Design of Gaze Estimation Model Based on Multiple Feature Fusion

Zhonghe Ren[1], Fengzhou Fang[1,2,#] and Rui Niu[1]

1 State Key Laboratory of Precision Measuring Technology and Instruments, Laboratory of Micro/Nano Manufacturing Technology (MNMT), Tianjin University, Tianjin 300072, China
2 Centre of Micro/Nano Manufacturing Technology (MNMT-Dublin), University College Dublin, Dublin 4, Ireland
# Corresponding Author / Email: fzfang@tju.edu.cn, TEL: +86-022-27407503

KEYWORDS: Gaze estimation, Artificial intelligence, Human–computer interaction, Metrology

Gaze estimation is a fundamental task in many applications of cognitive sciences, human–computer interaction, and robotics. In this study, a multiple-feature fusion method based on the gaze conduction mechanism is proposed to improve the performance of the gaze estimation model. Accurate pupil localization is a crucial requirement in gaze estimation. A novel multi-directional pixel array computation method is proposed, which utilizes the gray-scale features of eye morphology to calculate pupil localization intelligently. Then, a feature element matrix is established that includes the original images, region images, pupil coordinates, and head pose vectors. According to the gaze conduction mechanism and the feature element matrix, gaze estimation models with different input modes and network structures are designed, and model training and test experiments are carried out on a large-scale dataset. Experimental results show that optimizing the feature combination and fine-tuning the computational architecture can improve the performance of the gaze estimation model, which would enable the reduction of the model by incorporating the critical features and thus improve the performance and accessibility of the method. The average error is 1.63 cm on the GazeCapture dataset, which achieves comparable accuracy with state-of-the-art methods.

1. Introduction

Gaze is a non-verbal cue with many functions [1]. Gaze estimation uses mechanical, electronic, optical, and other detection methods to obtain the gaze information for indicating human attentiveness. It has recently received significant interest in computer vision due to its significance to many arising computing paradigms ranging from scientific research to commercial applications, such as cognitive sciences [2], human-computer interaction [3], driver attention detection [4], virtual reality [5] and robotics [6].

Appearance-based gaze estimation with a data-driven artificial intelligence model has significant application value in the era of big data and computing power. The current artificial intelligence models mainly employ deep learning algorithms to transform unstructured input data into valuable structured output information after training on large-scale datasets [7]. However, the purely data-driven methods built by deep learning techniques may suffer from a lack of interpretability, which prevents their applicability to critical scenarios [8]. For the next-generation artificial intelligence paradigm, researchers are reaching a consensus that it is imperative to combine symbolism and connectionism to establish an eXplainable Artificial Intelligence (XAI) theory and method to make data-driven models human-interpretable and trustworthy, which presents an opportunity for further advancing gaze estimation research [8][9]. Therefore, it is feasible and valuable exploratory research to optimize the gaze estimation model by the integrating strategy of data and knowledge to improve the performance of gaze estimation.

In this study, a multiple-feature fusion method is proposed to improve the performance of the gaze estimation model. Based on the constructed feature element matrix with multiple meta-features, a gaze conduction mechanism is proposed to guide the design of meta-feature combination modes for model input. Furthermore, the training and test experiments are carried out to verify the performance of the designed gaze estimation models.

2. Feature Element Matrix

Appearance-based gaze estimation methods typically treat the feature region images extracted from the original image captured by a camera on the human-computer interaction device as a high-dimensional vector and learn a regression mapping model from such vector to the gaze information through labeled training data. However, the original images captured by the cameras in unconstrained scenarios may contain many background elements and interference information. If the original images are directly input to the gaze estimation mapping model, large-scale labeled training samples and a large amount of training time are usually required, which restricts the model accuracy and is not conducive to model deployment. Therefore, to improve the training efficiency and test accuracy of the gaze estimation model, as well as optimize the model deployment, a more robust and interpretable computational paradigm should be studied: first, preprocess the original images for feature extraction and expression based on human knowledge, and then input the meticulously combined features into the deep learning-based model for training.

