2} \\ 1 \end{bmatrix} (3)$$

We first construct the camera coordinate system to the pixel coordinate system conversion relationship. $Z_{c1}$ is the distance from the camera to the human eye, and $u_1, v_1$ is the pupil key points in 2D image, and $X_{c1}, Y_{c1}, Z_{c1}$ represents the coordinates of the pupil point in the camera coordinate system, and $f_{x1}, f_{y1}, c_{x1}, c_{y1}$ is the camera intrinsics parameters. The same goes for Equation 2 and Equation 1. By combining equations, it is possible to construct a superdefinite system of equations to solve $X_{c1}, Y_{c1}, X_{c1}$.

The camera in the cabin is generally facing the driver, so the gaze point is outside the FOV area of the camera. To construct the gaze point coordinates in the camera coordinate system, an intermediate medium is needed to establish a conversion relationship. The checkerboard can be used as an auxiliary tool to establish the transformation relationship between camera coordinate system and checkerboard coordinate system. At the same time, the laser range finder is added to the checkerboard, and the laser light is perpendicular to the checkerboard plane. The transformation relationship between the laser point coordinate system and the checkerboard coordinate system can be obtained by simple translation vector calibration. Thus, the laser landing point is regarded as the sight landing point, which can be converted to the camera coordinate system through the checkerboard.

$$\begin{bmatrix} X_{c1} \\ Y_{c1} \\ Z_{c1} \\ 1 \end{bmatrix} = \begin{bmatrix} R_{cb} \quad T_{cb} \end{bmatrix} \begin{bmatrix} X_l \\ Y_l \\ d \end{bmatrix} (4)$$

where $X_{c1}, Y_{c1}, X_{c1}$ represents the coordinates of the laser point in the camera coordinate system, $R_{cb}, T_{cb}$ represents the transformation relationship from camera to checkerboard obtained by solving the PNP problem by detecting the checkerboard corners of the image and the known checkerboard 3D coordinates, $X_l, Y_l, d$ is the coordinates of the laser point in the checkerboard coordinate system, and $d$ represents the ranging result. From this, we can get a vector representation of the starting point to the gaze point, and calculate the angle with the coordinate axis.
Collecting large-scale labeled training datasets of gaze estimation is difficult and expensive. With only limited labeled datasets and a large amount of unlabeled data, we propose a semi-supervised multi-task gaze estimation method. A detailed of our method is given in the following.

As shown in Figure 2, the proposed Multi-task Gaze Estimation (MGE) model consists of three parts: backbone, angle branch and zone branch. The former one conducts gaze representation from the eyes image, while the latter two respectively attempt to produce gaze angle (pitch, yaw of each eye) and gaze zone (A-F partitions) according to the gaze representation.

Specifically, we introduce the ShuffleBlock with a GCBlock as the main component of MGE model. ShuffleBlock is an efficient network unit which can meet the needs of mobile real-time inference. GCBlock was proposed to reduce the parameters of the traditional self-attention module. As mentioned in [], the attention maps for different query keys are almost the same, therefore, GCBlock adopts query sharing mechanism to simplify the traditional self-attention module. Otherwise, a bottleneck was introduced to further reduce the parameters. The GCBlock can be formulated as follows:

$$z_i = x_i + W_\theta^2 \text{ReLU}(LN(W_\theta^1 \sum_{j=1}^{N_p} e^{W_k x_j} \sum_{m=1}^{N_p} e^{W_k x_m} x_j)),$$

(5)

DMS gaze estimation mainly includes two tasks, one is gaze angle estimation, which is used to predict the direction of gaze, and the other is gaze zone estimation, which is used to predict the area where people look. Both tasks are critical for DMS distraction detection. Our multi-task model uses the angle branch to regress the human eye gaze angle and the zone branch to classify the human eye landing area after the backbone, respectively.

It is difficult to obtain gaze training data. Different cars have different geometries and therefore different distribution of gaze zones, so it is necessary to collect special data for specific cars. For the gaze angle, its ground truth is very difficult to obtain. In the previous section, we introduced the solution for obtaining the ground truth of the gaze angle. Unlabeled data is much easier to obtain, including natural face images, public datasets (most only contain gaze angle labels), but these images’ styles are indeed very diverse, including near-infrared, visible light, etc., which isn’t completely matched the image style of a specific car DMS camera. To overcome the above problems, we propose a semi-supervised training method based on image style conversion as shown in Figure 3.

