Localizing the Gaze Target of a Crowd of People

Yuki Kodama[1], Yasutomo Kawanishi[1], Takatsugu Hirayama[2], Daisuke Deguchi[3], Ichiro Ide[1], Hiroshi Murase[1], Hidehisa Nagano[4], and Kunio Kashino[4]

[1] Graduate School of Informatics, Nagoya University, Aichi, Japan
Email: kodamay2@murase.is.i.nagoya-u.ac.jp
[2] Institutes of Innovation for Future Society, Nagoya University, Aichi, Japan
[3] Information Strategy Office, Nagoya University, Aichi, Japan
[4] NTT Communication Science Laboratories, NTT Corporation

Abstract. What target is focused on by many people? Analysis of the target is a crucial task, especially in a cinema, a stadium, and so on. However, it is very difficult to estimate the gaze of each person in a crowd accurately and simultaneously with existing image-based eye tracking methods, since the image resolution of each person becomes low when we capture the whole crowd with a distant camera. Therefore, we introduce a new approach for localizing the gaze target focused on by a crowd of people. The proposed framework aggregates the individually estimated results of each person's gaze. It enables us to localize the target being focused on by them even though each person's gaze localization from a low-resolution image is inaccurate. We analyze the effects of an aggregation method on the localization accuracy using images capturing a crowd of people in a tennis stadium under the assumption that all of the people are focusing on the same target, and also investigate the effect of the number of people involved in the aggregation on the localization accuracy. As a result, the proposed method showed the ability to improve the localization accuracy as it is applied to a larger crowd of people.

1 Introduction

Gaze estimation from an image is very useful for various applications. Gaze can tell us how a person looks into things to buy, which object s/he is interested in, and so on. Therefore, there are many studies focusing on human gaze estimation [1–5]. In a situation where many people gather in a space such as a cinema or a stadium, it is valuable to analyze where the people, i.e. a crowd of audience or spectators, are looking at. Here, each person's gaze is independent and there are many objects to be focused on by the people. So, we can assume that the more people look at a particular object simultaneously, the more potentially valuable, or interesting, the object is. Therefore, there is a strong demand to analyze a crowd of people looking at the same object simultaneously.

To analyze where a crowd of people is looking at, we need to observe all the members simultaneously. From the viewpoint of cost and convenience, rather than employing eye trackers, it is desired to estimate where they are looking at from an image capturing the whole crowd. One very simple solution is to capture

Fig. 1. Example of a situation that the proposed method assumes.

each person in high-resolution from many cameras. However, this is di cult to implement. Instead, capturing the whole crowd by one camera is a more realistic solution. However, in this case, the face image of each person becomes small in size, so we can only obtain their face images in low-resolution.

Therefore, our goal is to localize a gaze target focused on by a crowd of people from one image captured from a distant camera position (Figure 1). Since detailed information for each face or pair of eyes cannot be extracted from such an image, gaze estimation from the images with low-resolution will be inaccurate. Nevertheless, as shown in Figure 2, when many people in the crowd are focusing

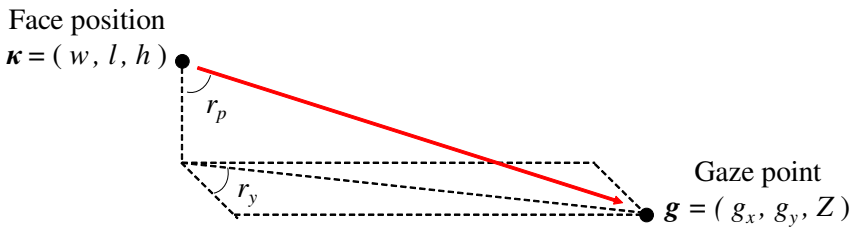


Fig. 4. Face and gaze positions in the world coordinate system.