A feature element matrix with multiple meta-features is proposed. Table 1 presents the composition and characteristics of the proposed feature element matrix. The data source is an original image captured by a camera on the human-computer interaction device. In the feature element matrix, a series of meta-features are considered fundamental features and are divided into three categories.

Table 1 Composition and characteristics of the proposed feature element matrix

Data source	Original image captured by a camera on the human-computer interaction device									
Category	Feature region images			Key marker points				High-level features		
Meta-features	Face	Eyes		Facial region corner points	Facial key landmarks	Eye periocular points		Eye pupil localization		Head pose estimation
Nomenclature	FRI	ERI-L	ERI-R	FRCP	FKL	EPP-L	EPP-R	EPL-L	EPL-R	HPE
How to obtain	Face detection	Eye detection		Face detection	Dlib toolkit			The proposed MPAC method		Perspective-n-Point
Expression format (size)	Image			Array						
				8	112	12	12	2	2	3

Category I refers to the feature region images, including the face region image (FRI) and eye region images (left eye region images, ERI-L; right eye region images, ERI-R).

Category II refers to the key marker points, including the facial region corner points (FRCP), facial key landmarks (FKL) and eye periocular points (left eye periocular points, EPP-L; right eye periocular points, EPP-R).

Category III refers to the high-level features, including the eye pupil localization (left eye pupil localization, EPL-L; right eye pupil localization, EPL-R) and head pose estimation (HPE).

FRI, ERI-L, and ERI-R are expressed in image format, while FRCP, FKL, EPP-L, EPP-R, EPL-L, EPL-R, and HPE are expressed in array format. The adopted conventional methods for obtaining the proposed meta-features include face detection, eye detection, Dlib toolkit, and Perspective-n-Point (PnP) [10].

Accurate pupil localization is a crucial requirement in gaze estimation. To perform robust and accurate pupil localization in near real-time during image preprocessing, a novel pupil localization method based on a multi-directional pixel array computation (MPAC) strategy is proposed, which utilizes the gray-scale features of eye morphology to calculate pupil localization intelligently. Figure 1 presents the schematic diagram of the proposed multi-directional pixel array computation method for eye pupil localization. To get target pixel landmarks, a series of searches in four directions need to be executed, as shown in Fig. 1. Each directional search is to find the localization of the minimum pixel value that first appears in the region containing one or more rows of pixels. Finally, the geometric center coordinates of all pixel landmarks are calculated as pupil coordinates, as shown in Equation (1):

$$(x,y)=\left(\frac{\sum_{n=1}^{N}x_n}{N},\frac{\sum_{n=1}^{N}y_n}{N}\right), \tag{1}$$

where N is the total number of pixel landmarks, and (x_n, y_n) are the coordinates of pixel landmarks. The calculated results (x, y) correspond to the pupil coordinates.

BioID benchmark dataset with pupil localization labels is considered complex and realistic because some samples were captured with poor-quality illumination. BioID can be used to evaluate the accuracy, robustness, and real-time performance of localization methods [11]. To assess the proposed MPAC method, experiments on the BioID dataset were implemented. The results show that the proposed pupil localization method can achieve state-of-the-art efficiency and accuracy, where 99.74% of the samples correspond to normalized error ≤ 0.20.

