A Method Utilizing RGB and Optical Flow Sequences: "Looking at me" Challenge

KeioEgo Team

Abstract

Egocentric videos are widely used in the analysis of social interactions. Detecting if someone is looking at the person with an egocentric camera during social interactions is an interesting and meaningful application. According to the baseline method of this challenge, it is possible to determine if a tracked person is "looking at me" in a frame with its first and next few frames. However, only the RGB image sequences are used as input in the baseline framework. In our case, we introduce the optical flow sequences to explicitly utilize the movement information of the tracked human faces. The optical flow images are generated beforehand by a state-of-the-art optical flow estimation method. Regarding the mean average precision (mAP), our proposed framework achieves 73.91% on the test set, outperforming the baseline method.

Method Description

1. Overall Framework

The feature extraction part in our proposed framework consists of an RGB branch and an optical flow branch as shown in Figure 1, the two branches don't share the same weights though their structures (ResNet18 + Bi-LSTM) are similar. To obtain the result of frame *t*, 7 continuous RGB frames and 6 continuous optical flow frames of one's face are input to the network. We then concatenate their extracted features to fuse them as the input of the fully connected (FC) output layers.

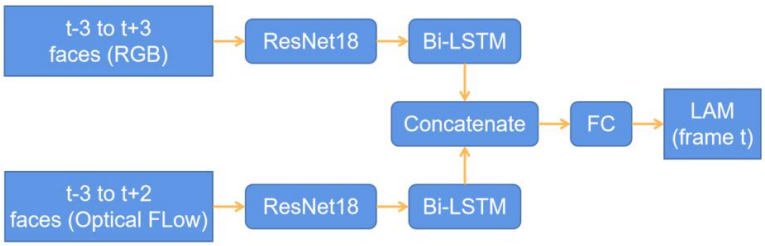

Figure 1. The overall framework of the proposed method.

2. Optical Flow Generation

For the purposes of saving time and computational resources when training and evaluating the network shown in Figure 1, we generated the optical flow images with the given RGB faces beforehand. FlowFormer [1] is a recently proposed optical flow estimation network whose outstanding performance has been proved on multiple datasets, therefore we used a trained FlowFormer to generate the optical flow of the tracked faces provided by Ego4D. For a 7-frame RGB input sequence, its corresponding optical flow sequence is of 6 frames, explicitly containing the movement information during this period.

Experiments and Discussion

We initialized the backbone of our network from Gaze360 [2]; as for the optical flow branch, their weights were loaded from the RGB branch before training. During the training procedure, the batch size was set as 64. We chose Adam as the optimizer, and the learning rate was set as 5x10-4. Besides, we introduced a learning rate scheduler, reducing the learning rate by half every 5 epochs.

We evaluated several trained models of the same network structure on the test set. As shown in Table 1, it can be observed that there is a gap between the mAP values on the validation set and the test set for each obtained model. The model which has the best performance on the validation set (80.95%) achieves 71.20% on the test set, while the highest mAP value on the test set is 73.91%. According to Table 1, it seems that the validation set doesn't perform well enough in helping screen models.

	mAP (%)								
Val	80.94	79.52	79.45	79.44	80.20	80.01	80.51	80.43	**80.95**
Test	73.00	71.73	71.19	**73.91**	71.28	71.15	73.12	71.90	71.20

Table 1. The mAP values corresponding to different trained models on the validation set and the test set.

Besides, we checked some examples of the validation set. In Figure 2a and 2b, the woman is not looking at the camera, but our model fails to recognize it for Figure 2a, we think it might be caused by the existence of glasses which makes her eyes not clear enough. In Figure 2c, our model predicts that "the man is not looking at me", which we think it's correct; however, the ground truth of this frame is "he is looking at me", and we are afraid that it might be a labeling mistake. In Figure 2d, we successfully predict that "he is looking at me".

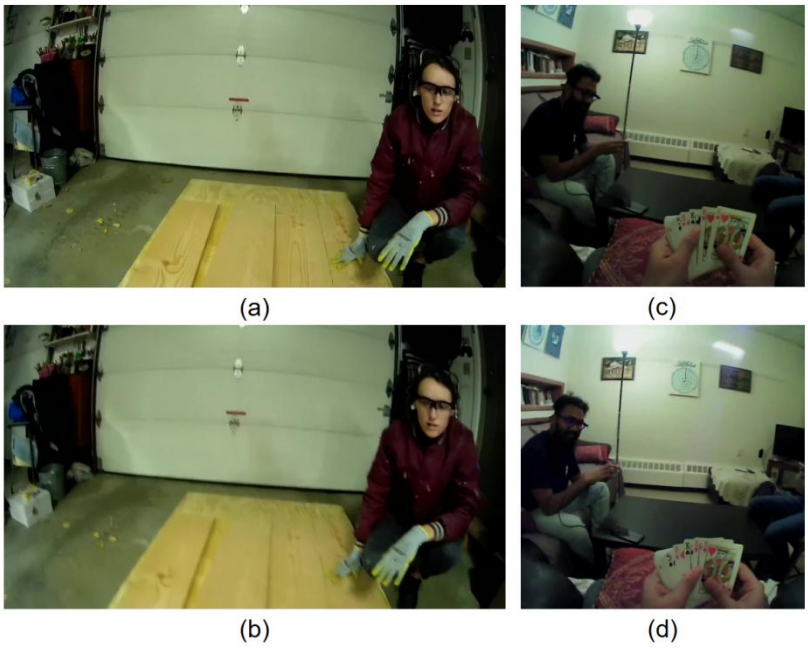

Figure 2. Examples of the validation set.

About KeioEgo Team

KeioEgo Team contains one student and two advisors:
Haowen Hu (hu_haowen@keio.jp), a doctoral student at the Graduate School of Science and Technology, Keio University;
Ryo Hachiuma (ryo-hachiuma@keio.jp), a project assistant professor at the Graduate School of Science and Technology, Keio University;
Hideo Saito (hs@keio.jp), a professor at the Graduate School of Science and Technology, Keio University.

References

[1] Huang Z, Shi X, Zhang C, et al. FlowFormer: A Transformer Architecture for Optical Flow[J]. arXiv preprint arXiv:2203.16194, 2022.
[2] Kellnhofer P, Recasens A, Stent S, et al. Gaze360: Physically unconstrained gaze estimation in the wild[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6912-6921.