For the problem of image style mismatch, we use GAN for image style conversion. VSANet is an unpaired image domain conversion network. We perform VSNet training between the unlabeled dataset and the collected labeled dataset to make the image style of the unlabeled data is consistent with the image style of the collected labeled dataset. Then we use the collected labeled dataset to train a pseudo-labeling model to label the unlabeled data. The large-scale pseudo-label data will be
used to train a pre-trained model, and then we use the collected data to finetune the pre-trained model to obtain the final multi-task gaze estimation model. By this training strategy, we have successfully solved the gap problem between heterogeneous data and made full use of unlabeled data to improve model performance and generalization.

4 Experiments

4.1 Dataset and Metrics

We consider the self-collected dataset to evaluate the performance of our model. The dataset contains 29,658 eye images of 150 different identities, all images are resized to size of 80 × 160. We take 1,000 images as the test set and the rest as the train set. The labels includes the pitch and yaw angles of the left and right eyes, and the partition (A-F) where the gaze point is located. In addition, we collected a large-scale unlabeled dataset which contains 2 million images, including visible light images and near-infrared images taken by different IR cameras. We use MAE to evaluate the performance of gaze angles, which can be formulated as:

$$MAE = \frac{g \cdot \hat{g}}{||g|| \cdot ||\hat{g}||},$$

where $g \in \mathbb{R}^3$ is the ground truth of gaze direction, and $\hat{g} \in \mathbb{R}^3$ is the predicted gaze direction. We use precision, and recall to evaluate the performance of gaze zone.

4.2 Implementation Details

During the training process, we use VSANet as the image style transfer model. We use smoothL1 as the angle regression loss and crossentropy as the zone classification loss. The total loss is shown in the following:

$$L_{total} = L_{angle} + \lambda L_{zone},$$

where $\lambda$ is 5. Adam optimizer is employed with a learning rate of 2e-3, the $\beta_1$ of 0.5 and the $\beta_2$ of 0.99. The batch-size is set to 128.

4.3 Quantitative Results

Table 1: Evaluation results on our dataset. ↑/↓ denotes higher/lower is better. **Bold** means the best.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE↓</th>
<th>Precision↑</th>
<th>Recall↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model1</td>
<td>7.75°</td>
<td>86.9%</td>
<td>79.9%</td>
</tr>
<tr>
<td>Model2</td>
<td>4.93°</td>
<td>92.6%</td>
<td>86.3%</td>
</tr>
<tr>
<td>Model3</td>
<td>4.84°</td>
<td><strong>94.0%</strong></td>
<td><strong>89.7%</strong></td>
</tr>
</tbody>
</table>

Table 1 report the quantitative results on our self-collecting dataset. Model1 is a small dataset training model, Model2 is a semi-supervised training model using unlabeled data directly, Model3
is a semi-supervised training model using GAN transformed unlabeled data. It can be seen that the performance of Model1 which train on small dataset is relatively poor, which may be mainly due to model overfitting. Using large-scale pseudo-label data for pre-training has greatly improved the performance. In addition, keeping the style of the pre-training dataset and the finetune dataset consistent is helpful for model training.

5 Conclusion

We propose a adaptable gaze acquisition system. It attempts to address the challenges in collecting labeled gaze data. We use the binocular camera and the laser rangefinder with checkerboard to obtain the start (pupil) and end (gaze point) of the gaze vector, we further calculate the pitch and yaw of the gaze from the vector. In view of the current situation of coexistence of small labeled datasets and large heterogeneous unlabeled datasets, we propose a multi-task gaze estimation method which utilizes image style conversion and Semi-Supervised training. It effectively improves the performance of the model training on small data.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
(a) Did you state the full set of assumptions of all theoretical results? [Yes]
(b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [No]
(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
(a) If your work uses existing assets, did you cite the creators? [N/A]
(b) Did you mention the license of the assets? [N/A]
(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
(d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]