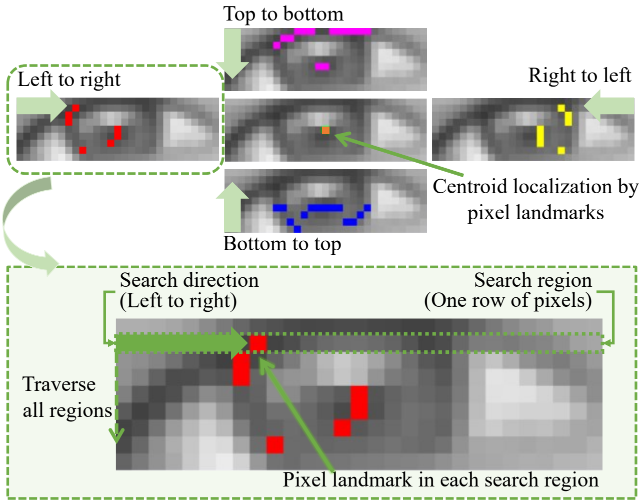
Top to bottom
Left to right
Right to left
Centroid localization by pixel landmarks
Bottom to top
Search direction (Left to right)
Search region (One row of pixels)
Traverse all regions
Pixel landmark in each search region

Fig. 1 Schematic diagram of the proposed multi-directional pixel array computation method

3. Gaze Conduction Mechanism

The meta-features in the proposed feature element matrix can be used as input data for the deep learning model for gaze estimation. Although each meta-feature may play different roles in the model, some may have inclusion or complementary relationships with other meta-features. The combination modes of meta-features should follow the motion conduction mechanism of human eye gazing behavior rather than arbitrary or random. Hence, a gaze conduction mechanism (GCM) is proposed to guide the design of meta-feature combination modes. Figure 2 presents the schematic diagram of the proposed gaze conduction mechanism. In an appearance-based gaze estimation scenario with free head motion, eyeball movement and head movement are the two main factors defining gaze direction. The coupling between eyeball movement and head movement is also a critical factor in mapping gaze direction.

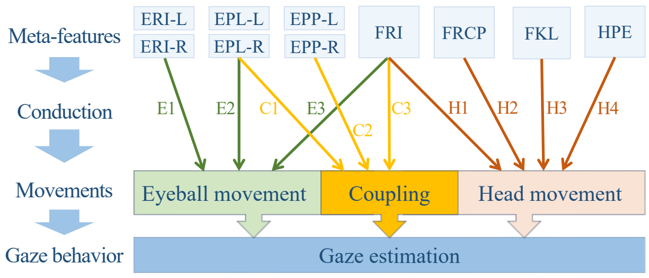

Fig. 2 Schematic diagram of the proposed gaze conduction mechanism

The main meta-features that characterize eyeball movement include eye region images and eye pupil localization, as marked with E1 and E2 in Fig. 2, respectively. Since the face region images also contain eye regions, the corresponding meta-feature can also be employed to characterize eye movement, as marked with E3 in Fig. 2. The coupling between eyeball movement and head movement can infer the location and orientation of the eyes relative to the head in the same frame. Eye pupil localization and eye periocular points are the primary meta-features to characterize the coupling, as marked with C1 and C2 in Fig. 2, respectively. Similarly, the face region images can also be employed to characterize the coupling, as marked with C3 in Fig. 2. The meta-features that characterize head movement, which infers the head pose relative to the camera, include the face region image, facial region corner points, facial key landmarks, and head pose estimation, as marked with H1, H2, H3, and H4 in Fig. 2, respectively.

Complying with the GCM, the combination modes of meta-features should be capable of comprehensively characterizing the features of eye movement, head movement, and coupling. Thus, the proposed GCM can serve as a scheduler for selecting and combining meta-features from the proposed feature element matrix to guide the design of various input modes in the gaze estimation model.

4. Gaze Estimation Model with Multiple-Feature Fusion

4.1 Model Design

A multiple-feature fusion method based on the gaze conduction mechanism is proposed to improve the performance of the gaze estimation model. The proposed gaze estimation model with regulatable input modes and network architectures consists of three modules, as shown in Fig. 3.

Module 1 is the input control module, which is used to selectively call input data from the proposed feature element matrix under the guidance of GCM. Thus, the model can couple multiple meta-features and change the input mode to test various combinations for exploring a relatively optimal combination of meta-features.