$p(\boldsymbol{g}|X;\Theta)$, where Θ represents the set of parameters of the gaze estimation. By assuming that the prior probability $p(\boldsymbol{g})$ of the gaze point $\boldsymbol{g}$ follows a uniform distribution, it can be replaced by a constant C. We also assume that $p(\boldsymbol{g}|x_i;\Theta)$, $x_i = 1...M$, are independent from each other. With these assumptions, Eq.(1) is derived from Bayes' theorem.

$$p(\boldsymbol{g}|X;\Theta) = \frac{1}{C}\prod_i p(\boldsymbol{g}|x_i;\Theta). \tag{1}$$

By taking the logarithm of this equation, the log likelihood of the gaze of the crowd can be calculated. Finally, localizing their gaze point $\hat{\boldsymbol{g}}$ can be formulated as

$$\hat{\boldsymbol{g}} = \arg\max_{\boldsymbol{g}} \sum_i \log(p(\boldsymbol{g}|x_i;\Theta)). \tag{2}$$

3.2 Training Phase

Firstly, from training images $\mathcal{I}_{\text{train}} = \{I_1, I_2, ...\}$, which are images capturing a crowd of people, faces are detected and cropped. For the face detection, we used a face detecter based on multi-task cascaded CNNs proposed by Zhang et al. [15]. The cropped face images are resized to $W \times H$ [pixels], and the resized face images $\{x_i\}_{i=1}^{Q}$ are used for training the gaze angle estimator. Q represents the number of the detected faces.

Then, the parameters Θ of the gaze angle estimator $f(x_i;\Theta)$ are learned by using training data $T = \{(x_i, \boldsymbol{r}_i)\}$, where $\boldsymbol{r}_i = (r_y, r_p)^T$ represents the gaze angle from the center of his/her face (face position $\boldsymbol{\kappa}_i$) with the yaw angle r_y and the pitch angle r_p of the gaze angle, as shown in Figure 4. Here, the gaze angle is calculated as the summation of face orientation and gaze direction angles. The function f is modeled as a Convolutional Neural Network (CNN) trained with this training data T. With regard to the loss function, we employ the mean squared error loss function defined as

$$L(\Theta) = \frac{1}{Q}\sum_{i=1}^{Q}(\boldsymbol{r}_i - f(x_i;\Theta))^2. \tag{3}$$

In the training phase, parameters which minimizes $L(\Theta)$,

$$\hat{\Theta} = \arg\min_{\Theta} L(\Theta) \tag{4}$$

are searched as the optimal solution using the Adaptive Momentum (Adam) method [16].

3.3 Localization Phase

The localization phase commences in the same fashion as in the training phase, with M faces detected from an input image capturing a crowd of people. The detected faces $\{d_j\}_{j=1}^{M}$ are resized to $W \times H$ [pixels], and input to the trained gaze angle estimator $f(x_i; \hat{\Theta})$. Then, a gaze angle

$$\boldsymbol{a}_j = (a_{jy}, a_{jp})^T = f(d_j; \hat{\Theta}) \tag{5}$$

is calculated for each face image. Here, a_{jy} represents the yaw angle of his/her gaze, and a_{jp} the pitch angle of his/her gaze. As a result, a gaze point $\boldsymbol{g}_j = (g_{jx}, g_{jy}, Z)^T$ can be calculated based on the gaze angle $\boldsymbol{a}_j$ and the face position $\boldsymbol{\kappa}_j = (w_j, l_j, h_j)$ of each person as the point where the vector from the position of the face center and the plane with height Z crosses.

In general, even if the result of the gaze angle estimation is not correct, the true angle is likely to be nearby the result. Therefore, we assume that $p(\boldsymbol{g}|d_j; \hat{\Theta})$ follows the normal distribution $\mathcal{N}(\boldsymbol{g}_j, \sigma^2)$. By assuming this, Eq.(2) is equivalent to

$$\hat{\boldsymbol{g}} = \arg\max_{\boldsymbol{g}} \left(-\sum_{j=1}^{M} \frac{(\boldsymbol{g} - \boldsymbol{g}_j)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2} \right) = \arg\min_{\boldsymbol{g}} \sum_{j=1}^{M} (\boldsymbol{g} - \boldsymbol{g}_j)^2, \tag{6}$$

where M is the number of the detected faces. Then, it can be solved as

$$\hat{\boldsymbol{g}} = \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{g}_j = \overline{\boldsymbol{g}}. \tag{7}$$

As a result, a common gaze point of the crowd $\overline{\boldsymbol{g}} = (\overline{g}_x, \overline{g}_y, Z)^T$ is output.

4 EVALUATION

4.1 Dataset Construction

Although there are some public datasets including face images with the ground truth of the gaze direction [9, 17–20], most of them were recorded under controlled laboratory conditions, and there is also no dataset including images capturing many people in a frame. Therefore, we constructed a dataset by capturing 96 participants including men/women with/without glasses in a tennis stadium.

(a) Bird-view of the tennis stadium.

(b) Side view of the tennis stadium.

Fig. 5. Capture setting.

The ratio of both conditions were half and half. All participants signed an agreement form that allows us to use the images capturing them for research purpose. Each participant sat on one of the prede ned spectator seats and was requested to focus on the same target on the tennis court ($Z = 0$) following instructions from a facilitator. These instructions did not include any restriction except for the focus target. All participants focused on the target in a natural way. Figure 5 shows the capture setting. We de ned the center of the tennis court as the origin of the world coordinates, the Y-axis directing to the spectator seats, and the Z-axis directing upwards from the origin, in a right-handed coordinate system.

Figure 6 shows the positions of the gaze targets, and Figure 7 shows the series of identi cation numbers assigned to the targets. The spectator seats are located toward the upper side of the gures. We de ned 22 targets on the tennis

was used for evaluation, alternately. There was no overlap in the participants and they were cross-validated.

In the training phase, all pairs of detected face images and their gaze angles in the training data were used. All pairs of mirrored face images and their gaze angles were also used for training. Specifically, 552,596 pairs for camera 1 and 556,882 pairs for camera 2 were used. Considering that if the people are sufficiently far from the cameras, the gaze angle of every person in each captured image can be approximated as the same, since in our dataset, the distances between participants and cameras were quite far compared with the distance between participants, we trained only one estimator to estimate the gaze angle of all the participants.

As the gaze angle estimator, we used LeNet-5 [21], which was originally designed for the classification of very low-resolution images. To change the task from classification to gaze angle estimation, the number of units in the output layer was modified to two units corresponding to $\boldsymbol{a}_j = (a_{jy}, a_{jp})^T$ and also the activation layer of the output was changed to hyperbolic tangent. The gaze angles were normalized to the range [0, 1].

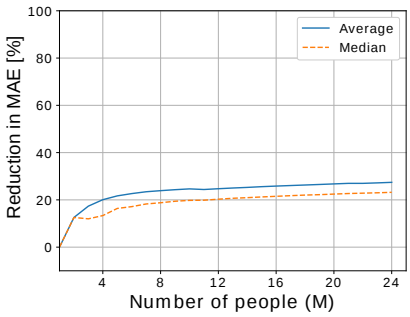
In the evaluation phase, all images capturing a crowd of people in the evaluation data were used. From each evaluation image, face images of many people are detected, and the trained gaze angle estimator outputs the gaze angle for each face image. The position of each participant's seat can be calculated from the capture settings. With an assumption that each person's face is positioned 75 cm above the seat level, the face position $\boldsymbol{\kappa}_j$ can also be calculated in the world coordinate system. Therefore, gaze points of all the face images are estimated based on the geometry in Figure 4, and are aggregated by the proposed method.

We analyzed the relationship between the number of people involved in the aggregation and the localization error. Concretely, we aggregated the estimated gaze points while increasing the number of face images from 1 to 24. If more than 24 face images were detected, those with higher scores in the face detection were preferentially selected. We evaluated the method by the Mean Absolute Error (MAE), which is calculated as the distance between the estimated gaze point of the crowd and the position of the ground-truth target.