Module 2 is the feature extraction module, which includes data preprocessing, CNN, and fully connected (FC) layers. Since the input data includes images and arrays, the module is designed with multiple channels for array-visual bimodal feature extraction based on multimodal deep-learning with only a one-stage training phase.

Module 3 is the feature fusion module, which is mainly composed of fully connected layers. The feature fusion module is employed to fuse data transmitted from multiple processing channels of the feature extraction module. Finally, the model output the predicted gaze point coordinates.

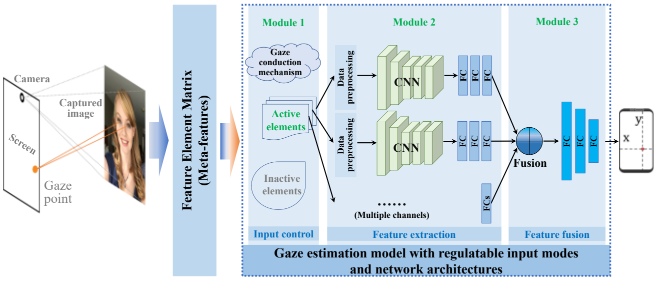

Fig. 3 Architectural design of gaze estimation model

4.2 Experiment

The proposed feature element matrix contains multiple meta-features, which can generate various combinations for model input through the input control module. In addition, optimizing the feature extraction networks is also an effective way to improve model performance. Consequently, several input modes and network structures are designed, and a series of grouping experiments are carried out on a small-scale sub-dataset from GazeCapture. Then, the optimized input modes and network structures are selected for training and testing on the large-scale GazeCapture dataset to verify and compare model performance.

GEM-Base is the original iTracker model [12]. GEM-Lite is the representative model that optimizes meta-feature combination and computational design without inputting face images in training. GEM-Full can achieve the best accuracy in the grouping experiment, but drawing face images into the input may make the model not light enough. Table 2 presents the meta-feature configuration in the input modes and corresponding FC layer dimension for feature fusion. Table 3 presents the comparison of test error and test multi-frame-error. Compared with GEM-Base, the test error and multi-frame-error of GEM-Lite are reduced by 20.25% and 20.53% respectively, while the test error and multi-frame-error of GEM-Full are reduced by 26.15% and 25.37% respectively.

Table 2 Meta-feature configuration in the input modes and corresponding FC layer dimension for feature fusion

Model	FRI	FRCP	ERI-L	ERI-R	FKL	EPP-L	EPP-R	EPL-L	EPL-R	HPE
GEM-Base	64	32	64	64	/	/	/	/	/	/
GEM-Lite	/	32	64	64	64	32	32	24	24	24
GEM-Full	64	32	64	64	64	32	32	24	24	24

Table 3 Comparison of test error and test multi-frame-error

Model	Test error	Test error in X direction	Test error in Y direction	Test multi-frame-error
GEM-Base	2.2201	1.278	1.514	2.0328
GEM-Lite	1.7546	0.980	1.223	1.6155
GEM-Full	1.6292	0.880	1.163	1.5170

Table 4 summarizes the performance of the state-of-the-art models and their improvement compared to iTracker as a baseline on the GazeCapture dataset. The results show that the gaze estimation model optimized with the proposed feature fusion method can achieve comparable accuracy as state-of-the-art methods on GazeCapture.