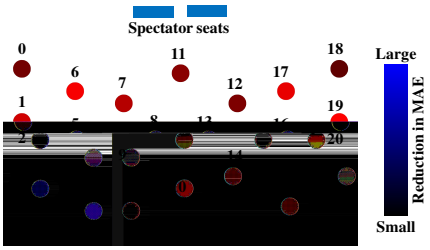
In the evaluation, we chose the result without aggregation (equivalent to the case when the number of people involved in the aggregation is only one) as the baseline. As a comparison method, we chose an aggregation method using median instead of average as in Eq.(7). By assuming that $p(\boldsymbol{g}|d_j;\hat{\Theta})$ follows the Laplace distribution, the solution of Eq.(2) can be calculated by the median of the estimated gaze points $\boldsymbol{g}_j$ for the detected face images d_j as similar to Eq.(6),

$$\hat{\boldsymbol{g}} = \underset{j=1,\ldots,M}{\text{median}}\, \boldsymbol{g}_j = \acute{\boldsymbol{g}}, \tag{8}$$

where the median is calculated as vector median [22]. As a result, a common gaze point of the crowd $\acute{\boldsymbol{g}} = (\acute{g}_x, \acute{g}_y, Z)^T$ is output.

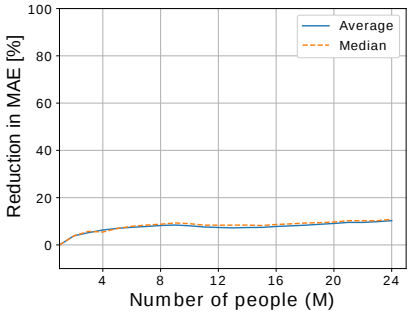


(a) Transition of the reduction in MAE.

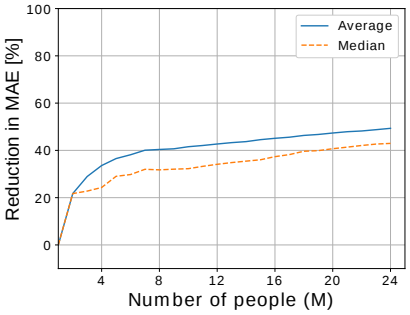


(b) Reduction in MAE for each target ($M = 24$).

Fig. 10. Results of the proposed aggregation method.



(a) Results of the worst case.



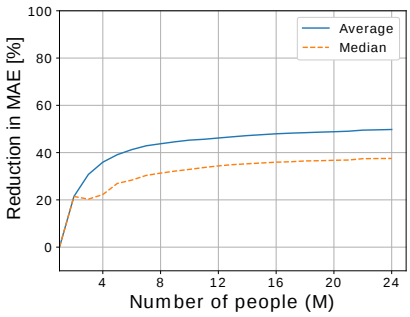
(b) Results of the best case.

Fig. 11. Results of the worst and the best reduction in MAE.

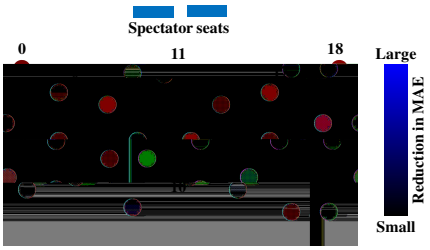
4.4 Results

Figure 10 (a) shows the transition of reduction in MAE while increasing the number of people involved in the aggregation. The score shows the reduction rate from the case where the number of people is one (baseline). The results reveal that both aggregation methods based on Eq.(7) and Eq.(8) improved the gaze localization accuracy by increasing the number of people, but the proposed method showed a larger reduction in MAE than the comparison method. The reduction in MAE from the baseline was 25.73 % by aggregating the estimation results from 24 people. In particular, while the MAE was 13.99 m for the baseline, it decreased to 10.39 m by aggregating the estimation results from 24 people.

Figure 10 (b) visualizes the reduction in MAE per target by aggregating the estimation results from 24 people. The color of each target indicates the reduction in MAE. Note that before the visualization, the reduction in MAE were normalized to the range [0, 1]. The worst reduction in MAE was obtained for target 21. Even so, the reduction in MAE reached 10.27 % by aggregating the estimation results from 24 people. Moreover, the best reduction in MAE



(a) Transition of the reduction in MAE.



(b) Reduction in MAE for each target ($M = 24$).

Fig. 12. Results along the X-axis.

which was obtained for target 19 reached as high as 49.33 % from the baseline. Figure 11 shows the results which showed the worst and the best reduction in MAE.

5 Discussions

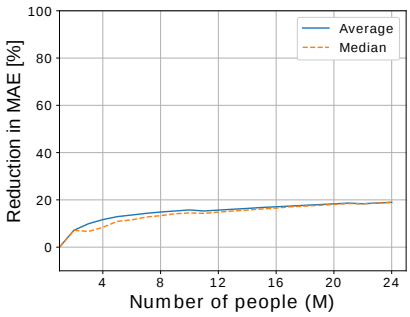
Figure 10 revealed that the gaze localization accuracy was improved by increasing the number of people involved in the aggregation. In this section, we confirm the details of the results.

5.1 Number of People Involved in the Aggregation

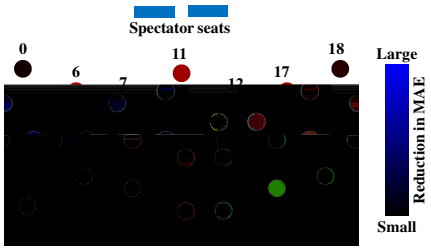
Although Figure 10 shows that the reduction in MAE increases in proportion to the number of people involved in the aggregation, the gain saturates at around 10 people. We consider that this is caused by the distribution of the gaze estimation results; although the proposed method expects that the estimation results isotropically distribute around the ground truth, this may not be true. In such a case, the aggregated results will not approach the ground truth but rather approach a biased center of the distribution. We will need to improve the aggregation method considering this problem in order to further improve the results.

5.2 Results along Each Axis

The reduction in MAE for each target showed different behaviors along each axis. Figures 12 (a) and 13 (a) show the results along the X-axis and the Y-axis of the world coordinates, respectively. The reduction in MAE for the X-axis and the Y-axis were 49.78 % and 18.95% by aggregating the estimation results from 24 people, respectively. We consider that the difference was caused by the significant difference of the ranges of the gaze angle; while the range of the yaw angle (corresponds to the estimation accuracy of the X-axis) was $[-74.02, 74.02]$,



(a) Transition of the reduction in MAE



(b) Reduction in MAE for each target ($M = 24$).

Fig. 13. Results along the Y-axis.

the range of the pitch angle (corresponds to the estimation accuracy of the Y-axis) was only between $[-20.09, -3.01]$, the latter being eight times smaller than the former. The appearance of the faces along the Y-axis changes little and the estimation is more difficult than that along the X-axis. Therefore, the estimation along the Y-axis had larger bias than that along the X-axis.

Figures 12 (b) and 13 (b) visualize the reduction in MAE for each target along the X-axis and the Y-axis by aggregating the estimation results from 24 people, respectively. These figures show that the reduction in MAE of the farther targets behave differently along the X-axis and the Y-axis; along the Y-axis, the reduction in MAE of the farther targets were low, while along the X-axis, the proposed method also improved that of the farther targets. We consider that this difference occurred since the proposed method is robust to the variance of gaze estimation results but not to the bias; along the Y-axis, the bias of the estimation results was too large to approach the ground truth, while along the X-axis, the proposed method could approach the ground truth.

In summary, if the estimation results have a too large bias, the proposed method could hardly approach the ground truth, since the proposed aggregation method did not have sufficient robustness to deal with it. An aggregation method with a higher robustness to a large bias should increase the reduction in MAE.