Table 4 Comparison with the state-of-the-art methods

Method	Error value (cm)	Error decrease ratio
iTracker [12]	2.04	Baseline
iTracker with augmentation [12]	1.86	8.82%
SAGE [13]	1.78	12.75%
SD [14][15]	1.81	11.27%
TAT [14]	1.77	13.24%
AFF-Net [16]	1.62	20.59%
GazeAttentionNet [17]	1.67	18.14%
GEM-Lite (Ours)	1.75	14.22%
GEM-Full (Ours)	1.63	20.10%

5. Conclusions

A multiple-feature fusion method is proposed to improve the performance of the gaze estimation model in this study. Based on the constructed feature element matrix with multiple meta-features, a gaze conduction mechanism is proposed to guide the design of meta-feature combination modes for model input. The proposed gaze estimation model with regulatable input modes and network architectures consists of an input control module, a feature extraction module, and a feature fusion module. Experimental results show that optimizing the feature combination and fine-tuning the computational architecture can improve the performance of the gaze estimation model, which would enable the reduction of the model by incorporating the critical features and thus improve the performance and accessibility of the method. The average error is 1.63 cm on the GazeCapture dataset, which achieves comparable accuracy with state-of-the-art methods. Furthermore, the data processing workload for edge computing and cloud computing can improve the practicability of the gaze estimation model in practical scenarios.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (NSFC) (No. 52035009), and the Tianjin Research Innovation Project for Postgraduate Students (No. 2021YJSB161).

REFERENCES

1. Yu, Y., & Odobez, J. M. (2020). Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7312-7322).
2. Kim, H. I., Kim, J. B., Lee, J. E., Lee, T. Y., & Park, R. H. (2016). Gaze estimation using a webcam for region of interest detection. *Signal, Image and Video Processing, 10*(5), 895-902.
3. Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based human–computer interaction. *Advances in physiological computing* (Springer), 39-65.
4. Kim, E., Ryu, H., Oh, H., & Kang, N. (2022). Safety monitoring system of personal mobility driving using deep learning. *Journal of Computational Design and Engineering, 9*(4), 1397-1409.
5. Chang, E., Kim, H. T., & Yoo, B. (2021). Predicting cybersickness based on user's gaze behaviors in HMD-based virtual reality. *Journal of Computational Design and Engineering, 8*(2), 728-739.
6. Qiu, Q., Zhu, J., Gou, C., & Li, M. (2022). Eye gaze estimation based on stacked hourglass neural network for aircraft helmet aiming. *International Journal of Aerospace Engineering, 2022*, 1-11.
7. Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.
8. Nascita, A., Montieri, A., Aceto, G., Ciuonzo, D., Persico, V., & Pescapé, A. (2021). XAI meets mobile traffic classification: Understanding and improving multimodal deep learning architectures. *IEEE Transactions on Network and Service Management*, *18*(4), 4225-4246.
9. Pan, Y. (2016). Heading toward artificial intelligence 2.0. *Engineering, 2*(4), 409-413.
10. Hsu, H.-W., Wu, T.-Y., Wan, S., Wong, W. H., & Lee, C.-Y. (2018). Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia, 21*(4), 1035-1046.
11. Jesorsky, O., Kirchberg, K. J., & Frischholz, R. W. (2001). Robust face detection using the hausdorff distance. In *International Conference on Audio-and Video-based Biometric Person Authentication* (pp. 90-95).
12. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2176-2184).
13. He, J., Pham, K., Valliappan, N., Xu, P., Roberts, C., Lagun, D., & Navalpakkam, V. (2019). On-device few-shot personalization for real-time gaze estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 1149-1158).
14. Guo, T., Liu, Y., Zhang, H., Liu, X., Kwak, Y., In Yoo, B., Han, J.-J., & Choi, C. (2019). A generalized and robust method towards practical gaze estimation on smart phone. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 1149-1158).
15. Yang, C., Xie, L., Su, C., & Yuille, A. L. (2019). Snapshot distillation: Teacher-student optimization in one generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2854-2863).
16. Bao, Y., Cheng, Y., Liu, Y., & Lu, F. (2021). Adaptive feature fusion network for gaze tracking in mobile tablets. In *25th International Conference on Pattern Recognition (ICPR)* (pp. 9936-9943).
17. Huang, H., Ren, L., Yang, Z., Zhan, Y., Zhang, Q., & Lv, J. (2022), Gazeattentionnet: Gaze estimation with attentions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2435-2439).