6 Conclusion

In this paper, we proposed a novel method for localizing a common gaze point focused on by a crowd of people. We also constructed a dataset including images capturing many people looking at a target. An evaluation using the dataset showed that the proposed method could improve the localization accuracy by aggregating the gaze estimation results from a crowd of people.

In future work, we plan to propose an aggregation method more robust to the estimation results with larger bias by employing a machine learning framework, and also deal with situations where multiple gaze targets exist. We also plan to extend the method to localize the gaze targets in the three-dimensional space.

References

1. Okamoto, K., Utsumi, A., Yamazoe, H., Miyashita, T., Abe, S., Takahashi, K., Hagita, N.: Classification of pedestrian behavior in a shopping mall based on LRF and camera observations. Proc. 12th IAPR Conf. on Machine Vision Applications (2011) 1–5
2. Fridman, L., Langhans, P., Lee, J., Reimer, B.: Driver gaze region estimation without use of eye movement. IEEE Intell. Syst. **31** (2016) 49–56
3. Yonetani, R., Kawashima, H., Hirayama, T., Matsuyama, T.: Gaze probing: Event-based estimation of objects being focused on. In: Proc. 20th IAPR Int. Conf. on Pattern Recognition. (2010) 101–104
4. Hirayama, T., Sumi, Y., Kawahara, T., Matsuyama, T.: Info-concierge: Proactive multi-modal interaction through mind probing. In: Proc. 3rd Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (2011) 1–10
5. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: Proc. 28th Int. Conf. on Neural Information Processing Systems. (2015) 199–207
6. Park, H.S., Jain, E., Sheikh, Y.: 3D gaze concurrences from head-mounted cameras. In: Proc. 25th Int. Conf. on Neural Information Processing Systems. (2012) 422–430
7. Park, H.S., Shi, J.: Social saliency prediction. In: Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition. (2015) 4777–4785
8. Sugano, Y., Zhang, X., Bulling, A.: Aggregaze: Collective estimation of audience attention on public displays. Proc. 29th ACM Symposium on User Interface Software and Technology (2016) 821–831
9. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops. (2017) 2299–2308
10. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition. (2016) 2176–2184
11. Ono, Y., Okabe, T., Sato, Y.: Gaze estimation from low resolution images. In: Proc. 1st Pacific Rim Conf. on Advances in Image and Video Technology. (2006) 178–188
12. Vasilescu, M.A.O., Terzopoulos, D.: Multilinear analysis of image ensembles: Tensorfaces. In: Proc. 7th European Conf. on Computer Vision. Volume 1. (2002) 447–460
13. Tawari, A., Mgelmose, A., Martin, S., Moeslund, T.B., Trivedi, M.M.: Attention estimation by simultaneous analysis of viewer and view. Proc. 17th IEEE Int. Conf. on Intelligent Transportation Systems (2014) 1381–1387
14. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building Rome in a day. Comm. ACM **54** (2011) 105–112
15. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23** (2016) 1499–1503
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. 3rd Int. Conf. for Learning Representations. (2015) 1–15
17. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition. (2015) 4511–4520

18. Ariz, M., Bengoechea, J., Villanueva, A., Cabeza, R.: A novel 2D/3D database with automatic face annotation for head tracking and pose estimation. Comput. Vis. Image Underst. **148** (2016) 201–210
19. Hong, X., Zhao, G., He, Q., Chai, X., Holappa, J., Chen, X., Pietikinen, M.: The Oulu multi-pose eye gaze (OMEG) dataset. Proc. 19th Scandinavian Conf. on Image Analysis (2015) 418–427
20. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap database: Data description and gaze tracking evaluation benchmarks. Proc. 2014 Symposium on Eye Tracking Research and Applications (2014) 255–258
21. Le Cun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86** (1998) 2278–2324
22. Astola, J., Haavisto, P., Neuvo, Y.: Vector median filters. Proc. IEEE **78** (1990) 678–689